

MolScreen

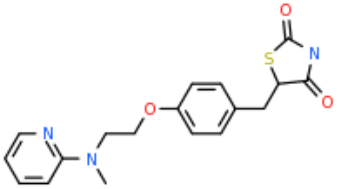
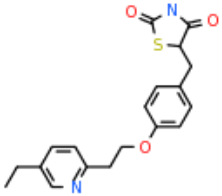
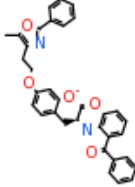
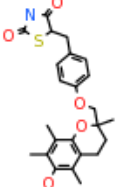
Polo Lam, Ph.D.

Senior Research Scientist

MolSoft LLC

MolScreen

- Objective: Self-contained prediction models for **MolSoft ICM Screening**
- Currently: ~3280 models for ~1300 targets (continual expansion/improvement)

	mol		MolPPARG	MolPPARA	MolPPARD
1		racemic	6.892	5.253	5.298
2		racemic	6.165	5.117	ND
3		chiral	7.93	6.348	5.626
4		racemic	6.076	6.655	4.912

Usage

- Compounds -> Which Targets?
- Target -> What compounds?
- Profiling: Multi-Targets vs Multi-Cpds
- Drug Re-purposing

2 Categories of Models in ICM

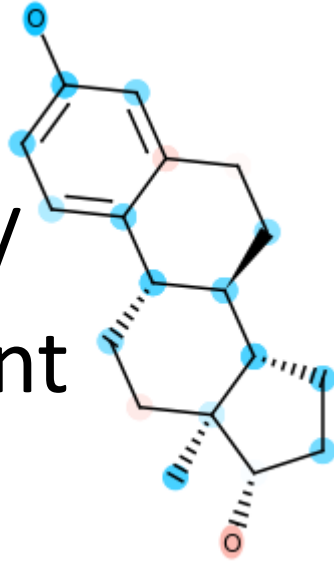
- ADMET (**mcp**), Property Models
 - CACO2, hERG, HALFLIFE, LD50, CYP, Tox21, etc
 - Properties like, Regression/Classification
- **5 Different types of Activity Models**
 - ~3280 models against ~1300 targets
 - Fingerprint (**kcc, eca**), 3D Atomic Properties Field (**dfz**), 4D Docking/3D-QSAR (**dpc**), 3D APF/3D-QSAR (**dfa**)

ADMET (Miscellaneous Chemical Property **mcp**) Models

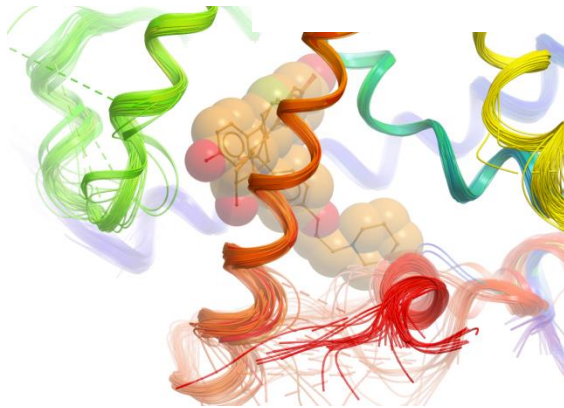
- Currently 38 models, mostly from PubChem data
- All validated by external test set (20% of data set aside)
- Regression Models, Mean external test set Q2: 0.7
 - CACO2, PAMPA permeability
 - LD50 (mg/kg), Half-life (hr)
- Classification Models, Median external test set AUC: 84%
 - hERG, PGP inhibitor, PGP substrate, PAINS
 - Cytochrome P450 1A2, 2C19, 2C9, 2D6, 3A4
 - 25 Tox21 Classifier, including Estrogen Agonist/Antagonist, Genotoxicity, Aromatase, etc

5 Types of Activity Models in ICM

2D QSAR/
Fingerprint
(**kcc/eca**)



3D Atomic
Property
Field (**dfz**)

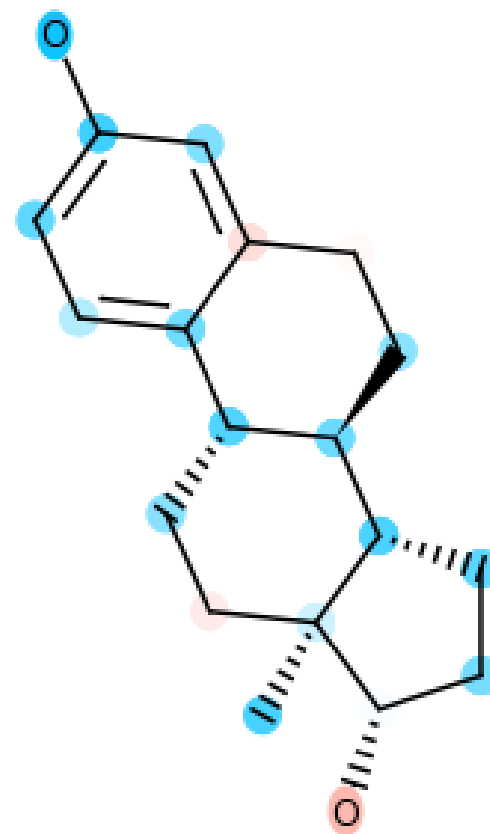


APF/3D-QSAR
(**dfa**)

Docking/3D-QSAR (**dpc**)

2D QSAR/Fingerprint (**kcc**)

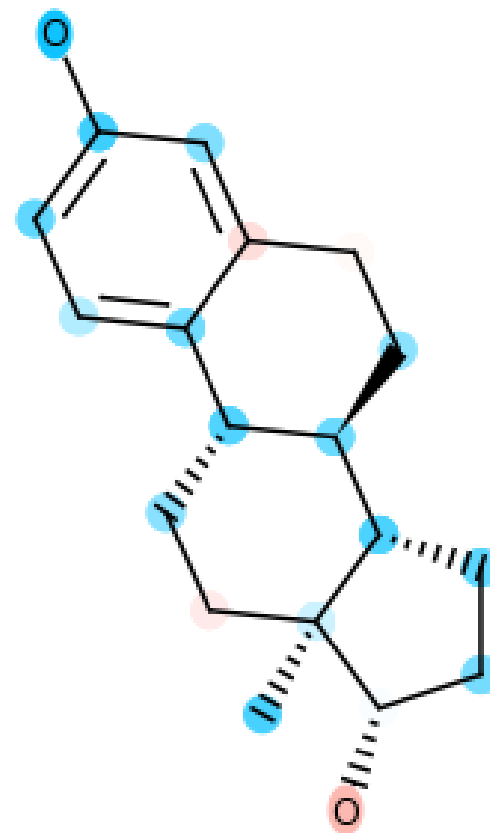
- **kcc**: **K**ernel **C**hemical fingerprint **C**lassification/Activity
- Currently: 999 mammalian models
- Training set: ChEMBL Ki, IC50, EC50, Drugbank assignment
- Median size: 245 ligands
- **All Models' Validation: 20% of ChEMBL set aside as external set vs Approved drugs decoy**
- Median external Q^2 : 0.52
- Median external AUC: 97%



2D QSAR/Fingerprint (kcc) Method

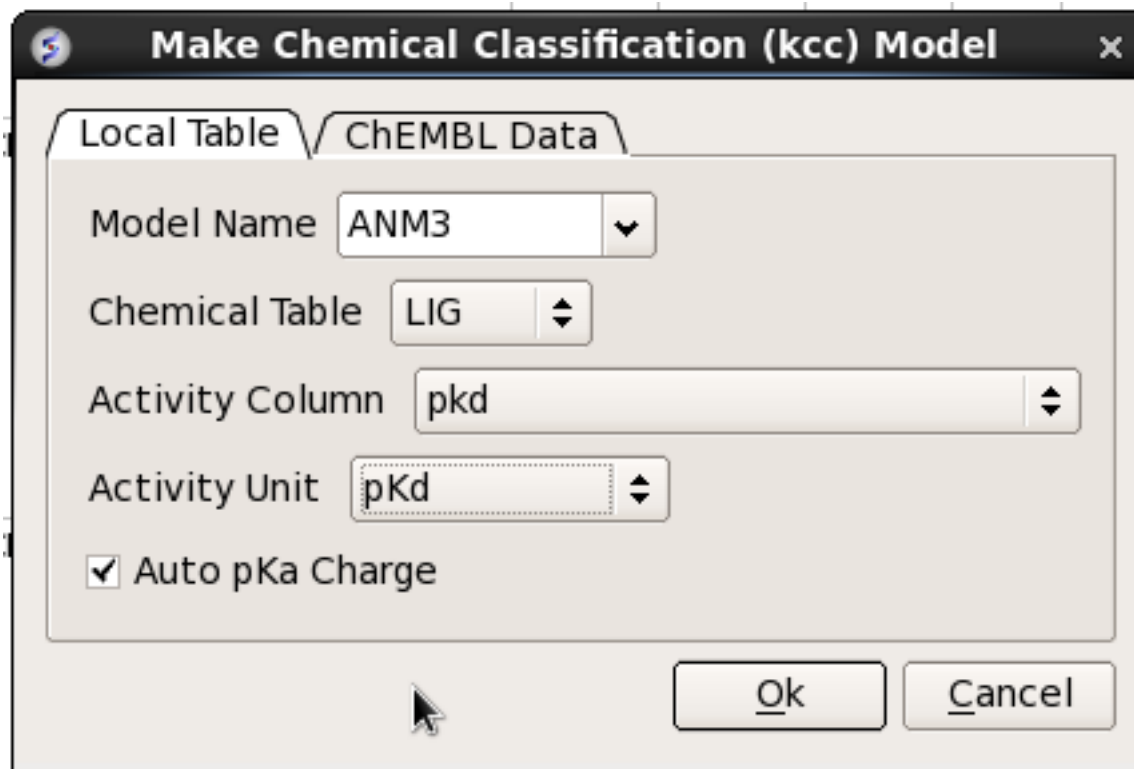
Training:

- Cluster Actives by fingerprint
- Add 40k ChEMBL actives decoy
- Kernel function to each cluster -> probability score (**kcc/MolClass Score**)
- Partial Least Square Regression for each cluster + Kernel Regression (**kca/MolpKd Score**)
- **MolScore**: combine **MolpKd** and **MolSimilarity** to known binders



Make Custom kcc Model

- Input: 2D table w/ Activity column (pKd/uM,nM etc)



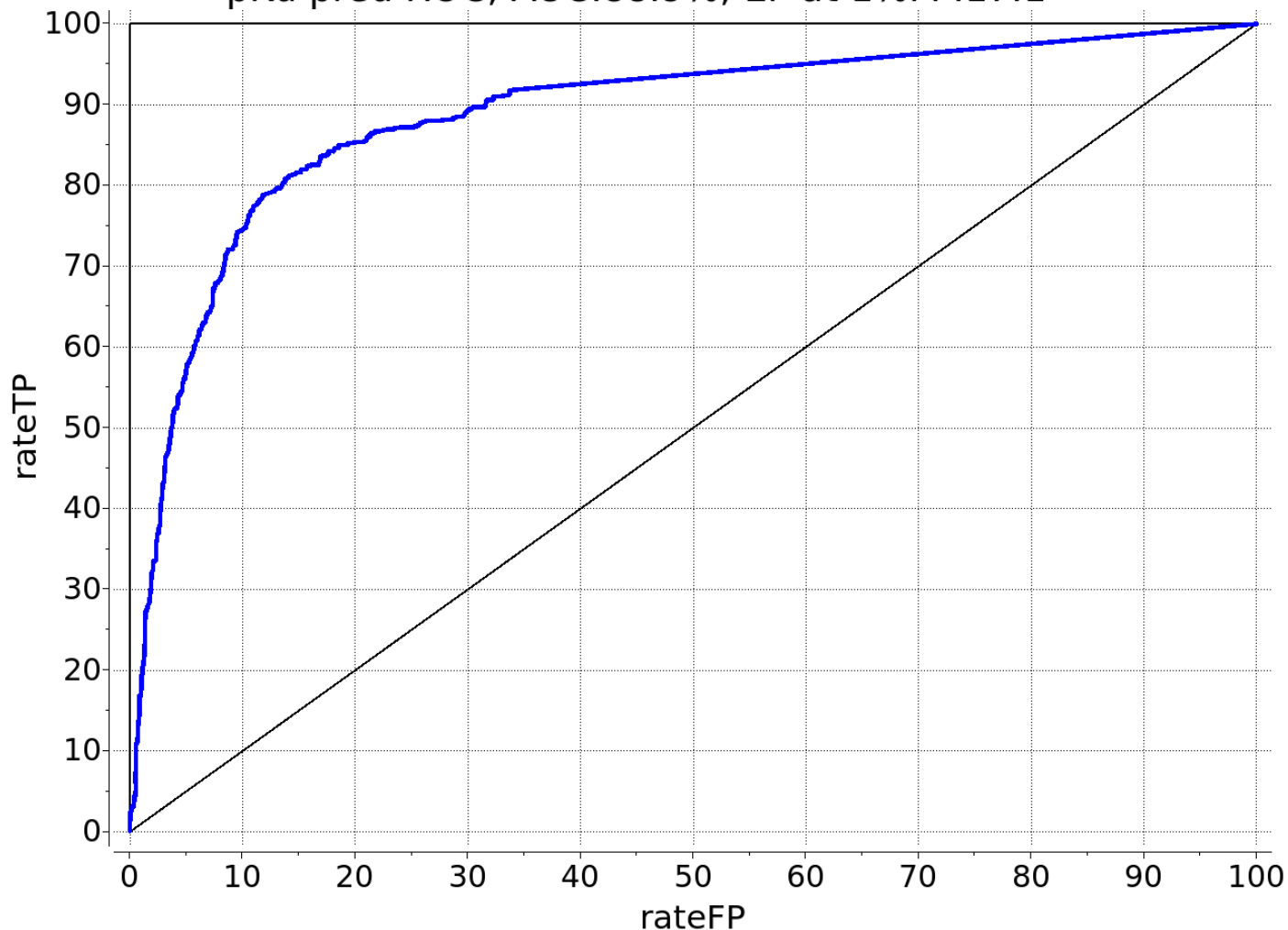
2D QSAR/Fingerprint (kcc) Output

T_tmp/msLigandModel		mol	chk	Target	MolScore	MolPkd	MolPkd Error	MolClass	MolSim	modelnearestCpd	predictType	modelAUC	modelpKdAUC	modelRMSE	modelQ2	Target Name
5		<input type="checkbox"/>	KCNH2	8.0	8.4	1.1	0.94	0.96		kcc	87.10	88.86	0.7	0.48	Potassium voltage-gated channel member 2	
6		<input type="checkbox"/>	KCNH2	7.6	8.3	1.1	0.34	0.94		kcc	87.10	88.86	0.7	0.48	Potassium voltage-gated channel member 2	
7		<input type="checkbox"/>	KCNH2	7.6	8.4	1.1	0.04	0.94		kcc	87.10	88.86	0.7	0.48	Potassium voltage-gated channel member 2	
8		<input type="checkbox"/>	KCNH2	8.6	8.8	1.1	0.97	0.98		kcc	87.10	88.86	0.7	0.48	Potassium voltage-gated channel member 2	
9		<input type="checkbox"/>	KCNH2	6.7	7.3	1.1	0.02	0.94		kcc	87.10	88.86	0.7	0.48	Potassium voltage-gated channel member 2	
10		<input type="checkbox"/>	KCNH2	7.4	8.7	1.1	0.03	0.90		kcc	87.10	88.86	0.7	0.48	Potassium voltage-gated channel member 2	
		<input type="checkbox"/>	KCNH2	7.3	8.7	1.1	0.19	0.90		kcc	87.10	88.86	0.7	0.48	Potassium voltage-gated channel member 2	

Performance 2D QSAR/fingerprint

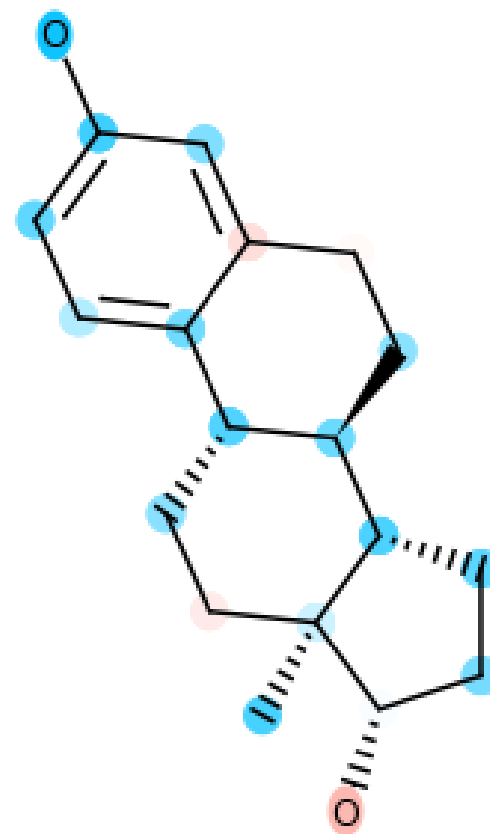
hERG: External test set vs decoy: 3505 compounds

pKd pred ROC, AUC:88.9%, EF at 1%FP:17.1



2D QSAR/Fingerprint (**eca**)

- **eca**: **E**xtended Kernel **C**hemical fingerprint **A**ctivity
- Currently: 409 mammalian models
- Training set: ChEMBL Ki, IC50, EC50, Drugbank assignment
- Median size: 211 ligands
- **All Models' Validation: 25% of ChEMBL set**
- Median external Q^2 : 0.65
- Median external AUC: 95%



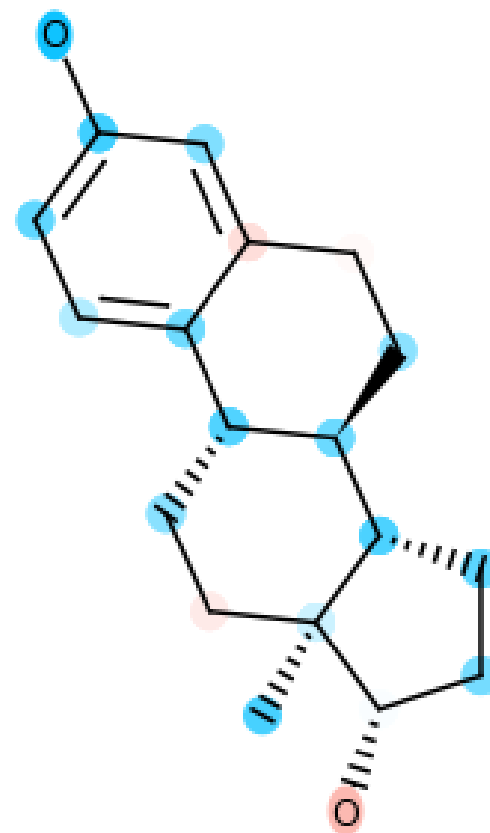
2D QSAR/Fingerprint (**eca**) Method

Differences between kcc and eca Method:

- ChEMBL coverage for some targets might be spotty
- kcc only use data from that target
- eca use data from related targets
- kcc has lower FP rate, lower sensitivity for some not well covered targets
- eca has higher sensitivity, higher FP rate

Training:

- Find all related targets
- Kernel Regression (**MolpKd Score**)
- **MolScore**: combine **MolpKd** and **MolSimilarity** to known binders



3D Atomic Property Field (**dfz**)

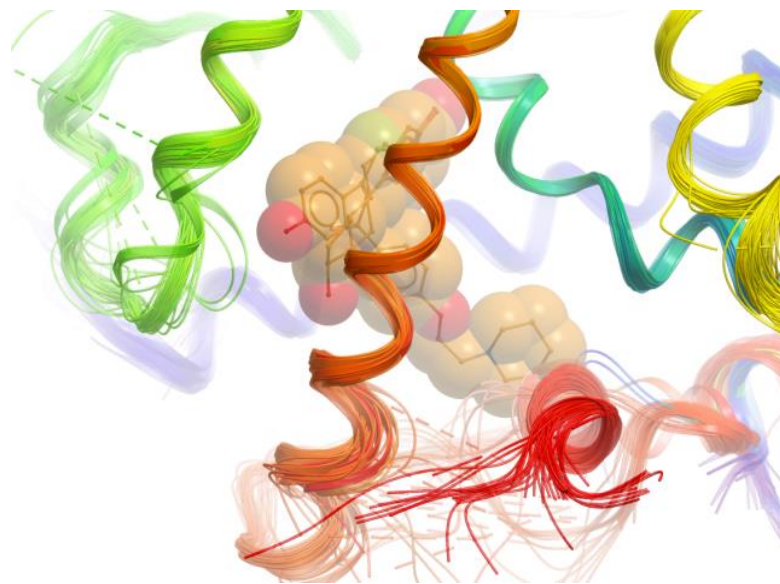
- **dfz**: Docking to ligand **F**ield **Z**-score prediction model
- Currently: 504 mammalian models
- Pocketome ligands/custom alignment as APF template
- ChEMBL cpds for validation
- Median AUC: 92%, 139 cpds vs decoy
- Superseded by superior **dfa** and **dpc** models
- **dfz** as backup when ligand data is insufficient



Giganti, D. *et al.* Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment. *J Chem Inf Model* **50**, 992–1004 (2010).

Docking/3D QSAR (**dpc**) model

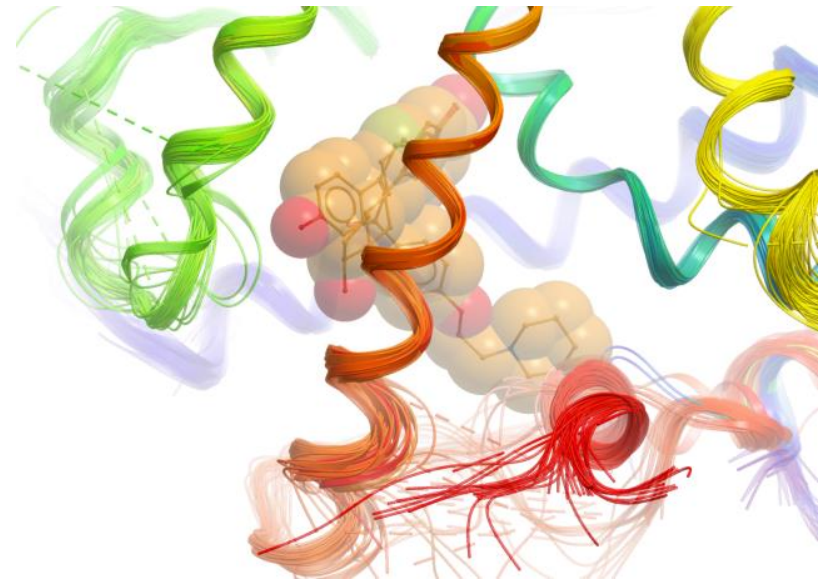
- **dpc**: **D**ocking to **P**ocket **C**lassification/Activity
- Currently: 343 mammalian models w/ AUC > 80%
- Training set: ChEMBL Ki, IC50, EC50, Drugbank assignment
- Median size: 307 ligands
- Median external Q^2 : 0.53
- Median external AUC: 95%



Docking/3D QSAR (**dpc**) Method

Training:

- Pocketome -> Clustering of pocket residues
- 4D Docking w/ co-crystallized ligand as APF template
- Docking Score -> Probability score (**dpc/MolClass** score)
- 3D QSAR training of Activity- > (**dpa/MolpKd**)
- **MolScore**: combine **MolpKd** and **MolSimilarity** to known binders



Make Custom dpc Model

- From Either: 1. Docking Project; 2. Protein object (+Pocketome); 3. Pure Pocketome

Make Docking/SAR (dpc) Model

From Docking Project | From Receptor Object | From Pocketome Entry

Model Name: ANM3

Project Directory: /home/pololam/icm/project/Nov

Receptor object: Graphical Selection (2 obj)

Optional Pocketome Entry: ANM3_HUMAN_209_531

Chemical Table: LIG

Activity Column: mean

Activity Unit: pKd

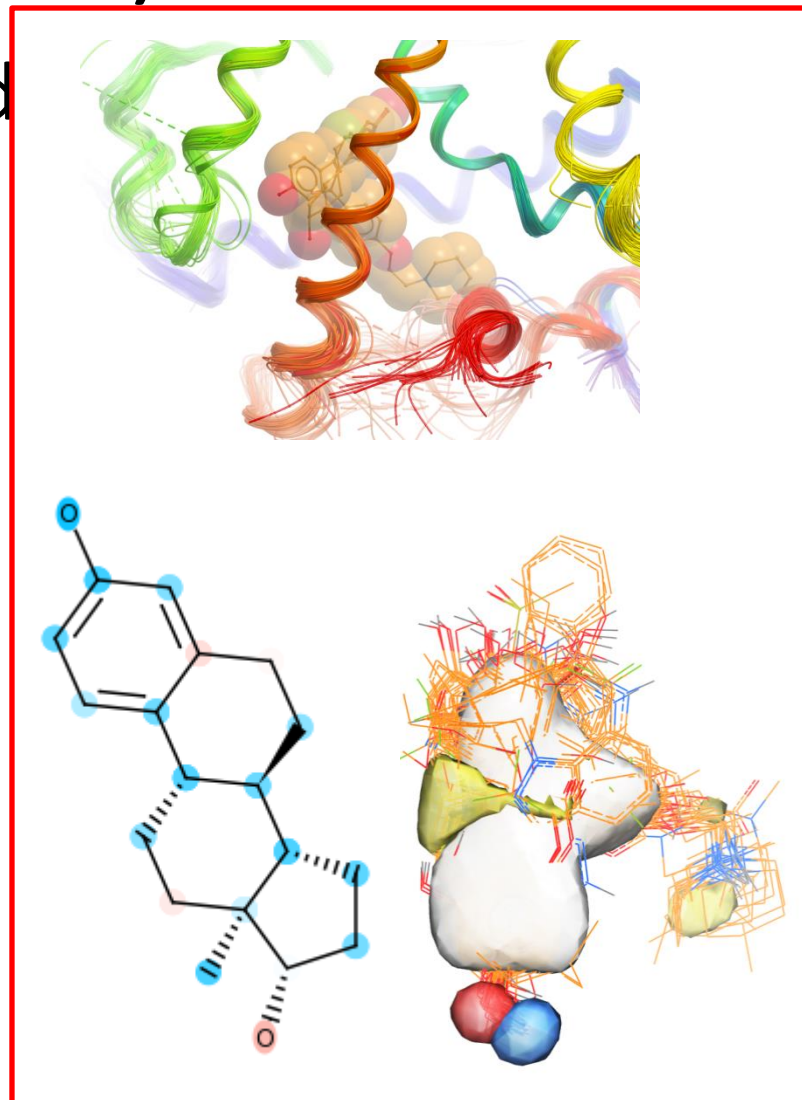
Auto pKa Charge

Hint

If optional Pocketome entry is specified, selected representative will be added to user supplied receptor object
Docking project and dpc model will be written in the Project Directory, previous version will be overwritten

APF/3D QSAR (**dfa**) Model

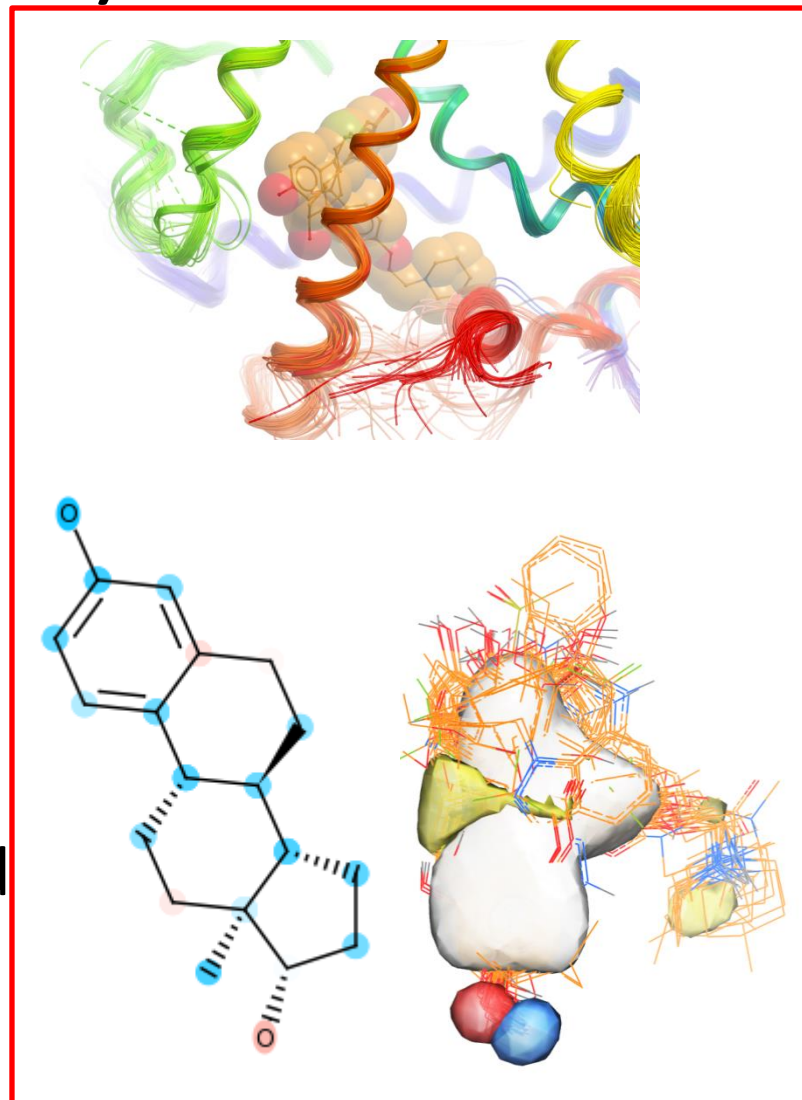
- **dfa**: **D**ocking to Ligand **F**ield Classification/**A**ctivity
- Currently: 612 mammalian models w/ AUC > 80%
- Training set: ChEMBL Ki, IC50, EC50, Drugbank assignment
- Median size: 270 ligands
- Median external Q^2 : 0.65
- Median external AUC: 96%



APF/3D QSAR (**dfa**) Method

Training:

- Also from Pocketome -> 4D Docking + Ligand APF template
- Cpd align to ligand template -> cluster by 3D poses
- APF Score -> Probability Score (**dfc/MolClass** score)
- 3D-QSAR training for each cpd cluster (**dfa/MolpKd** score)
- **MolScore**: combine **MolpKd** and **MolSimilarity** to known binders



Make Custom dfa Model

- Either: 1. 2D mol-> Align to 3D poses 2. Docking Project/Protein Object/Pocketome

Make APF/SAR (dfa) Model

From Ligand 3D Poses | From Docking Project | From Receptor Object | From Pocketome Entry

Model Name: ANM3

Project Directory: /home/pololam/icm/project/Novi [Browse]

Chemical Table: LIG

Activity Column: mean

Activity Unit: pKd

Auto pKa Charge

Hint

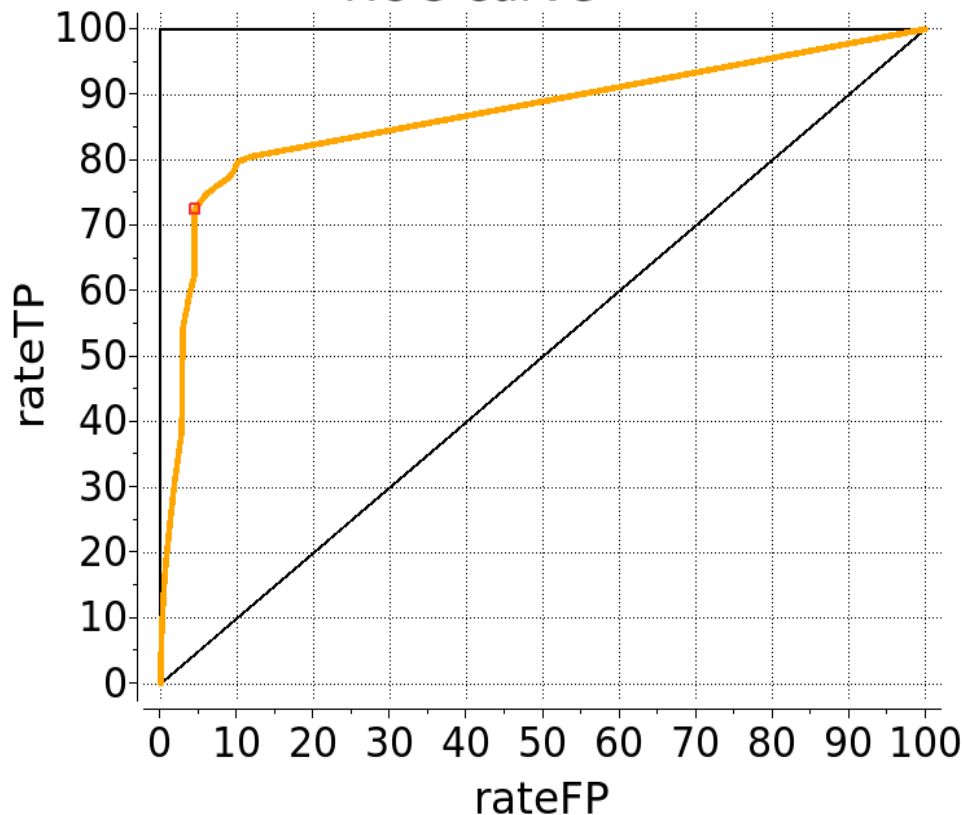
If Ligand table is 3D, poses will be used as ligand template.
If it is 2D, ligand will be aligned in 3D first using APF method
dfa model will be written in the Project Directory, previous version will be overwritten

Ok Cancel

Improving MolpKd: MolScore

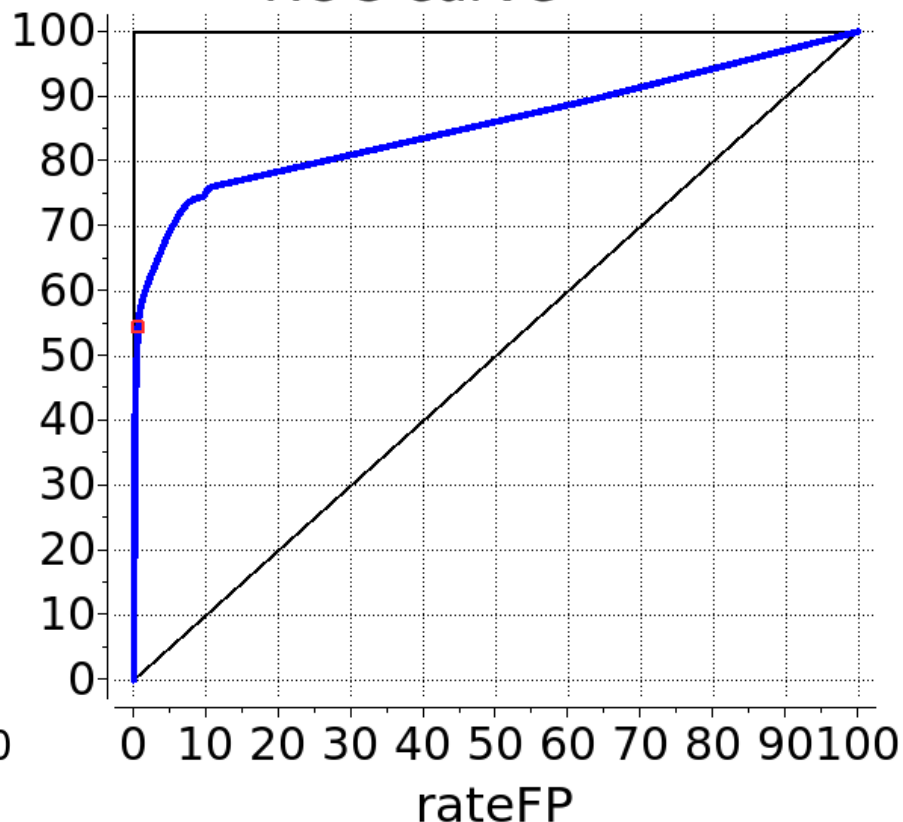
(3.4M approved drugs – Model pairs)

ROC curve



MolpKd, cutoff: 5., top 5%
Sensitivity: 72%, Precision 13%

ROC curve



New MolScore, cutoff: 3., top 1%
AUC: 86%, NSA: 69%
Sensitivity: 55%, Precision 45%

Usage Consideration

- **kcc/eca** fingerprint model:
 - Very fast (thousands of cpds in min)
 - Highly accurate if Tanimoto Similarity ≤ 0.2
- **dfa** APF/3D-QSAR model:
 - Accuracy extend beyond fingerprint similarity
 - Flexible, w/ or w/o protein structure
- **dpc** Docking/3D-QSAR model:
 - Accurate Docking pose due to 4D docking w/ Ligand APF template
 - Rationalize Ligand/Pocket interactions

Custom Learning Models Considerations

- Learn Global 2D **kcc** model
 - Suitable for differentiating actives from random cpds due to added decoy
- Learn Local 4D/2D **dfa/dpc** model
 - Suitable for improving SAR series
 - Local model
 - Shorten training time
 - Might not differentiate against random cpds