Machine Learning - Based q-RASPR Modeling of Power Conversion Efficiency of Organic Dyes in Dye-Sensitized Solar Cells

Souvik Pore, Arkaprava Banerjee, Kunal Roy*

Drug Theoretics and Chemoinformatics Laboratory,

Department of Pharmaceutical Technology, Jadavpur University,

188 Raja S C Mullick Road, 700032, Kolkata, India

Supplementary Materials SI-3

Corresponding author

Prof. Kunal Roy, Phone: +91 9831594140; Fax: +91-33-2837-1078;

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

A partial dependence plot shows the marginal effect of a feature (or two features) on the predicted outcome of a machine learning model. This plot can suggest the dependence interaction between two features. In case of an interaction with the other feature, a distinct vertical pattern of coloring will be seen.

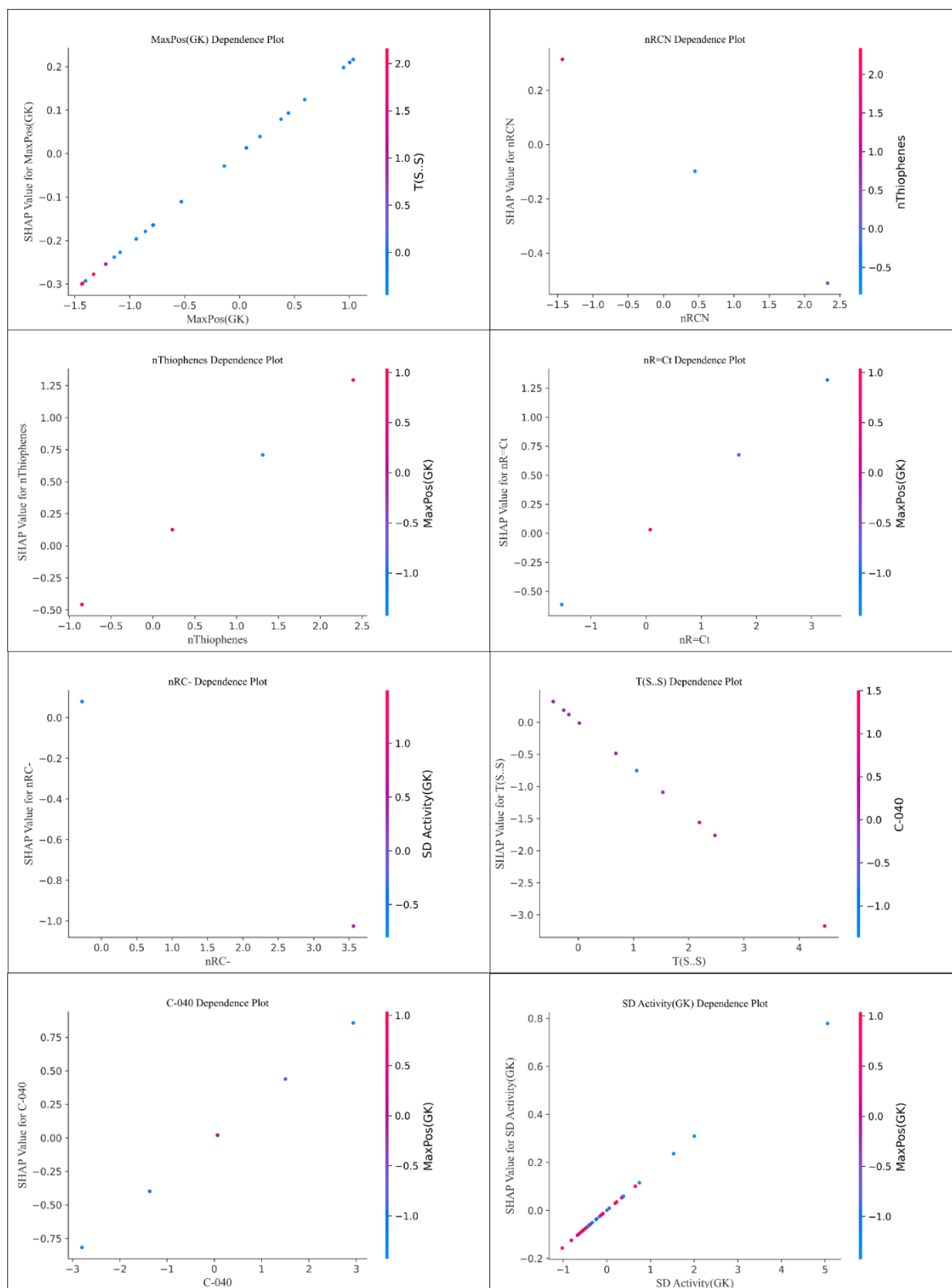# Dependence Plots for the Coumarin Dataset:

XGBoost:

Fig. S13. SHAP dependence plot of the XGBoost model for the Coumarin Dataset. In this plot, the feature value is represented along the x axis and the corresponding shap values are represented along the y-axis. Interaction with other descriptors is represented through the color gradient.
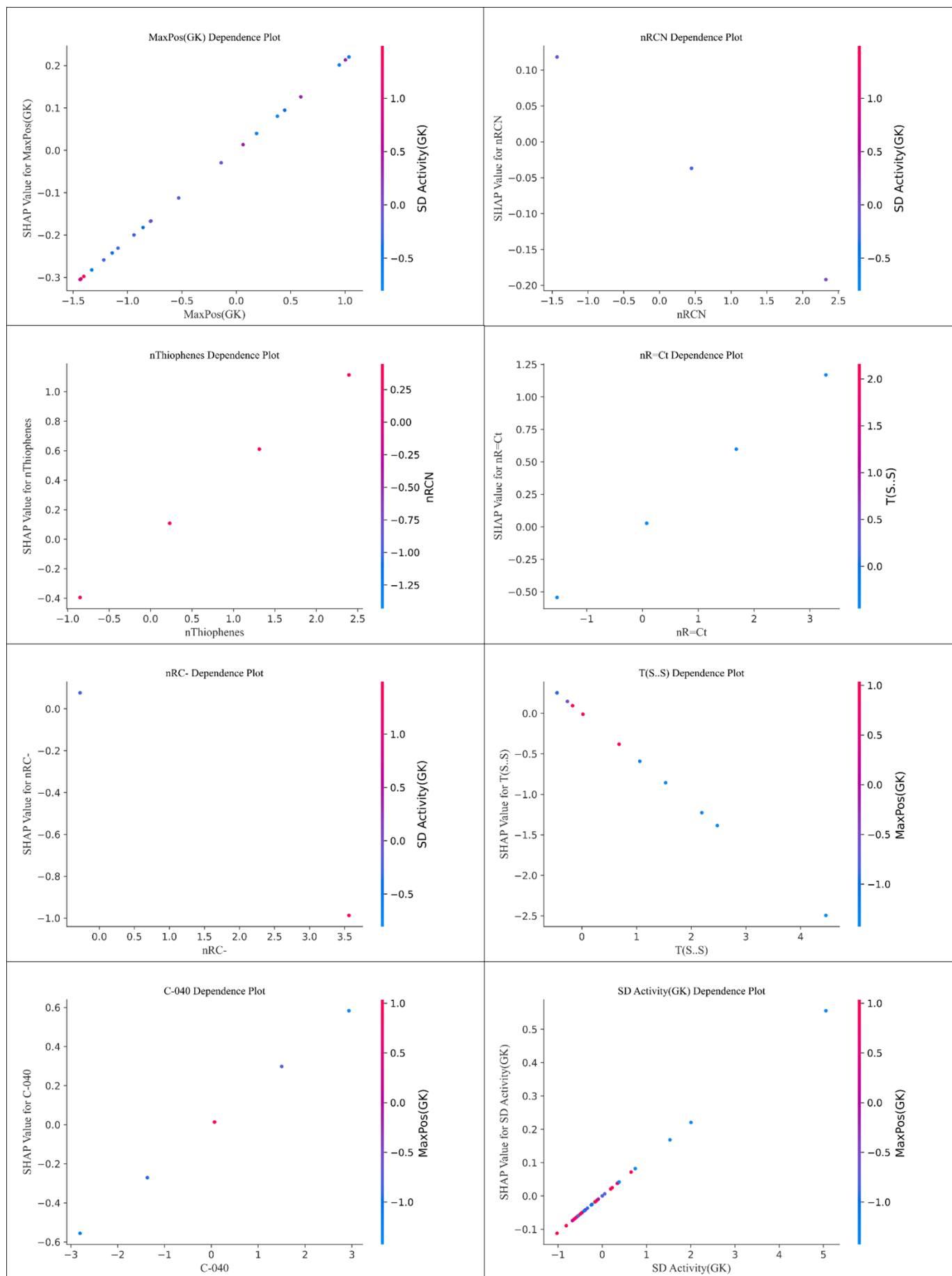
Ridge Regression:

Fig. S14. SHAP dependence plot of the Ridge Regression model for the Coumarin Dataset. In this plot, the feature value is represented along the x axis and the corresponding shap values are represented along the y-axis. Interaction with other descriptors is represented through the color gradient.
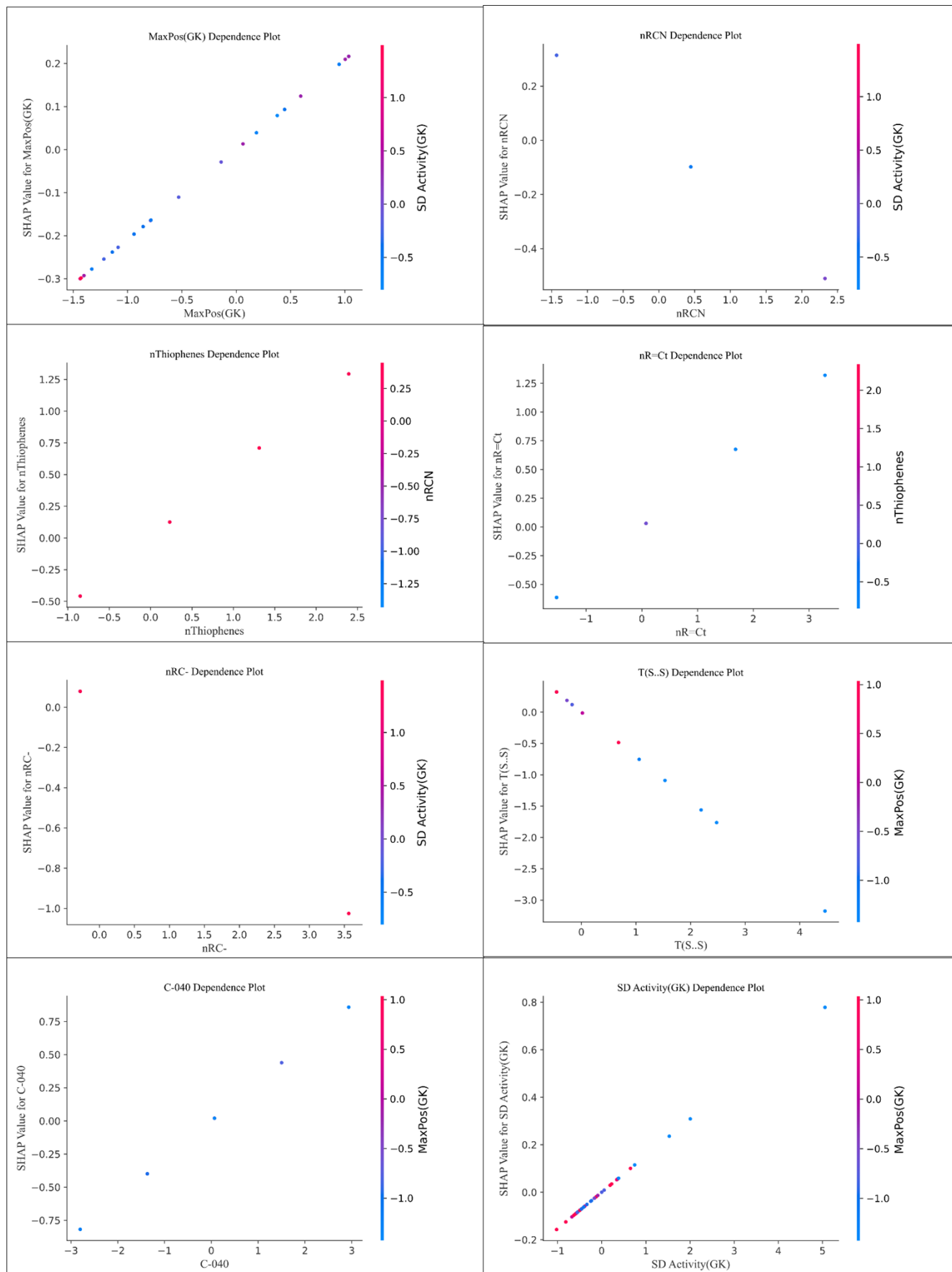
PLS:

Fig. S15. SHAP dependence plot of the PLS model for the Coumarin Dataset. In this plot, the feature value is represented along the x axis and the corresponding shap values are represented along the y-axis. Interaction with other descriptors is represented through the color gradient.
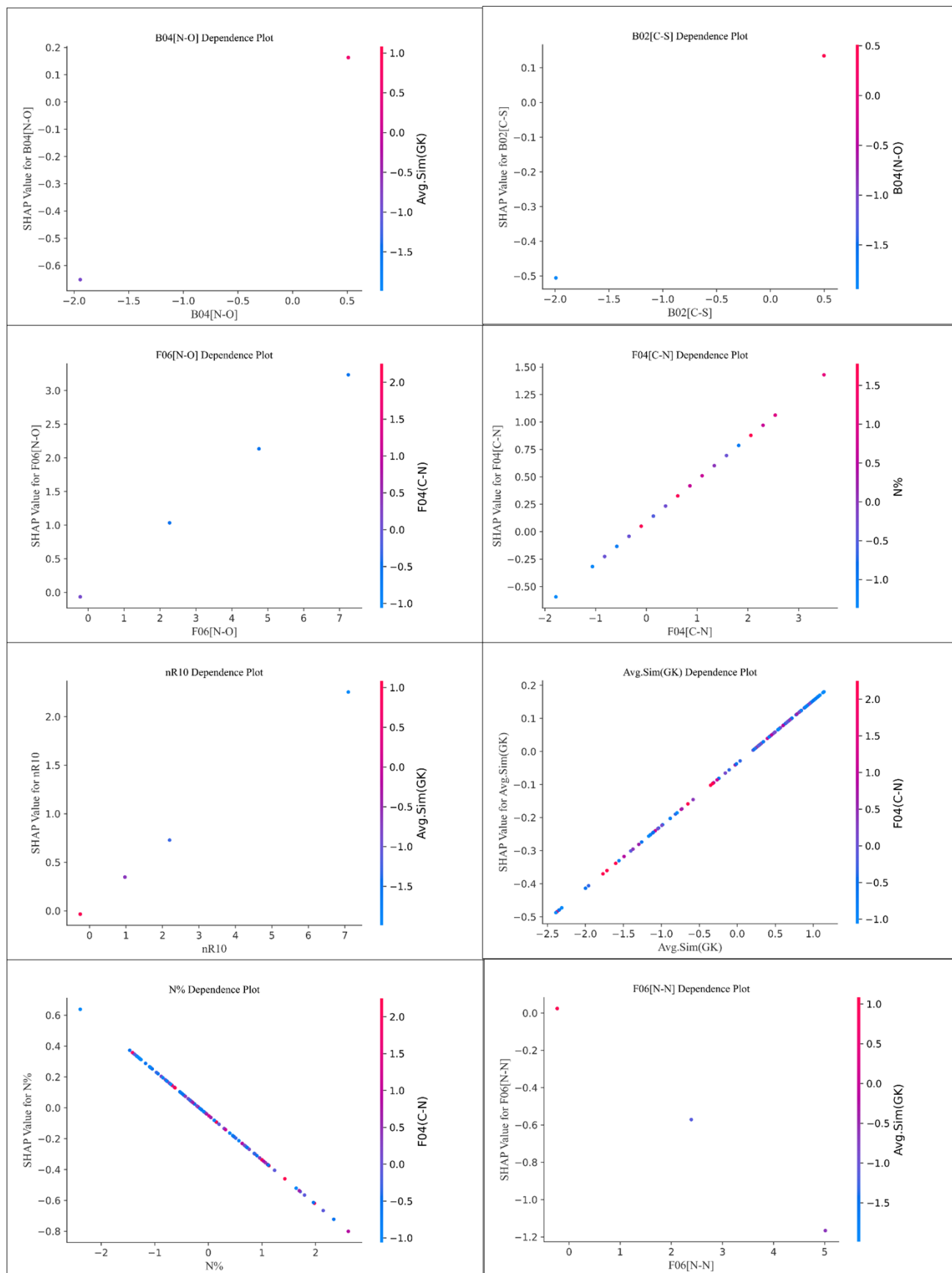
Dependence Plots for the Carbazole Dataset:

XGBoost:

Fig. S16. SHAP dependence plot of the XGBoost model for the Carbazole Dataset. In this plot, the feature value is represented along the x axis and the corresponding shap values are represented along the y-axis. Interaction with other descriptors is represented through the color gradient.
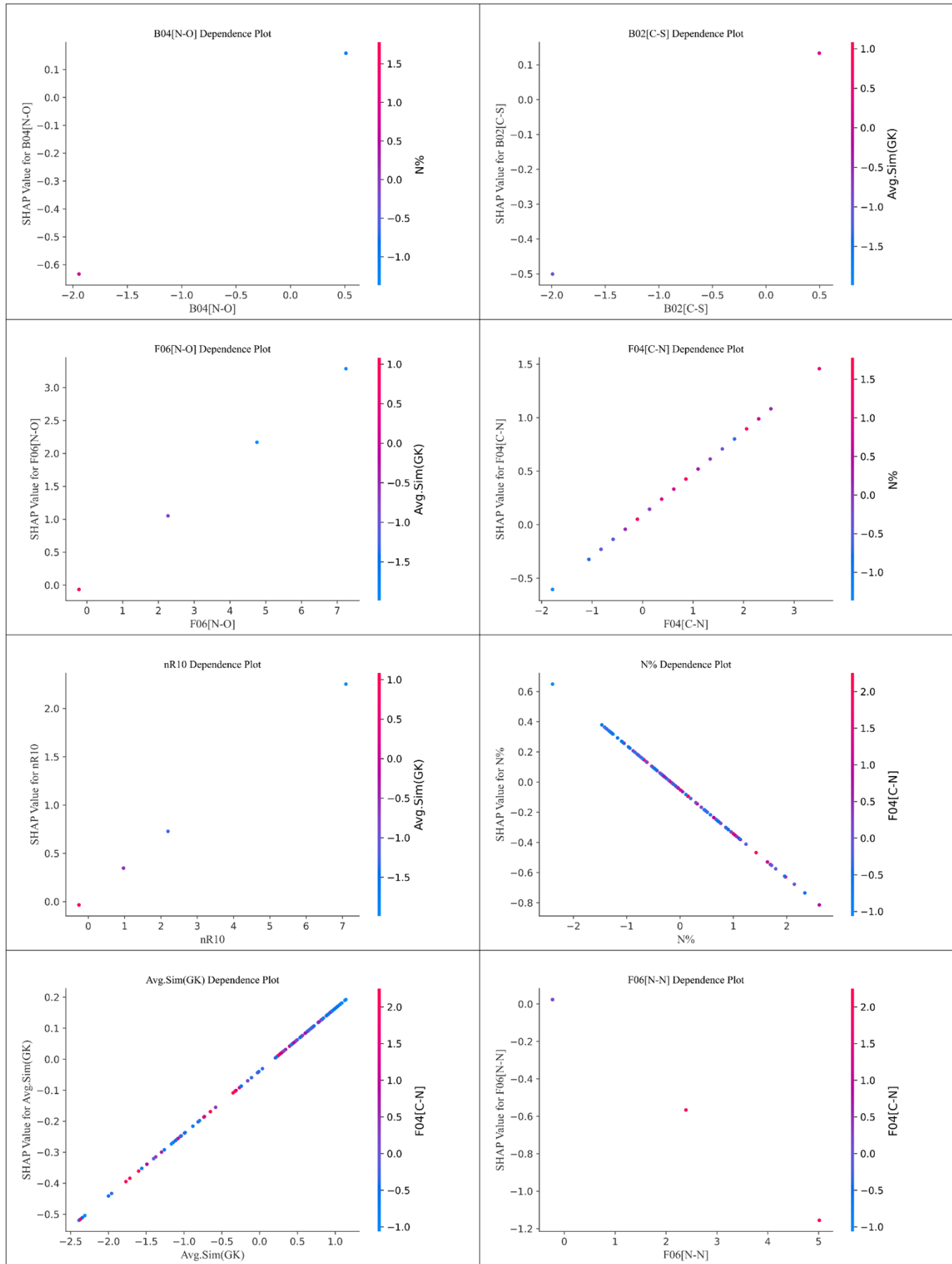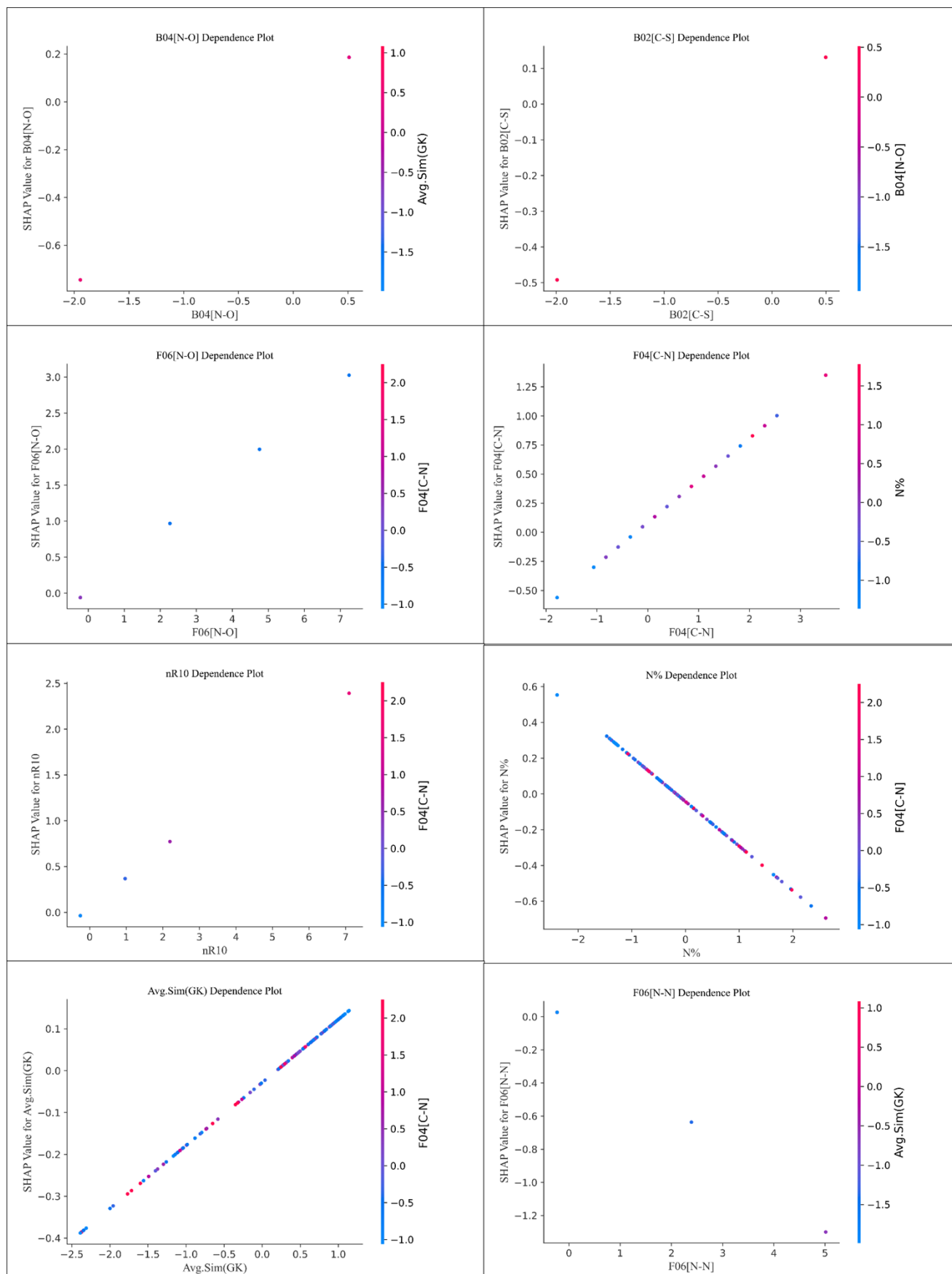
Ridge Regression:

Fig. S17. SHAP dependence plot of the Ridge Regression model for the Carbazole Dataset. In this plot, the feature value is represented along the x axis and the corresponding shap values are represented along the y-axis. Interaction with other descriptors is represented through the color gradient.

PLS:

Fig. S18. SHAP dependence plot of the PLS model for the Carbazole Dataset. In this plot, the feature value is represented along the x axis and the corresponding shap values are represented along the y-axis. Interaction with other descriptors is represented through the color gradient.
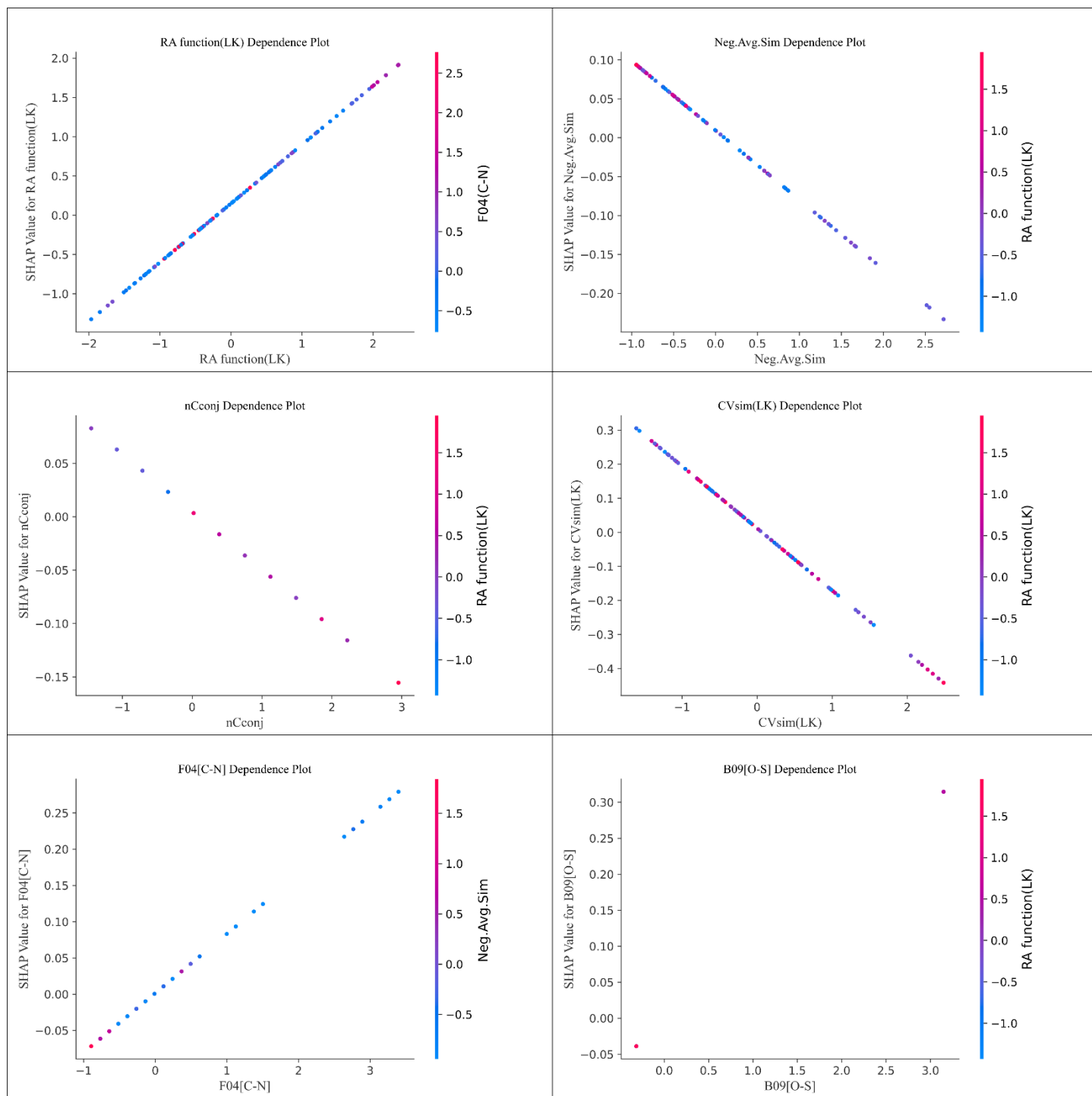
Dependence Plots for the Indoline Dataset:

XGBoost:

Fig. S19. SHAP dependence plot of the XGBoost model for the Indoline Dataset. In this plot, the feature value is represented along the x axis and the corresponding shap values are represented along the y-axis. Interaction with other descriptors is represented through the color gradient.
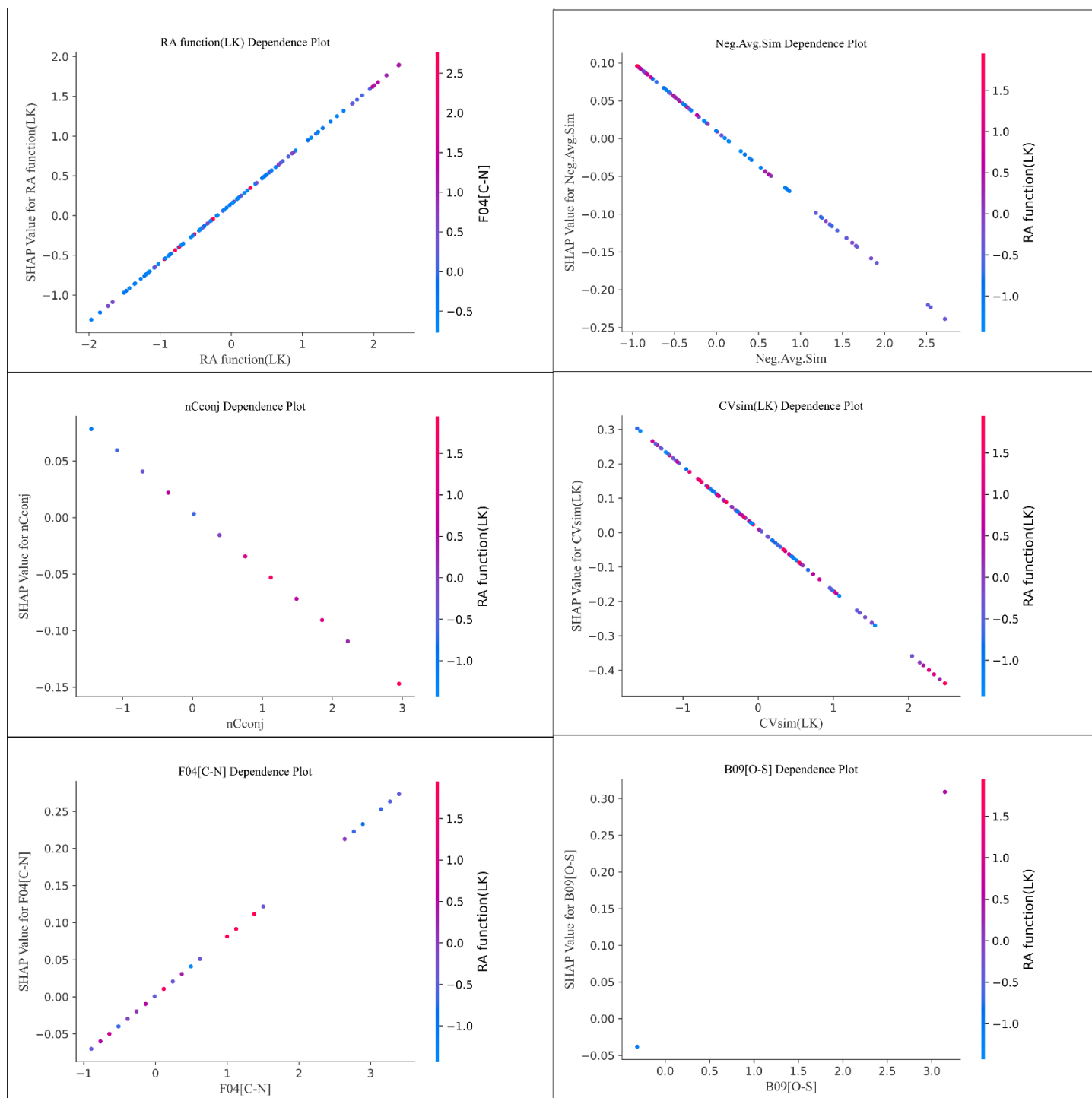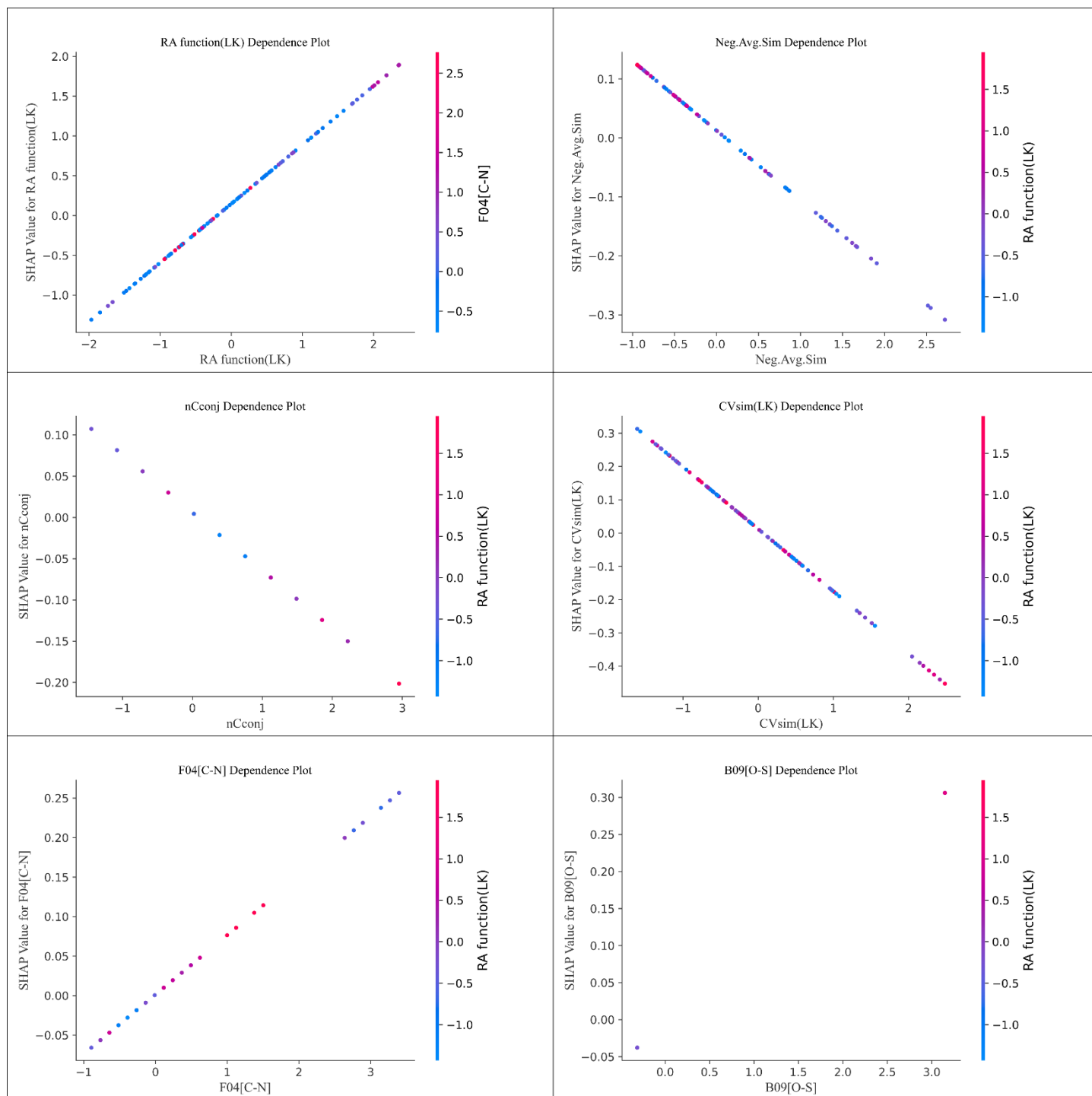
Ridge Regression:

Fig. S20. SHAP dependence plot of the Ridge Regression model for the Indoline Dataset. In this plot, the feature value is represented along the x axis and the corresponding shap values are represented along the y-axis. Interaction with other descriptors is represented through the color gradient.

PLS:

Fig. S21. SHAP dependence plot of the PLS model for the Indoline Dataset. In this plot, the feature value is represented along the x axis and the corresponding shap values are represented along the y-axis. Interaction with other descriptors is represented through the color gradient.

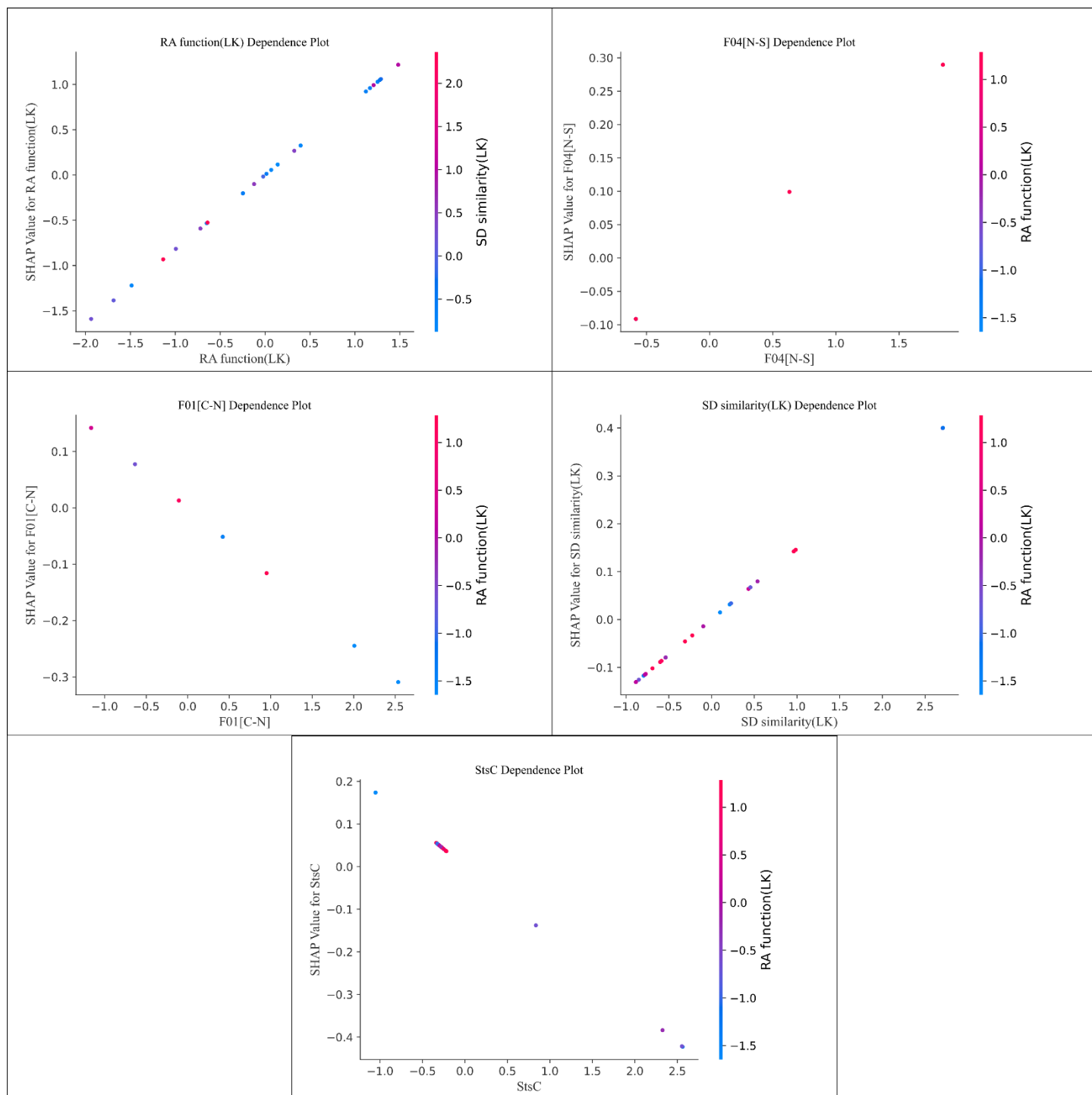# Dependence Plots for the Diphenylamine Dataset:

XGBoost:

Fig. S22. SHAP dependence plot of the XGBoost model for the Diphenylamine Dataset. In this plot, the feature value is represented along the x axis and the corresponding shap values are represented along the y-axis. Interaction with other descriptors is represented through the color gradient.
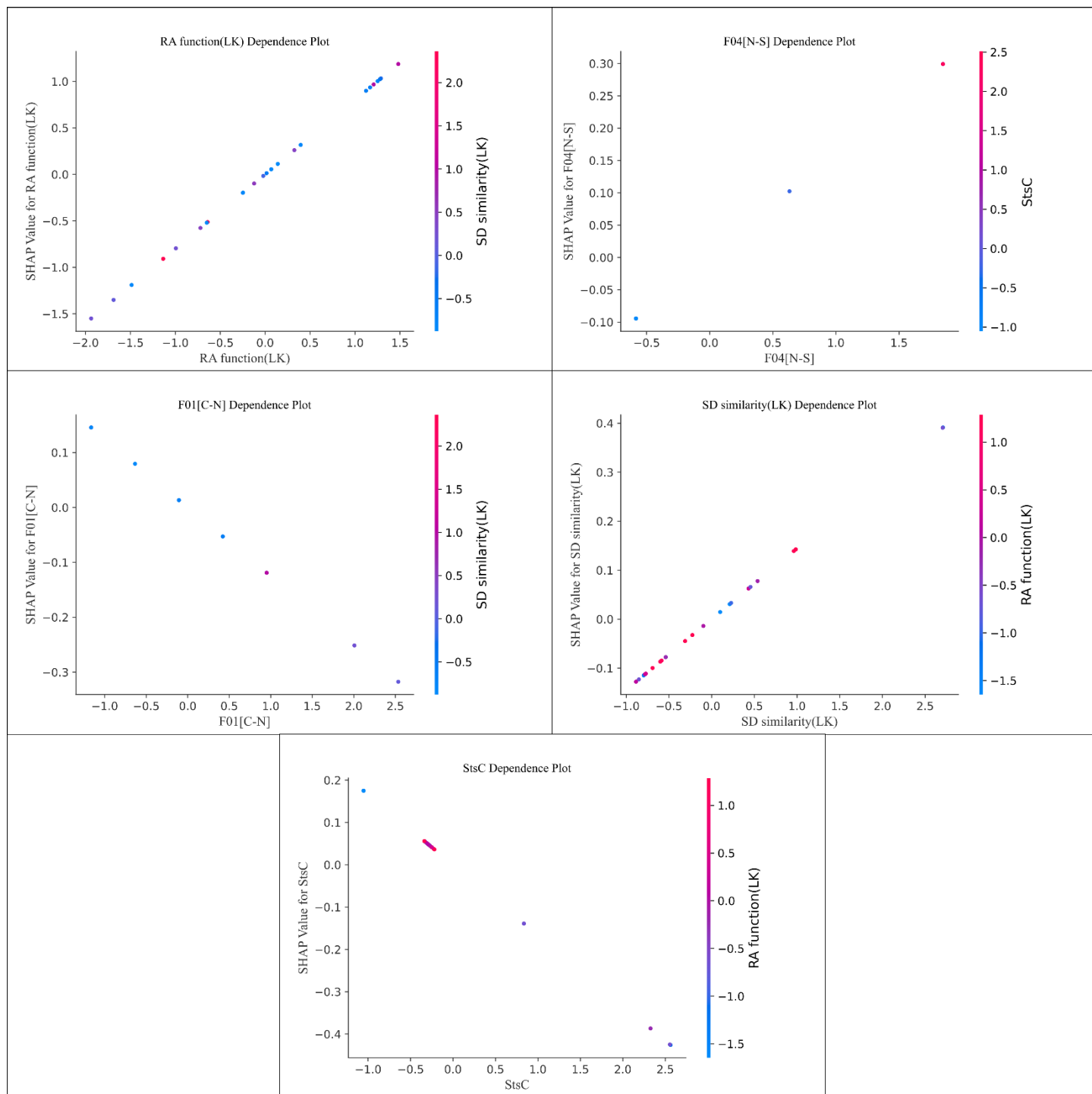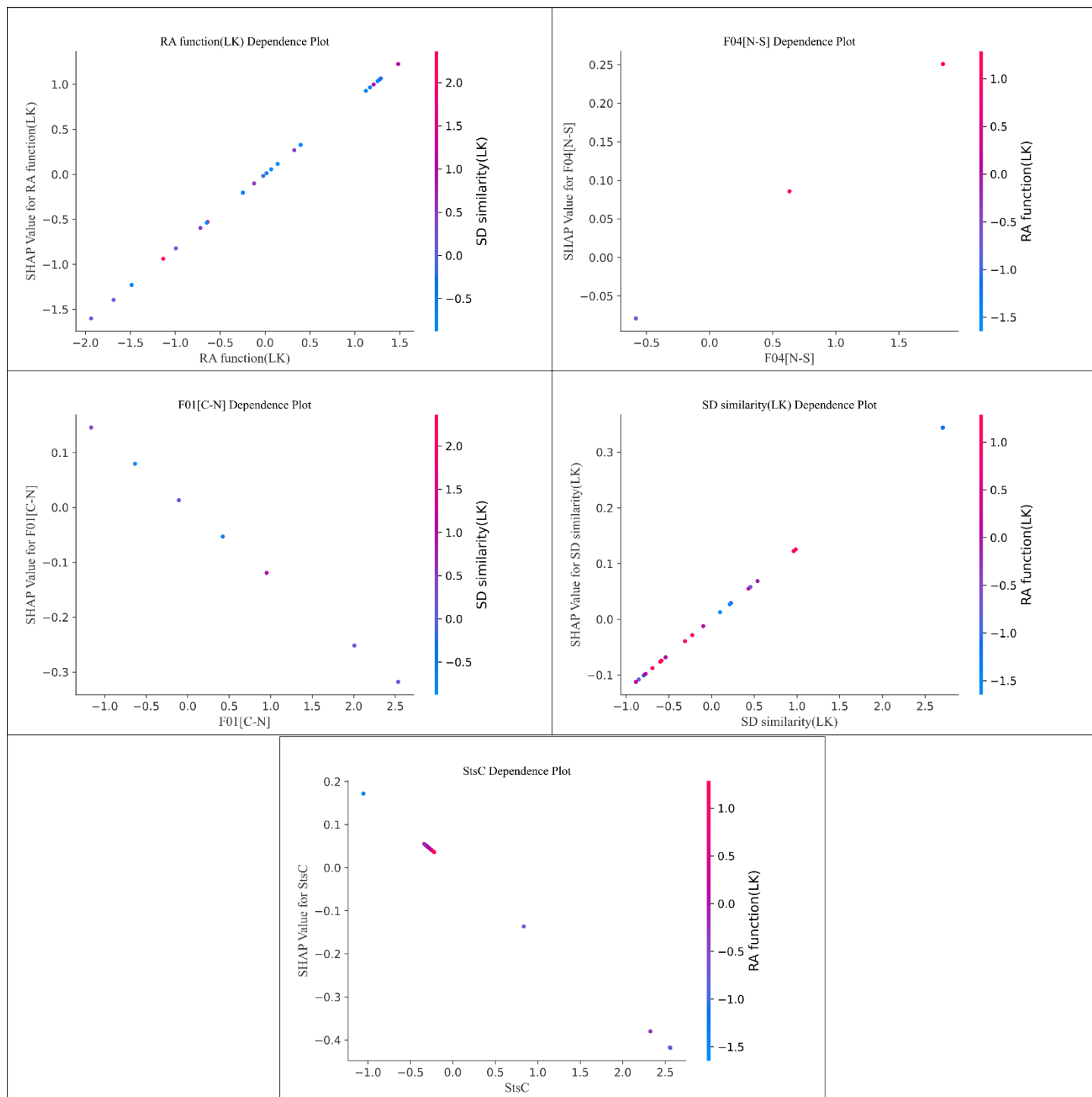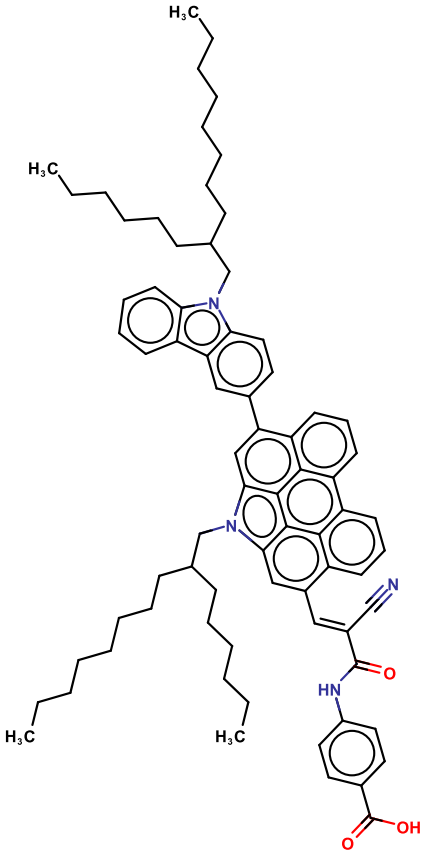
Ridge Regression:

Fig. S23. SHAP dependence plot of the Ridge Regression model for the Diphenylamine Dataset. In this plot, the feature value is represented along the x axis and the corresponding shap values are represented along the y-axis. Interaction with other descriptors is represented through the color gradient.

PLS:

Fig. S24. SHAP dependence plot of the PLS model for the Diphenylamine Dataset. In this plot, the feature value is represented along the x axis and the corresponding shap values are represented along the y-axis. Interaction with other descriptors is represented through the color gradient.

Table S1. List of designed dyes and their predicted PCE values

Compounds	Structure	Predicted PCE Values			SA Score
		PLS	Ridge Regression	XGBoost	
Carbazoles					
NCA1		14.15	14.02	13.97	8.93

NCA2		15.22	15.09	15.06	9.55
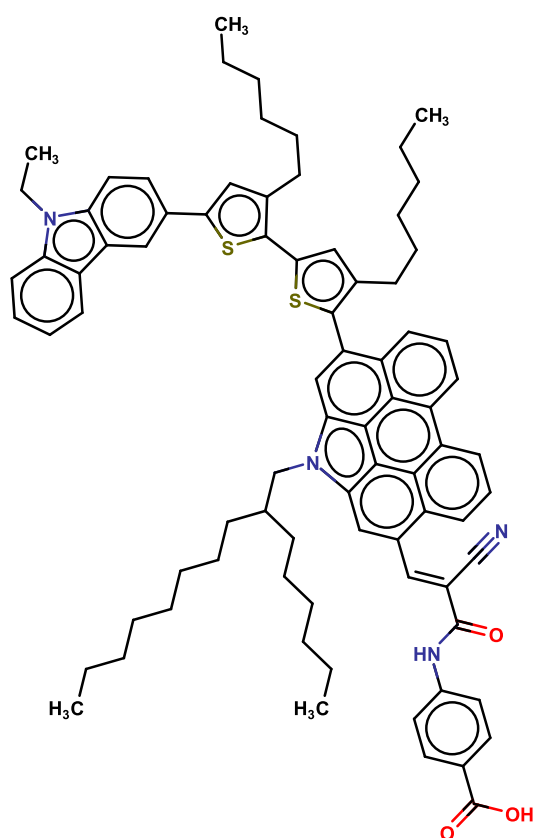
NCA3		15.79	15.73	15.69	10
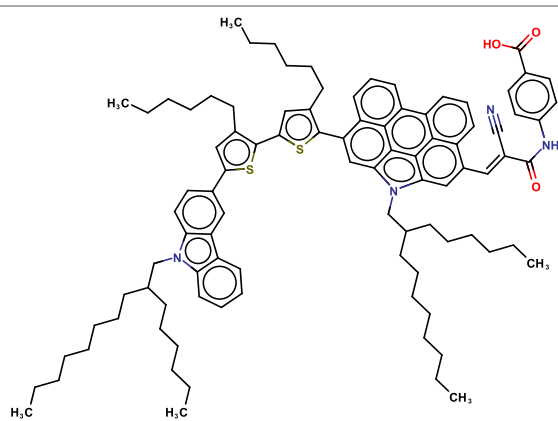
NCA4		18.18	18.33	18.24	9.61

| **NCA5** | | 11.47 | 11.89 | 11.81 | 9.61 |

Coumarins
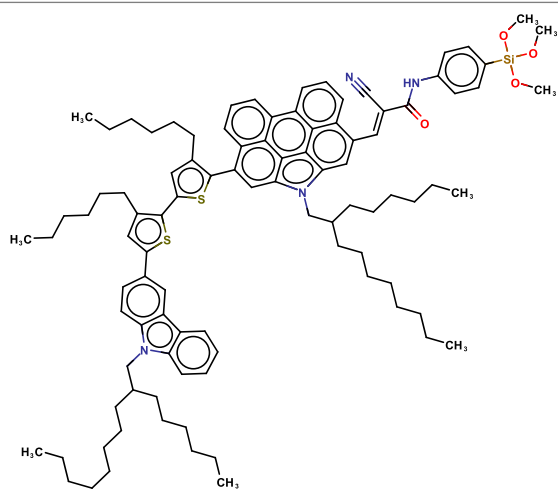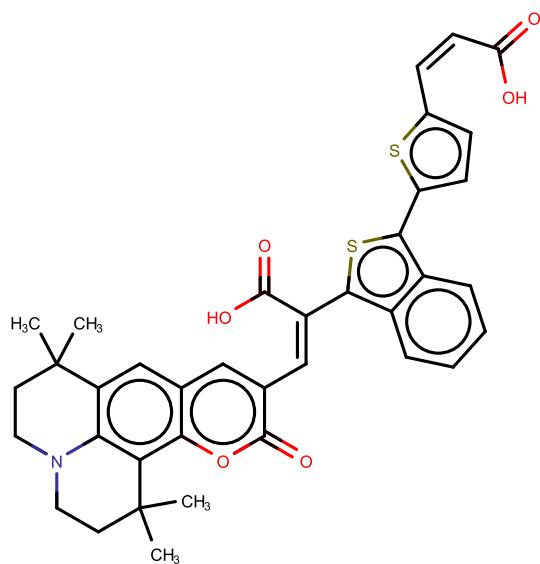
| **NCO1** | | 5.67 | 5.1 | 5.67 | 4.85 |

| NCO2 | 5.67 | 5.1 | 5.67 | 5.09 |

| NCO3 | 6.74 | 6.03 | 6.74 | 5.21 |

| NCO4 | 6.74 | 6.03 | 6.74 | 5.21 |
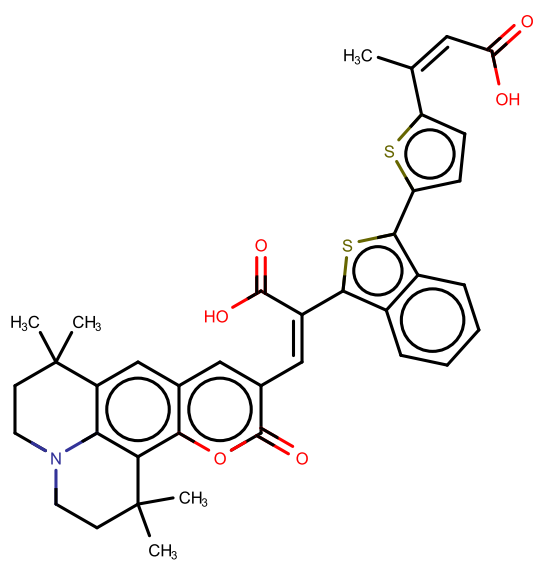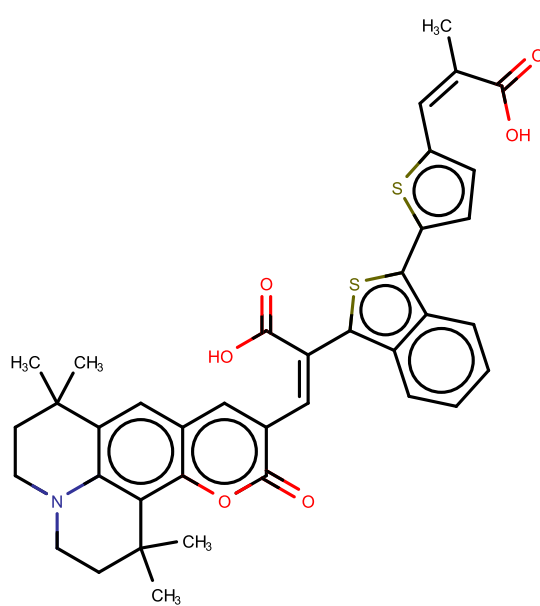
NCO5		6.6	5.94	6.6	4.74
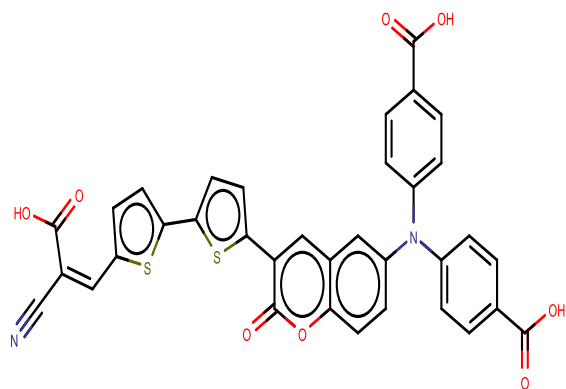
Diphenylamines

NDI1		7.06	7.05	7.1	4.86
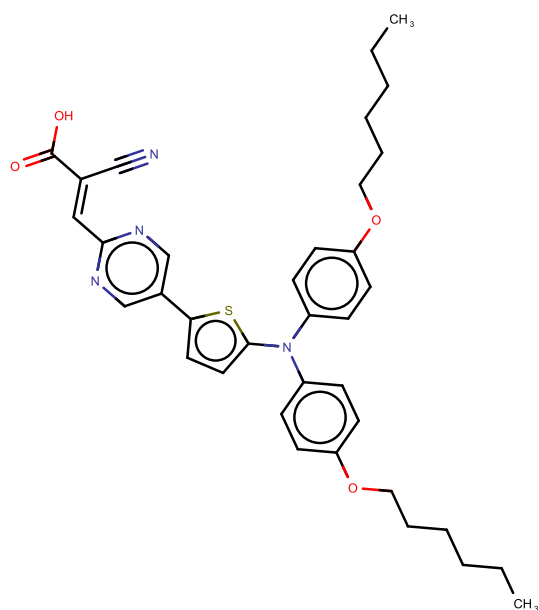
NDI2		6.41	6.33	6.4	6.44
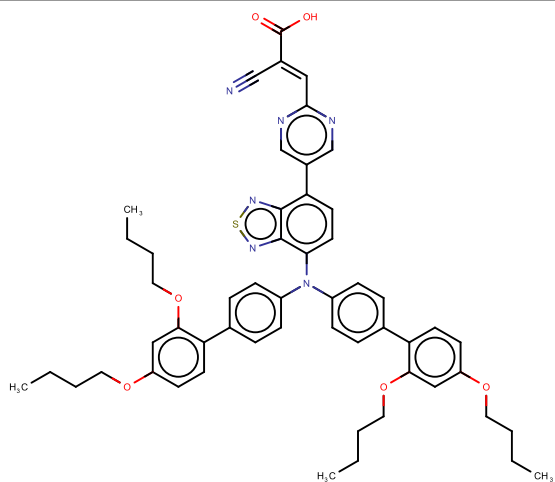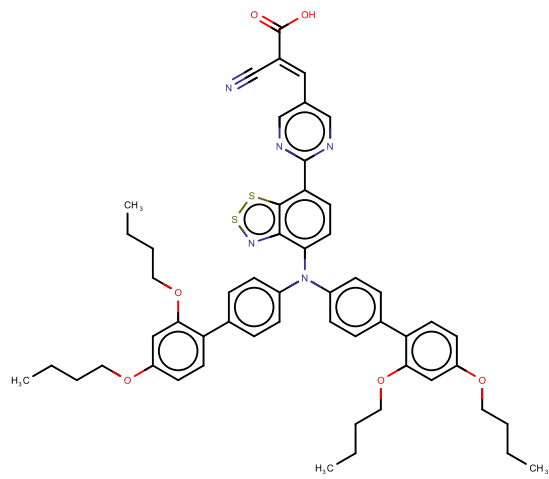
NDI3		7.62	7.75	7.77	7.8
NDI4		7.07	7.22	7.22	4.2
NDI5		7.16	7.16	7.19	9.42
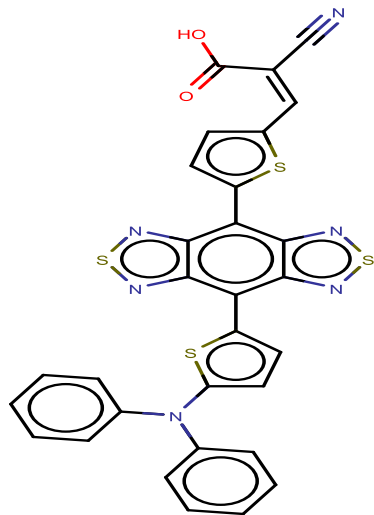
Indolines
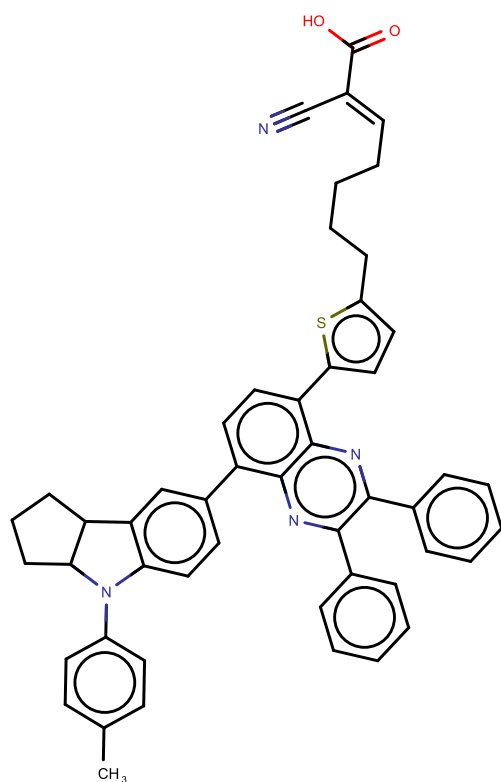
NIN1	6.39	6.34	6.36	6.25
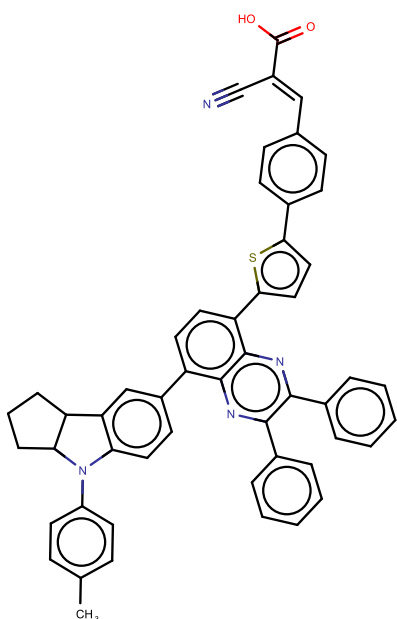
NIN2	6.39	6.34	6.36	6.14

| NIN3 | 8.22 | 8.15 | 8.19 | 6.58 |

| NIN4 | 8.29 | 8.2 | 8.23 | 7.54 |

| NIN5 | 8.49 | 8.41 | 8.45 | 7.71 |

SA Score: Synthetic Accessibility score