

## **Interpretable machine learning framework for predicting the reactivity of trifluoromethylating reagents**

Vaneet Saini<sup>\*1</sup>, Shivansh Kanwar

*Department of Chemistry & Centre for Advanced Studies in Chemistry, Panjab University, Chandigarh 160014,*

*India.* Email: [vsaini@pu.ac.in](mailto:vsaini@pu.ac.in)

Orcid ID: [0000-0002-8186-5166](https://orcid.org/0000-0002-8186-5166)

Table S1. Number of molecules in the train and test set from each type.

Set	Type of reagent	Train molecules	Test molecules	Total
Type of reagents	Chalcogenium salts	48	13	61
	Sulfoximines	9	3	12
	HIR	20	6	26
				<b>99</b>

Table S2. Molecules with highest and lowest TC+DA values.

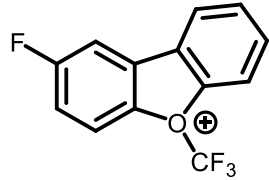
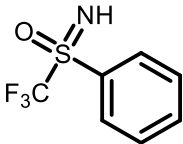
S. No	Type	Structure	TC+DA
1.	Chalcogenium salt		26.7
2.	Sulfoximines		89.4

Table S3. . Hyperparameters for NN and ET models, along with the performance metrics obtained from training with 5 descriptors.

S. No.	Model	Hyperparameters	Cross-val R <sup>2</sup>	Train R <sup>2</sup>	Test R <sup>2</sup>	Test RMSE
1.	NN	{ Input_layer = 64, dropout_layer = 0.1, hidden_layer1 = 128, dropout_layer1 = 0.1, hidden_layer2 = 128, dropout_layer2 = 0.1, optimizer = "Adam", loss_function = "mse", epochs = 200, batch_size = 32, lr = 0.001, activation function = "ReLU" }	0.847	0.962	0.943	3.84
2.	NN*	{ Input_layer = 96, dropout_layer = 0.2, hidden_layer1 = 256, dropout_layer1 = 0.5, hidden_layer2 = 128, dropout_layer2 = 0.1, optimizer = "Adam", loss_function = "mse", epochs = 100, batch_size = 32, lr = 0.01, activation function = "ReLU" }	0.900	0.945	0.956	3.43
3.	ET	{bootstrap=False, max_depth=None, max_features=1.0,	0.867	1.0	0.943	3.75

		min_samples_leaf=1, min_samples_split=2, n_estimators=100, random_state = 42}				
--	--	--	--	--	--	--

\*represents tuned parameters.

Table S4 Actual and predicted TC<sup>+</sup>DA values for the NN model trained on the 5 descriptors.

ID	TOR	Smiles	Set	TC+DA values	
				Actual	Predicted
Ea1b	Chalc · Salts	<chem>FC([S+])C2=CC=C(OC)C=C2C3=C1C=CC=C3)(F)F</chem>	Train	45.5	44.1
Ea1c	Chalc · Salts	<chem>FC([S+])C2=CC=C(C)C=C2C3=C1C=CC=C3)(F)F</chem>	Train	44.1	45.2
Ea1d	Chalc · Salts	<chem>FC1=CC=C([S+])(C(F)(F)F)C2=C3C=CC=C2)C3=C1</chem>	Train	42.1	43.3
Ea1e	Chalc · Salts	<chem>C1C1=CC=C([S+])(C(F)(F)F)C2=C3C=CC=C2)C3=C1</chem>	Train	41	45.6
Ea1g	Chalc · Salts	<chem>FC([S+])C2=CC=C(C(F)(F)F)C=C2C3=C1C=CC=C3)(F)F</chem>	Train	39.2	44.2
Ea1h	Chalc · Salts	<chem>FC([S+])C2=CC=C(C#N)C=C2C3=C1C=CC=C3)(F)F</chem>	Train	37.9	44.1
Ea1i	Chalc · Salts	<chem>FC([S+])C2=CC=C([N+])([O-])=O)C=C2C3=C1C=CC=C3)(F)F</chem>	Train	36.7	37.8
Ea1j	Chalc · Salts	<chem>[H]C1=C(C)C=C2C([S+])(C(F)(F)F)C3=C2C=C(C)C([H])=C3[H])=C1[H]</chem>	Train	45.2	47.4
Ea1k	Chalc · Salts	<chem>[H]C1=C(C)C=C2C([S+])(C(F)(F)F)C3=C2C=C([H])C([H])=C3C)=C1[H]</chem>	Train	44.5	45.4
Ea1l	Chalc · Salts	<chem>[H]C1=C(C)C=C2C([S+])(C(F)(F)F)C3=C2C=C([H])C([H])=C3[H])=C1C</chem>	Train	44.9	45.4
Ea1o	Chalc · Salts	<chem>[H]C1=C([H])C([H])=C([S+])(C(F)(F)F)C2=C3C=C([H])C([N+])([O-])=O)=C2[H])C3=C1</chem>	Train	37.5	37.8
Ea1p	Chalc · Salts	<chem>[H]C1=C([N+])([O-])=O)C([H])=C([S+])(C(F)(F)F)C2=C3C=C([H])C([N+])([O-])=O)=C2[H])C3=C1</chem>	Train	32.3	30.8
Ea1r	Chalc · Salts	<chem>[H]C1=C(C)C=C2C([S+])(C(F)(F)F)C3=C2C=C([H])C([N+])([O-])=O)=C3[H])=C1C</chem>	Train	39.3	38.1
Ea2a	Chalc · Salts	<chem>[H]C1=CC=C([O+])(C(F)(F)F)C2=C3C=CC=C2)C3=C1</chem>	Train	27.8	27.8
Ea2b	Chalc · Salts	<chem>FC([O+])C2=CC=C(C(C)(C)C)C=C2C3=C1C=CC=C3)(F)F</chem>	Train	29.1	27.7
Ea2c	Chalc · Salts	<chem>FC1=CC=C([O+])(C(F)(F)F)C2=C3C=CC=C2)C3=C1</chem>	Train	26.7	28.8
Ea3a	Chalc · Salts	<chem>[H]C1=CC=C2C([Se+])(C(F)(F)F)C3=C2C=CC=C3)=C1</chem>	Train	51.6	43.8
Ea3b	Chalc · Salts	<chem>FC([Se+])C2=CC([N+])([O-])=O)=CC=C2C3=C1C=CC=C3)(F)F</chem>	Train	43.3	44.2
Ea4a	Chalc · Salts	<chem>[H]C1=CC=C2C([Te+])(C(F)(F)F)C3=C2C=CC([H])=C3)=C1</chem>	Train	59.7	53.3
Ea4b	Chalc · Salts	<chem>FC([Te+])C2=CC([N+])([O-])=O)=CC=C2C3=C1C=C([N+])([O-])=O)C=C3)(F)F</chem>	Train	52.2	53.5
Ea5a	Chalc · Salts	<chem>[H]C(C=C1)=CC=C1[S+](C(F)(F)F)C2=CC=CC=C2</chem>	Train	50.4	48.6

Ea5b	Chalc Salts	FC([S+](C1=CC=CC=C1)C2=CC=C(OC)C=C2)(F)F	Train	52.2	49.3
Ea5c	Chalc Salts	FC([S+](C1=CC=CC=C1)C2=CC=C(C)C=C2)(F)F	Train	51.1	51.7
Ea5d	Chalc Salts	FC(C=C1)=CC=C1[S+](C(F)(F)F)C2=CC=CC=C2	Train	49.5	47.4
Ea5e	Chalc Salts	ClC(C=C1)=CC=C1[S+](C(F)(F)F)C2=CC=CC=C2	Train	48.5	45.5
Ea5f	Chalc Salts	BrC(C=C1)=CC=C1[S+](C(F)(F)F)C2=CC=CC=C2	Train	48.3	44.1
Ea5h	Chalc Salts	FC([S+](C1=CC=CC=C1)C2=CC=C(C#N)C=C2)(F)F	Train	45.4	46.5
Ea5k	Chalc Salts	FC([S+](C1=CC=C(C)C=C1)C2=CC=C(C)C=C2)(F)F	Train	52.7	53.1
Ea5l	Chalc Salts	FC(C=C1)=CC=C1[S+](C(F)(F)F)C2=CC=C(F)C=C2	Train	49.3	46.9
Ea5m	Chalc Salts	FC1=CC(F)=CC=C1[S+](C(F)(F)F)C2=CC=CC=C2	Train	45.7	46.4
Ea5n	Chalc Salts	ClC(C=C1)=CC=C1[S+](C(F)(F)F)C2=CC=C(Cl)C=C2	Train	46.8	44.0
Ea5o	Chalc Salts	FC([S+](C1=CC=C(OC2=CC=CC=C2)C=C1)C3=CC=C(OC4=CC=CC=C4)C=C3)(F)F	Train	52.9	49.0
Ea5q	Chalc Salts	FC([S+](C1=CC=CC=C1)C2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)(F)F	Train	43.6	45.4
Ea5r	Chalc Salts	FC([S+](C1=CC=C([N+])([O-])=O)C=C1)C2=CC=C([N+])([O-])=O)C=C2)(F)F	Train	39.3	30.6
Ea5s	Chalc Salts	FC(C=C1)=CC=C1[S+](C(F)(F)F)C2=CC=CC([N+])([O-])=O)C=C2	Train	44.4	41.2
Ea5t	Chalc Salts	ClC(C=C1)=CC=C1[S+](C(F)(F)F)C2=CC=C(OC)C=C2	Train	50.5	46.9
Ea5u	Chalc Salts	ClC(C=C1)=CC=C1[S+](C(F)(F)F)C2=CC=C(C)C=C2C	Train	49.1	49.6
Ea6b	Chalc Salts	C[C@]12C(C3=CC4=CC=CC=C4[S+]3C(F)(F)F)=C[C@@H](C2(C)C)CC1	Train	41.9	43.2
Ea6c	Chalc Salts	FC([S+]1C2=CC=CC=C2C=C1C3=CC=CC=C3)(F)F	Train	39.5	45.3
Ea6e	Chalc Salts	FC([S+]1C2=CC=CC=C2C=C1C3=CC=C(C)C=C3)(F)F	Train	40	47.8
Ea6f	Chalc Salts	ClC1=CC=CC=C1C2=CC3=CC=CC=C3[S+]2C(F)(F)F	Train	39.4	43.5
Ea6g	Chalc Salts	BrC1=CC(C2=CC3=CC=CC=C3[S+]2C(F)(F)F)=CC=C1	Train	38.7	42.1
Ea6h	Chalc Salts	BrC(C=C1)=CC=C1C2=CC3=CC=CC=C3[S+]2C(F)(F)F	Train	38.8	42.1
Ea6i	Chalc Salts	IC(C=C1)=CC=C1C2=CC3=CC=CC=C3[S+]2C(F)(F)F	Train	39.1	40.5

Ea7a	Chalc Salts	FC([S+]1C2=CC=CC=C2OC3=C1C=CC=C3)(F)F	Train	49.8	44.8
Ea8a	Chalc Salts	FC([S+]1C2=CC=CC=C2C3=C1C=C(S(=O)([O-])=O)C=C3)(F)F	Train	43.2	44.0
Ea8b	Chalc Salts	FC([S+]1C2=CC=C(C)C=C2C3=C1C=C(S(=O)([O-])=O)C(C)=C3)(F)F	Train	45.9	45.4
Ea9a	Chalc Salts	O=C([C-](C(OC)=O)[S+](C(F)(F)F)C1=CC=CC=C1)OC	Train	72.4	74.6
Eb1b	Sulf. Salts	O=S(C1=CC=CC=C1)(C(F)(F)F)=NC(C2=CC=CC=C2)=O	Train	73.1	73.7
Eb1c	Sulf. Salts	O=S(C1=CC=CC=C1)(C(F)(F)F)=NC(C(F)(F)F)=O	Train	65	67.2
Eb1d	Sulf. Salts	O=S(C1=CC=CC=C1)(C(F)(F)F)=NS(C2=CC=CC=C2)(=O)=O	Train	67.9	68.3
Eb1e	Sulf. Salts	O=S(C1=CC=CC=C1)(C(F)(F)F)=NS(=O)(C(F)(F)F)=O	Train	58.8	63.2
Eb2a	Sulf. Salts	[H]C1=C([H])C=C(S(=NC2=O)(C(F)(F)F)=O)C2=C1	Train	65.7	65.5
Eb2b	Sulf. Salts	[H]C1=C([N+](O)=O)C=C2C(S(=NC2=O)(C(F)(F)F)=O)C=C1	Train	60.3	61.2
Eb2c	Sulf. Salts	[H]C1=C([N+](O)=O)C=C(S(=NC2=O)(C(F)(F)F)=O)C2=C1	Train	60.4	61.2
Eb3a	Sulf. Salts	[H]C1=C([H])C=C(S(=NS2(=O)=O)(C(F)(F)F)=O)C2=C1	Train	60.7	60.7
Eb4a	Sulf. Salts	O=S(C1=CC=CC=C1)(C(F)(F)F)=[N+](C)C	Train	45.3	42.1
Ec1b	HIR	O=C1C2=C(I(C(F)(F)F)O1)C=CC(OC)=C2	Train	54.6	61.2
Ec1c	HIR	O=C1C2=C(I(C(F)(F)F)O1)C=CC(C)=C2	Train	54.3	59.2
Ec1d	HIR	FC1=CC2=C(I(C(F)(F)F)OC2=O)C=C1	Train	52.5	52.4
Ec1e	HIR	C1C1=CC2=C(I(C(F)(F)F)OC2=O)C=C1	Train	52	52.3
Ec1f	HIR	BrC1=CC2=C(I(C(F)(F)F)OC2=O)C=C1	Train	52	52.3
Ec1g	HIR	O=C1C2=C(I(C(F)(F)F)O1)C=CC(C(F)(F)F)=C2	Train	51.6	52.9
Ec1h	HIR	O=C1C2=C(I(C(F)(F)F)O1)C=CC(C#N)=C2	Train	50.8	52.4
Ec2b	HIR	COC1=CC2=C(I(C(F)(F)F)OC2(C)C)C=C1	Train	79.9	84.0
Ec2d	HIR	FC1=CC2=C(I(C(F)(F)F)OC2(C)C)C=C1	Train	78.3	80.2
Ec2e	HIR	C1C1=CC2=C(I(C(F)(F)F)OC2(C)C)C=C1	Train	78.1	74.9
Ec2f	HIR	BrC1=CC2=C(I(C(F)(F)F)OC2(C)C)C=C1	Train	77.9	78.3
Ec2g	HIR	FC(C1=CC2=C(I(C(F)(F)F)OC2(C)C)C=C1)(F)F	Train	77.5	80.7
Ec2i	HIR	O=[N+](C1=CC2=C(I(C(F)(F)F)OC2(C)C)C=C1)[O-]	Train	76.4	80.4
Ec3a	HIR	[H]C1=CC2=C(I(C(F)(F)F)OC2(C(F)(F)F)C(F)(F)F)C=C1	Train	63.6	64.1
Ec3b	HIR	CC1=CC2=C(I(C(F)(F)F)OC2(C(F)(F)F)C(F)(F)F)C=C1	Train	64.4	69.4
Ec3c	HIR	FC(I1C2=C([C@ @])(C3=CC=CC=C3)(C)O1)C=CC=C2)(F)F	Train	74.7	78.0
Ec3d	HIR	FC(I1C2=C([C@ @])(CC)(C)O1)C=CC=C2)(F)F	Train	80.7	83.5
Ec3e	HIR	FC(I1C2=C([C@ @])(C(C)C)(C)O1)C=CC=C2)(F)F	Train	81	86.0
Ec3f	HIR	FC(I1C2=C([C@ @])(C3CCCCC3)(C)O1)C=CC=C2)(F)F	Train	81.4	87.5
Ec4a	HIR	[H]C1=C2C(C34CCCC2C3)=C(I(OC4(C)C)C(F)(F)F)C=C1	Train	81	84.6
Ea1a	Chalc Salts	[H]C1=CC=C([S+](C(F)(F)F)C2=C3C=CC=C2)C3=C1	Val	42.7	43.6
Eb1a	Sulf. Salts	O=S(C1=CC=CC=C1)(C(F)(F)F)=N	Val	89.4	83.5

Ea5p	Chalc Salts	<chem>FC([S+](C1=CC=C(OC(F)(F)F)C=C1)C2=CC=C(OC(F)(F)F)C=C2)(F)F</chem>	Val	46.5	46.7
Ea1f	Chalc Salts	<chem>BrC1=CC=C([S+](C(F)(F)F)C2=C3C=CC=C2)C3=C1</chem>	Val	40.8	41.5
Eb3c	Sulf. Salts	<chem>[H]C1=C([N+](O)=O)C=C(S(=NS2(=O)=O)(C(F)(F)F)=O)C2=C1</chem>	Val	55.5	57.1
Ea5g	Chalc Salts	<chem>FC([S+](C1=CC=CC=C1)C2=CC=C(C(F)(F)F)C=C2)(F)F</chem>	Val	46.7	47.7
Ec2h	HIR	<chem>N#CC1=CC2=C(I(C(F)(F)F)OC2(C)C)C=C1</chem>	Val	76.8	77.5
Ea6j	Chalc Salts	<chem>FC(C=C1F)=CC=C1C2=CC3=CC=CC=C3[S+]2C(F)(F)F</chem>	Val	38.7	44.9
Ea1n	Chalc Salts	<chem>[H]C1=C(C)C=C2C([S+](C(F)(F)F)C3=C2C=C(C)C([H])=C3C)=C1C</chem>	Val	46.2	48.3
Ea1q	Chalc Salts	<chem>[H]C1=C(C)C=C2C([S+](C(F)(F)F)C3=C2C=C([H])C(S(F)(F)(F)F)=C3[H])=C1</chem>	Val	39.6	46.1
Ec1i	HIR	<chem>O=C1C2=C(I(C(F)(F)F)O1)C=CC([N+](O)=O)=C2</chem>	Val	50.1	52.5
Ea1m	Chalc Salts	<chem>[H]C1=C(C)C([H])=C([S+](C(F)(F)F)C2=C3C=C([H])C(C)=C2[H])C3=C1</chem>	Test	43.9	47.4
Ea5i	Chalc Salts	<chem>FC([S+](C1=CC=CC=C1)C2=CC=C([N+](O)=O)C=C2)(F)F</chem>	Test	44.1	41.1
Ea5j	Chalc Salts	<chem>FC([S+](C1=CC=CC=C1)C2=CC=CC([N+](O)=O)=C2)(F)F</chem>	Test	45.3	41.1
Ea6a	Chalc Salts	<chem>FC([S+]1C2=CC=CC=C2C=C1C3CC3)(F)F</chem>	Test	41.9	45.7
Ea6d	Chalc Salts	<chem>FC([S+]1C2=CC=CC=C2C=C1C3=CC=C(OC)C=C3)(F)F</chem>	Test	40.5	45.7
Ea8c	Chalc Salts	<chem>FC([S+]1C2=CC([N+](O)=O)=CC=C2C3=C1C=C(S(=O)(O)=O)C=C3)(F)F</chem>	Test	37.9	41.2
Eb3b	Sulf. Salts	<chem>[H]C1=C([N+](O)=O)C=C2C(S(=NS2(=O)=O)(C(F)(F)F)=O)=C1</chem>	Test	55.1	57.1
Ec1a	HIR	<chem>[H]C1=CC2=C(I(C(F)(F)F)OC2=O)C=C1</chem>	Test	53.9	52.6
Ec2a	HIR	<chem>[H]C1=CC2=C(I(C(F)(F)F)OC2(C)C)C=C1</chem>	Test	79.7	81.5
Ec2c	HIR	<chem>CC1=CC2=C(I(C(F)(F)F)OC2(C)C)C=C1</chem>	Test	79.8	84.4
Ec3g	HIR	<chem>[H]C1=CC2=C(I(C(F)(F)F)OC23CCCCC3)C=C1</chem>	Test	81.8	84.6



## Feature Selection and Dimensionality Reduction

In machine learning applied to chemical data, feature reduction is an essential step for improving model performance, generalization, and interpretability. High-dimensional descriptor sets, such as those computed by Mordred, often contain redundant, irrelevant, or noisy features, which can lead to overfitting and decreased model transparency. To address this, we employed a combination of unsupervised statistical filtering and supervised feature importance analysis.

### 1. Variance Threshold Filtering

We first removed descriptors with zero or near-zero variance across the dataset. Features that do not vary across molecules do not carry meaningful discriminative information and can add unnecessary complexity to the model. Eliminating such descriptors reduces dimensionality and computational burden without affecting the predictive signal.

### 2. Correlation-Based Filtering

To minimize multicollinearity, we applied pairwise Pearson correlation analysis to the remaining descriptors. For descriptor pairs with a correlation coefficient  $|r| > 0.8$ , one descriptor was removed based on redundancy across chemical categories. This process ensures that each retained feature contributes uniquely to the model's decision-making and prevents overlapping signals from diluting feature importance metrics.

### 3. Extra Trees-Based Feature Importance

Beyond unsupervised filtering, we applied a supervised feature selection strategy using the Extra Trees (ET) algorithm. The ET model, an ensemble of randomized decision trees, provides a built-in feature importance score based on each descriptor's contribution to reducing prediction error (impurity) across decision paths. After training on the full descriptor set, the ET model identified the most influential features for predicting  $TC^{+}DA$  values. This method considers not only the statistical variation of features but also their relevance to the target property, capturing nonlinear relationships that may be missed by correlation-based filters alone.

### 4. Layered approach

In order to identify the optimal number of descriptors to use in training the model, a layered approach was employed. In this approach the features were added sequentially to the dataset, and

performance metrics were tracked for the models. With these five descriptors, the NN model reached peak performance, achieving cross-validation, training, and test  $R^2$  scores of 0.847, 0.962, and 0.943, respectively. However, the subsequent addition of other descriptor resulted in a dip in the cross-validation  $R^2$  score, with only slight increments observed in the training and test scores. Further incorporation of additional descriptors, whether 7, 8, or even 20 in total, led to pronounced overfitting, evidenced by training  $R^2$  scores nearing 1 while significant gaps emerged between training and test performances. Therefore, five descriptors were identified as the optimal feature set, enabling the NN model to achieve high predictive accuracy while keeping the model architecture simple and free from unnecessary complexity.

From this analysis, a subset of five interpretable and chemically meaningful descriptors was selected. These included features related to electrostatic potential, bond energies, atomic environments, and molecular shape, all of which correlate with known mechanistic influences on trifluoromethylation reactivity. By combining statistical and model-driven selection methods, we ensured that the final descriptor set retained the most relevant features for both predictive performance and chemical insight.

### **Perspective on Feature Reduction**

Feature selection not only enhances model efficiency and accuracy but also serves as a tool for uncovering mechanistic patterns in chemical data. However, it must be applied with caution. Overzealous reduction can discard descriptors that, while correlated, may provide complementary mechanistic information. Similarly, reliance on a single feature selection criterion can introduce bias if the underlying relationships are nonlinear or context-specific.

In our approach, combining filtering techniques with supervised importance scoring allowed us to mitigate these risks. The retained features were interpretable and aligned with known reactivity trends, thus reinforcing confidence in the model and offering practical guidance for future reagent design.

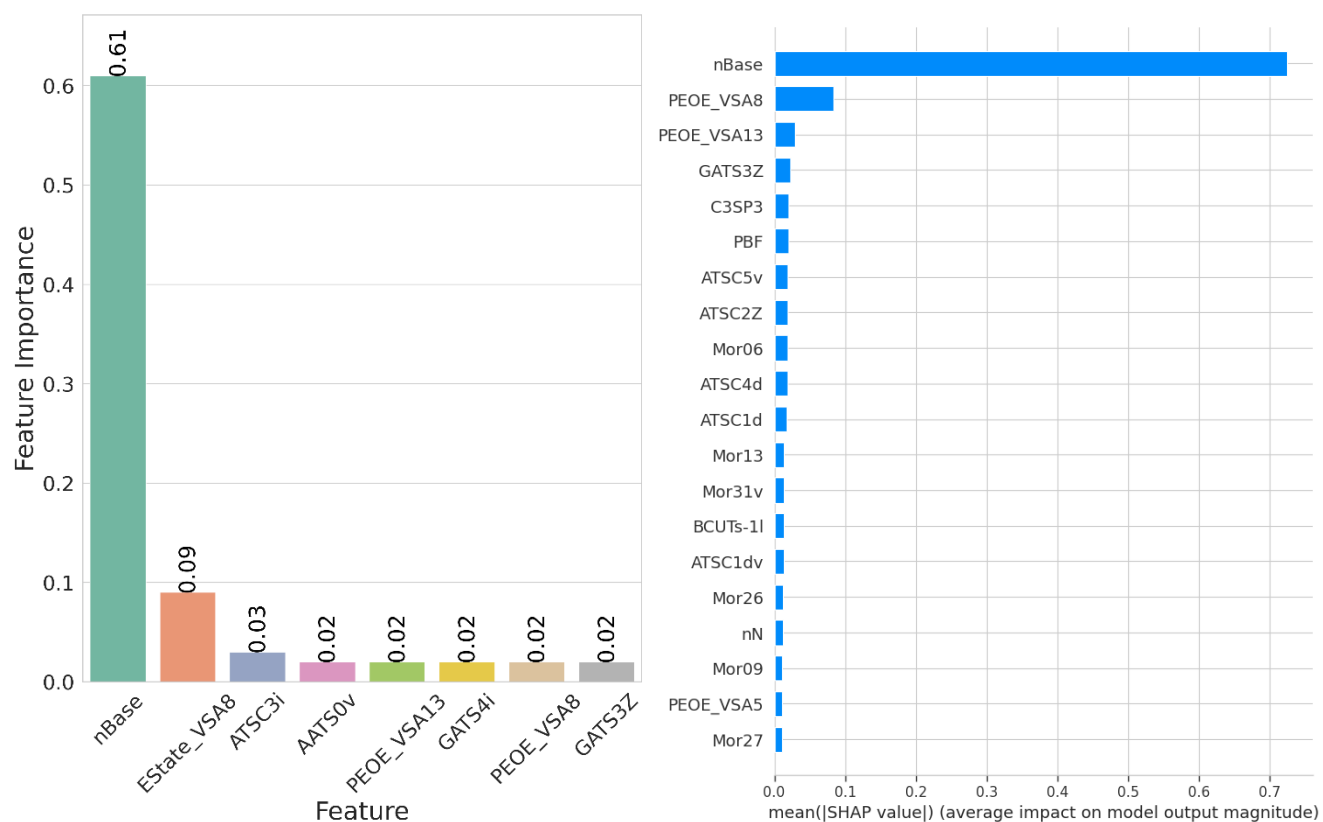


Figure S1 Comparison of different feature importance algorithm. A) Extra tree based feature importance chart. B) Feature importance based on SHAP values.