Supporting Information

Hybrid Supervised and Unsupervised Machine Learning Approach for Identifying Nucleoside Drugs Using Nanopore Readouts

Sneha Mittal, † Milan Kumar Jena, † Biswarup Pathak*, †

†Department of Chemistry, Indian Institute of Technology (IIT) Indore, Indore, Madhya Pradesh, 453552, India
*E-mail: biswarup@iiti.ac.in

1. Rotational Effects Driven Structural Optimization

Considering the dynamic nature of nucleoside drug molecules during translocation through the nanopore, we have considered possible rotations from 0° to 90° (in the steps of 30°) around the x-axis in the yz-plane for each considered nucleoside drug molecule, as shown in **Figure S1.** In our investigation, each drug molecule is first placed in the plane with the nanopore device, considered as 0° configuration. We then rotated the drug molecule with respect to this orientation around the x-axis in the step of 30° and achieved configurations 30°, 60°, and 90°. To this end, we have four orientations for each nucleoside drug molecule, resulting in a total of 32 optimizations corresponding to 8 nucleoside drugs.

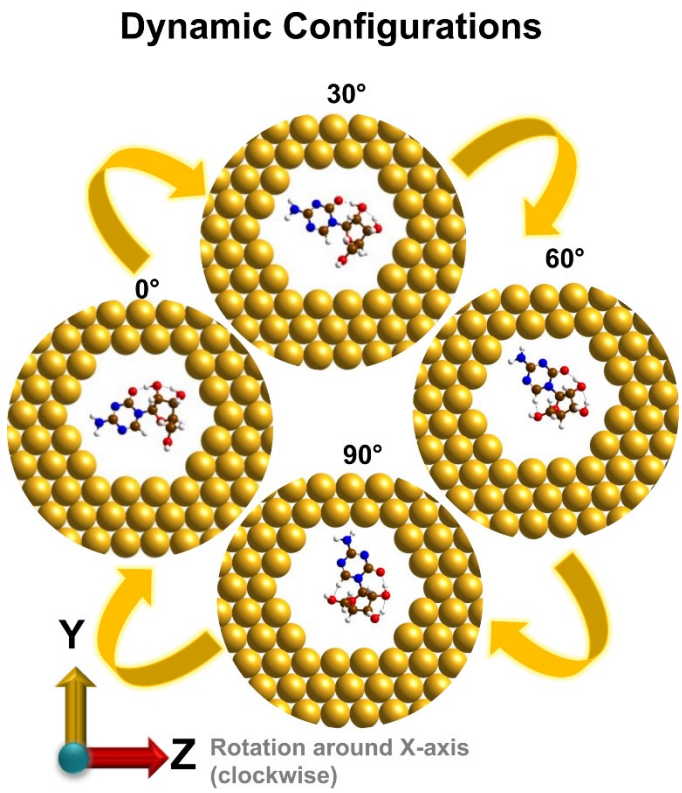

Figure S1: Representative orientations of 5-Azacytidine nucleoside drug inside the nanopore illustrated corresponding to in-plane rotations from 0° to 90° in the steps of 30° around the x-axis in the yz-plane. Atom color code: Au (yellow), C (brown), H (white), N (blue), F (purple), and O (red).

After optimizing the considered drug molecules over the dynamic configuration space, we evaluated the relative energy values for each configuration, as shown in **Table S1**. The energetically most favorable configurations are chosen as the most likely geometry of nucleoside drug molecules during translocation through the nanopore.

Table S1. Relative energies (in eV) of the nanopore-drug systems when nucleoside drug molecules are interacting with the nanopore edges in different orientations (0°, 30°, 60°, 90°), as shown in **Figure S1**.

S.No.	Nucleoside Drugs	0°	30°	60°	90°
1.	5-Azacytidine	0.21	0.21	**0.00**	0.04
2.	5-Fluorouridine	0.17	0.13	0.09	**0.00**
3.	6-Azauridine	**0.00**	0.07	0.14	0.02
4.	GS-441524	0.54	0.4	**0.00**	0.28
5.	Loxoribine	0.52	0.01	0.20	**0.00**
6.	Mizoribine	0.08	0.01	**0.00**	0.02
7.	Ribavirin	0.24	0.30	**0.00**	0.12
8.	Zebularine	1.17	0.01	**0.00**	0.31

2. Nanopore Transmission Readouts for the Most Likely Configuration

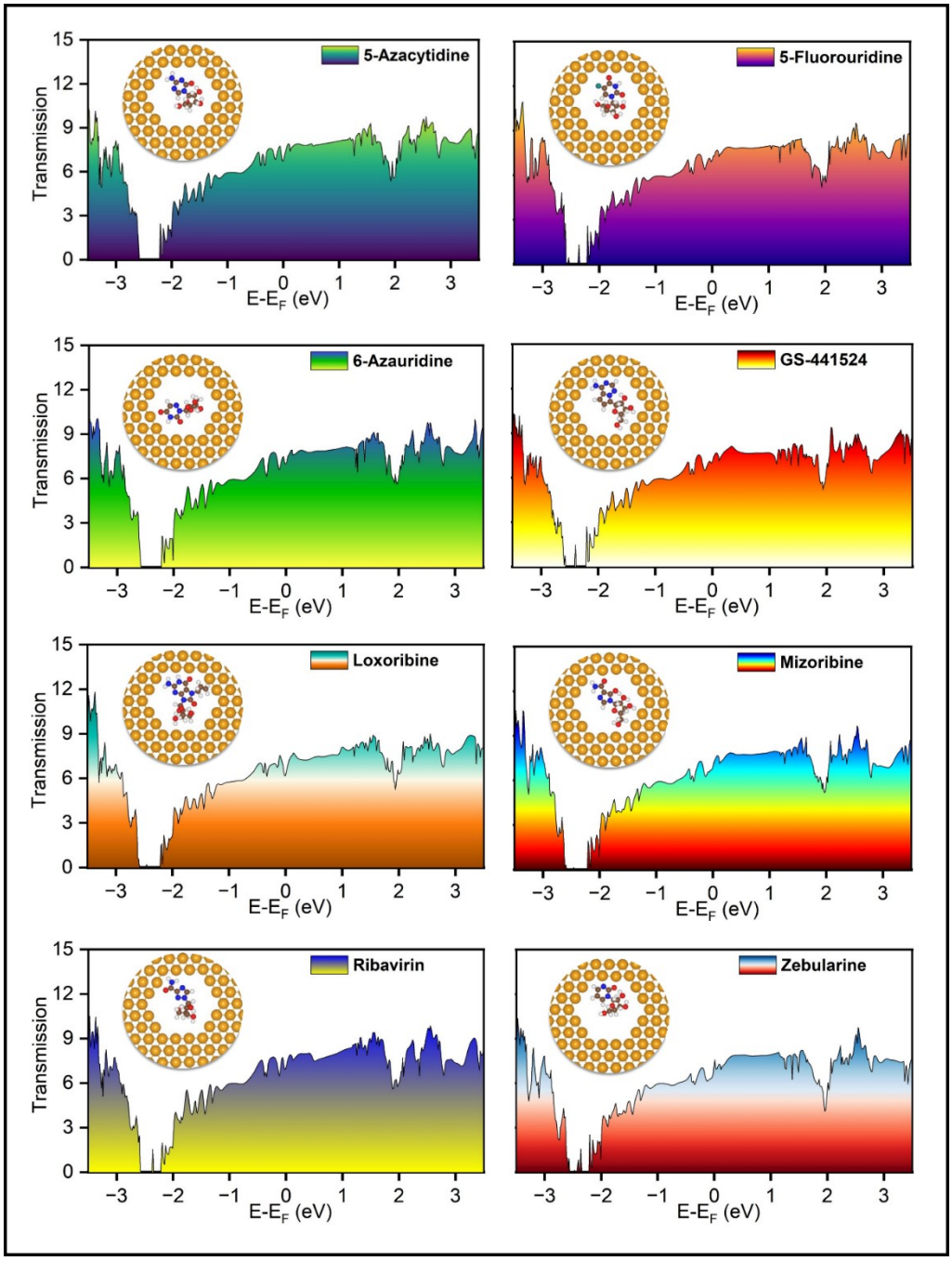

Figure S2. Nanopore transmission readouts for the most likely configuration of nanopore-nucleoside drug systems in the energy window of -3.5 to $+3.5$ eV. This selected energy window allows for the detection of subtle changes in transmission values across a wider energy range, encompassing both valance and conduction bands. The most likely configuration of nucleoside drugs inside the nanopore is shown in the respective inset. Atom color code: Au (yellow), C (brown), H (white), N (blue), F (purple), and O (red).

3. Impact of Orientational Fluctuations

To study the impact of orientational fluctuations on transmission fingerprints, we have performed electronic transport calculations over a large dynamic configuration space including both in-plane and out-of-plane orientational fluctuations. In our analysis, to mimic the in-plane rotation dynamics of drug molecules, we have considered a total of four orientations (0°, 30°, 60°, 90°) as shown in **Figure S3a** and to mimic the out-of-plane translation dynamics, we considered two configurations (+1.0 Å and -1.0 Å) as shown in **Figure S3b**. To address the effect of out-of-plane translation dynamics, we have translated the drug molecule in both upward (+1.0 Å) and downward (-1.0 Å) directions from the initial position (0.0 Å) along the x-axis in the yz-plane.

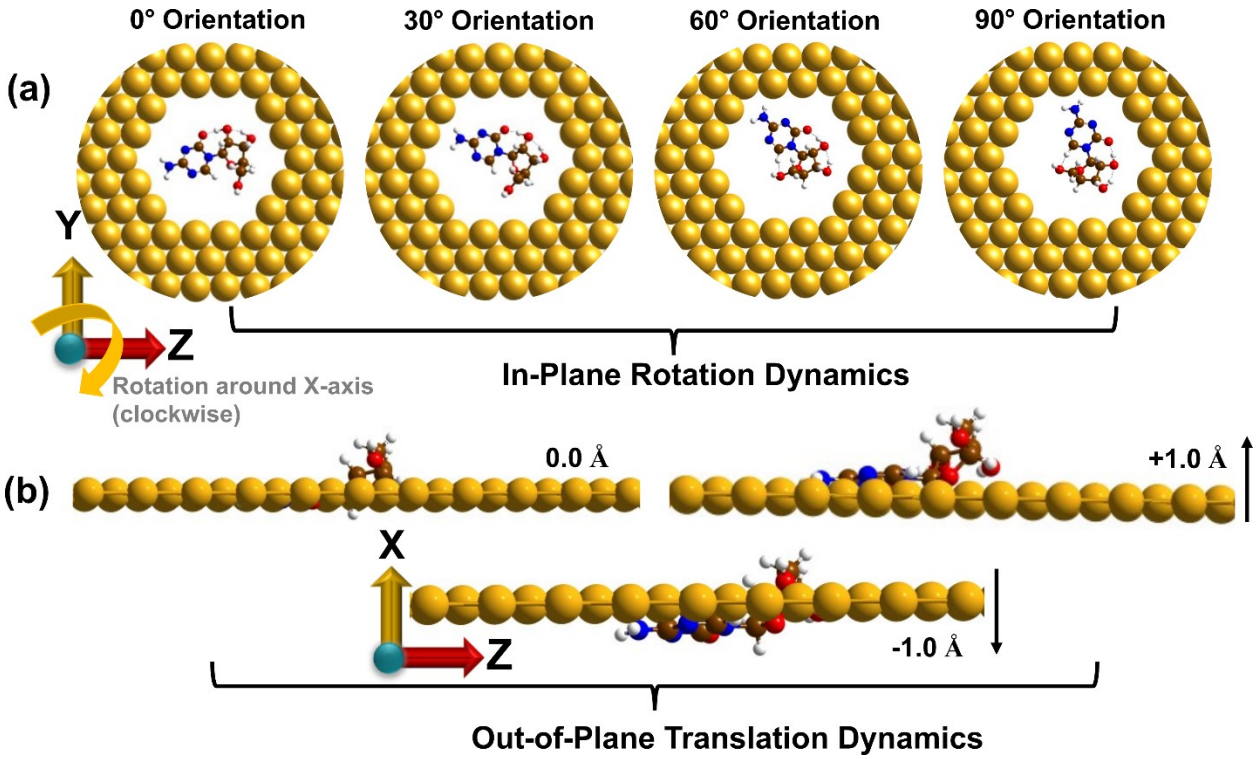

Figure S3: (a) Representative orientations of 5-Azacytidine nucleoside drug inside the nanopore illustrated corresponding to rotation from 0° to 90° (in the step of 30°) around the x-axis in the yz-plane and **(b)** representation of 5-Azacytidine nucleoside drug translated out-of-plane along the x-axis in the yz-plane, in both positive and negative directions by ± 1.0 Å. Atom color code: Au (yellow), C (brown), H (white), N (blue), F (purple), and O (red).

4. Rotation and Translation Dynamics with Molecular Orbital Wavefunctions

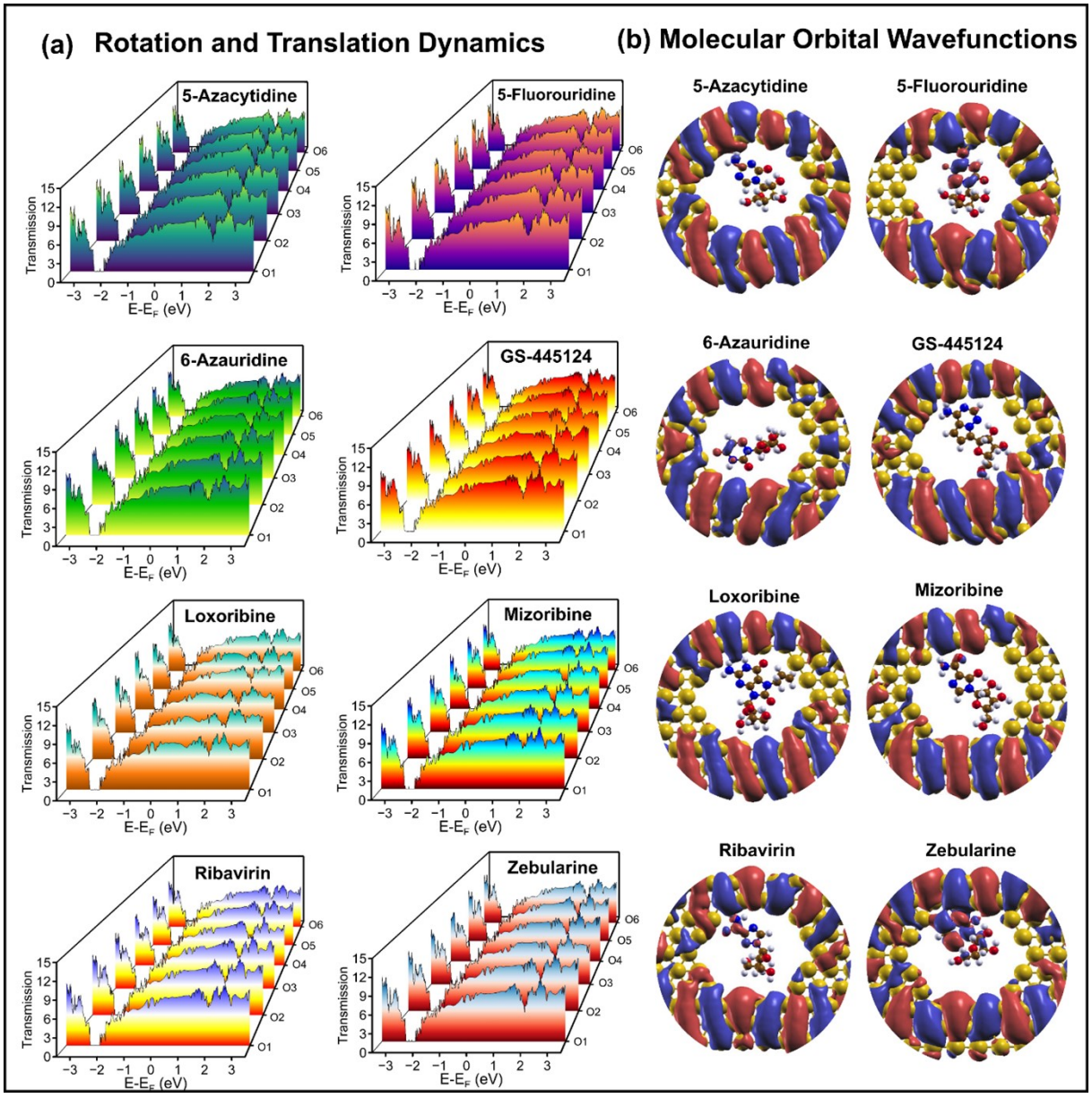

Figure S4. (a) Nanopore transmission readouts of nanopore-nucleoside drug systems in the energy window of -3.5 to $+3.5$ eV with different rotation and translation dynamics. The configurations undergoing rotational fluctuations are represented as O1, O2, O3, and O4 corresponding to rotations 0°, 30°, 60°, and 90°, respectively, and those with translational fluctuations are represented as O5 and O6 corresponding to translations +1.0 Å and -1.0 Å, respectively and **(b)** isosurface plots (isosurface value is 0.005 e/Å^3) of the molecular orbitals (MOs) responsible for the sharp transmission peaks of nanopore-nucleoside drug systems. The negative and positive lobes are shown in blue and red colors, respectively. Atom color code: Au (yellow), C (brown), H (white), N (blue), F (purple), and O (red).

5. Features Selection for Supervised Machine Learning Framework

In our pursuit of building a generalized ML framework, our emphasis lies in the careful selection of features derived from the transmission-energy profiles of nucleoside drugs. Recognizing the profound correlation between transmission and energy, a total of 13 features have been selected (refer to **Table S2**). These chosen features exhibit a substantial correlation with the nucleoside drug, forming a crucial foundation for the efficiency of our approach.

Table S2. A detailed description of the input features selected for supervised machine learning.

S.No.	Features	Description
1.	F1	Transmission (T)
2.	F2	$\frac{T}{T_{max}}$, where T_{max} is the maximum transmission value
3.	F3	$\frac{T}{T_{min}}$, where T_{min} is the minimum transmission value
4.	F4	$\frac{T}{T_{avg}}$, where T_{avg} is the average transmission value
5.	F5	Height (H); Difference between transmission values of two consecutive transmission peaks
6.	F6	Levels (L), Ratio of transmission values of two consecutive transmission peaks
7.	F7	$\frac{T}{Height}$; Height normalized transmission values
8.	F8	$\frac{T}{Level}$; Level normalized transmission values
9.	F9	$\frac{E}{Transmission}$; Transmission normalized energy values
10.	F10	$\frac{E}{Height}$; Height normalized energy values
11.	F11	$\frac{E}{Level}$; Level normalized energy values
12.	F12	$\frac{E}{T/Height}$; Ratio of energy values and height normalized transmission

values

13.	F13	$\frac{E}{T/Level}$	; Ratio of energy values and Level normalized transmission values

6. Spearman's Rank Correlation Matrix

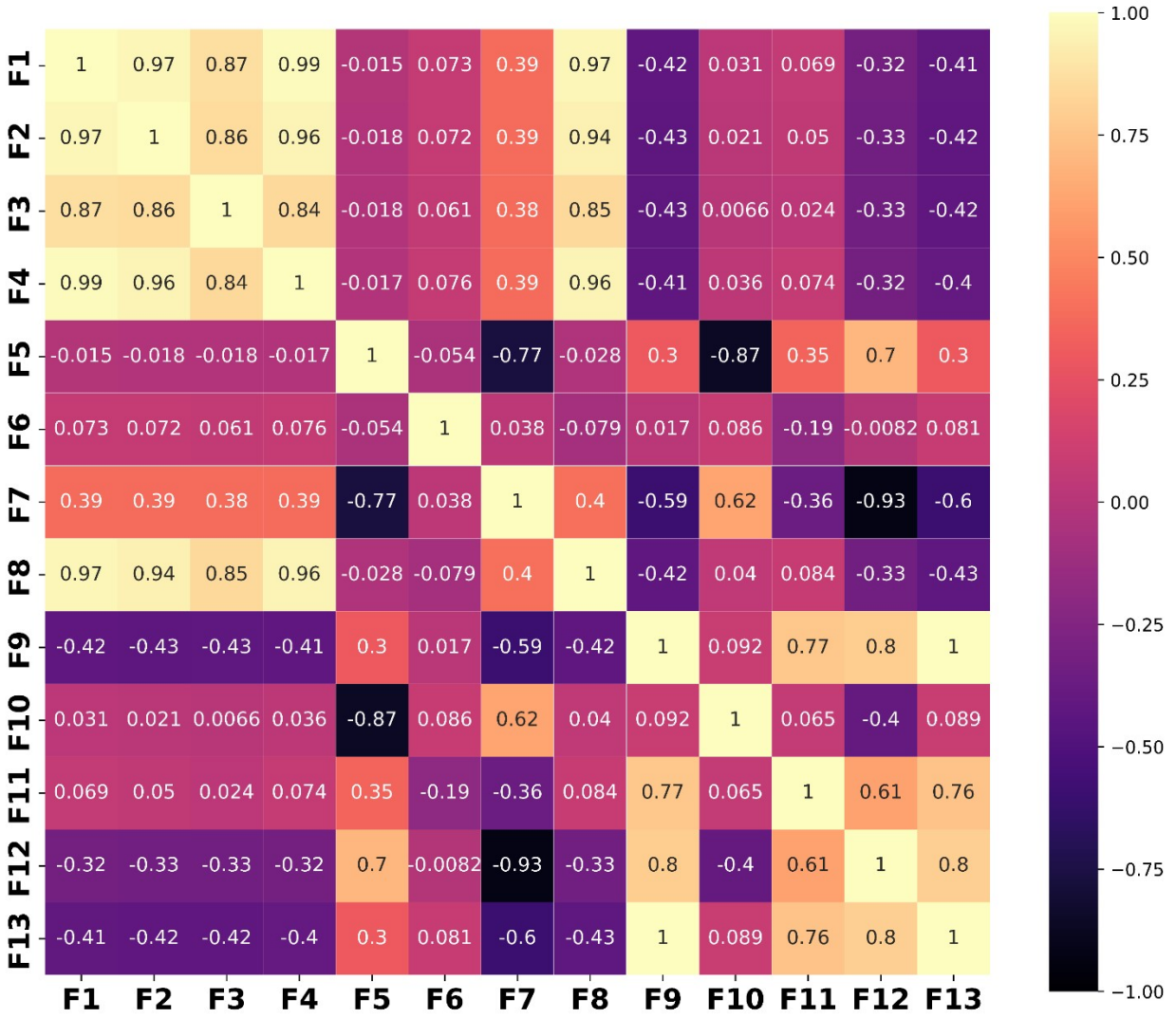

Figure S5. Spearman's rank correlation matrix illustrating Spearman's rank correlation coefficients (SCCs) between each possible pair of selected input feature vectors. A coefficient close to 1, -1, and 0 indicates a strong positive, strong negative, and weak or no monotonic relationship between the features, respectively.

7. Pearson's Correlation Matrix

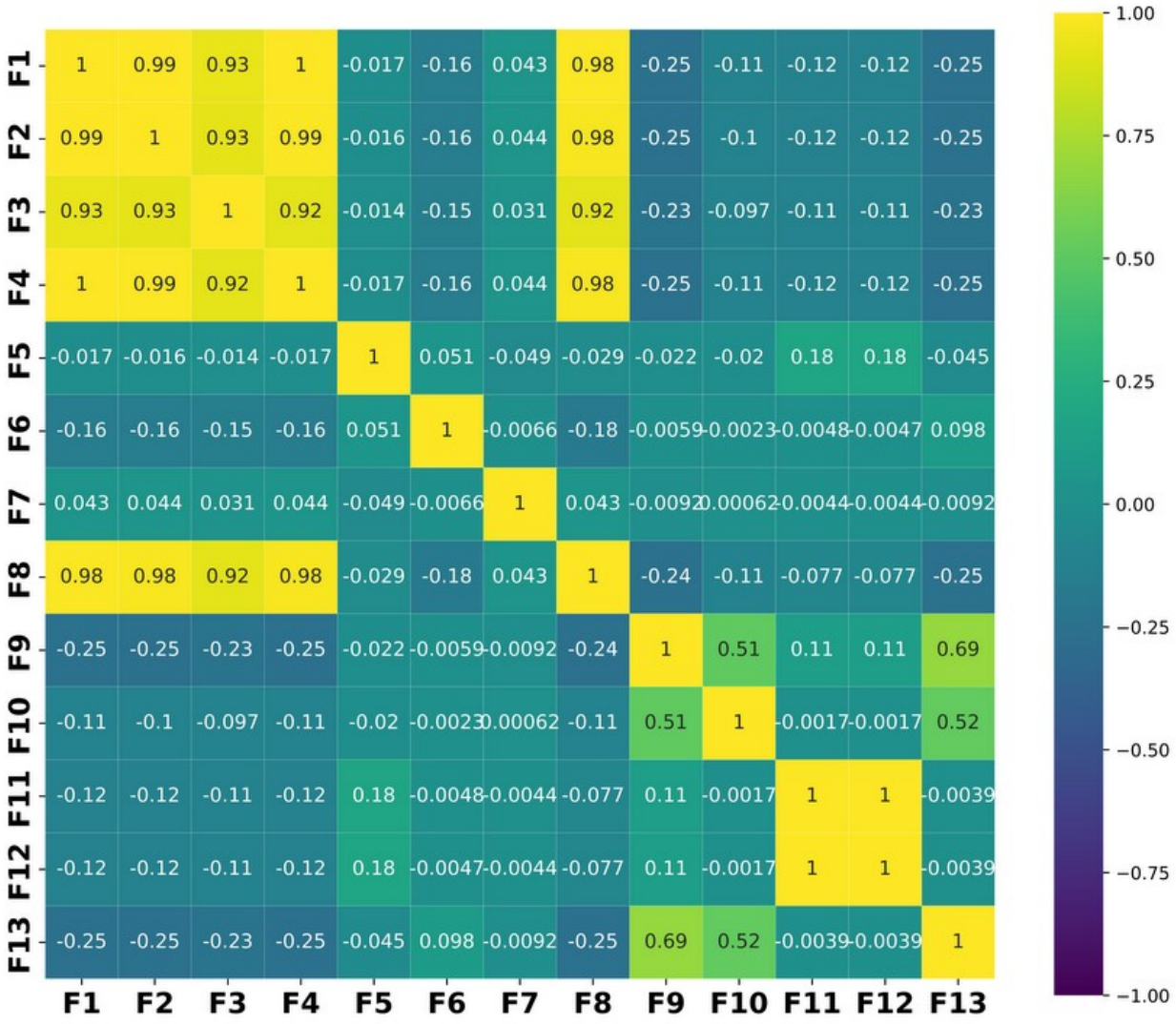

Figure S6. Pearson's correlation matrix illustrating standard Pearson's correlation coefficients (PCCs) among selected input feature vectors, offering a detailed perspective on the linear relationships between each pair. A coefficient close to 1, -1, and 0 indicates a strong positive, strong negative, and weak or no linear correlation between the input features.

8. Details of Tuned Hyperparameters

Table S3. Tuned hyperparameters for selected supervised ML classification algorithms. Hyperparameter tuning has been performed by employing the Randomized Search CV, as implemented in the scikit-learn package.

S.No.	ML Algorithms	Optimized Hyperparameters
1.	Random Forest Classification (RFC)	n_estimators= 100, random_state=400, min_samples_split= 2, min_samples_leaf= 1, max_features= None, max_depth= 32, criterion='entropy'
2.	Decision Tree Classification (DTC)	criterion='entropy',max_features=None,random_state=292, min_samples_split= 2, min_samples_leaf= 1, max_depth= 32
3.	K-Nearest Neighbor Classification (KNN)	n_neighbors= 18, weights='uniform', metric='euclidean', p=2
4.	Logistic Regression (LR)	penalty= 'none', solver= 'newton-cg', max_iter= 200, C=30, random_state=42
5.	Support Vector Machine Classification (SVM)	C=1000, cache_size=200, decision_function_shape='ovr', degree=5, gamma='scale', kernel='poly', max_iter=-1, shrinking=True, tol=0.001
6.	Extreme Gradient Boosting Classification (XGBC)	subsample= 1, objective= 'multi:softmax', num_class= 8, n_estimators= 1000, min_child_weight= 1, max_depth= 4, learning_rate= 0.3, gamma= 0.0001, colsample_bytree= 1.0
7.	Naive Bayes Classification (NBC)	priors=None, var_smoothing=0.05
8.	AdaBoost Classification (ADBC)	Base_estimator = DecisionTreeClassifier() , n_estimators= 200, learning_rate= 0.1

9. Train and Validation Accuracy

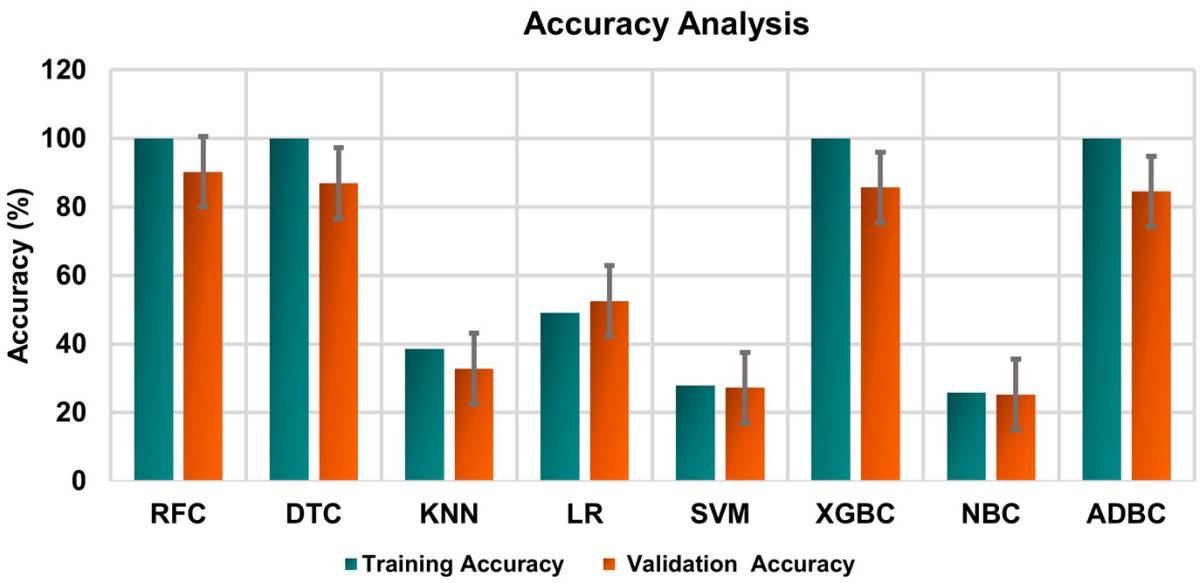

Figure S7. Train and test accuracy (%) for selected supervised ML classification algorithms, RFC, DTC, KNN, LR, SVM, XGBC, NBC, and ADBC.

10. 10-Fold Cross-Validation

Table S4. Mean and validation accuracy for each fold of 10-fold cross-validation using the optimized ML models, RFC, DTC, KNN, LR, SVM, XGBC, NBC, and ADBC.

Fold	RFC	DTC	KNN	LR	SVM	XGBC	NBC	ADBC
1	89.3	84.0	27.8	43.8	25.5	85.8	24.8	85.0
2	90.8	87.8	33.8	48.3	28.5	86.8	22.0	87.3
3	88.8	88.3	28.3	46.0	26.0	87.0	26.3	85.3
4	90.3	87.3	32.3	51.8	26.3	86.0	22.0	83.5
5	88.5	86.5	31.3	45.0	28.3	85.3	22.5	83.8
6	88.8	82.3	33.5	48.8	28.5	84.0	24.3	80.0
7	88.8	84.0	32.0	38.0	23.5	85.5	21.5	85.3
8	88.8	86.3	29.5	48.5	27.0	84.0	20.3	85.3
9	88.8	85.3	34.5	45.5	28.8	86.0	27.5	85.5
10	91.3	86.8	34.0	49.0	28.5	86.8	28.0	86.8
Mean Accuracy ± Standard Deviation	8 9.4 ± 1.(	8 5.8 ± 1.'	3 1.7 ± 2.-	4 6.5 ± 3.:	2 7.1 ± 1.:	8 5.7 ± 1.:	23.9 ± 2.7	84.8 ± 2.3
Validation Accuracy	90.2	86.9	32.8	49.0	27.2	85.7	25.0	85.2

11. HOMO Energy Values

Table S5. HOMO energy values (in a.u.) of the nucleoside drug molecules.

S.No.	Nucleoside Drugs	HOMO (a.u.)
1.	5-Azacytidine	-0.27
2.	5-Fluorouridine	-0.26
3.	6-Azauridine	-0.27
4.	GS-441524	-0.24
5.	Loxoribine	-0.21
6.	Mizoribine	-0.22
7.	Ribavirin	-0.27
8.	Zebularine	-0.26

12. ML Calling of Nucleoside Drugs in 0° Configuration

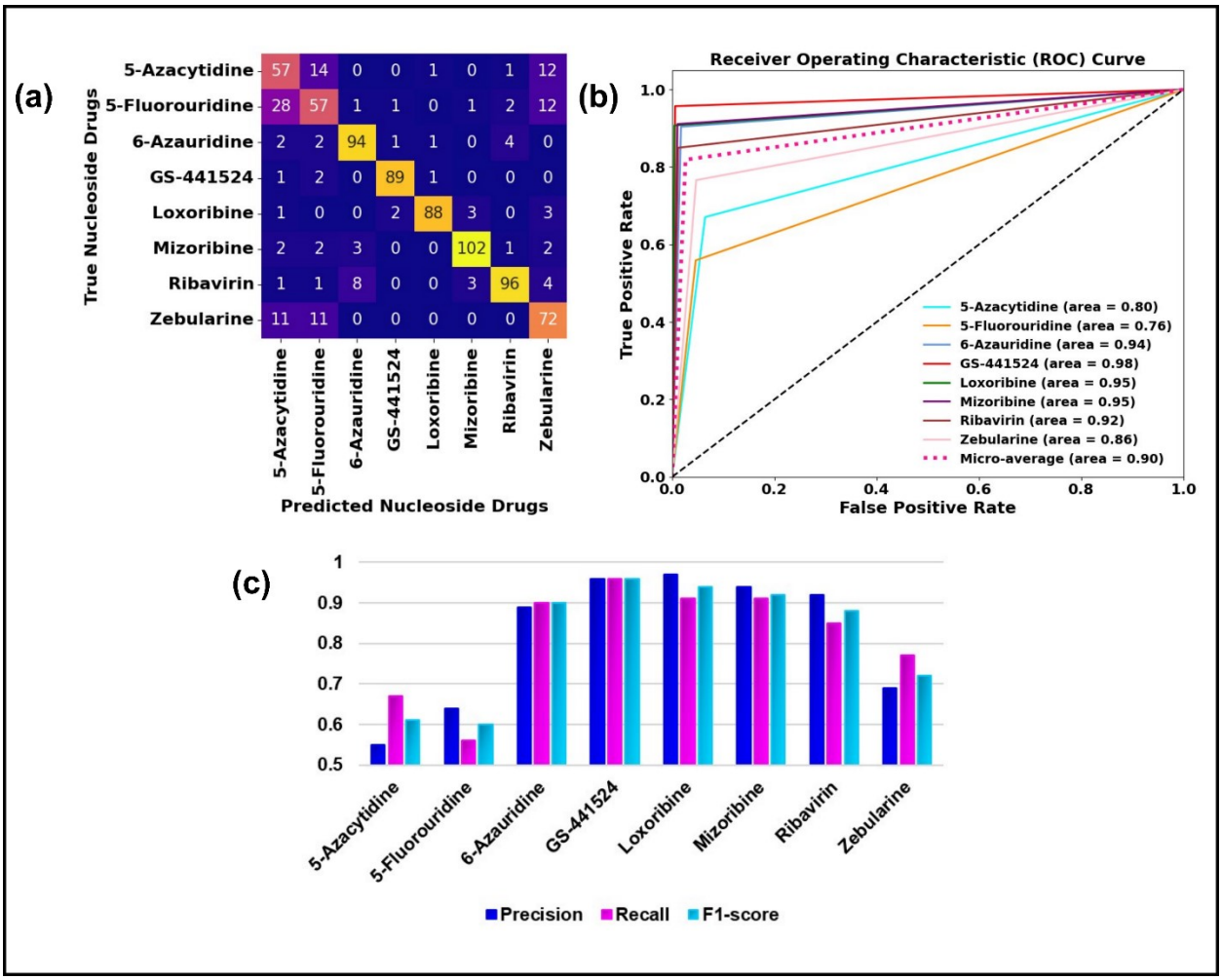

Figure S8. **(a)** Confusion matrix for RFC_1 calling of nucleoside drugs in the 0° configuration inside the nanopore, **(b)** receiver operating characteristic (ROC) curve illustrating the performance of RFC algorithm in distinguishing different classes of nucleoside drugs, and **(c)** classification report enclosing parameters precision, recall, and f1-score.

13. ML Calling of Nucleoside Drugs in 30° Configuration

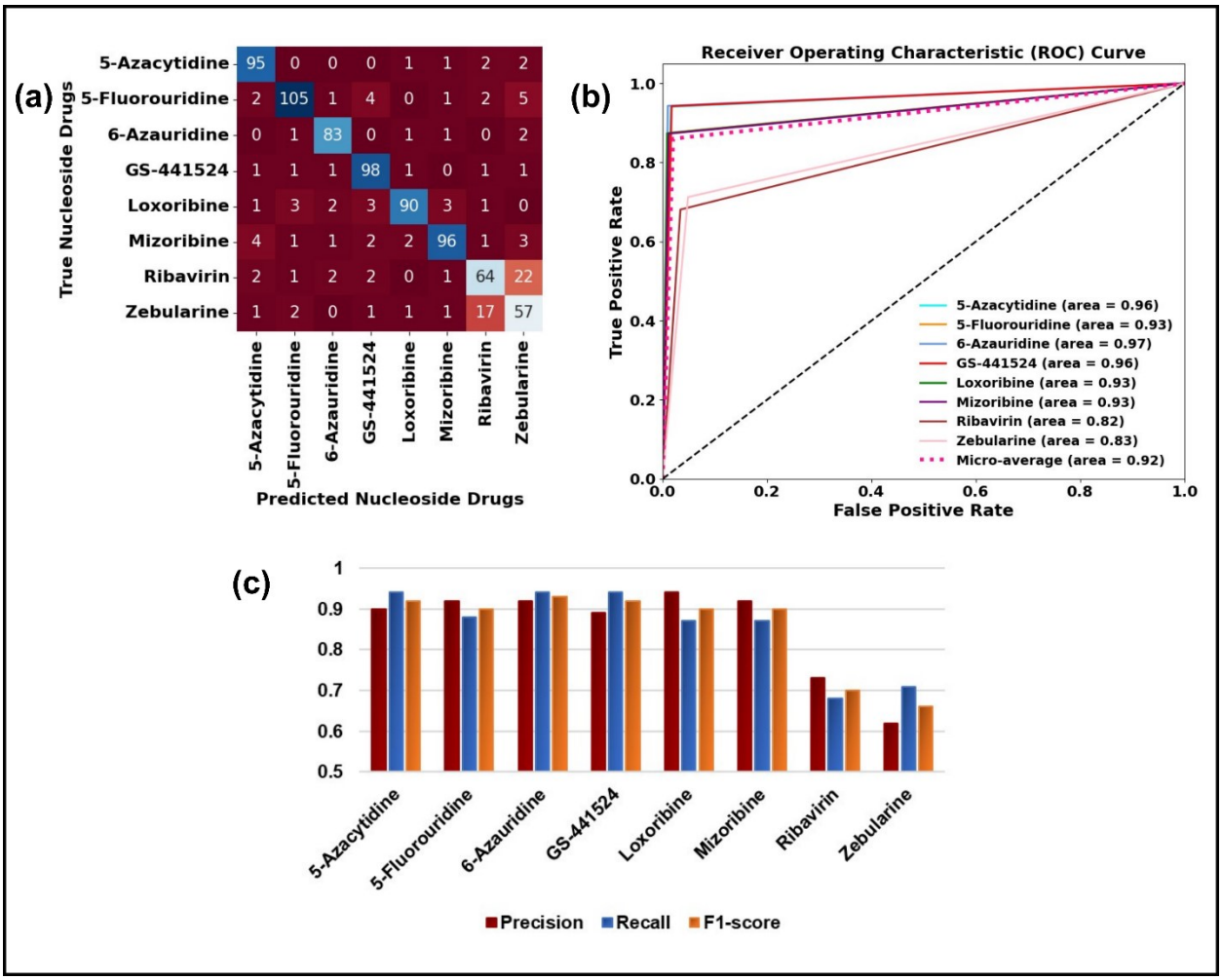

Figure S9. **(a)** Confusion matrix for RFC_2 calling of nucleoside drugs in the 30° configuration inside the nanopore, **(b)** receiver operating characteristic (ROC) curve illustrating the performance of RFC algorithm in distinguishing different classes of nucleoside drugs, and **(c)** classification report enclosing parameters precision, recall, and f1-score.

14. ML Calling of Nucleoside Drugs in 60° Configuration

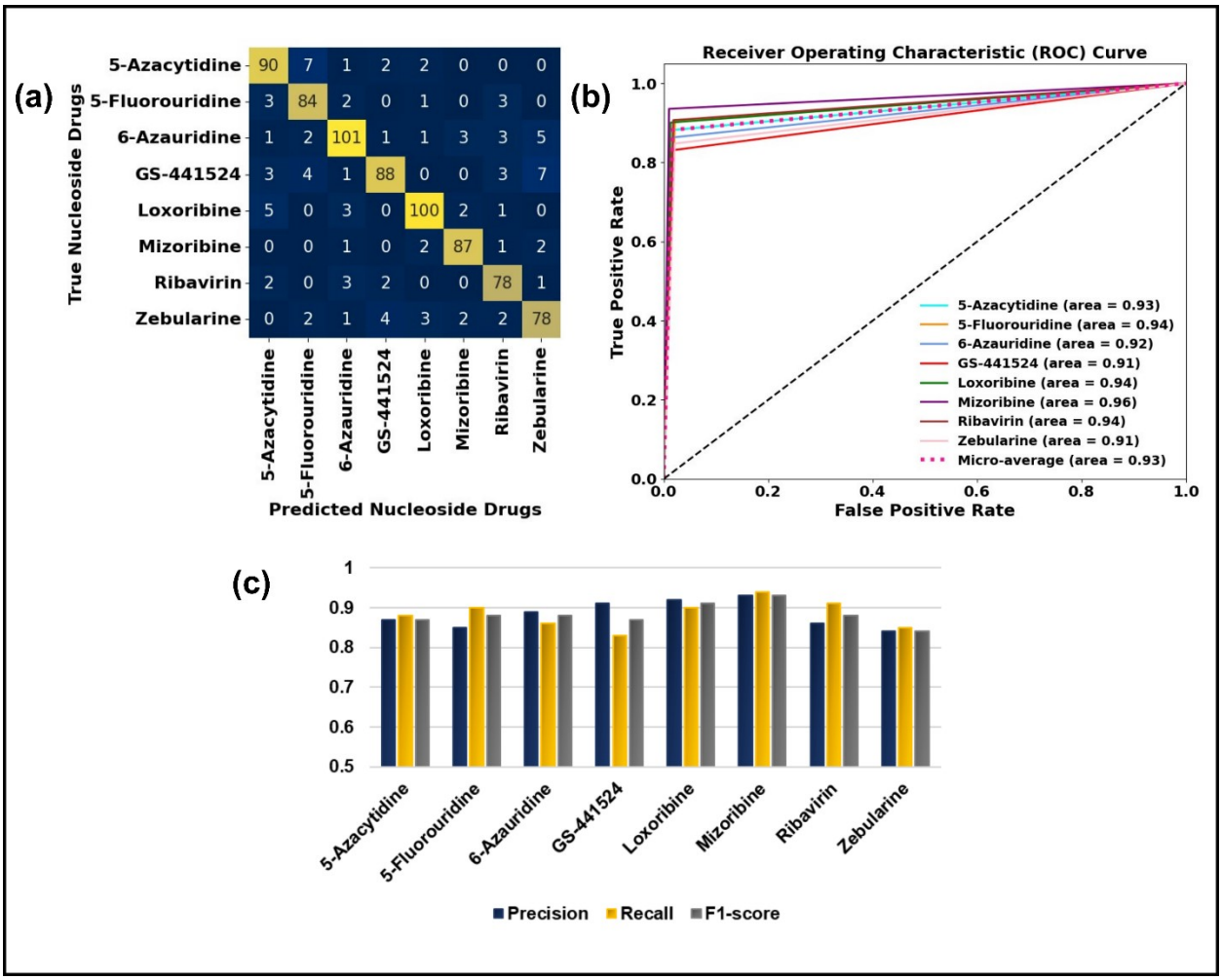

Figure S10. **(a)** Confusion matrix for RFC_3 calling of nucleoside drugs in the 60° configuration inside the nanopore, **(b)** receiver operating characteristic (ROC) curve illustrating the performance of RFC algorithm in distinguishing different classes of nucleoside drugs, and **(c)** classification report enclosing parameters precision, recall, and f1-score.

15. ML Calling of Nucleoside Drugs in 90° Configuration

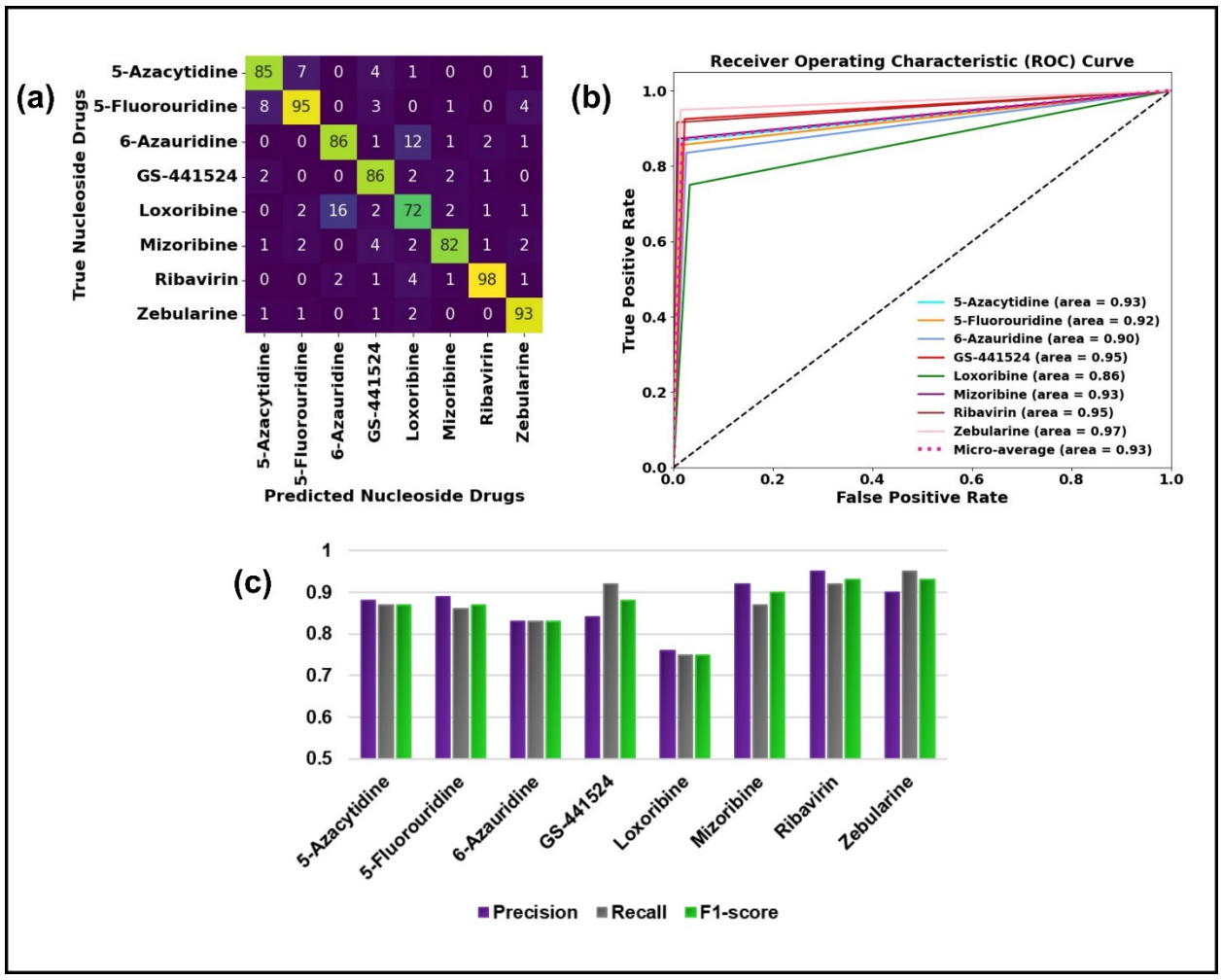

Figure S11. **(a)** Confusion matrix for RFC_4 calling of nucleoside drugs in the 90° configuration inside the nanopore, **(b)** receiver operating characteristic (ROC) curve illustrating the performance of RFC algorithm in distinguishing different classes of nucleoside drugs, and **(c)** classification report enclosing parameters precision, recall, and f1-score.

16. ML Calling of Nucleoside Drugs in +1.0 Å Configuration

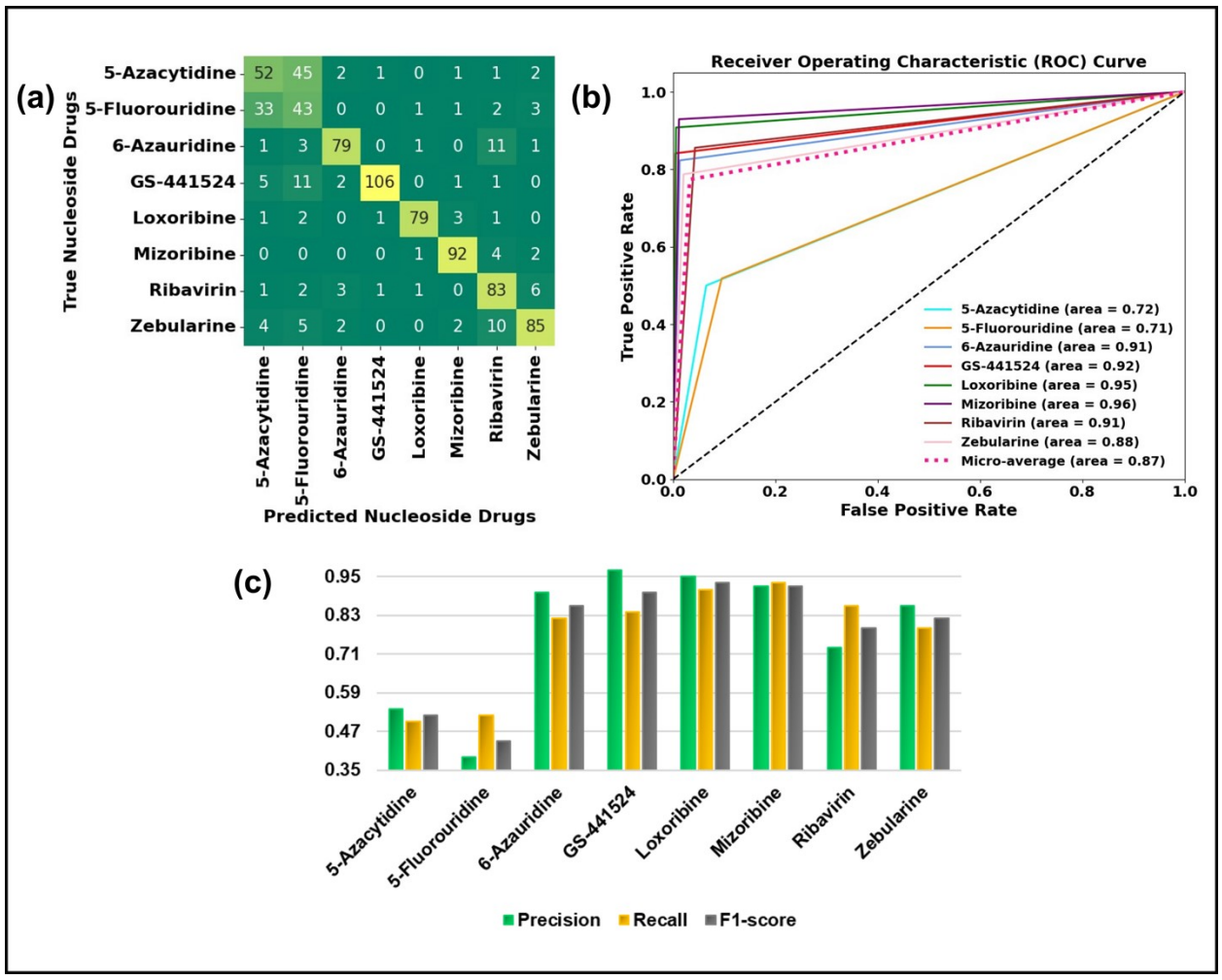

Figure S12. **(a)** Confusion matrix for RFC_5 calling of nucleoside drugs in the +1.0 Å configuration inside the nanopore, **(b)** receiver operating characteristic (ROC) curve illustrating the performance of RFC algorithm in distinguishing different classes of nucleoside drugs, and **(c)** classification report enclosing parameters precision, recall, and f1-score.

17. ML Calling of Nucleoside Drugs in -1.0 Å Configuration

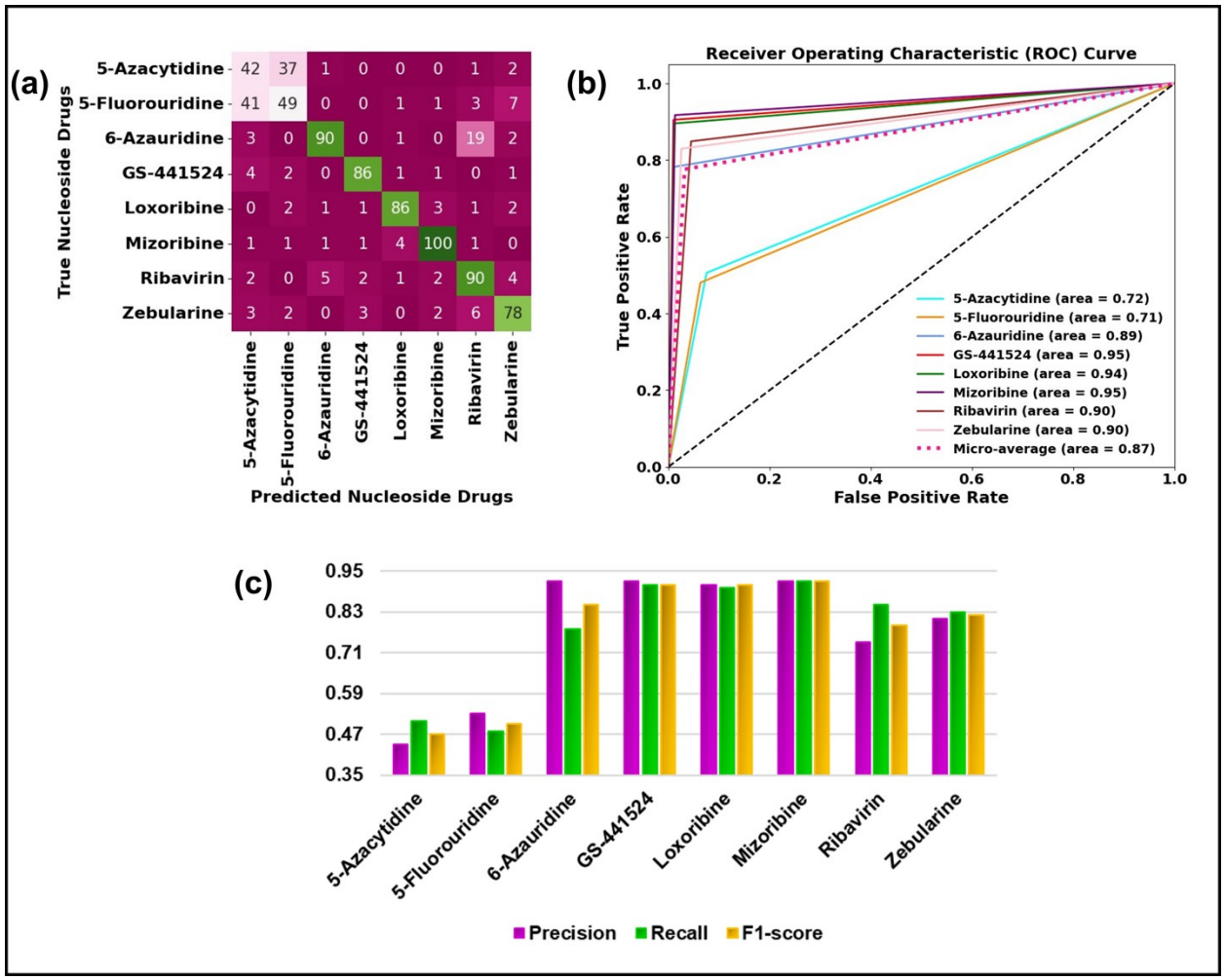

Figure S13. **(a)** Confusion matrix for RFC_6 calling of nucleoside drugs in the -1.0 Å configuration inside the nanopore, **(b)** receiver operating characteristic (ROC) curve illustrating the performance of RFC algorithm in distinguishing different classes of nucleoside drugs, and **(c)** classification report enclosing parameters precision, recall, and f1-score.

18. 10-Fold Cross-Validation

Table S6. Mean and validation accuracy for each fold of 10-fold cross-validation using the algorithms, RFC_1, RFC_2, RFC_3, RFC_4, RFC_5, RFC_6 for calling nucleoside drugs in configurations 0°, 30°, 60°, 90°, +1.0 Å and − 1.0 Å, respectively.

Fold	RFC_1	RFC_2	RFC_3	RFC_4	RFC_5	RFC_6
1	80.5	83.0	91.3	82.8	74.0	79.0
2	81.8	85.0	89.5	88.3	77.8	77.5
3	83.5	88.3	90.0	88.5	80.0	78.8
4	78.3	82.8	89.3	83.3	78.8	76.5
5	81.8	86.5	88.3	86.8	80.0	78.5
6	79.0	85.0	87.8	87.5	77.8	79.3
7	80.5	81.0	87.8	86.5	73.8	78.5
8	76.3	87.0	87.8	87.3	79.8	75.5
9	79.3	84.3	86.0	84.8	79.5	80.3
10	82.0	86.0	89.8	87.3	73.3	74.8
Mean Accuracy ± Standard Deviation	8 0.3 ± 2.1	8 4.9 ± 2.2	8 8.7 ± 1.5	8 6.3 ± 2.0	7 7.5 ± 2.8	7 7.9 ± 1.8
Test Accuracy	81.9	86.0	88.3	87.2	77.4	77.6

19. Interpretability of ML Calling in 0° Configuration

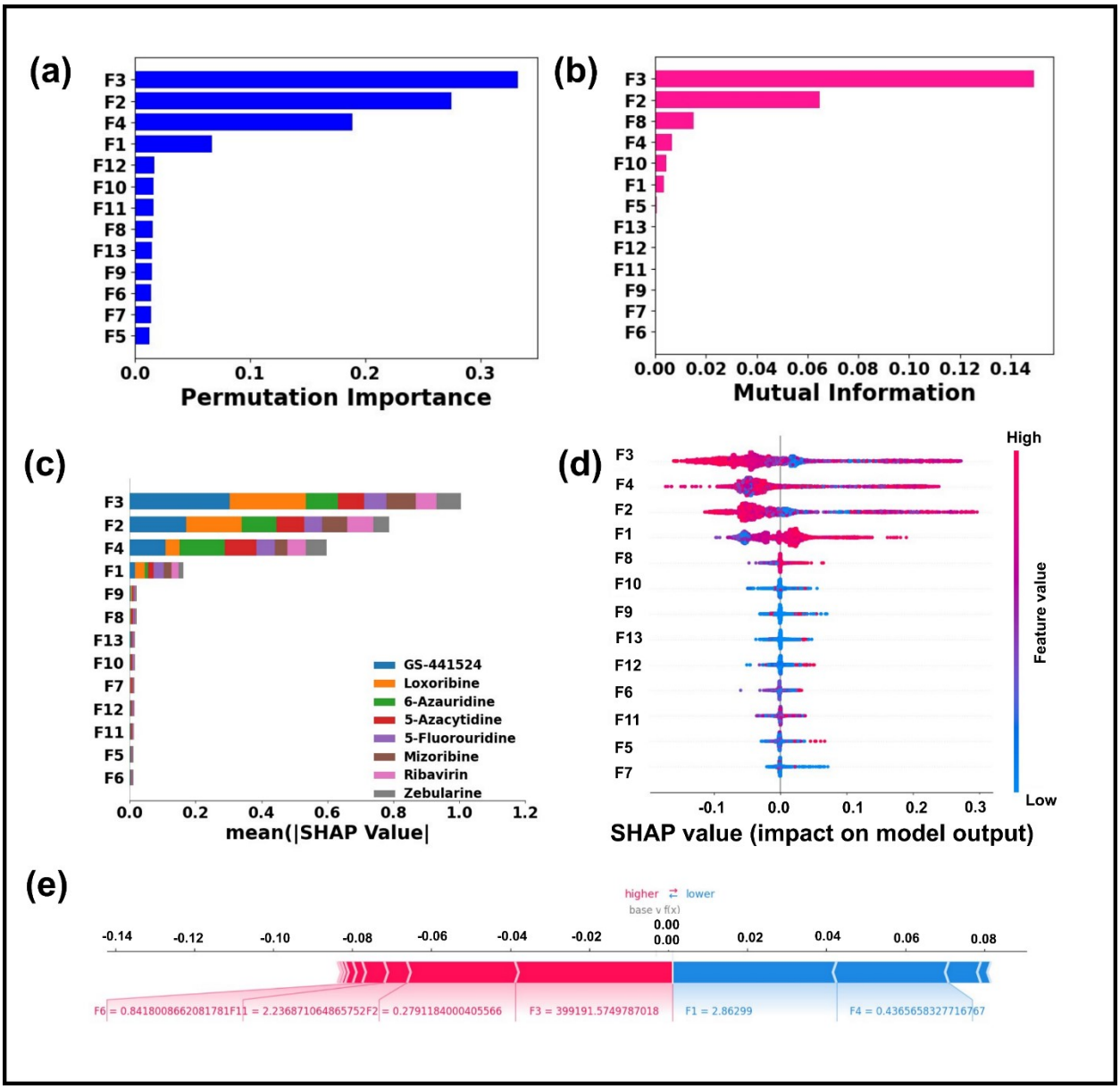

Figure S14. (a) Global permutation feature importance plot, **(b)** mutual information feature importance plot, **(c)** SHAP summary bar plot illustrating the contribution of each input feature in

the prediction of an individual drug class, **(d)** SHAP beeswarm plot illustrating the contribution of each feature toward every prediction made by the algorithm, and **(e)** SHAP summary force plot illustrating the contribution of input features toward a single prediction.

20. Interpretability of ML Calling in 30° Configuration

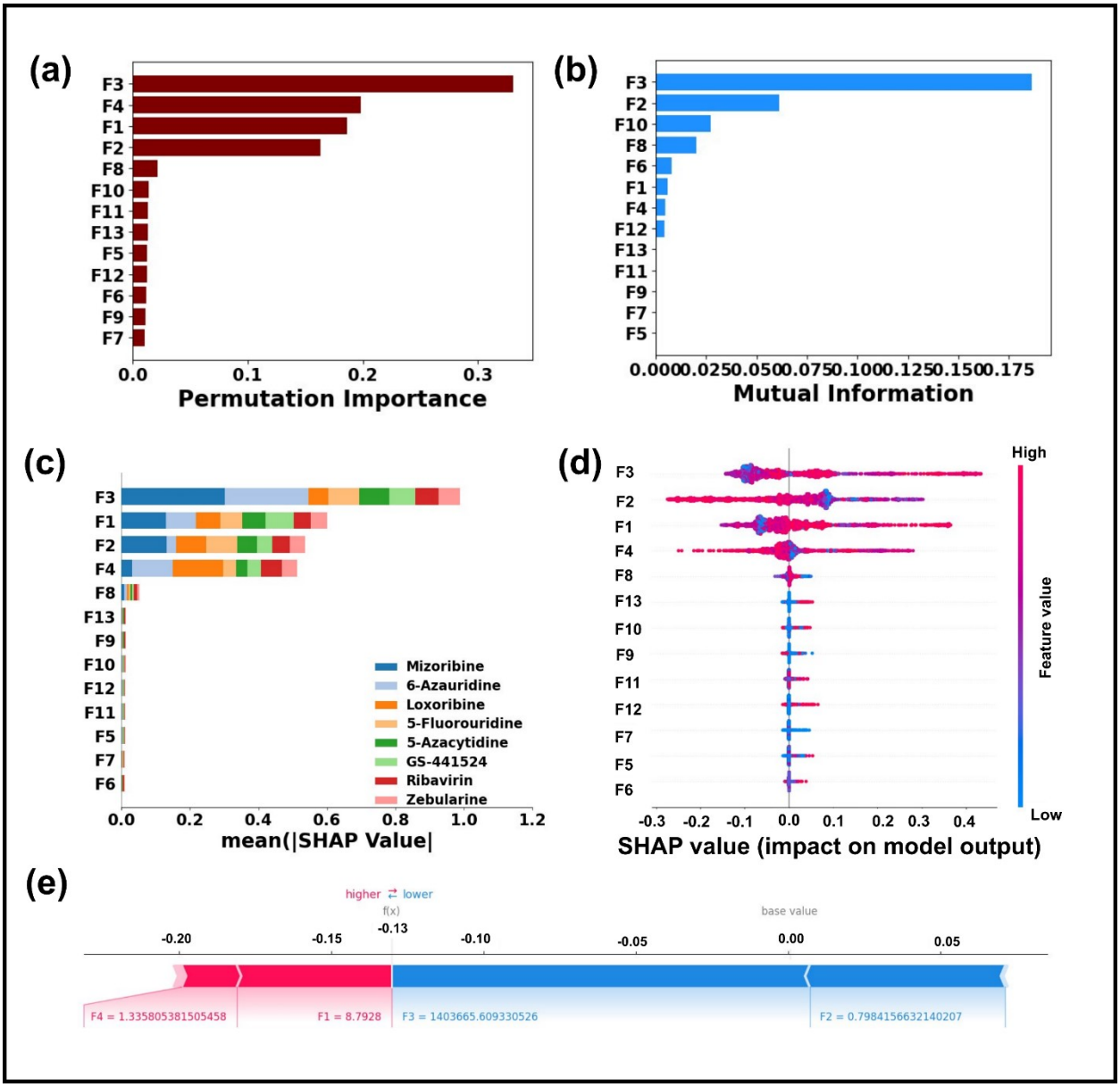

Figure S15. (a) Global permutation feature importance plot, **(b)** mutual information feature importance plot, **(c)** SHAP summary bar plot illustrating the contribution of each input feature in

the prediction of an individual drug class, **(d)** SHAP beeswarm plot illustrating the contribution of each feature toward every prediction made by the algorithm, and **(e)** SHAP summary force plot illustrating the contribution of input features toward a single prediction.

21. Interpretability of ML Calling in 60° Configuration

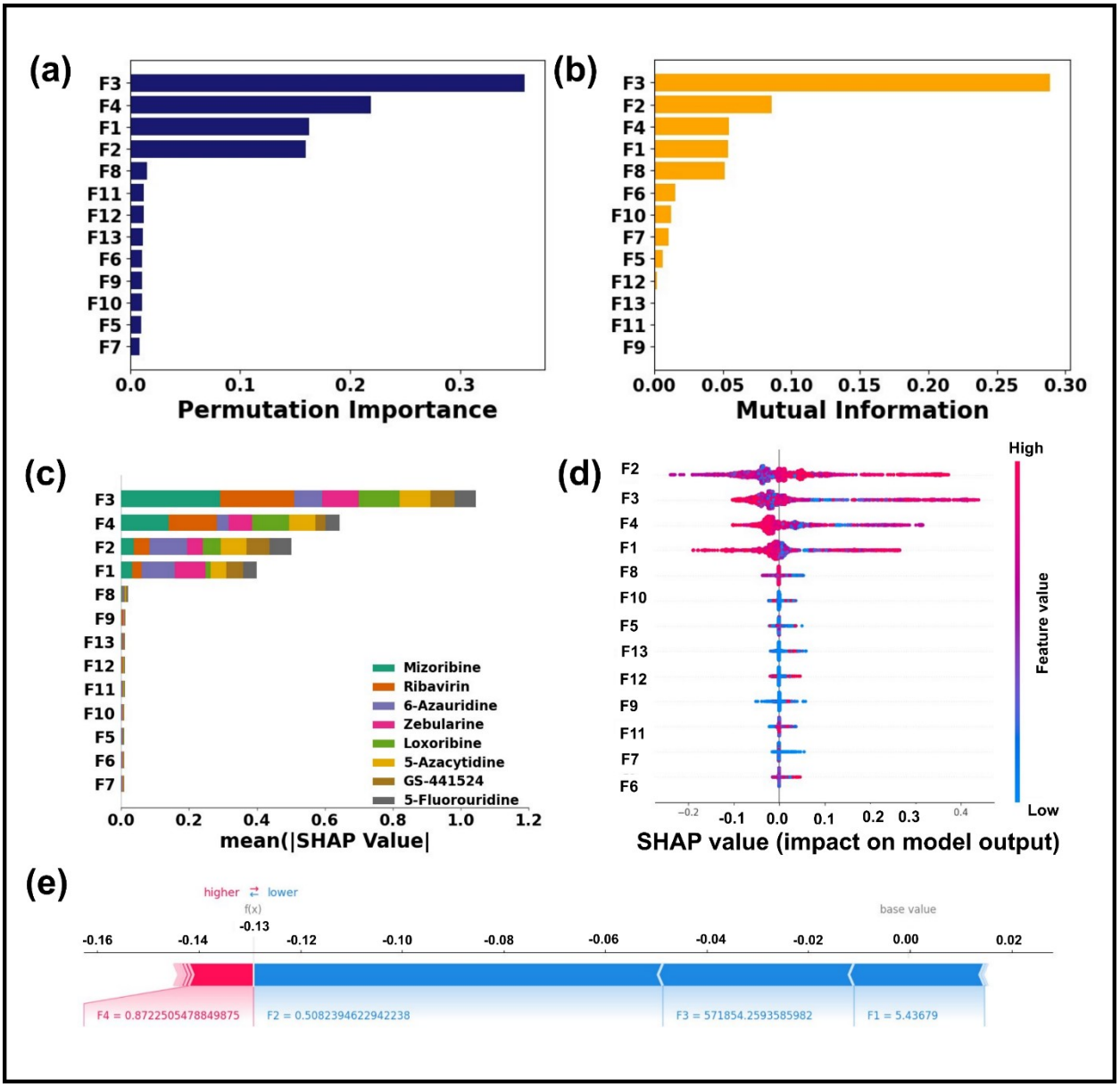

Figure S16. (a) Global permutation feature importance plot, **(b)** mutual information feature importance plot, **(c)** SHAP summary bar plot illustrating the contribution of each input feature in the prediction of an individual drug class, **(d)** SHAP beeswarm plot illustrating the contribution of each feature toward every prediction made by the algorithm, and **(e)** SHAP summary force plot illustrating the contribution of input features toward a single prediction.

22. Interpretability of ML Calling in 90° Configuration

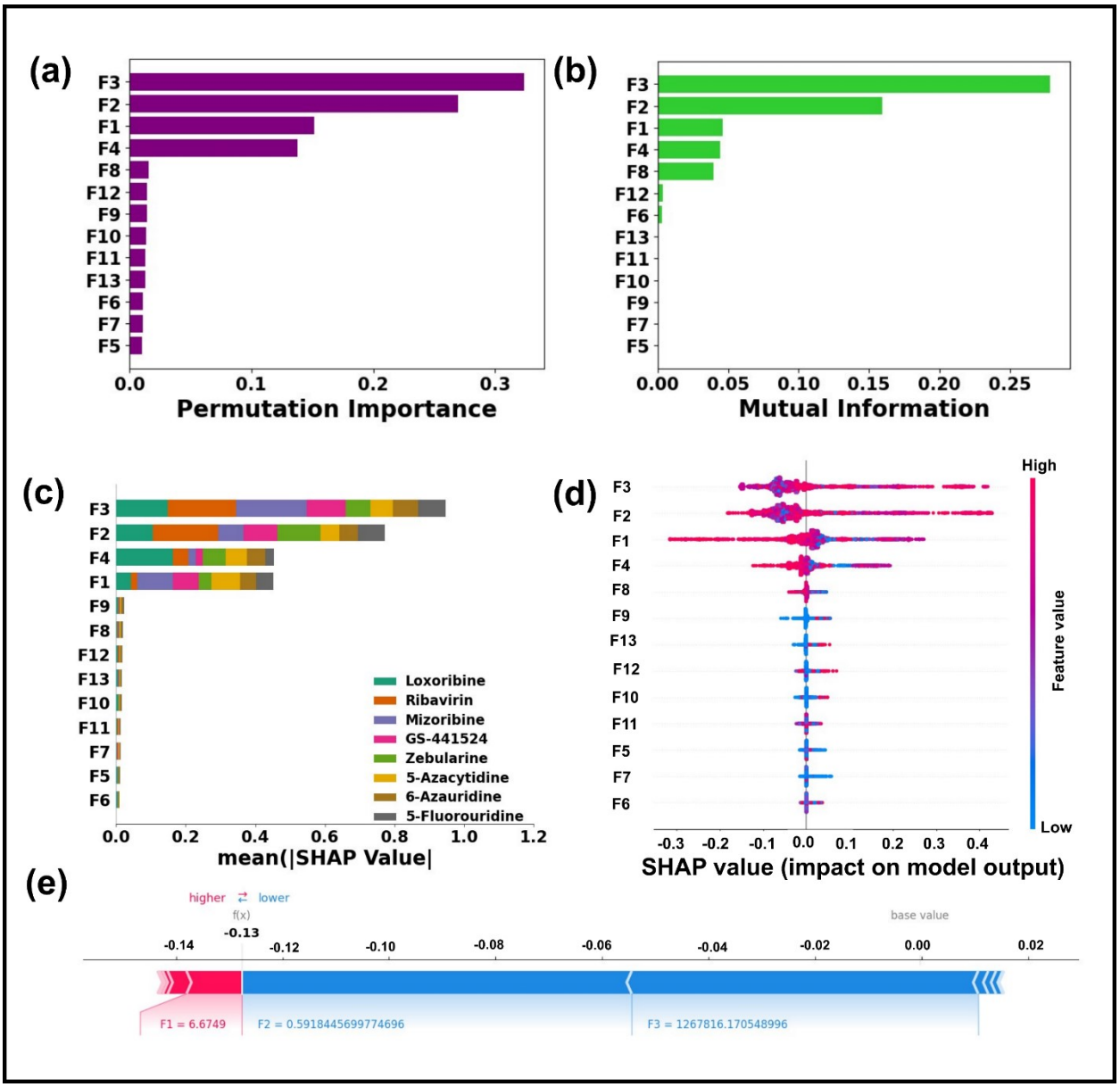

Figure S17. (a) Global permutation feature importance plot, **(b)** mutual information feature importance plot, **(c)** SHAP summary bar plot illustrating the contribution of each input feature in the prediction of an individual drug class, **(d)** SHAP beeswarm plot illustrating the contribution of each feature toward every prediction made by the algorithm, and **(e)** SHAP summary force plot illustrating the contribution of input features toward a single prediction.

23. Interpretability of ML Calling in +1.0 Å Configuration

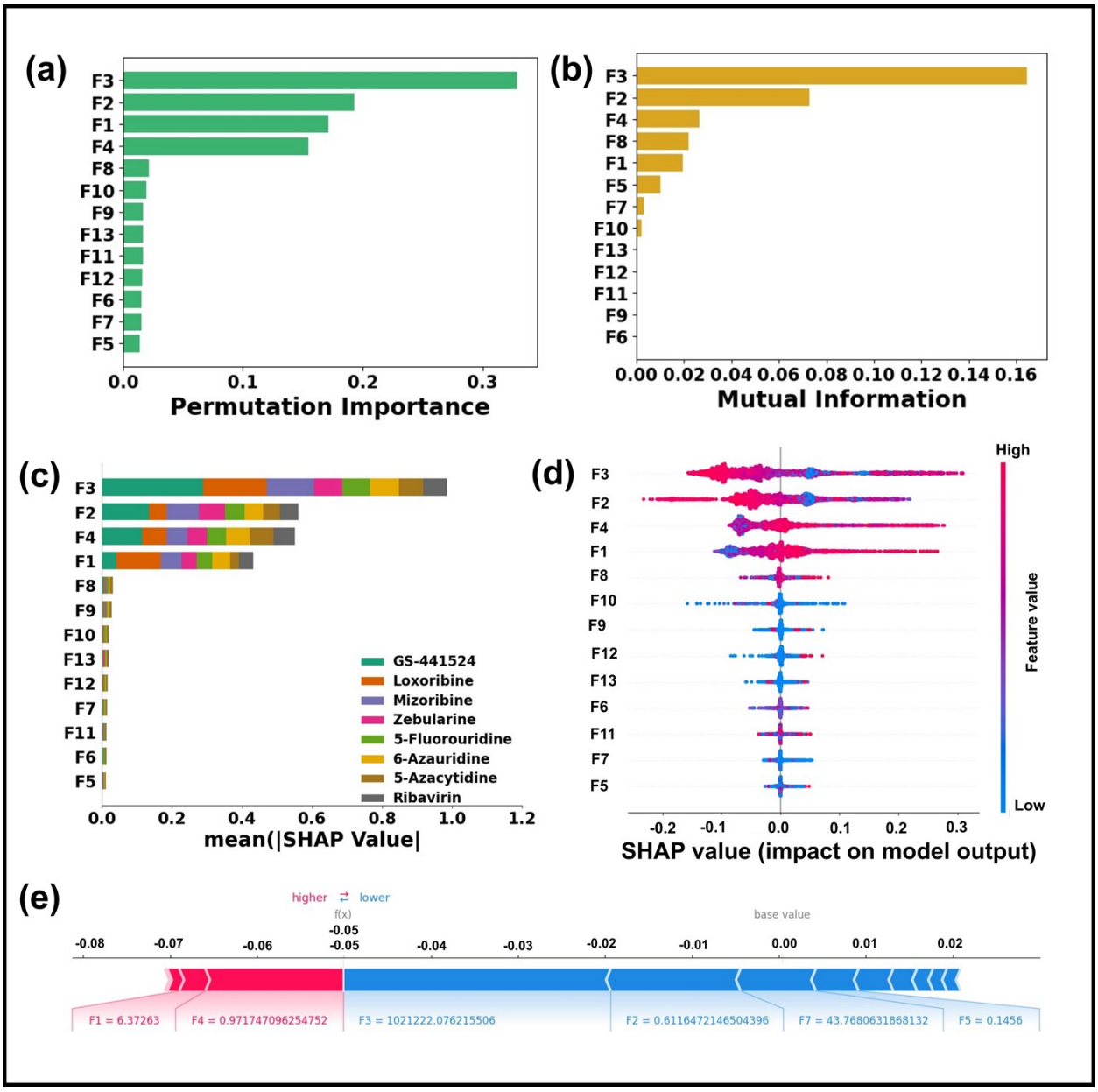

Figure S18. (a) Global permutation feature importance plot, **(b)** mutual information feature importance plot, **(c)** SHAP summary bar plot illustrating the contribution of each input feature in the prediction of an individual drug class, **(d)** SHAP beeswarm plot illustrating the contribution of each feature toward every prediction made by the algorithm, and **(e)** SHAP summary force plot illustrating the contribution of input features toward a single prediction.

24. Interpretability of ML Calling in -1.0 Å Configuration

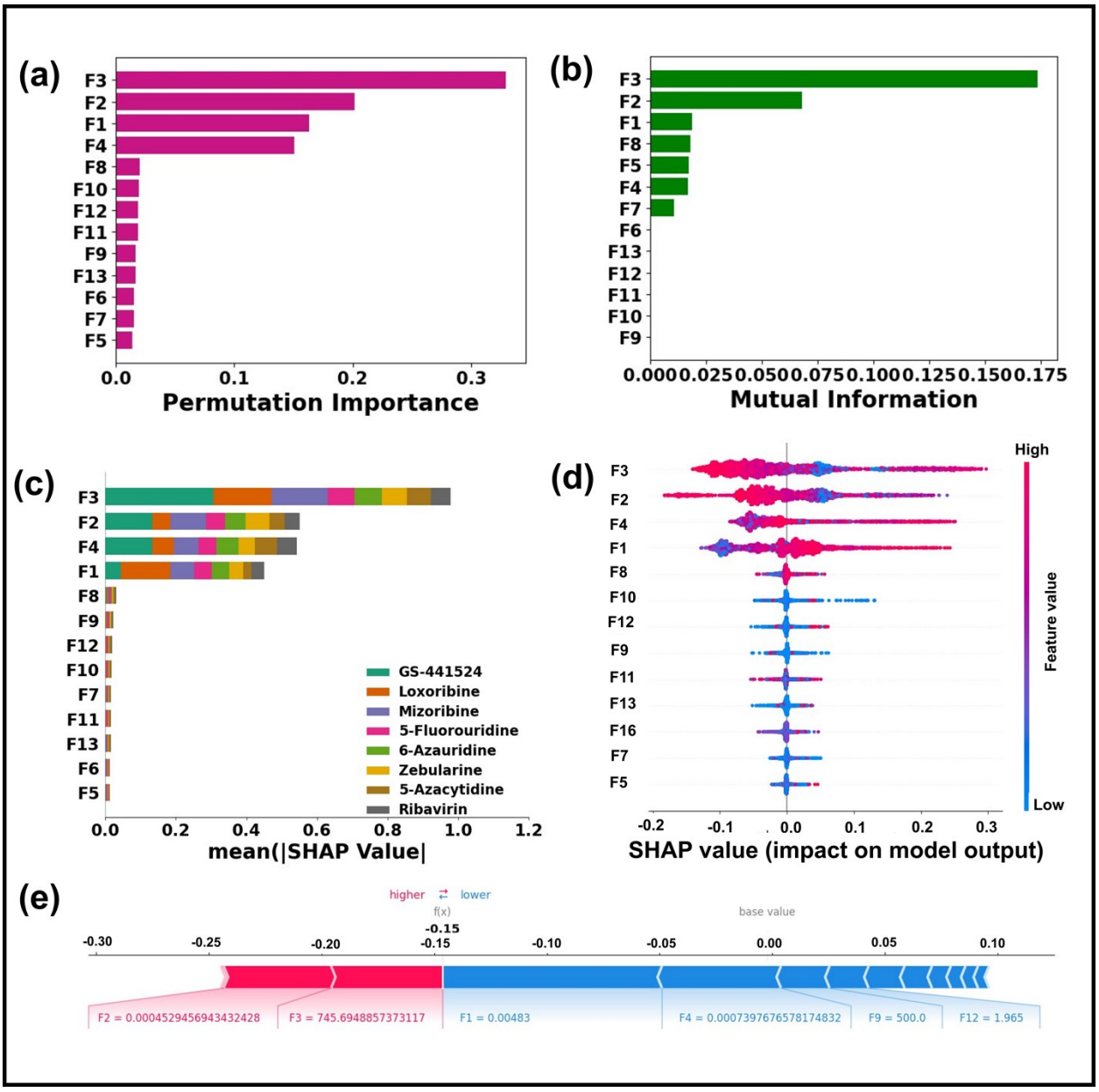

Figure S19. (a) Global permutation feature importance plot, **(b)** mutual information feature importance plot, **(c)** SHAP summary bar plot illustrating the contribution of each input feature in the prediction of an individual drug class, **(d)** SHAP beeswarm plot illustrating the contribution of each feature toward every prediction made by the algorithm, and **(e)** SHAP summary force plot illustrating the contribution of input features toward a single prediction.

25. Unsupervised ML Algorithms and Tuned Hyperparameters

Table S7. A detailed description of optimized hyperparameters for selected unsupervised ML algorithms.

S.No.	ML Algorithms	Optimized Hyperparameters
1.	K-Means Clustering	n_clusters=8, init='k-means++', n_init=10, max_iter=300, tol=0.0001, verbose=0, random_state=None, copy_x=True, algorithm='auto'
2.	Hierarchical Agglomerative Clustering	n_clusters=8, linkage='ward'
3.	DBSCAN Clustering	eps=0.5, min_samples=5
4.	MeanShift Clustering	bandwidth=0.4
5.	BIRCH Clustering	branching_factor = 50, threshold=0.5, n_clusters=8
6.	Gaussian Mixture Clustering	n_components=4, random_state=0

26. Unsupervised ML Clustering

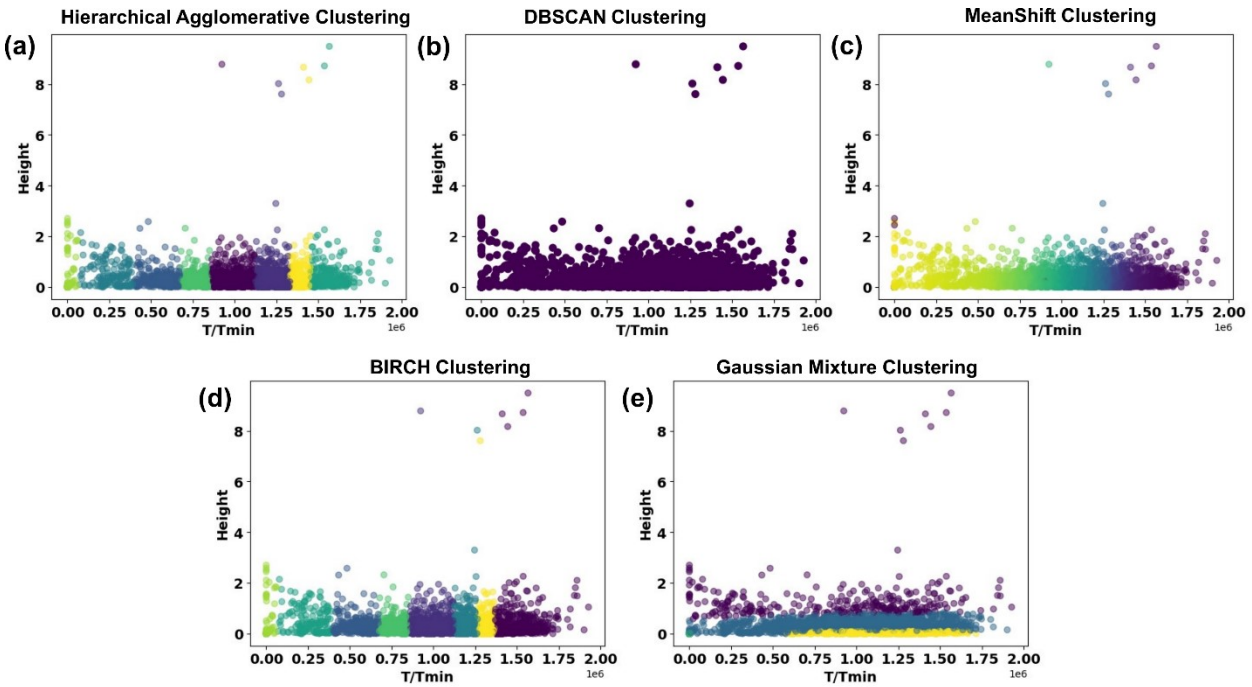

Figure S20. Clustering performance of unsupervised ML algorithms, hierarchical agglomerative clustering, DBSCAN clustering, meanshift clustering, BIRCH clustering, and gaussian mixture clustering to identify the nanopore events of 8 nucleoside drugs.

26. Unsupervised Performance Metrics Evaluation

Table S8. Performance metrics evaluation for selected unsupervised ML algorithms.

S.No.	Unsupervised ML Algorithms	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
1.	K-Means Clustering	0.57	0.47	28021.7
2.	Hierarchical Agglomerative Clustering	0.51	0.49	22090.5
5.	BIRCH Clustering	0.51	0.49	22090.5