Electronic Supplementary Material (ESI) for Journal of Materials Chemistry A.

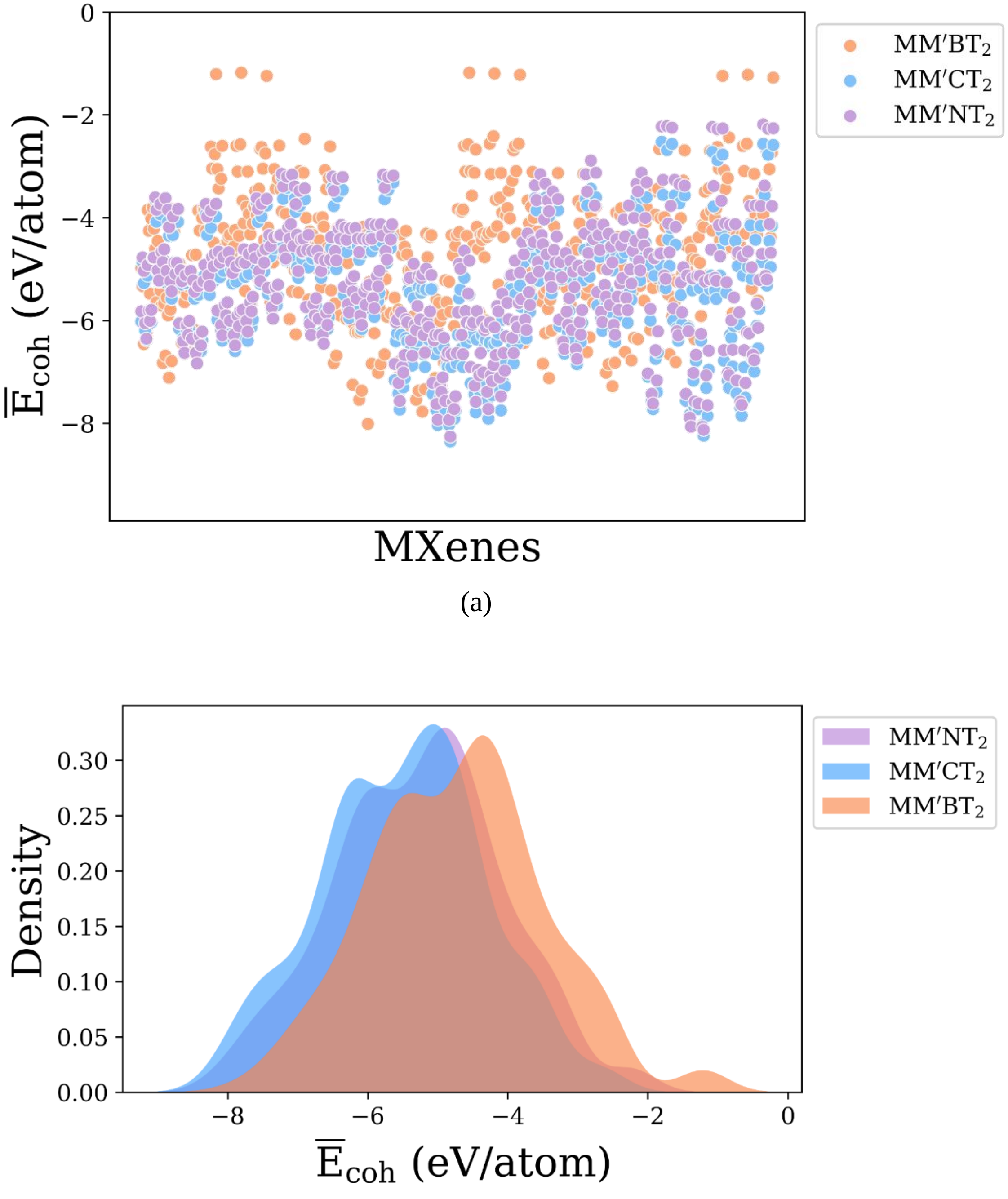
Supplementary Information

Fusing machine learning strategy with density functional theory to hasten the discovery of 2D MXene based catalysts for hydrogen generation

B. Moses Abraham[1†], Priyanka Sinha[1†], Prosun Halder[1] and Jayant K. Singh[1,2*]
[1]Department of Chemical Engineering, Indian Institute of Technology Kanpur, Kanpur, 208016, India.
[2]Prescience Insilico Private Limited, Bangalore, 560049, India.

*E-mail: jayantks@iitk.ac.in

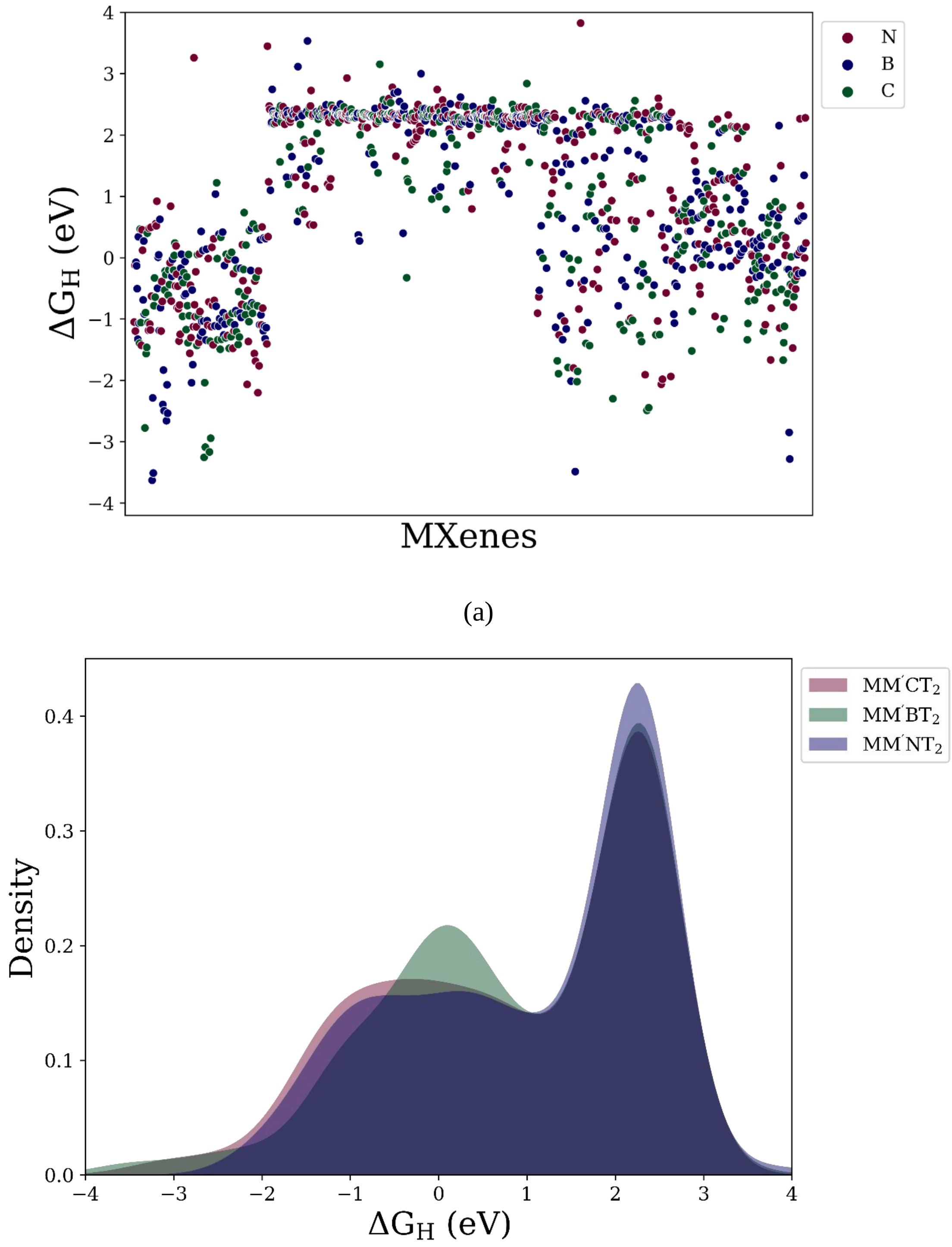
†Equal contribution



(a)

(b)

Figure S1: (a) Computed normalized cohesive energies $\overline{\mathrm{E}}_{\mathrm{coh}}$ (eV/atom) and (b) corresponding distribution for the randomly selected 1,125 MM'XT$_2$-type MXenes with respect to different X-layers (X = B, C or N).



(a)

(b)

Figure S2: (a) DFT computed hydrogen adsorbed Gibbs free energies (ΔG_H) for randomly selected 1,125 MM'XT$_2$-type MXenes and (b) their distribution with respect to different X-layers (X = B, C or N)

Table S1: List of primary features, including atomistic, structural and electronic indicators.

SYMBOL	**ATOMISTIC FEATURES**
$W_X, W_M, W_{M'}, W_T$	Atomic weight
$N_X, N_M, N_{M'}, N_T$	Atomic number
$P_X, P_M, P_{M'}, P_T$	Period number
$G_X, G_M, G_{M'}, G_T$	Group number
$V_X, V_M, V_{M'}, V_T$	Valence electron
$IE_X, IE_M, IE_{M'}, IE_T$	First ionization potential
$r_M, r_{M'}, r_T$	Radius
$EA_M, EA_{M'}, EA_T$	Electron affinity
$MP_M, MP_{M'}, MP_T$	Melting point
$BP_M, BP_{M'}, BP_T$	Boiling point
$\chi_M, \chi_{M'}$	Electronegativity
X	X-atom type
T	Termination type

SYMBOL	**STRUCTURAL FEATURES**
W_{sur}, W	Surface weight, normalized surface weight
A_{sur}	Surface area
ρ_{sur}	Surface density
$IE_D = IE_M - IE_{M'}$	First ionization potential difference

SYMBOL	**ELECTRONIC FEATURES**
LT	Layer thickness
$E_{coh}, \bar{E}_{coh}$	Cohesive energy, normalized cohesive energy
$l_{M-T}, l_{M'-T}, l_{M-X}, l_{M-X}$	Bond lengths

$d_{M-M}, d_{M-M'}, d_{T-T}, d_{T_1-T_2}, d_{X-X}$	Distances (nearest)
d_{NN}	Distance b/w nearest neighbors
dbc	d-band center
WF	Work function

Table S2: Statistical functions for each of the γ properties that are used to expand the primary features.

FEATURE	DESCRIPTION	FORMULA
$\bar{\gamma}$	Average value	$\sum_{i=0}^{n} \gamma_i / N$
$\tilde{\gamma}$	Average weighted value	$\sum_{i=0}^{n} \gamma_i n_i / N$
γ_{max}	Maximum value	$Max(\gamma_i)$
γ_{min}	Minimum value	$Min(\gamma_i)$
γ_{σ}	Standard deviation with respect to average	$\sqrt{\sum_{i=0}^{n} \frac{(\bar{\gamma} - \gamma_i)^2}{N}}$
γ_{σ^2}	Variance with respect to average	$\sum_{i=0}^{n} \frac{(\bar{\gamma} - \gamma_i)^2}{N}$
γ^2	Squared value	γ_i^2

Table S3: Machine learning models and their description.

ABBREVIATION	MODEL	TYPE	DESCRIPTION
ABR	AdaBoost Regressor	Ensemble	'Adaptive Boosting', fits a sequence of weak learning models
GBR	Gradient Boosting Regressor	Ensemble	Builds additive model in forward stage-fashion
KNR	K Neighbors Regressor	Neighbors	Based on K-nearest neighbors
KRR	Kernel Ridge	Kernel Ridge	Combines ridge (L2) penalty with kernel trick
LAS	Lasso	Linear	Trained with L1 penalty
RDG	Ridge Regression	Linear Model	Trained with L2 penalty
RFR	Random Forest Regressor	Ensemble	Meta estimator fitting a number of classifying decision trees
PLS	Partial Least Squares	Cross Decomposition	Regularized linear regression, similar to Lasso
ENR	Elastic Net Regressor	Linear	Uses penalties from both lasso (L1) and ridge (L2) regressions

Table S4: Mean absolute error (MAE) and coefficient of determination (R^2) after cross-validation for various feature subsets of nine different models.

MODEL	**FEATURE SUBSET**	**MAE**	**R^2**
ABR	1	0.635 ± 0.055	0.672 ± 0.092
	2	0.724 ± 0.074	0.558 ± 0.096
	3	0.709 ± 0.079	0.657 ± 0.117
	1 + 2	0.698 ± 0.063	0.561 ± 0.066
	1 + 3	0.597 ± 0.055	0.711 ± 0.101
	2 + 3	0.602 ± 0.092	0.727 ± 0.057
	1 + 2 + 3	0.767 ± 0.027	0.576 ± 0.034
GBR	1	0.512 ± 0.105	0.726 ± 0.154
	2	0.662 ± 0.071	0.546 ± 0.069
	3	0.585 ± 0.123	0.701 ± 0.068
	1 + 2	0.675 ± 0.06	0.549 ± 0.086
	1 + 3	0.472 ± 0.052	0.772 ± 0.098
	2 + 3	0.514 ± 0.077	0.776 ± 0.083
	1 + 2 + 3	0.424 ± 0.05	0.771 ± 0.066
KRR	1	0.592 ± 0.076	0.712 ± 0.13
	2	0.715 ± 0.043	0.546 ± 0.068
	3	0.799 ± 0.144	-1.205 ± 4.603
	1 + 2	0.666 ± 0.049	0.589 ± 0.045

	1 + 3	0.995 ± 0.39	-0.502 ± 2.228
	2 + 3	0.77 ± 0.107	0.573 ± 0.096
	1 + 2 + 3	0.696 ± 0.085	0.18 ± 0.077
KNR	1	1.25 ± 0.126	0.003 ± 0.161
	2	0.812 ± 0.039	0.401 ± 0.091
	3	1.449 ± 0.101	-0.247 ± 0.119
	1 + 2	0.921 ± 0.039	0.239 ± 0.08
	1 + 3	1.453 ± 0.161	-0.286 ± 0.212
	2 + 3	1.458 ± 0.135	-0.268 ± 0.135
	1 + 2 + 3	1.319 ± 0.066	0.191 ± 0.09
LAS	1	0.512 ± 0.105	0.726 ± 0.154
	2	1.017 ± 0.067	0.294 ± 0.066
	3	1.096 ± 0.081	0.247 ± 0.129
	1 + 2	0.728 ± 0.088	-0.541 ± 0.083
	1 + 3	0.787 ± 0.071	0.546 ± 0.106
	2 + 3	1.062 ± 0.129	0.291 ± 0.074
	1 + 2 + 3	0.639 ± 0.047	0.614 ± 0.068
RFR	1	0.507 ± 0.074	0.728 ± 0.094
	2	0.709 ± 0.068	0.472 ± 0.102
	3	0.571 ± 0.067	0.724 ± 0.086
	1 + 2	0.714 ± 0.061	0.459 ± 0.094

	1 + 3	0.447 ± 0.085	0.776 ± 0.09
	2 + 3	0.493 ± 0.108	0.781 ± 0.061
	1 + 2 + 3	0.382 ± 0.025	0.806 ± 0.061
RDG	1	0.586 ± 0.092	0.711 ± 0.087
	2	0.713 ± 0.063	0.547 ± 0.06
	3	0.609 ± 0.053	0.674 ± 0.155
	1 + 2	0.668 ± 0.061	0.587 ± 0.052
	1 + 3	0.622 ± 0.117	0.554 ± 0.46
	2 + 3	0.605 ± 0.099	0.679 ± 0.075
	1 + 2 + 3	0.563 ± 0.039	0.686 ± 0.058
ENR	1	0.693 ± 0.068	0.633 ± 0.05
	2	0.993 ± 0.066	0.32 ± 0.047
	3	1.084 ± 0.089	0.279 ± 0.087
	1 + 2	0.717 ± 0.034	0.544 ± 0.043
	1 + 3	0.701 ± 0.084	0.622 ± 0.095
	2 + 3	1.034 ± 0.087	0.328 ± 0.109
	1 + 2 + 3	0.634 ± 0.042	0.618 ± 0.069
PLS	1	0.582 ± 0.078	0.714 ± 0.077
	2	0.914 ± 0.062	0.371 ± 0.083
	3	1.025 ± 0.071	0.342 ± 0.087
	1 + 2	0.682 ± 0.071	0.575 ± 0.054

	1 + 3	0.637 ± 0.063	0.667 ± 0.127
	2 + 3	0.925 ± 0.064	0.43 ± 0.113
	1 + 2 + 3	0.704 ± 0.062	0.569 ± 0.068

Where,

Set 1 = Atomistic Features

Set 2 = Surface Features (Structural + Electronic)

Set 3 = Statistical Features

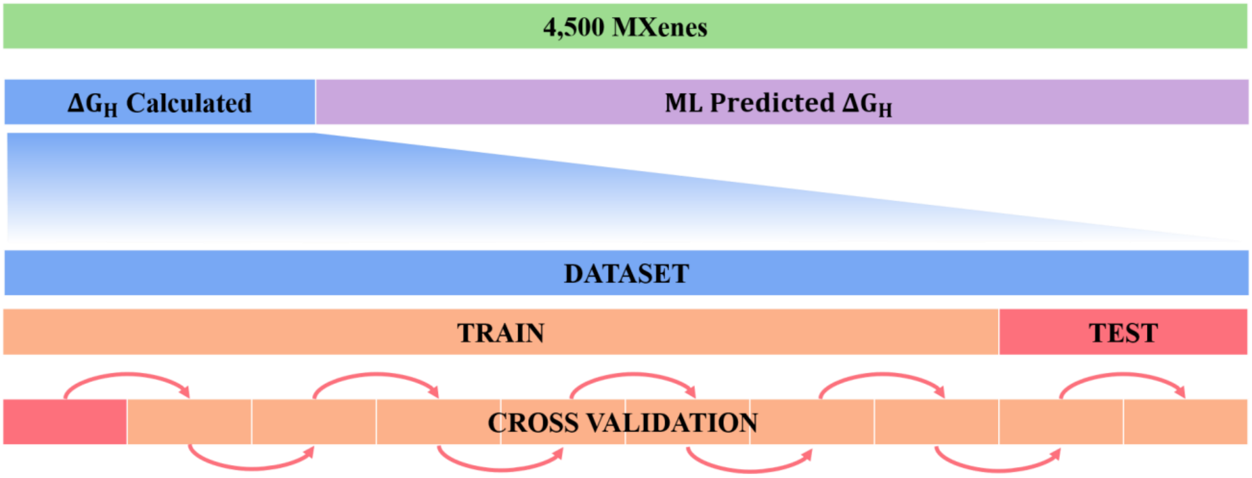
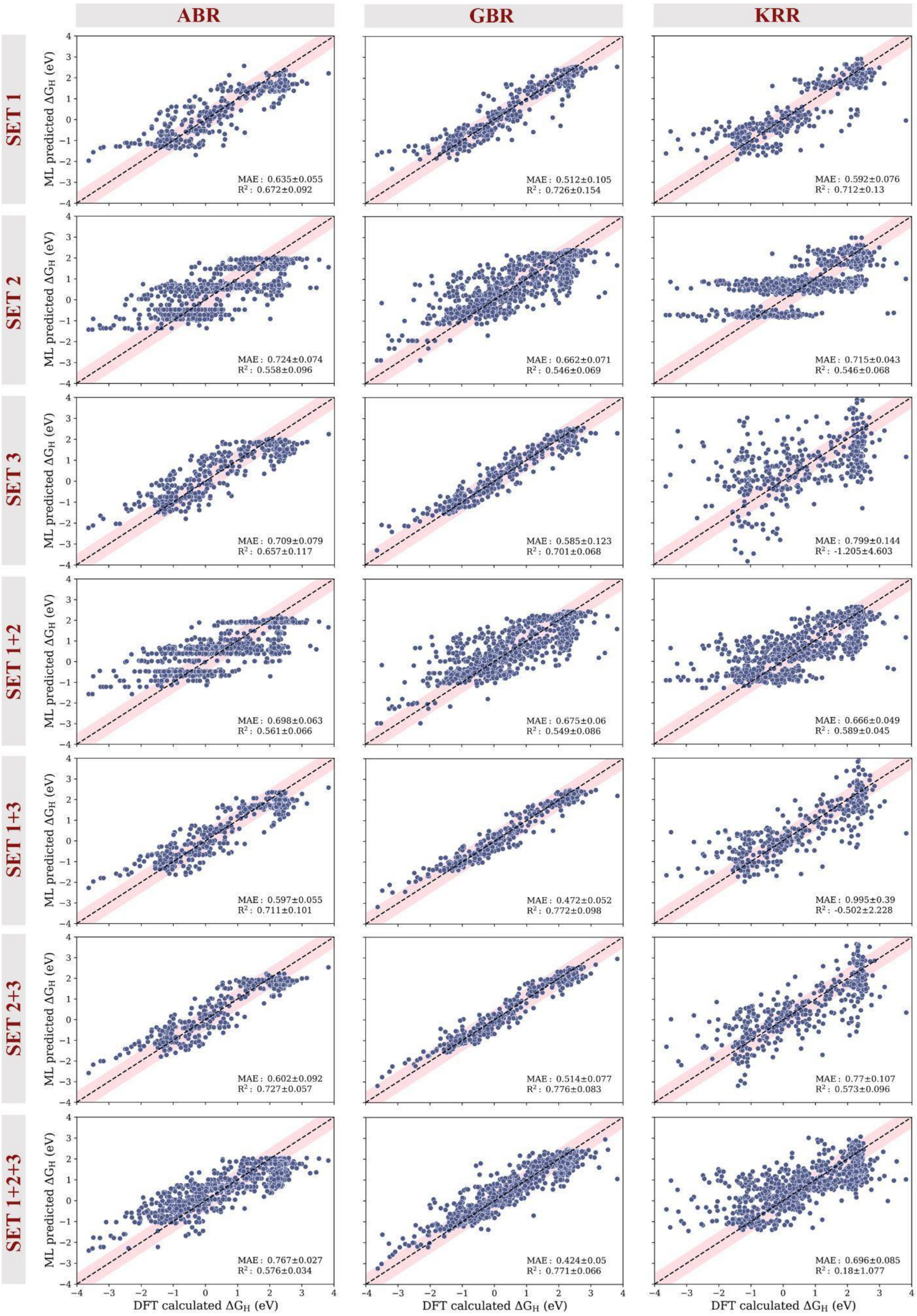
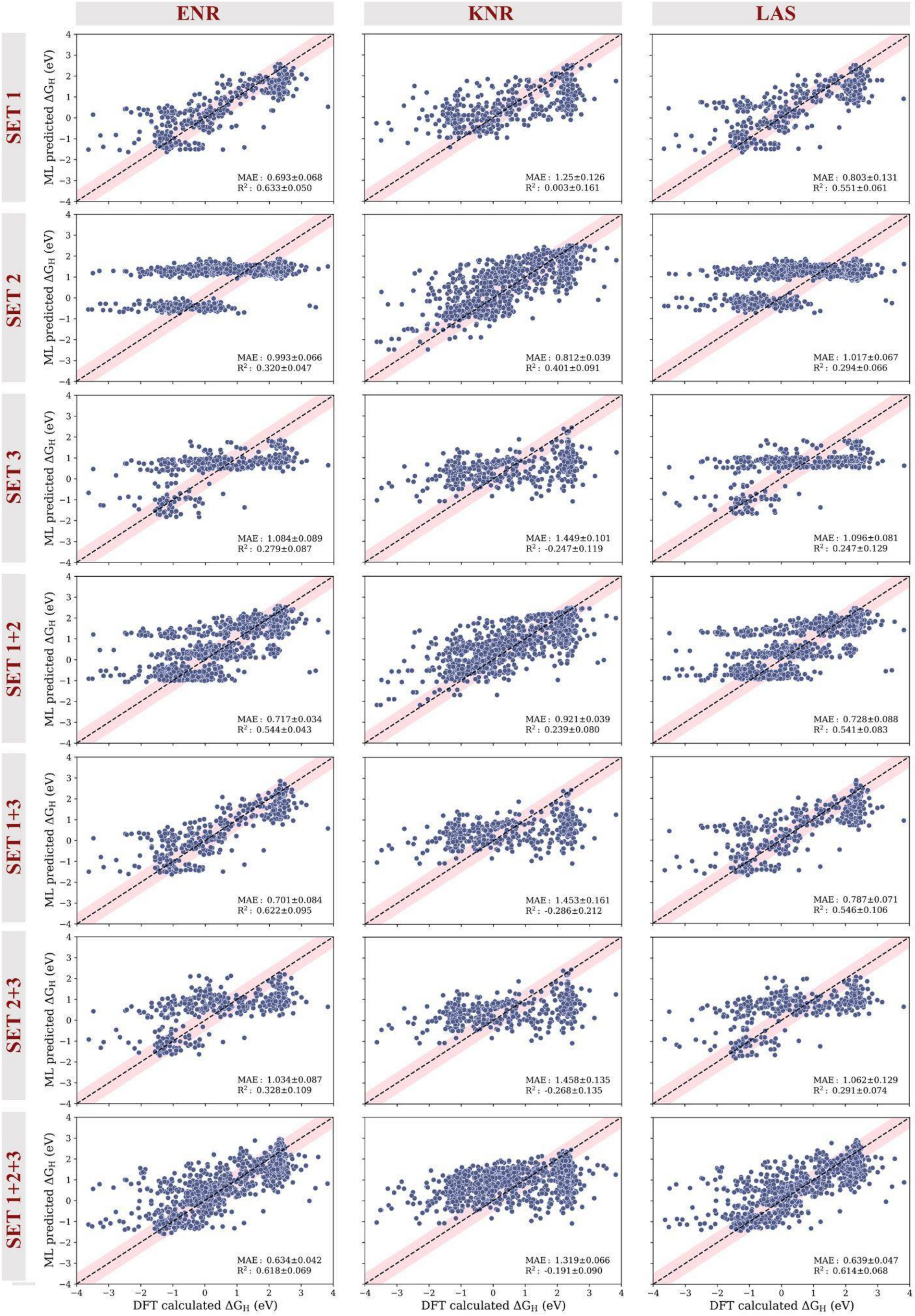


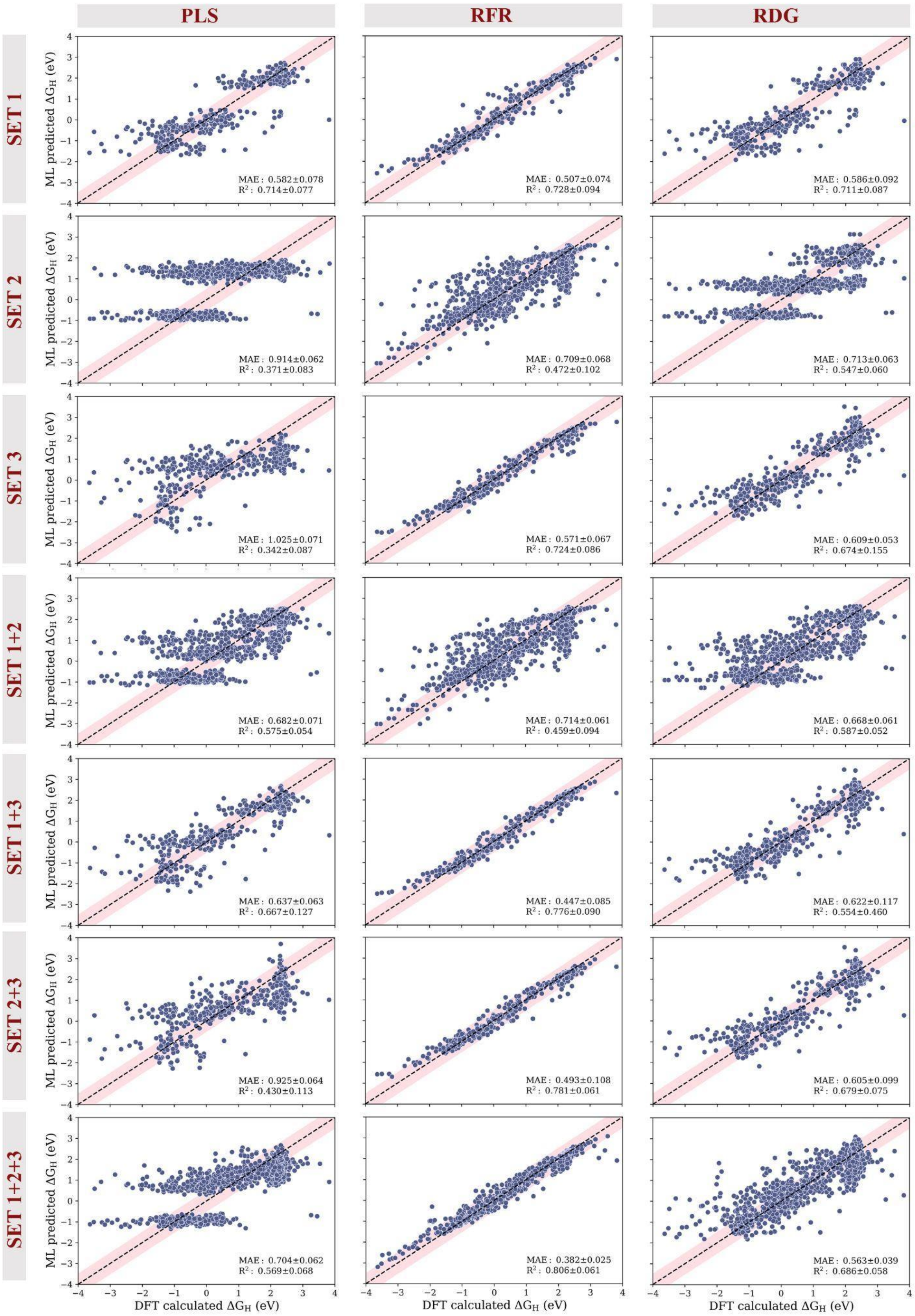
Figure S3: Pictorial representation of data distribution for training-testing and cross validation.


ABR
GBR
KRR
SET 1
SET 2
SET 3
SET 1+2
SET 1+3
SET 2+3
SET 1+2+3
ML predicted ΔG_H (eV)
DFT calculated ΔG_H (eV)
MAE: 0.635±0.055
R^2: 0.672±0.092
MAE: 0.512±0.105
R^2: 0.726±0.154
MAE: 0.592±0.076
R^2: 0.712±0.13
MAE: 0.724±0.074
R^2: 0.558±0.096
MAE: 0.662±0.071
R^2: 0.546±0.069
MAE: 0.715±0.043
R^2: 0.546±0.068
MAE: 0.709±0.079
R^2: 0.657±0.117
MAE: 0.585±0.123
R^2: 0.701±0.068
MAE: 0.799±0.144
R^2: -1.205±4.603
MAE: 0.698±0.063
R^2: 0.561±0.066
MAE: 0.675±0.06
R^2: 0.549±0.086
MAE: 0.666±0.049
R^2: 0.589±0.045
MAE: 0.597±0.055
R^2: 0.711±0.101
MAE: 0.472±0.052
R^2: 0.772±0.098
MAE: 0.995±0.39
R^2: -0.502±2.228
MAE: 0.602±0.092
R^2: 0.727±0.057
MAE: 0.514±0.077
R^2: 0.776±0.083
MAE: 0.77±0.107
R^2: 0.573±0.096
MAE: 0.767±0.027
R^2: 0.576±0.034
MAE: 0.424±0.05
R^2: 0.771±0.066
MAE: 0.696±0.085
R^2: 0.18±1.077
−4
−3
−2
−1
0
1
2
3
4

(a)


ENR
KNR
LAS
SET 1
SET 2
SET 3
SET 1+2
SET 1+3
SET 2+3
SET 1+2+3
ML predicted ΔG_H (eV)
DFT calculated ΔG_H (eV)
−4
−3
−2
−1
0
1
2
3
4
MAE : 0.693±0.068
R^2 : 0.633±0.050
MAE : 1.25±0.126
R^2 : 0.003±0.161
MAE : 0.803±0.131
R^2 : 0.551±0.061
MAE : 0.993±0.066
R^2 : 0.320±0.047
MAE : 0.812±0.039
R^2 : 0.401±0.091
MAE : 1.017±0.067
R^2 : 0.294±0.066
MAE : 1.084±0.089
R^2 : 0.279±0.087
MAE : 1.449±0.101
R^2 : -0.247±0.119
MAE : 1.096±0.081
R^2 : 0.247±0.129
MAE : 0.717±0.034
R^2 : 0.544±0.043
MAE : 0.921±0.039
R^2 : 0.239±0.080
MAE : 0.728±0.088
R^2 : 0.541±0.083
MAE : 0.701±0.084
R^2 : 0.622±0.095
MAE : 1.453±0.161
R^2 : -0.286±0.212
MAE : 0.787±0.071
R^2 : 0.546±0.106
MAE : 1.034±0.087
R^2 : 0.328±0.109
MAE : 1.458±0.135
R^2 : -0.268±0.135
MAE : 1.062±0.129
R^2 : 0.291±0.074
MAE : 0.634±0.042
R^2 : 0.618±0.069
MAE : 1.319±0.066
R^2 : -0.191±0.090
MAE : 0.639±0.047
R^2 : 0.614±0.068

(b)



(c)

Figure S4: Parity plots for various sets using different ML models.

Table S5: Mean absolute error (MAE) and coefficient of determination (R^2) for the training and testing data of set 1 + 2 + 3 using different ML models.

MODELS	TRAIN MAE	TEST MAE	TRAIN R^2	TEST R^2
ABR	0.666	0.702	0.701	0.574
GBR	0.294	0.421	0.913	0.753
KNR	1.083	1.342	0.215	0.005
KRR	0.707	0.625	0.561	0.621
LAS	0.647	0.578	0.619	0.668
RDG	0.52	0.568	0.746	0.672
RFR	0.144	0.388	0.973	0.776
PLS	0.717	0.647	0.576	0.592
ENR	0.642	0.573	0.629	0.676

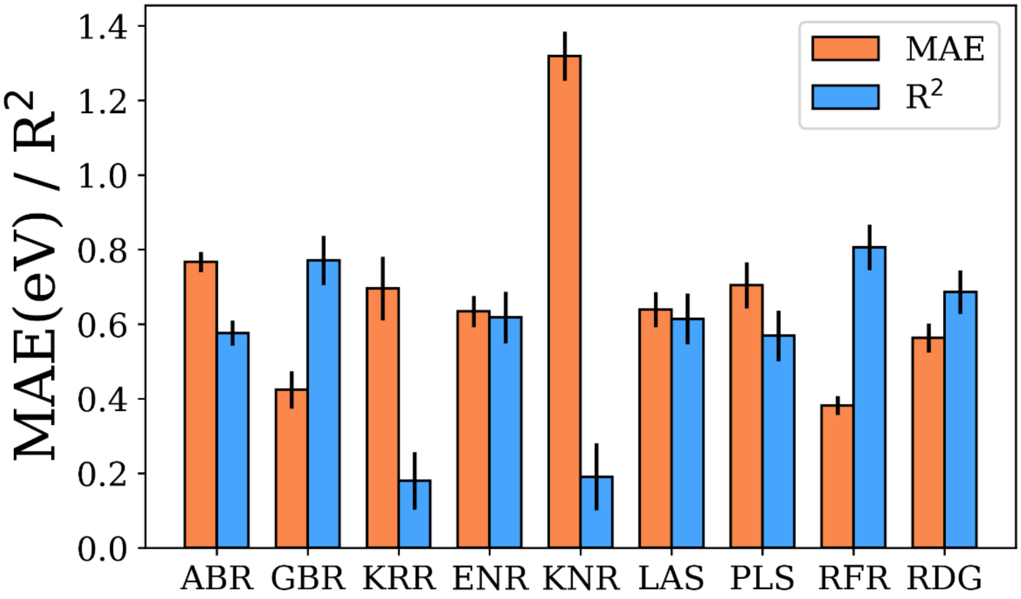


Figure S5: Mean absolute error (MAE) and coefficient of determination (R^2) after cross-validation for set 1 + 2 + 3 using different ML models.

Table S6: List of hyperparameters selected after using randomized search CV.

METHOD	HYPERPARAMETERS
Random Forest	n_estimators = 1000, min_samples_leaf = 2, max_features = 15, max_depth = 500, bootstrap = True
Gradient Boosting	n_estimatos = 400, min_samples_leaf = 10, max_features = ‘sqrt’, max_depth = 1000, learning_rate = 0.015

Table S7: Feature elimination using recursive feature elimination (RFE), hyperparameter optimization (HO), and leave-one-out (LOO) approach for RFR and GBR models. Here ‘K’ refers to the number of folds in cross-validation.

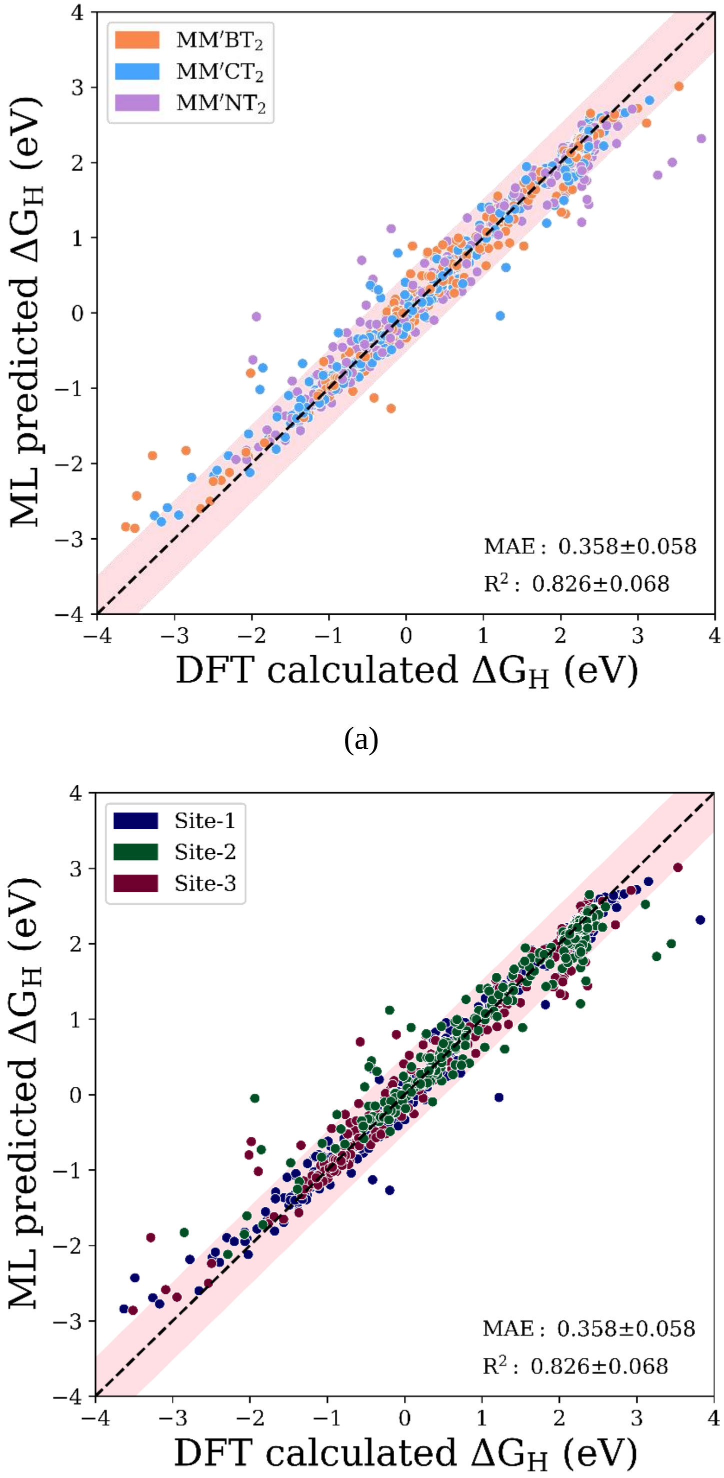
MODEL	APPROACH	K	NO. OF FEATURES	R^2	MAE
RFR	RFE	10	24	0.817	0.374
	LOO	20	15	0.820	0.367
	LOO	20	11	0.778	0.418
GBR	RFE	10	30	0.814	0.371
	LOO	20	19	0.826	0.358
	LOO	20	16	0.466	0.723

Table S8: Top seven features for GBR model after RFE-HO-LOO parameterization

SYMBOLS	**FEATURES**
V_T	Valence electron of termination
dbc_{σ^2}	d-band center variance with respect to average
EA_T	Electron affinity of termination
BP_T	Boiling point of termination
$(EA_T)_{\sigma^2}$	Electron affinity of termination variance with respect to average
MP_T	Melting point of termination
IE_T	Ionization enthalpy of termination

Table S9: Top seven features for RFR model after RFE-HO-LOO parameterization

SYMBOLS	**FEATURES**
V_T	Valence electron of termination
BP_T	Boiling point of termination
MP_T	Melting point of termination
dbc_{σ}	d-band center standard deviation with respect to average
LT	Layer thickness
$(EA_T)_{\sigma^2}$	Electron affinity of termination variance with respect to average
$d_{M-M'}$	Distance between inner metal and outer metal



(a)

(b)

Figure S6: Parity plots of GBR model after RFE-HO-LOO approach with respect to (a) X-layer (X = B, C, N) and (b) adsorption sites. The pink-shaded region indicates a deviation of up to 0.5 eV.

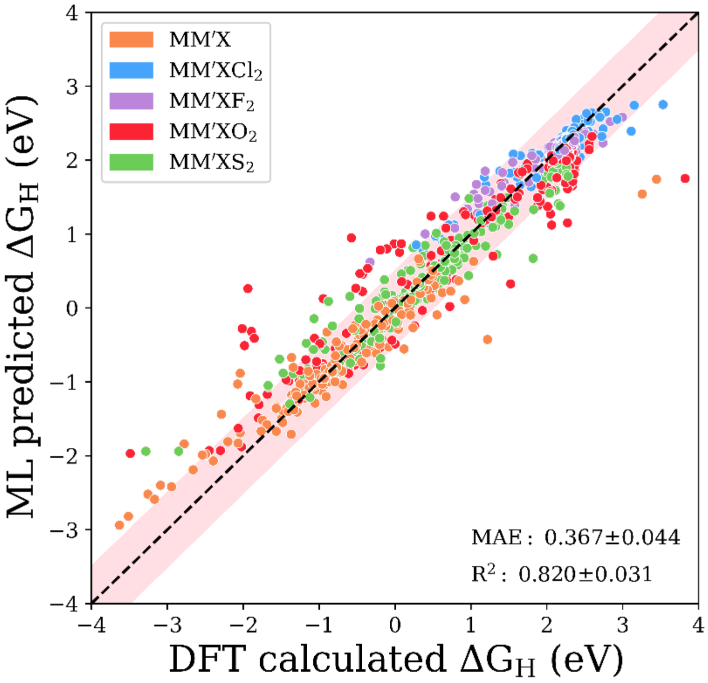
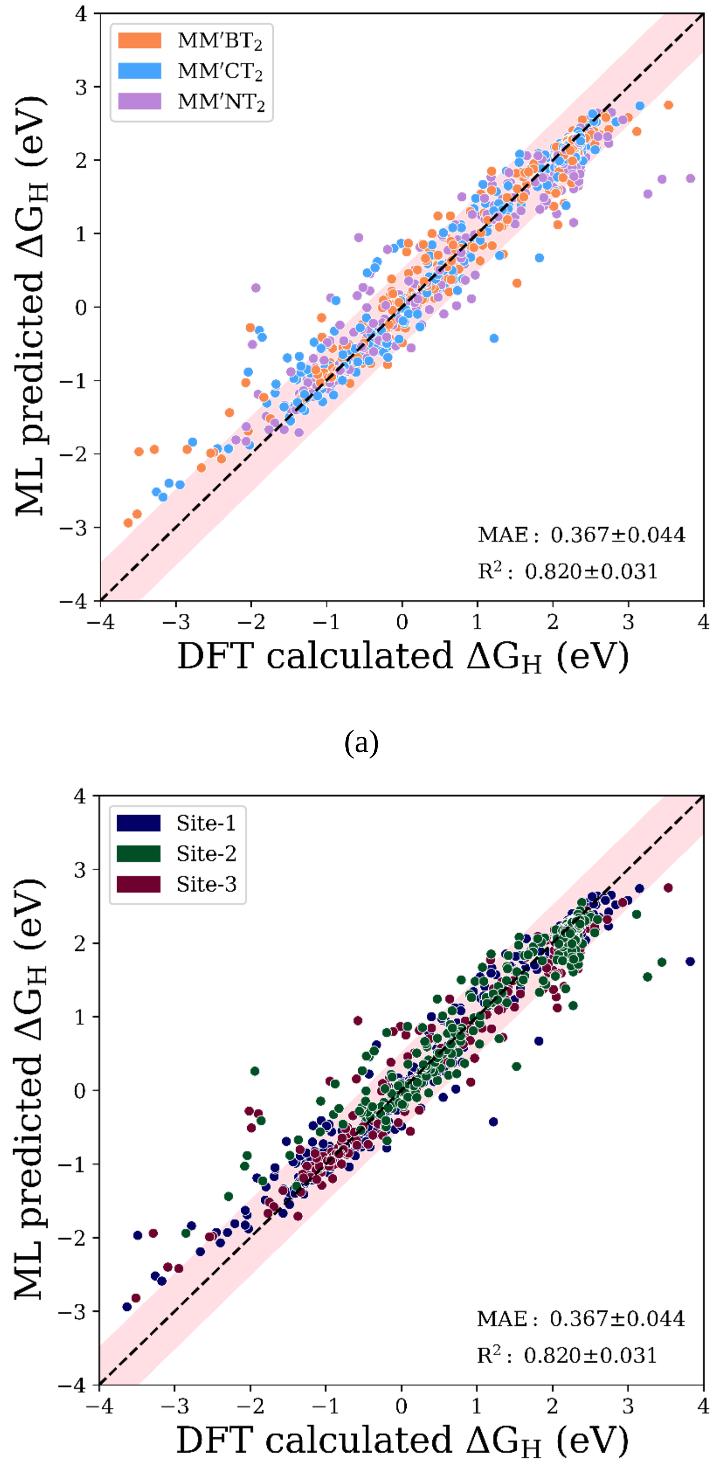


Figure S7: Parity plots of RFR model after RFE-HO-LOO approach with respect to terminations. The pink-shaded region indicates a deviation of up to 0.5 eV.



(a)

(b)

Figure S8: Parity plots of RFR model after RFE-HO-LOO approach with respect to (a) X-layers (X = B, C or N) and (b) adsorption sites. The pink-shaded region indicates a deviation of up to 0.5 eV.

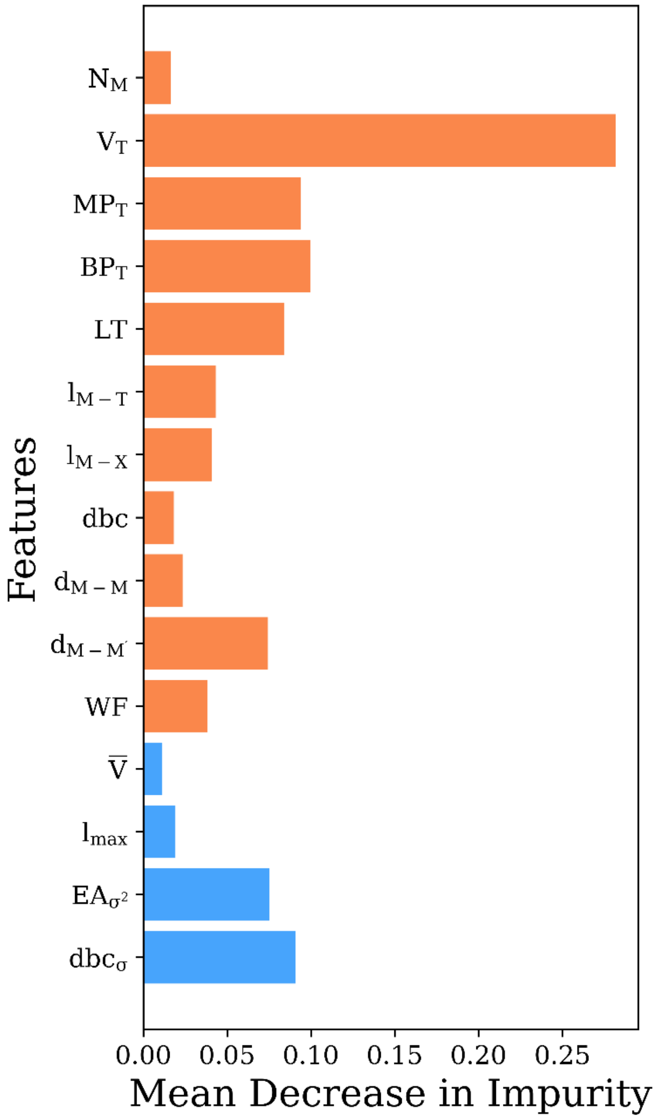
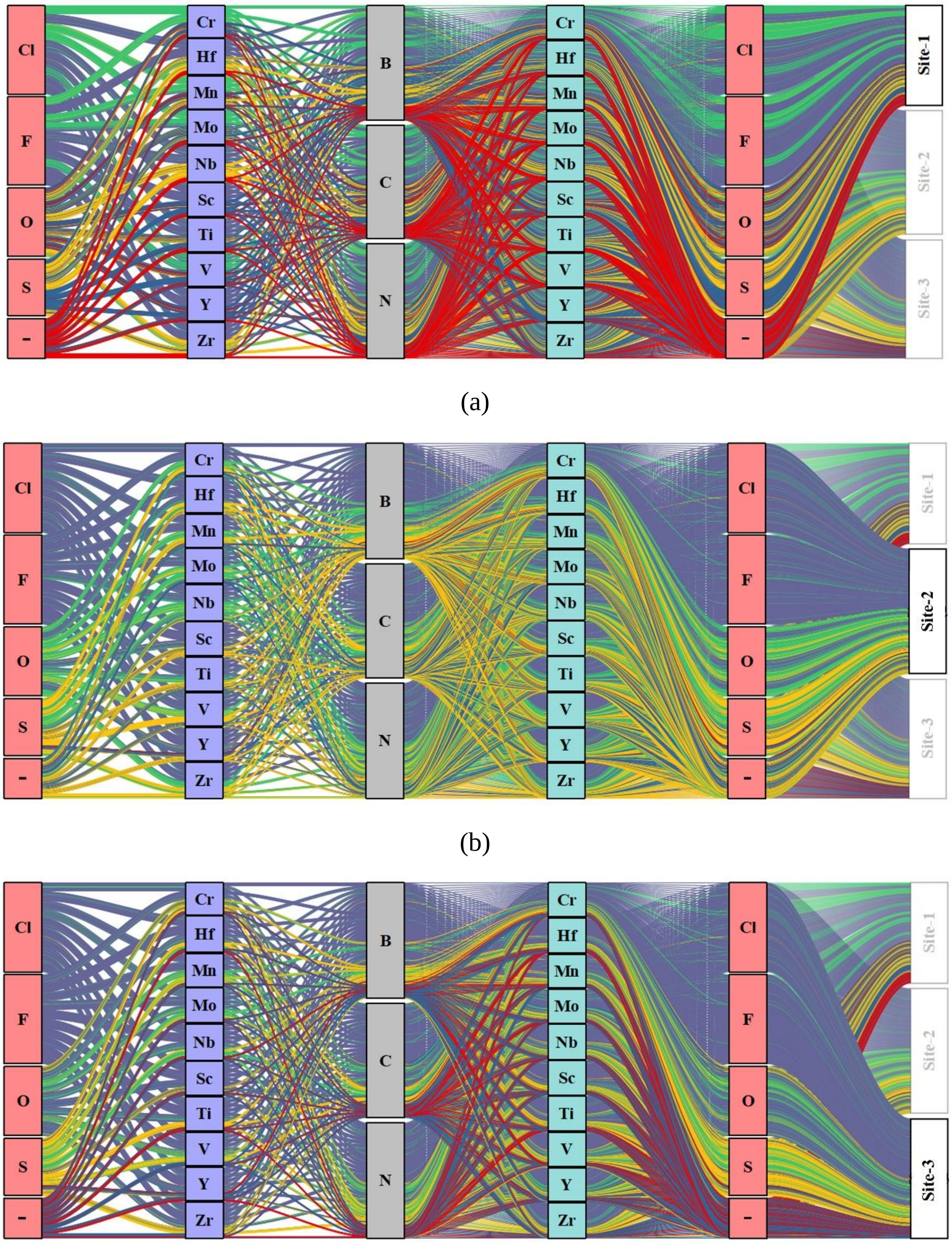


Figure S9: Feature importance using mean decrease in impurity on the RFR model after RFE-HO-LOO parameterization that were evaluated via 20-fold cross-validation. Orange and blue colors indicate primary and statistical features, respectively.

Table S10: Top 30 MM'XT$_2$-type MXenes with better stability and high HER activity predicted using GBR Model after RFE-HO-LOO parameterization.

MXENES	$\bar{E}_{coh}$	ML PREDICTED ΔG_H
CrNbNO$_2$-1	-7.729	-0.063
MnNbNO$_2$-1	-6.901	-0.097
NbMoBO$_2$-1	-7.539	-0.087
NbMoCO$_2$-1	-7.931	-0.048
NbVCO$_2$-1	-7.722	0.056
TiNbBO$_2$-1	-7.117	0.046
YMoNO$_2$-1	-7.389	-0.083
CrMoBO$_2$-3	-6.824	-0.058
CrVNO$_2$-3	-7.297	0.084
CrYNO$_2$-3	-7.123	0.021
MoCrC-2	-7.426	-0.050
MoCrN-2	-7.559	0.002
MoNbC-2	-8.066	0.097
MoNbNO$_2$-2	-8.030	0.084
NbCrB-2	-6.791	-0.078
NbCrC-2	-7.614	0.007
NbTiC-2	-7.063	0.061
NbTiN-2	-7.309	0.015
NbYN-2	-6.957	0.017
TiCrN-2	-6.780	-0.098
TiMoC-2	-7.019	0.062
TiMoN-2	-7.171	-0.090
TiVN-2	-6.893	-0.061
VMoB-2	-6.849	0.088
VNbC-2	-7.641	-0.040
VYN-2	-6.611	-0.017
YCrN-2	-6.621	0.008
YNbC-2	-6.769	0.022
YNbN-2	-6.994	-0.008
YYNO$_2$-2	-6.819	-0.0570



(a)

(b)

(c)

Figure S10: Alluvial Diagram showing (a) Site-1, (b) Site-2, and (c) Site-3 of ML Predicted $\Delta \mathrm{G}_{\mathrm{H}}$ of 4,500 MXenes. Blue, red and yellow color links represent positive, negative and close to zero ΔG_H value respectively.