Finding High Temperature Superconductors using ML

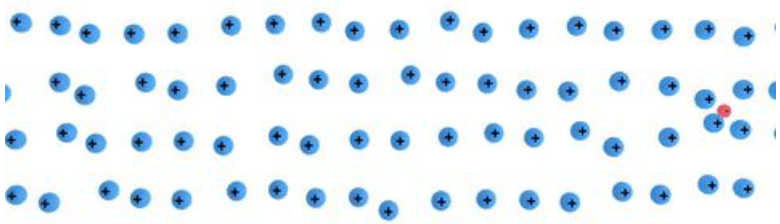
June 10th 2026

Amalie F. Davidsen, Rasmus B. Madsen, Frida B. Nielsen, Silas B. Schack and Hugo Schreckenberg
(Group 5)

Final Project: Applied Machine Learning 2026
NBI University of Copenhagen

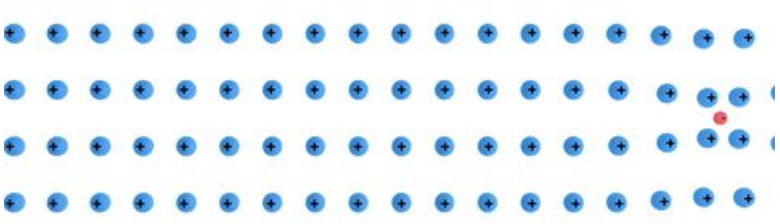
What is a superconductor ?

Normal Metal



$$U = R\,I$$

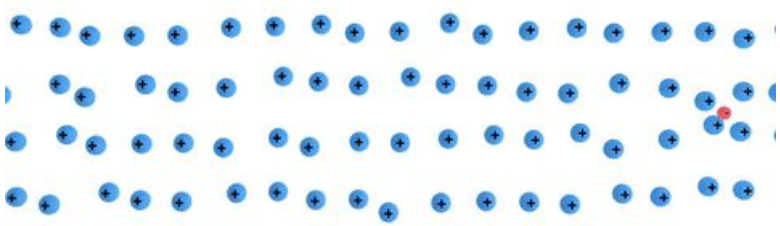
Superconductor



$$R = 0$$

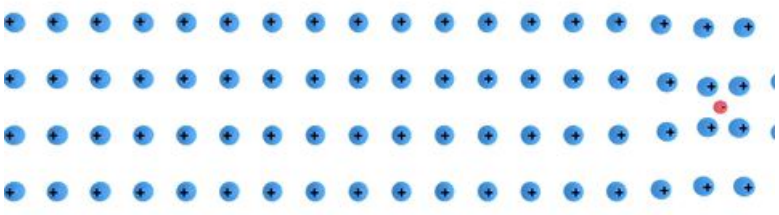
What is a superconductor ?

Normal Metal



$$U = R\,I$$

Superconductor



$$R = 0$$

$$\longrightarrow T \leq T_c \leq -122^{\circ}C$$

Finding a higher temperature superconductor

Cost of the experiments

EUROPEAN SPALLATION SOURCE

https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.skanska.hu/what-we-do/projects/126197/ESS%2C-Lund&ved=0CBoQ3YkBahcKEwjY9o_fntyUAxUAAAAAHQAAAAAQBg&opi=89978449

MAX IV

https://www.google.com/url?sa=t&source=web&rct=j&url=https://en.wikipedia.org/wiki/MAX_IV_Laboratory&ved=0CBoQ3YkBahcKEwiInb_wh9yUAxUAAAAAHQAAAAAQlQ&opi=89978449

Finding a higher temperature superconductor

Cost of the experiments

EUROPEAN SPALLATION SOURCE

https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.skanska.hu/what-we-do/projects/126197/ESS%2C-Lund&ved=0CBoQ9YkBahcKEwjY9o_fhtyUAxUAAAAAHQAAAAAQBg&opi=89978449

MAX IV

https://www.google.com/url?sa=t&source=web&rct=j&url=https://en.wikipedia.org/wiki/MAX_IV_Laboratory&ved=0CBoQ3YkBahcKEwiInb_wh9yUAxUAAAAAHQAAAAAQIQ&opi=89978449

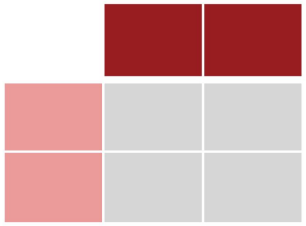
Variety of the materials

$\sim 10^{100}$ estimated materials

Walsh A 2015 The quest for new functionality Nat. Chem.

SuperCon dataset

SuperCon
Superconductors (SC)



→ **149 features**
+ Label (T_c)

↓
13 661 Entries

Statistics about **individual** atoms' properties in the chemical formula

Range atomic number, mean valence electrons…

High-T_c superconductor candidates proposed by machine learning

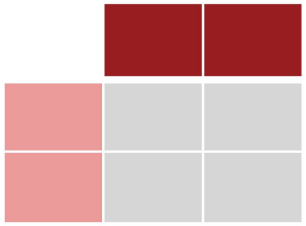
Siwoo Lee[1,2], Jason Hattrick-Simpers[2,3], Young-June Kim[4] and O Anatole von Lilienfeld[1,2,3,4,5,*]

[1] Department of Chemistry, University of Toronto, St. George campus, Toronto, ON, Canada
[2] Acceleration Consortium, University of Toronto, St. George campus, Toronto, ON, Canada
[3] Department of Materials Science & Engineering, University of Toronto, St. George campus, Toronto, ON, Canada
[4] Department of Physics, University of Toronto, St. George campus, Toronto, ON, Canada
[5] Vector Institute for Artificial Intelligence, Toronto, ON, Canada
[*] Author to whom any correspondence should be addressed.

Mach. Learn.: Sci. Technol. 6 (2025) 035052

SuperCon dataset

SuperCon
Superconductors (SC)

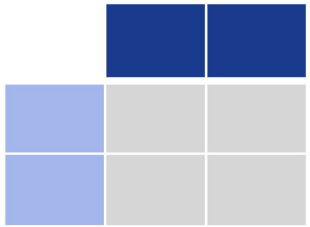


→ **149 features + Label (T_c)**

↓

13 661 Entries

Materials Project
SC + non-SC



→ **149 features**

↓

153 209 Entries

Statistics about **individual** atoms' properties in the chemical formula

Range atomic number,
mean number of valence electrons…

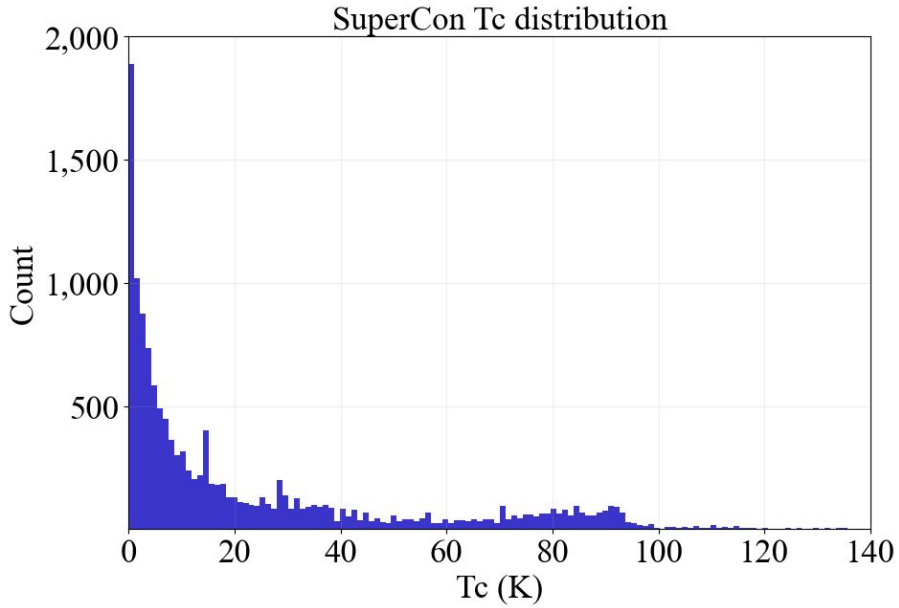
High-T_c superconductor candidates proposed by machine learning

Siwoo Lee[1,2], Jason Hattrick-Simpers[2,3], Young-June Kim[4] and O Anatole von Lilienfeld[1,2,3,4,5,*]

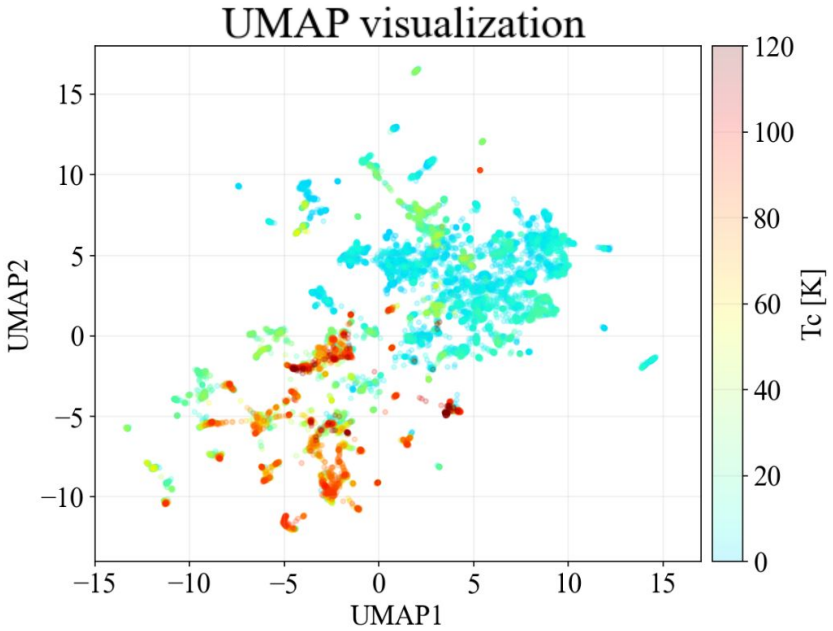
[1] Department of Chemistry, University of Toronto, St. George campus, Toronto, ON, Canada
[2] Acceleration Consortium, University of Toronto, St. George campus, Toronto, ON, Canada
[3] Department of Materials Science & Engineering, University of Toronto, St. George campus, Toronto, ON, Canada
[4] Department of Physics, University of Toronto, St. George campus, Toronto, ON, Canada
[5] Vector Institute for Artificial Intelligence, Toronto, ON, Canada
[*] Author to whom any correspondence should be addressed.

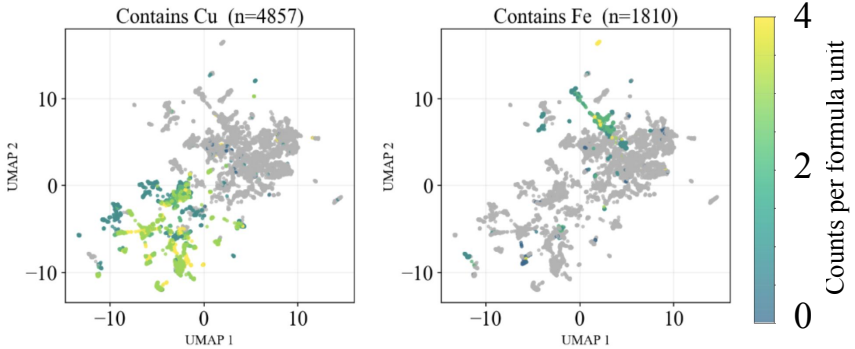
Mach. Learn.: Sci. Technol. 6 (2025) 035052

Visualization of SuperCon data

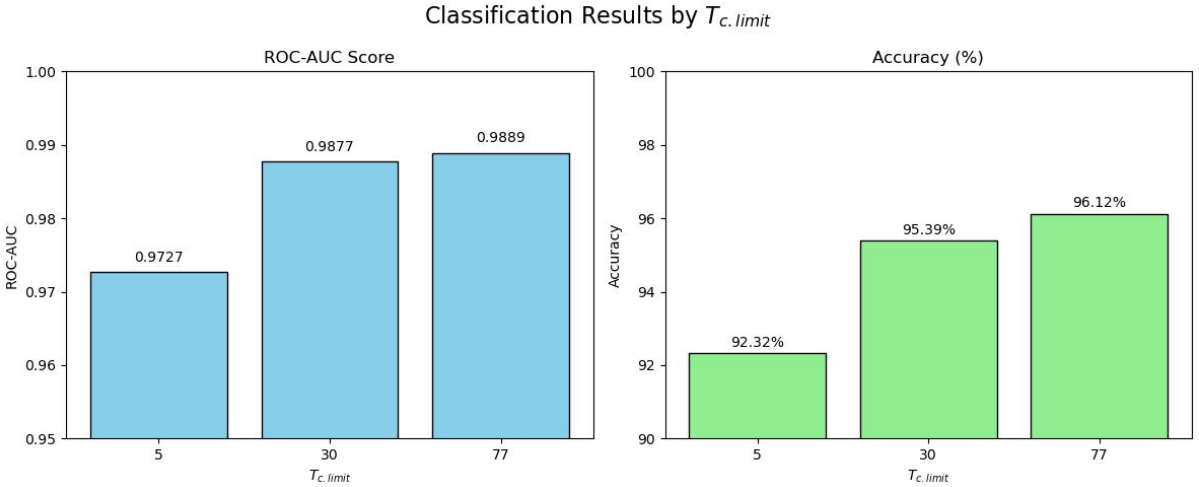


Visualization of SuperCon data

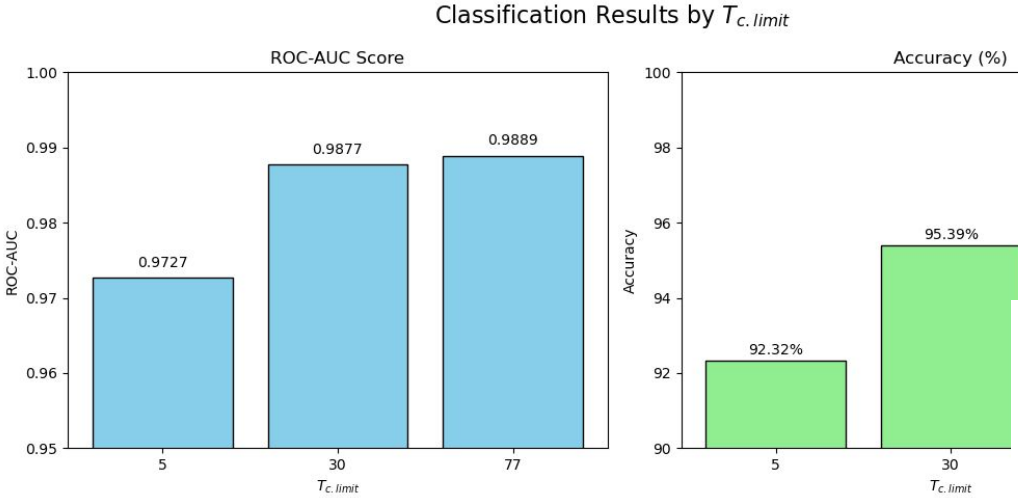


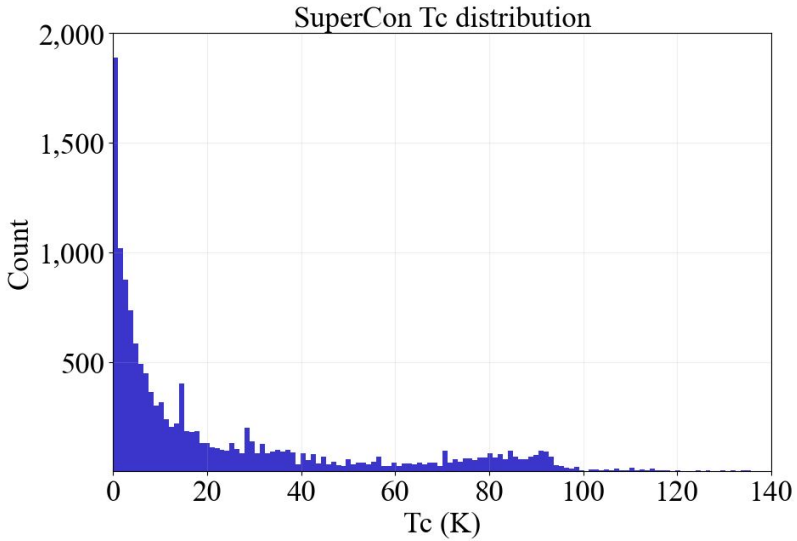


Classification result using XGBoost

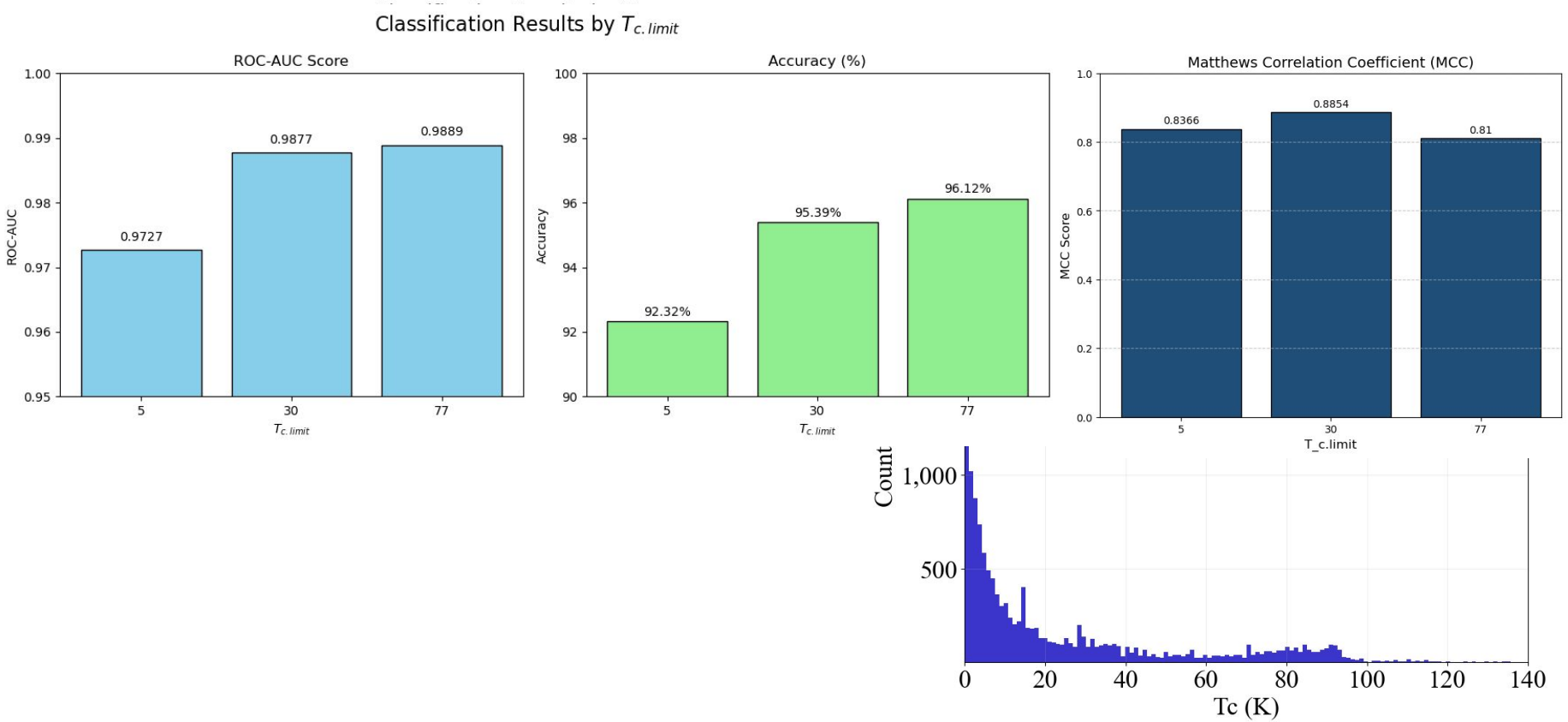


Classification result using XGBoost



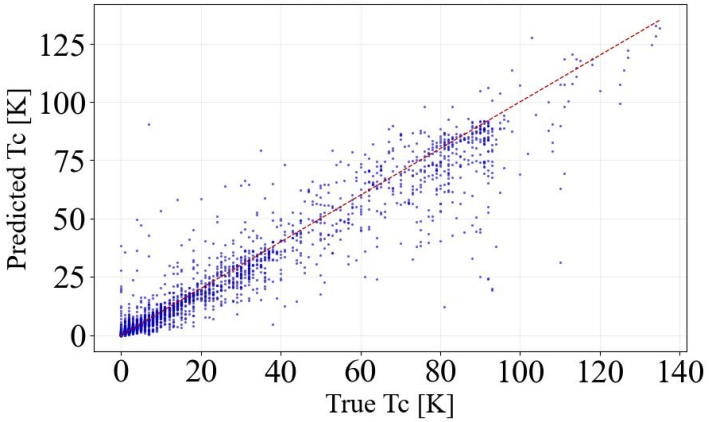


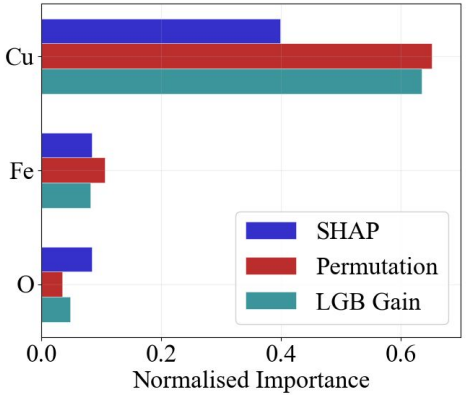
Classification result using XGBoost

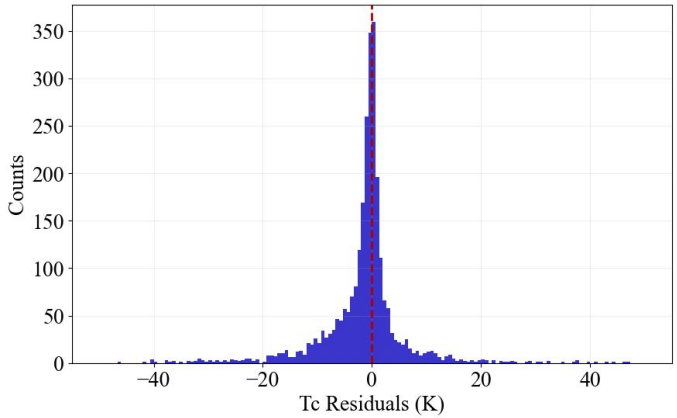


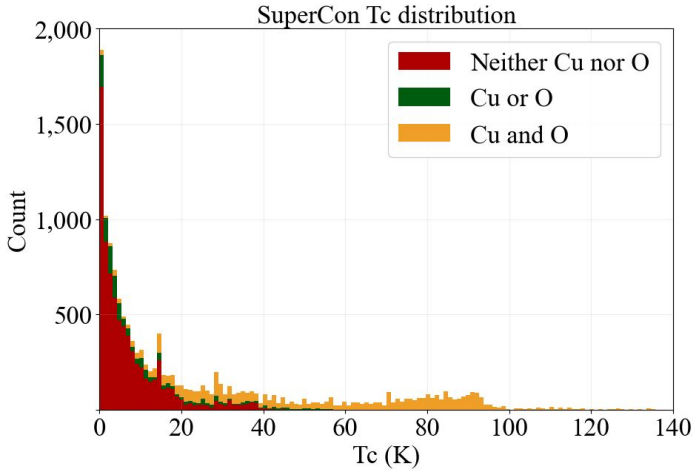
Regression: LGBM

Regression
SuperCon data (120 features)
R^2=0.91, RMSE=8.90 K









Predicting superconductors

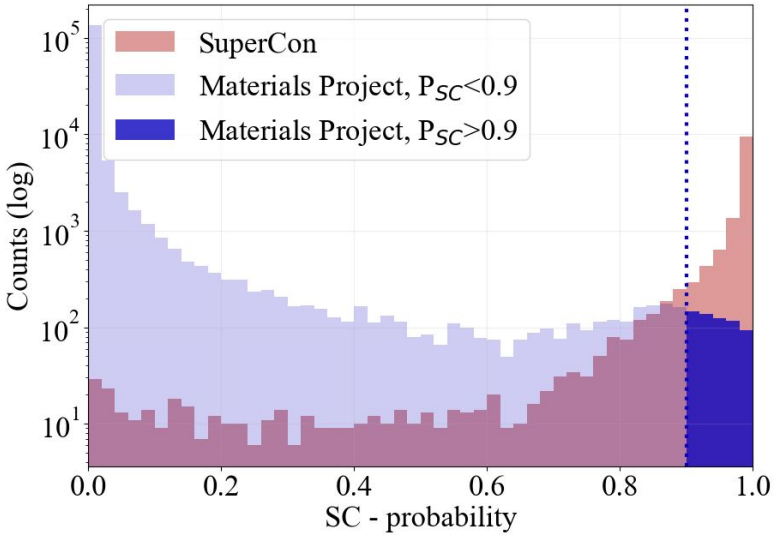
Materials Project —*Assume*→ Non SC
SuperCon → SC

Which 'Non-SC' have the highest probability of being SC?

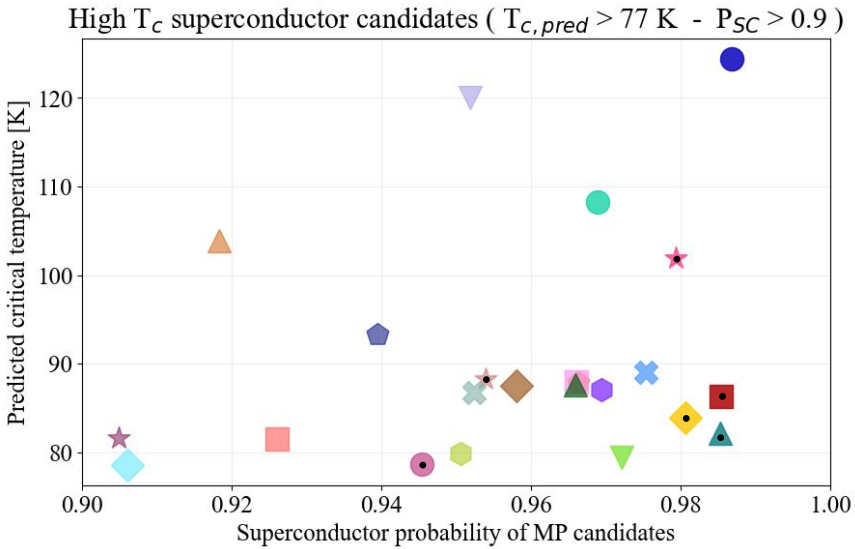
MP SC candidates, highest to lowest SC probability

Ba2Ca2Cu3HgO8	Ba2TmCu3O7
Ba10CaY4(Cu3O7)5	Ba2DyCu3O7
Ba3CaLa2(Cu3O7)2	Ba2SmCu3O7
Ba10GdY4(Cu3O7)5	Ba2Ca3TlCu4O11
Sr2Ca2Cu3BiO8	Ba4Y2Cu7O15
Ba2YCu3O7	Sr4Ca2Cu4Bi4IO16
Ba8Y4(Cu4O9)3	Ba2GdCu3O7
Ba2EuCu3O7	Ba2CaTlCu2O7
Ba2Ca2Tl2Cu3O10	Ba2CaCu2HgO6
Ba2ErCu3O7	Ba2LuCu3O7
Ba2HoCu3O7	Ba8Gd4(Cu4O9)3
Ba2NdCu3O7	

XGB Classification
SuperCon vs Materials Project:
AUC=0.993, Accuracy=97.3%



Only T_c>77 K

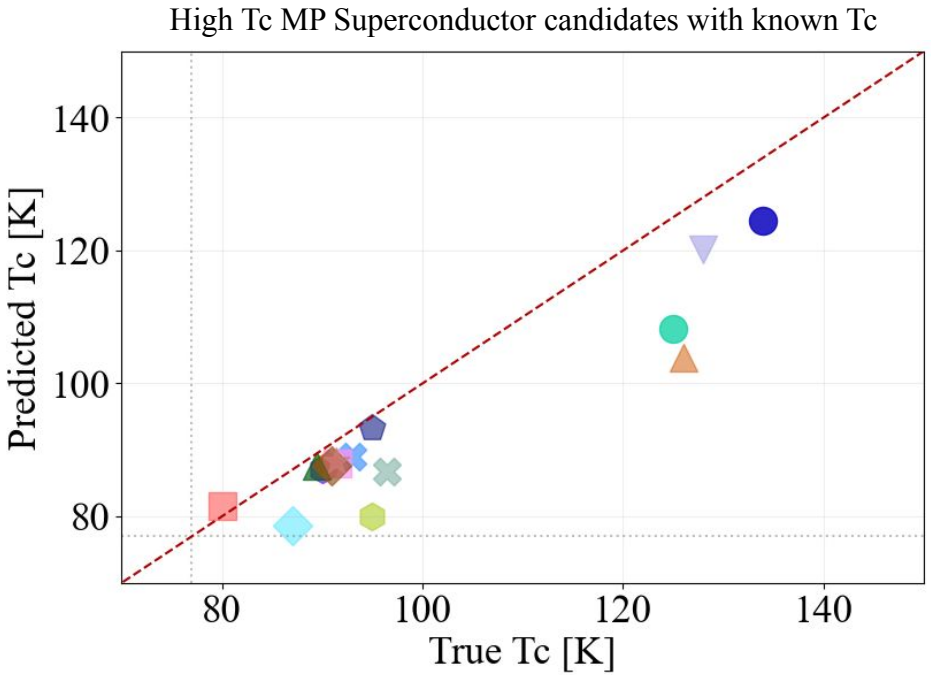


Predicting superconductors

Trained reg. model on MP SC candidates:
R^2=0.69, RMSE=9.3K

MP SC candidates, highest to lowest SC probability

- Ba2Ca2Cu3HgO8
- ? Ba10CaY4(Cu3O7)5
- ? Ba3CaLa2(Cu3O7)2
- ? Ba10GdY4(Cu3O7)5
- ? Sr2Ca2Cu3BiO8
- Ba2YCu3O7
- ? Ba8Y4(Cu4O9)3
- Ba2EuCu3O7
- Ba2Ca2Tl2Cu3O10
- Ba2ErCu3O7
- Ba2HoCu3O7
- Ba2NdCu3O7
- Ba2TmCu3O7
- Ba2DyCu3O7
- Ba2SmCu3O7
- Ba2Ca3TlCu4O11
- Ba4Y2Cu7O15
- ? Sr4Ca2Cu4Bi4IO16
- Ba2GdCu3O7
- Ba2CaTlCu2O7
- Ba2CaCu2HgO6
- Ba2LuCu3O7
- ? Ba8Gd4(Cu4O9)3
- 77 K threshold

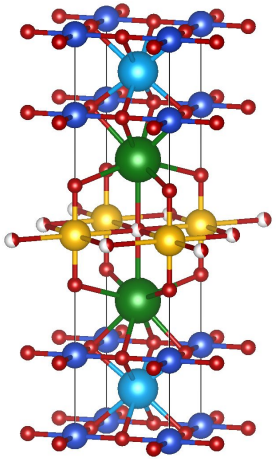


→ 23 High Tc superconductor candidates,
of which 7 (to our knowledge) have not been studied yet

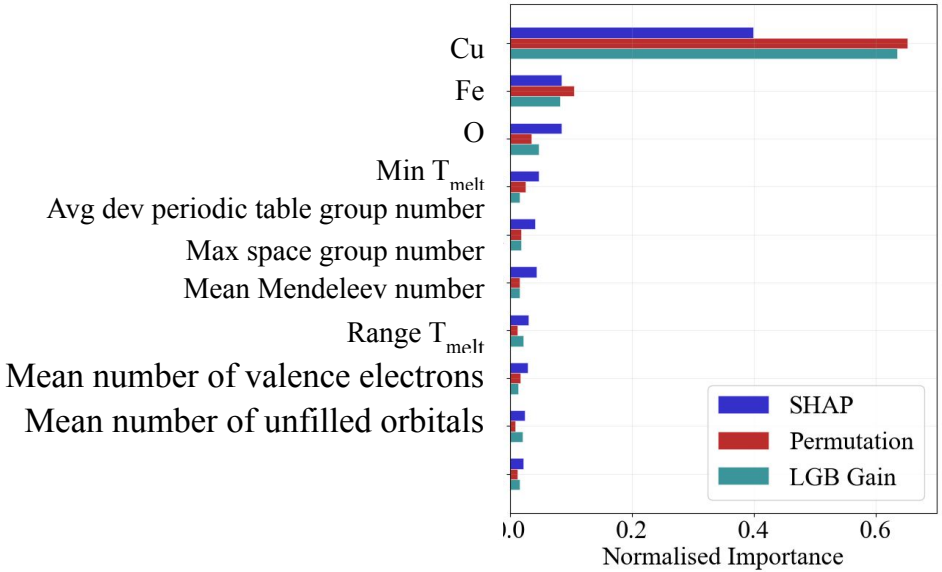
More physically informed predictions

- Pooled features → ignores spatial properties
- Crystal structures → Spatial information

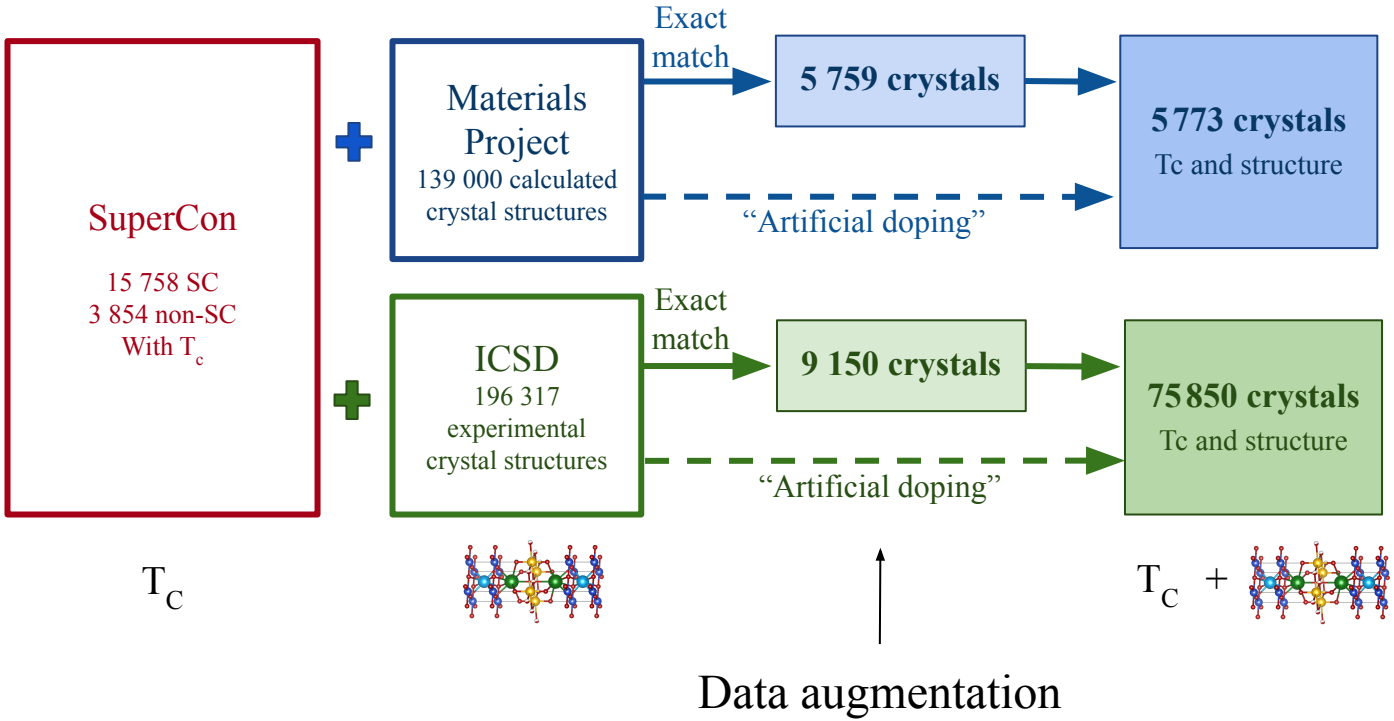
$Ba_2Ca_2Cu_3HgO_{8.4}$ →



Regression model:
Most important features



3DSC: Dataset



Graph creation

1. Define the nodes

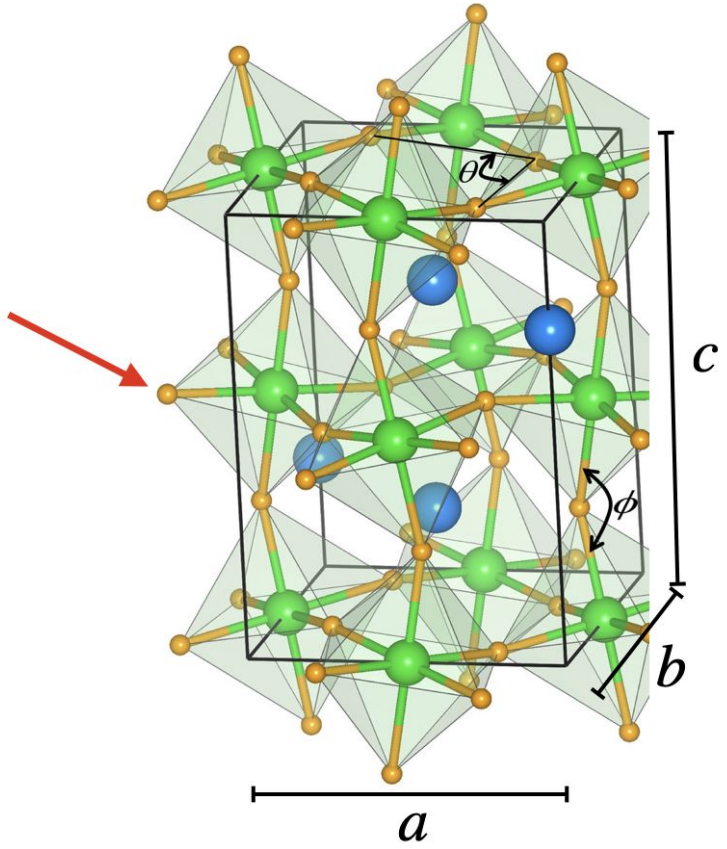
atomic number

atomic mass

occupancy

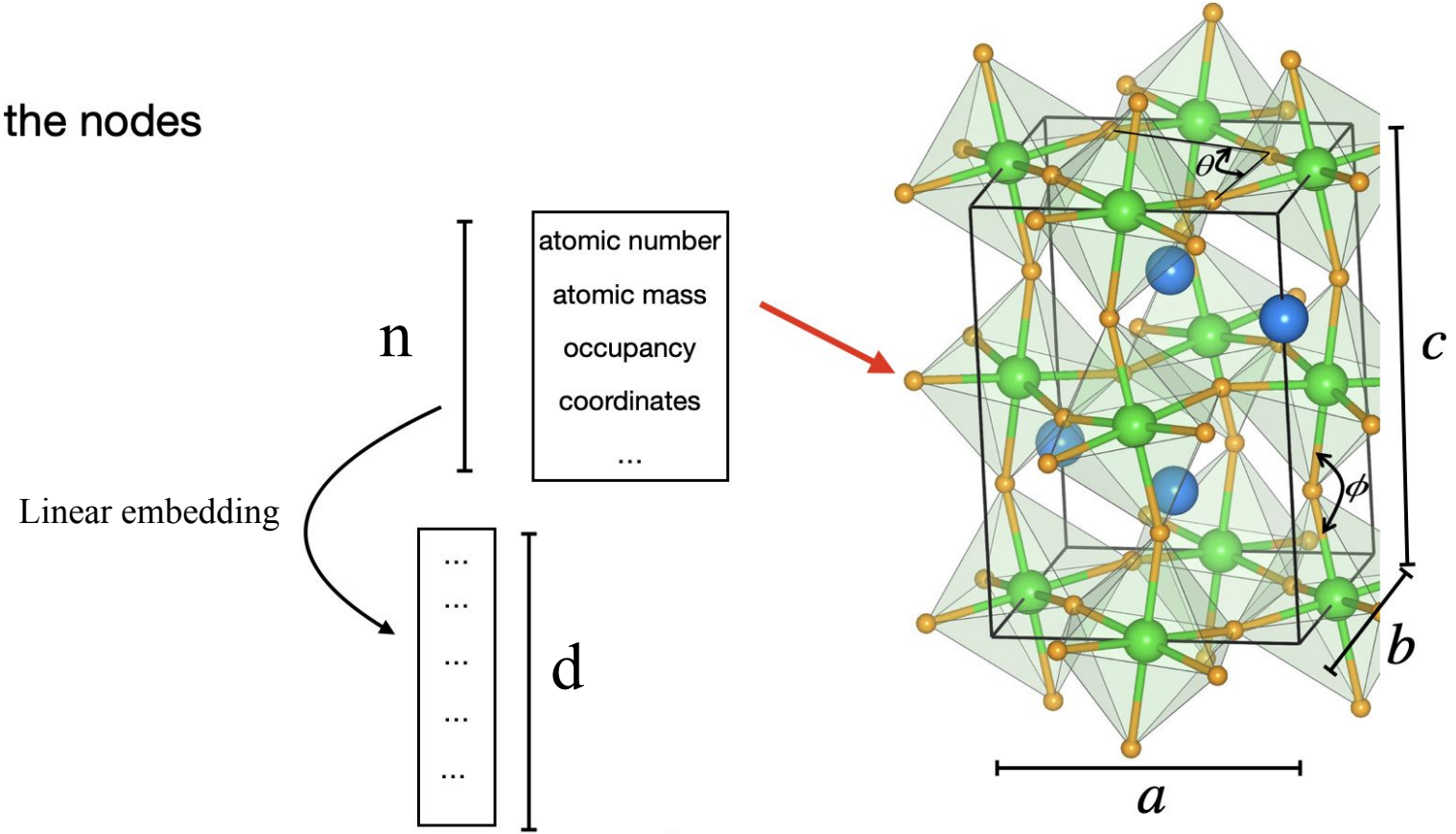
coordinates

...



Graph creation

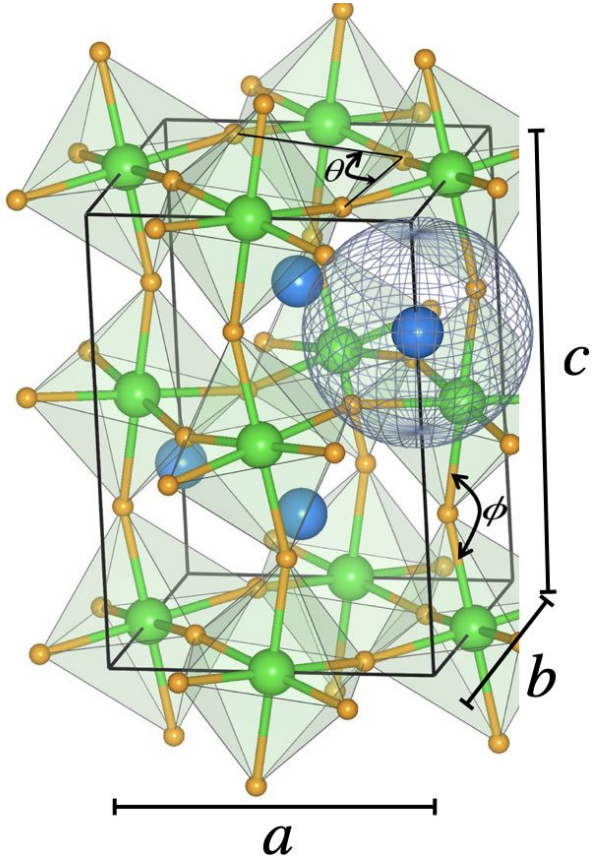
1. Define the nodes



Graph creation

1. Define the nodes

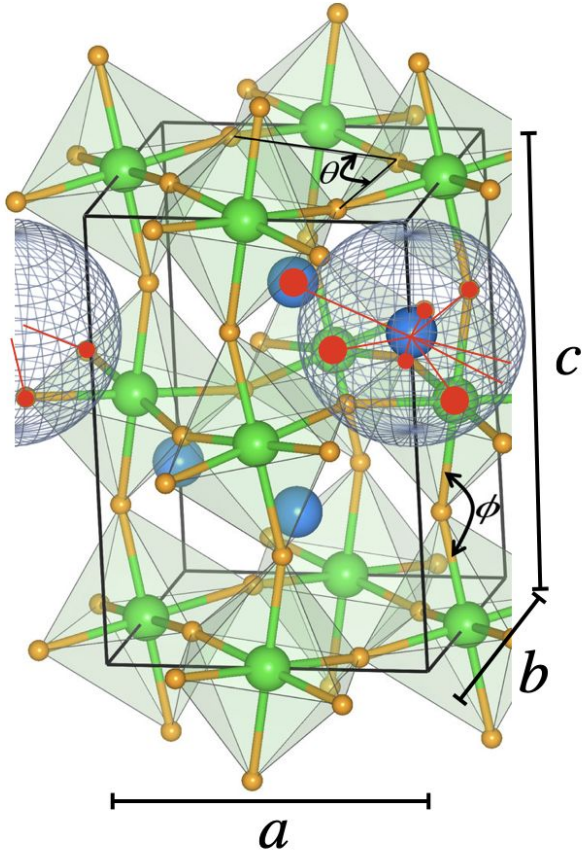
2. Define the Edges



Graph creation

1. Define the nodes

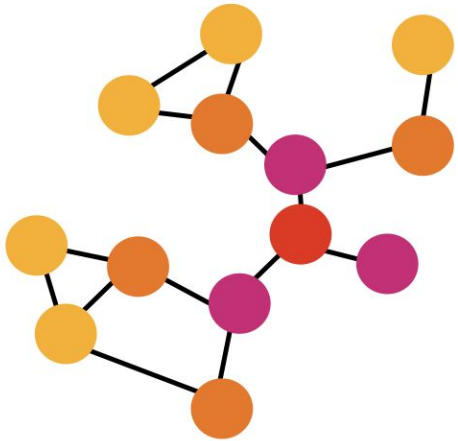
2. Define the Edges

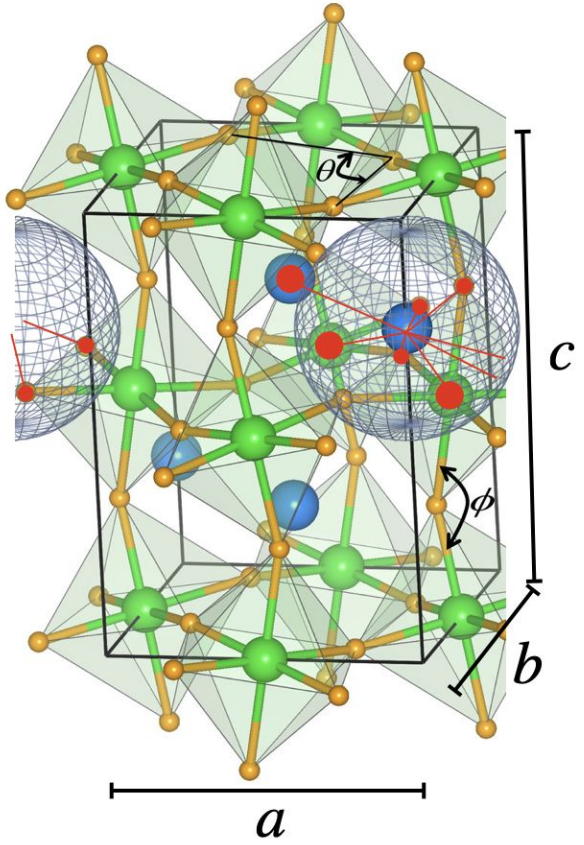


Graph creation

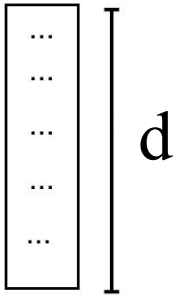
1. Define the nodes

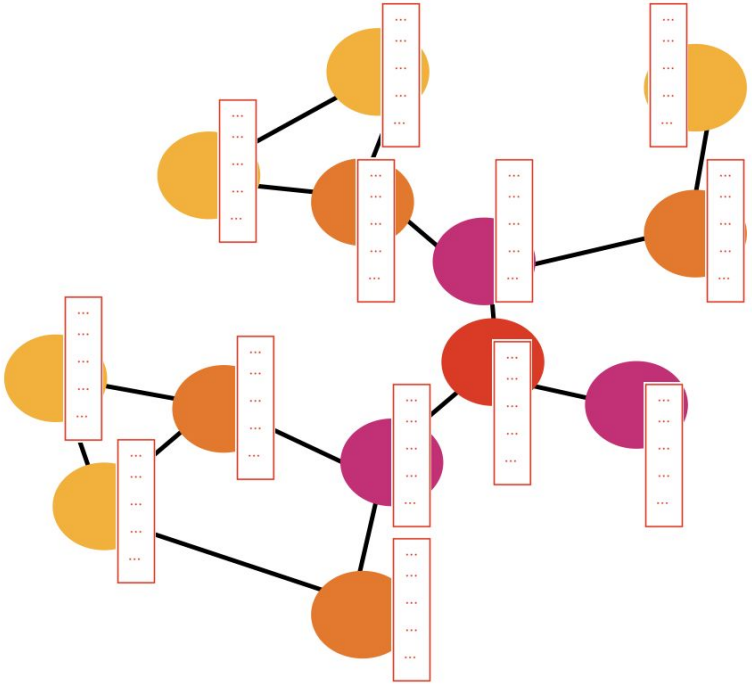
2. Define the Edges



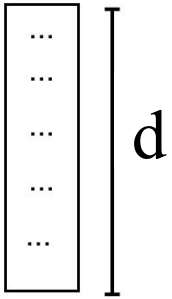


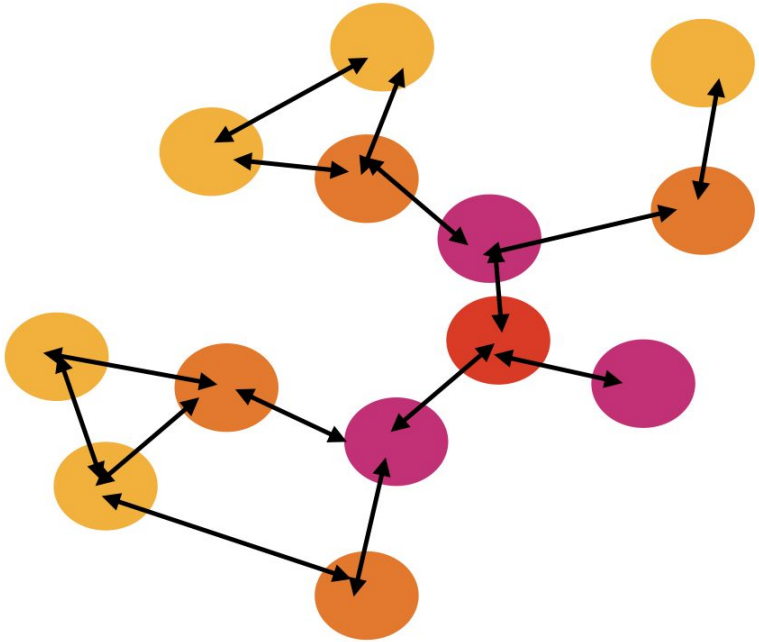
Graph neural network



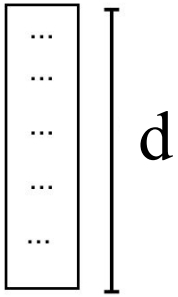


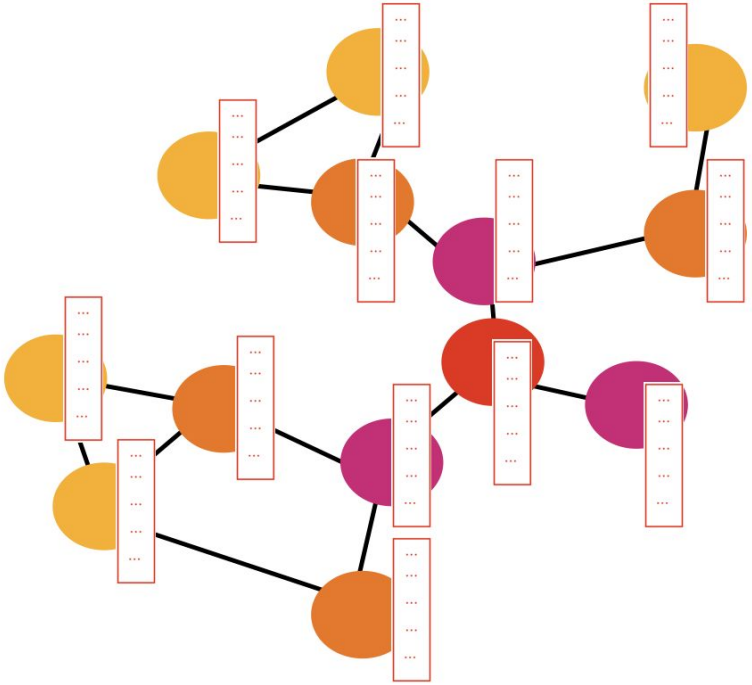
Graph neural network





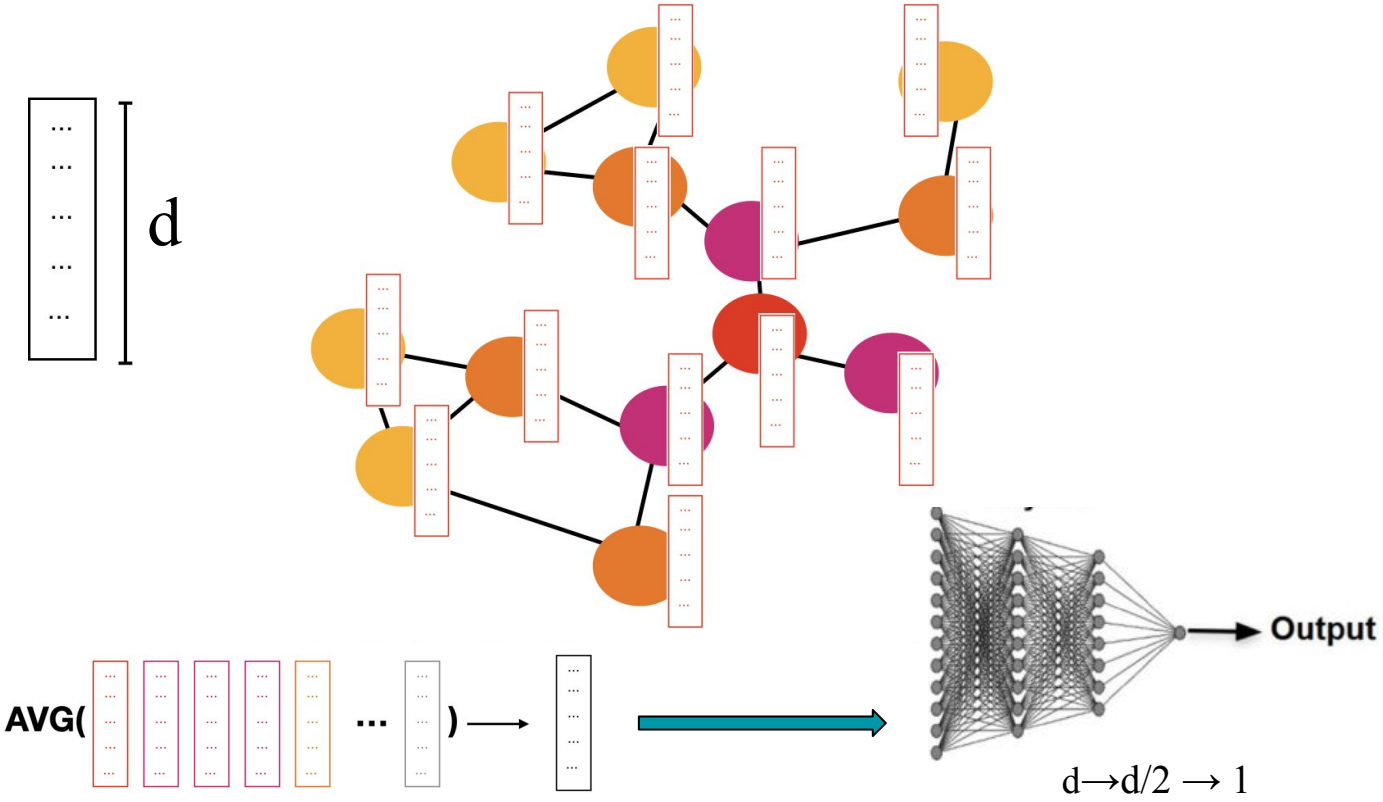
Graph neural network





X times, for each convolution

Graph neural network



Classification: MP

- Classifying above/below Tc = 7 K
  - Class split around 1:2
- Trying different features
- Model has 4 convolutional layers, with hidden dimension of 64

Examples of features

94 +19 node (atomic) features (disordered sites are averaged):

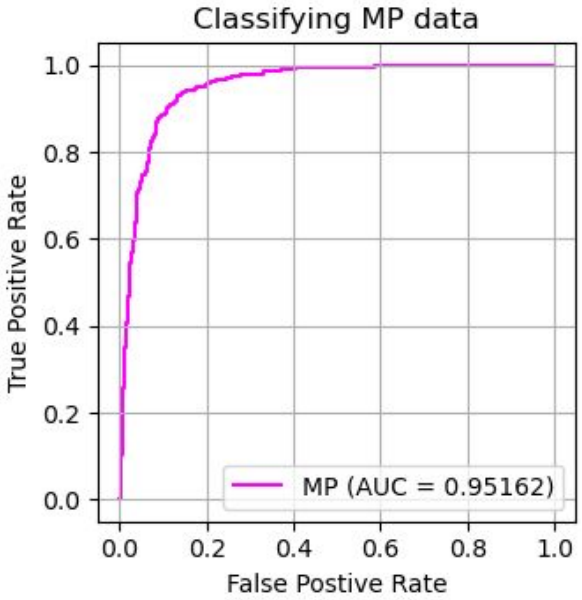
Z (one-hot encoder), mass, fractional coordinates, valence electrons…

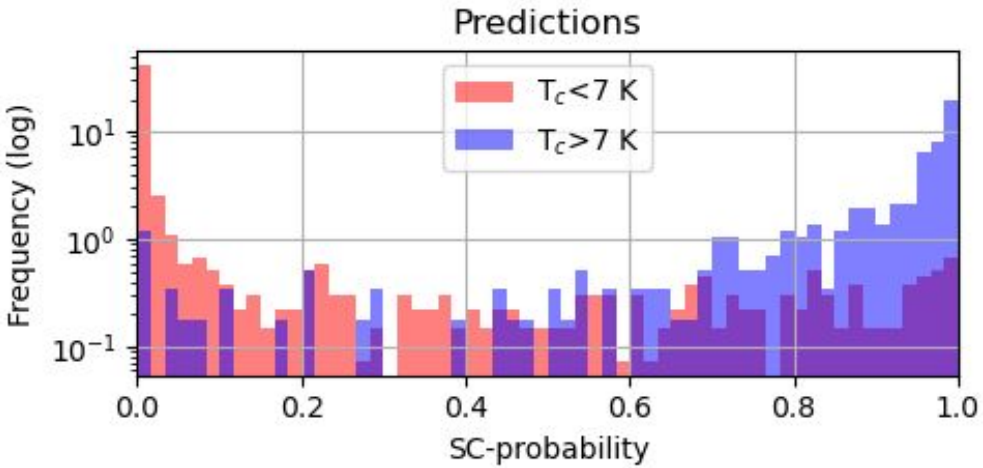
Edge features:

Distance (Gaussian expanded)

Classification: MP

- Classifying above/below Tc = 7 K
  - Class split around 1:2
- Trying different hand-crafted features
- Model has 4 convolutional layers, with hidden dimension of 64

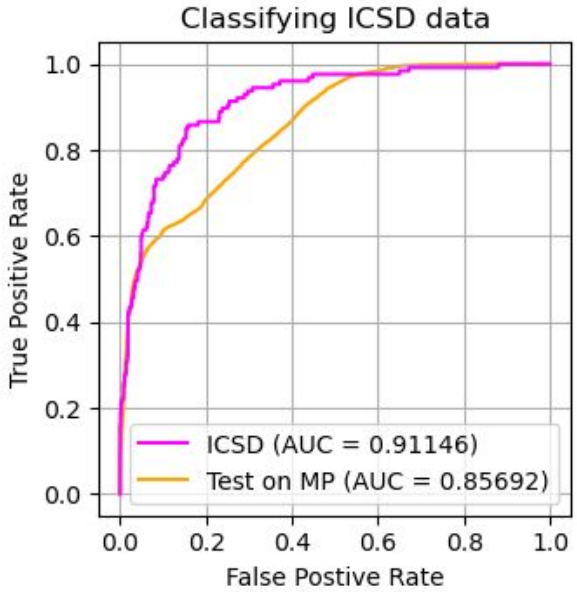


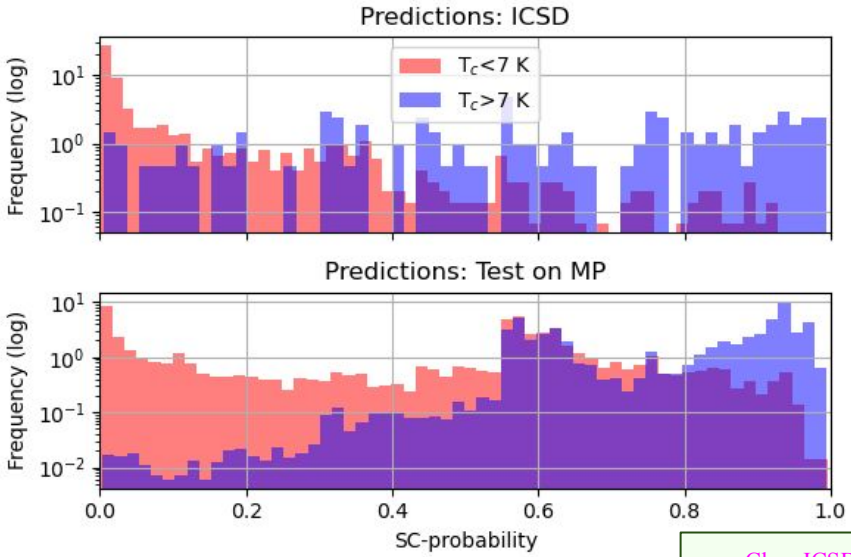


Class MPv: AUC = 0.952

Classification: ICSD

- More structures: Many are doped variants
  - Class split around 2:1 (reversed compared to MP)
- Model has 2 convolutional layers, with hidden dimension of 16
- Test + Final prediction on data from MP





Class ICSD: AUC = 0.911
Class ICSD on MP: AUC=0.857

Regression using a GCNN

Node features

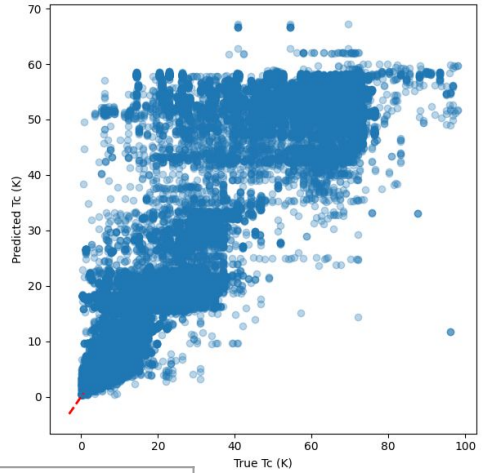
Edge features

Z, row, group, occ, electronegativity, atomic radius

Distance: Gaussian expansion

Disorder

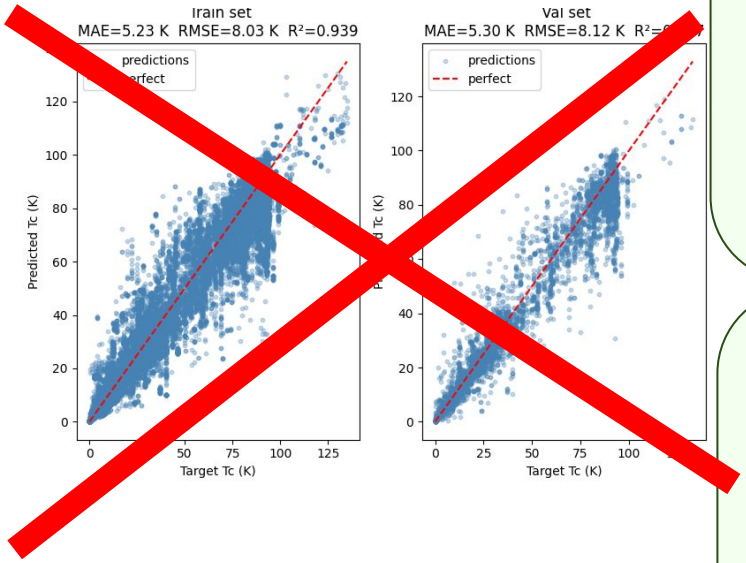
First edition



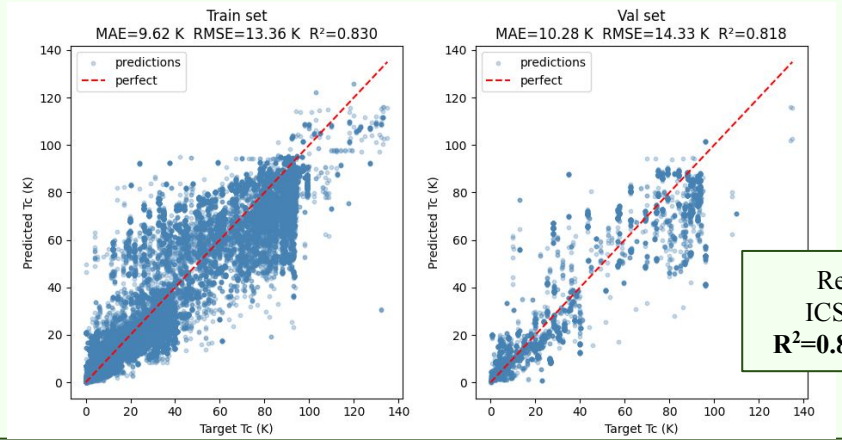
	First edition	“Improved” edition
Node embedding	Linear	Linear → BatchNorm → SiLu
Convolution step	x = Conv(x)	x = x + Conv(x)
Activation function	ReLu	SiLu
Depth	2	3

Regression

No grouping of data

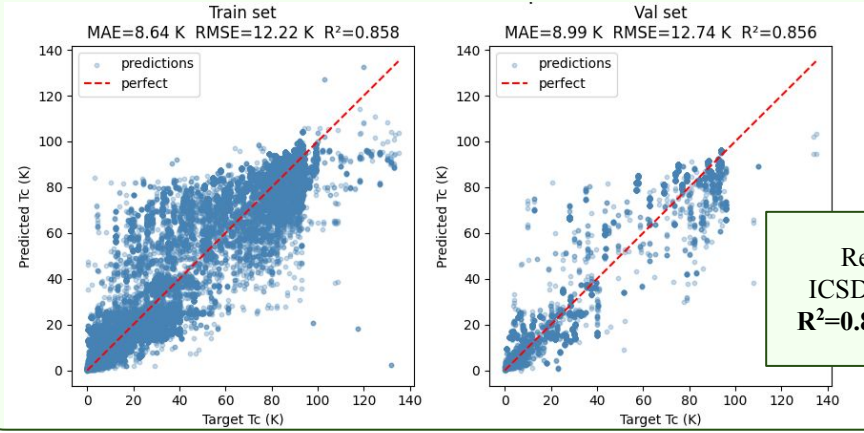


Data grouped - weighted



Regression: GCNN
ICSD data - weighted
R^2=0.818, RMSE=14.33 K

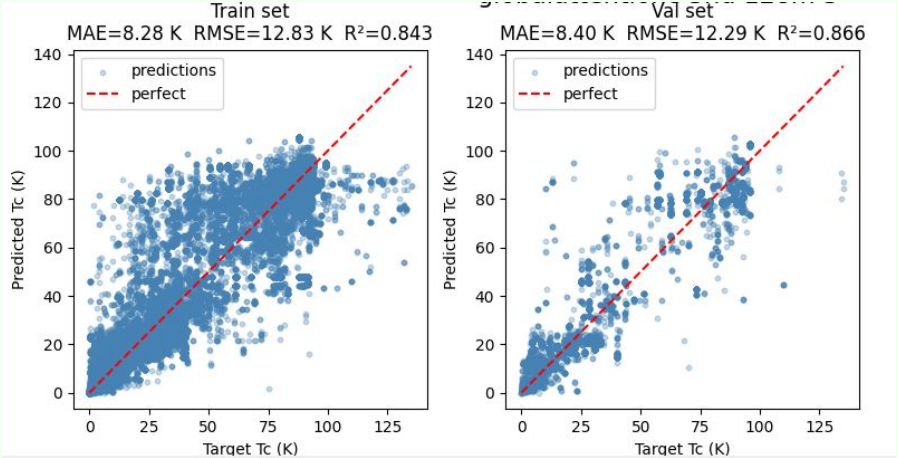
Data grouped - unweighted



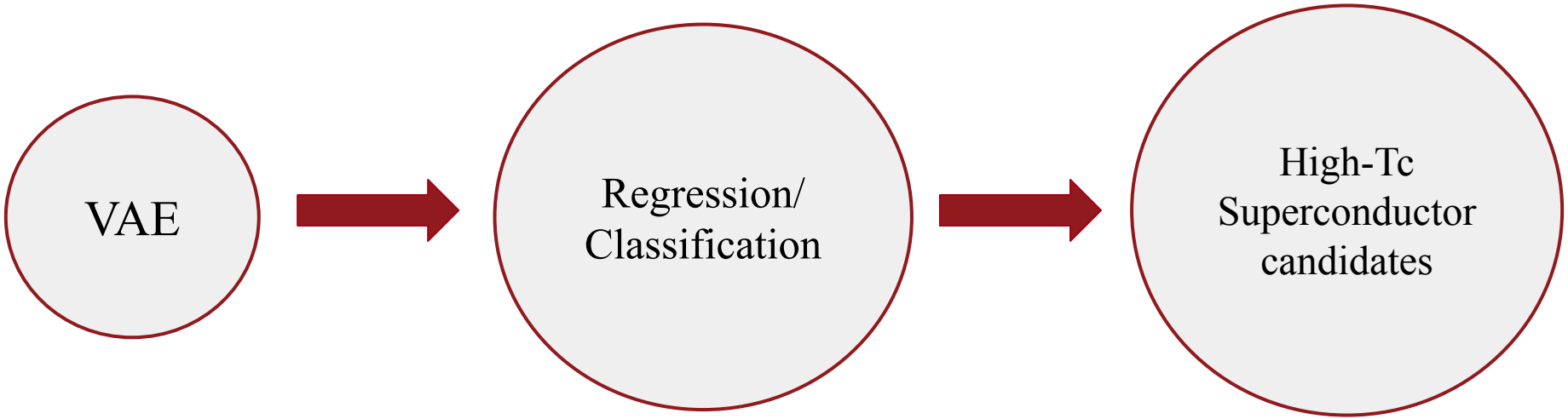
Regression: GCNN
ICSD data - not weighted
R^2=0.856, RMSE=12.74 K

Last minute attempts at improvements - Global attention

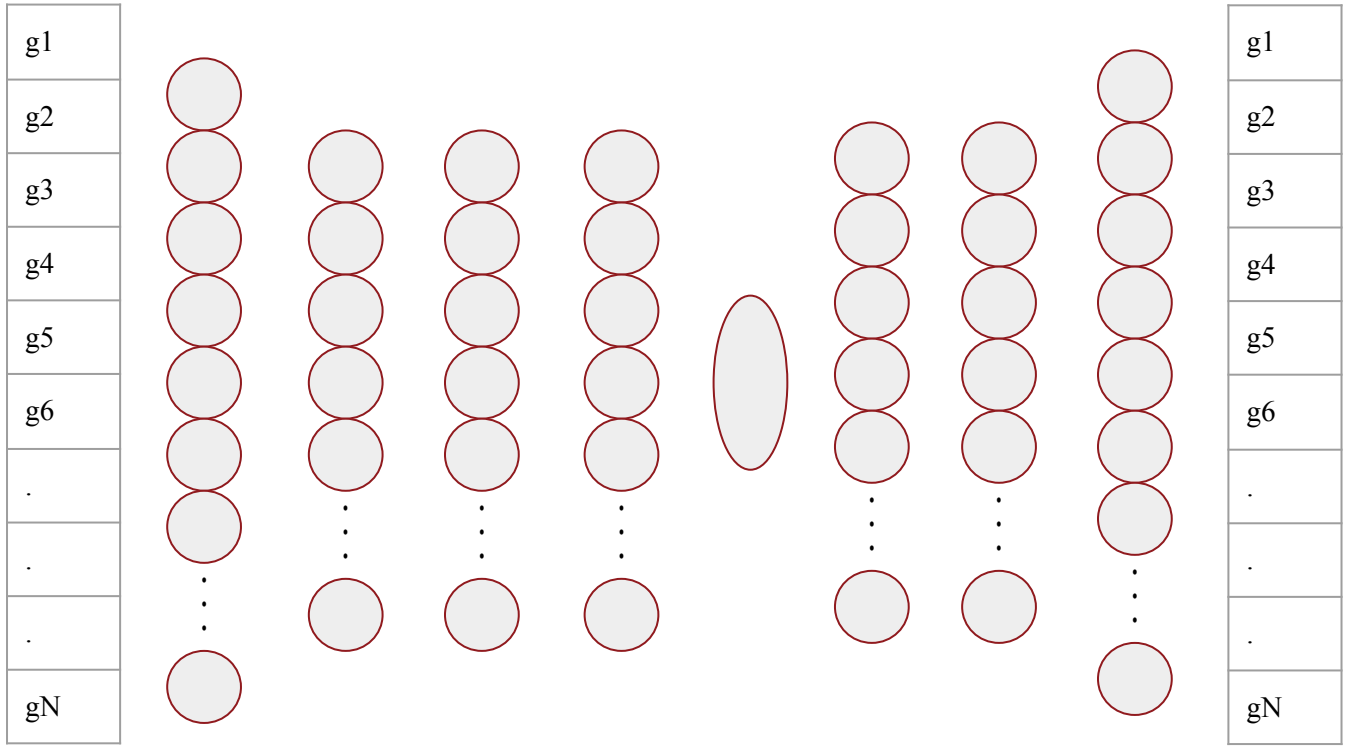
Global mean pooling → Global attention



Regression: GCNN + Global attention
ICSD data - not weighted
R^2=0.866, RMSE=12.29 K

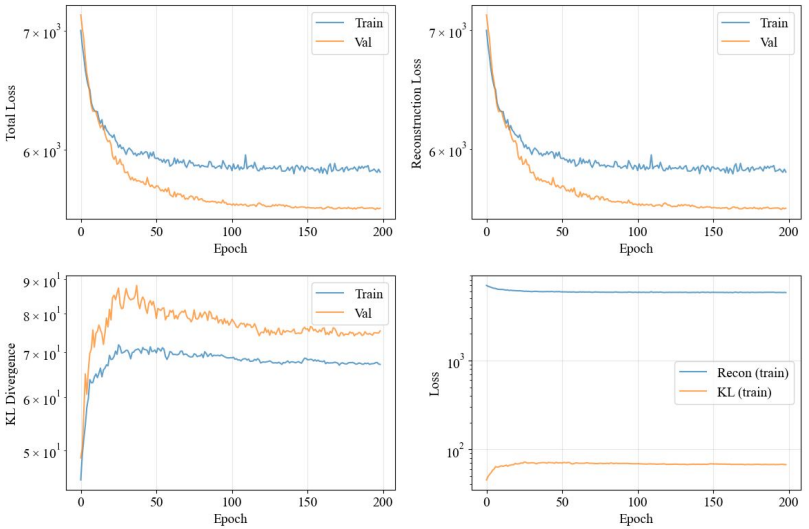

VAE
Regression/
Classification
High-Tc
Superconductor
candidates

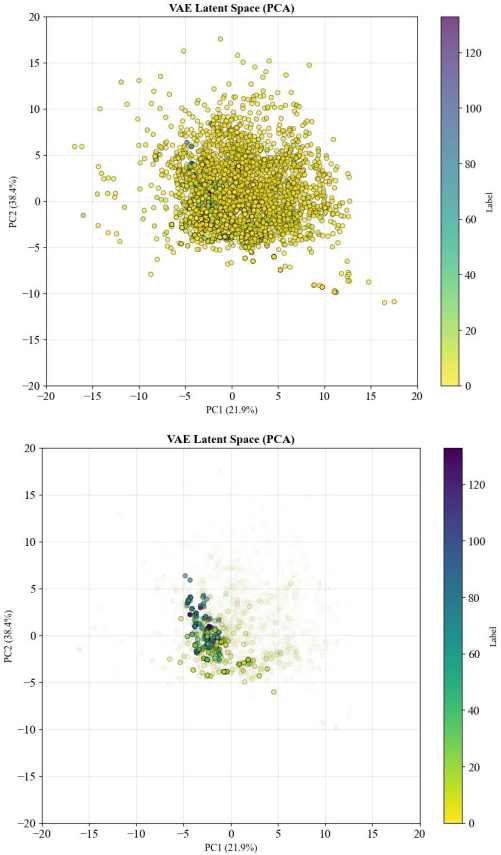
Variational Auto Encoder

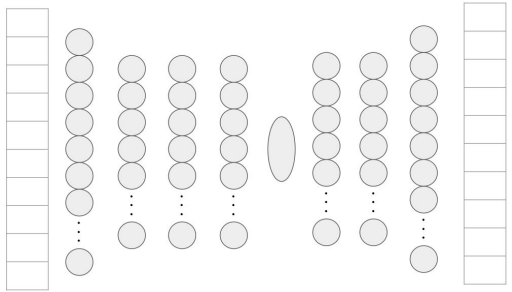


Variational Auto Encoder - MP

Node-level MSE: 0.52 ± 0.51
Median MSE: 0.33
Max MSE: 4.18
KL divergence: 104 ± 86







$$Loss = v \cdot Loss_{nodes} + w \cdot Loss_{edges} + x \cdot KL$$

Conclusion

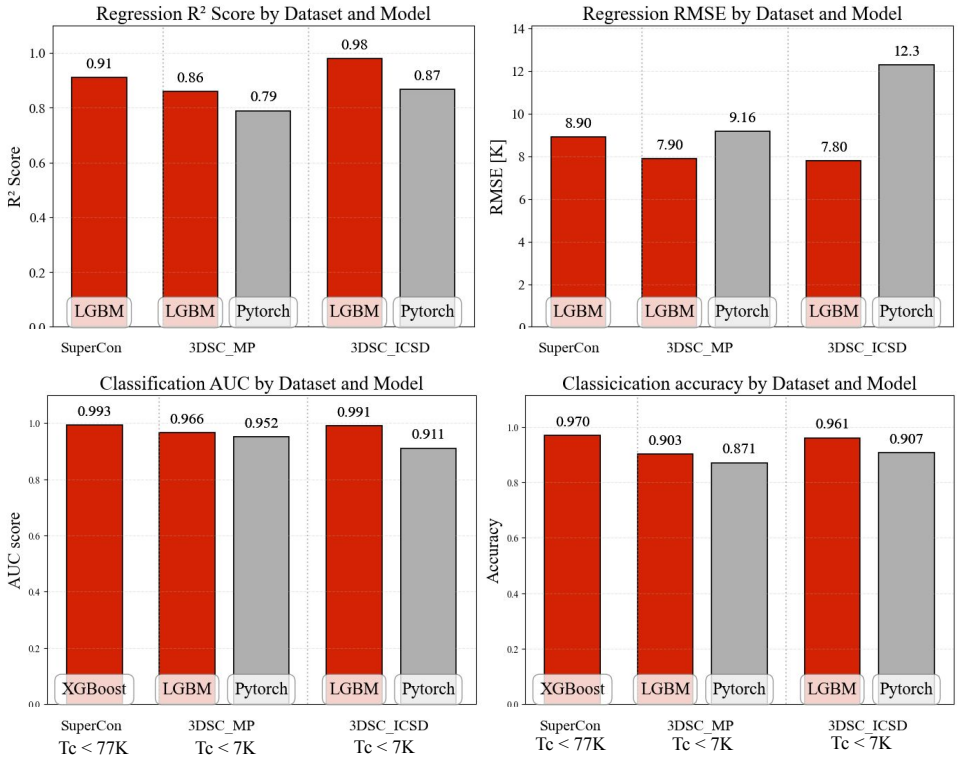
Tackling challenging data

Predicting Tc

Extrapolation is a challenge

Graphs: Proof of concept

Toolbox utilized



Thank you for listening!

Appendix

All code is available at
https://github.com/AmalieFalkenberg/Applied_Machine_Learining.git

References describing used data:

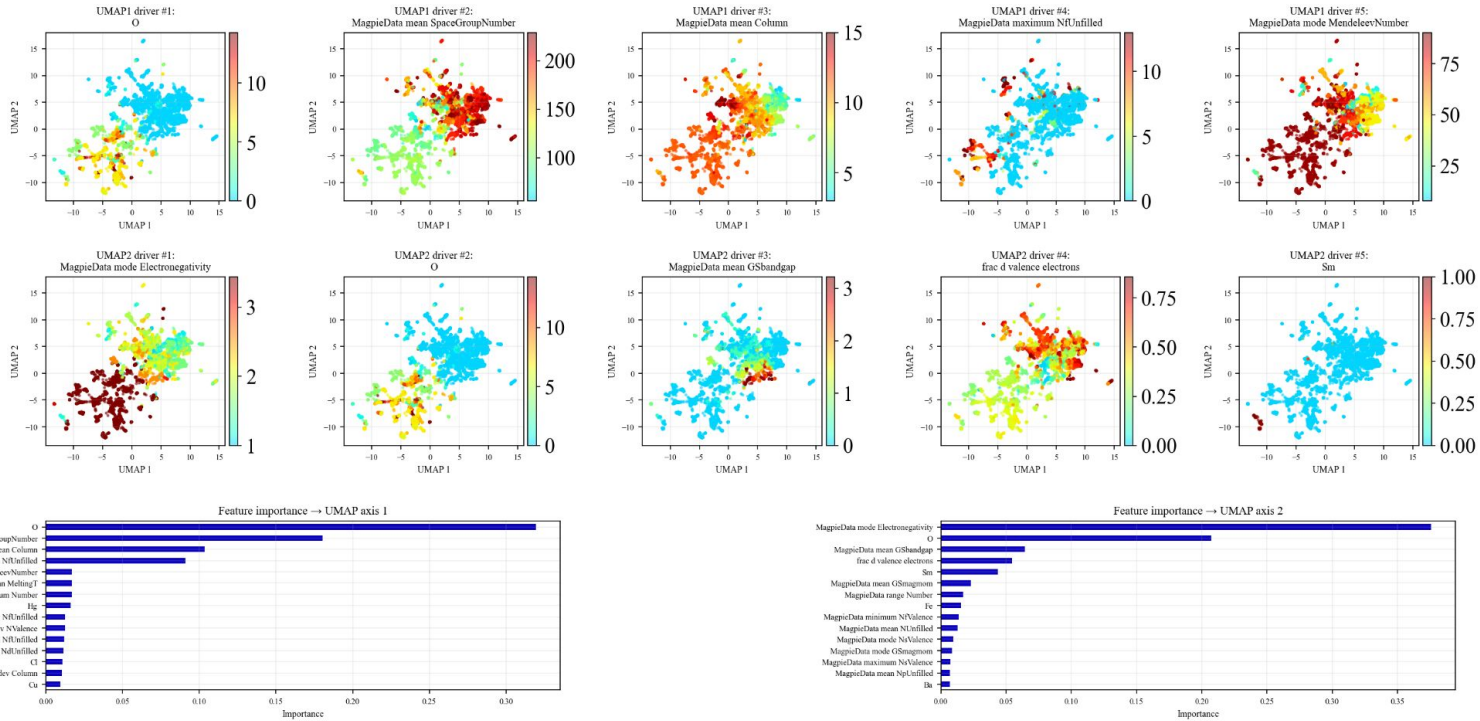
Siwoo Lee et al 2025 Mach. Learn.: Sci. Technol. 6 035052

Timo Sommer et al 2023 Nature: Scientific data 10:816

I. SuperCon

I.2. Visualization

UMAP Feature importance



I.3. Regression

XGBoost

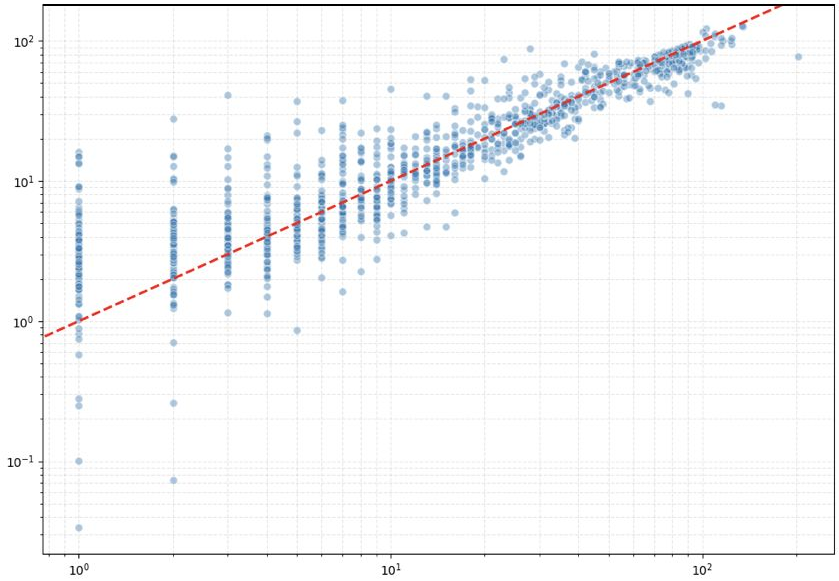
Training: 75%

Validating: 15%

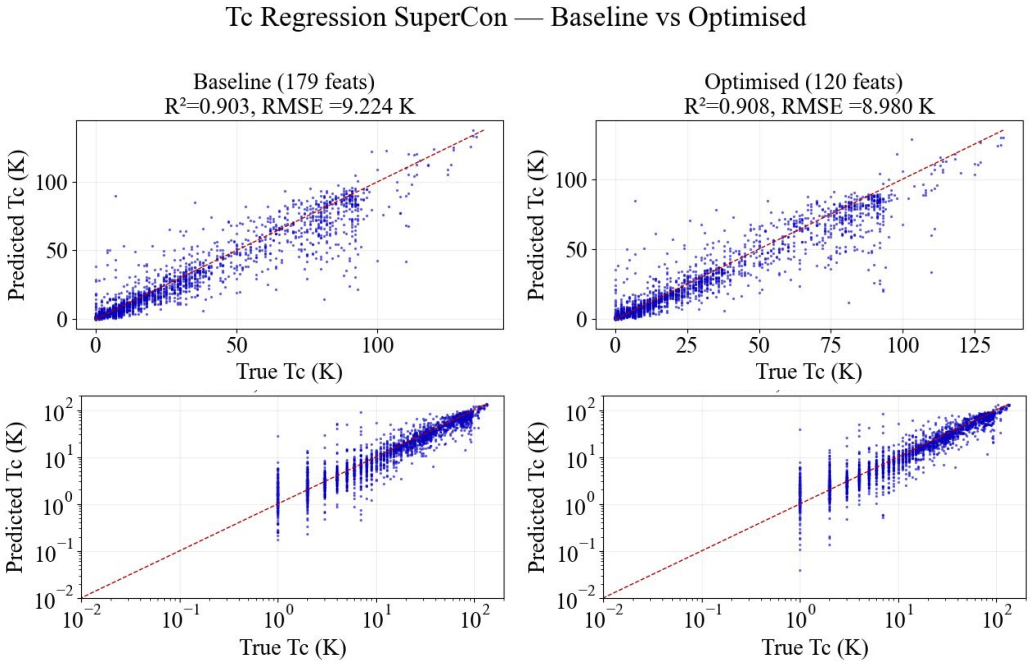
Testing: 10%

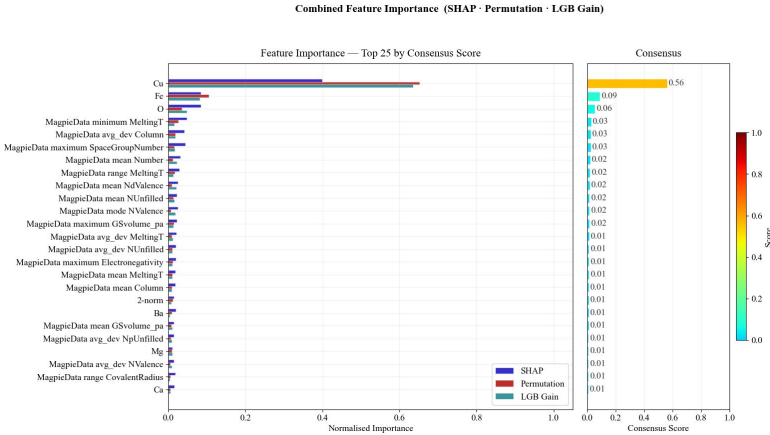
Article result : $\sigma = 10K$

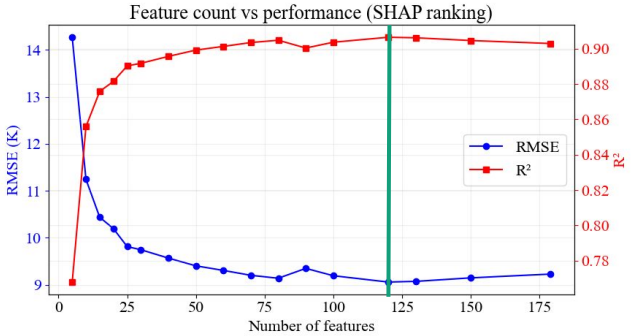
Final results: $\sigma = 9.5K$



I.3. Regression





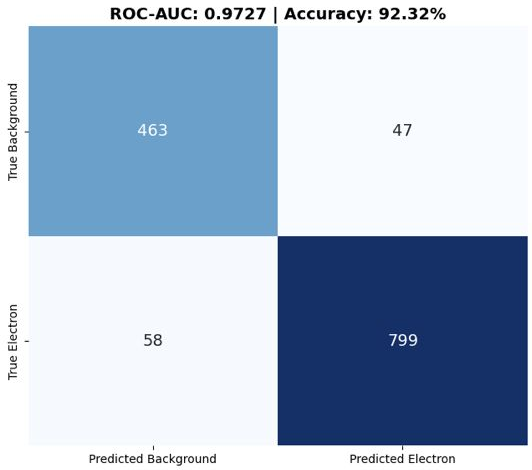


Plots of SuperCon regression presented in main slides (LGBM with 120 features)

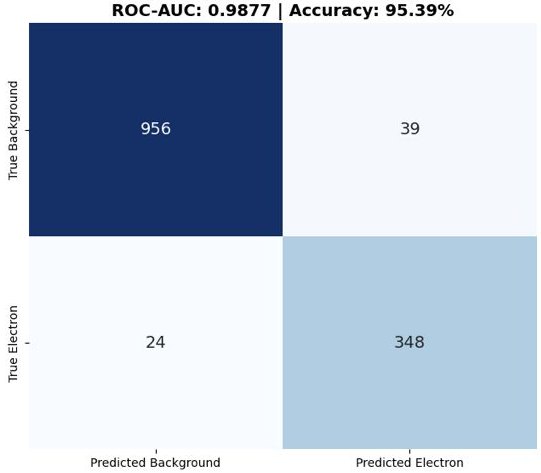
I.4. Classification

Classification results

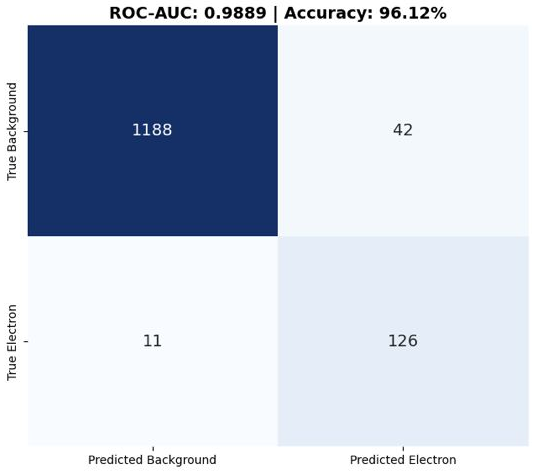
$T_{c.limit} = 5$



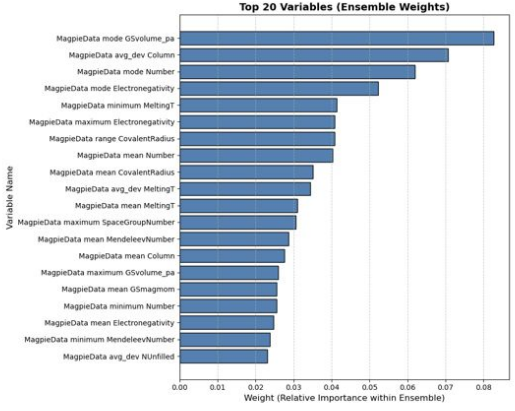
$T_{c.limit} = 30$

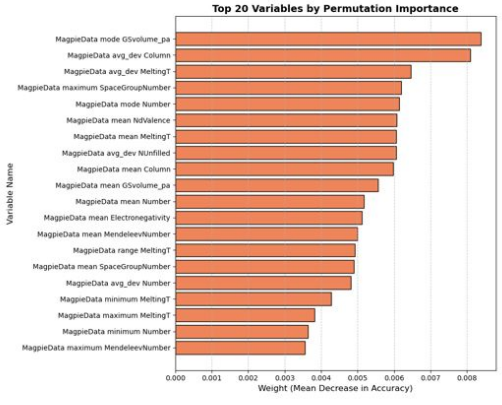


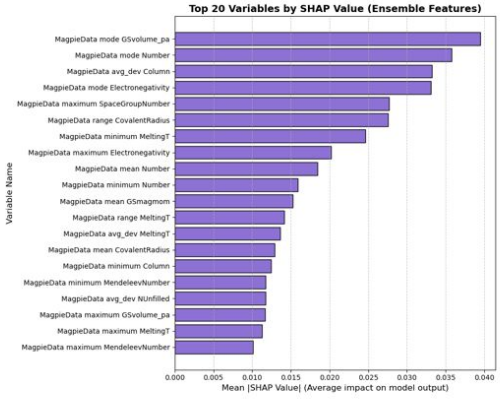
$T_{c.limit} = 77$



I.5. Feature importance







$$T_{c.limit} = 5$$

	Forward_Rank	Variable	Accuracy
0	1	MagpieData mode Number	0.840001
1	2	MagpieData mean MendeleevNumber	0.867799
2	3	MagpieData avg_dev Column	0.882800
3	4	MagpieData maximum MendeleevNumber	0.894000
4	5	MagpieData mean MeltingT	0.898200
5	6	MagpieData avg_dev Number	0.899000
6	7	MagpieData minimum Number	0.898401
7	8	0-norm	0.899401
8	9	MagpieData mean CovalentRadius	0.900801
9	10	MagpieData mean Electronegativity	0.900001
10	11	MagpieData mean GSvolume_pa	0.900201
11	12	MagpieData avg_dev MeltingT	0.901601
12	13	MagpieData maximum Electronegativity	0.898000
13	14	MagpieData mean NdValence	0.900000
14	15	MagpieData maximum GSvolume_pa	0.901401
15	16	MagpieData mean Column	0.901801

	Variable	RFE_Rank
0	MagpieData avg_dev Column	1
1	MagpieData mode GSvolume_pa	2
2	MagpieData mode Number	3
3	MagpieData avg_dev MeltingT	4
4	MagpieData mean MendeleevNumber	5
5	MagpieData mode Electronegativity	6
6	MagpieData mean MeltingT	7
7	MagpieData mean Number	8
8	MagpieData minimum MeltingT	9
9	MagpieData maximum SpaceGroupNumber	10
10	MagpieData mean Electronegativity	11
11	MagpieData mean Column	12
12	MagpieData maximum Electronegativity	13
13	MagpieData mean GSmagmom	14
14	MagpieData mean GSvolume_pa	15
15	MagpieData maximum GSvolume_pa	16

I.5. Feature importance

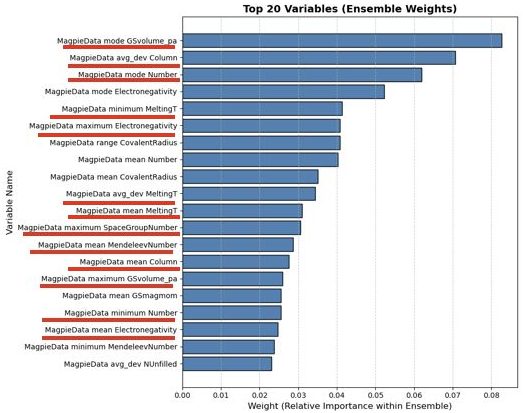
What atoms : Copper - Oxygen - Iron

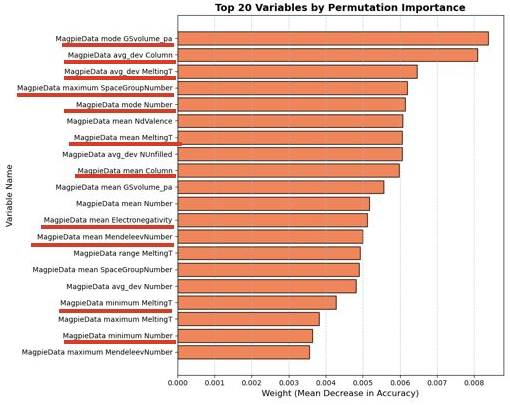
Atom configuration : GS_volume - space group number - colavent radius

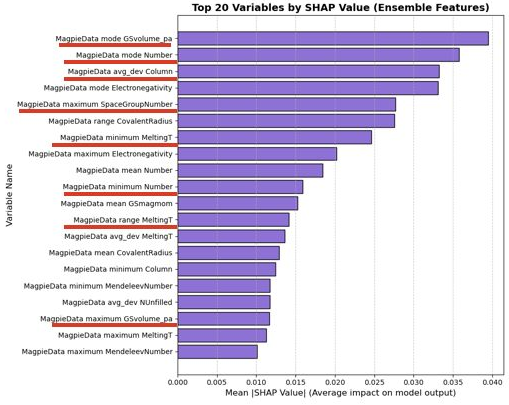
Macroscopic properties : Melting temperature - electronegativity

I.5. Feature importance

Which variables are important ?







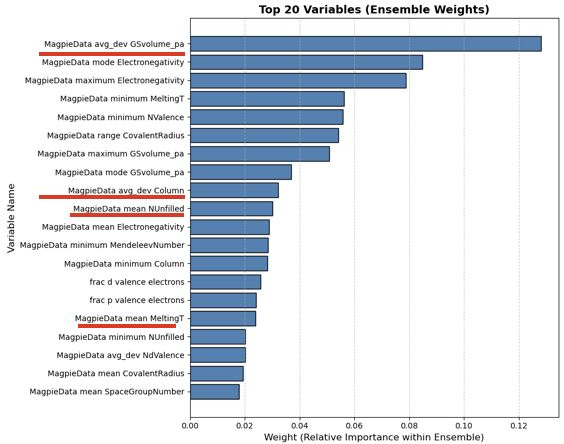
$$T_{c.limit} = 5$$

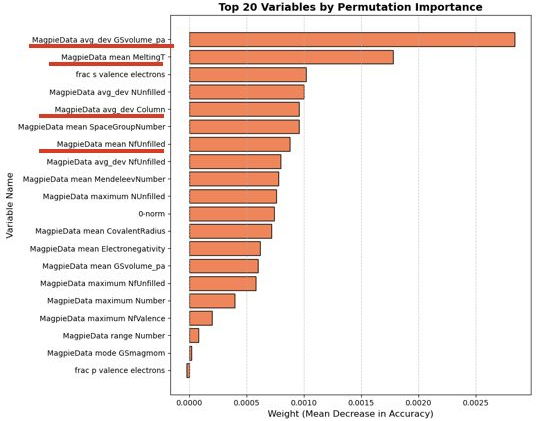
	Forward_Rank	Variable	Accuracy
0	1	MagpieData mode Number	0.840001
1	2	MagpieData mean MendeleevNumber	0.867799
2	3	MagpieData avg_dev Column	0.882800
3	4	MagpieData maximum MendeleevNumber	0.894000
4	5	MagpieData mean MeltingT	0.898200
5	6	MagpieData avg_dev Number	0.899000
6	7	MagpieData minimum Number	0.898401
7	8	0-norm	0.899401
8	9	MagpieData mean CovalentRadius	0.900801
9	10	MagpieData mean Electronegativity	0.900001
10	11	MagpieData mean GSvolume_pa	0.900201
11	12	MagpieData avg_dev MeltingT	0.901601
12	13	MagpieData maximum Electronegativity	0.898000
13	14	MagpieData mean NdValence	0.900000
14	15	MagpieData maximum GSvolume_pa	0.901401
15	16	MagpieData mean Column	0.901801

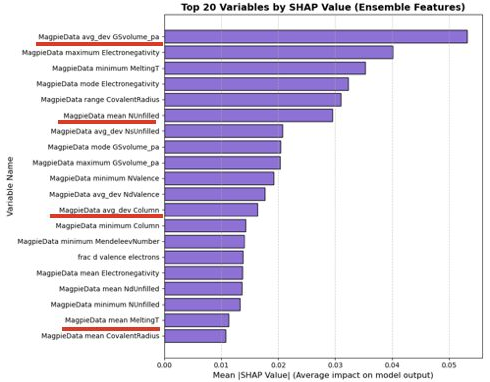
	Variable	RFE_Rank
0	MagpieData avg_dev Column	1
1	MagpieData mode GSvolume_pa	2
2	MagpieData mode Number	3
3	MagpieData avg_dev MeltingT	4
4	MagpieData mean MendeleevNumber	5
5	MagpieData mode Electronegativity	6
6	MagpieData mean MeltingT	7
7	MagpieData mean Number	8
8	MagpieData minimum MeltingT	9
9	MagpieData maximum SpaceGroupNumber	10
10	MagpieData mean Electronegativity	11
11	MagpieData mean Column	12
12	MagpieData maximum Electronegativity	13
13	MagpieData mean GSmagmom	14
14	MagpieData mean GSvolume_pa	15
15	MagpieData maximum GSvolume_pa	16

I.5. Feature importance

Which variables are important ?







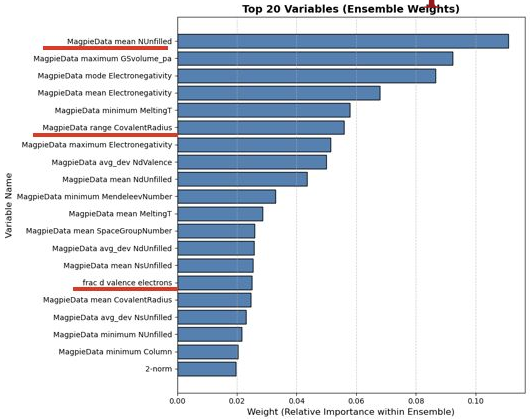
$$T_{c.limit} = 30$$

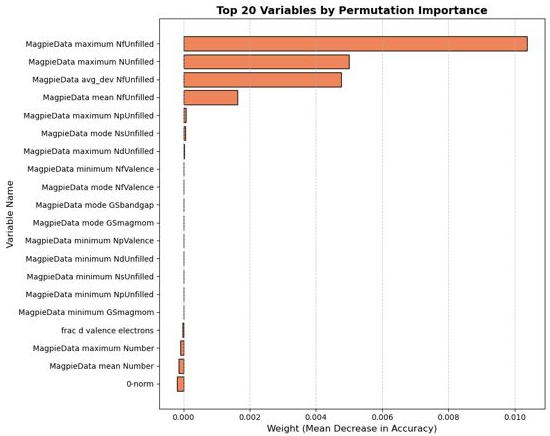
	Variable	Perm_Weight
0	MagpieData avg_dev GSvolume_pa	0.00284
1	MagpieData mean MeltingT	0.00178
2	frac s valence electrons	0.00102
3	MagpieData avg_dev NUnfilled	0.00100
4	MagpieData avg_dev Column	0.00096
5	MagpieData mean SpaceGroupNumber	0.00096
6	MagpieData mean NfUnfilled	0.00088
7	MagpieData avg_dev NfUnfilled	0.00080
8	MagpieData mean MendeleevNumber	0.00078
9	MagpieData maximum NUnfilled	0.00076
10	0-norm	0.00074
11	MagpieData mean CovalentRadius	0.00072
12	MagpieData mean Electronegativity	0.00062
13	MagpieData mean GSvolume_pa	0.00060
14	MagpieData maximum NfUnfilled	0.00058
15	MagpieData maximum Number	0.00040

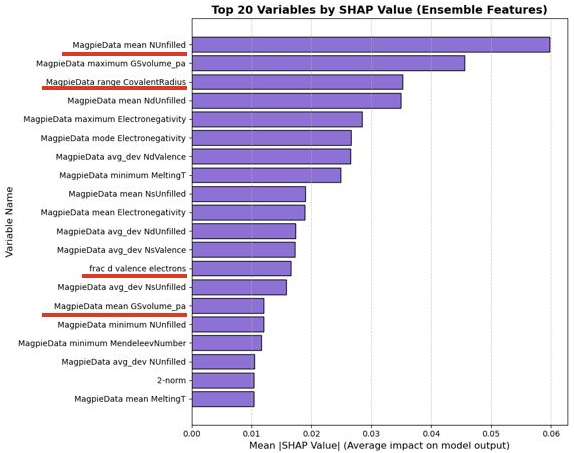
	Variable	RFE_Rank
0	MagpieData avg_dev GSvolume_pa	1
1	MagpieData mode Electronegativity	2
2	MagpieData maximum Electronegativity	3
3	MagpieData maximum GSvolume_pa	4
4	MagpieData mean NUnfilled	5
5	frac d valence electrons	6
6	MagpieData minimum MeltingT	7
7	MagpieData minimum NValence	8
8	MagpieData mean SpaceGroupNumber	9
9	MagpieData mode GSvolume_pa	10
10	MagpieData avg_dev Column	11
11	MagpieData avg_dev NdValence	12
12	MagpieData range CovalentRadius	13
13	MagpieData mean MeltingT	14
14	frac s valence electrons	15
15	MagpieData mean NdUnfilled	16

I.5. Feature importance

Which variables are important ?







$$T_{c.limit} = 77$$

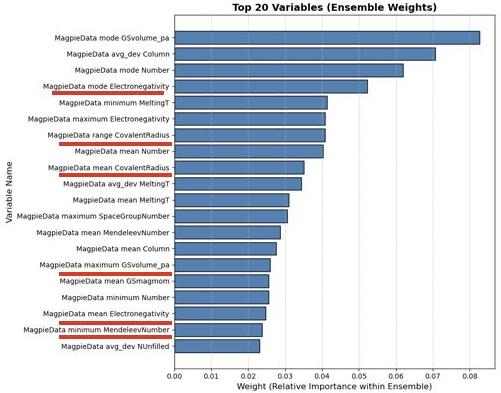
Forward_Rank	Variable	Accuracy
1	MagpieData mean NfUnfilled	0.892200
2	MagpieData avg_dev NfUnfilled	0.893601
3	MagpieData maximum NfUnfilled	0.895601
4	MagpieData minimum GSmagmom	0.896001
5	MagpieData minimum NsUnfilled	0.895601
6	MagpieData range CovalentRadius	0.862000
7	MagpieData avg_dev NUnfilled	0.885999
8	frac d valence electrons	0.899599
9	MagpieData maximum NpUnfilled	0.903399
10	MagpieData maximum NdUnfilled	0.905200
11	MagpieData mean Electronegativity	0.909000
12	MagpieData mean Number	0.905800
13	MagpieData maximum NUnfilled	0.909600
14	MagpieData mean CovalentRadius	0.909199
15	MagpieData maximum CovalentRadius	0.906399
16	MagpieData mean GSvolume_pa	0.909801

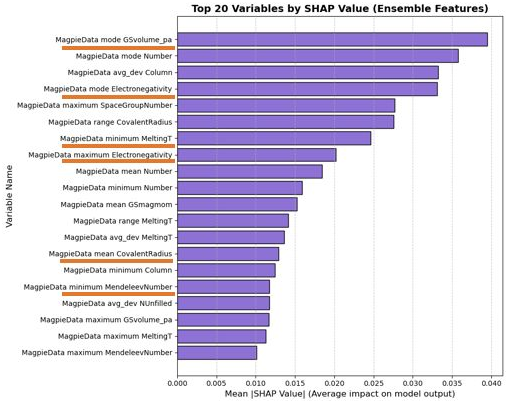
	Variable	RFE_Rank
0	MagpieData mean NUnfilled	1
1	MagpieData maximum GSvolume_pa	2
2	MagpieData mode Electronegativity	3
3	MagpieData range CovalentRadius	4
4	MagpieData mean Electronegativity	5
5	MagpieData avg_dev NdValence	6
6	MagpieData minimum MeltingT	7
7	MagpieData mean NsUnfilled	8
8	frac d valence electrons	9
9	MagpieData mean NdUnfilled	10
10	MagpieData avg_dev NsUnfilled	11
11	MagpieData mean GSvolume_pa	12
12	MagpieData mean MeltingT	13
13	MagpieData mean CovalentRadius	14
14	MagpieData avg_dev NdUnfilled	15
15	MagpieData avg_dev NsValence	16

I.5. Feature importance

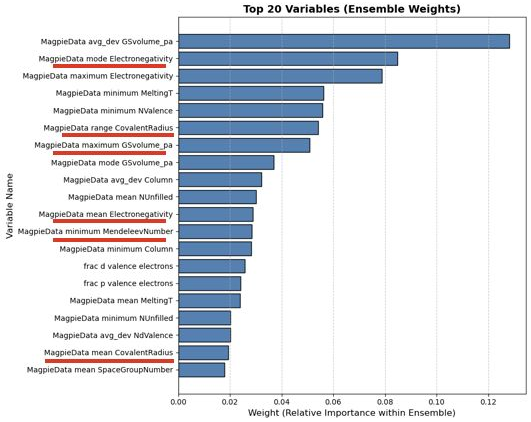
Which variables are important ?

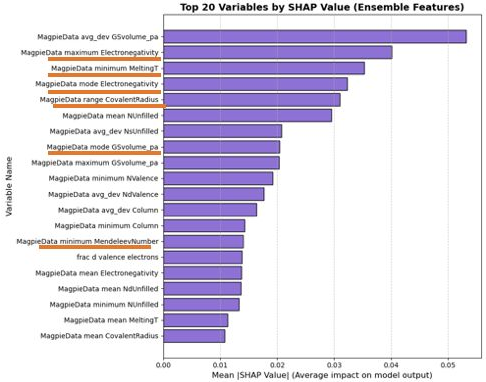
$T_{c.limit} = 5$



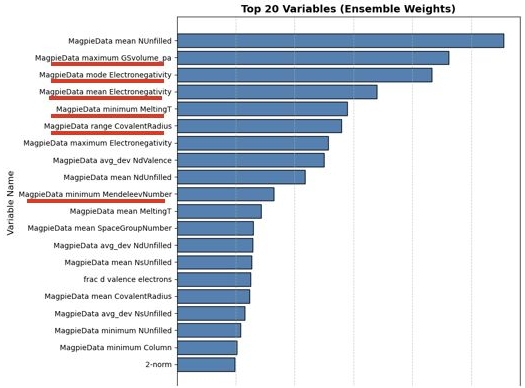


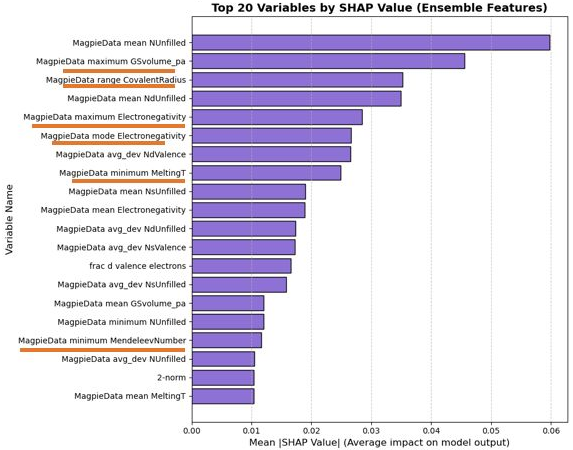
$T_{c.limit} = 30$





$T_{c.limit} = 77$





I.5. Feature importance

Which variables are important ?

$T_{c.limit} = 5$

Forward_Rank	Variable	Accuracy
1	MagpieData mode Number	0.840001
2	MagpieData mean MendeleevNumber	0.867799
3	MagpieData avg_dev Column	0.882800
4	MagpieData maximum MendeleevNumber	0.894000
5	MagpieData mean MeltingT	0.898200
6	MagpieData avg_dev Number	0.899000
7	MagpieData minimum Number	0.898401
8	0-norm	0.899401
9	MagpieData mean CovalentRadius	0.900801
10	MagpieData mean Electronegativity	0.900001
11	MagpieData mean GSvolume_pa	0.900201
12	MagpieData avg_dev MeltingT	0.901601
13	MagpieData maximum Electronegativity	0.898000
14	MagpieData mean NdValence	0.900000
15	MagpieData maximum GSvolume_pa	0.901401
16	MagpieData mean Column	0.901801

	Variable	RFE_Rank
0	MagpieData avg_dev Column	1
1	MagpieData mode GSvolume_pa	2
2	MagpieData mode Number	3
3	MagpieData avg_dev MeltingT	4
4	MagpieData mean MendeleevNumber	5
5	MagpieData mode Electronegativity	6
6	MagpieData mean MeltingT	7
7	MagpieData mean Number	8
8	MagpieData minimum MeltingT	9
9	MagpieData maximum SpaceGroupNumber	10
10	MagpieData mean Electronegativity	11
11	MagpieData mean Column	12
12	MagpieData maximum Electronegativity	13
13	MagpieData mean GSmagmom	14
14	MagpieData mean GSvolume_pa	15
15	MagpieData maximum GSvolume_pa	16

$T_{c.limit} = 30$

	Variable	Perm_Weight
0	MagpieData avg_dev GSvolume_pa	0.00284
1	MagpieData mean MeltingT	0.00178
2	frac s valence electrons	0.00102
3	MagpieData avg_dev NUnfilled	0.00100
4	MagpieData avg_dev Column	0.00096
5	MagpieData mean SpaceGroupNumber	0.00096
6	MagpieData mean NfUnfilled	0.00088
7	MagpieData avg_dev NfUnfilled	0.00080
8	MagpieData mean MendeleevNumber	0.00078
9	MagpieData maximum NUnfilled	0.00076
10	0-norm	0.00074
11	MagpieData mean CovalentRadius	0.00072
12	MagpieData mean Electronegativity	0.00062
13	MagpieData mean GSvolume_pa	0.00060
14	MagpieData maximum NfUnfilled	0.00058
15	MagpieData maximum Number	0.00040

	Variable	RFE_Rank
0	MagpieData avg_dev GSvolume_pa	1
1	MagpieData mode Electronegativity	2
2	MagpieData maximum Electronegativity	3
3	MagpieData maximum GSvolume_pa	4
4	MagpieData mean NUnfilled	5
5	frac d valence electrons	6
6	MagpieData minimum MeltingT	7
7	MagpieData minimum NValence	8
8	MagpieData mean SpaceGroupNumber	9
9	MagpieData mode GSvolume_pa	10
10	MagpieData avg_dev Column	11
11	MagpieData avg_dev NdValence	12
12	MagpieData range CovalentRadius	13
13	MagpieData mean MeltingT	14
14	frac s valence electrons	15
15	MagpieData mean NdUnfilled	16

$T_{c.limit} = 77$

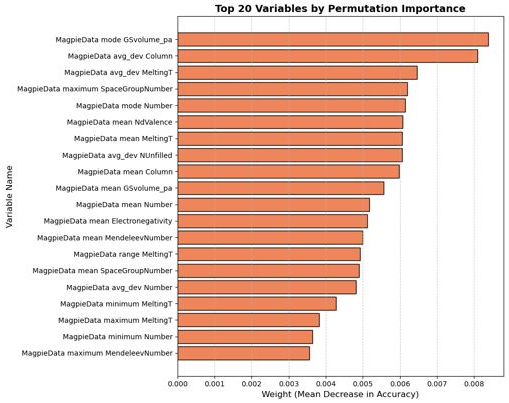
Forward_Rank	Variable	Accuracy
1	MagpieData mean NfUnfilled	0.892200
2	MagpieData avg_dev NfUnfilled	0.893601
3	MagpieData maximum NfUnfilled	0.895601
4	MagpieData minimum GSmagmom	0.896001
5	MagpieData minimum NsUnfilled	0.895601
6	MagpieData range CovalentRadius	0.862000
7	MagpieData avg_dev NUnfilled	0.885999
8	frac d valence electrons	0.899599
9	MagpieData maximum NpUnfilled	0.903399
10	MagpieData maximum NdUnfilled	0.905200
11	MagpieData mean Electronegativity	0.909000
12	MagpieData mean Number	0.905800
13	MagpieData maximum NUnfilled	0.909600
14	MagpieData mean CovalentRadius	0.909199
15	MagpieData maximum CovalentRadius	0.906399
16	MagpieData mean GSvolume_pa	0.909801

	Variable	RFE_Rank
0	MagpieData mean NUnfilled	1
1	MagpieData maximum GSvolume_pa	2
2	MagpieData mode Electronegativity	3
3	MagpieData range CovalentRadius	4
4	MagpieData mean Electronegativity	5
5	MagpieData avg_dev NdValence	6
6	MagpieData minimum MeltingT	7
7	MagpieData mean NsUnfilled	8
8	frac d valence electrons	9
9	MagpieData mean NdUnfilled	10
10	MagpieData avg_dev NsUnfilled	11
11	MagpieData mean GSvolume_pa	12
12	MagpieData mean MeltingT	13
13	MagpieData mean CovalentRadius	14
14	MagpieData avg_dev NdUnfilled	15
15	MagpieData avg_dev NsValence	16

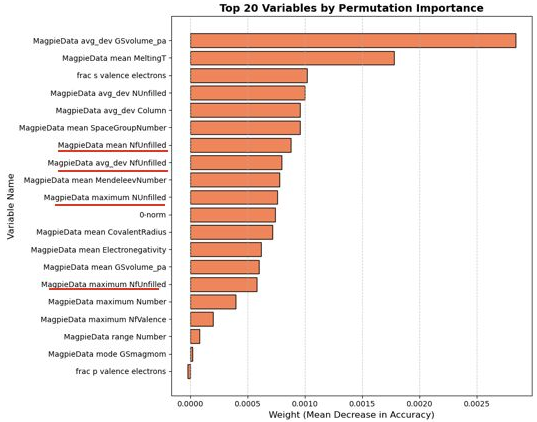
I.5. Feature importance

Which variables are important ?

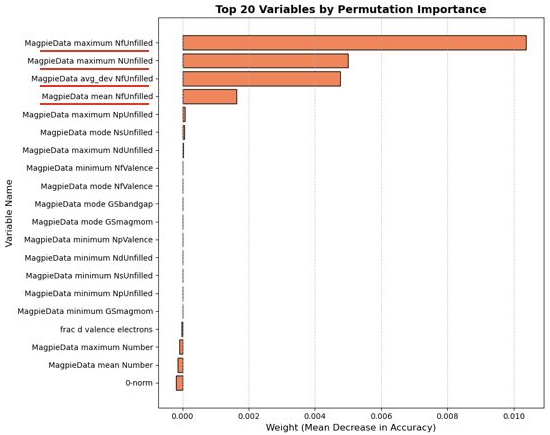
$T_{c.limit} = 5$

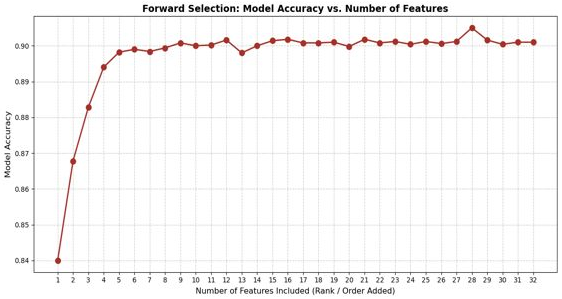


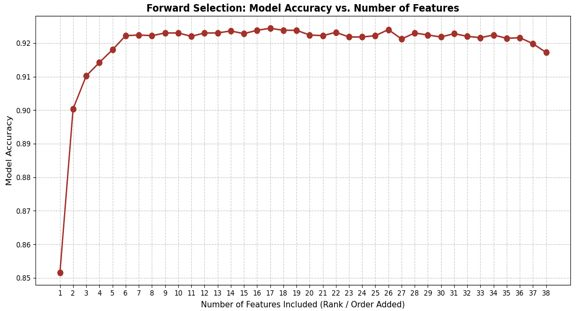
$T_{c.limit} = 30$

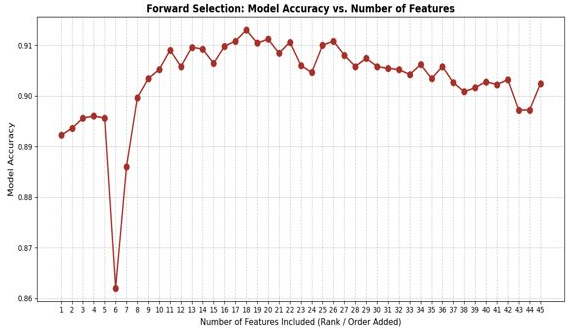


$T_{c.limit} = 77$





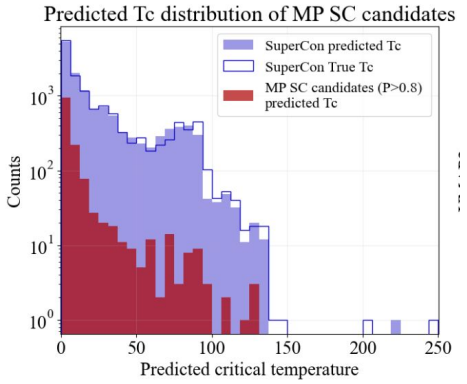


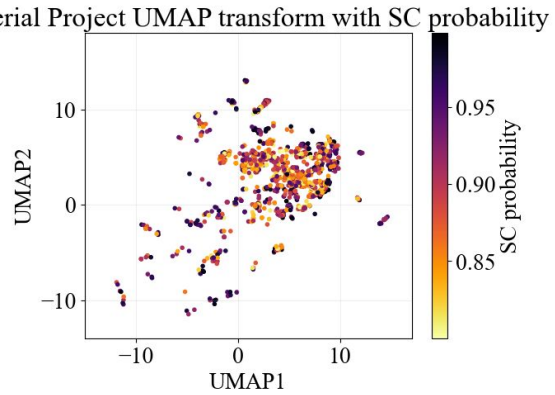


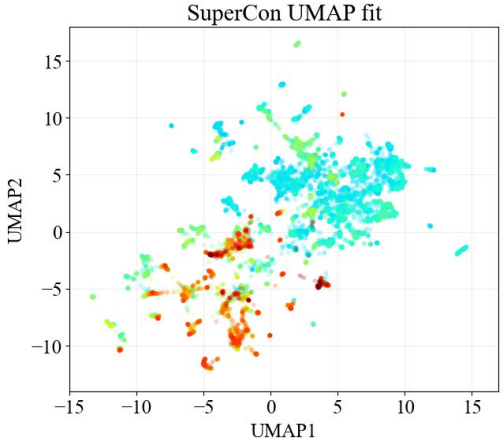
I.6. Prediction High SC candidates from Materials Project

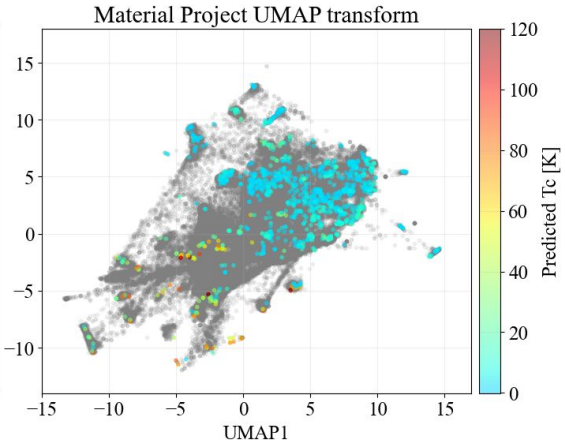
SC candidates are visualized using the SuperCon UMAP presented in main slides

MP entries with $P_{Superconductor}$>0.8 (colored by predicted Tc) plotted on top of all MP entries (grey). The SuperCon data is plotted to the left, colored by the true Tc.







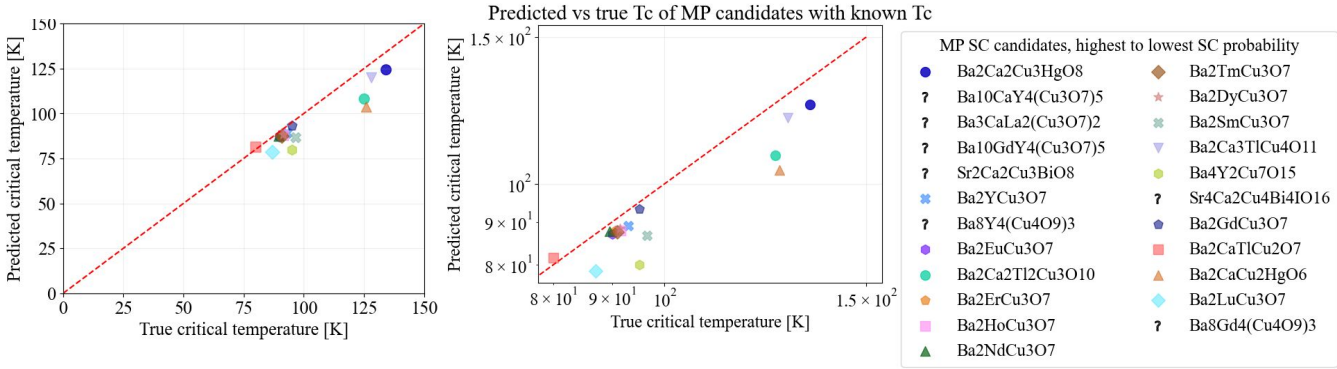


I.6. Prediction High SC candidates from Materials Project

formula	Tc_pred	SC_prob	Tc_true
Ba2Ca2Cu3HgO8	124.433788	0.986904	134.0
Ba10CaY4(Cu3O7)5	86.267455	0.985506	-1.0
Ba3CaLa2(Cu3O7)2	82.213114	0.985318	-1.0
Ba10GdY4(Cu3O7)5	83.941220	0.980621	-1.0
Sr2Ca2Cu3BiO8	101.927806	0.979355	-1.0
Ba2YCu3O7	89.068514	0.975423	93.0
Ba8Y4(Cu4O9)3	79.489492	0.972126	-1.0
Ba2EuCu3O7	87.103466	0.969430	90.0
Ba2Ca2Tl2Cu3O10	108.224302	0.968994	125.0
Ba2ErCu3O7	87.819944	0.966323	90.5
Ba2HoCu3O7	87.909878	0.966081	91.5
Ba2NdCu3O7	87.667679	0.965958	89.5
Ba2TmCu3O7	87.558059	0.958122	90.9
Ba2DyCu3O7	88.300886	0.953925	91.4
Ba2SmCu3O7	86.760238	0.952358	96.5
Ba2Ca3TlCu4O11	120.090212	0.951845	128.0
Ba4Y2Cu7O15	79.918825	0.950629	95.0
Sr4Ca2Cu4Bi4IO16	78.721598	0.945462	-1.0
Ba2GdCu3O7	93.353197	0.939499	95.0
Ba2CaTlCu2O7	81.525488	0.926054	80.0
Ba2CaCu2HgO6	103.853234	0.918371	126.0
Ba2LuCu3O7	78.600486	0.906098	87.0
Ba8Gd4(Cu4O9)3	81.621837	0.904996	-1.0

MP high Tc Superconductor candidates. None of them are in the SuperCon dataset. Finding articles (or not) about each compound lead us to be able to state a true Tc for some of the candidates. For the candidates where we could not find any Tc information, Tc_true=-1

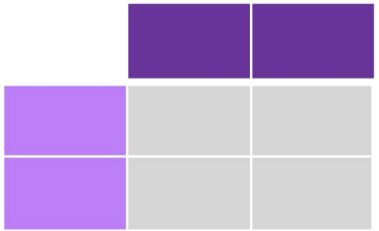
The compounds with known Tc are plotted below on lin-lin- (left), and log-log- (right) scale.



II. 3DSC

II.1.Data description

3DSC MP

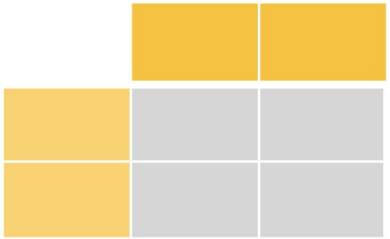


→ 76 features
+ Structure
+ Label (T_c)

↓

5 773 Entries

3DSC ICSD



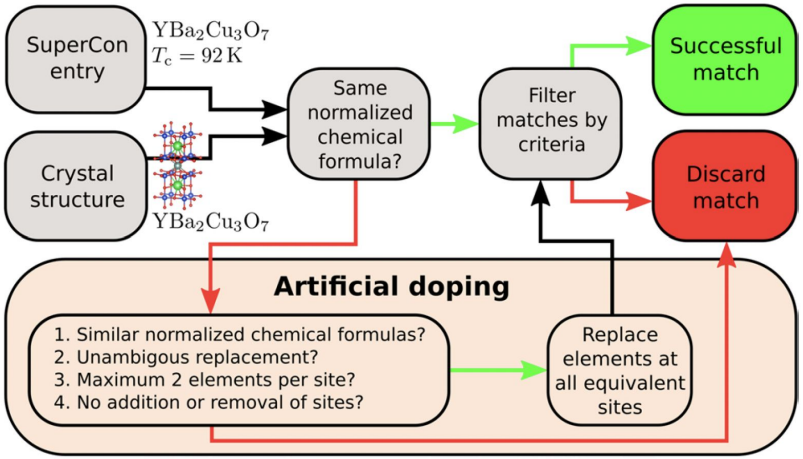
→ 63 features
+ Structure
+ Label (T_c)

↓

75 850 Entries

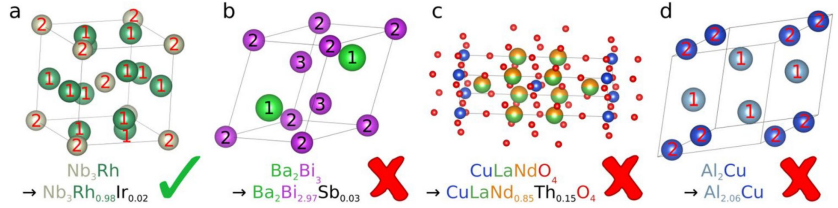
II.1. Data description

$$\Delta_{\text{totrel}} = \frac{2\sum_i |x_{\text{sc,i}} - x_{\text{cry,i}}|}{\sum_i x_{\text{sc,i}} + x_{\text{cry,i}}}.$$



Artificial Doping:

a) Everything is fine, compound is a successful match
b) Fails criterium 2 (Ambiguous replacement, symmetry different B sites)
c) Fails criterium 3 (+2 elements per site)
d) Fails criterium 4 (Addition of sites)



Example on successful match:

$CuLa_{1.95}Nd_{0.05}O_4$ with Tc — SuperCon

+ Structure of $CuLa_2O_4$ — Materials Project

Replacing 0.05 x La with Nd in CIF
"Artificial doping"

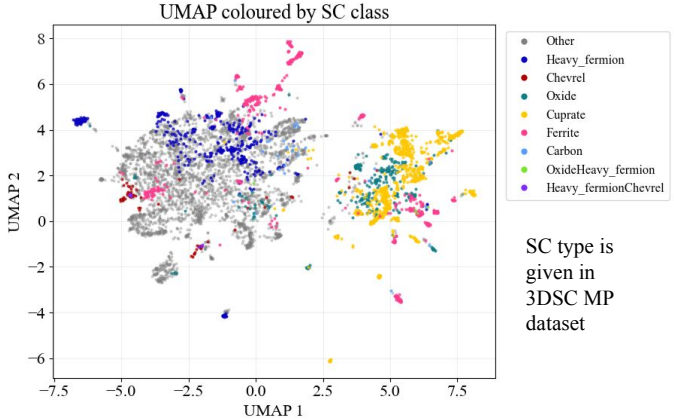
Approximated structure of $CuLa_{1.95}Nd_{0.05}O_4$ with Tc — 3DSC MP dataset

Rankring criteria:
MP: lowest E_hull + lowest Δ_{totrel}
(2 equivalent= both are kept)

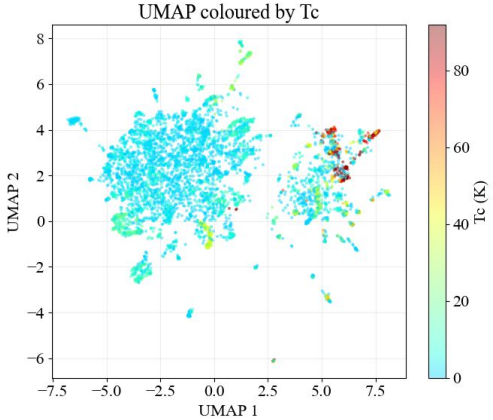
ICSD: If multiple temperature of crystal is reported, all are kept (→ Many structures)

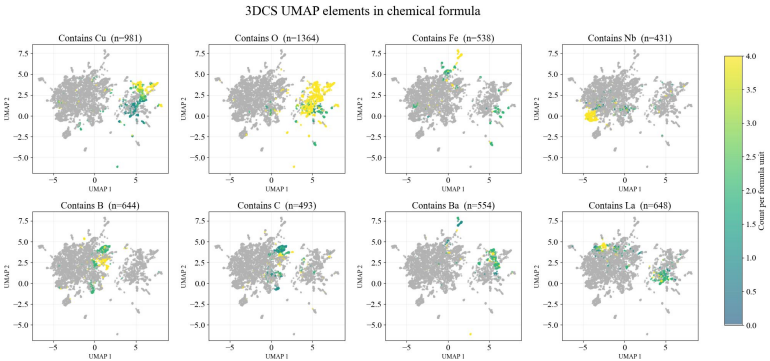
Lee et al 2025: https://iopscience.iop.org/article/10.1088/2632-2153/ae04c1

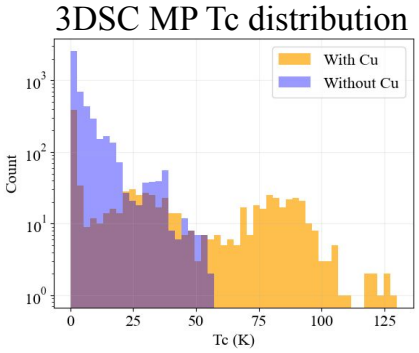
II.2. Visualization of 3DSC



SC type is given in 3DSC MP dataset

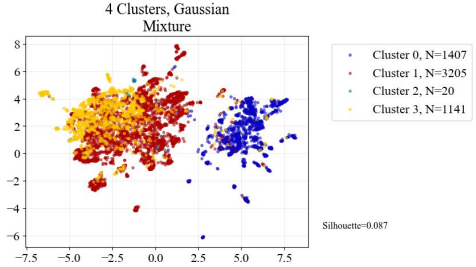


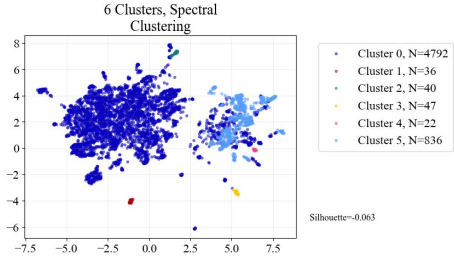


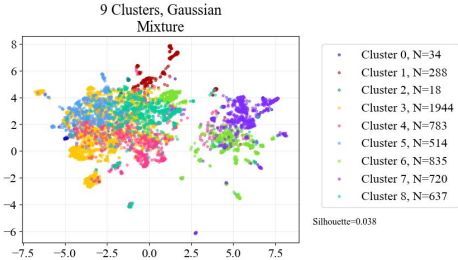


Clustering

No additional useful information was obtained from clustering, tendency to cluster after elements (Cu), could not find a 'High Tc' cluster more generally

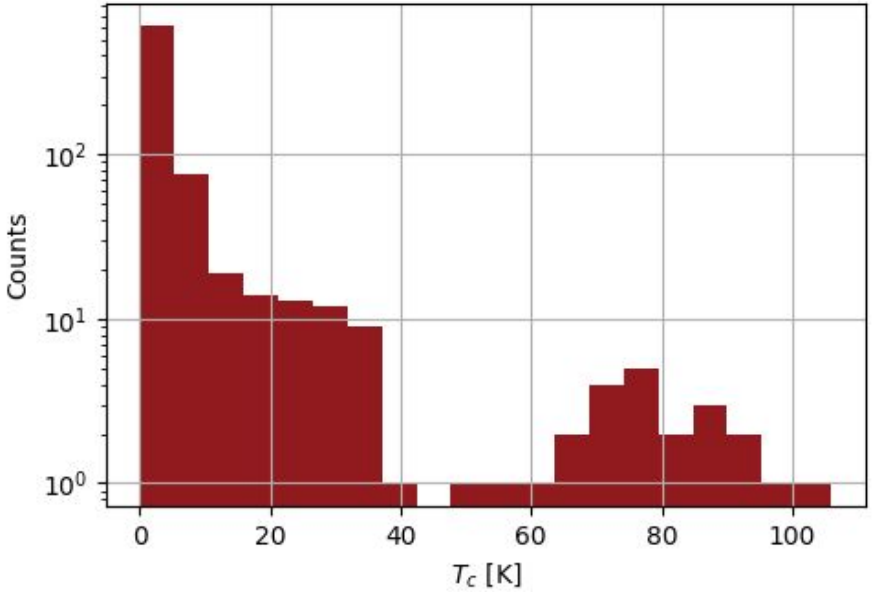




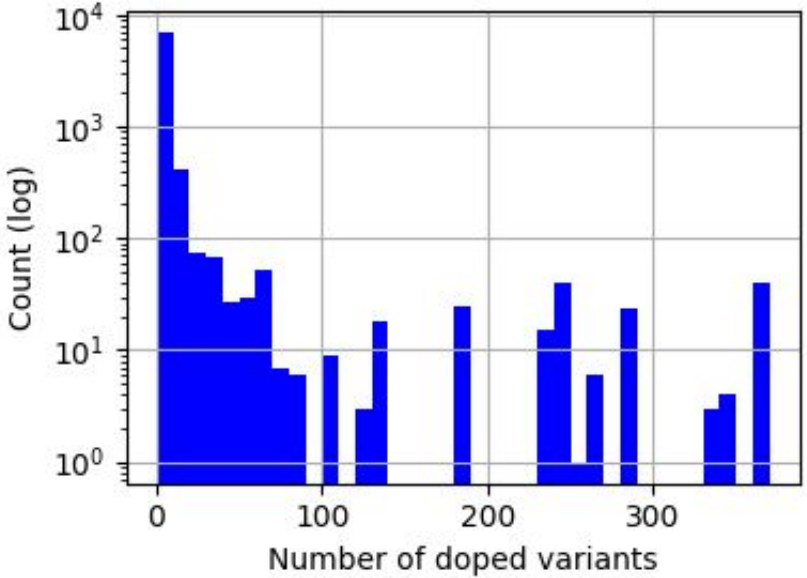


II.2. Visualization of 3DSC MP

Distribution of critical temperatures for materials found in MP but not in ICSD

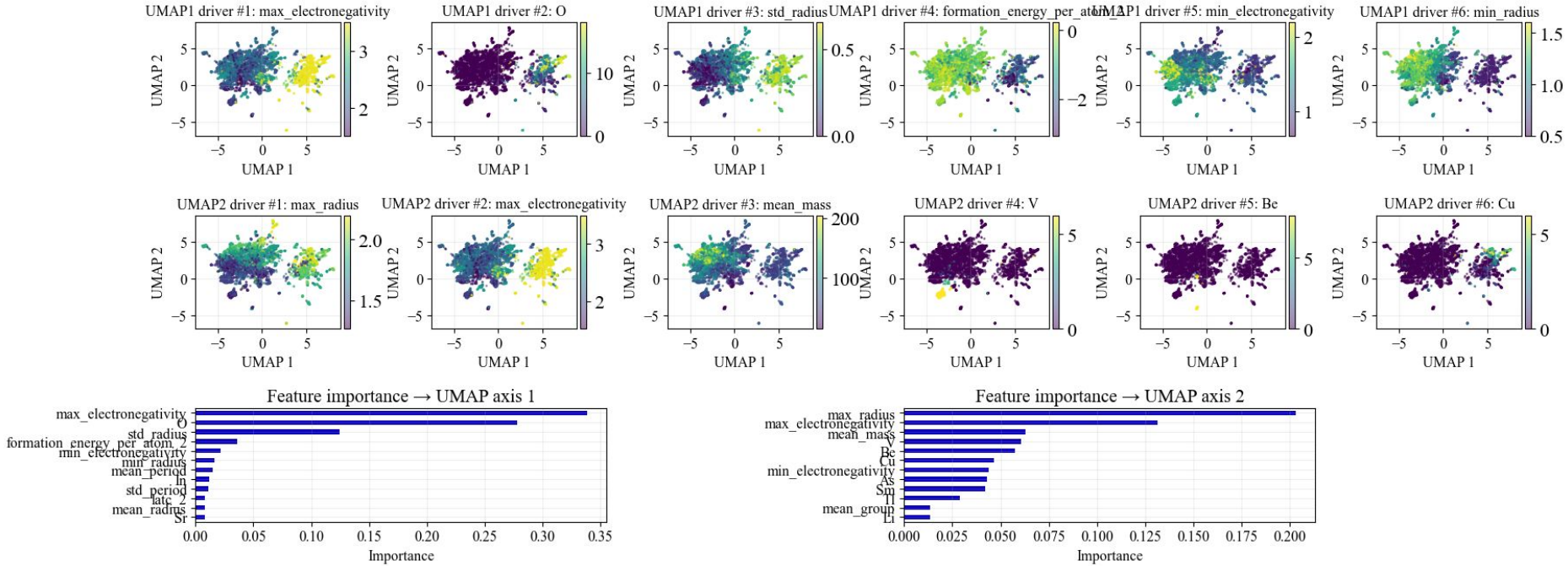


Number of artificially doped crystals per original CIF file



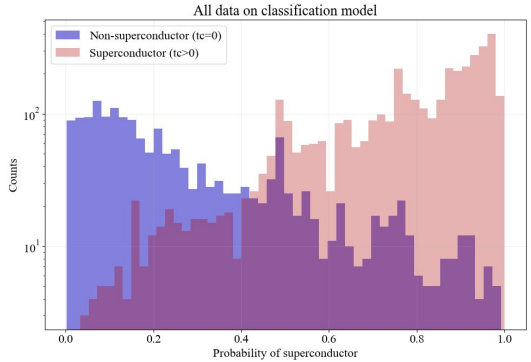
II.2. Visualization of 3DSC MP

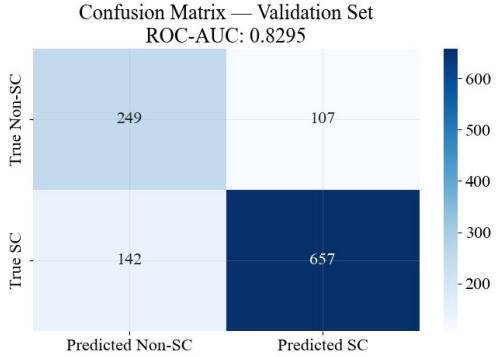
UMAP Feature importance



II.2. 3DSC MP: LGBM Classification

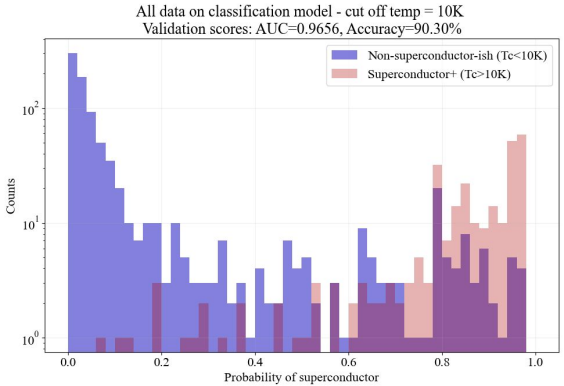
Tc= 0 vs Tc>0K:
All data is plotted, scores are calculated from held out data.

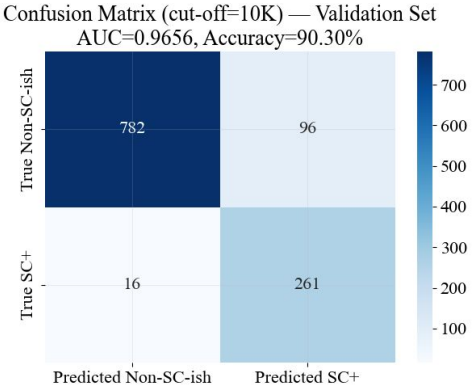




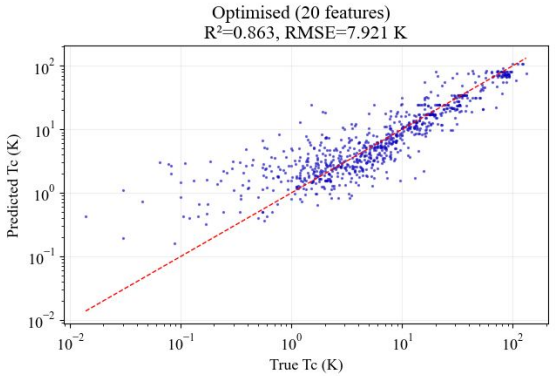
All ICSD data from here on is weighted to counteract imbalance in data augmentation

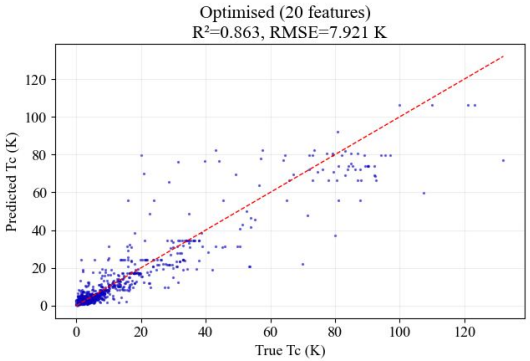
Tc < 10K vs Tc>10K:
Only held out data is plotted, scores are calculated from held out data as well.

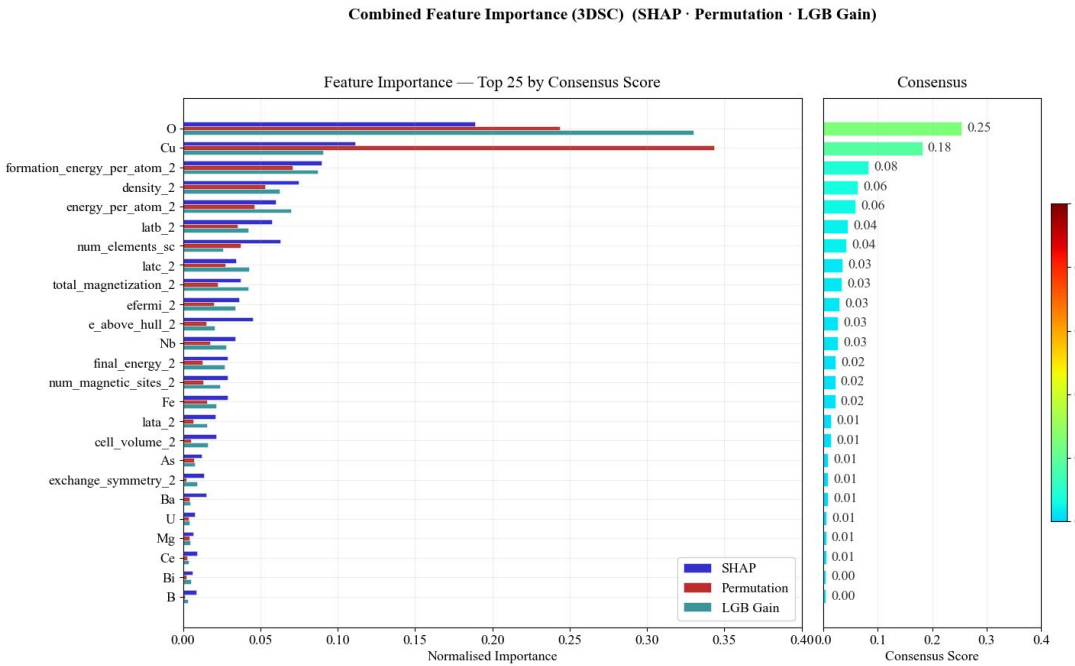




II.2. 3DSC MP: Regression



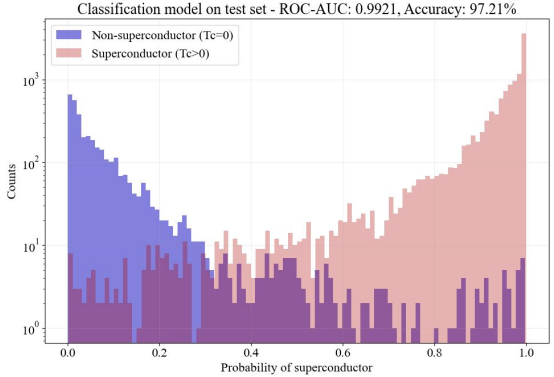


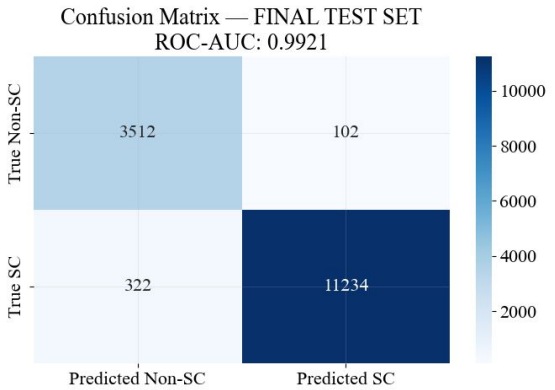


II.2. 3DSC ICSD: LGBM Classification

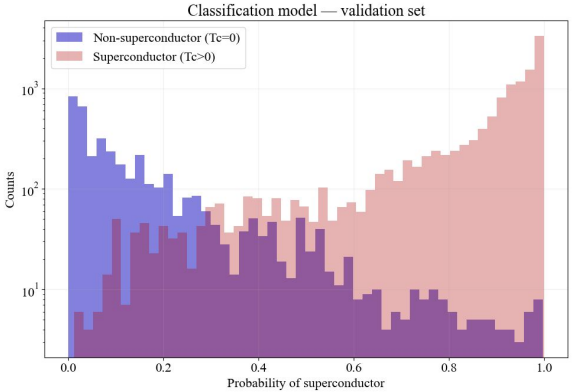
Grouped vs non grouped: On held out test set

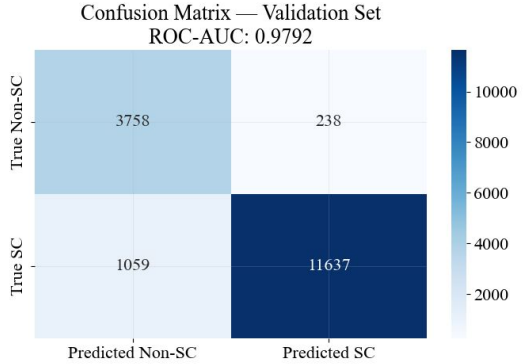
Not grouped
Tc= 0 vs Tc>0K:
Held out data is plotted and scores is calculated from this





Grouped
Tc= 0 vs Tc>0K:
Held out data is plotted and scores is calculated from this



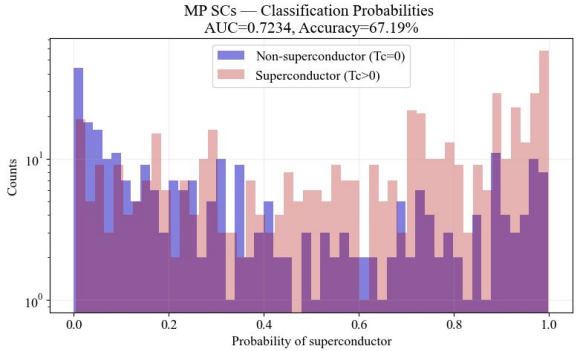


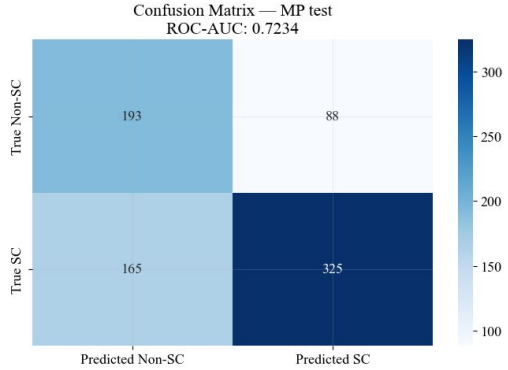
Testing whether grouping of data is important (so all artificially doped samples coming from the same original structure, is all in the test/train/validation set.)

II.2. 3DSC ICSD: LGBM Classification

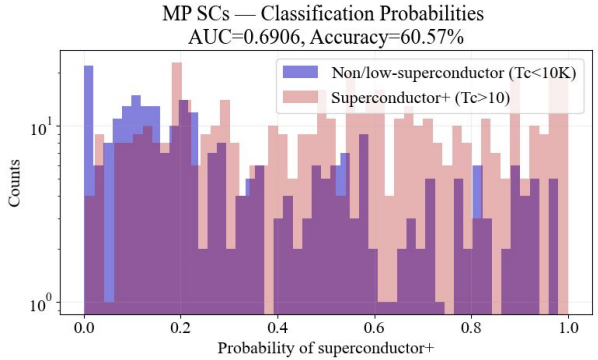
Grouped vs non grouped: On (real) MP structures (not in ICSD)

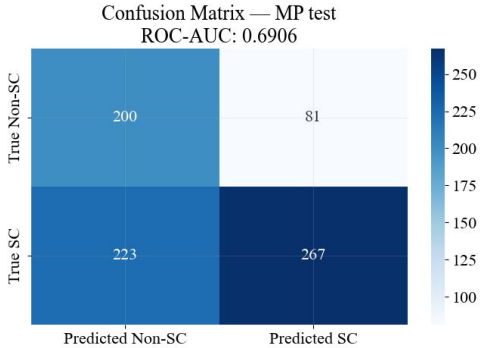
Not grouped
Tc= 0 vs Tc>0K:
Held out data is plotted and scores is calculated from this





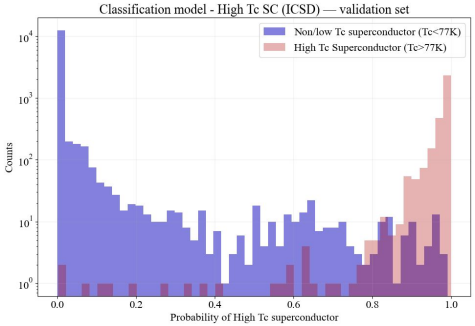
Grouped
Tc= 0 vs Tc>0K:
Held out data is plotted and scores is calculated from this

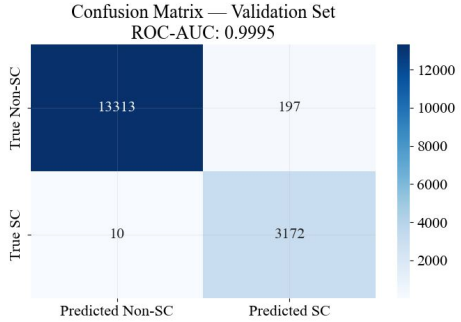




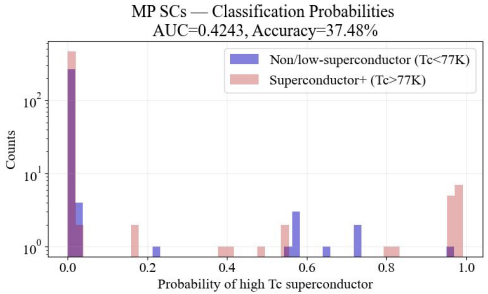
II.2. 3DSC ICSD: LGBM Classification **Grouped**

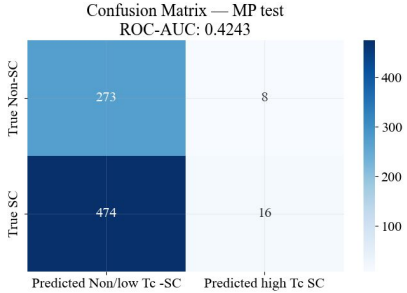
On held out ICSD data
Tc< 77K vs Tc>77K:
Held out data is plotted and scores is calculated from this





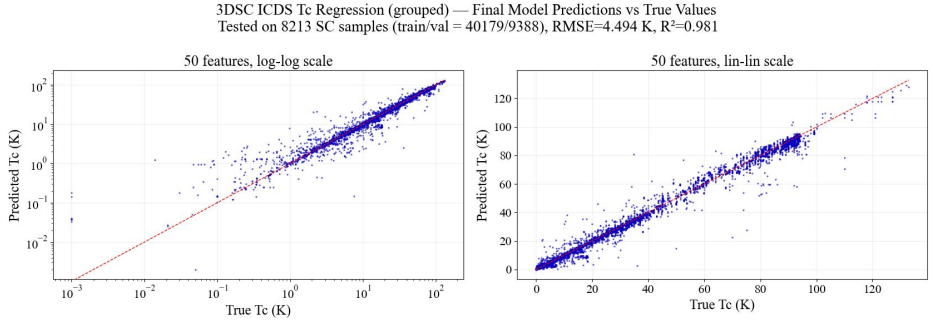
On (real) MP structures not in ICSD
Tc< 77K vs Tc>77K:



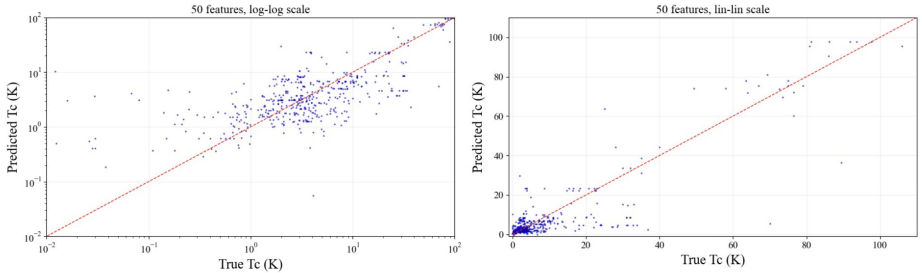


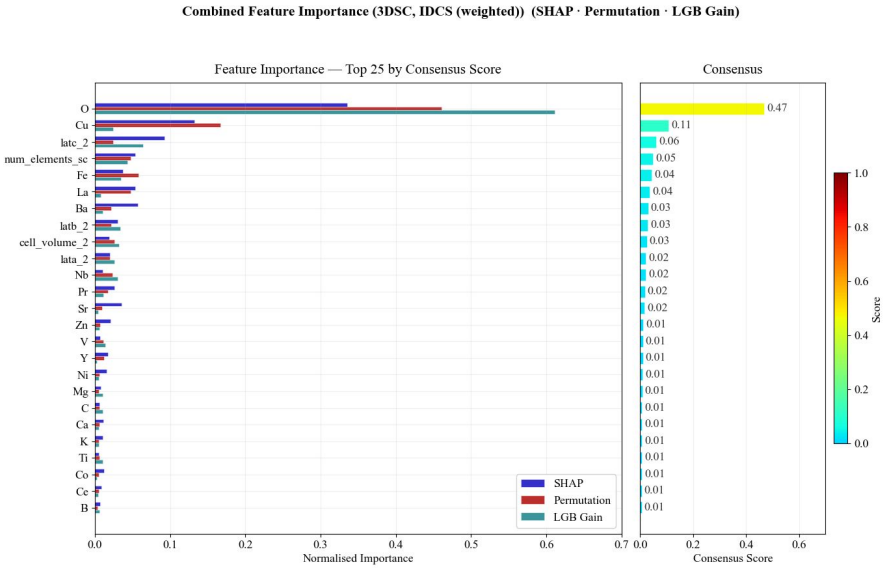
Conclusion:
Due to the very large ICSD data, the classification results are quite good when just testing on th held out data, but applying it to the MP entries, the AUC and accuracy becomes a lot worse. This is alarming, and might be a sign of overtraining (which we tries to avoid by grouping the data).

II.2. 3DSC ICSD: Regression (grouped)



On MP test:

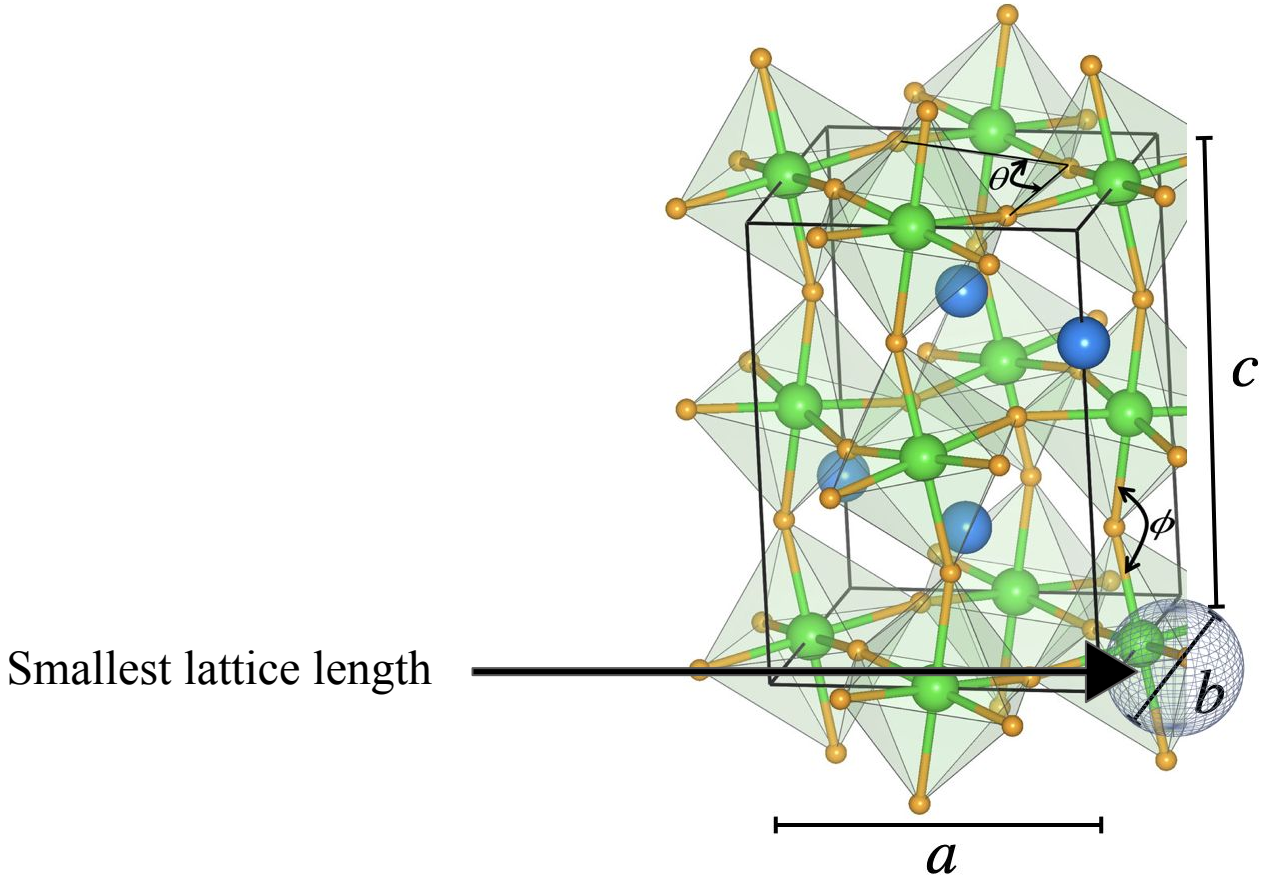




II.3. Graph presentation

3DSC neural network algorithm

1. Define the nodes
2. Define the Edges

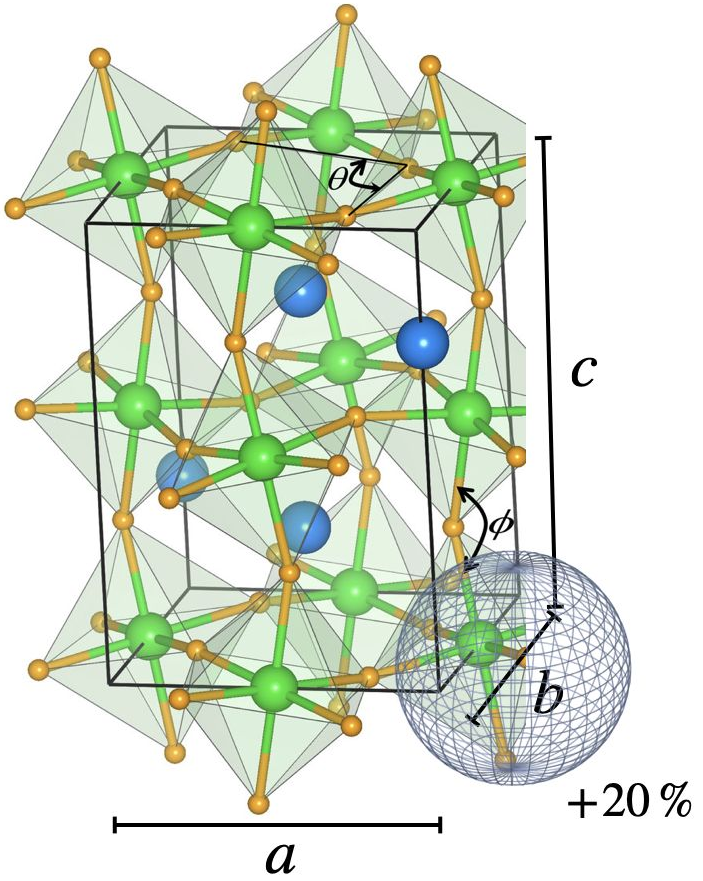


II.3. Graph presentation

3DSC neural network algorithm

1. Define the nodes

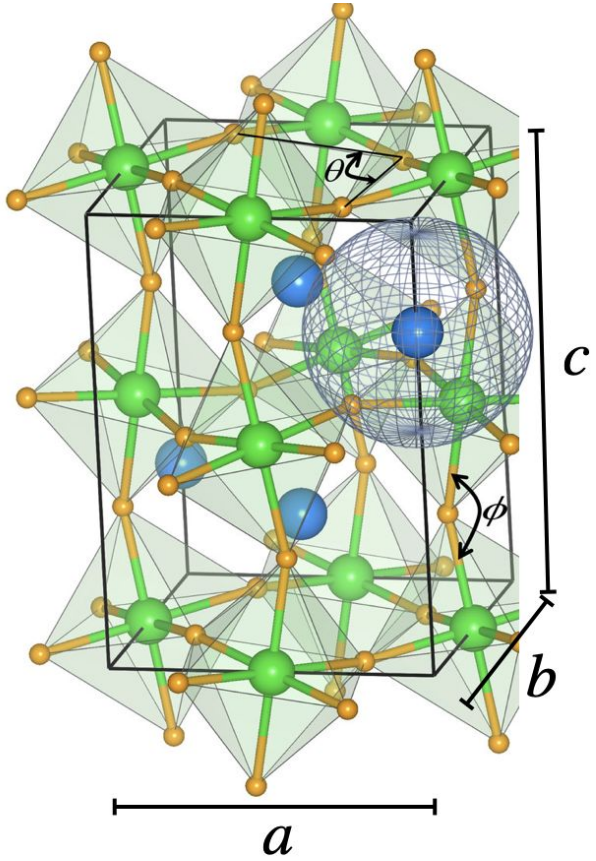
2. Define the Edges



II.3.Graph presentation

1. Define the nodes

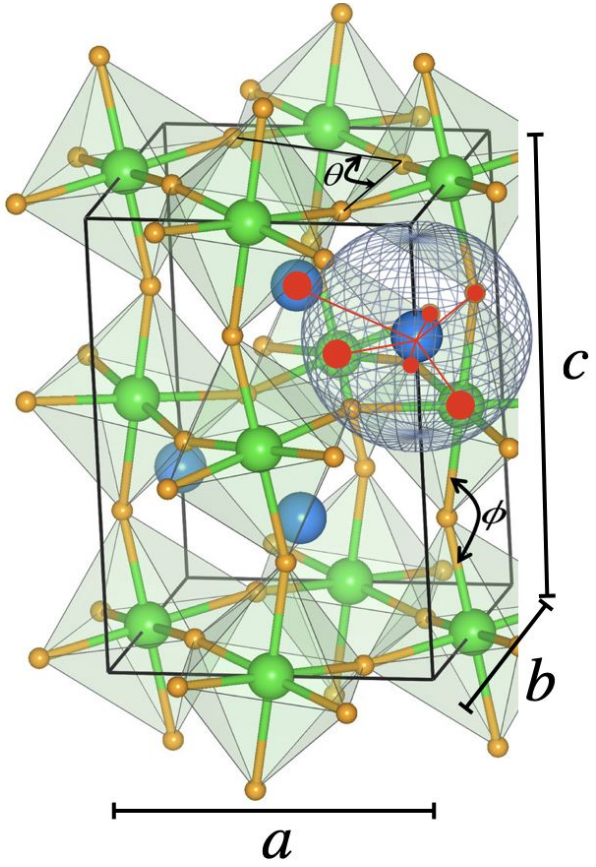
2. Define the Edges



II.3.Graph presentation

1. Define the nodes

2. Define the Edges



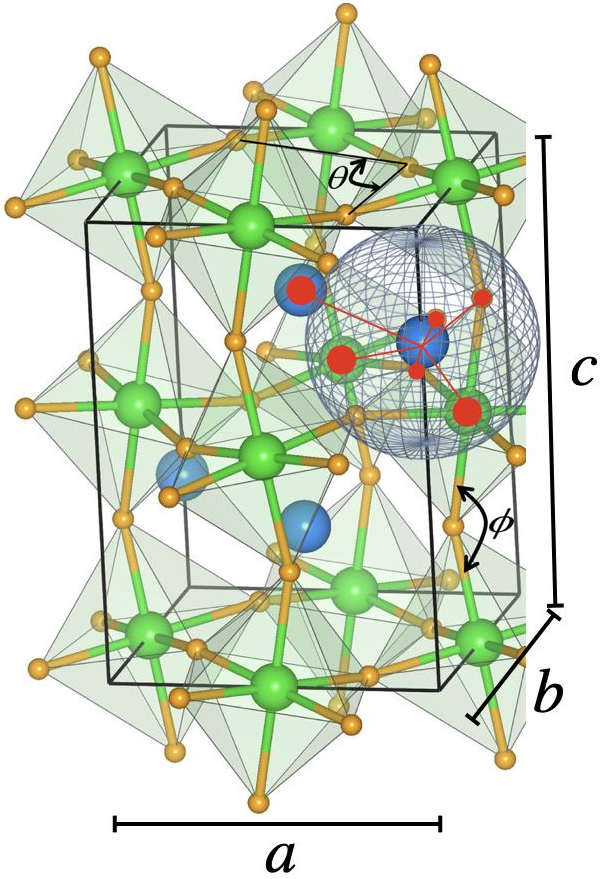
II.3. Graph presentation

3DSC neural network algorithm

1. Define the nodes

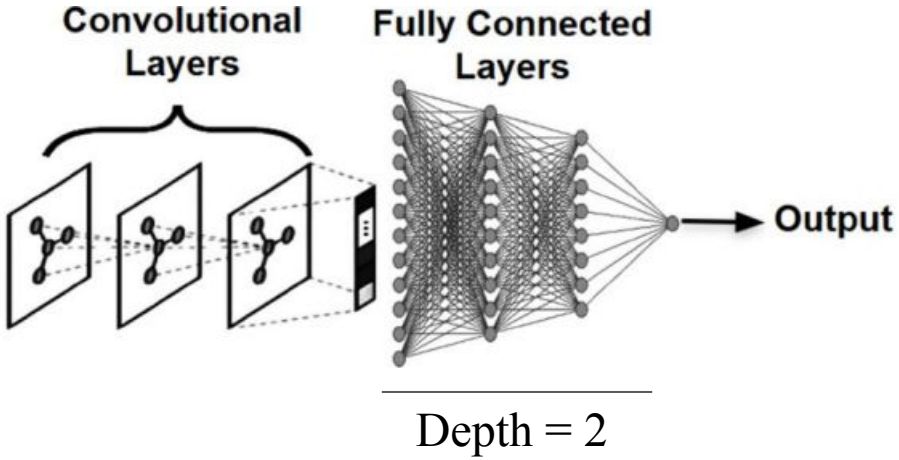
2. Define the Edges

$$d \xrightarrow{\text{Gaussian Expansion}} [tab]$$



II.3. Graph Convolutional neural network

Package: pytorch geometric



Nodes features (size n)

Linear embedding

Embedded features (size d)

Convolution (xtimes) + global mean pooling

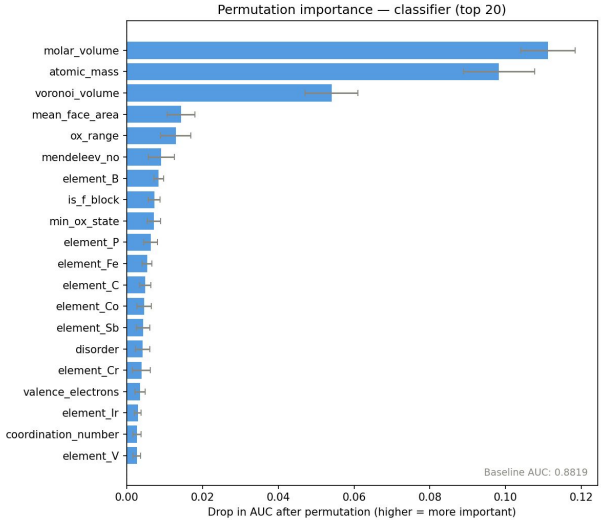
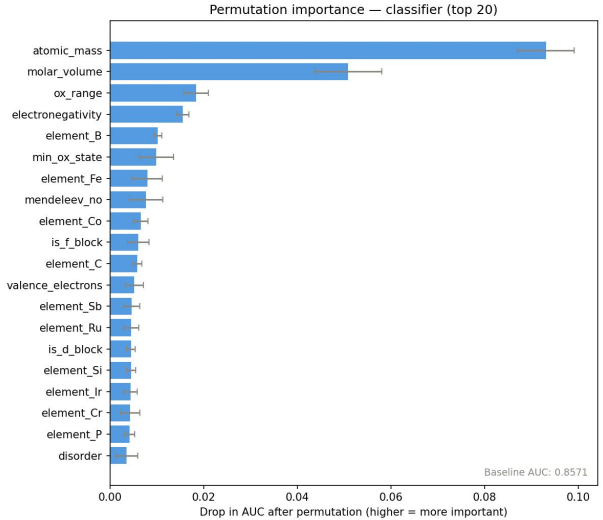
Crystal vector (size d)

Neural Network $d \rightarrow d/2 \rightarrow 1$

Critical temperature (Tc)

II.4. 3DSC MP Classification (GCNN)

Results of permutation importance. The computational cost of including the env. makes it desirable to test whether we can exclude this feature.

MP, incl. env.	MP, ex. env.
	

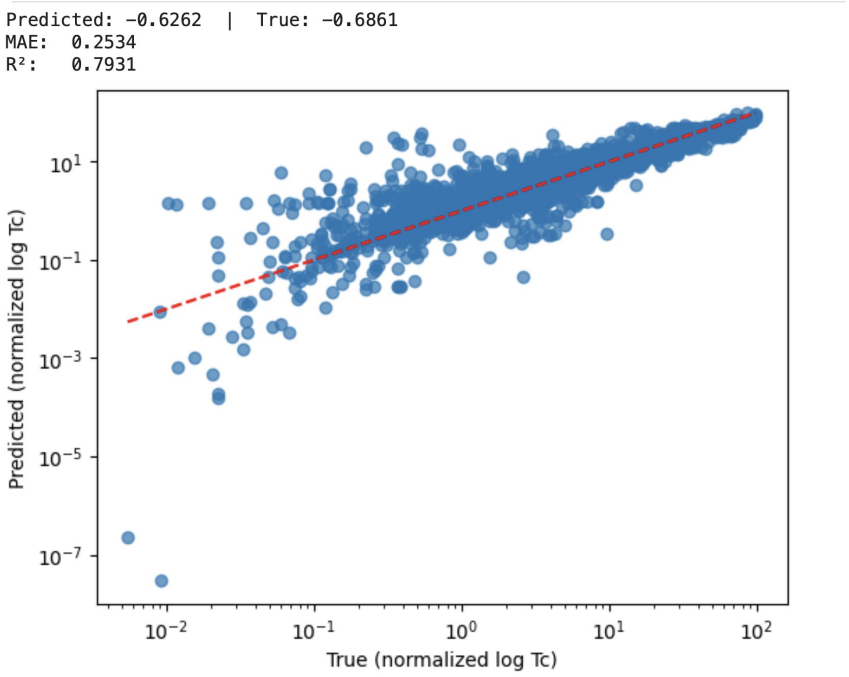
II.4. 3DSC MP Classification (GCNN)

Resulting AUC for different models

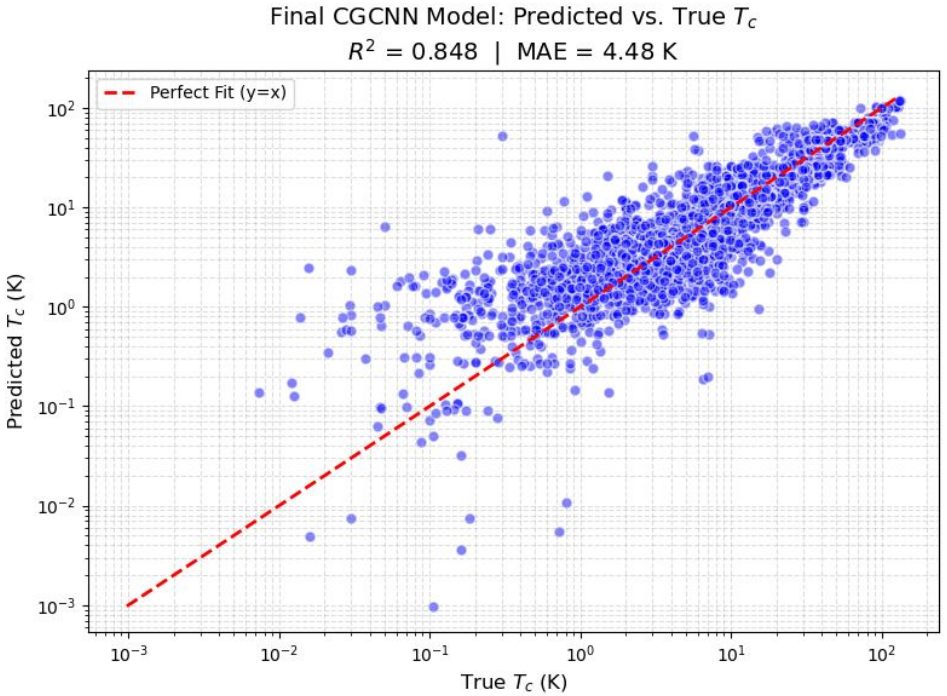
Train/Test data	AUC
Full MP (Incl. env.)	0.923
Full MP (Ex. env.)	0.916
ICSD	0.911
Train on ICSD, tested on MP	0.857

II.5. Regression: MP 3DSC (GCNN)

Small data set - we use k-fold cross validation to train our neural network. with a 10% validation set.



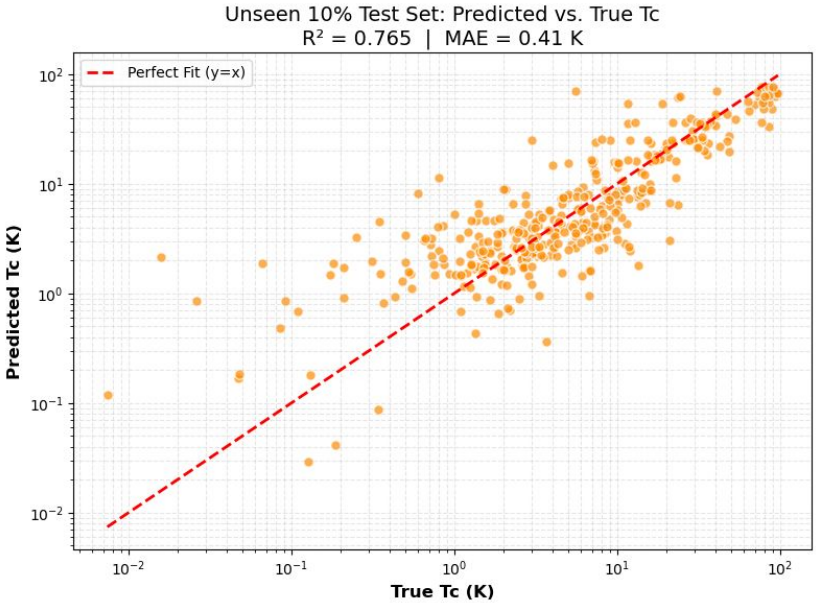
without k-fold

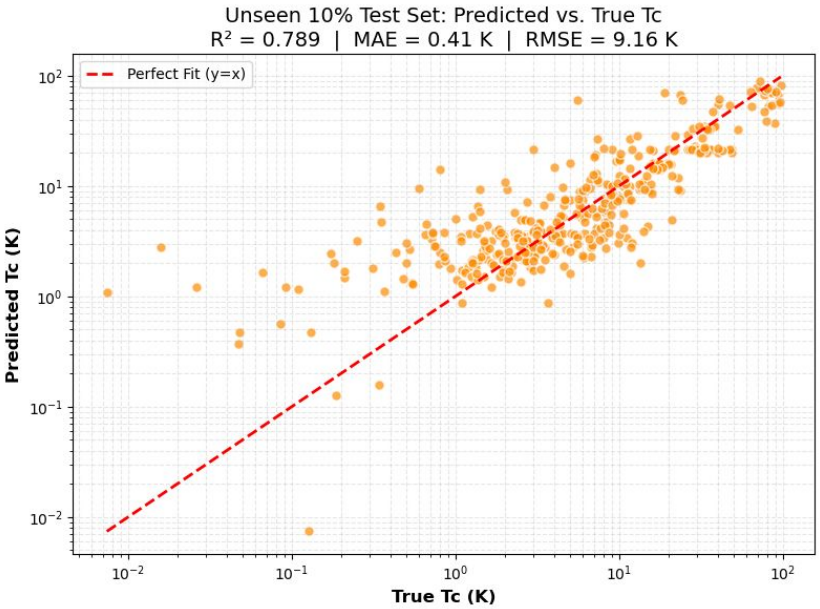


with k-fold

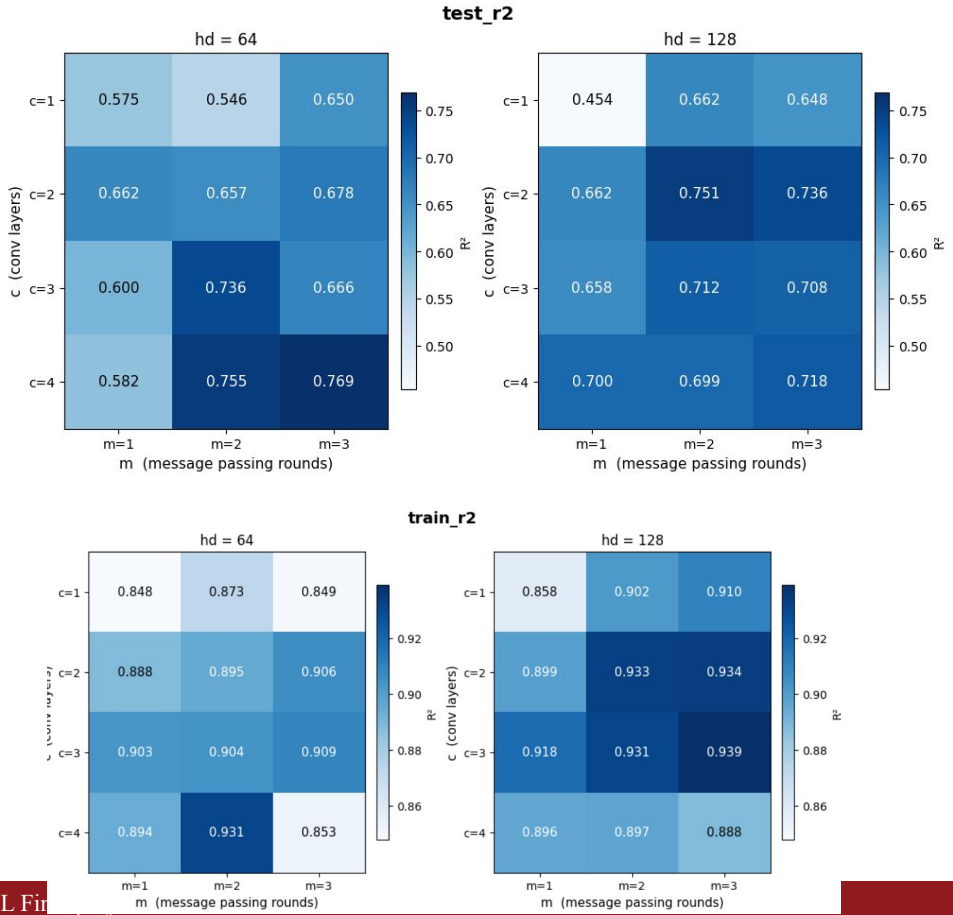
II.5. Regression (GCNN)

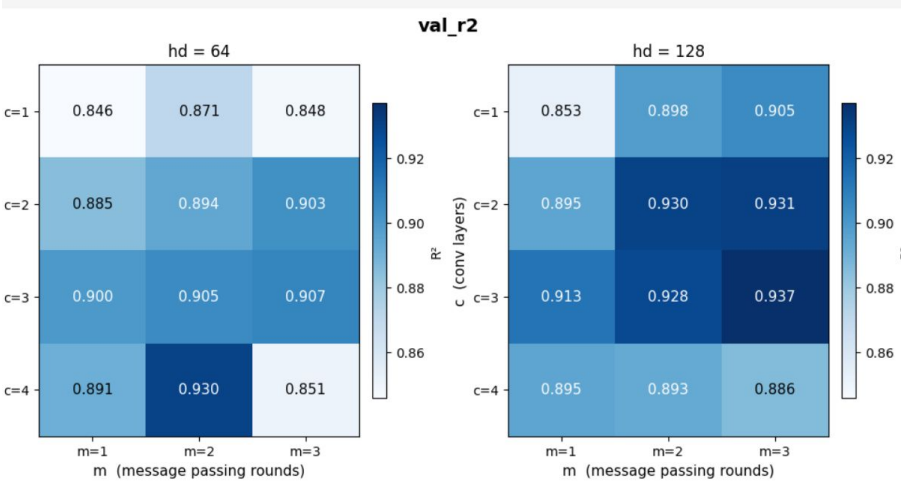
We use MatGL's pre-trained MEGNet to predict the Formation Energy for every crystal in our dataset, then we use this value in our Neural Network



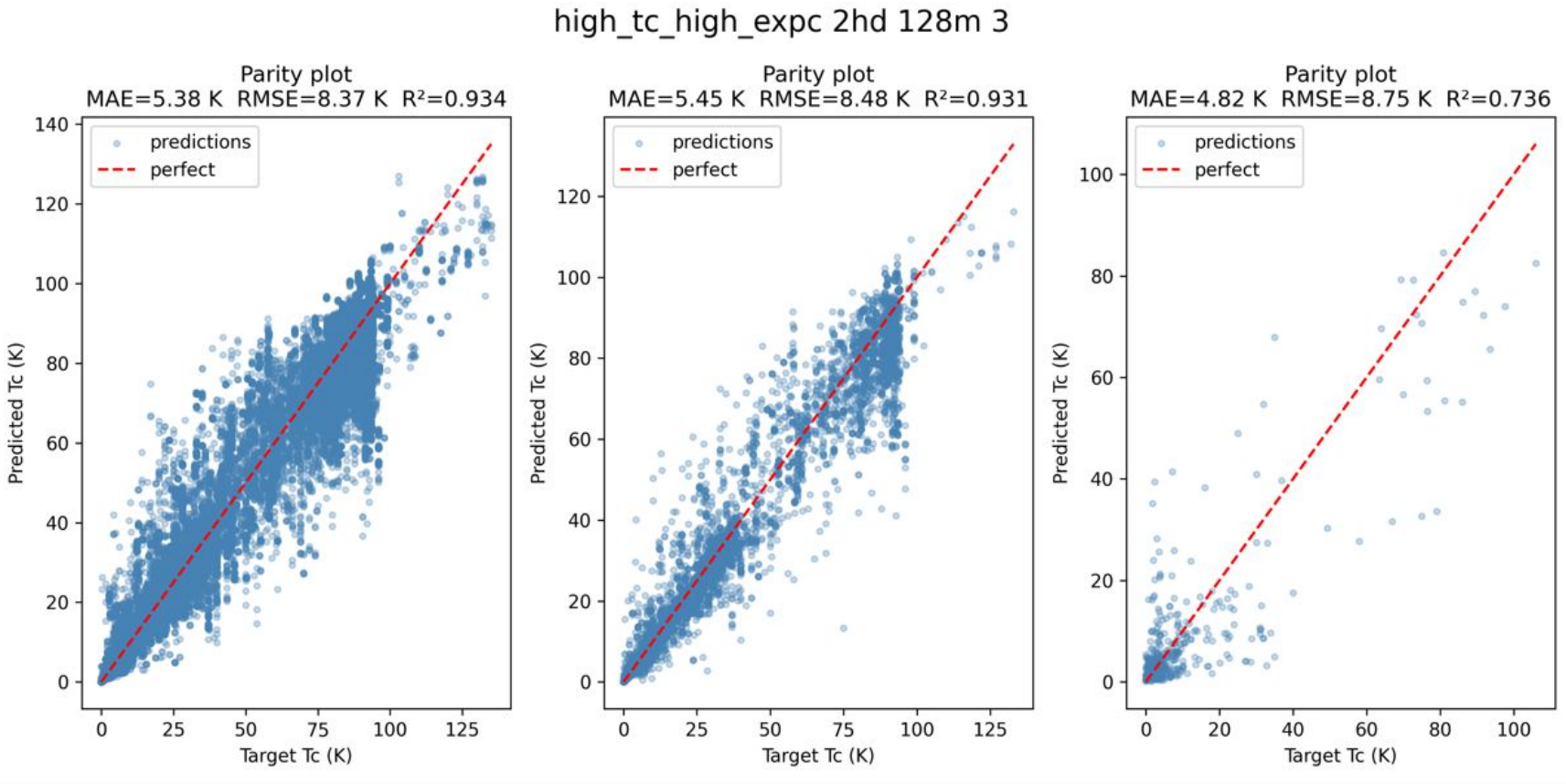


II.5. Hyper parameter optimization for upgraded model (GCNN)



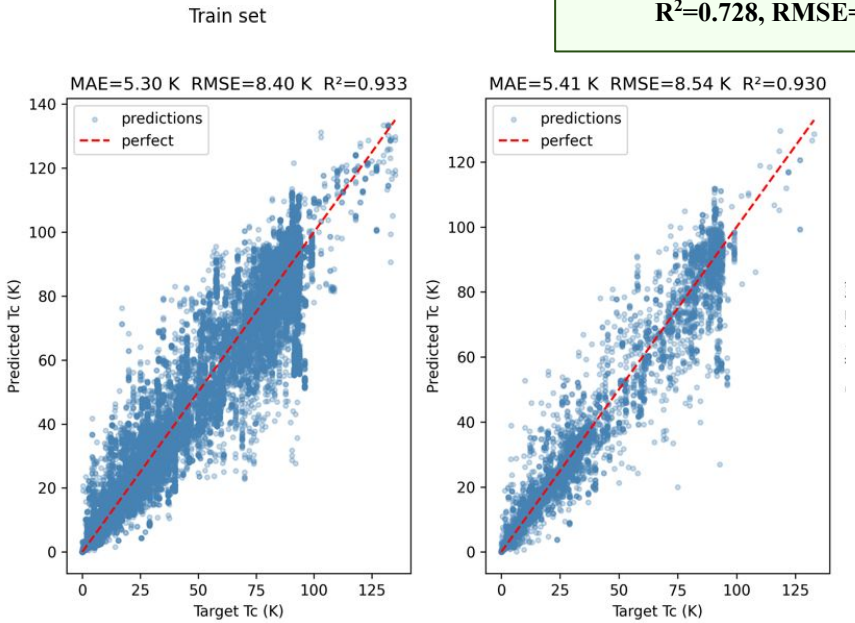


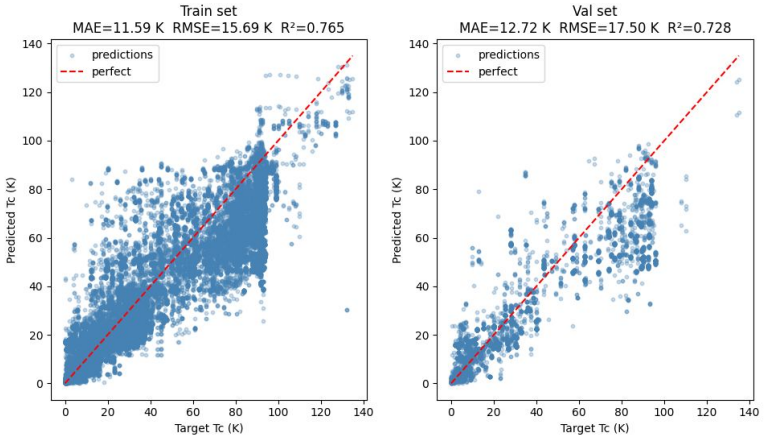
II.5. Not dividing training and validation before synthetic doping (GCNN)



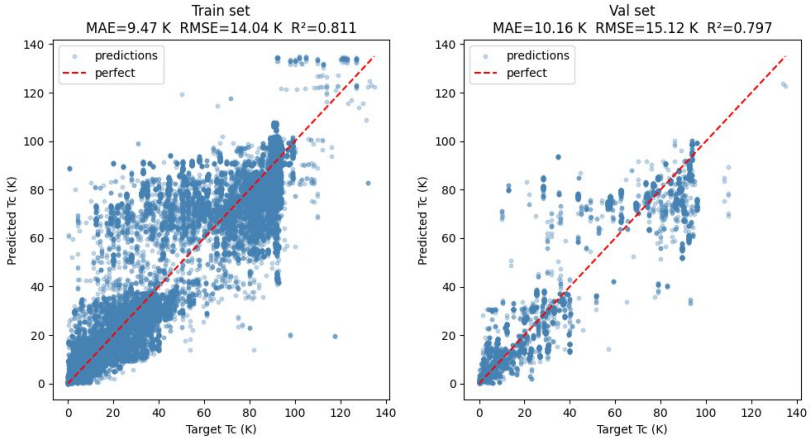
II.5. Regression (GCNN)

Regression: GCNN
ICSD data - weighted
R^2=0.728, RMSE=17.5 K



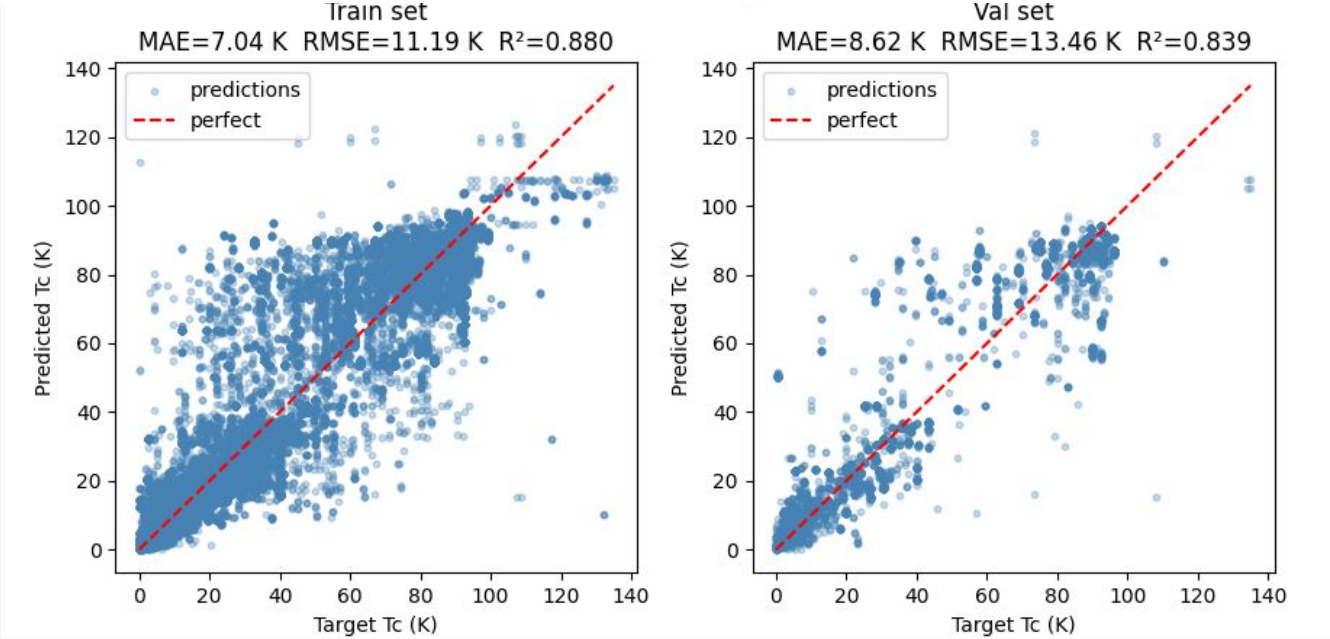


Regression: GCNN
ICSD data - not weighted
R^2=0.797, RMSE=15.12 K



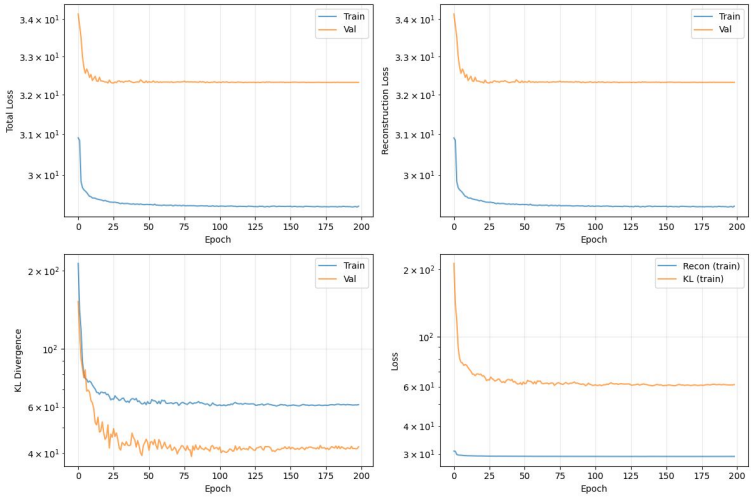
II.5. Regression (GCNN)

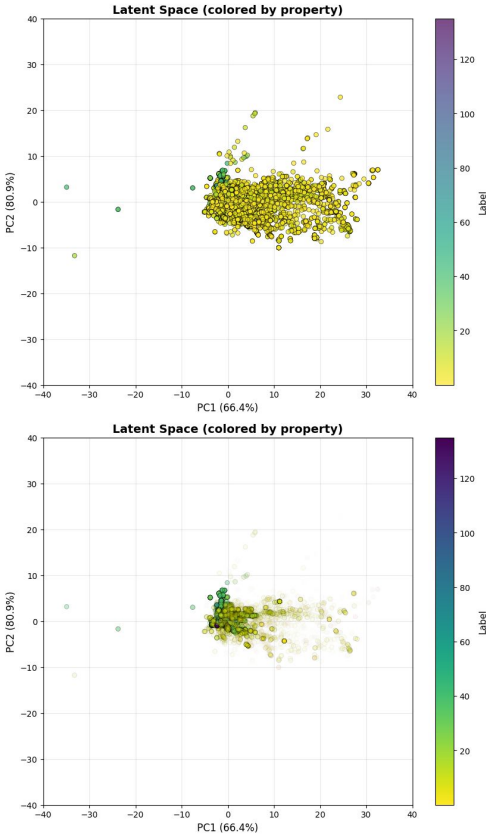
Overtrained GCNN with global attention

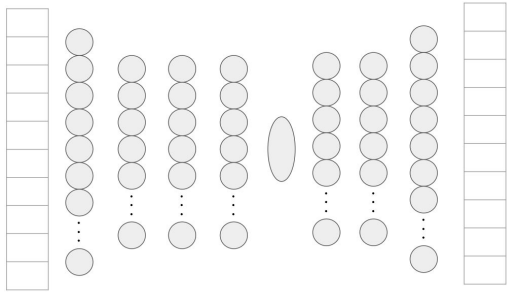


II.6. Variational Auto Encoder - ICSD

Node-level MSE: 7.7501 ± 13.4069
Median MSE: 1.4539
Max MSE: 108.9094
KL divergence: 136.37 ± 134.79







II.7. Challenges

Weighting loss $Loss = v \cdot Loss_{nodes} + w \cdot Loss_{edges} + x \cdot KL$

Encoder - Decoder difference

Graph to material

Fun