



LigandNet: A Suite of Ligand Based Predictive Models for Activity and Property Prediction

Suman Sirimulla



@ssirimulla

Dept. of Pharmaceutical Sciences

School of Pharmacy

The University of Texas at El Paso

Classical SFs

$$p = \sum_{m=1}^M w_m x_m$$

DOCK (force-field SF):

$$E_{\text{bind}} \equiv \sum_{k=1}^K \sum_{l=1}^L \left(\frac{A_{kl}}{d_{kl}^{12}} - \frac{B_{kl}}{d_{kl}^6} \right) + \sum_{k=1}^K \sum_{l=1}^L \left(332 \frac{q_k q_l}{\epsilon(d_{kl}) d_{kl}} \right)$$

PMF (knowledge-based SF):

$$\text{PMF} \equiv \sum_i \sum_j \sum_{k=1}^{K_j} \sum_{l=1}^{L_i} H_{kl}(d_{kl})$$

X-Score (empirical SF):

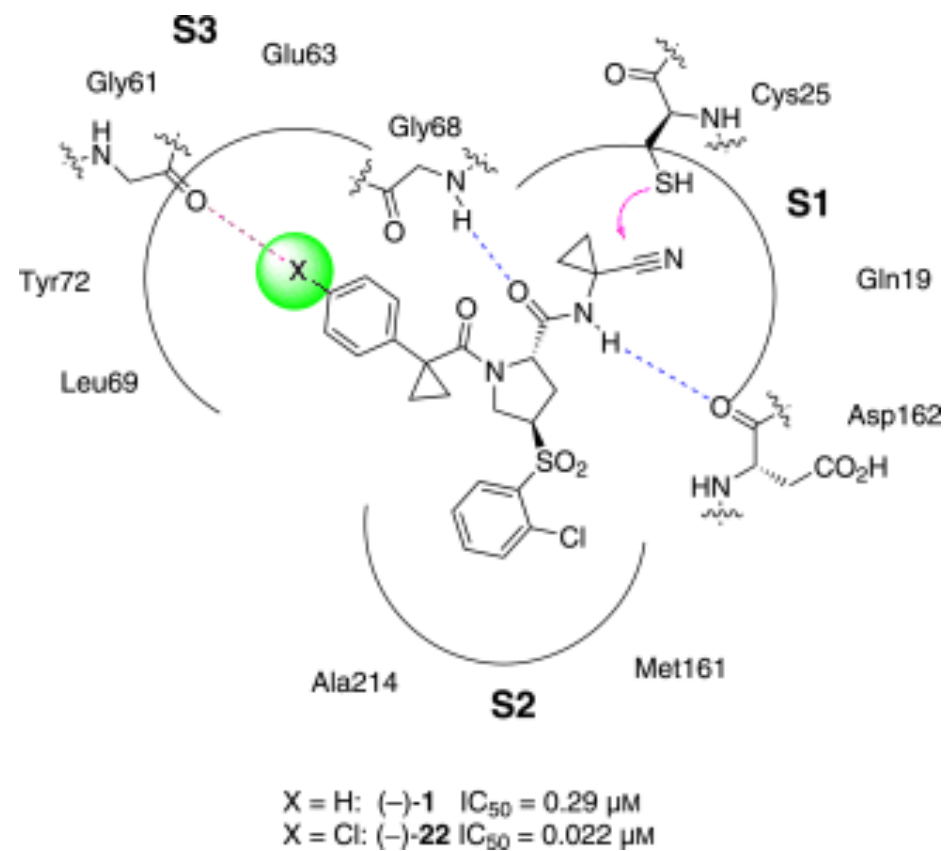
$$\Delta G_{\text{bind}} \equiv w_0 + w_1 \Delta G_{\text{vdW}} + w_2 \Delta G_{\text{HBonds}} + w_3 \Delta G_{\text{rotor}} + w_4 \Delta G_{\text{hydrophobic}}$$

Machine-learning SFs

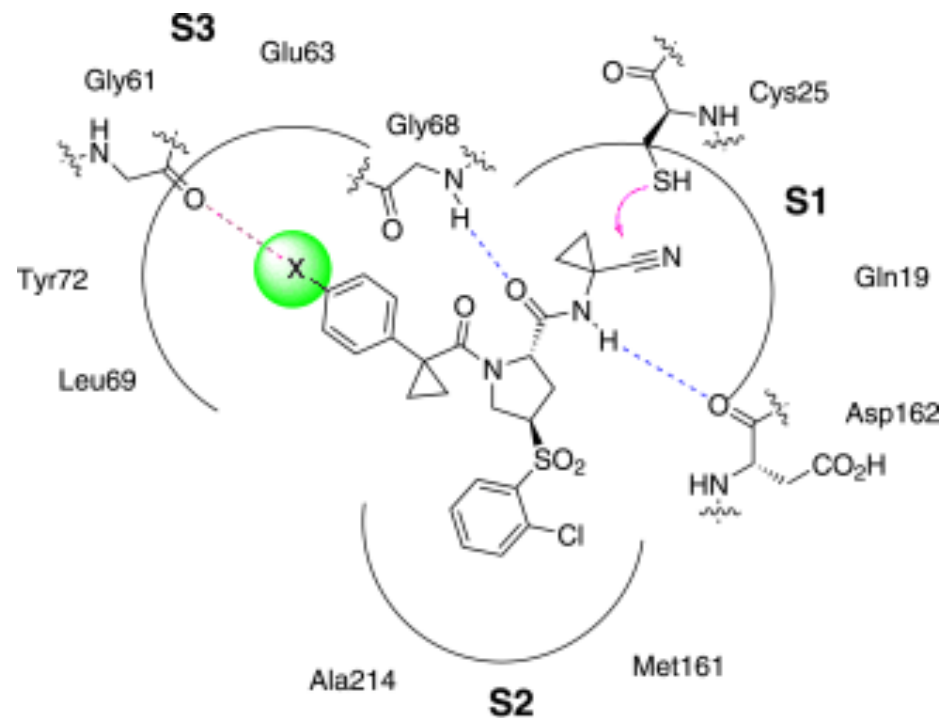
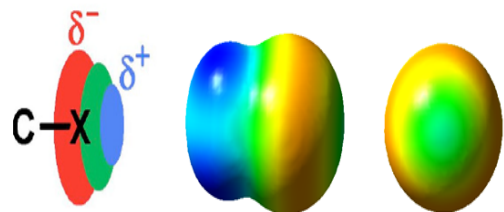
$$p = f_{RF}(x_m)$$

$$x_{ij} \equiv \sum_{k=1}^{K_j} \sum_{l=1}^{L_i} \Theta(d_{\text{cutoff}} - d_{kl})$$

Systematic Investigation of Halogen Bonding in Protein–Ligand Interactions



Systematic Investigation of Halogen Bonding in Protein–Ligand Interactions



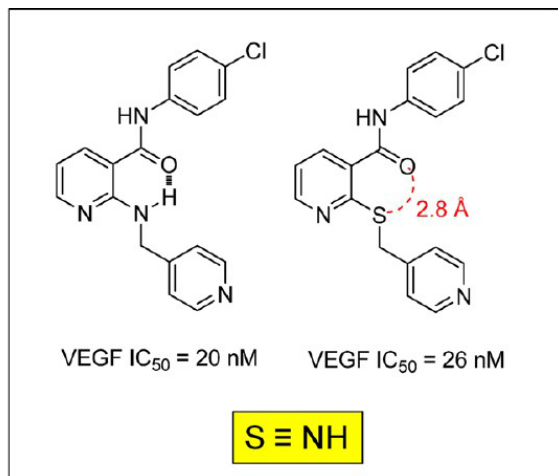
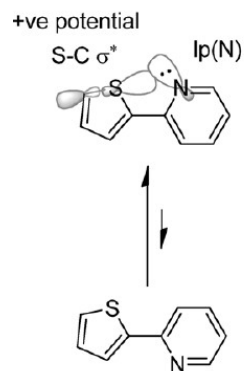
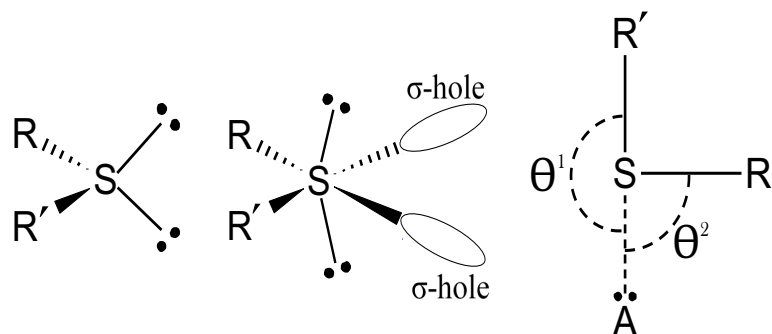
X = H: (–)-**1** IC₅₀ = 0.29 μM
 X = Cl: (–)-**22** IC₅₀ = 0.022 μM

$$\begin{aligned} \text{gauss}_1(d) &= w_{\text{gauss}_1} e^{-(d/0.5)^2} \\ \text{gauss}_2(d) &= w_{\text{gauss}_2} e^{-((d-3)/2)^2} \\ \text{repulsion}(d) &= \begin{cases} w_{\text{repulsion}} d^2 & d < 0 \\ 0 & d \geq 0 \end{cases} \end{aligned}$$

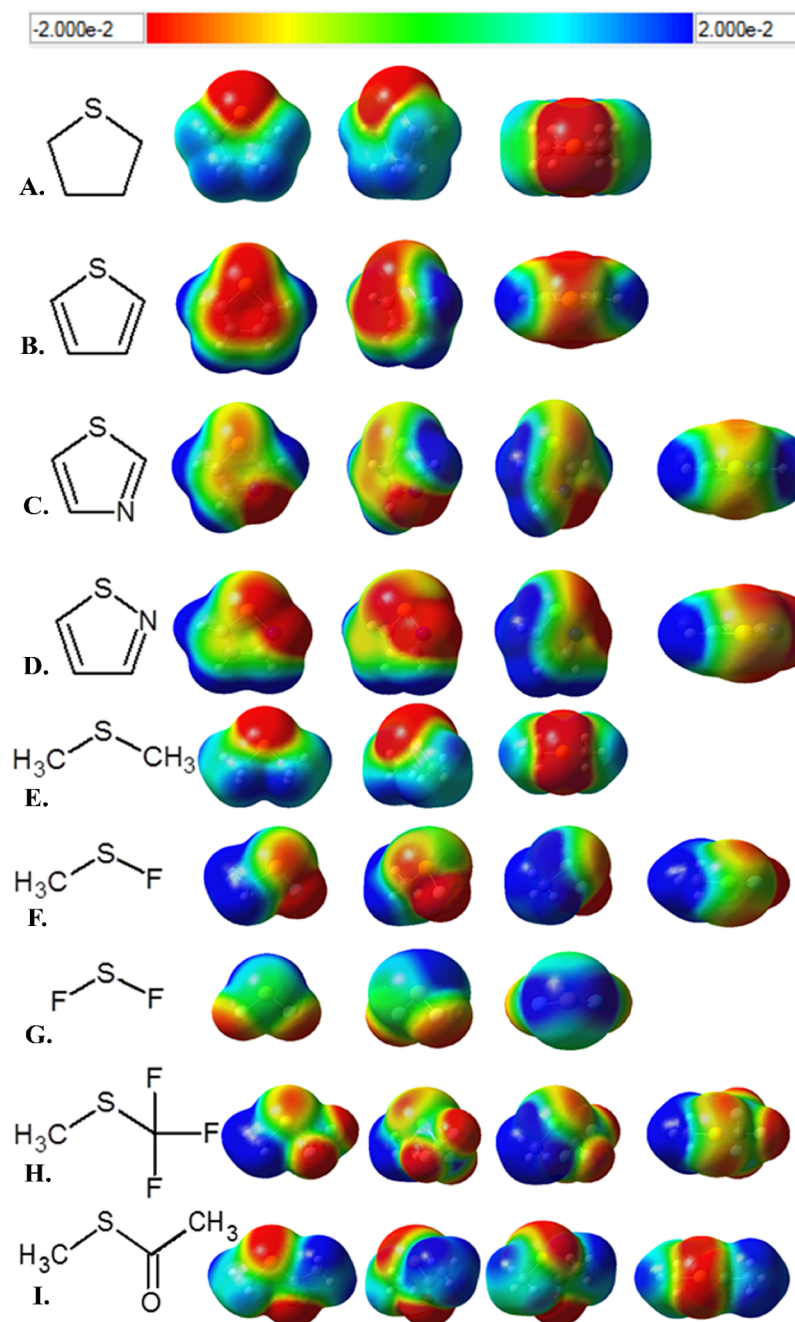
$$\text{hydrophobic}(d) = \begin{cases} w_{\text{hydrophobic}} & d < 0.5 \\ 0 & d > 1.5 \\ w_{\text{hydrophobic}}(1.5 - d) & \text{otherwise} \end{cases}$$

$$\text{hbond}(d) = \begin{cases} w_{\text{hbond}} & d < -0.7 \\ 0 & d > 0 \\ w_{\text{hbond}}(-\frac{10}{7}d) & \text{otherwise} \end{cases}$$

Autodock Vina
 Scoring function



J. Med. Chem., **2015**, 58 (11), pp 4383–4438



Sigma hole scoring function (SHSF)

- SHSF scoring function is represented as follows:

$$E = W\Phi D$$

- W is the weight
- Cl=0.265, Br=0.320, I= 0.400, S=0.320
- Φ is the angle factor
- D is the distance factor

Angle factor (Φ)

Halogen

$$Z_{X,\Theta} = A \cos \alpha + B$$

$$\alpha = 180^\circ - \Theta$$

$Z_{X,\Theta}$ is the effective charge on the halogen at Θ angle.

Sulfur

1 when $\geq 160^\circ$
 135° - 160° is linearly interpolated
 0 when $< 135^\circ$

Distance factor (D)

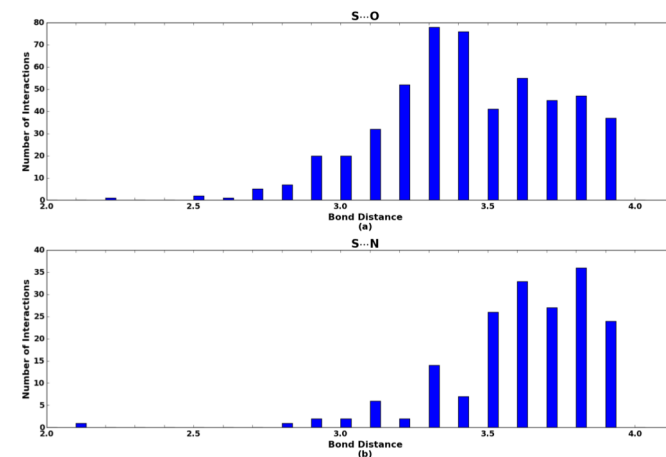
Halogen

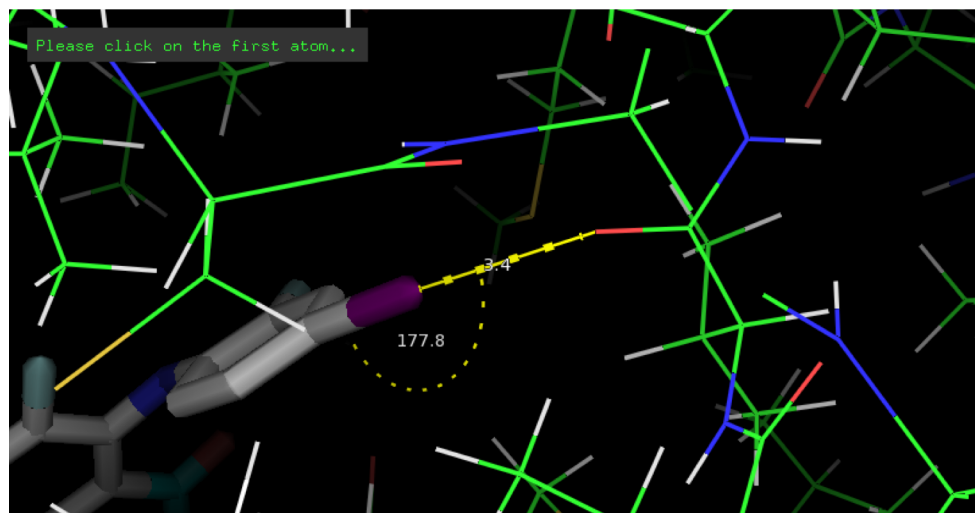
1 when $d < -\delta \text{ \AA}$
 Linearly interpolated between
 0 when $d > 0 \text{ \AA}$

Sulfur

1 when $d < \text{sum} - 0.2 \text{ \AA}$
 Linearly interpolated between
 0 when $d > 0 \text{ \AA}$

Aromatic Sulfur---Acceptor distance data





```

Detected 8 CPUs
Reading input ... done.
Setting up the scoring function ... done.
Affinity: -8.23605 (kcal/mol)
Intermolecular contributions to the terms, before weighting:
  gauss 1      : 86.61536
  gauss 2      : 1296.41056
  repulsion    : 1.31628
  hydrophobic  : 45.60111
  Hydrogen     : 0.79994
  Halogen (Cl): 0.00000
  Halogen (Br): 0.00000
  Halogen (I) : 0.99014

```

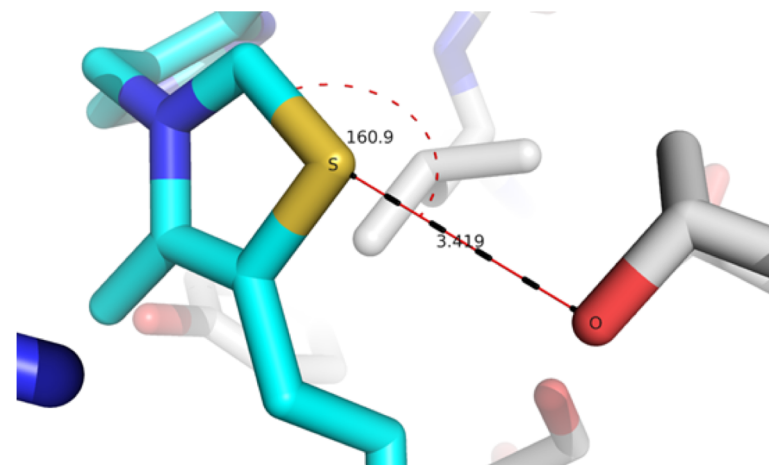
Angle- 177.8

Distance- 3.4 Ang

Overlap δ = -0.545 Ang

Angle term Φ = 0.998

Distance term D = 0.992

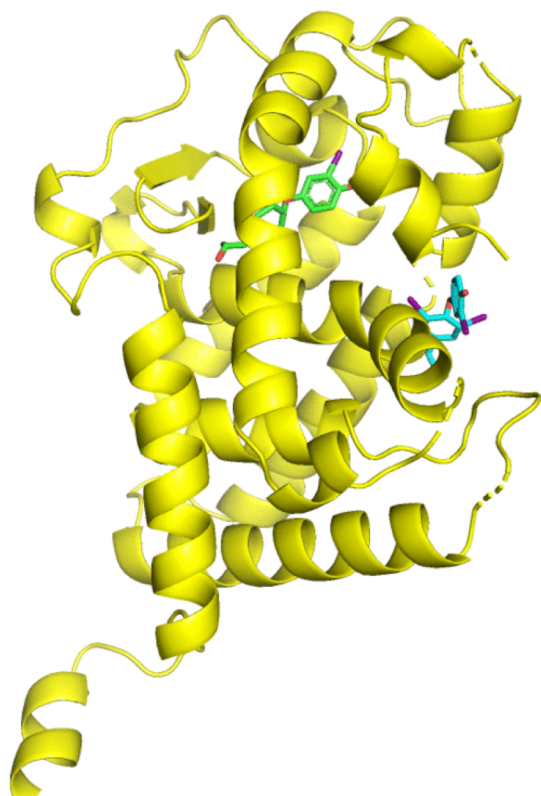


```

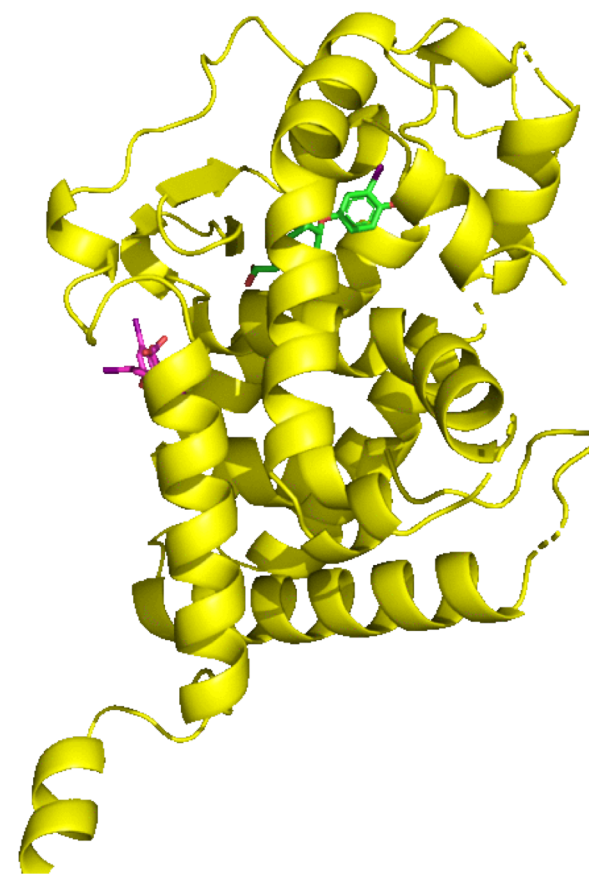
Detected 8 CPUs
Reading input ... done.
Setting up the scoring function ... done.
Affinity: -7.96214 (kcal/mol)
Intermolecular contributions to the terms, before weighting:
  gauss 1      : 115.94439
  gauss 2      : 1225.20303
  repulsion    : 5.79224
  hydrophobic  : 21.83337
  Hydrogen     : 8.28174
  Halogen (Cl): 0.00000
  Halogen (Br): 0.00000
  Halogen (I) : 0.00000
  Sulfur       : 1.00000

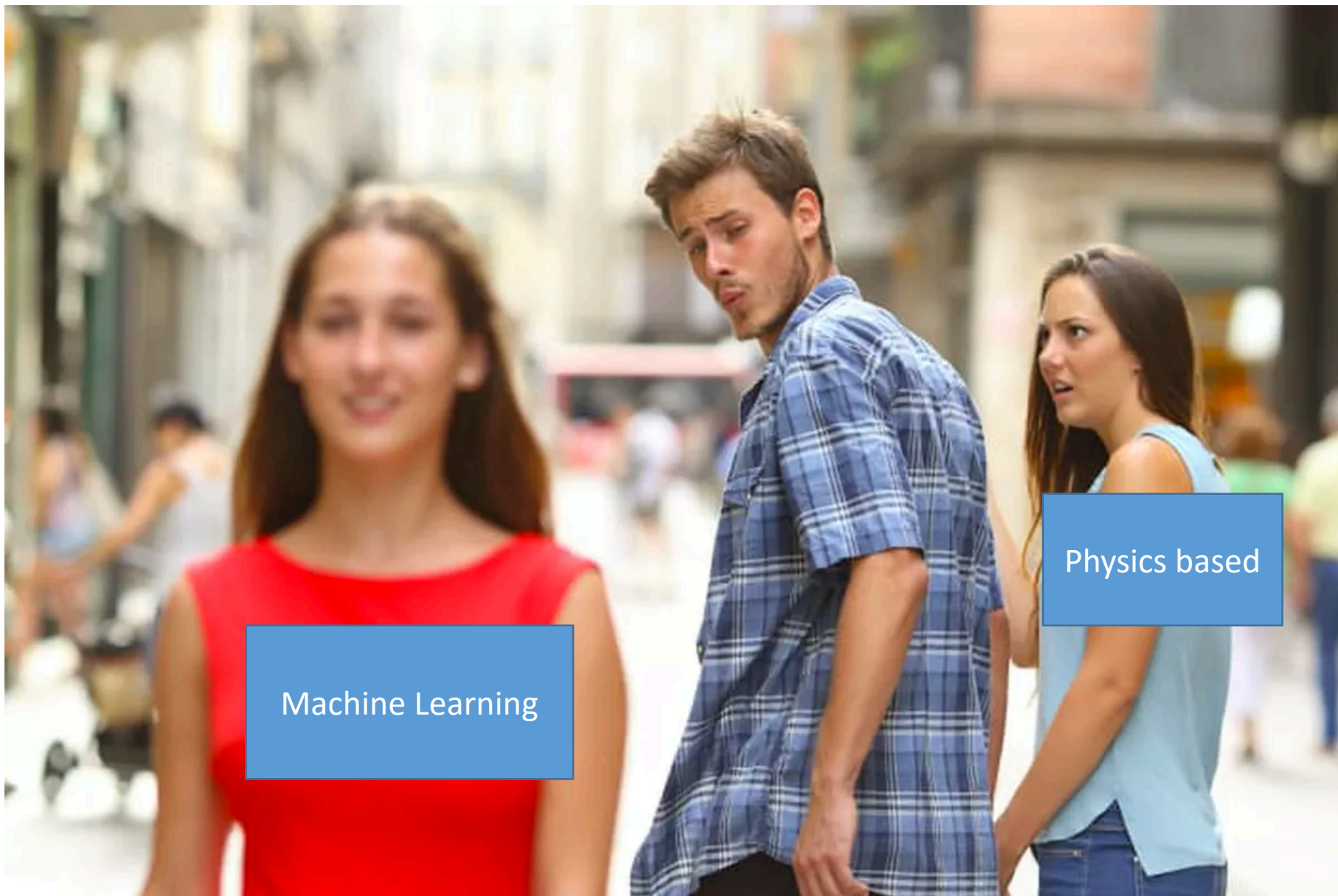
```

VinaXB



Vina





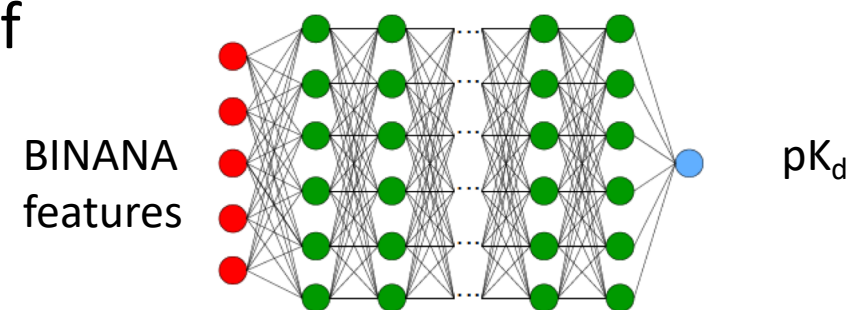
Machine Learning

Physics based

DLSCORE

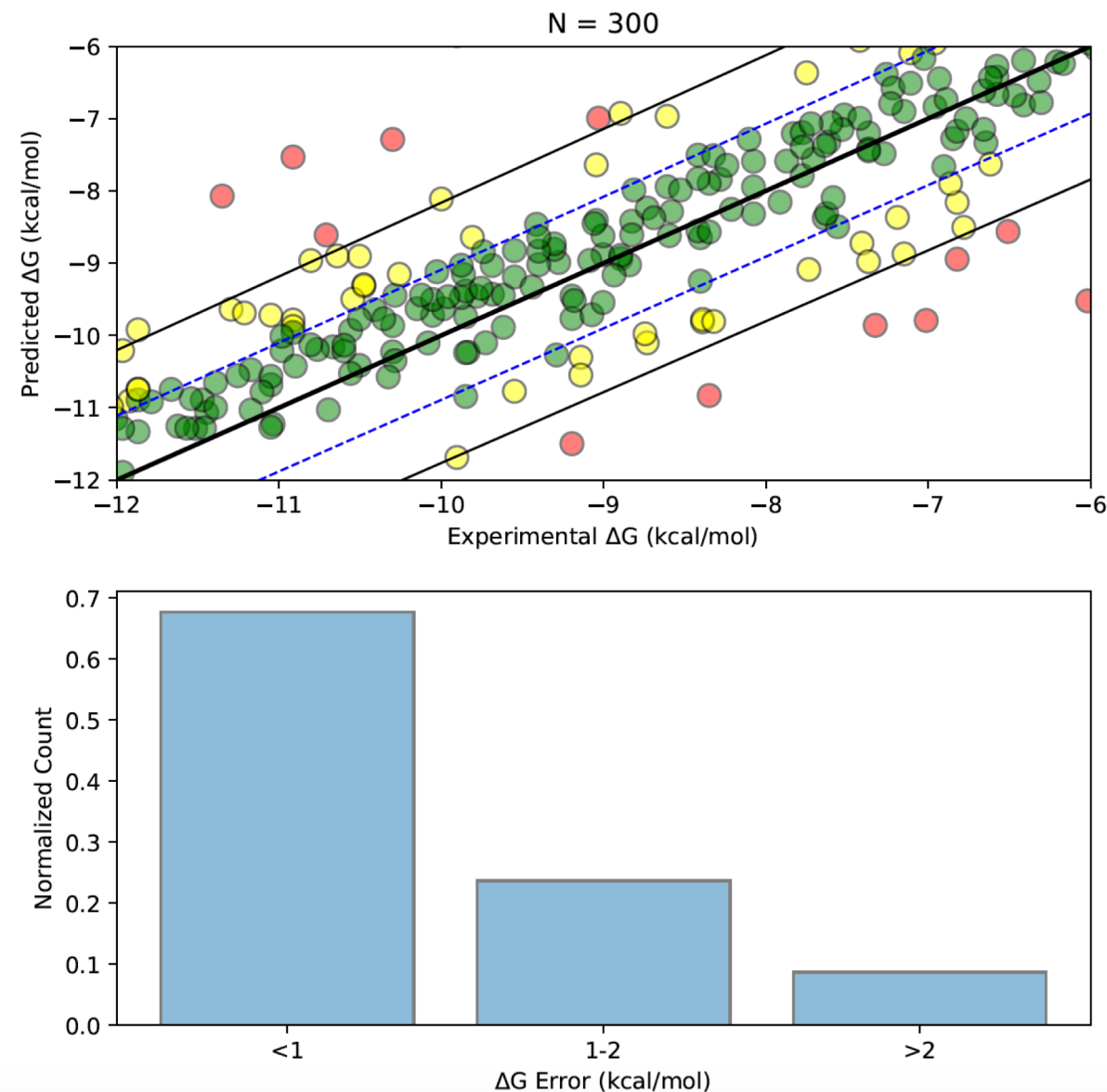
Data set and Input layers

- We used 'refined-set 2017' (4057 protein-ligand complexes) from PDBbind for the study.
- We calculate BINding ANALyzer (BINANA) features using NNSCORE2.0. There were total of 348 features.
- These features represent **Close Contacts, Electrostatic Interactions, Binding-Pocket Flexibility, Hydrophobic Contacts, Hydrogen Bonds, Salt Bridges, π Interactions**
- To generate descriptors, we first converted pdb files to pdbqt (both proteins and ligands) by using MGL tools.



DL Performance (using 10 networks)

Parameter	Result
RMSE	1.15 kcal/mol
MAE	0.86 kcal/mol
Max possible correlation	0.98
Pearson R	0.82 95%C.I [0.78,0.84]
Spearman Rho	0.90
Kendall tau	0.74



Hassan, Mahmudulla; Mogollon, Daniel Castaneda; Fuentes, Olac; sirimulla, suman (2018): DLSCORE: A Deep Learning Model for Predicting Protein-Ligand Binding Affinities. ChemRxiv. Preprint. <https://doi.org/10.26434/chemrxiv.6159143.v1>



Tweet



suman sirimulla @ssirimulla · Mar 7, 2018



Brain Goldman @VertexPharma :I read google created an AI that beats the world champoion in go. Why can't we write an AI program to design a drug? #OECUP @OpenEyeSoftware



1



1



4



↳ You Retweeted



Olexandr Isayev

@olexandr



Replying to @ssirimulla @jchodera and 2 others

We do;) Paper currently in revision.

arxiv.org/abs/1711.10907

10:50 AM · Mar 7, 2018 · [Twitter for iPhone](#)

1 Retweet 2 Likes



suman sirimulla @ssirimulla · Mar 7, 2018



Replying to @olexandr @jchodera and 2 others

also wanna tag this other paper arxiv.org/pdf/1704.07555...



COMPUTATIONAL BIOLOGY

Deep reinforcement learning for de novo drug design

Mariya Popova^{1,2,3}, Olexandr Isayev^{1*}, Alexander Tropsha^{1*}

Olivecrona et al. *J Cheminform* (2017) 9:48
DOI 10.1186/s13321-017-0235-x

 Journal of Cheminformatics

Cc
Th
rig
ex
Ar
foi
of

RESEARCH ARTICLE

Open Access



Molecular de-novo design through deep reinforcement learning

Marcus Olivecrona^{*}, Thomas Blaschke, Ola Engkvist and Hongming Chen

ACS
central
science

Cite This: *ACS Cent. Sci.* 2018, 4, 268–276

Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

Rafael Gómez-Bombarelli,^{†,‡,§} Jennifer N. Wei,^{‡,§} David Duvenaud,^{‡,§} José Miguel Hernández-Lobato,^{‡,§} Benjamín Sánchez-Lengeling,[‡] Dennis Sheberla,[‡] Jorge Aguilera-Iparraguirre,[‡] Timothy D. Hirzel,[‡] Ryan P. Adams,^{‡,||} and Alán Aspuru-Guzik^{*,‡,||}

BRIEF COMMUNICATION

<https://doi.org/10.1038/s41587-019-0224-x>

nature
biotechnology

Deep learning enables rapid identification of potent DDR1 kinase inhibitors

Alex Zhavoronkov^{1*}, Yan A. Ivanenkov¹, Alex Aliper¹, Mark S. Veselov¹, Vladimir A. Aladinskiy¹, Anastasiya V. Aladinskaya¹, Victor A. Terentiev¹, Daniil A. Polykovskiy¹, Maksim D. Kuznetsov¹, Arip Asadulaev¹, Yury Volkov¹, Artem Zholus¹, Rim R. Shayakhmetov¹, Alexander Zhebrak¹, Lidiya I. Minaeva¹, Bogdan A. Zagribelnyy¹, Lennart H. Lee², Richard Soll², David Madge², Li Xing², Tao Guo² and Alán Aspuru-Guzik^{3,4,5,6}

Cell

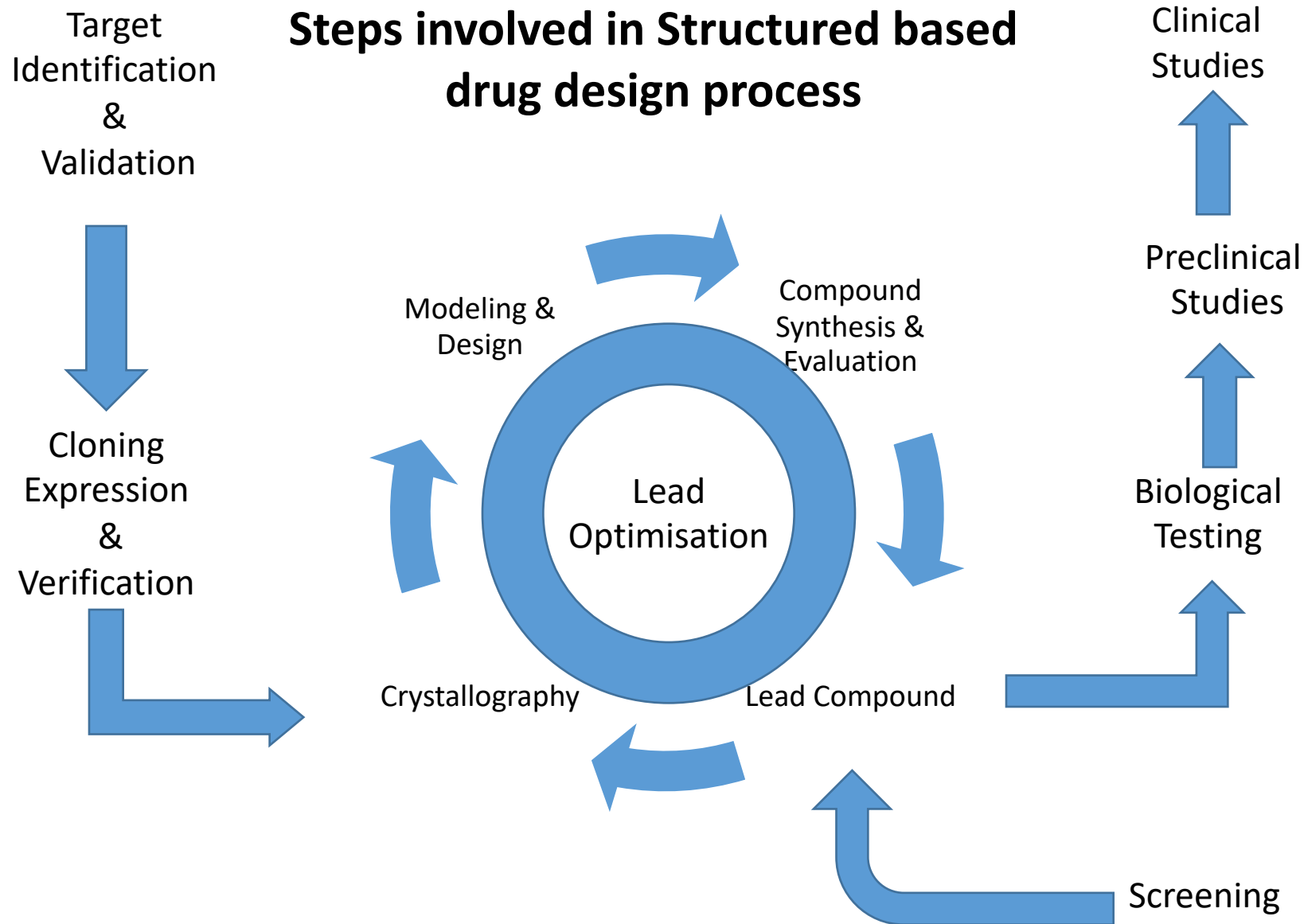
Article

A Deep Learning Approach to Antibiotic Discovery

Jonathan M. Stokes^{1,2,3}, Kevin Yang^{3,4,10}, Kyle Swanson^{3,4,10}, Wengong Jin^{3,4}, Andres Cubillos-Ruiz^{1,2,5}, Nina M. Donghia^{1,5}, Craig R. MacNair⁶, Shawn French⁶, Lindsey A. Carfrae⁶, Zohar Bloom-Ackerman^{2,7}, Victoria M. Tran², Anush Chiappino-Pepe^{5,7}, Ahmed H. Badran², Ian W. Andrews^{1,2,5}, Emma J. Chory^{1,2}, George M. Church^{5,7,8}, Eric D. Brown⁶, Tommi S. Jaakkola^{3,4}, Regina Barzilay^{3,4,9,*} and James J. Collins^{1,2,5,8,9,11,*}



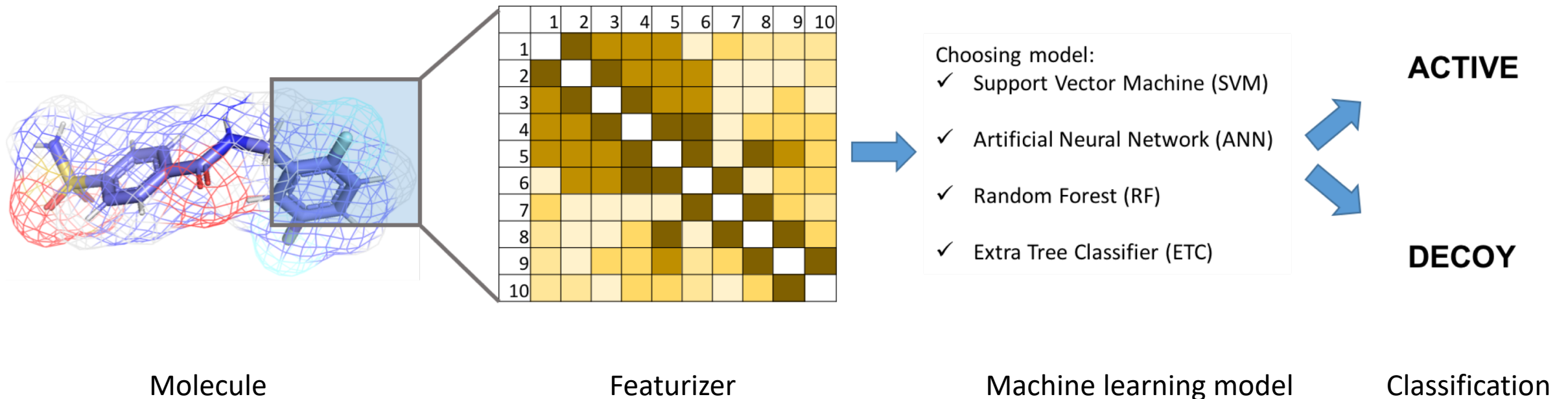
Research Article



Effort → to decrease the drug discovery timeline by reducing lead optimization iterative process using effective cheminformatic filters.

LigandNet

- A Machine Learning (ML) tool
 - Uses algorithms to predict active and decoy ligand with a specific human protein.



Pharos Database

Table 1 Summary of the protein classification, actives, decoys and machine learning models generated. The third column refers to the number of files with 20 or more active ligands found. The fourth indicates the amount of files with decoys less than 20. The last column refers to the number of machine learning models generated.

Protein type	<i>Total</i>	≥ 20 actives	≤ 20 decoys	<i>MLmodels</i>
GPCR	207	152	7	145
Nuclear Factor	38	23	0	23
Kinase	337	141	0	134
Enzyme	590	300	0	247
Ion Channel	154	59	0	57
Epigenetic	65	34	0	34
Transporter	72	30	0	27
Transcription Factor	22	6	0	5
TF and Epigenetic	7	2	0	2
Other	103	87	0	79
Total	1,595	834	7	753

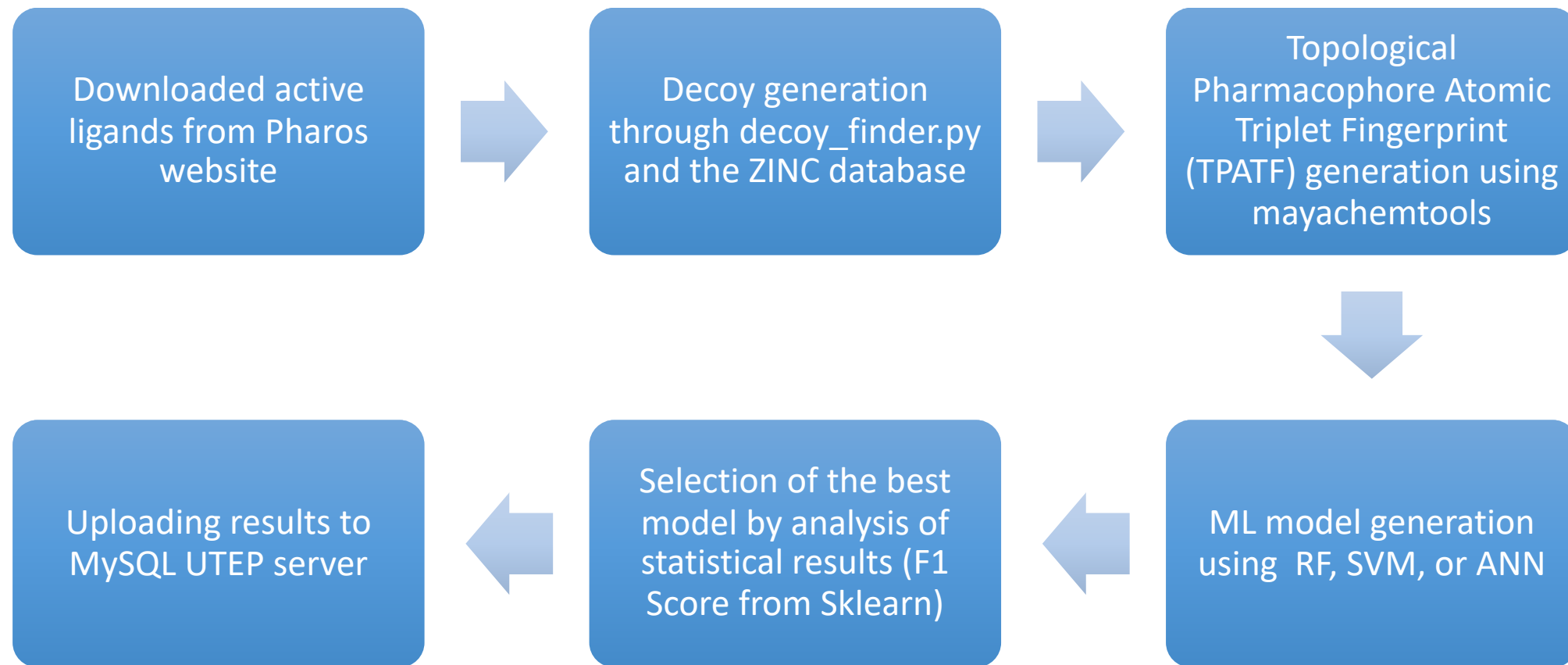


Tools used/required

- Anaconda
- Jupyter Notebook Platform
- Python 3.5 or above
- Python modules (including scikit-learn, rdkit, pandas, tqdm, os, pickle, numpy, pandas, shutil)
- Mayachemtools featurizer (TPATF fingerprints)

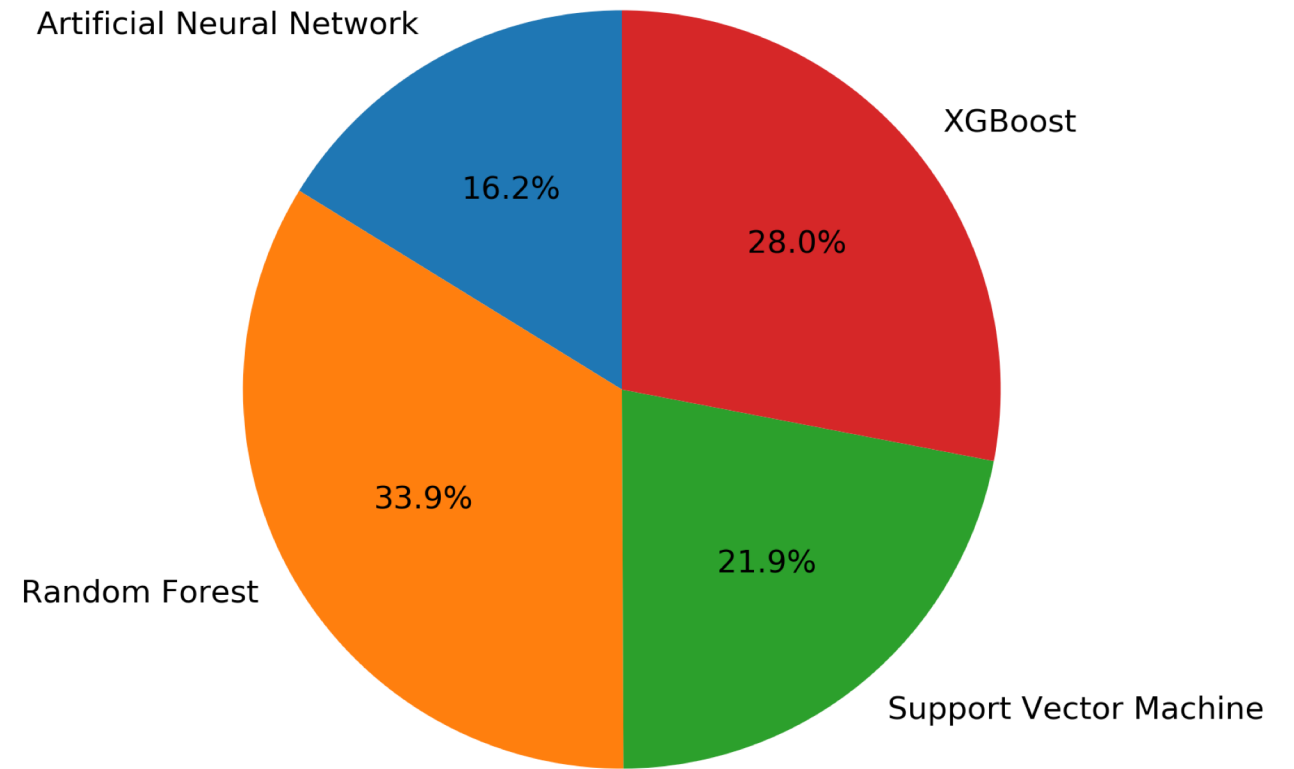


LigandNet (database)

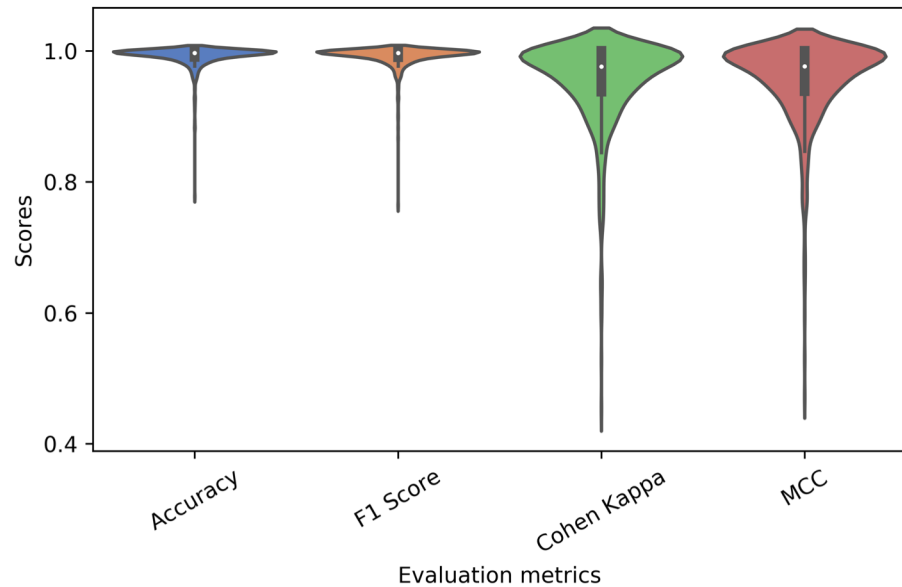


Machine Learning Models

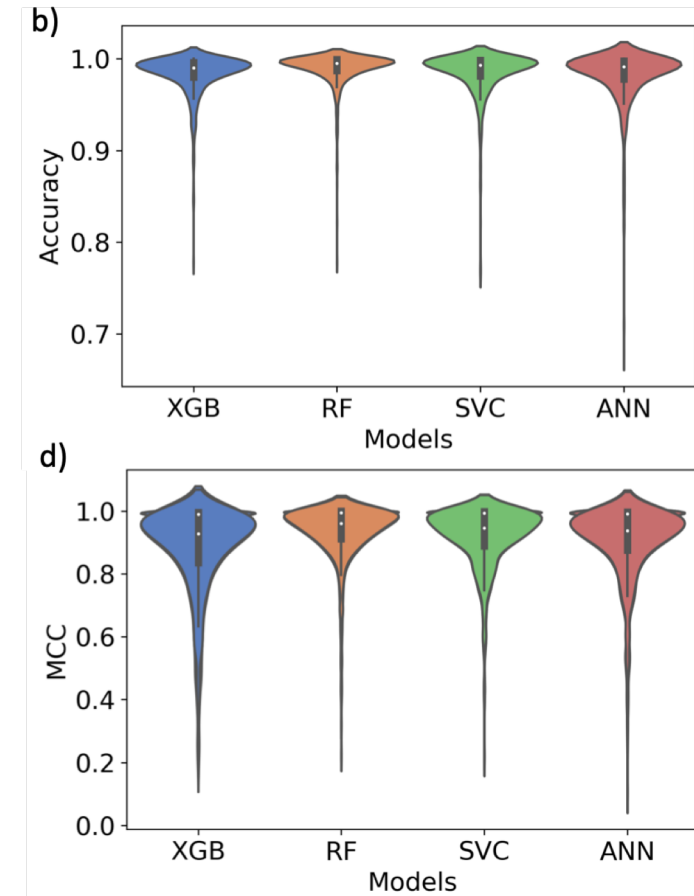
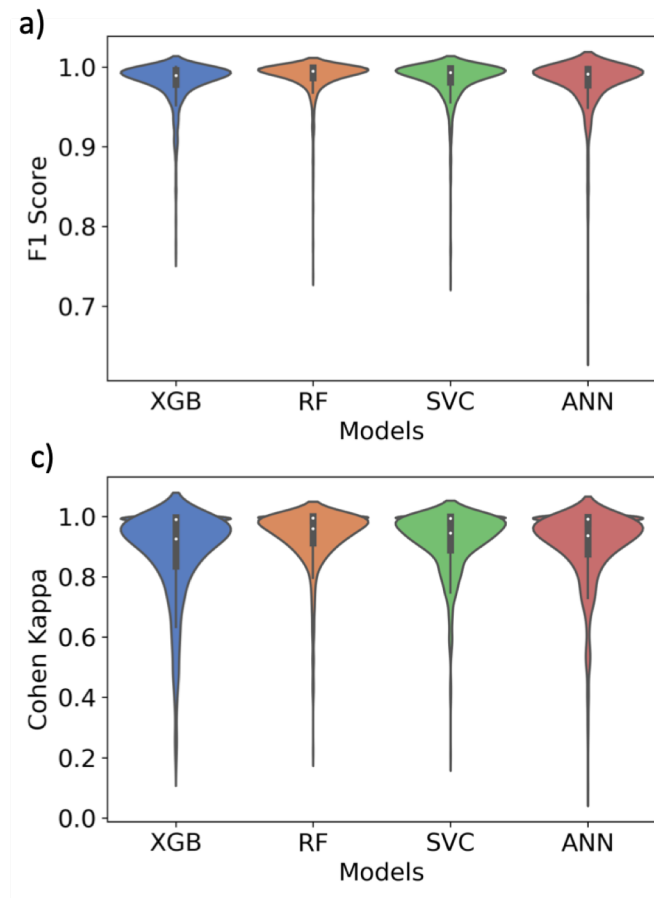
- XGBoost Classifier
- Random Forest Classifier
- MLP Classifier (Feed-forward network)
- Support Vector Classifier



Model evaluation



Scores evaluated by all of the models

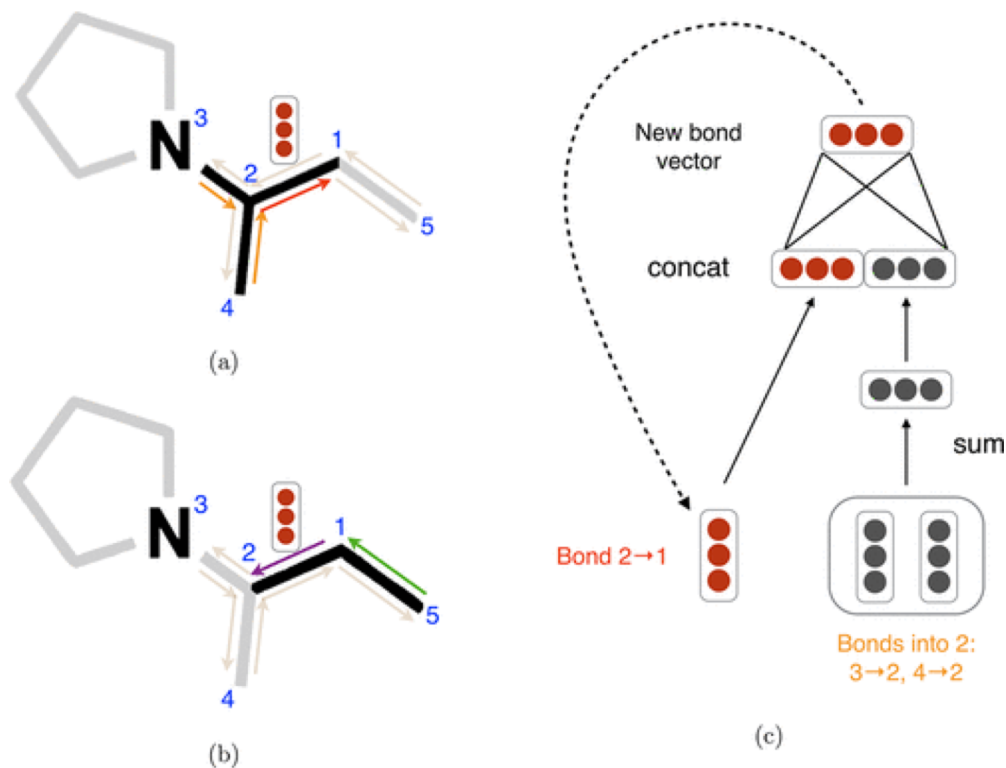
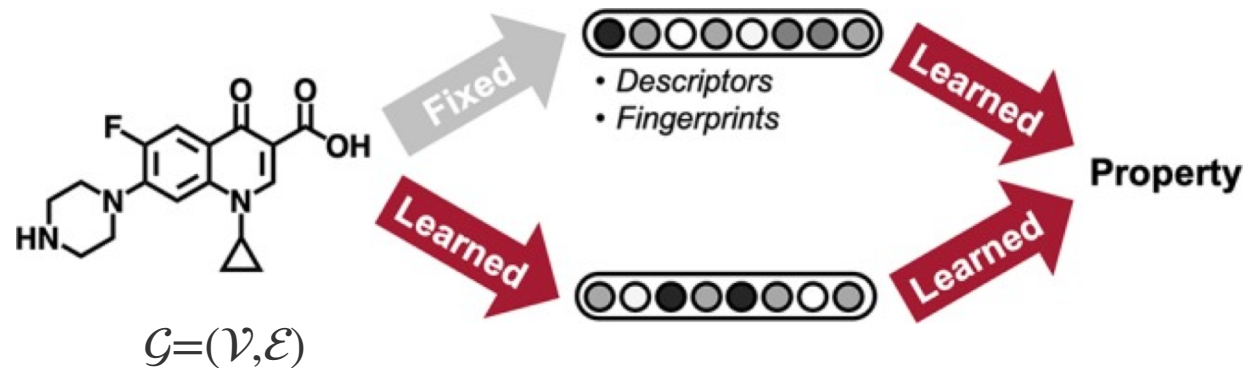


Performance of models against various evaluation metrics a) F1 score b) accuracy c) Cohen's kappa and d) MCC score



Graph
convolution
Neural Networks

Fingerprint/
Descriptor
based

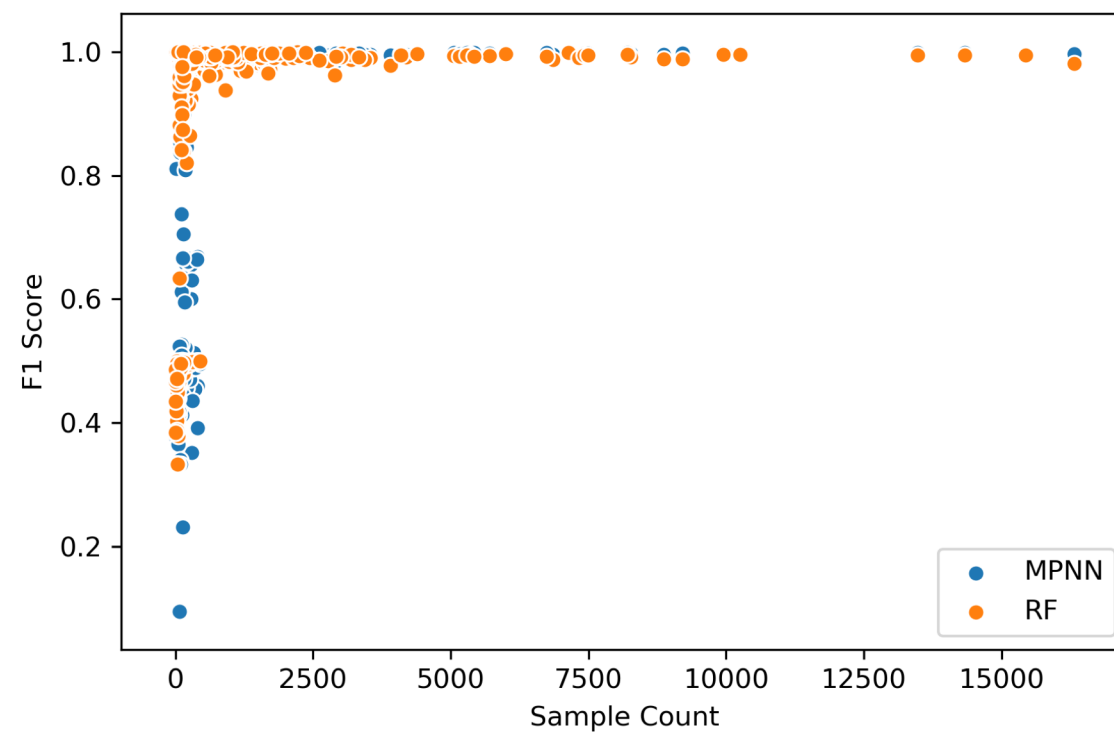
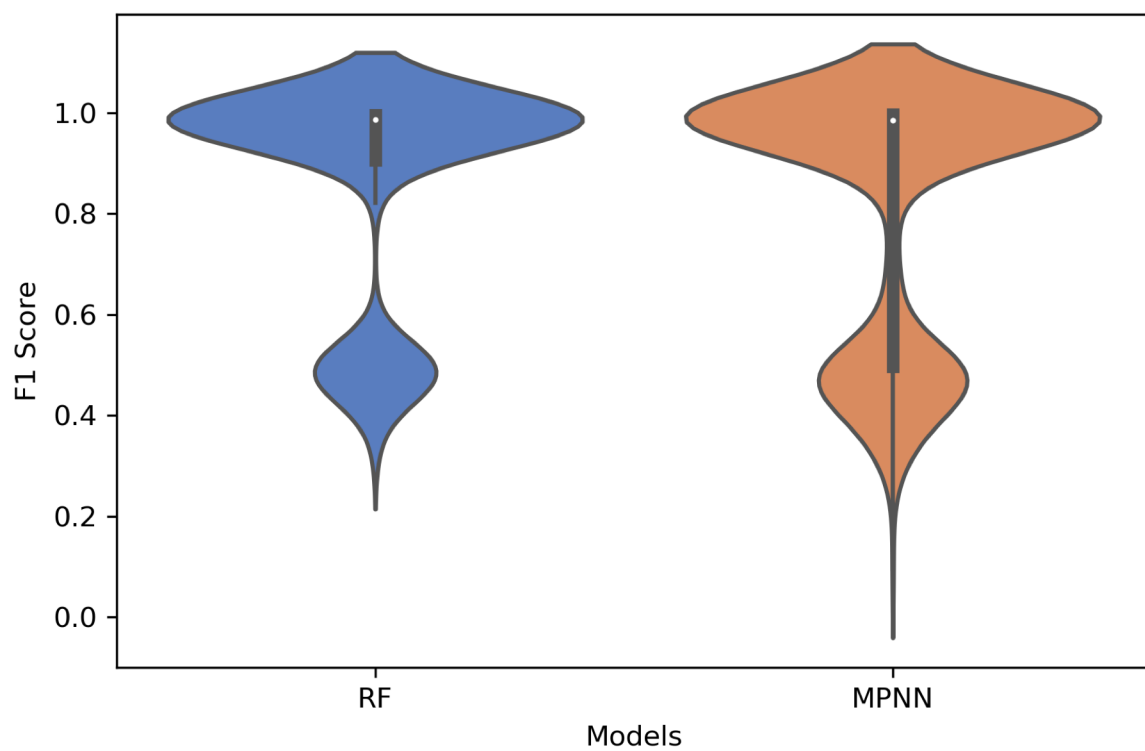


$$m_{vw}^{t+1} = \sum_{k \in [N(v) \setminus w]} M_t(x_v, x_k, h_{kv}^t)$$

$$h_{vw}^{t+1} = U_t(h_{vw}^t, m_{vw}^{t+1})$$

D-MPNN operates on hidden states h_{vw}^t and messages m_{vw}^t instead of on node based hidden states h_v^t and messages m_v^t

RF vs D-MPNN on LigandNet models



Webportal

Ligand Net

Home Contact

Ligand Activity Prediction

Provide a SMILES string:

CC(=O)N/C1=N/N=C([S]1)[S](N)(=O)=O

Example: CC(=O)N/C1=N/N=C([S]1)[S](N)(=O)=O

SUBMIT

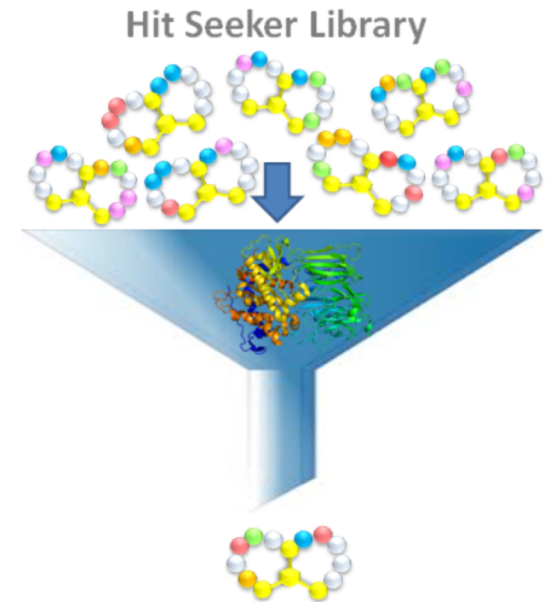
drugdiscovery.utep.edu/ligandnet

Virtual Screening

Computational technique used in drug discovery to search libraries of small molecules in order to identify those structures which are most likely to bind to a drug target, typically a protein receptor or enzyme.

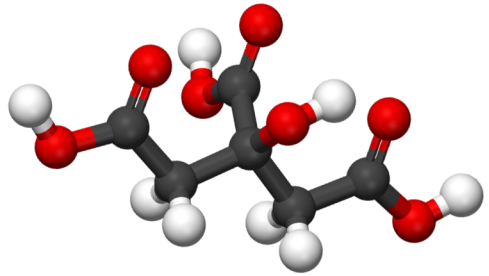
Predicting the number of ligands that will bind to each generated model from Pharos against:

- **Enamine database**
- **Chembridge database**

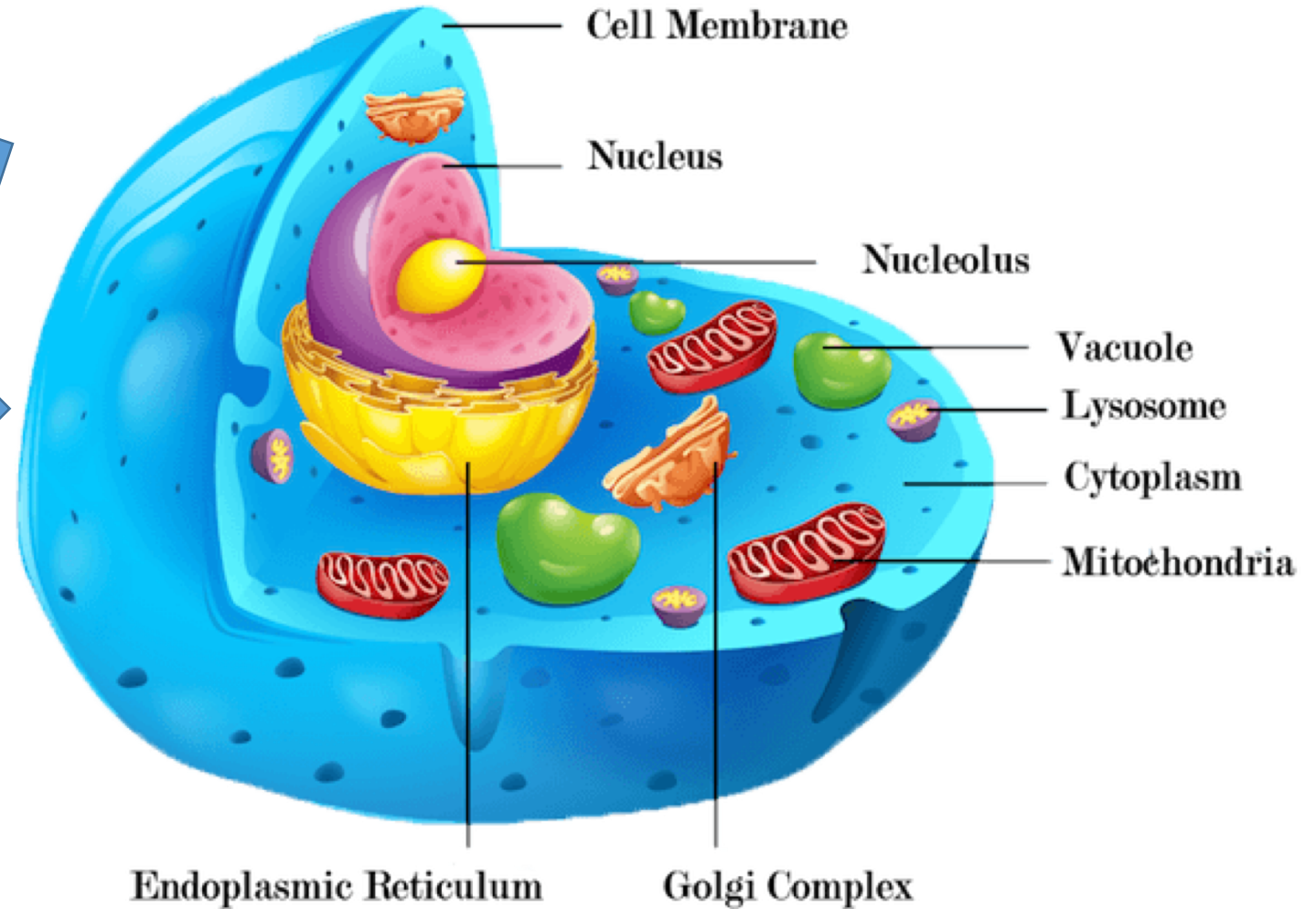


Prediction of ligand activity at Subcellular Location

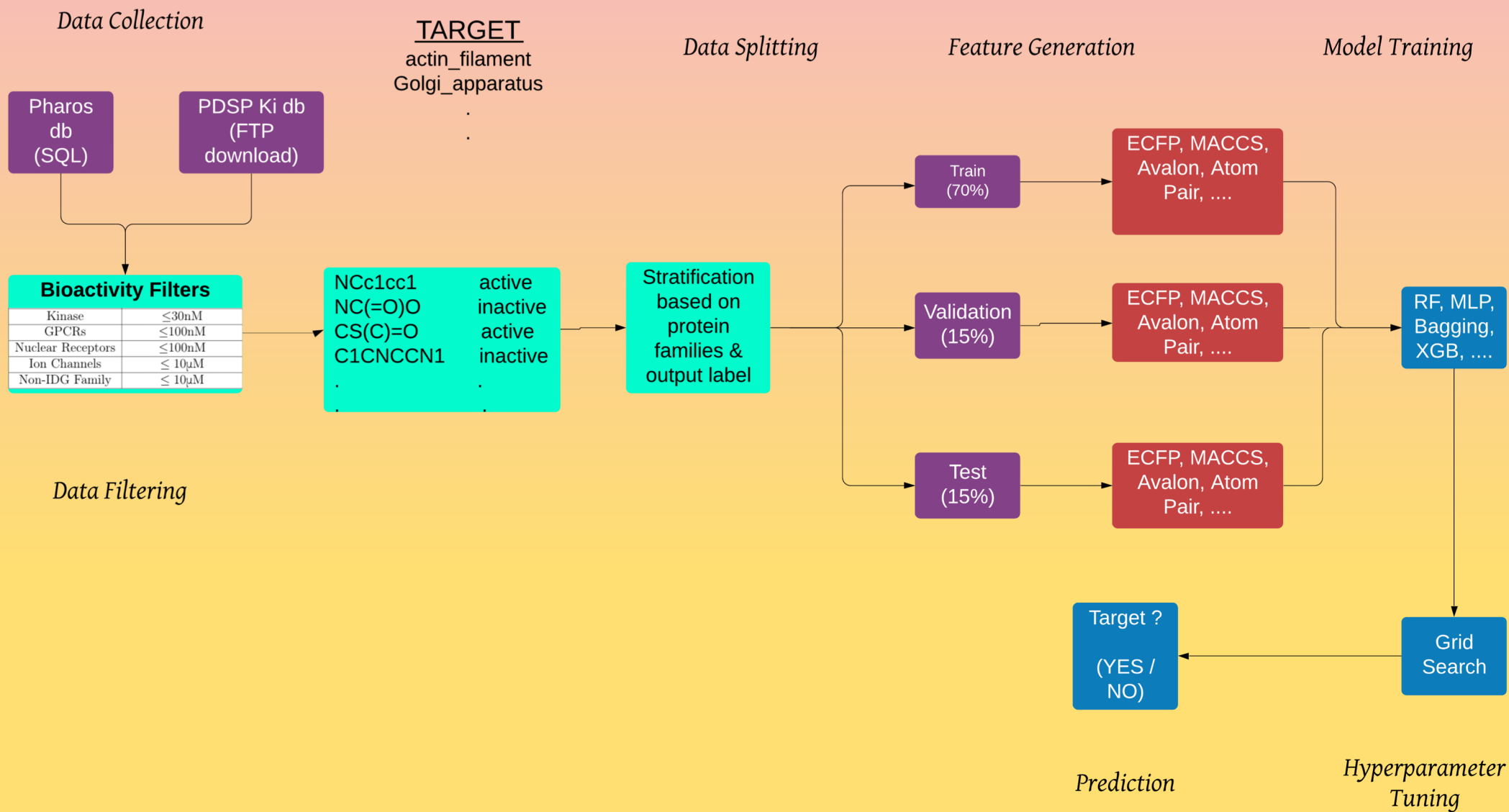
Chemical Entity



Subcellular Location



Workflow



1. Data Collection

- a) Data Collected from Pharos and PDSP database, using Knime and python
- b) Data for 2,167 Uniprot IDs found.
- c) Each Uniprot had several smiles pattern
- d) Total subcellular location = 44
- e) 4 out of these 44 had datapoints<10, which were dropped.

Target	Dataset_Size	Actives	Inactives	No_Protein_Fam
actin_filaments	225049	5175	219874	12
cell_junctions	225049	6184	218865	12
centrosome	225049	2464	222585	12
cytokinetic_bridge	225049	832	224217	12
cytoplasmic_bodies	225049	270	224779	12
cytoskeleton	225049	3132	221917	12
cytosol	225049	60633	164416	12
endoplasmic_reticulum	225049	6955	218094	12
endosomal_membrane	225049	75	224974	12
endosome	225049	1770	223279	12
er_lumen	225049	4599	220450	12
er_membrane	225049	8637	216412	12
focal_adhesions	225049	5228	219821	12
golgi_apparatus	225049	16932	208117	12
golgi_lumen	225049	4222	220827	12
golgi_membrane	225049	582	224467	12
intermediate_filaments	225049	500	224549	12
lysosomal_membrane	225049	397	224652	12
lysosome	225049	1539	223510	12
membrane_protein	225049	439	224610	12
microtubule_organizing_center	225049	1600	223449	12
microtubules	225049	2729	222320	12
midbody	225049	1281	223768	12
mitochondria	225049	14383	210666	12
mitochondrial_membrane	225049	3096	221953	12
mitochondrial_protein	225049	2715	222334	12
mitotic_spindle	225049	927	224122	12
nuclear_bodies	225049	2203	222846	12
nuclear_membrane	225049	8427	216622	12
nuclear_speckles	225049	6355	218694	12
nucleoli	225049	6612	218437	12
nucleoli_(fibrillar_center)	225049	1753	223296	12
nucleoplasm	225049	28154	196895	12
nucleus	225049	29811	195238	12
peroxisome	225049	810	224239	12
plasma_membrane	225049	131862	93187	12
rods_and_rings	225049	485	224564	12
secreted	225049	9670	215379	12
vesicle_membrane	225049	847	224202	12
vesicles	225049	47212	177837	12

Format of Data Collected

S UniProt	S Subcell_Location	S Smiles
A0AVT1	Nucleoplasm Cytosol	<chem>NS(=O)(=O)OC[C@H]1O[C@H]([C@H](O)[C@@H]1O)n2cnc3c(N[C@H]4</chem>
A4D1B5	Vesicles	<chem>CC(NC(=O)Cc1cc(F)cc(F)c1)C(=O)NC2N=C(c3cccc3)c4cccc4N(C)C2:</chem>
A4D1B5	Vesicles	<chem>Fc1ccc(OCC(=O)Nc2ccc3CC4CCC(Cc3c2)C4NS(=O)(=O)c5ccc(Cl)s5)cc</chem>
A5X5Y0	Plasma membrane	<chem>[H]C1(CCN(CCCOC2=CC=C(F)C=C2)CC1([H])OC)NC(=O)C1=C(OC)C=</chem>
A5X5Y0	Plasma membrane	<chem>CCN(CC)CCNC(=O)C1=C(OC)C=C(N)C(Cl)=C1</chem>
B1AH88	Mitochondrial membrane	<chem>CCN(C(=O)Cn1c2nc(ncc2n(c1=O)C)c1cccc1)Cc1cccc1</chem>
B1AH88	Mitochondrial membrane	<chem>CCN(C(=O)Cn1c2nc(ncc2n(c1=O)C)c1cccc1)Cc1cccc1</chem>
B2RXH2	Nucleus	<chem>O=C1NC(=Nc2cnccc12)Oc3cnn(CC4CCCC4)c3</chem>
B2RXH2	Nucleus	<chem>O=C1NC(=Nc2cnccc12)Oc3cnn(Cc4cccc4)c3</chem>
B2RXH2	Nucleus	<chem>N[C@@H]1C[C@H]1c2ccc(NC(=O)c3ccnc(c3)c4cc(ccn4)C(=O)O)cc2.O</chem>
B2RXH2	Nucleus	<chem>CC(C)c1ccc(OC2=Nc3cnccc3C(=O)N2)cn1</chem>

2. Data Preprocessing

- a) Split output variable by '|'
- b) Convert *actin filament* to *actin filaments* for consistency.
- c) Drop duplicate smiles pattern
- d) Data filtered according to the following protein family criteria:
 - 1) **Kinases**: $\leq 30\text{nM}$
 - 2) **GPCRs**: $\leq 100\text{nM}$
 - 3) **Nuclear Receptors**: $\leq 100\text{nM}$
 - 4) **Ion Channels**: $\leq 10\mu\text{M}$
 - 5) **Non-IDG Family Targets**: $\leq 1\mu\text{M}$

3. Data Labelling

- a) Split the data into 2 parts : +ve and -ve.
- b) Label all the +ve as '1', remove duplicate smiles.
- c) Label -ve as '0'.
- d) No -ve data should lie in +ve set.
- e) Concatenate the two.

Format of Labelled Data

S Label	S Smiles	S Subcell_L...	S UniProt
negative	<chem>CC(C)C[C@H](NC(=O)[C@H](CCc1ccccc1)CP(=O)(O)CCC(=O)NCc2ccccc2)C(=O)...</chem>	vesicles	P08254
negative	<chem>CCCCC[C@@H](N[C@@H](CCc1ccccc1)C(=O)N[C@@H](CC(C)C)C(=O)Nc2ccccc...</chem>	vesicles	P08254
negative	<chem>COC(=O)CCP(=O)(O)C[C@@H](CCc1ccccc1)C(=O)N[C@@H](CC(C)C)C(=O)Nc2cc...</chem>	vesicles	P08254
negative	<chem>Sc1nnc(NC(=O)N[C@@H](Cc2ccccc2)C(=O)N3CCN(CC3)c4ccccn4)s1</chem>	vesicles	P08254
negative	<chem>CNC(=O)[C@H](Cc1ccccc1)NC(=O)[C@H](CCc2ccccc2)N[C@@H](CCNS(=O)(=O)...</chem>	vesicles	P08254
negative	<chem>Sc1nnc(NC(=O)N[C@@H](Cc2ccccc2)C(=O)NCc3ccccc3)s1</chem>	vesicles	P08254
positive	<chem>CN(Cc1cc(Cl)cc(C2=CC(=C(C#N)C(=O)N2)c3cc(ccc3Cl)C(F)(F)F)c1O)C(=O)CCN4...</chem>	cytokinetic b...	O15392
positive	<chem>C[C@]12CC[C@H]3[C@@H](CCc4cc(O)ccc34)[C@@H]1CC[C@@H]2O</chem>	cytokinetic b...	P11474
positive	<chem>Oc1ccc(Cl)cc1C2=CC(=C(C#N)C(=O)N2)c3cc(ccc3Cl)C(F)(F)F</chem>	cytokinetic b...	O15392
positive	<chem>CC(C)n1ncnc1c2nc3c(CCOc4cc(ccc34)C5CCNCC5)s2</chem>	cytokinetic b...	O00329
positive	<chem>COc1ncc(cc1C(F)(F)F)N2CCc3ncnc(O[C@H]4CCN(C4)C(=O)CC(C)(C)C)c3C2</chem>	cytokinetic b...	O00329
positive	<chem>CNC(=O)c1ccc2c(OCCc3cc(sc23)C(=O)N(C)c4ccc(cc4Cl)C(=O)N(C)C)c1</chem>	cytokinetic b...	O00329

4. FP Generation

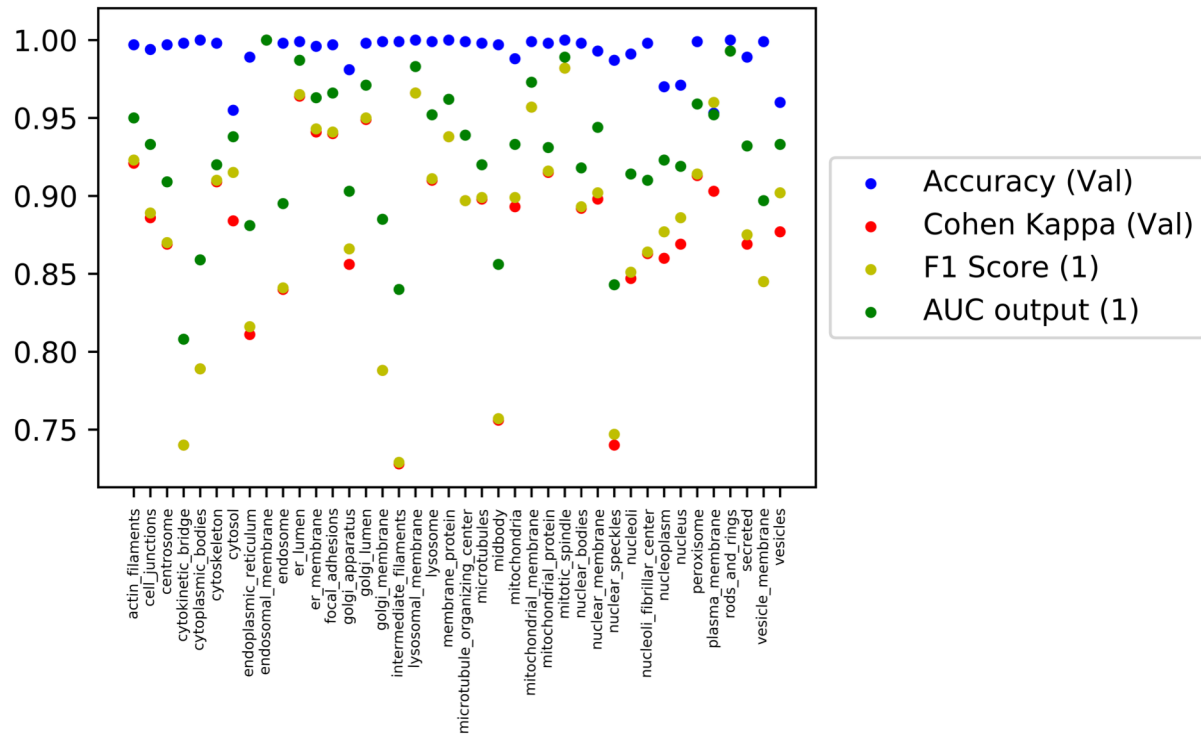
- a) 19 FPs, of 3 different kinds (Circular, Path-based and Substructure keys) and Descriptors generated.
- b) RDKit module used here.
- c) Each of the FP converted to a bit vector of either 1024 or 16,384 length.

5. Model Training

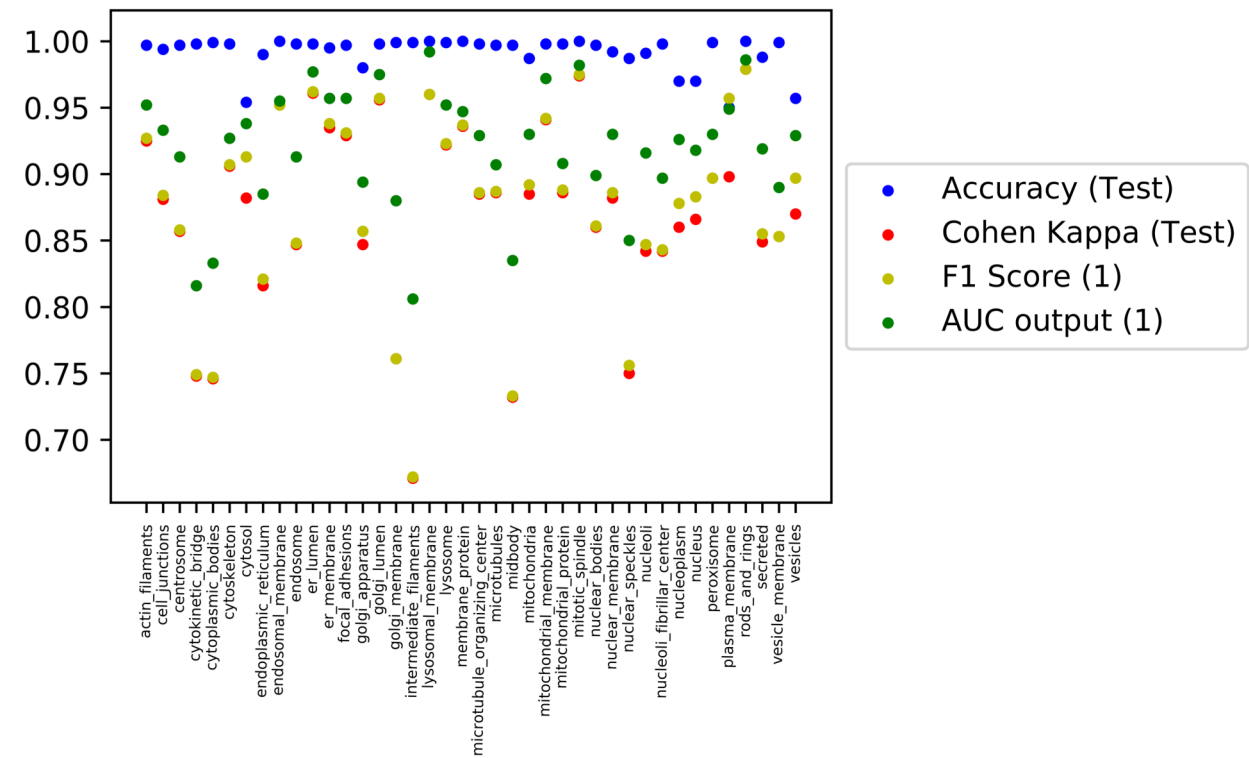
- a) 28 sklearn classifier models used
and $20 * 28 * 40 = 22,400$ models trained.
- b) Models were first trained in their default settings, after which grid search was used.
- c) Train:Val:Test split :: 70:15:15 (for all models)

Best Model Val Scores and their performance on Test data

Val Results for 40 best models

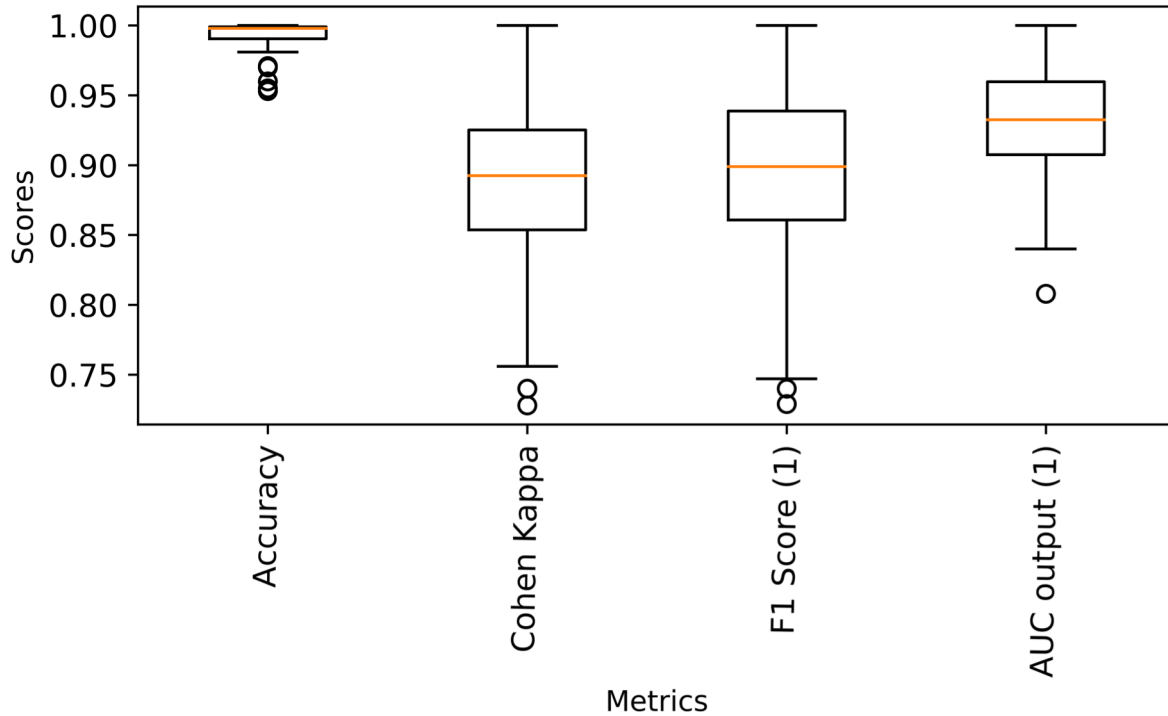


Test Results, based on best 40 validation models

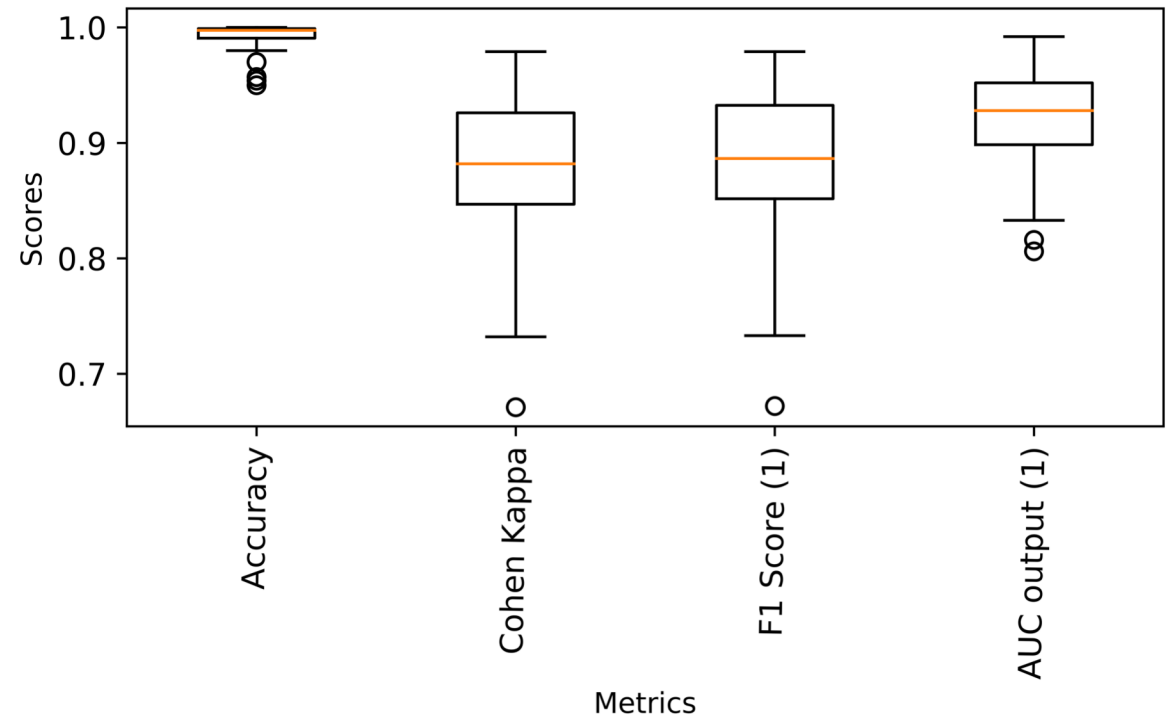


Performance of Models on Validation and Test sets

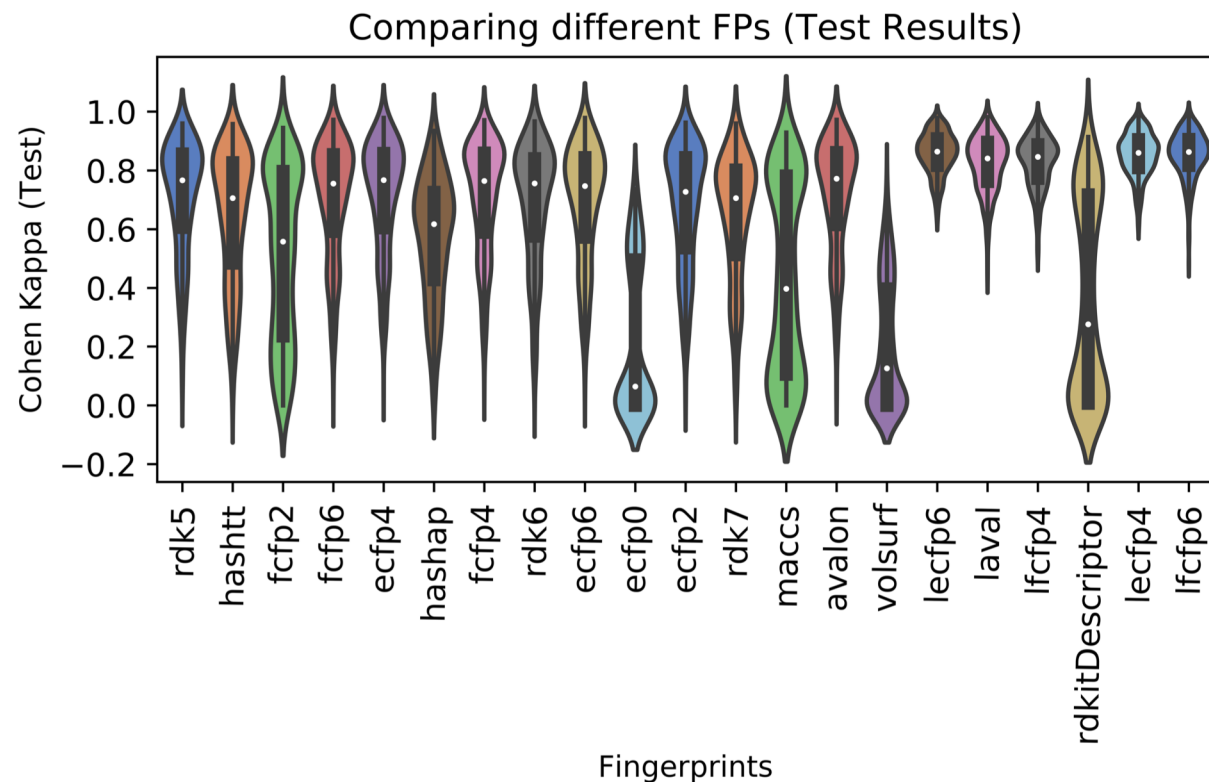
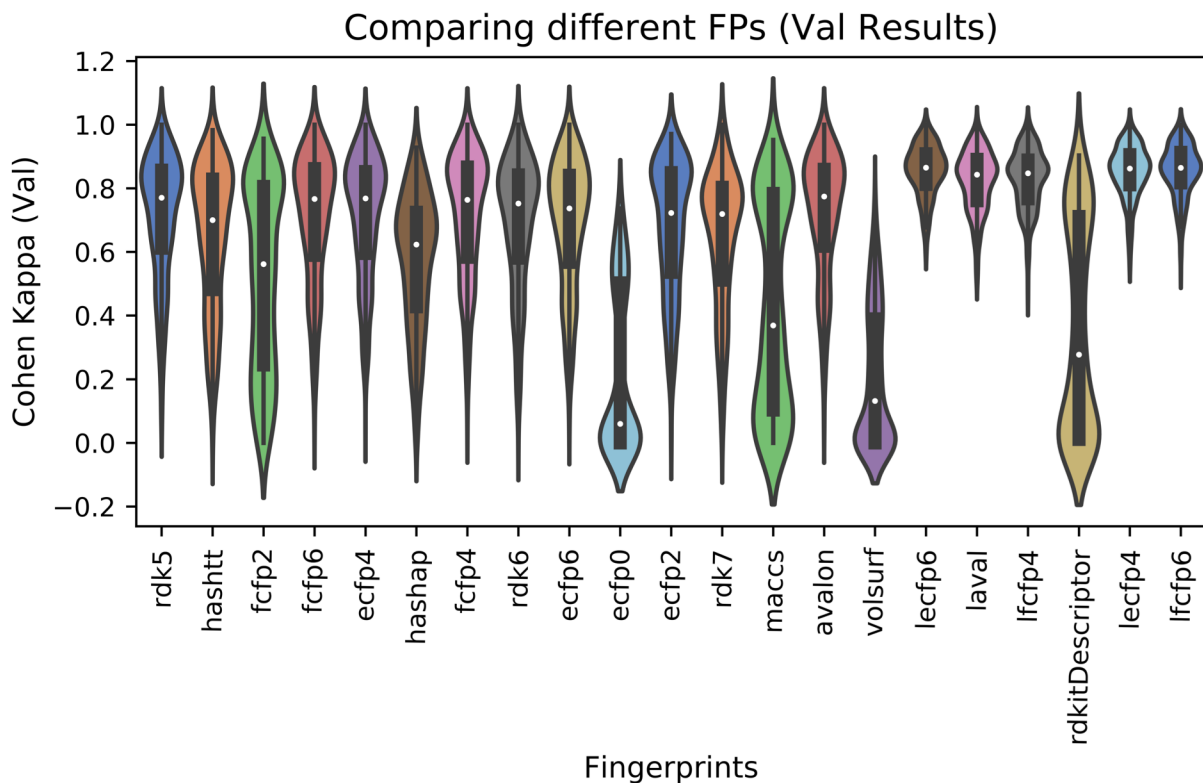
Val Results for 40 best models



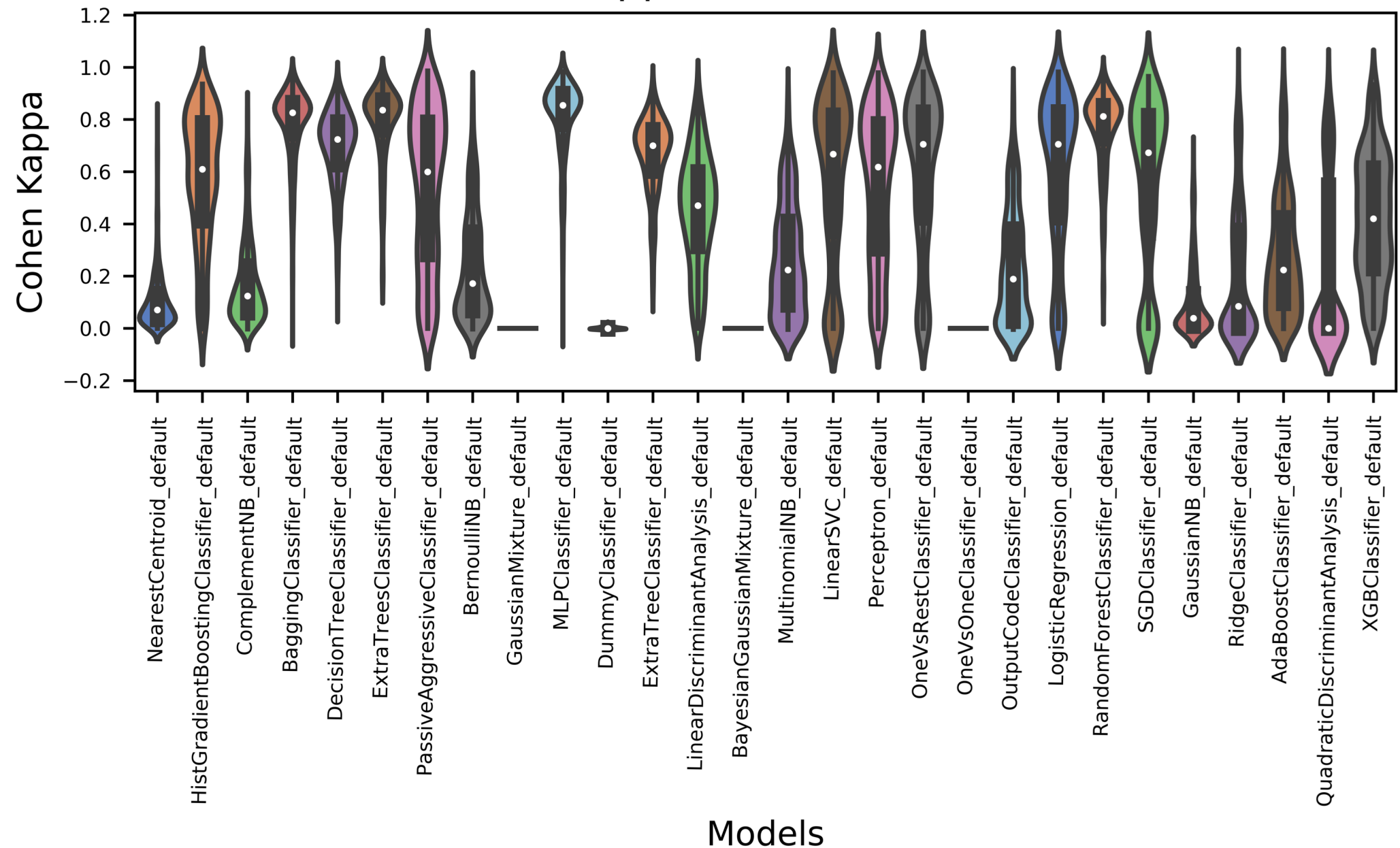
Test Results based on 40 best Val models



Performance of Features on validation and test sets

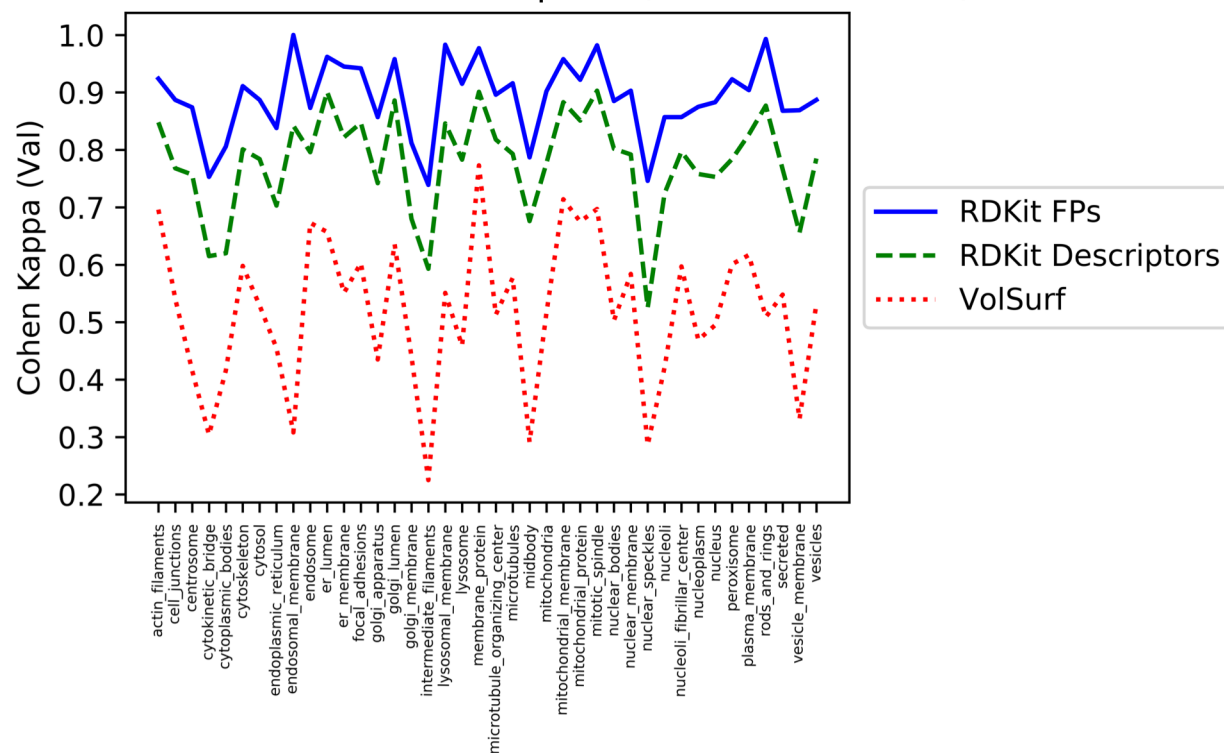


Overall Cohen Kappa (Test) for various classifiers

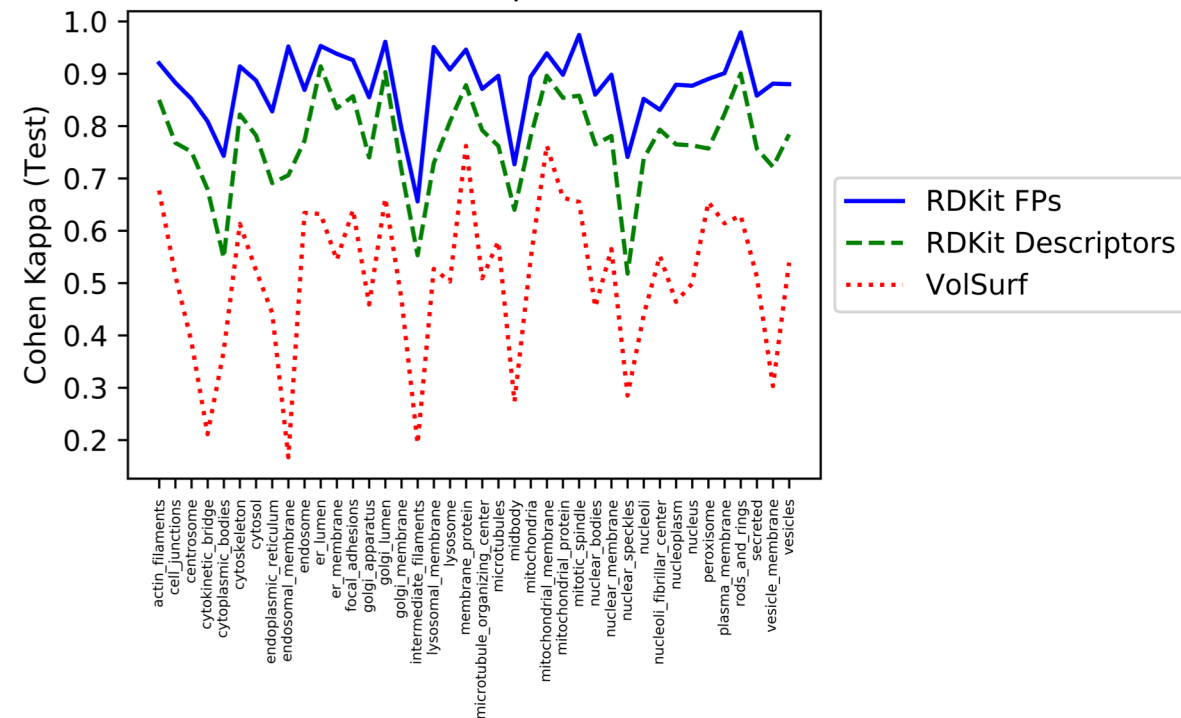


Comparison of VolSurf Descriptors vs RDKit Features

FPs vs VolSurf vs RDKit Descriptors (Best Models - Val)



FPs vs VolSurf vs RDKit Descriptors (Best Models - Test)



Volsurf- 128 ADME relevant descriptors

OpenDMPK

Objectives:

- To establish an opensource tool kit for prediction of DMPK properties.
- Boosting prediction accuracy with advanced algorithms.

Xenobiotic Metabolizing Enzymes

Phase I "Oxygenases"

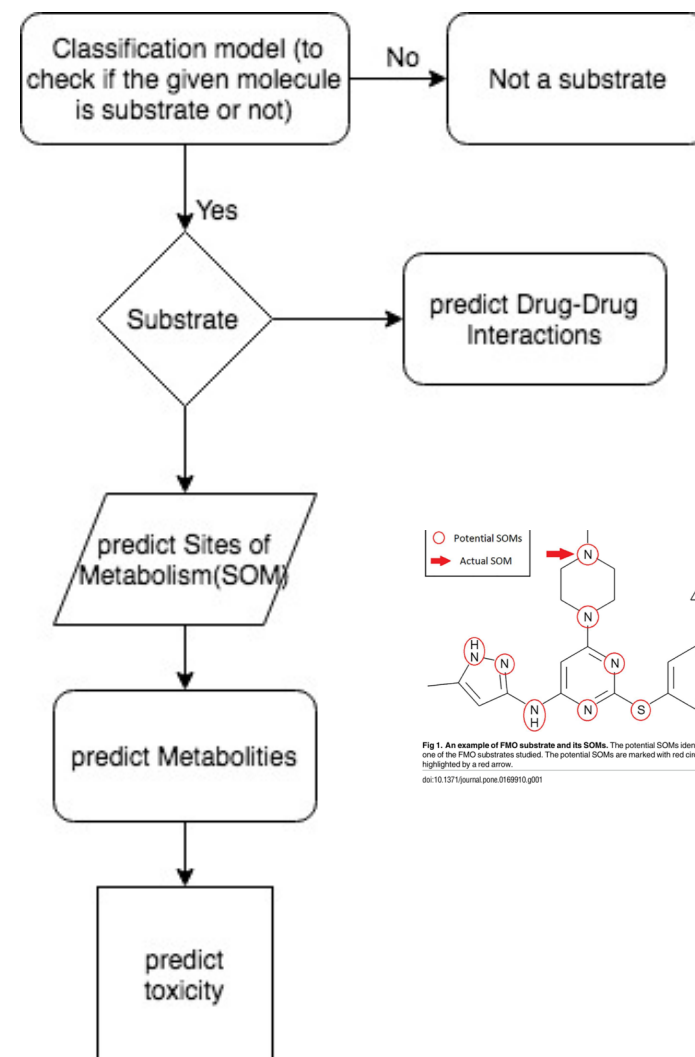
- Cytochrome P450 (P450 or CYP)
- Flavin-containing-monooxygenase(FMO) -----> N and S oxidation
- Epoxide hydrolases(mEH, sEH)

Phase 2 "Transferases"

- Sulfotransferases(SULT)
- UDP-glucuronosyltransferases(UGT)
- Glutathione-S-transferases(GST)
- N-acetyltransferases(NAT)
- Methyltransferases(MT)

Other Enzymes

- Alcohol dehydrogenases
- Aldehyde dehydrogenases
- NADPH-quinone oxidoreductase(NQO)



openDMPK: open source tool for drug metabolism and pharmacokinetic properties prediction

Absorption

- 1) Human Oral Bioavailability (HOB)
- 2) Human Intestine Absorption (HIA)
- 3) Caco2Permeability (Caco2)

Distribution

- 1) Plasma protein binding (PPB)
- 2) P-glycoprotein substrate (PgpS)
- 3) P-glycoprotein inhibitor (Pgpi)
- 4) Blood-brain barrier penetration (BBB)
- 5) Volume Distribution (VD)
- 6) Central Nervous System Permeability (CNSP)

Metabolism

- 1) Cytochrome P450 (CYP)
 - a) Cyp2c9 substrate
 - b) Cyp2d6 substrate

Excretion

- 1) Total Clearance (TC)

Toxicity

1. Eye Corrosion (EC)
2. Eye Irritation (EI)
3. Hepatotoxicity (HEP)
4. Honeybee Toxicity (HBT)
5. Avian Toxicity (AT)
6. Fish Aquatic Toxicity (AT)
7. Caco2Permeability (Caco2)
8. Crustacean Aquatic Toxicity (CAT)
9. Tetrahymena Pyriformis Toxicity (TPT)
10. Ames Mutagenesis (Ames)
11. Biodegradation (Biod)
12. Micronucleus assay (MN)
13. Human either-a-go-go inhibition (hEGG)
14. Oral Rat Toxicity Lowest Observed Adverse Effect (LOAE)

openDMPK: open source tool for drug metabolism and pharmacokinetic properties prediction

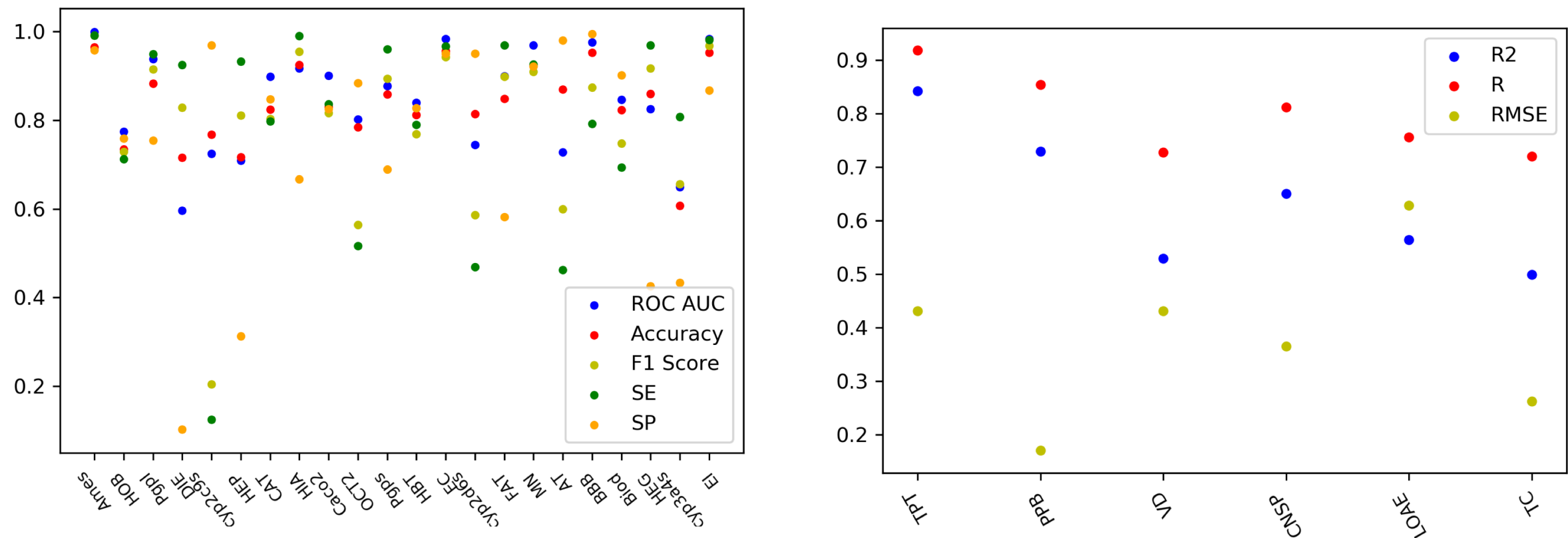
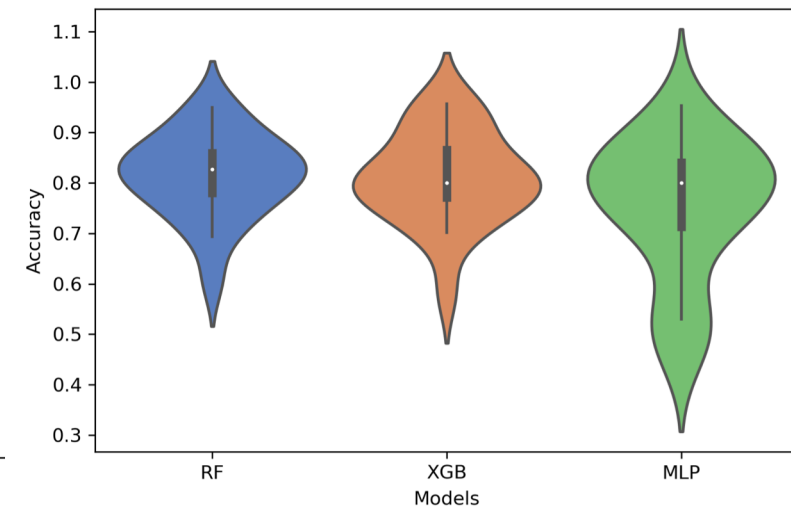
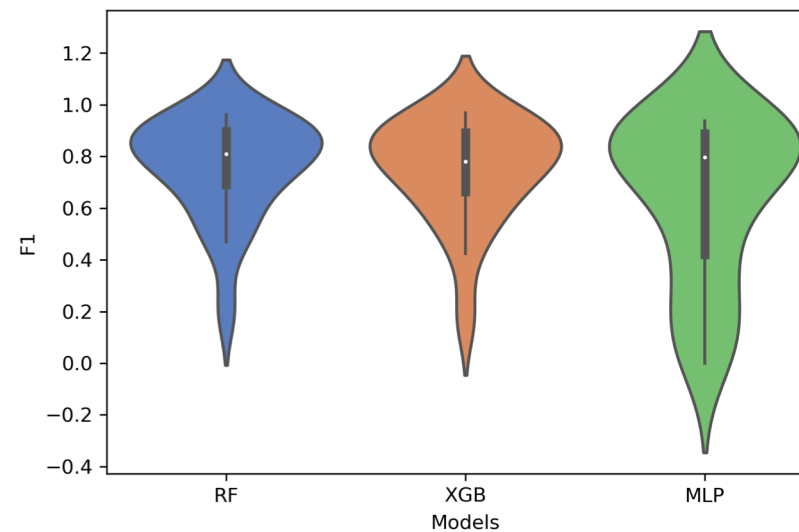
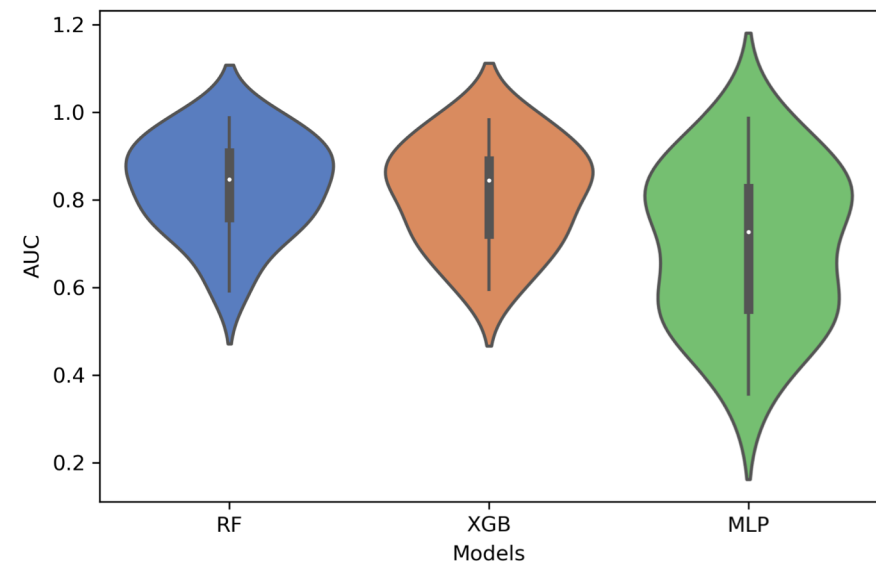
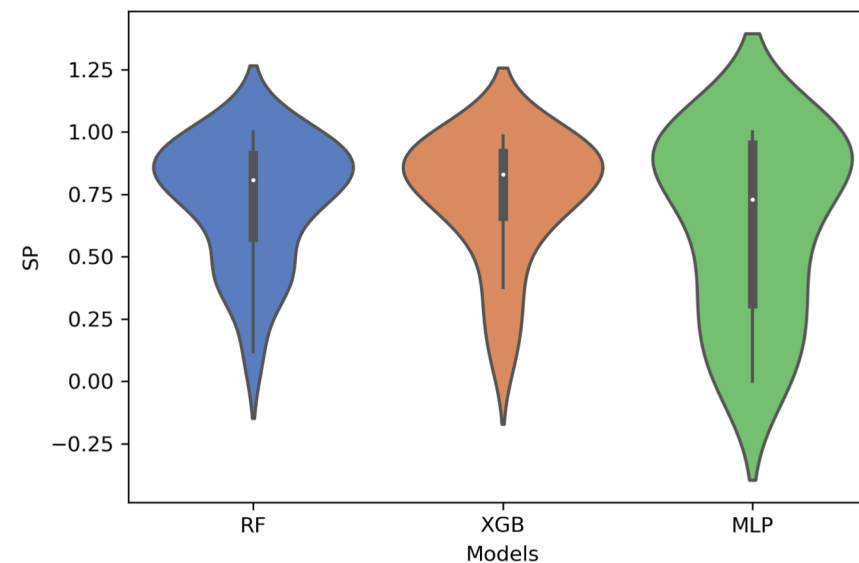
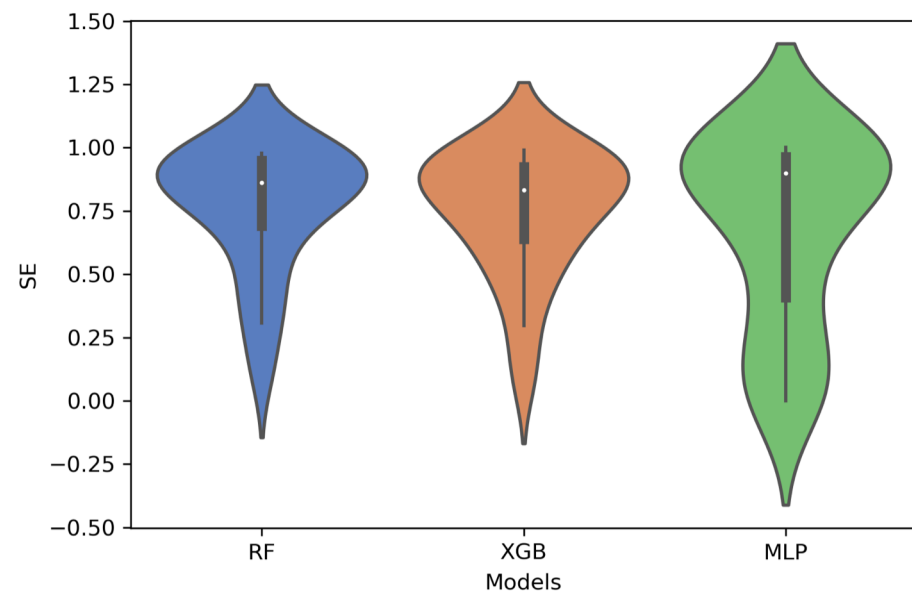


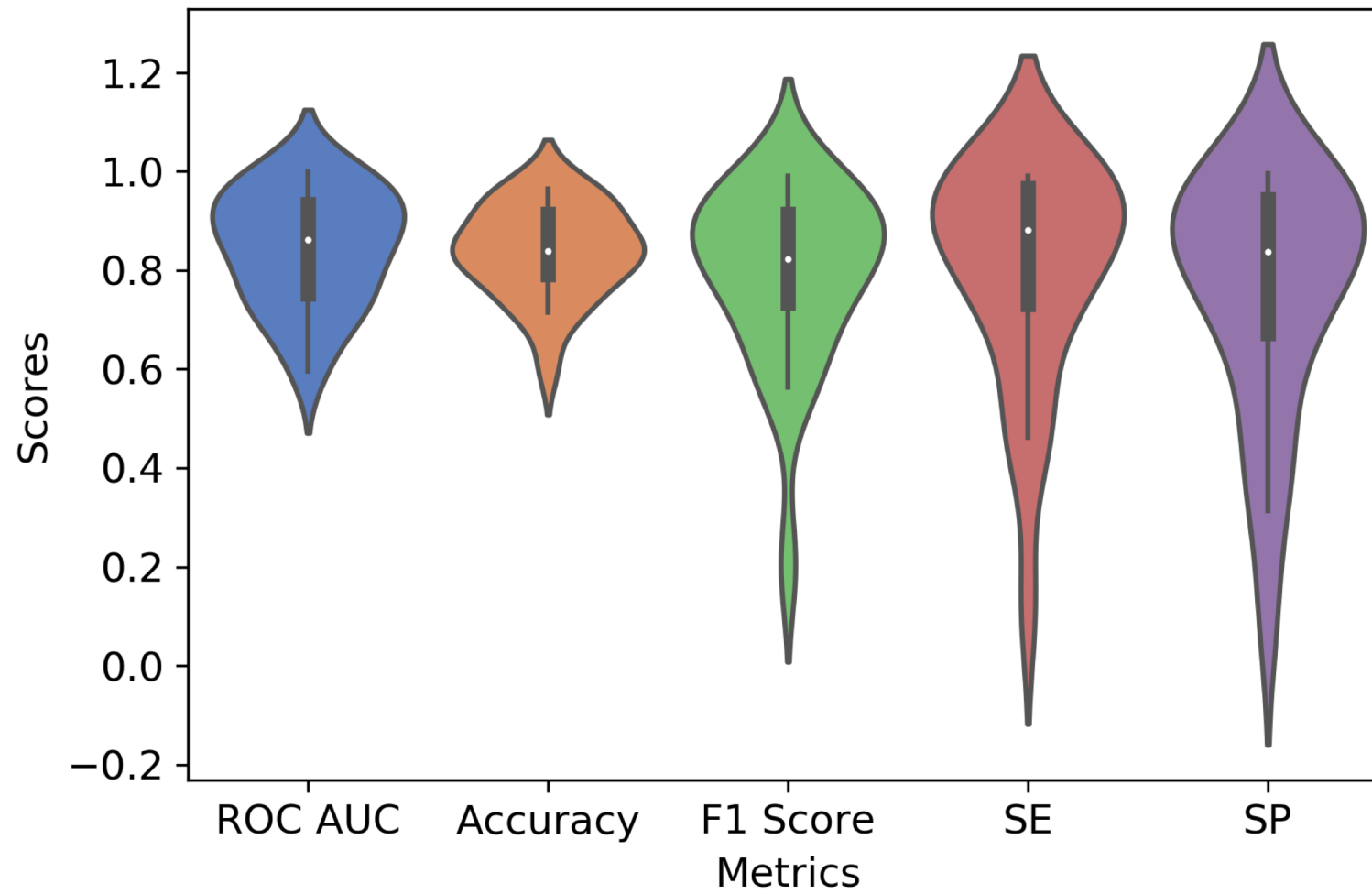
Fig. Scatter plots of (a) Classification models (left) (b) Regression models (right)



Comparison of RF, XGB, and MLP using (a) AUC (b) F1 score (c) Accuracy

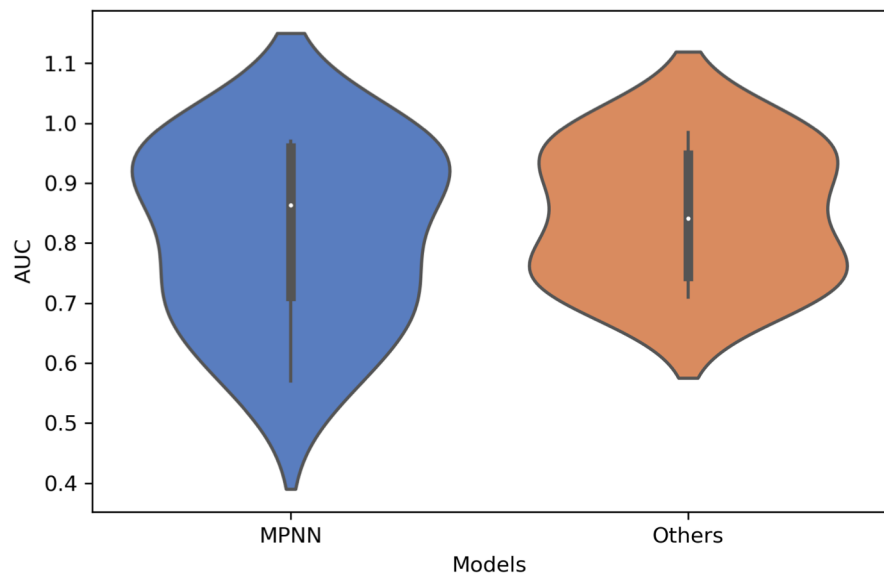
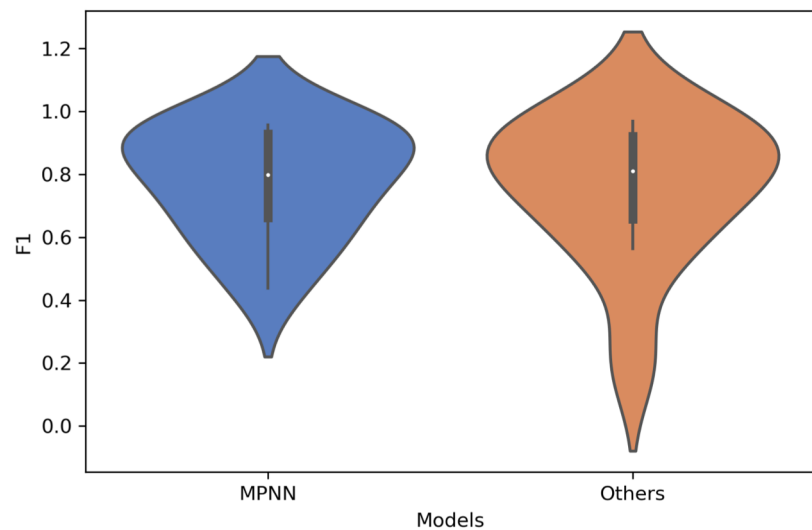


Comparison of RF, XGB, and MLP using (a) SE (b) Specificity (SP)

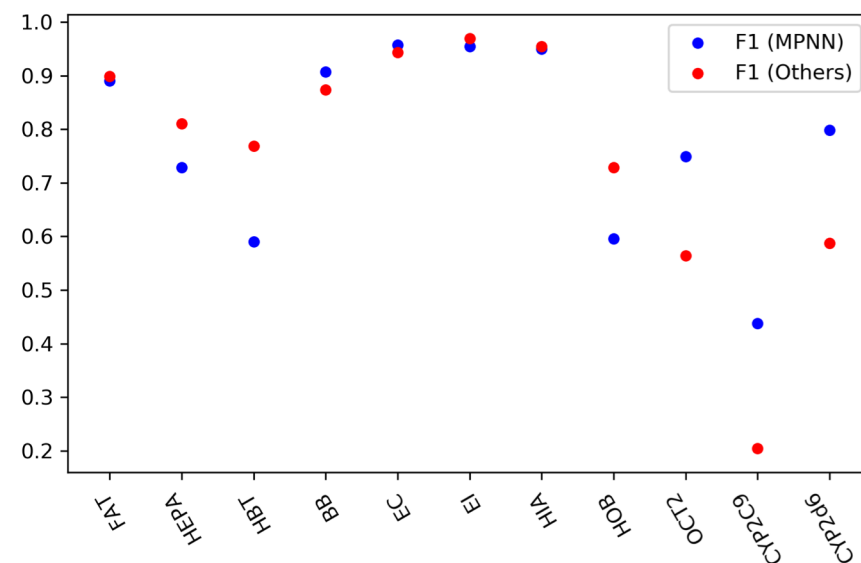
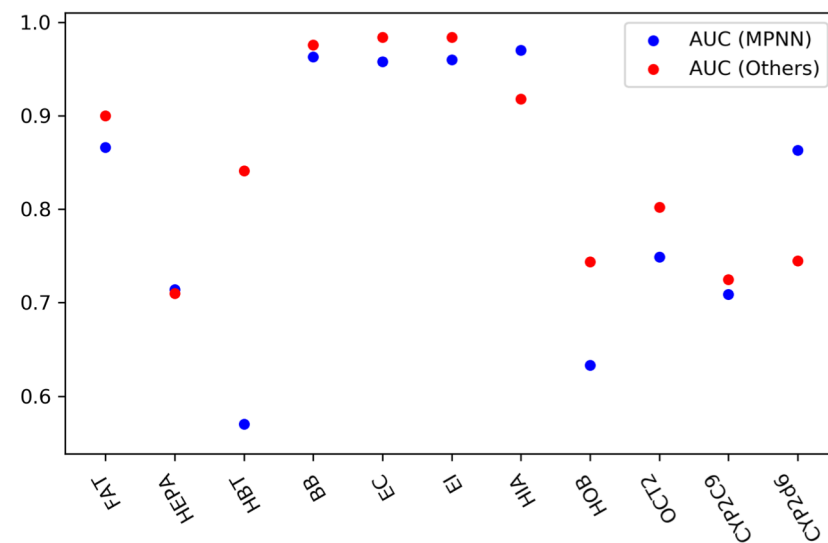


Violin plots of best classification models for AUC, F1, SE, and SP

**Comparison of D-MPNN and Others
using F1 Score (top) and AUC (bottom)**



Comparison of D-MPNN and others (best models) (a) ROC-AUC (top) (b) F1 (bottom)



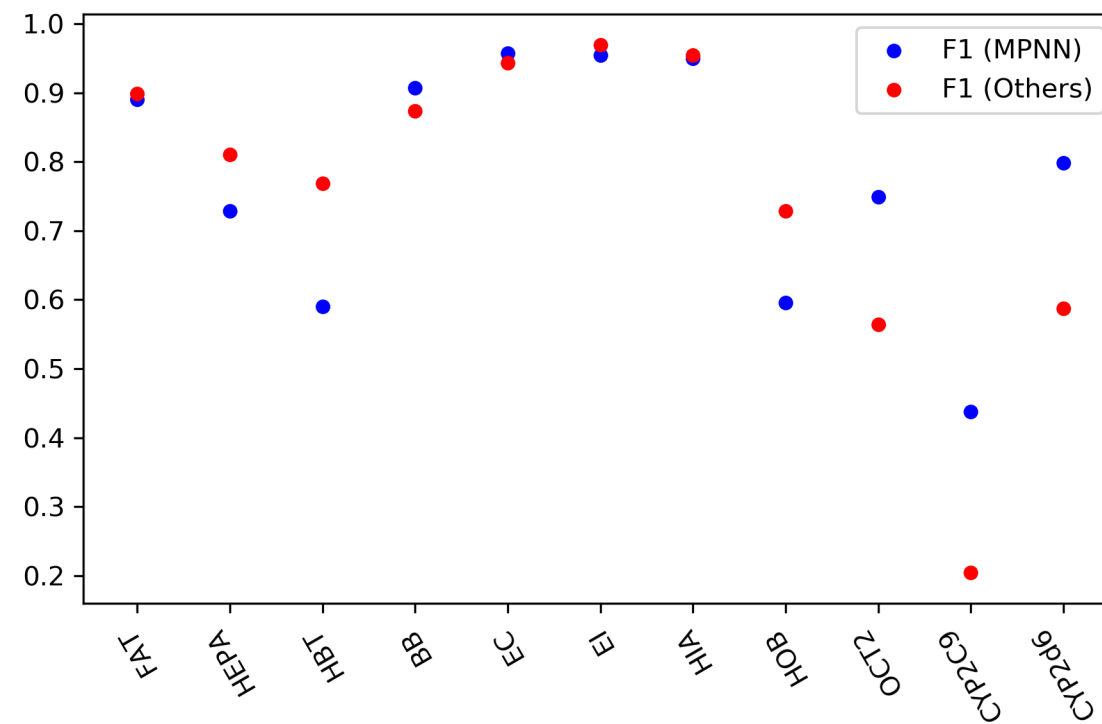
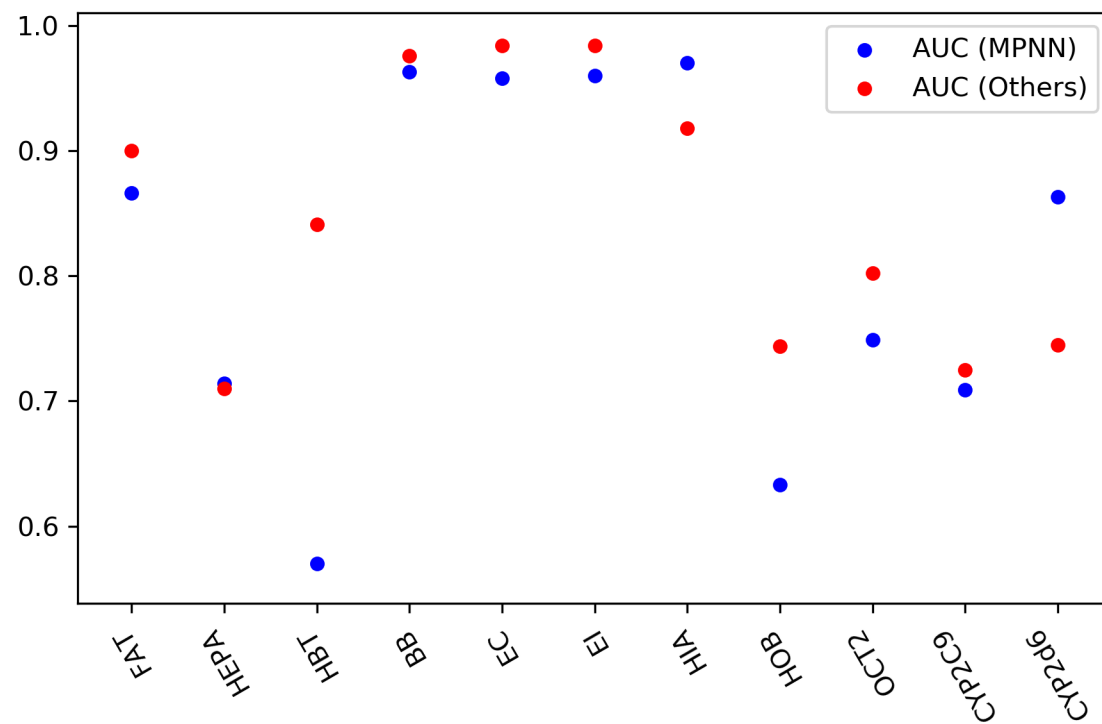


Fig: Comparison of GCNN and others (best models) (a) ROC-AUC (left) (b) F1 (right)

Webportal

OpenDMPK - Open Drug Metabolism and Pharmacokinetic Properties

[Home](#) [Contact](#)

Drug Metabolism Prediction

Provide a SMILES string:

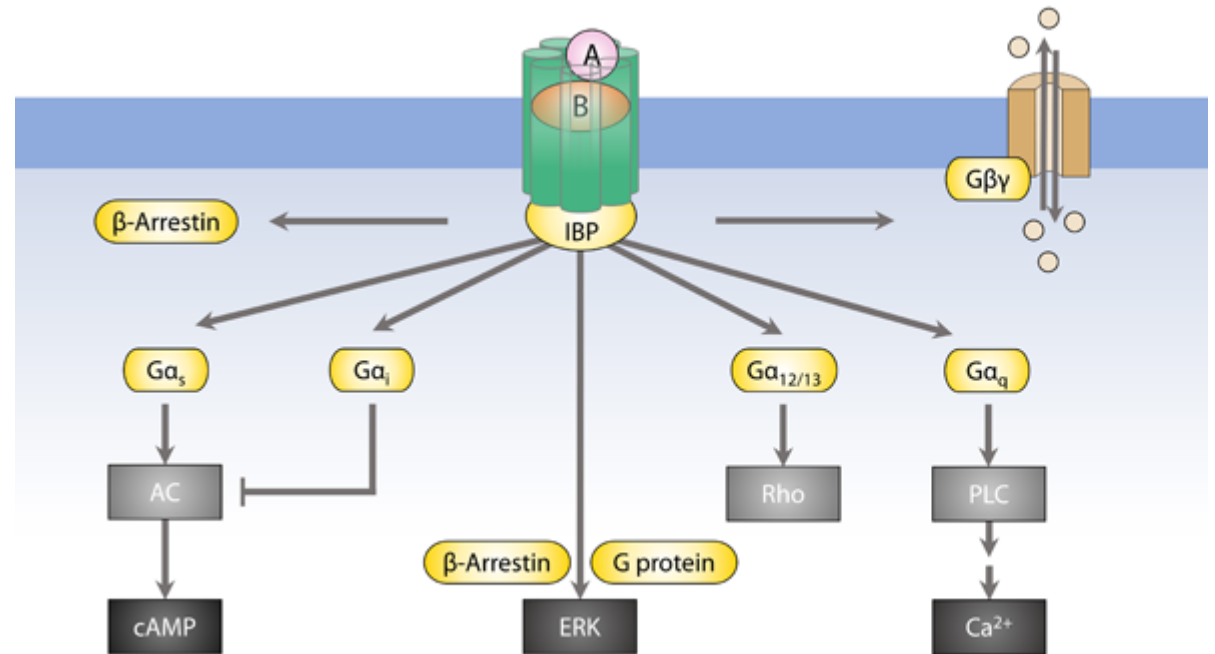
Example: COC(=O)C(C1=CC=CC=C1Cl)N2CCC3=C(C2)C=CS3

SUBMIT

drugdiscovery.utep.edu/opendmpk/v2

Drug-Related GPCR Signaling Bias

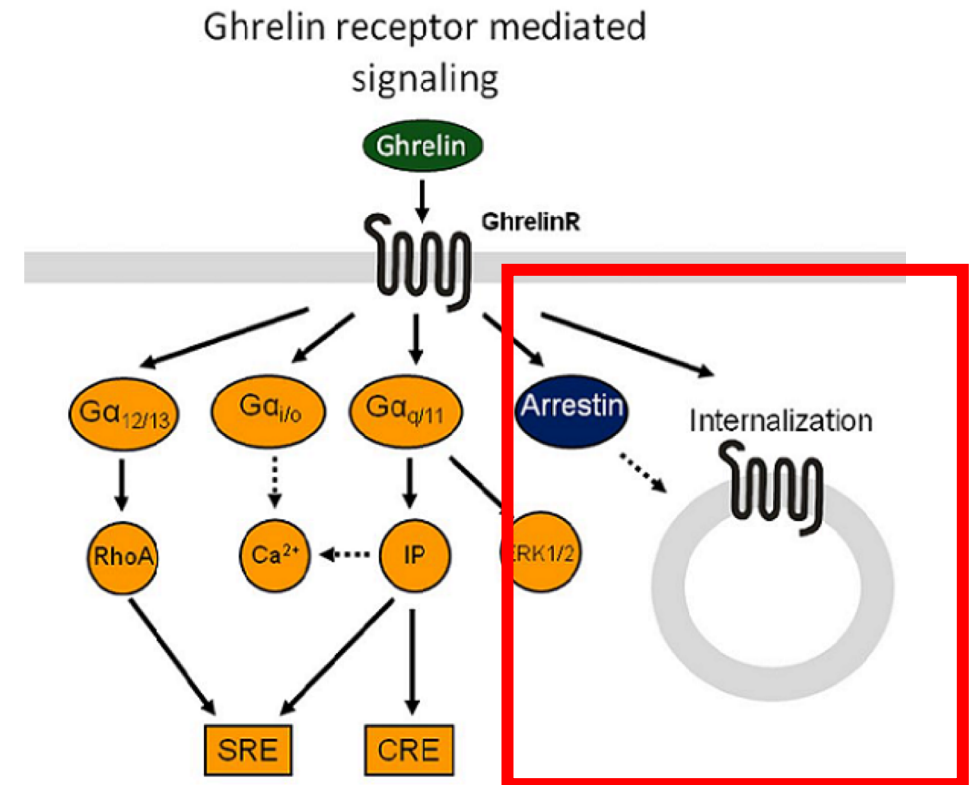
- Most drug targets are G-Protein Coupled Receptors (GPCRs)
- GPCRs can activate many downstream players
- Effective drug therapies minimize unwanted side effects



BiasDB: A Comprehensive Database for Biased GPCR Ligands

Why Should We Care?

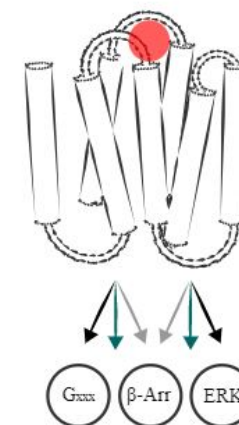
- Consider a pathway with many branchpoints
- The β -Arrestin pathway internalizes the Ghrelin receptor
- Such mechanism downregulates Ghrelin's own action
- Might want to prevent self downregulation



- BiasedDB groups molecules by bias
- Cases are grouped by 4 main bias categories
 - G-Protein, G-Protein Selective, β -Arrestin, ERK
- Biased categories are further subdivided into specific biases
 - E.g. G_{α} over G_i

BiasDB

A manually curated database of biased GPCR ligands



BiasDB^[1] is a manually curated database containing all published biased GPCR ligands. BiasDB currently contains **654 bias cases** of signaling bias representing **489 individual ligands** for **61 receptors**. We provide information about the chemical structure, target receptor, type of bias, assay categories used for bias determination, reference ligand and literature source. BiasDB is a resource for medicinal chemists, pharmacologists and researchers interested in biased GPCR signaling.

Data Search

Structure Search

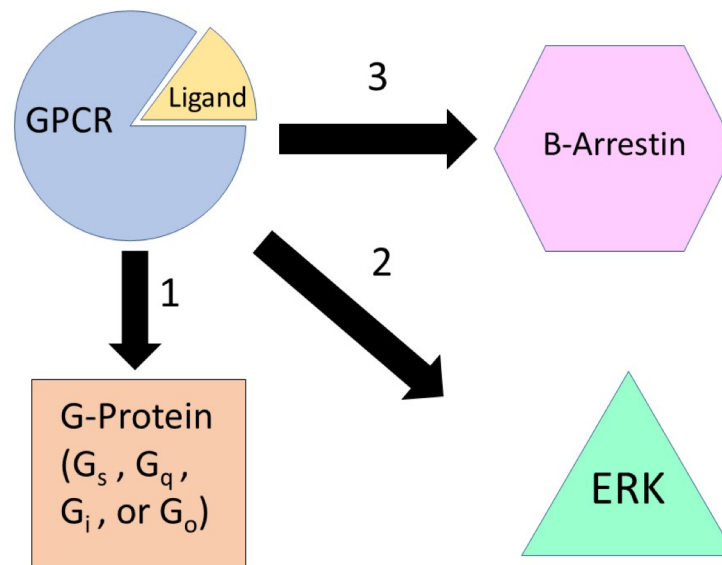
Visualize categories

Our Goal

Describe specific pharmacophores that contribute to biased signaling

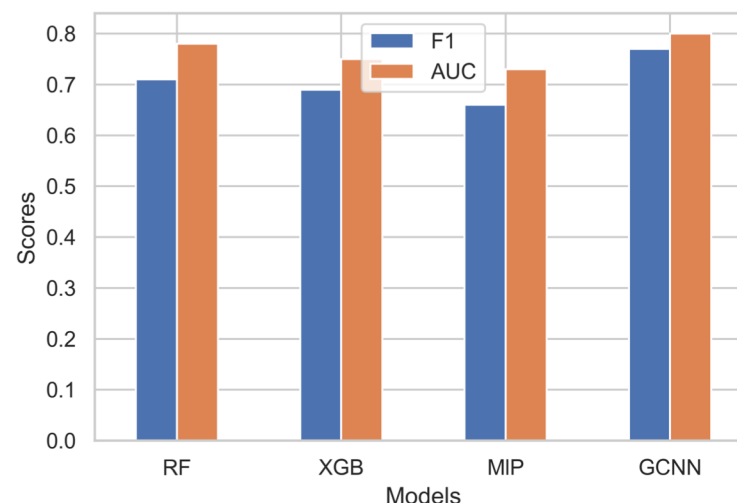
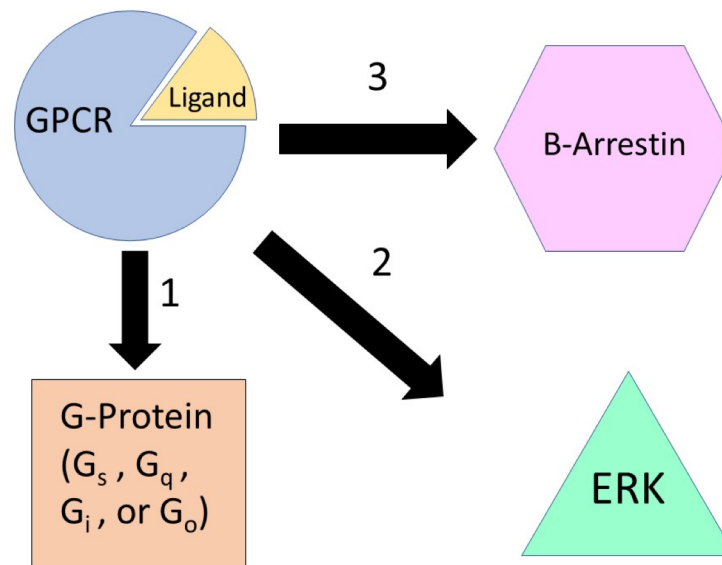
Our Approach So Far

1. Group BiasDB data into three categories: G-Protein, β -Arrestin, and ERK
2. Obtain molecular descriptors for ligands using RDKit
3. Train and test four machine learning models on known biased cases
 - Random Forrest, XGBoost, Multilayer Perception, Graph Convolution Neural Networks



Our Approach So Far

1. Group BiasDB data into three categories: G-Protein, β -Arrestin, and ERK
2. Obtain molecular descriptors for ligands using RDKit
3. Train and test four machine learning models on known biased cases
 - Random Forrest, XGBoost, Multilayer Perception, Graph Convolution Neural Networks



Acknowledgments



- Ant Nicholls (Openeye)
- **Lab Members**
 - Mahmudulla Hassan (Graduate student)
 - Govinda KC(Graduate student)
 - Jason Sanchez (Graduate student)
 - Srijan Verma (Visiting Scholar-BITS PILANI)
 - Denise Cano (UNM summer intern)
 - Ibrahim Salama (undergraduate student)
 - Claire Govia (undergraduate student)
- **Collaborators**
 - Tudor I. Oprea (UNM)
 - Giovanni Bocci
 - Jeremy yang

