Supplementary Materials

Improved hardness prediction for reduced-activation high-entropy alloys by incorporating symbolic regression and domain adaptation on small datasets

Hao Pan[1,2,3], Mingjie Zheng[1,2,*], Xiaochen Li[1,4,*], Shijun Zhao[3,5]

[1]Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, Anhui, China.
[2]University of Science and Technology of China, Hefei 230026, Anhui, China.
[3]Department of Mechanical Engineering, City University of Hong Kong, Hong Kong 999077, China.
[4]School of Physics and Electronic Engineering, Jining University, Qufu 273155, Shandong, China.
[5]City University of Hong Kong Shenzhen Research Institute, Shenzhen 518057, Guangdong, China.

[*]**Correspondence to:** Prof. Mingjie Zheng, Hefei Institutes of Physical Science, Chinese Academy of Sciences, 350 Shushanhu Road, Hefei 230031, Anhui, China. E-mail: mingjie.zheng@inest.cas.cn; Dr. Xiaochen Li, School of Physics and Electronic Engineering, Jining University, 1 Xingtan Road, Qufu 273155, Shandong, China, E-mail: xiaochen.li@jnxy.edu.cn

1. Supplementary information on generation of element-based features

Supplementary Table 1. List of the elemental parameters for the construction of the initial features

Parameters	Symbol	Parameters	Symbol
Atomic number	Z	Melting point	T_{m}
Atomic radius	R_{a}	1st ionization energy	I_1
Covalent radius	R_{c}	Molar entropy	S_0
Valance electron concentration	VEC	Molar heat capacity at constant pressure	C_{p}
Electronegativity (Pauling)	χ_{p}	Bulk modulus	B
Electronegativity (Allred-Rochow)	χ_{ar}	Elastic modulus	E
Work function	Fw	Shear modulus	G
Molar fusion enthalpy	H_{fus}	Poisson's ratio	υ
Molar vaporization enthalpy	H_{vap}	Elastic constant	C_{ij}
Molar sublimation enthalpy	H_{atm}	Linear expansion coefficient	L_{t}
Thermal expansion coefficient	K	Cohesive energy	E_{c}

$$\mathrm{M}X = \sum_i C_i X_i \tag{S1}$$

$$\mathrm{V}X = \sum_i C_i \left(1 - \frac{X_i}{\mathrm{M}X}\right)^2 \tag{S2}$$

$$\mathrm{F}X = \max\left(C_i\left(1 - \frac{X_i}{\mathrm{M}X}\right)^2\right) - \min\left(C_i\left(1 - \frac{X_i}{\mathrm{M}X}\right)^2\right) \tag{S3}$$

$$\mathrm{D}X = \sum_i \sum_{j,j\neq i} C_i C_j \left|X_i - X_j\right| \tag{S4}$$

$$H_{\mathrm{mix}} = 4\sum_{i=1,i\neq j}^{n} C_i C_j \Delta H_{ij} \tag{S5}$$

$$S_{\mathrm{mix}} = R\sum_{i=1}^{n} C_i ln(C_i) \tag{S6}$$

$$w^6 = \left(\sum_{i=1}^{n} C_i\, Fw_i\right)^6 \tag{S7}$$

where variables C_{i} and X_{i} represent the element concentration and the property of the

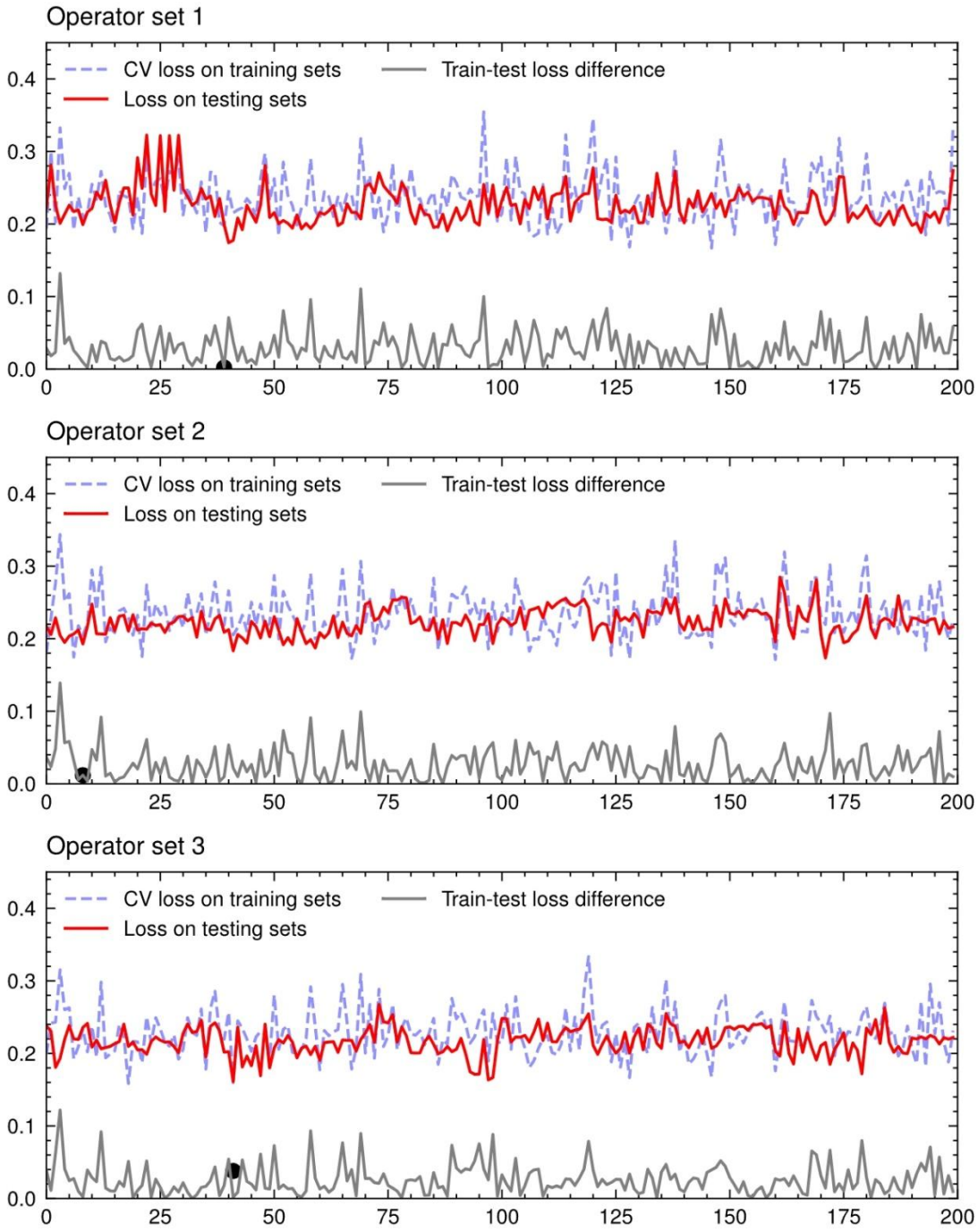
specific element, respectively. R is the gas constant, which equals to 8.3145 J/(mol·K). ΔH_{ij} refers to the enthalpy term with two interact constituent elements i and j, the values of ΔH_{ij} is from the literature [1].

2. Supplementary information on generation of symbolic regression (SR) features

Supplementary Table 2. Hyperparameters in SR

Parameter	Meaning	Value
Population size	The number of features generated in each generation	500
Generations	The maximum number of generations	50
Hall of fame	Number of features stored in the collection of best features across generations	50
Number of features	Number of features selected from feature generator	10
Function set	Mathematical operators	See the footnote*
Pc	Crossover probability	(0.6,0.9),step = 0.1
Ps	Subtree mutation probability	[(0.93-Pc)/3, (1-Pc)/3] step = 0.01
Ph	Hoist mutation probability	Ps
Pp	Point mutation probability	1-Pc-Ps-Ph
Depth	Depth of tree	[2,5]
Parsimony	Penalty factor for the complex features	0.001

*Function sets include #1: [add($+$), sub($-$), mul($\times$), div($\div$), inv($1/x$), log($\log x$), exp($\exp x$), min($\min(x_1, x_2)$), max($\max(x_1, x_2)$), sqrt($\sqrt{x}$)], #2: [add($+$), sub($-$), mul($\times$), div($\div$)] and #3: [add($+$), sub($-$), mul($\times$), div($\div$), inv($1/x$), log($\log x$), exp($\exp x$), sqrt($\sqrt{x}$)]



Supplementary Figure 1. Cross-validation and testing loss of ridge regression model using features from the best SR feature generator in each iteration. Black dots marked the selected SR generator using different operator sets in Supplementary Table 2.

3. Supplementary information on construction of machine learning (ML) models.

Supplementary Table 3. The search range of hyperparameters for ML models

Algorithms	Hyperparameters
DTR	max_depth = [None, 5, 10, 20],min_samples_split = [2, 5, 10], min_samples_leaf = [1, 2, 4, 8], max_leaf_nodes = [None, 10, 20]
ETR	n_estimators = [50, 100, 200], max_depth = [None, 10, 20], , max_leaf_nodes = [None, 10, 20] min_samples_split = [2, 5, 10], min_samples_leaf = [1, 2, 4, 8]
GBR	n_estimators = [10, 20, 50, 100, 200, 300, 500], max_leaf_nodes = [None, 10, 20], learning_rate = [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}], max_depth = [None, 3, 5, 7, 9, 11], min_samples_split = [2, 5, 10], min_samples_leaf = [1, 2, 4]
KNR	n_neighbors = [3, 5, 7], weights = [‘uniform’, ‘distance’]
RFR	n_estimators = [10, 20, 50, 100, 200, 300, 500], max_depth = [None, 5, 10, 20], min_samples_split = [2, 5, 10], min_samples_leaf = [1, 2, 4, 8], max_leaf_nodes = [None, 10, 20],
Ridge	alpha = [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10]
SVR	C = [0.1, 1, 10], gamma = [‘scale’, ‘auto’], kernel = [‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’], degree = [2, 3, 4], epsilon = [0.1, 0.2, 0.5]

Supplementary Table 4. Hyperparameters of best ML models of each feature set

Feature subset	ML algorithm	Hyperparameters
NoGPF	ETR	max_depth=None, max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2, n_estimators=100,
GPFOnly	RFR	max_depth=10, max_leaf_nodes=20, min_samples_split=2, min_samples_leaf=1, n_estimators=50,
GPFCombined	GBR	max_depth=10, max_leaf_nodes=20, min_samples_leaf=2, min_samples_split=10, n_estimators=200,

REFERENCES

[1] A. Takeuchi, A. Inoue, Mater. Trans. 46 (2005) 2817–2829.