SUPPORTING INFORMATION

Evaluating the Use of Graph Neural Network and Transfer Learning for Oral Bioavailability Prediction

Sherwin S.S. Ng, Yunpeng Lu

School of Chemistry, Chemical Engineering and Biotechnology, Nanyang Technological University, 21 Nanyang Link, Singapore 637371.

A. Hyperparameters for Random Forests models

All Random Forest models were optimised using Optuna[1] in 30 evaluations and five-fold cross validation method using oral bioavailability train dataset.

Table S1. Best parameters for Random Forest models.

	n_estimators	**max_depth**
Molecular Descriptors	76	96
Morgan Fingerprints	89	40
RDKit Fingerprints	28	11
MACCSkeyys	85	46

B. Hyperparameters for GNN models

All GNN models were optimised using Optuna[1] in 30 evaluations and five-fold cross validation method using oral bioavailability train dataset.

Table S2. Best parameters for GIN model.

Hyperparameters	**Values**
num_layers	1
hidden_size	66
learning_rate	0.00889495369073538

Table S3. Best parameters for Graph Transformer model.

Hyperparameters	**Values**
num_layers	2
hidden_size	439
n_heads	1
dropout	0.269754753387312
learning_rate	0.007890910361468965

Table S4. Best parameters for Vertical GNN model.

Hyperparameters	**Values**
num_gin_layers	2
num_graph_trans_layer	2
hidden_size	122
n_heads	2
dropout	0.36738054656589025
learning_rate	0.00452976319043267

C. Hyperparameters for Transfer Learning GNN Models

All Transfer Learning GNN models were optimised using Optuna[1] in 30 evaluations using solubility train and validation dataset.

Table S5. Best parameters for Transfer Learning Vertical GNN model.

Hyperparameters	**Values**
num_gin_layers	2
num_graph_trans_layer	2
hidden_size	245
n_heads	1
dropout	0.30146027310173296
learning_rate	0.0012649520485726895

Table S6. Best parameters for Transfer Learning Pre-Trained Vertical GNN model.

Hyperparameters	**Values**
learning_rate	0.00012649520485726895
es_trigger	15

D. List of Molecular Descriptors

Table S7. Descriptions of the 45 molecular descriptors used to build random forest model to predict oral bioavailability.

Molecular Descriptors	Categories	Details
MaxEStateIndex	Basic EState descriptors	States the maximum EState index
MinEStateIndex	Basic EState descriptors	States the minimum EState index
MaxAbsEStateIndex	Basic EState descriptors	States the maximum absolute EState index
MinAbsEStateIndex	Basic EState descriptors	State the minimum absolute EState index
qed	quantitative estimation of drug-likeness	States the weighted sum of ADS-mapped properties
MolWt	General descriptors	States the average molecular weight of a molecule
HeavyAtomMolWt	General descriptors	States the average molecular weight of a molecule with the removal of hydrogen atoms
ExactMolWt	General descriptors	States the exact molecular weight of a molecule
NumValenceElectrons	General descriptors	States the number of valence electrons a molecule possesses
MaxPartialCharge	General descriptors	States the maximum partial charge
MinPartialCharge	General descriptors	States the minimal partial charge
MaxAbsPartialCharge	General descriptors	States the maximum absolute partial charge
MinAbsPartialCharge	General descriptors	States the minimal absolute partial charge
FpDensityMorgan1	General descriptors	Morgan fingerprint, radius 1
FpDensityMorgan2	General descriptors	Morgan fingerprint, radius 2
FpDensityMorgan3	General descriptors	Morgan fingerprint, radius 3
BCUT2D_MWHI	BCUT descriptors	Highest eigenvalue of Burden matrix weighted by atomic masses
BCUT2D_MWLOW	BCUT descriptors	Lowest eigenvalue of Burden matrix weighted by atomic masses
BCUT2D_CHGHI	BCUT descriptors	the highest eigenvalue of Burden matrix weighted by gasteiger charges
BCUT2D_CHGLO	BCUT descriptors	the lowest eigenvalue of Burden matrix weighted by gasteiger charges
BCUT2D_LOGPHI	BCUT descriptors	the highest eigenvalue of Burden matrix weighted by Crippen LogP
BCUT2D_LOGPLOW	BCUT descriptors	the lowest eigenvalue of Burden matrix weighted by Crippen LogP

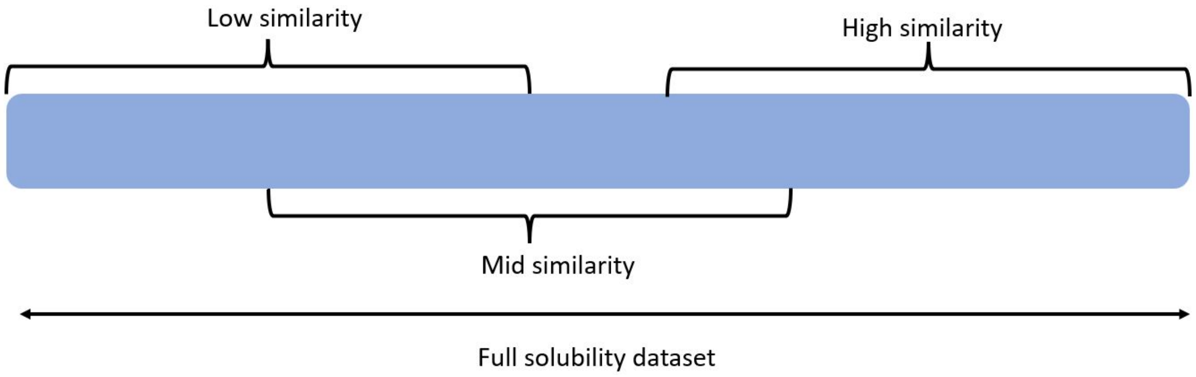
BCUT2D_MRHI	BCUT descriptors	the highest eigenvalue of Burden matrix weighted by Crippen MR
BCUT2D_MRLOW	BCUT descriptors	the lowest eigenvalue of Burden matrix weighted by Crippen MR
BalabanJ	Topological/topochemical descriptors	Balaban's J value
BertzCT	Topological/topochemical descriptors	A topological index meant to quantify "complexity"
Chi0	Topological/topochemical descriptors	From equations (1), (9) and (10) of reference 2
Chi0n	Topological/topochemical descriptors	Similar to Hall Kier Chi0v, but uses nVal instead of valence.
Chi0v	Topological/topochemical descriptors	From equations (5),(9) and (10) of reference 2
Chi1	Topological/topochemical descriptors	From equations (1),(11) and (12) of reference 2
Chi1n	Topological/topochemical descriptors	Similar to Hall Kier Chi1v, but uses nVal instead of valence
Chi1v	Topological/topochemical descriptors	From equations (5),(11) and (12) of reference 2
Chi2n	Topological/topochemical descriptors	Similar to Hall Kier Chi2v, but uses nVal instead of valence.
Chi2v	Topological/topochemical descriptors	From equations (5),(15) and (16) of reference 2
Chi3n	Topological/topochemical descriptors	Similar to Hall Kier Chi3v, but uses nVal instead of valence.
Chi3v	Topological/topochemical descriptors	From equations (5),(15) and (16) of reference 2
HallKierAlpha	Topological/topochemical descriptors	The Hall-Kier alpha value for a molecule
Ipc	Topological/topochemical descriptors	The information content of the coefficients of the characteristic polynomial of the adjacency matrix of a hydrogen-suppressed graph of a molecule
Kappa1	Topological/topochemical descriptors	Hall-Kier Kappa1 value
Kappa2	Topological/topochemical descriptors	Hall-Kier Kappa2 value
Kappa3	Topological/topochemical descriptors	Hall-Kier Kappa3 value
LabuteASA	MOE-like approximate molecular surface area descriptors	Labute's Approximate Surface Area (ASA from MOE)
HeavyAtomCount	Lipinski parameters for molecules	States the number of heavy atoms a molecule

MolLogP	Atom-based calculation of LogP and MR using Crippen's approach	Wildman-Crippen LogP value
MolMR	Atom-based calculation of LogP and MR using Crippen's approach	Wildman-Crippen MR value

E. Solubility Dataset Splitting Strategy

Firstly, we calculated the Tanimoto similarity scores between the solubility dataset and oral bioavailability test dataset. The molecules were then arranged in order from the smallest to the largest according to the Tanimoto similarity scores. The first 5000 molecules were classified as low similarity, the 2501th molecule to 7500th molecules were classified as low similarity and 4845th molecule to the last molecule were classified as high similarity. Thus, creating 3 datasets of different similarity level. This is a similar method adopted from Farsi[3] and inspired from k-fold cross-validation methodology where overlapping train datasets are formed from splitting thus generating more permutation and hence more datasets for training purposes.

Figure S1. Splitting strategy for solubility dataset.



F. Prediction Performance of pre-trained models during five-fold cross-validation

Table S8. Prediction performance for oral bioavailability prediction during five-fold cross-validation comparing different pre-training epochs [a]

Number of pre-training epochs	Metrics	Data Similarity Level			
		Low (5000)	Mid (5000)	High (5000)	Mid (9844)
20	Log Loss	0.648±0.028	0.655±0.019	0.637±0.033	0.633±0.032
	Acc	0.613±0.069	0.620±0.059	0.632±0.069	0.622±0.056
	F1 Score	0.633±0.154	0.571±0.296	**0.682±0.109**	0.632±0.136
	AUC-ROC	0.602±0.059	0.559±0.085	**0.646±0.056**	0.648±0.056
40	Log Loss	0.642±0.025	0.644±0.032	**0.640±0.032**	0.625±0.043
	Acc	0.628±0.054	0.623±0.053	0.637±0.054	**0.642±0.070**
	F1 Score	**0.635±0.142**	0.564±0.281	0.679±0.102	**0.642±0.142**
	AUC-ROC	0.627±0.042	0.625±0.073	0.627±0.059	0.655±0.081
60	Log Loss	**0.632±0.037**	**0.641±0.032**	**0.640±0.027**	**0.636±0.044**
	Acc	**0.634±0.060**	**0.629±0.053**	**0.639±0.066**	0.636±0.074
	F1 Score	0.631±0.147	**0.651±0.128**	0.671±0.118	0.636±0.150
	AUC-ROC	**0.662±0.061**	**0.651±0.045**	**0.646±0.051**	**0.662±0.071**

[a] Prediction performance for oral bioavailability train dataset during five-fold cross-validation reported in mean ± standard deviation. Models were pre-trained with different number of epochs (20, 40, 60). Bold value represents the best score across different epoch level.

G. Prediction Performance of Transfer Learning Model across different similarity levels

Table S9. Prediction Performance of Transfer Learning Model across different similarity levels [b]

Similarity (Size)	Low (5000)	Mid (5000)	High (5000)	Mid (9844)
Log Loss	0.588±0.066	0.532±0.034	0.531±0.033	**0.520±0.042**
Acc	0.718±0.053	0.729±0.025	**0.755±0.022**	0.737±0.020
F1 Score	0.760±0.066	0.789±0.016	**0.809±0.016**	0.782±0.026
AUC-ROC	0.746±0.041	0.799±0.027	**0.801±0.024**	0.795±0.044

[b] Vertical GNN models were pre-trained with solubility dataset of different similarity level for 60 epochs. Prediction performance using oral bioavailability test dataset were reported in mean ± standard deviation. Bold values represent the best score across different similarity level.

H. SHAP Analysis of Random Forest Models

Figure S2. Beeswarm plot of top 20 important molecular descriptors for Random Forest model towards oral bioavailability prediction using oral bioavailability test dataset. Analysis done on model developed from the second fold dataset produced using five-fold cross-validation.

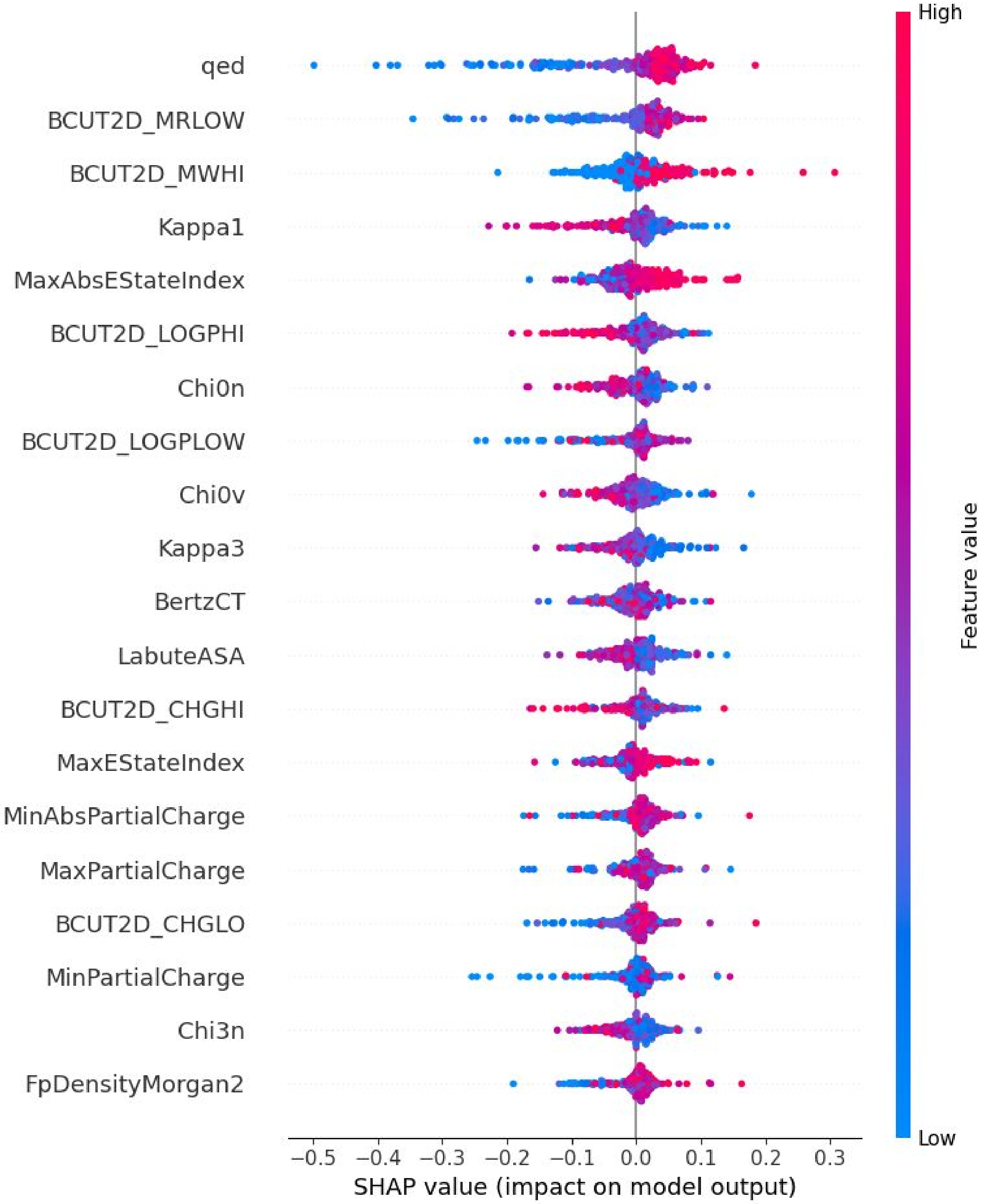


Figure S3. Beeswarm plot of top 20 important molecular descriptors for Random Forest model towards oral bioavailability prediction using oral bioavailability test dataset. Analysis done on model developed from the third fold dataset produced using five-fold cross-validation.

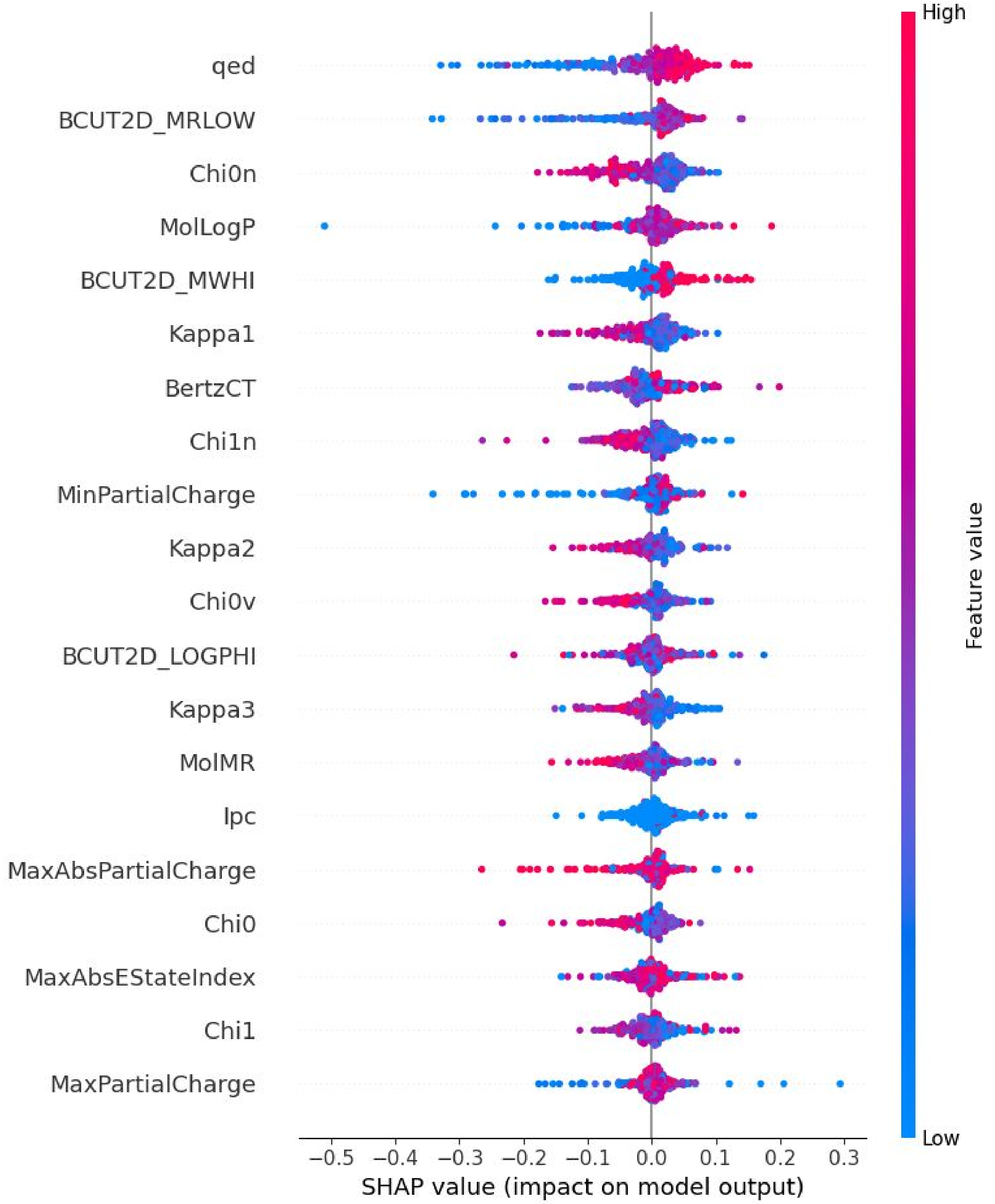


Figure S4. Beeswarm plot of top 20 important molecular descriptors for Random Forest model towards oral bioavailability prediction using oral bioavailability test dataset. Analysis done on model developed from the fourth fold dataset produced using five-fold cross-validation.

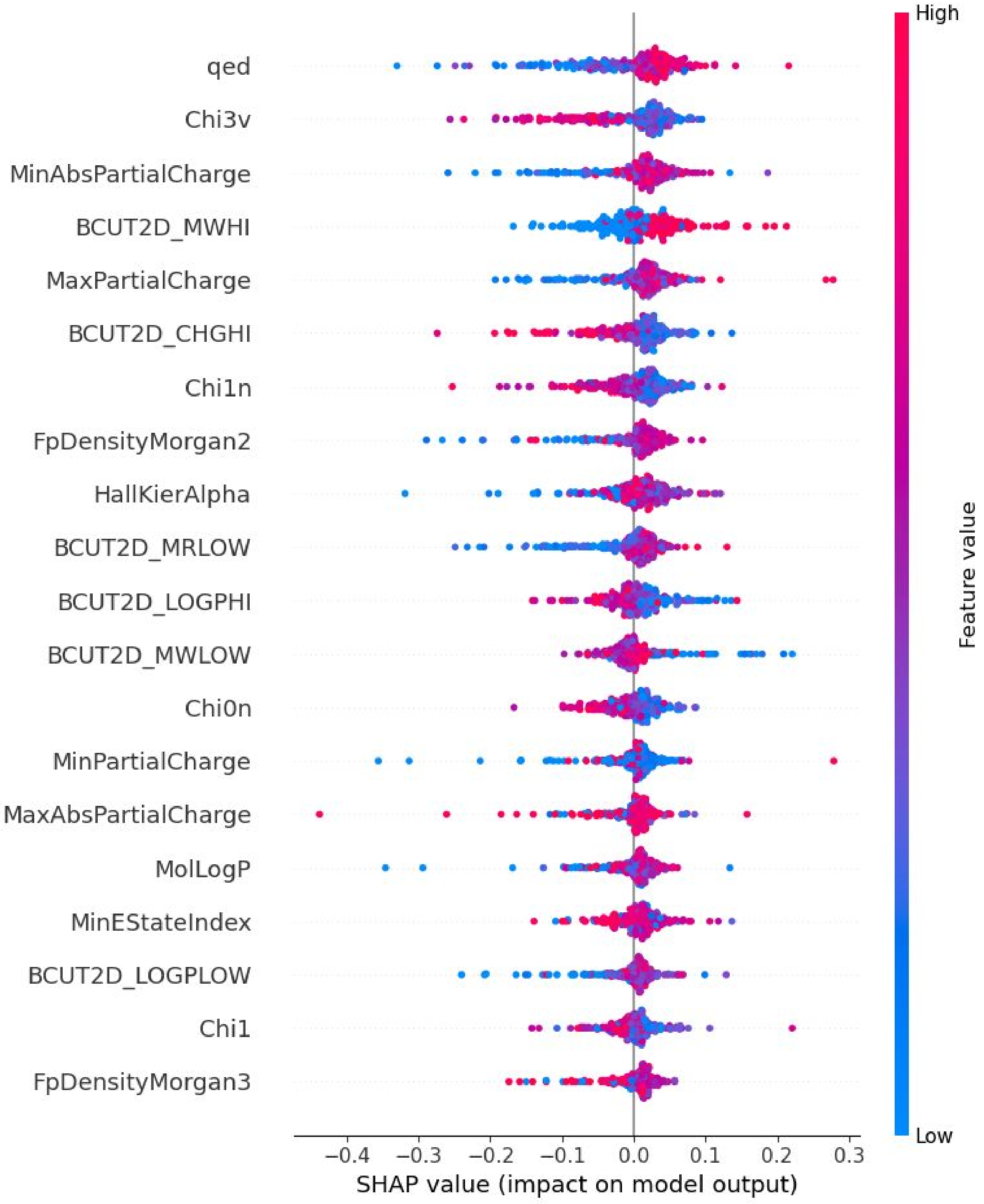


Figure S5. Beeswarm plot of top 20 important molecular descriptors for Random Forest model towards oral bioavailability prediction using oral bioavailability test dataset. Analysis done on model developed from the fifth fold dataset produced using five-fold cross-validation.

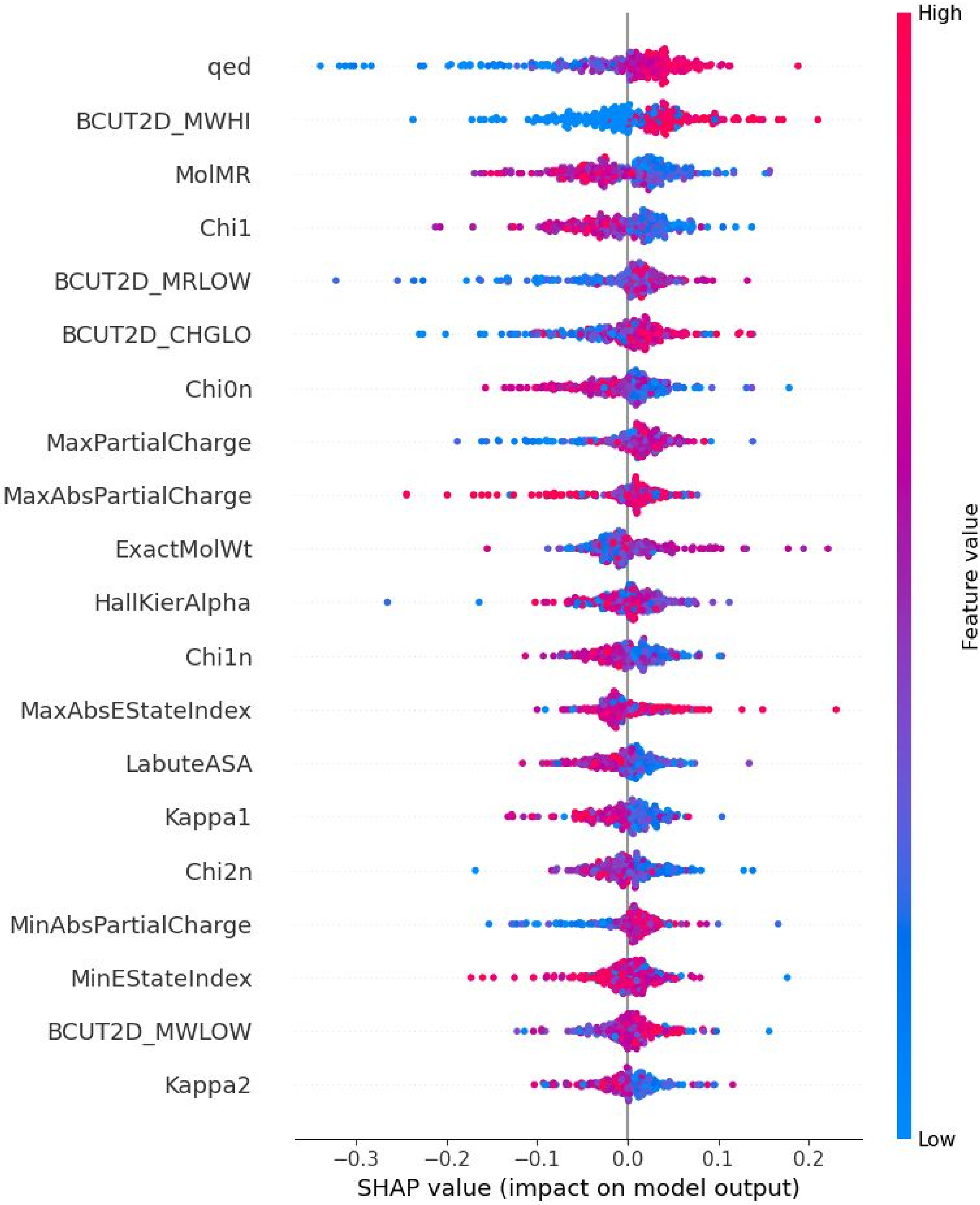


Figure S6. Global feature importance bar plot highlighting top 10 most important molecular descriptors. Absolute mean for that feature is taken over all the given sample. Analysis done on model developed from the first fold dataset produced using five-fold cross-validation.

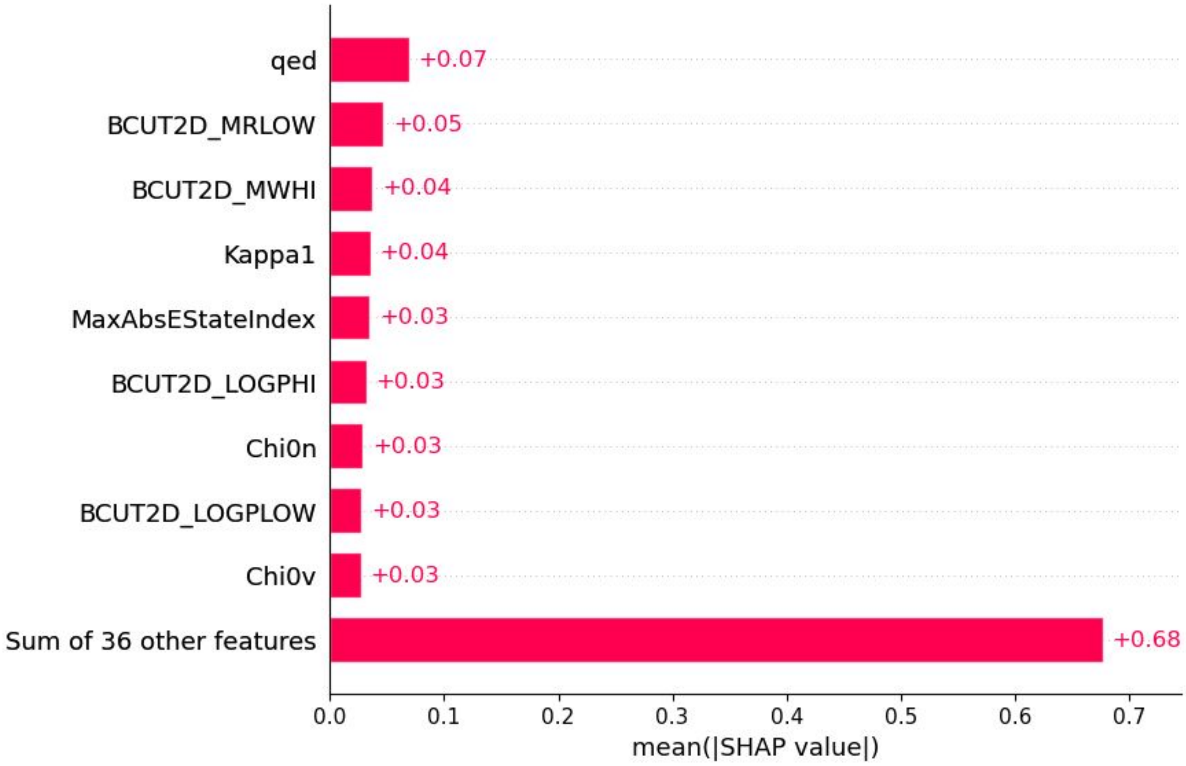


Figure S7. Global feature importance bar plot highlighting top 10 most important molecular descriptors. Absolute mean for that feature is taken over all the given sample. Analysis done on model developed from the second fold dataset produced using five-fold cross-validation.

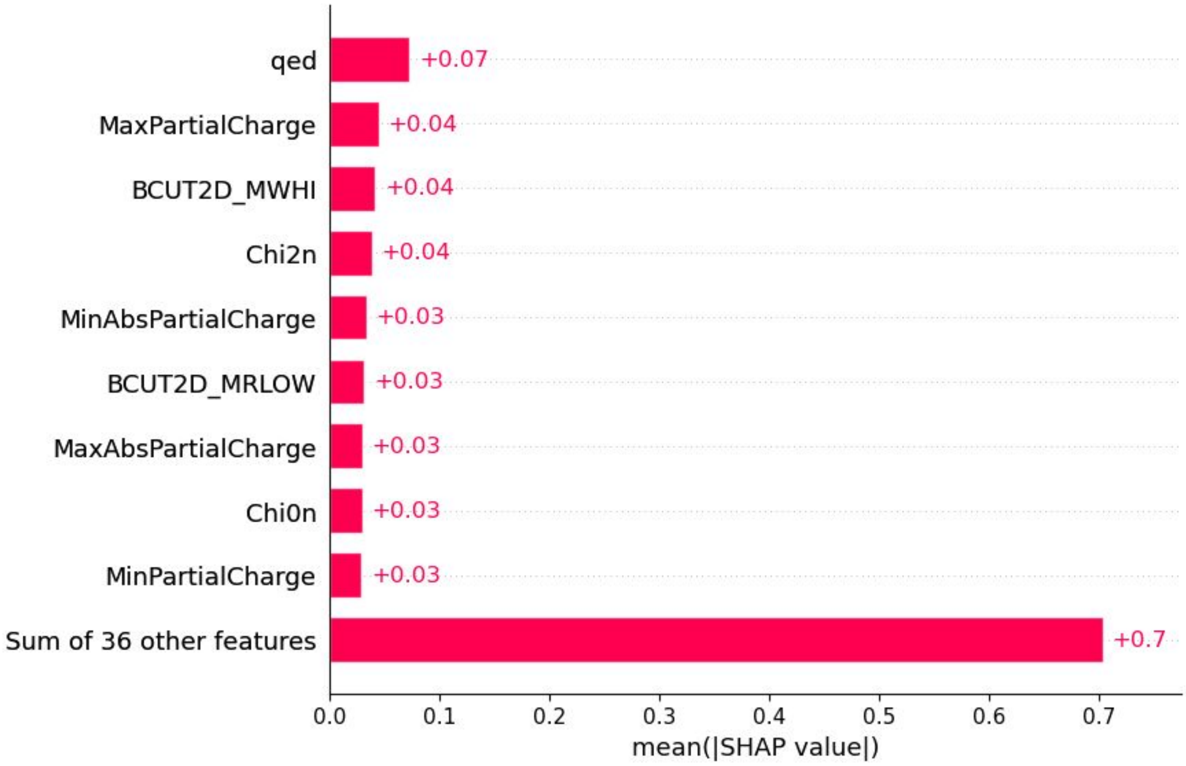


Figure S8. Global feature importance bar plot highlighting top 10 most important molecular descriptors. Absolute mean for that feature is taken over all the given sample Analysis done on model developed from the third fold dataset produced using five-fold cross-validation.

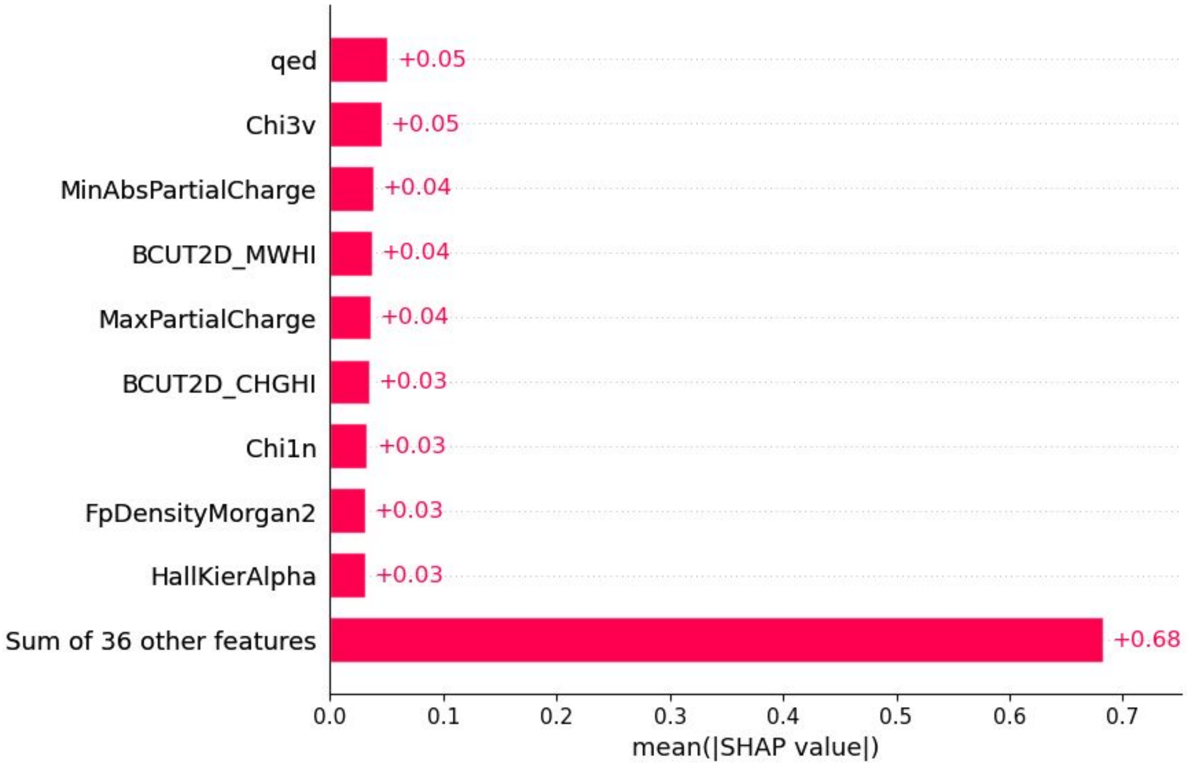


Figure S9. Global feature importance bar plot highlighting top 10 most important molecular descriptors. Absolute mean for that feature is taken over all the given sample. Analysis done on model developed from the fourth fold dataset produced using five-fold cross-validation.

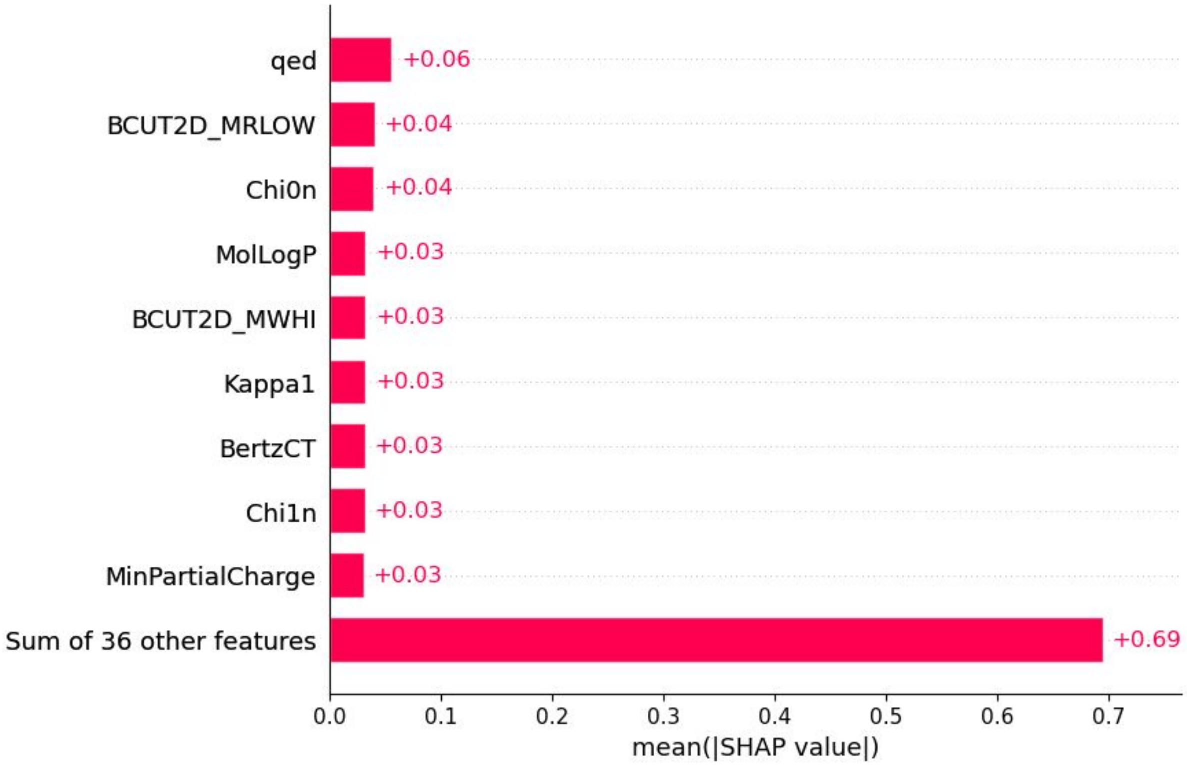
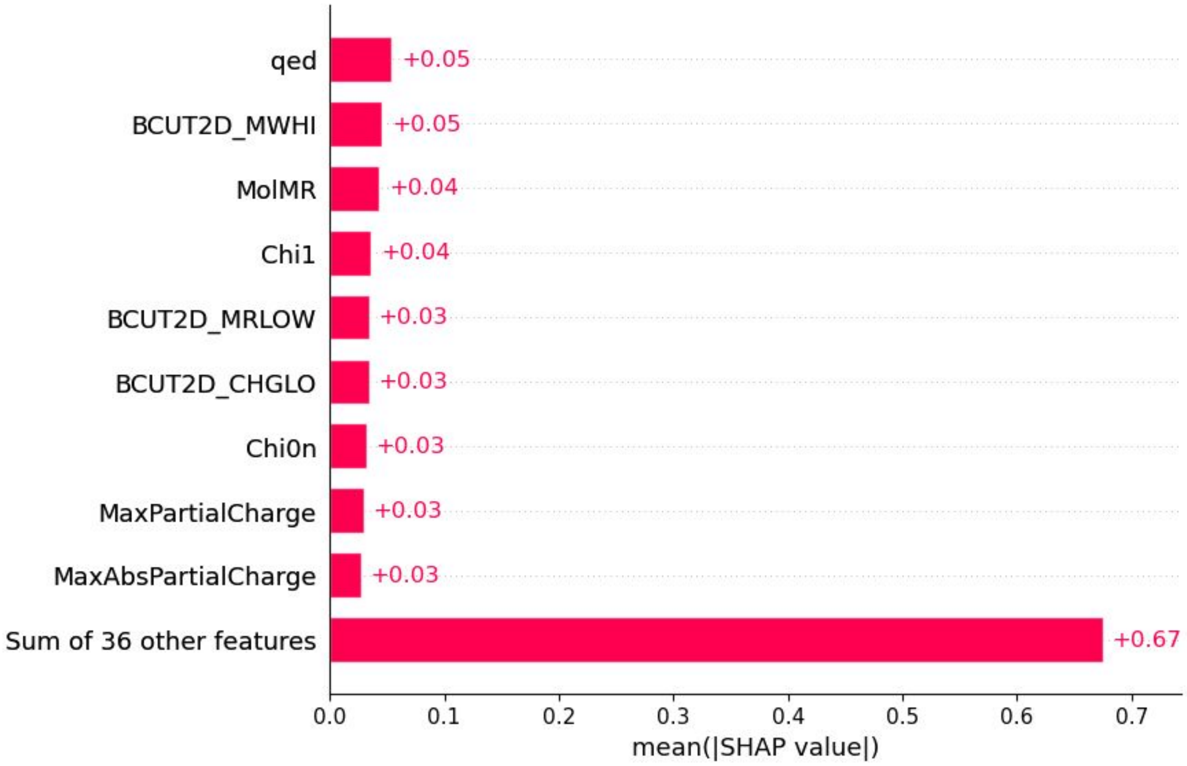


Figure S10. Global feature importance bar plot highlighting top 10 most important molecular descriptors. Absolute mean for that feature is taken over all the given sample. Analysis done on model developed from the fifth fold dataset produced using five-fold cross-validation.



References

1. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-Generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; ACM: New York, NY, USA, 2019.
2. Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. *Reviews in Computational Chemistry.* **2007**, 367–422.
3. Farsi, M. Application of Ensemble RNN Deep Neural Network to the Fall Detection through IoT Environment. *Alex. Eng. J.* **2021**, *60*, 199–211.