Machine Learning - driven insights for predicting the impact of nanoparticles on the functionality of biomolecules, Illustrated by the case of DNA Damage-Inducible Transcript 3 (CHOP) inhibitors

Electronic Supplementary Material

Machine Learning - driven insights for predicting the impact of nanoparticles on the functionality of biomolecules, Illustrated by the case of DNA Damage-Inducible Transcript 3 (CHOP) inhibitors

Mariya L. Ivanova[1,*], Nicola Russo[1], Gueorgui Mihaylov[2], Konstantin Nikolic[1]

Author affiliations
[1]School of Computing and Engineering, University of West London, London, UK
[2]Haleon, London, UK
*Corresponding author mariya.ivanova@uwl.ac.uk

Tables

Table ESM1 Metrics of ML models based on Dataset 1 (13C NMR concise data)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
SVM	0.779	0.798	0.747	0.772	0.779
RandomForest	0.713	0.842	0.524	0.646	0.713
GradientBoost	0.706	0.716	0.685	0.700	0.706
Decision	0.634	0.663	0.544	0.598	0.634
K-nearest	0.634	0.760	0.392	0.517	0.634

Table ESM2. Five-fold cross-validation for ML based on Dataset 1 (13C NMR concise data)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8642	0.0470	[0.8331, 0.8294, 0.8161, 0.9189, 0.9237]
SVM	0.8127	0.0122	[0.8129, 0.8019, 0.7975, 0.8311, 0.8203]
Decision	0.7686	0.0528	[0.7351, 0.7238, 0.7187, 0.8263, 0.8391]
K-nearest	0.7386	0.0399	[0.7085, 0.7143, 0.6964, 0.783, 0.7905]
GradientBoost	0.7223	0.0072	[0.7295, 0.7238, 0.709, 0.7272, 0.722]

Table ESM3 Metrics of ML models based on Dataset 2 (13C NMR extensive data)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
SVM	0.793	0.826	0.741	0.782	0.793
RandomForest	0.711	0.854	0.510	0.639	0.711
GradientBoost	0.673	0.696	0.613	0.652	0.673
Decision	0.653	0.683	0.572	0.623	0.653
K-nearest	0.644	0.789	0.392	0.524	0.644

Table ESM4. Five-fold cross-validation for ML based on Dataset 2 (13C NMR extensive data)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8635	0.0555	[0.8196, 0.8189, 0.8163, 0.9284, 0.9344]
SVM	0.8455	0.0192	[0.8366, 0.8264, 0.8276, 0.8714, 0.8656]
Decision	0.7868	0.0511	[0.7518, 0.7401, 0.7437, 0.8481, 0.8504]
K-nearest	0.7287	0.0405	[0.7005, 0.7003, 0.6879, 0.7678, 0.787]
GradientBoost	0.6792	0.0049	[0.6813, 0.6743, 0.6727, 0.6854, 0.6822]

Table ESM5 Metrics of ML models based on Dataset 3 (1H NMR concise data)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
SVM	0.677	0.691	0.638	0.664	0.677
GradientBoost	0.677	0.680	0.667	0.673	0.677
RandomForest	0.649	0.678	0.566	0.617	0.649
Decision	0.618	0.643	0.528	0.580	0.618
K-nearest	0.613	0.669	0.445	0.535	0.613

Table ESM6. Five-fold cross-validation for ML based on Dataset 3 (1H NMR concise data)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.7690	0.0328	[0.7483, 0.7433, 0.7357, 0.8071, 0.8108]
Decision	0.7342	0.0388	[0.7058, 0.6995, 0.7027, 0.7773, 0.7858]
K-nearest	0.6891	0.0277	[0.6705, 0.667, 0.6627, 0.717, 0.7282]
SVM	0.6872	0.0024	[0.6895, 0.6905, 0.6849, 0.6859, 0.6849]
GradientBoost	0.6835	0.0045	[0.6923, 0.6813, 0.6804, 0.6832, 0.6802]

Table ESM7 Metrics of ML models based on Dataset 4 (1H NMR extensive data)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
SVM	0.695	0.715	0.649	0.680	0.695
GradientBoost	0.679	0.690	0.649	0.669	0.679
RandomForest	0.663	0.754	0.484	0.589	0.663
K-nearest	0.602	0.661	0.419	0.513	0.602
Decision	0.595	0.615	0.506	0.555	0.595

Table ESM8. Five-fold cross-validation for ML based on Dataset 4 (1H NMR extensive data)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8205	0.0548	[0.7693, 0.7833, 0.775, 0.8851, 0.8896]
Decision	0.7500	0.0589	[0.7048, 0.6998, 0.7015, 0.8166, 0.8273]
SVM	0.7341	0.0152	[0.7205, 0.728, 0.7175, 0.7505, 0.754]
K-nearest	0.6910	0.0423	[0.663, 0.6535, 0.6532, 0.743, 0.7422]
GradientBoost	0.6857	0.0078	[0.672, 0.692, 0.6824, 0.6927, 0.6894]

Table ESM9 Metrics of ML models based on Dataset 5 (13C NMR concise data & 1H NMR concise data

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
SVM	0.778	0.796	0.747	0.771	0.778
RandomForest	0.737	0.842	0.582	0.689	0.737
GradientBoost	0.705	0.715	0.680	0.697	0.705
K-nearest	0.656	0.719	0.513	0.599	0.656
Decision	0.634	0.663	0.544	0.598	0.634

Table ESM10. Five-fold cross-validation for ML based on Dataset 5 (13C NMR concise data & 1H NMR concise data)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8608	0.0420	[0.8289, 0.8271, 0.8238, 0.9102, 0.9142]
SVM	0.7921	0.0036	[0.7903, 0.7918, 0.7875, 0.7985, 0.792]
Decision	0.7777	0.0461	[0.7356, 0.7481, 0.7375, 0.8278, 0.8396]
K-nearest	0.7386	0.0360	[0.7103, 0.7133, 0.7045, 0.7868, 0.778]
GradientBoost	0.7276	0.0040	[0.7313, 0.7318, 0.7217, 0.729, 0.724]

Table ESM11 Metrics of ML models based on Dataset 6 (13C NMR extensive data & 1H NMR concise data)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
SVM	0.764	0.790	0.719	0.753	0.764
RandomForest	0.747	0.844	0.605	0.705	0.747
GradientBoost	0.694	0.699	0.681	0.690	0.694
K-nearest	0.676	0.759	0.516	0.614	0.676
Decision	0.668	0.698	0.593	0.641	0.668

Table ESM12. Five-fold cross-validation for ML based on Dataset 6 (13C NMR extensive data & 1H NMR concise data)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8722	0.0447	[0.8356, 0.8339, 0.8376, 0.9284, 0.9254]
SVM	0.7879	0.0084	[0.7806, 0.7866, 0.7775, 0.7995, 0.7953]
Decision	0.7879	0.0469	[0.7498, 0.7561, 0.7435, 0.8448, 0.8453]
K-nearest	0.7496	0.0354	[0.727, 0.7195, 0.7165, 0.786, 0.799]
GradientBoost	0.7103	0.0032	[0.7078, 0.7158, 0.7077, 0.7122, 0.708]

Table ESM13 Metrics of ML models based on Dataset 7 (13C NMR concise data & 1H NMR extensive data)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
SVM	0.778	0.796	0.747	0.771	0.778
RandomForest	0.714	0.828	0.540	0.654	0.714
GradientBoost	0.711	0.722	0.687	0.704	0.711
K-nearest	0.641	0.736	0.440	0.550	0.641
Decision	0.616	0.639	0.532	0.581	0.616

Table ESM14. Five-fold cross-validation for ML based on Dataset 7 (13C NMR concise data & 1H NMR extensive data)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8611	0.0464	[0.8271, 0.8261, 0.8168, 0.9132, 0.9222]
SVM	0.8161	0.0176	[0.8084, 0.8026, 0.7968, 0.8441, 0.8286]
Decision	0.7636	0.0570	[0.7108, 0.7283, 0.7132, 0.8268, 0.8388]
GradientBoost	0.7321	0.0107	[0.724, 0.7308, 0.7177, 0.7412, 0.7467]
K-nearest	0.7275	0.0451	[0.6985, 0.6865, 0.6874, 0.7833, 0.7818]

Table ESM15 Metrics of ML models based on Dataset 8 (13C NMR extensive data & 1H NMR extensive data)

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
SVM	0.792	0.824	0.742	0.781	0.792
RandomForest	0.713	0.853	0.515	0.642	0.713
GradientBoost	0.665	0.690	0.600	0.642	0.665
Decision	0.645	0.678	0.550	0.608	0.645
K-nearest	0.640	0.776	0.392	0.521	0.640

Table ESM16. Five-fold cross-validation for ML based on Dataset 8 (13C NMR extensive data & 1H NMR extensive data)

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8610	0.0591	[0.8144, 0.8134, 0.8106, 0.9312, 0.9354]
SVM	0.8441	0.0195	[0.8324, 0.8294, 0.8236, 0.8714, 0.8636]
Decision	0.7898	0.0502	[0.7513, 0.7556, 0.7402, 0.8483, 0.8534]
K-nearest	0.7317	0.0418	[0.7, 0.7058, 0.6884, 0.7748, 0.7895]
GradientBoost	0.6809	0.0130	[0.6673, 0.6765, 0.6709, 0.7035, 0.6862]

Table ESM17 Metrics of ML models based on Dataset 2 (13C NMR extensive data) and molecular features performed without PCA

	1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
2	RandomForest	0.830	0.880	0.764	0.818	0.830
3	GradientBoost	0.810	0.831	0.780	0.805	0.810
0	SVM	0.803	0.845	0.743	0.790	0.803
1	Decision	0.765	0.794	0.715	0.753	0.765
4	K-nearest	0.673	0.725	0.558	0.631	0.673

Table ESM18 Five-fold cross-validation for ML based on Dataset 2 (13C NMR extensive data) and molecular features performed without PCA

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8399	0.0057	[0.8311, 0.8464, 0.8397, 0.8454, 0.8367]
GradientBoost	0.8247	0.0047	[0.8162, 0.8297, 0.8262, 0.8277, 0.8238]
SVM	0.7887	0.0050	[0.7831, 0.7956, 0.7915, 0.7903, 0.7828]
Decision	0.7799	0.0110	[0.7752, 0.7962, 0.7744, 0.7885, 0.7654]
K-nearest	0.7687	0.0026	[0.767, 0.7723, 0.7664, 0.7715, 0.7664]

Table ESM19 Metrics of ML models based on Dataset 2 (13C NMR extensive data) and molecular features performed with PCA

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
RandomForest	0.828	0.882	0.757	0.815	0.828
GradientBoost	0.827	0.849	0.796	0.822	0.827
Decision	0.774	0.811	0.714	0.759	0.774
K-nearest	0.677	0.691	0.641	0.665	0.677
SVM	0.508	0.545	0.101	0.171	0.508

Table ESM20 Five-fold cross-validation for ML based on Dataset 2 (13C NMR extensive data) and molecular features performed with PCA

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.9021	0.0292	[0.8762, 0.8779, 0.8832, 0.9247, 0.9484]
Decision	0.8490	0.0286	[0.8211, 0.8299, 0.8291, 0.8697, 0.8951]
GradientBoost	0.8376	0.0075	[0.8284, 0.8314, 0.8371, 0.8494, 0.8416]
K-nearest	0.7329	0.0230	[0.713, 0.716, 0.715, 0.7521, 0.7685]
SVM	0.5139	0.0194	[0.5009, 0.5006, 0.5111, 0.5049, 0.5521]

Table ESM21 Metrics of ML models based on molecular features only

1.Algorithm	2.Accuracy	3.Precision	4.Recall	5.F1	6.ROC
GradientBoost	0.807	0.826	0.778	0.801	0.807
RandomForest	0.803	0.850	0.737	0.790	0.804
SVM	0.764	0.857	0.634	0.729	0.764
Decision	0.752	0.785	0.694	0.737	0.752
K-nearest	0.729	0.739	0.708	0.723	0.729

Table ESM22. Five-fold cross-validation for ML based on molecular features only

1.Algorithm	2.Mean CV Score	3.Standard Deviation	4.List of CV Scores
RandomForest	0.8871	0.0286	[0.866, 0.8676, 0.8593, 0.9132, 0.9296]
Decision	0.8416	0.0296	[0.8183, 0.8165, 0.8189, 0.8679, 0.8863]
GradientBoost	0.8099	0.0038	[0.8053, 0.8165, 0.8082, 0.8087, 0.811]
K-nearest	0.7810	0.0221	[0.7615, 0.7688, 0.7589, 0.8071, 0.8084]
SVM	0.7660	0.0031	[0.7626, 0.7696, 0.767, 0.7686, 0.762]

Table ESM 23. Classification report of RFC based on IUPAC encoded data

	precision	recall	f1-score	support
Active (target 1)	0.67	0.77	0.71	2000
Inactive (target 0)	0.72	0.62	0.67	2000
accuracy			0.69	4000
macro avg	0.70	0.69	0.69	4000
weighted avg	0.70	0.69	0.69	4000

Figures

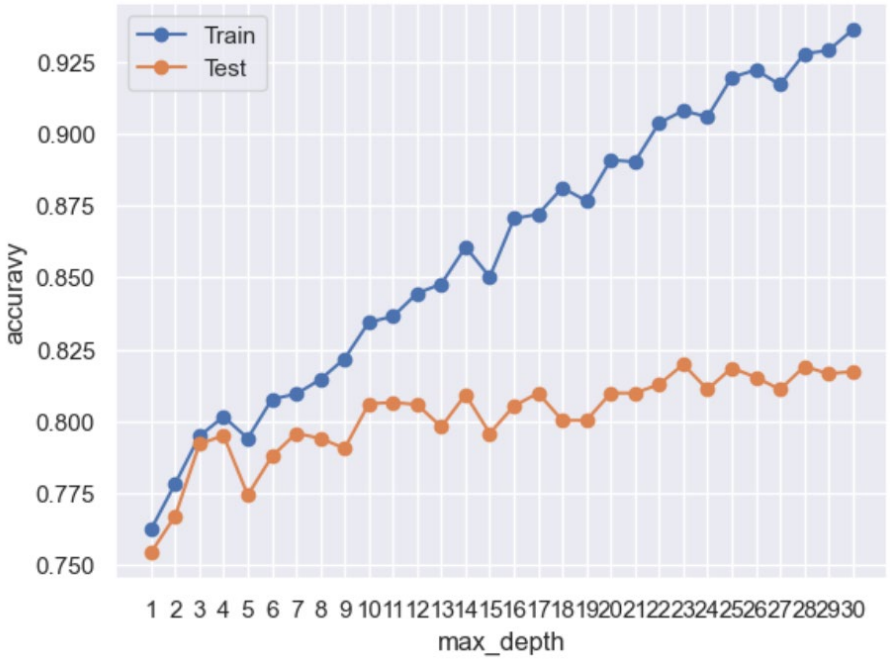

Figure ESM 1. Scrutinising for overfitting of RFC
based on Dataset 2 integrated with molecular feature

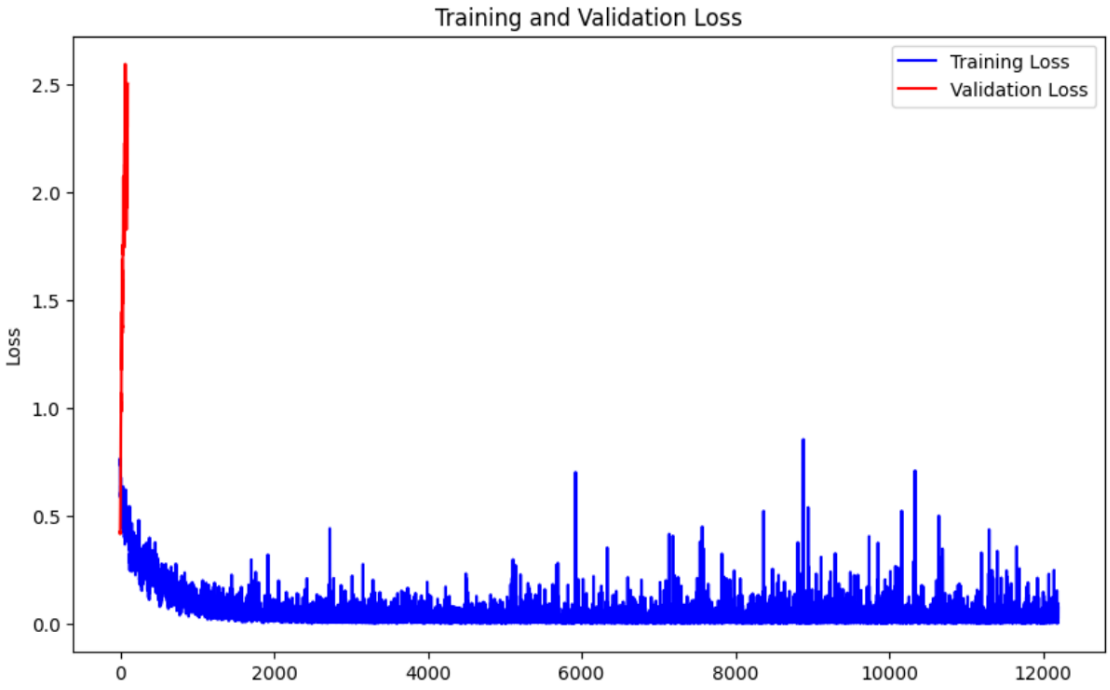

Figure ESM 2. Overfitting check of DNN optimised by Optuna.
Final Training Loss: 0.07168064266443253;
Final Validation Loss: 2.5014856861483667;
Loss Difference (Validation - Training): 2.4298050434839342;
Potential Overfitting Detected.

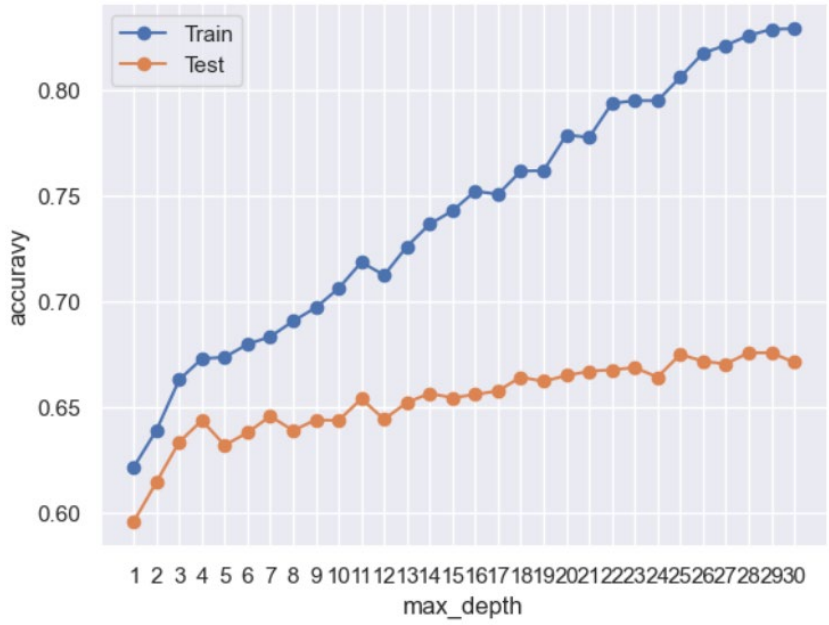

Figure ESM3. Scrutinise for overfitting of RFC based on IUPAC encoded data

Figure ESM4. Scrutinise for overfitting of RFC
based on IUPAC encoded data hyperparameter tuned by Optuna

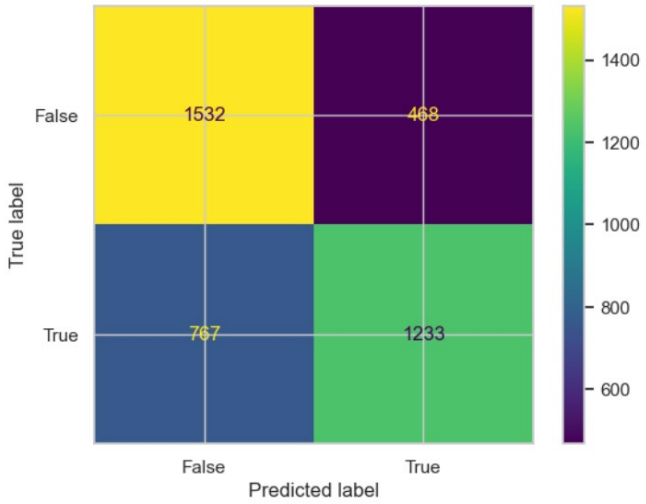

Figure ESM 5. Confusion matrix of ML model based on IUPAC encoded data

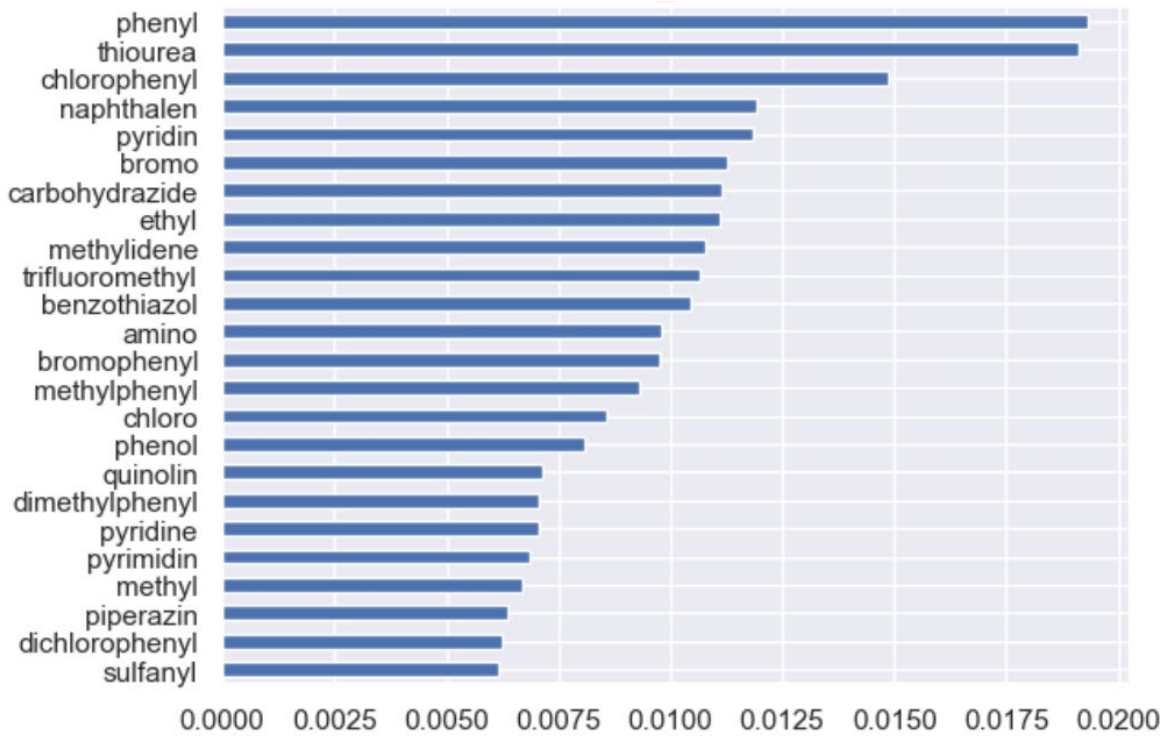

Figure ESM6. Descending order of the functional groups of RFC based on Dataset 2& molecular feature obtained by feature importance algorithm

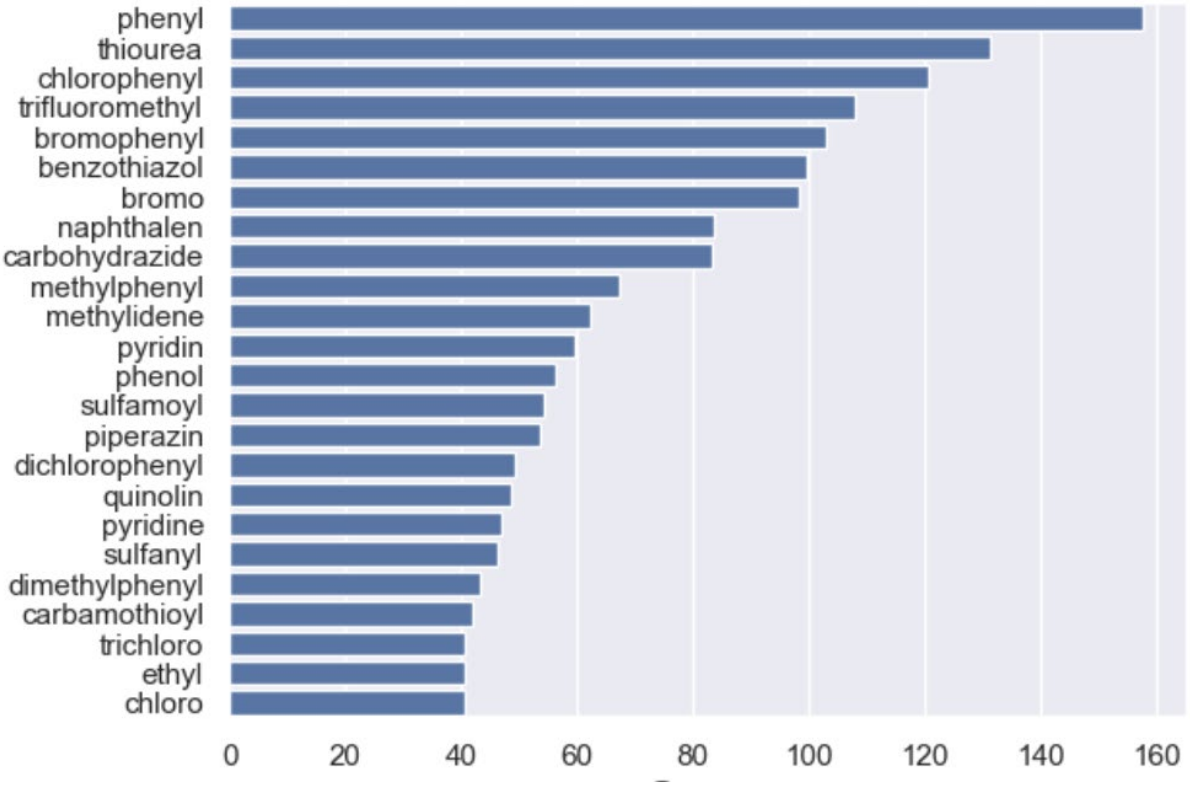

Figure ESM7. Descending order of the functional groups of RFC based on Dataset 2 & molecular features obtained by the permutation importance algorithm