Electronic Supplementary Material (ESI) for Energy Advances.

Predicting performance and stability parameters of energetic materials (EMs) using the ML-based q-RASPR approach

Shubham Kumar Pandey and Kunal Roy*

Department of Pharmaceutical Technology

Jadavpur University, Kolkata 700032, India

Email: kunal.roy@jadavpuruniversity.in

Supplementary Materials SI-2

*Corresponding author

Prof. Kunal Roy, Phone: +91 98315 94140;

Fax: +91-33-2837-1078

Email: kunalroy_in@yahoo.com; kunal.roy@jadavpuruniversity.in

Machine learning methods

a) **Ridge regression**: It is a popular technique used to address multicollinearity in MLR models without removing any independent variables. The method involves adding a small amount of bias or penalty to improve predictions. This technique is known as Tikhonov regularization and is vital in reducing model complexity. The mathematical equation of ridge regression is: $L(x,y) = Min(\sum_{i=1}^{n}(y_i - w_i x_i)^2 + \lambda\sum_{i=1}^{n}(w_i)^2)$ where w_i is weightage of each feature and λ is penalty term.

b) **Support Vector Machine (SVM)**: Support Vector Machines (SVM) is a machine learning algorithm that can be used for both classification and regression problems. The primary objective of SVM is to draw a decision boundary between observations to predict outcomes. In the case of nonlinear SVM, the data is transformed into a feature space using a kernel function before mapping with the response. This technique is also known as Support Vector Regression (SVR). The mathematical equation for SVM (non-linear) is represented as follows: $\hat{y} = w^T\phi(X) + b$, where $\hat{y}$ is predictions, w is the vector of weights, X is a vector of input features, Φ is a kernel function and b is bias Support Vector Machines (SVM) methods are used in both two-dimensional and higher-order spaces with a large number of features. In SVM, the method considers both margins and hyperplanes for predictions. The margin refers to the area between the decision boundary and the closest training compound, while the hyperplane is used to predict class boundaries. The margin is represented by the following equation: $margin = \frac{1}{w^T w}$. SVM tries to maximize the distance between the two closest training compounds on either side of the decision boundary.

c) **Linear Support Vector Machine (LSVM):** The LSVM (Linear Support Vector Machine) algorithm is a machine learning algorithm used for the classification of data. It is a powerful and popular tool in various fields, including image recognition, natural language processing, and bioinformatics. The LSVM algorithm involves mapping the input data domain to a response data space, where the data can be linearly classified without any transformation. This is accomplished by finding the hyperplane that best separates the data points of different classes. The algorithm aims to find the hyperplane that maximizes the margin, which is the distance between the hyperplane and the closest data points of each class. The generalized equation for LSVM is: $\hat{y} = w^T X + b$.

d) **Random forest (RF):** Random forest (RF) is a machine learning algorithm that combines the outcomes of multiple decision tree models to provide more accurate and stable predictions. This approach helps to overcome the overfitting problem common with decision tree models. RF is based on an ensemble learning method called Bagging (Bootstrap Aggregating), which is a resampling technique applied to a dataset. In bootstrapping, observations are randomly selected with replacement, and random feature subsets are chosen. Bagging creates a large number of datasets by bootstrapping the original dataset, builds multiple decision tree models using these datasets, and finally takes the average of their predictions.

e) **Gradient boosting (GB):** Boosting is a machine learning technique that combines multiple weak learners to form a strong learner. Gradient boosting (GB) is a specific type of boosting method that builds decision trees sequentially, with each subsequent tree trying to correct the errors of its predecessor.

f) **XGBoost:** The XGBoost algorithm was developed by researchers at the University of Washington as a way to improve upon the Gradient Boosting (GB) algorithm. GB becomes time-consuming when dealing with thousands of features, as it searches for the best way to split the node of a decision tree across all possible options. XGBoost overcomes this by taking into account the distribution of features across all data points in a single leaf node, which reduces the search space. While it can't generate multiple decision trees in parallel, it can generate multiple branches of a decision tree simultaneously.

g) **Adaboost:** AdaBoost is a powerful ensemble learning technique primarily used for classification tasks but also can be applied to regression tasks. It operates by combining multiple weak classifiers to create a strong classifier. The essence of AdaBoost lies in its ability to adaptively adjust the weights of misclassified instances, allowing subsequent weak learners to focus more on difficult examples, thus improving overall performance.

During each iteration, AdaBoost assigns weights to each training instance based on its classification accuracy in the previous iteration. Misclassified instances are assigned higher weights, effectively forcing subsequent weak learners to focus more on them.

QSPR model development

A 10-descriptor MLR model for decomposition temperature (T_{dec}) was selected after the feature selection process by performing a grid-search using the Best Subset Selection tool v2.1 available from http://teqip.jdvu.ac.in/QSAR_Tools/. The same descriptor set was used to develop the final PLS QSAR model with 5 latent variables (LVs) which are optimized by LOO Q^2. The equation for the model is given in **Table S1.** The training set of the melting point (T_m) temperature data set was subjected to a forward step-wise feature selection process to enlist the prominent features closely related to the melting point. A 29-descriptor MLR QSPR model was developed to predict the melting point temperature of the compounds. The MLR equation for the model is shown in **Table S1.** The feature selection of the density data set was performed through step-wise selection using the training set. After the feature selection process, a 6-descriptor MLR model was prepared and further, PLS regression was used to develop the QSPR model with 5 LVs. The PLS equation of the model is given in **Table S1.** For the enthalpy of formation (ΔH_f°), a step-wise feature selection process was performed after the division of the data set. The pool of descriptors so obtained from the step-wise selection was then used to develop several MLR models through a grid-search approach using a java based tool Best Subset Selection tool v2.1 available from http://teqip.jdvu.ac.in/QSAR_Tools/. An 11-descriptor MLR model was selected based on the cross-validation result (Q^2_{LOO}), and further with the same set of descriptors, a PLS QSPR model was developed with 3 LVs. The PLS equation is given in **Table S1.**

Table S1: Model equations and validation metrics of the developed QSPR models

Property	**Model equation**	**Training set metrics**	**Test set metrics**
$\mathbf{T_{dec}}$ (PLS model)	T_{dec} $= 436.990 + 3.952 \times C\% - 142.26$ $nArNO2 + 24.399 \times C - 005 - 25.5$ $Descriptors = 10, LVs = 5$	$n_{training} = 424$ $R^2 = 0.578$ $Q^2_{LOO} = 0.557$ $MAE_{tr} = 45.257$ $RMSE_C = 57.971$	$n_{test} = 141$ $Q^2_{F1} = 0.621$ $Q^2_{F2} = 0.621$ $MAE_{te} = 44.919$ $RMSE_P = 54.814$
$\mathbf{T_m}$ (MLR model)	T_m $= 291.1 + 13.46 \times Ui\ +\ 22.98$ $\times AMW\ -\ 0.212 \times T(N..O)\ +$ $\times nR$ $= Cp\ -\ 4.25 \times F07[C - N]\ -\ 3$ $MaxssCH2\ +\ 11.57 \times N - 072$ $-\ 14.8 \times F02[O - Cl]\ +\ 86 \times N$ $Descriptors = 29$	$n_{training} = 14750$ $R^2 = 0.679$ $Q^2_{LOO} = 0.676$ $MAE_{tr} = 39.633$ $RMSE_C = 51.686$	$n_{test} = 4917$ $Q^2_{F1} = 0.670$ $Q^2_{F2} = 0.670$ $MAE_{te} = 39.626$ $RMSE_P = 52.501$
Density (PLS model)	$Density = 1.235 + 0.120 \times AMW - 1.409 \times Mp + 0.015 \times nX - 0.008$ $Descriptors = 6, LVs = 5$	$n_{training} = 9604$ $R^2 = 0.924$ $Q^2_{LOO} = 0.922$ $MAE_{tr} = 0.037$ $RMSE_C = 0.053$	$n_{test} = 3201$ $Q^2_{F1} = 0.928$ $Q^2_{F2} = 0.928$ $MAE_{te} = 0.037$ $RMSE_P = 0.051$

$\Delta H_f{}^\circ$ (PLS model)	$\Delta H_f{}^\circ$ $=- 25.420 - 196.661 \times nF - 71 \times O - 058 + 57.671 \times F01[N -$ $Descriptors = 11, LVs = 3$	$n_{training} = 1924$ $R^2 = 0.967$ $Q_{LOO}^{2} = 0.966$ $MAE_{tr} = 53.553$ $RMSE_C = 78.571$	$n_{test} = 643$ $Q_{F1}^{2} = 0.932$ $Q_{F2}^{2} = 0.931$ $MAE_{te} = 47.903$ $RMSE_P = 67.412$

Table S2: Definitions of descriptors of the QSPR models

Descriptors	Definition
C%	Percentage of C atoms
B01[O-O]	Presence/absence of O – O at topological distance 1
B03[N-O]	Presence/absence of N – O at topological distance 3
Hy	Hydrophilic factor
LOGP99	Wildmann-Crippen octanol-water partition coeff. (logP)
NArNO2	Number of nitro groups (aromatic)
C-005	CH3X
nN	Number of N atoms
B01[N-N]	Presence/absence of N – N at topological distance 1
B01[N-O]	Presence/absence of N – O at topological distance 1
Ui	Unsaturation index
nHDon	Number of donor atoms for H-bonds (N and O)
Rbrid	Ring bridge count
B03[C-O]	Presence/absence of C – O at topological distance 3
NArCOOH	Number of carboxylic acids (aromatic)
AMW	Average molecular weight
T(N..O)	Sum of topological distances between N..O
Rprim	Ring perimeter
nRCOOH	Number of carboxylic acids (aliphatic)
F10[C-O]	Frequency of C – O at topological distance 10
NdssC	Number of atoms of type dssC
nR=Cp	Number of terminal primary C(sp2)
F07[C-N]	Frequency of C – N at topological distance 7
minsssB	Mimimum sssB
MLOGP2	Squared Moriguchi octanol-water partition coeff. (logp^2)
Mi	Mean first ionization potential (scaled on Carbon atom)
nCbH	Number of unsubstituted benzene C(sp2)
MaxssCH2	Maximum ssCH2
N-072	RCO-N< / >N-X=X
O%	Percentage of O atoms
F05[O-O]	Frequency of O – O at topological distance 5
F10[C-C]	Frequency of C – C at topological distance 10
B02[C-C]	Presence/absence of C – C at topological distance 2

F02[O-Cl]	Frequency of O – Cl at topological distance 2
NssssN$^+$	Number of atoms of type ssssN$^+$
StN	Sum of tn E-states
F10[O-O]	Frequency of O – O at topological distance 10
nOHs	Number of secondary alcohols
Mp	Mean atomic polarizability (scaled on Carbon atom)
nX	Number of halogen atoms
X%	Percentage of halogen atoms
MCD	Molecular cyclized degree
NRS	Number of ring systems
nF	Number of Fluorine atoms
F01[C-O]	Frequency of C – O at topological distance 1
nCsp3	Number of sp3 hybridized Carbon atoms
nCIC	Number of rings (cyclomatic number)
F01[N-F]	Frequency of N – F at topological distance 1
F01[N-N]	Frequency of N – N at topological distance 1
O-058	=O
NsOH	Number of atoms of type sOH
NdsCH	Number of atoms of type dsCH
nCsp	Number of sp hybridized Carbon atoms

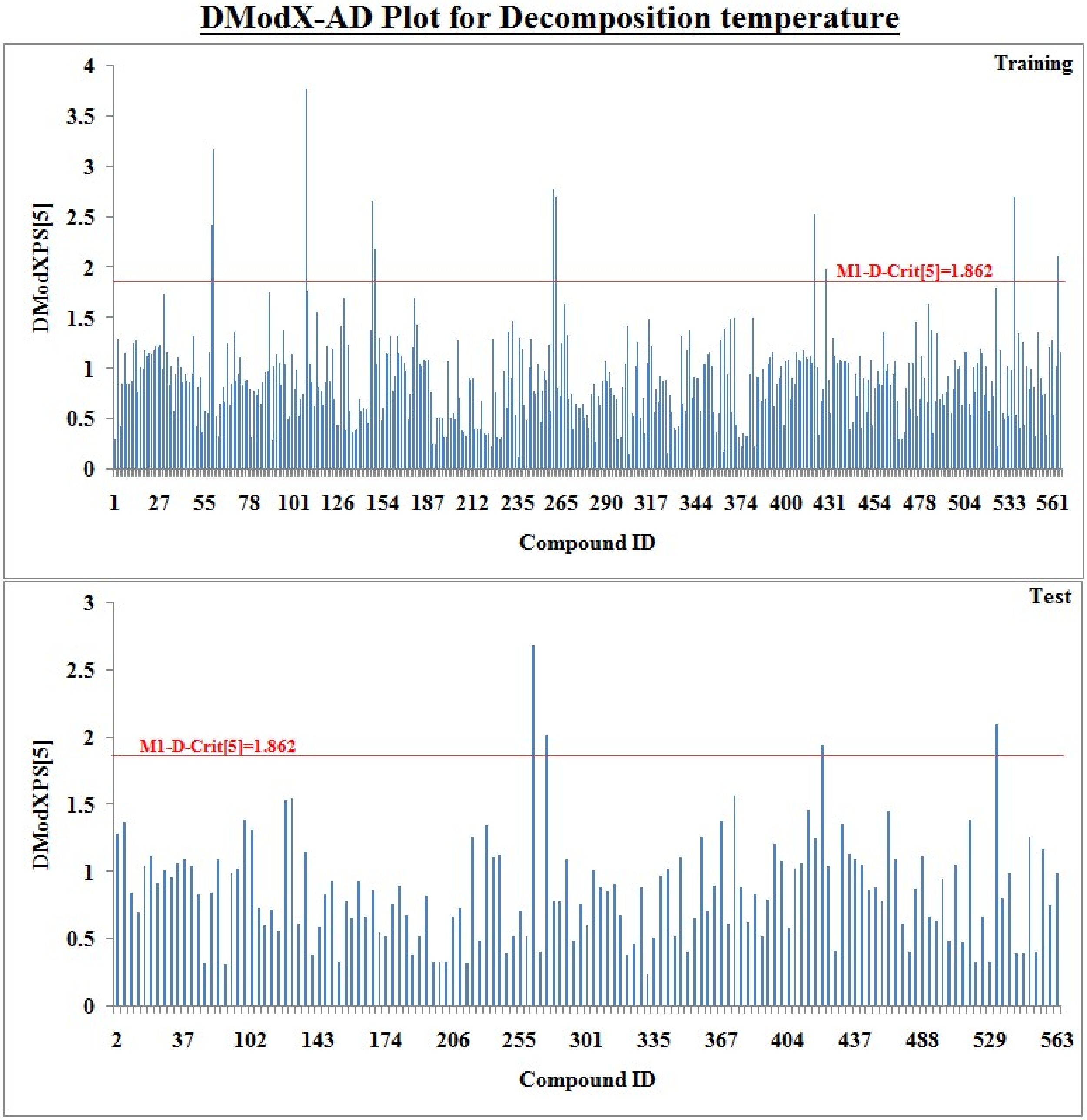


Figure S1: AD plot for T_{dec}

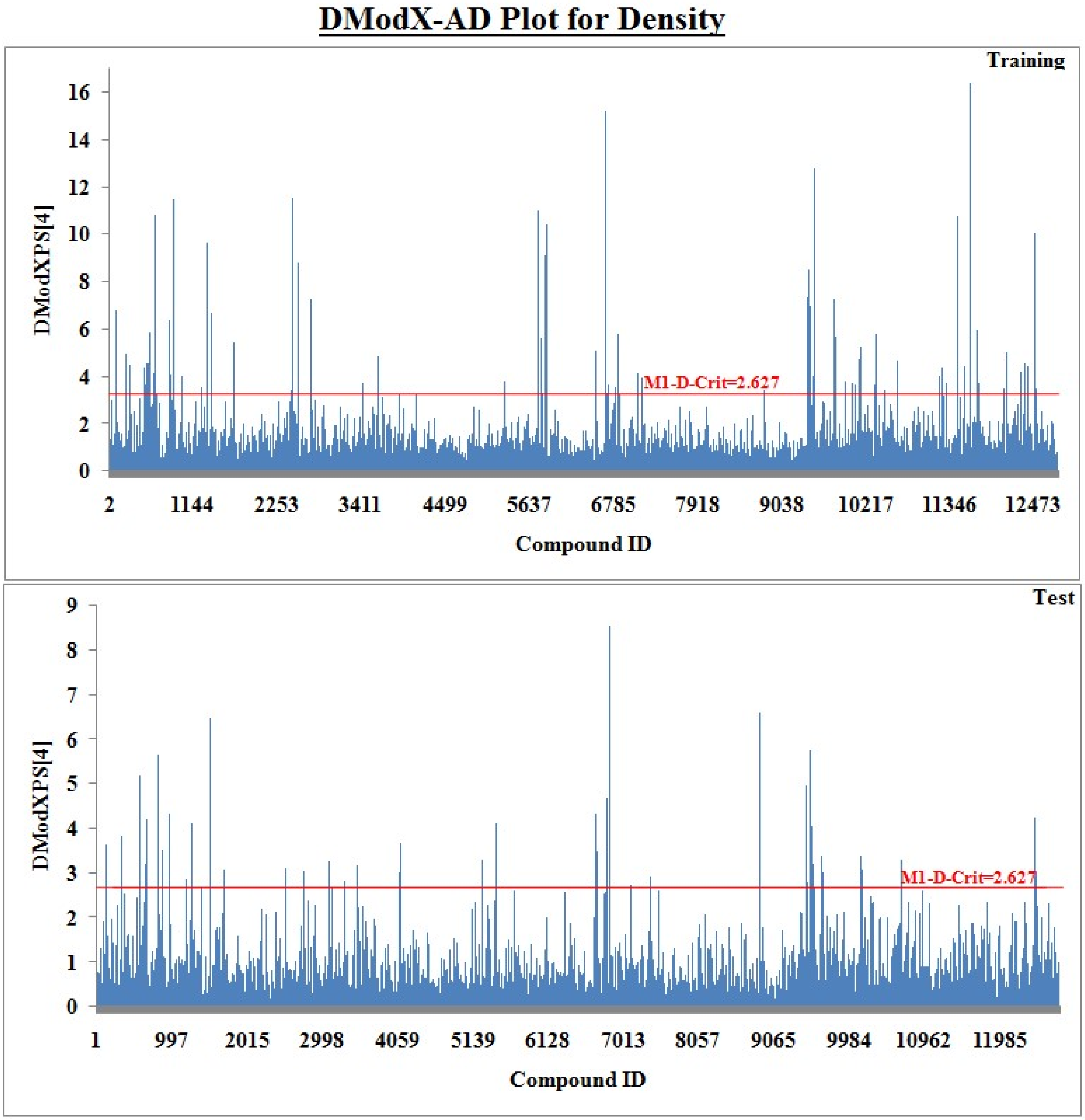


Figure S2: AD plot for Density

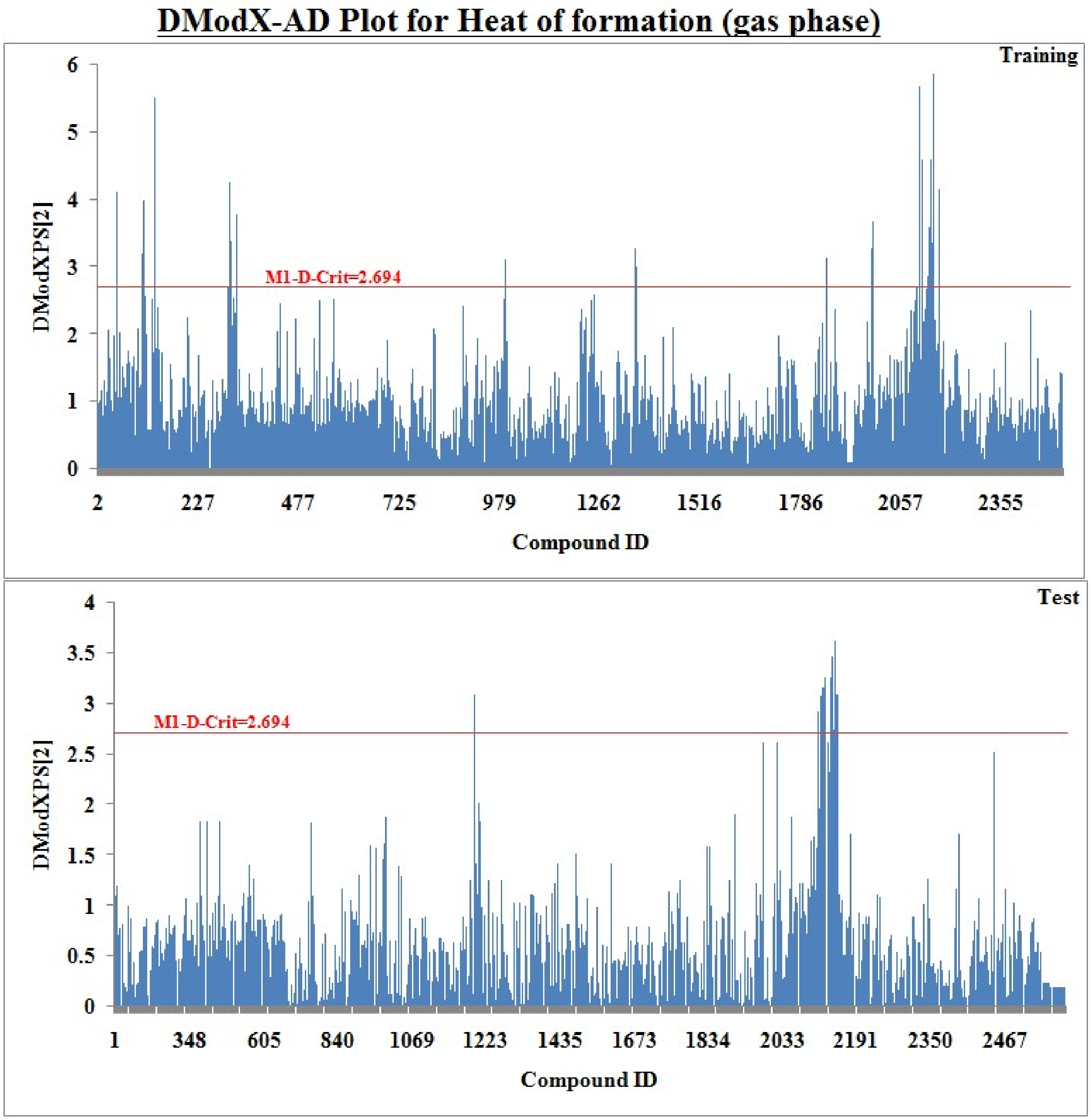


Figure S3: AD plot for ΔH_f°

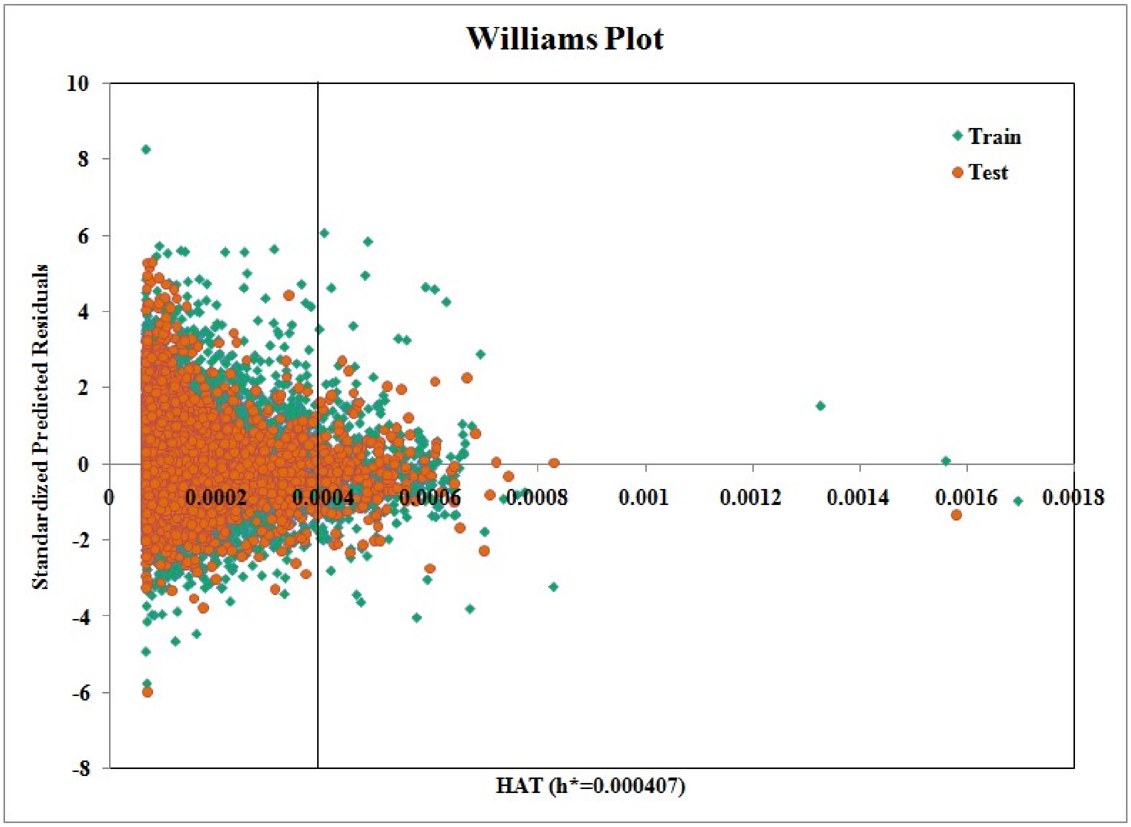


Figure S4: Williams plot for T_m

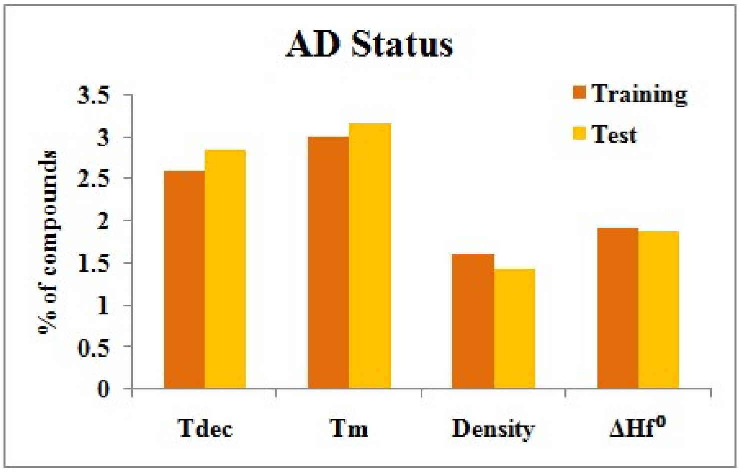


Figure S5: AD status for individual models. It represents the percentage (%) of compounds as outliers in training and test sets of the respective model.

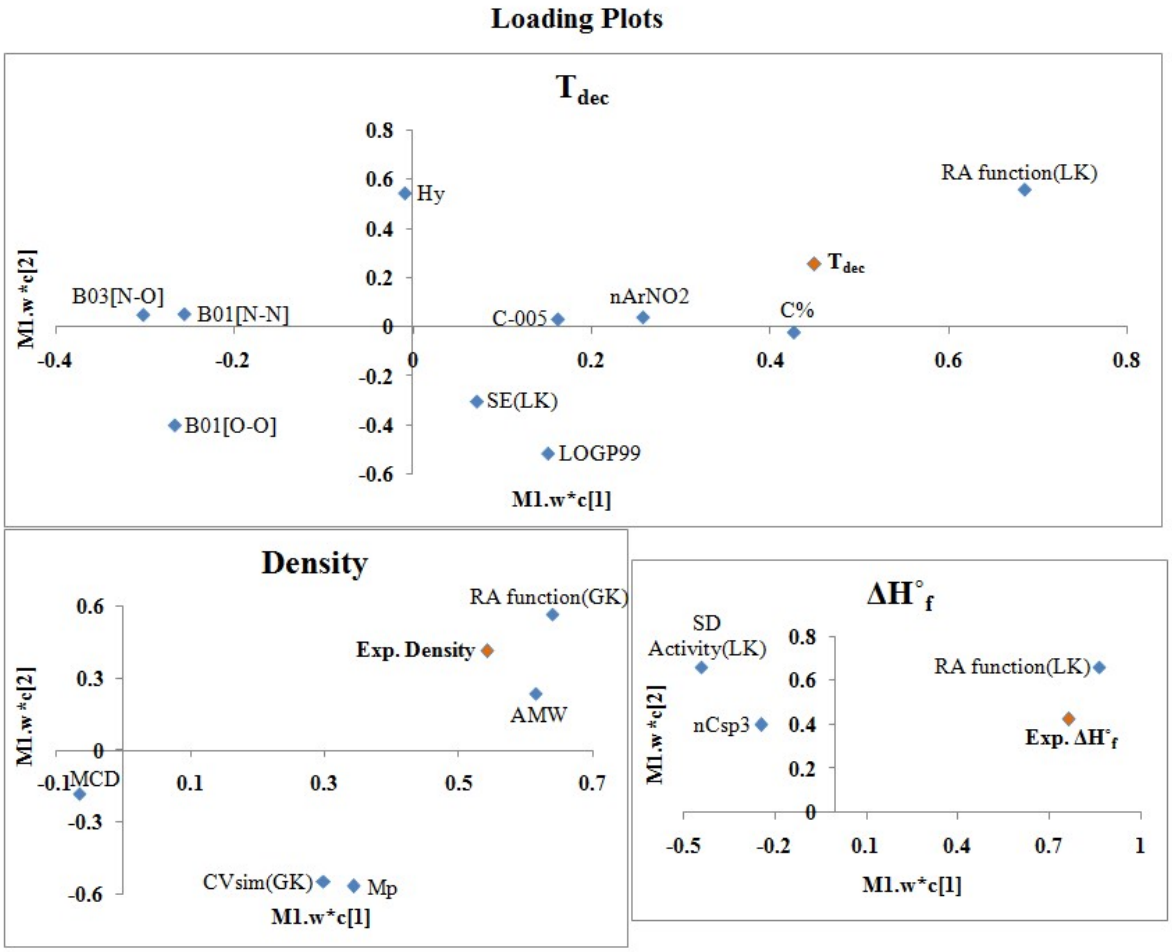


Figure S6: Loading Plots for different PLS q-RASPR models

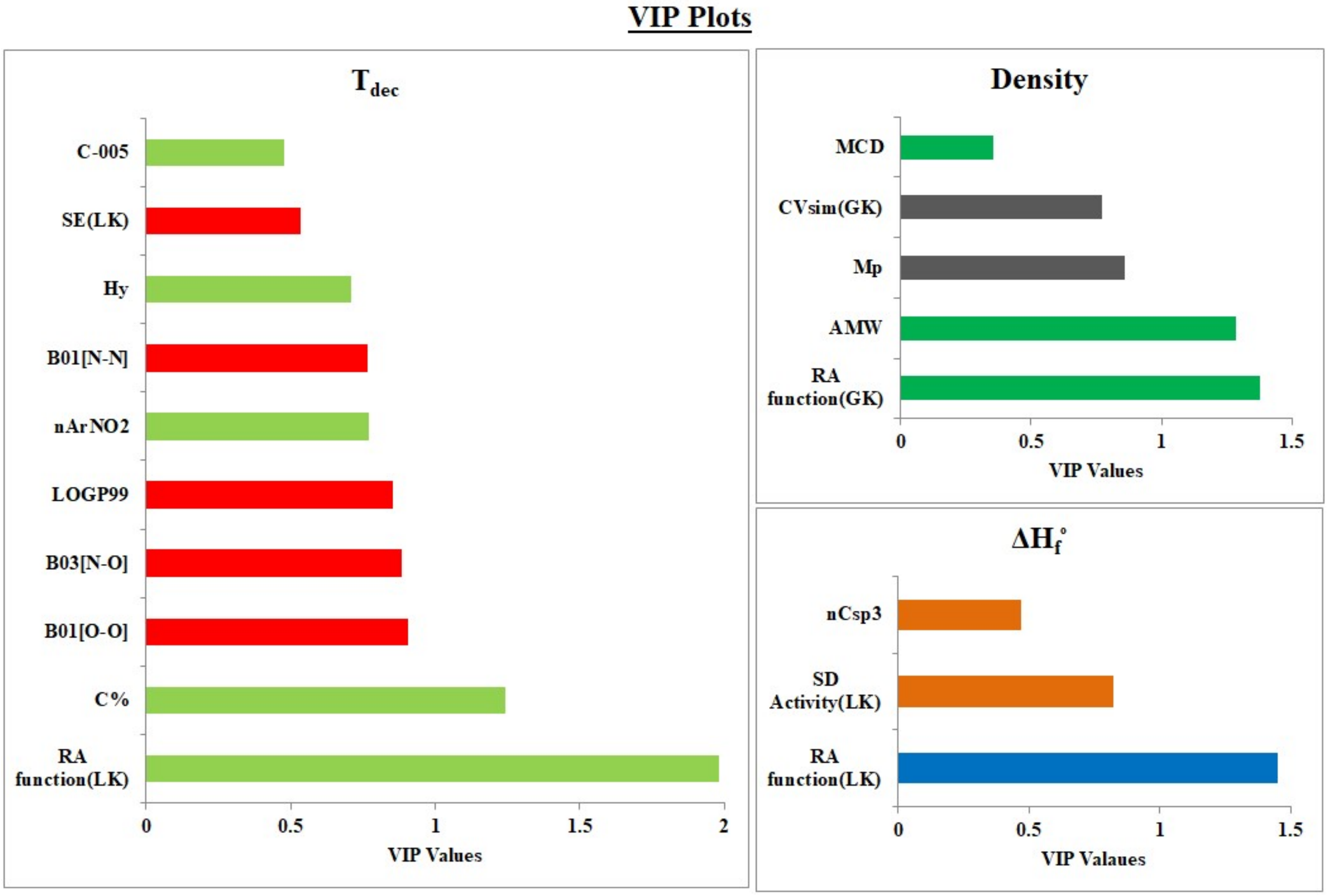


Figure S7: VIP plots for different PLS models

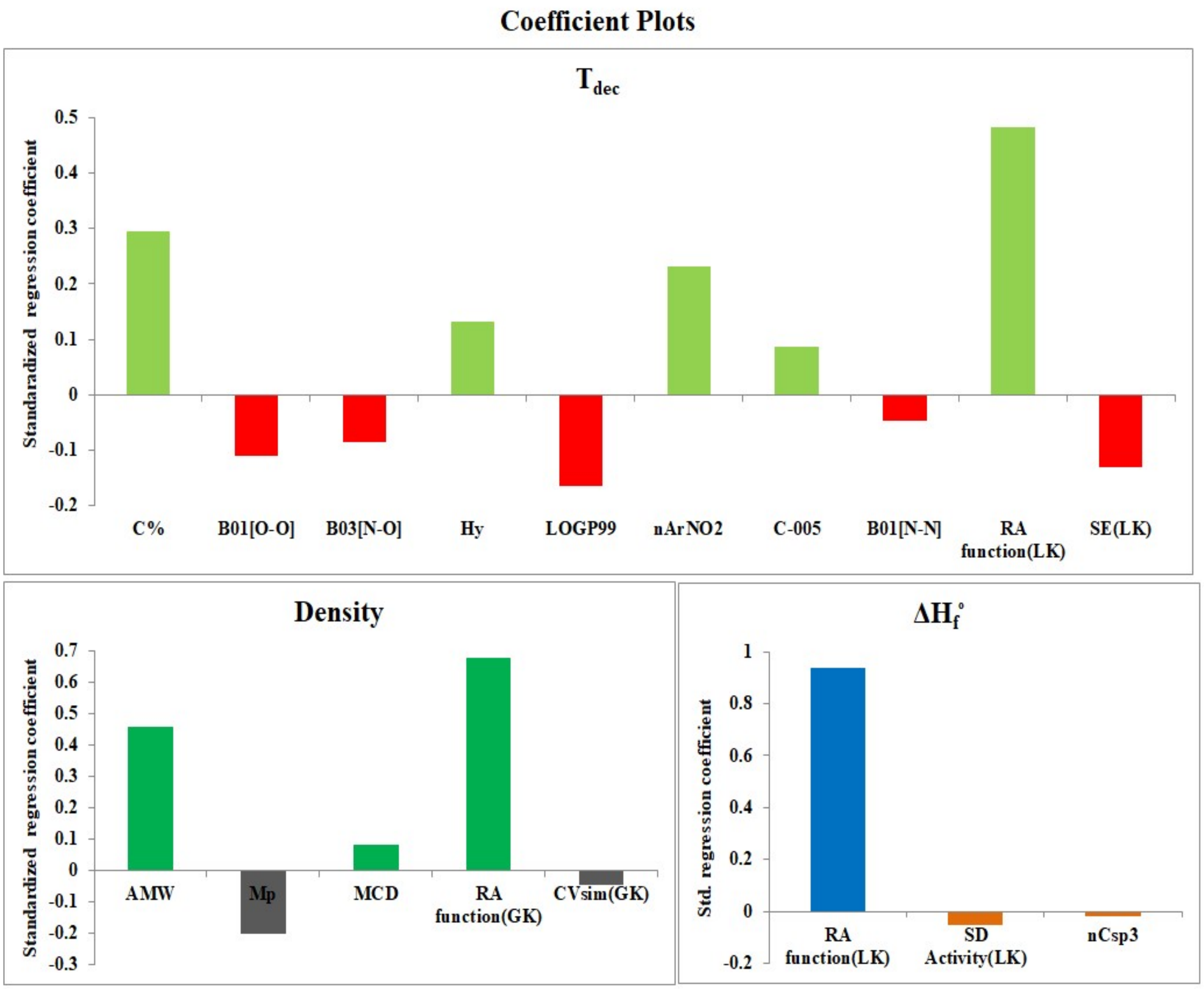
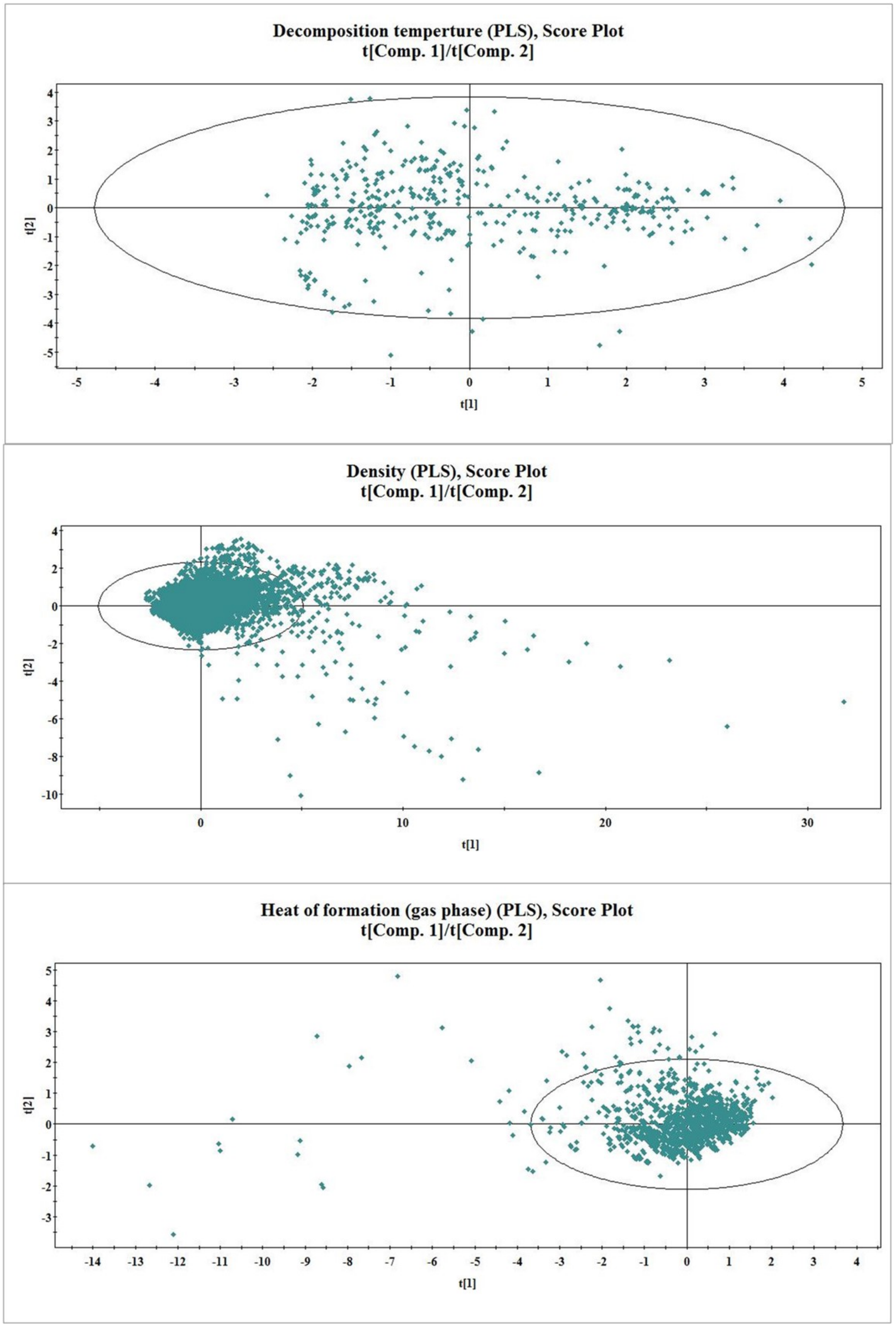


Figure S8: Coefficient Plots for each PLS model



Decomposition temperture (PLS), Score Plot
t[Comp. 1]/t[Comp. 2]

Density (PLS), Score Plot
t[Comp. 1]/t[Comp. 2]

Heat of formation (gas phase) (PLS), Score Plot
t[Comp. 1]/t[Comp. 2]

Figure S9: PLS Score Plots for respective models

Table S3: Comparison between the performances of different q-RASPR models for decomposition temperature (T_{dec})

T_{dec}	**Training set statistics**						**Test set statistics**			
Models	R^2	Q^2_{LOO}	MAE_C	MAE_C± SEM (5-foldCV)	MAE_C± SEM (10-foldCV)	$RMSE_C$	Q^2_{F1}	Q^2_{F2}	MAE_P	$RMSE_P$
RF	0.935	0.527	0.187	0.54 ± 0.036	0.53 ± 0.035	0.254	0.633	0.633	0.477	0.604
AB	0.632	0.496	0.505	0.58 ± 0.036	0.56 ± 0.028	0.606	0.564	0.564	0.557	0.658
GB	0.853	0.559	0.295	0.54 ± 0.036	0.53 ± 0.038	0.383	0.594	0.594	0.507	0.635
XGB	0.937	0.501	0.189	0.56 ± 0.040	0.55 ± 0.035	0.250	0.591	0.591	0.523	0.637
SVM	0.687	0.544	0.409	0.54 ± 0.031	0.54 ± 0.032	0.559	0.674	0.674	**0.456**	0.569
LSVM	0.613	0.605	0.469	0.49 ± 0.031	0.48 ± 0.028	0.621	0.662	0.662	0.468	0.574
RR	0.621	0.600	0.474	0.50 ± 0.027	0.49 ± 0.028	0.615	0.674	0.674	0.468	0.569
PLS	0.620	0.600	0.474	0.49 ± 0.027	0.49 ± 0.028	0.616	**0.676**	**0.676**	0.463	**0.567**

Table S4: Comparison between the performances of different q-RASPR models for density (Den)

Density Models	**Training set statistics**							**Test set statistics**			
	R^2	$R^2 \pm$ SEM (5-fold CV)	$R^2 \pm$ SEM (10-fold CV)	MAE_C	$MAE_C \pm$ SEM (5-fold CV)	$MAE_C \pm$ SEM (10-fold CV)	$RMSE_C$	Q^2_{F1}	Q^2_{F2}	MAE_P	$RMSE_P$
RF	0.991	0.92 ± 0.004	0.92 ± 0.006	0.066	0.19±0.009	0.19±0.006	0.931	0.936	0.931	0.182	0.250
AB	0.913	0.89 ± 0.013	0.88 ± 0.009	0.224	0.23±0.004	0.23±0.006	0.295	0.905	0.905	0.227	0.305
GB	0.947	0.92 ± 0.004	0.92 ± 0.006	0.172	0.19±0.004	0.19±0.006	0.230	0.932	0.932	0.184	0.257
XGB	0.911	0.87 ± 0.004	0.88 ± 0.009	0.205	0.23±0.009	0.22±0.009	0.298	0.905	0.905	0.215	0.303
SVM	0.915	0.87 ± 0.022	0.88 ± 0.016	0.172	0.19±0.009	0.19±0.009	0.292	0.916	0.916	0.178	0.286
LSVM	0.940	0.93± 0.004	0.92 ± 0.003	0.178	0.18±0.004	0.18±0.006	0.247	**0.939**	**0.939**	**0.177**	0.245
RR	0.940	0.93 ± 0.004	0.93 ± 0.006	0.179	0.18±0.004	0.18±0.006	0.244	**0.939**	**0.939**	0.178	**0.243**
PLS	0.940	0.93 ± 0.004	0.92 ± 0.006	0.180	0.18±0.004	0.18±0.006	0.246	**0.939**	**0.939**	0.180	0.244

Table S5: Comparison between the performance of different q-RASPR models forthe heat of formation (ΔH_f°)

ΔH_f°	**Training set statistics**								**Test set statistics**			
Models	R^2	Q^2_{LOO}	$R^2 \pm$ SEM (5-fold CV)	$R^2 \pm$ SEM (10-fold CV)	MAE_C	$MAE_C \pm$ SEM (5-foldCV)	$MAE_C \pm$ SEM (10-foldCV)	$RMSE_C$	Q^2_{F1}	Q^2_{F2}	MAE_P	$RMSE_P$
RF	0.991	0.934	0.86 ± 0.004	0.87 ± 0.013	0.054	0.18± 0.0031	0.17± 0.028	0.096	0.913	0.913	0.123	0.1758
AB	0.926	0.905	0.82 ± 0.022	0.83 ± 0.016	0.190	0.22± 0.027	0.21± 0.022	0.271	0.879	0.879	0.156	0.207
GB	0.968	0.933	0.88 ± 0.009	0.88 ± 0.016	0.118	0.17± 0.027	0.16± 0.025	0.180	0.925	0.925	0.114	0.163
XGB	0.935	0.897	0.82 ± 0.027	0.79 ± 0.028	0.146	0.20± 0.036	0.20± 0.028	0.255	0.899	0.899	0.137	0.189
SVM	0.827	0.761	0.74 ± 0.094	0.79 ± 0.054	0.154	0.21± 0.058	0.15± 0.044	0.416	0.928	0.928	0.110	0.159
LSVM	0.942	0.942	0.91 ± 0.013	0.90 ± 0.013	0.141	0.14± 0.018	0.19±0.019	0.240	0.930	0.930	**0.108**	0.157
RR	0.943	0.942	0.91 ± 0.013	0.90 ± 0.013	0.142	0.14± 0.018	0.14± 0.016	0.239	**0.931**	**0.931**	**0.108**	**0.156**
PLS	0.943	0.942	0.91 ± 0.013	0.90 ± 0.013	0.143	0.15± 0.018	0.14± 0.016	0.239	**0.931**	**0.931**	0.109	**0.156**