Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supporting Information

Optimal Machine Learning Feature Selection for Assessing the Mechanical Properties of a Zeolite Framework

Namjung Kim[1,*] and Kyoungmin Min[2,*]

[1]Department of Mechanical Engineering, Gachon University, 1342 Seongnamdaero, Sujeong-gu, Seongnam, Gyeonggi-do 13120, Republic of Korea

[2]School of Mechanical Engineering, Soongsil University, 369 Sangdo-ro, Sangdo-dong, Dongjak-gu, Seoul 06978, Republic of Korea

[*]Corresponding author: namjungk@gachon.ac.kr (N. Kim), kmin.min@ssu.ac.kr (K. Min)

Featurizer	Number of features	Prediction accuracy (MAE)	
		K_{VRH} (GPa)	G_{VRH} (GPa)
M (mean)	896	13.88 ± 2.65	11.00 ± 6.15
M (median)	896	13.71 ± 2.59	11.01 ± 6.14
M (zero)	896	13.62 ± 2.51	10.76 ± 6.24
Z + M(zero)	928	11.60 ± 2.57	10.47 ± 6.20

Table S1. Prediction accuracy change with respect to the different configurations of features.

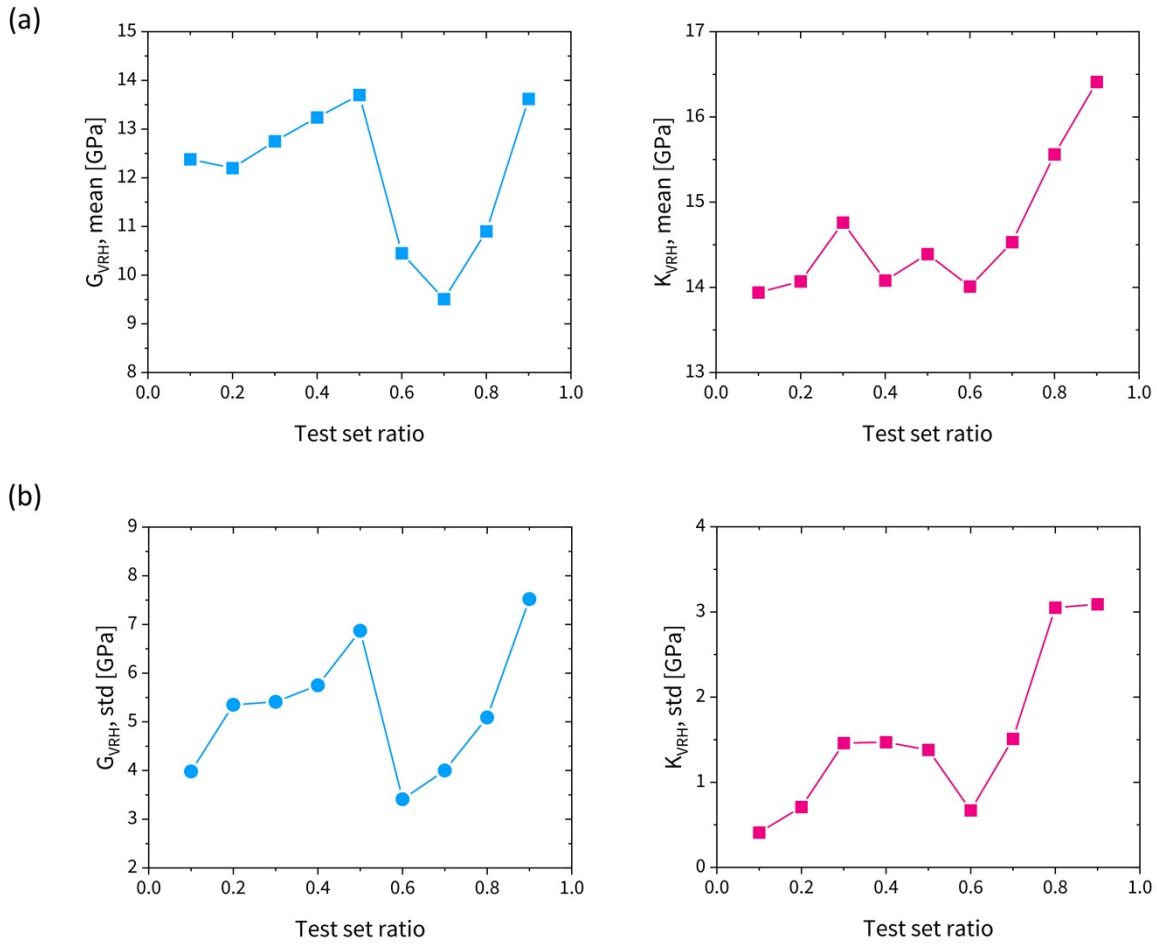

Figure S1. Test set ratio dependent prediction accuracy and the uncertainty change using Z descriptors

Featurizer	# of features	
	Available	Chosen
Density	3	3
GlobalSymmetryFeatures	1	0
ColumbMatrix	12	3
XRDPowderPattern	96	13
RadialDistributionFunction	50	1
AGNIFingerprints	53	5
AverageBondAngle	3	1
ChemEnvSiteFingerprint	6	3
CoordinationNumber	2	1
CrystalNNFingerprint	16	2
GaussianSymmFunc	16	0
GeneralizedRDF	20	4
OPSiteFingerprint	45	6
Zeolite	32	3
Sum	356	45

Table S2. The number of initial and the chosen features (45) whose feature importance is larger than zero for prediction of G_{VRH}.

Featurizer	# of features	
	Available	Chosen
Density	3	0
GlobalSymmetryFeatures	1	1
ColumbMatrix	12	10
XRDPowderPattern	96	72
RadialDistributionFunction	50	25
AGNIFingerprints	53	39
AverageBondAngle	3	2
ChemEnvSiteFingerprint	6	5
CoordinationNumber	2	2
CrystalNNFingerprint	16	11
GaussianSymmFunc	16	9
GeneralizedRDF	20	12
OPSiteFingerprint	45	37
Zeolite	32	24
Sum	356	249

Table S3. The number of initial and the chosen features (249) whose feature importance is larger than zero for prediction of K_{VRH}.

Feature	Feature Importance		Sum
	G_{VRH}	K_{VRH}	
ChemEnvSiteFingerprintmeanA2	1.00	0.79	1.79
CoordinationNumbermeanCN_VoronoiNN	0.33	1.00	1.33
Zeolite_density	0.56	0.64	1.20
Zeolite_largest_included_sphere_free	0.11	0.93	1.04
XRDPowderPatternxrd_105	0.56	0.43	0.98
OPSiteFingerprintstd_devsquarecoplanarCN_4	0.33	0.57	0.90
AGNIFingerPrintstd_devAGNIdirxeta289e00	0.44	0.36	0.80
CoulombMatrixcoulombmatrixeig3	0.22	0.57	0.79
Zeolite_SiOSi_hmean	0.22	0.57	0.79
OPSiteFingerprintmeanq4CN_11	0.22	0.50	0.72
OPSiteFingerprintmeanq4CN_10	0.11	0.57	0.68
XRDPowderPatternxrd_85	0.33	0.21	0.55
XRDPowderPatternxrd_37	0.11	0.43	0.54
OPSiteFingerprintstd_devq6CN_9	0.22	0.29	0.51
CrystalNNFingerprintmeanwtCN_1	0.11	0.36	0.47
AGNIFingerPrintmeanAGNIdirzeta104e01	0.22	0.21	0.44
GeneralizedRDFstd_devGaussiancenter70width10	0.22	0.21	0.44
AGNIFingerPrintstd_devAGNIdirzeta680e00	0.33	0.07	0.40
CoulombMatrixcoulombmatrixeig1	0.33	0.07	0.40
XRDPowderPatternxrd_77	0.22	0.14	0.37
XRDPowderPatternxrd_64	0.22	0.14	0.37
GeneralizedRDFstd_devGaussiancenter60width10	0.22	0.14	0.37
AGNIFingerPrintmeanAGNIdirzeta123e00	0.22	0.07	0.29
GeneralizedRDFmeanGaussiancenter50width10	0.22	0.07	0.29
OPSiteFingerprintmeantetrahedralCN_4	0.11	0.14	0.25
XRDPowderPatternxrd_82	0.11	0.14	0.25
CoulombMatrixcoulombmatrixeig0	0.11	0.07	0.18
OPSiteFingerprintmeanbent120degreesCN_2	0.11	0.07	0.18
XRDPowderPatternxrd_103	0.11	0.07	0.18
ChemEnvSiteFingerprintstd_devT4	0.11	0.07	0.18
RadialDistributionFunctionradialdistributionfunctiond_490	0.11	0.07	0.18
XRDPowderPatternxrd_108	0.11	0.07	0.18
XRDPowderPatternxrd_80	0.11	0.07	0.18

Table S4. Summation of feature importance whose value is larger than zero for predicting both of G_{VRH} and K_{VRH}.

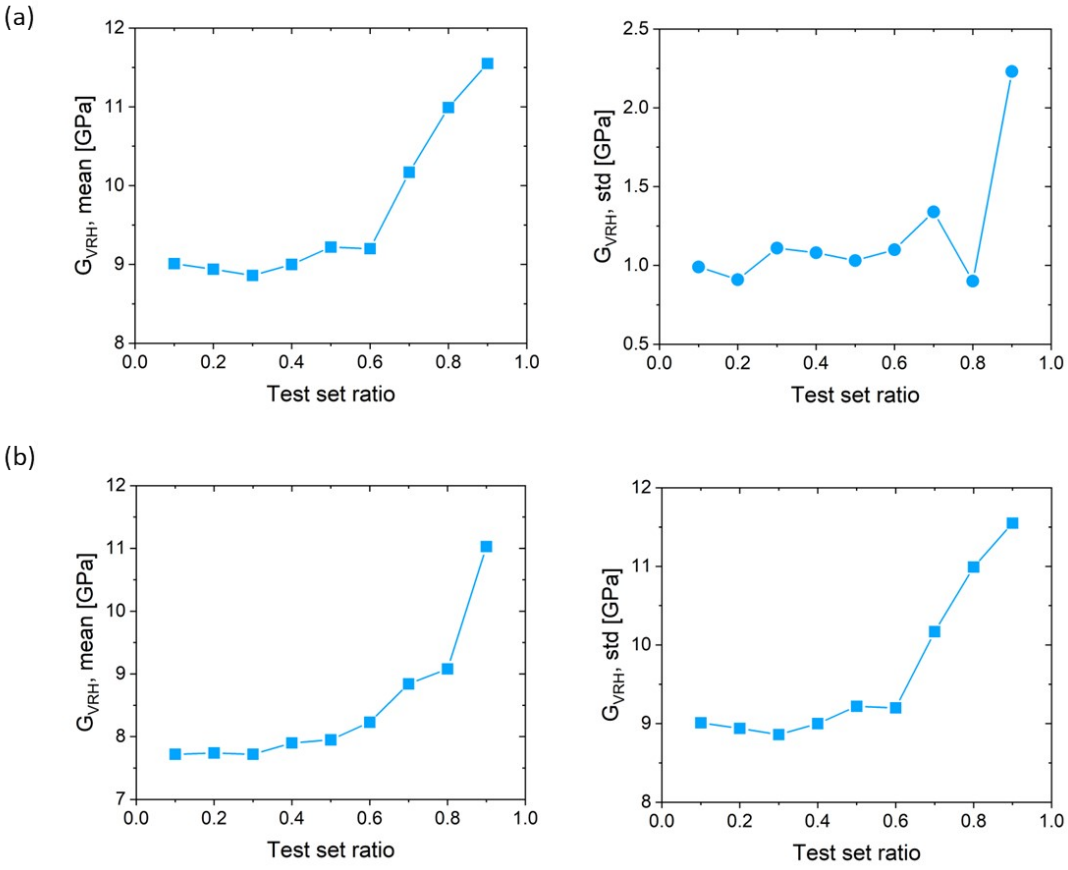

Figure S2. Test set ratio dependent prediction accuracy and uncertainty change using Z descriptors (left column) and Z+M descriptors (right column) without abnormal data