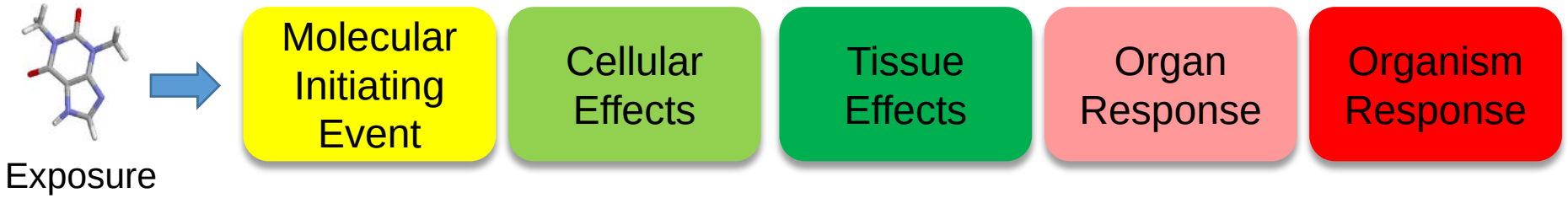
Using BigData to Build QSAR Models for Neurological Proteins

Yaroslav Chushak, PhD
Sr. Scientist
Henry M. Jackson Foundation
711 HPW/RHMO
Air Force Research Laboratory

Adverse Outcome Pathway



- Neurotoxicity - adverse effect on the functioning of the nervous system
- Neurotoxicity of chemicals depends on their interaction with neurological targets

ToxCast *vs* ChEMBL Database

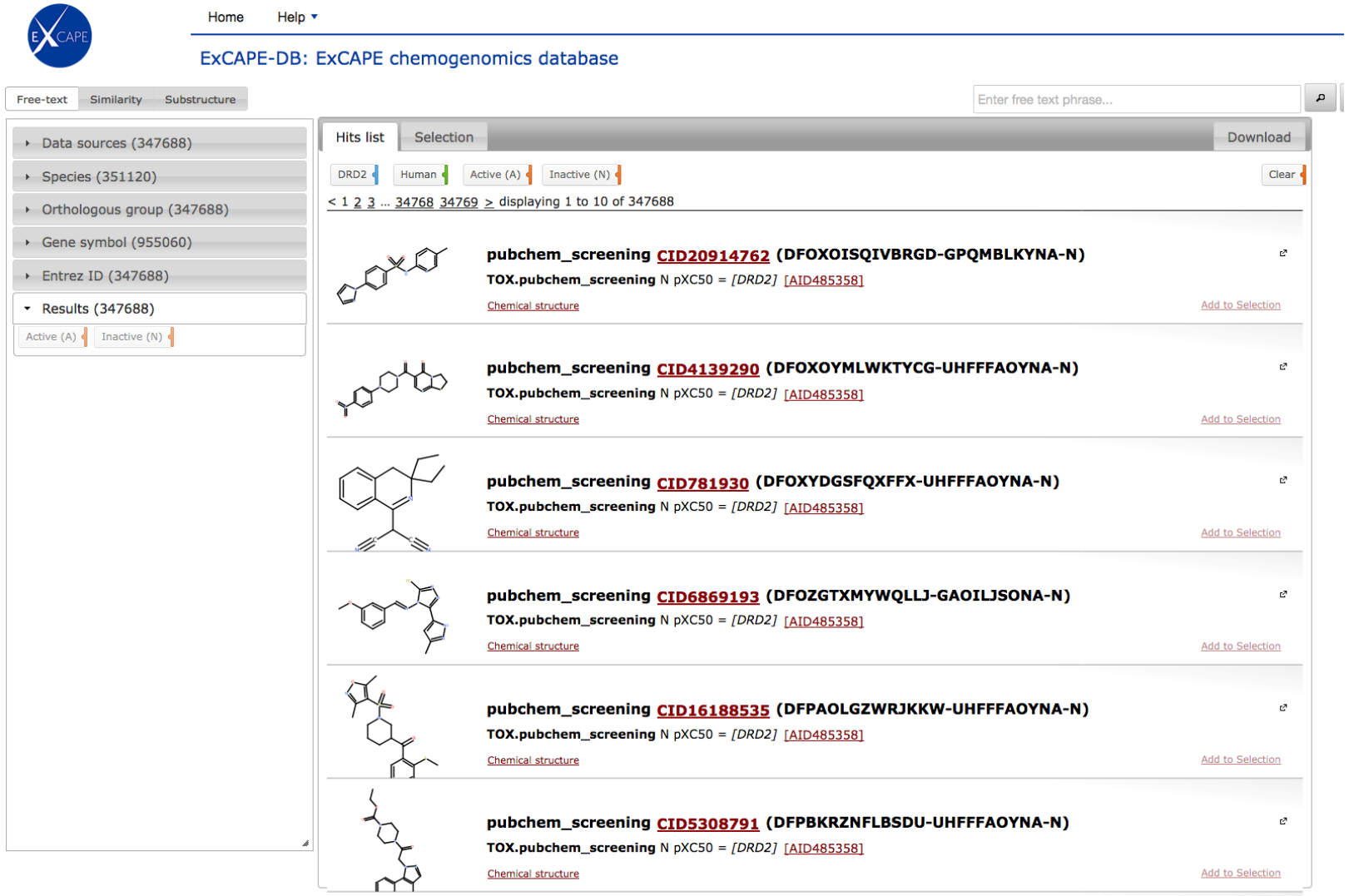
- In many cases ChEMBL has x100 active compound comparing with ToxCast

- ToxCast data are not integrated into ChEMBL database

	Gene	Protein Name	UniProt	#Active ToxCast	#Active ChEMBL
2	ACHE	Acetylcholinesterase	P22303	87	6823
3	ADORA1	Adenosine receptor A1	P30542	95	8027
4	ADORA2A	Adenosine receptor A2a	P29274	66	8730
5	ADRA1A	Alpha-1A adrenergic receptor	P35348	65	2206
6	ADRA1B	Alpha-1B adrenergic receptor	P35368	51	1927
7	ADRA2A	Alpha-2A adrenergic receptor	P08913	74	2024
8	ADRB1	Beta-1 adrenergic receptor	P08588	69	2951
9	ADRB2	Beta-2 adrenergic receptor	P07550	40	4777
10	BCHE	Butyrylcholinesterase	P06276	82	3199
11	CACNA1A	Voltage-dependent P/Q-type calcium channel subun	O00555	94	3
12	CACNA1B	Voltage-dependent N-type calcium channel subunit a	Q00975	2	515
13	CHRM1	Muscarinic acetylcholine receptor M1	P11229	54	4598
14	CHRM2	Muscarinic acetylcholine receptor M2	P08172	75	3656
15	CHRM3	Muscarinic acetylcholine receptor M3	P20309	85	3532
16	CHRM4	Muscarinic acetylcholine receptor M4	P08173	81	2382
17	CHRM5	Muscarinic acetylcholine receptor M5	P08912	60	2231
18	CHRNA2	Neuronal acetylcholine receptor subunit alpha-2	Q15822	26	17
19	CHRNA7	Neuronal acetylcholine receptor subunit alpha-7	P36544	31	1198
20	DRD1	D(1A) dopamine receptor	P21728	107	3096
21	DRD2	D(2) dopamine receptor	P14416	72	9594
22	DRD4	D(4) dopamine receptor	P21917	57	3979
23	GABBR1	Gamma-aminobutyric acid type B receptor subunit 1	Q9UBS5	4	24
24	GABRA1	Gamma-aminobutyric acid receptor subunit alpha-1	P14867	58	118
25	GABRA5	Gamma-aminobutyric acid receptor subunit alpha-5	P31644	8	66
26	GLRA1	Glycine receptor subunit alpha-1	P23415	5	118
27	GRIA1	Glutamate receptor 1	P42261	14	128
28	GRIK1	Glutamate receptor ionotropic, kainate 1	P39086	4	260
29	GRIN1	Glutamate receptor ionotropic, NMDA 1	Q05586	21	459
30	GRM1	Metabotropic glutamate receptor 1	Q13255	15	1359
31	GRM5	Metabotropic glutamate receptor 5	P41594	2	3520
32	HTR1A	5-hydroxytryptamine receptor 1A	P08908	58	5198
33	HTR2A	5-hydroxytryptamine receptor 2A	P28223	47	6156
34	HTR2C	5-hydroxytryptamine receptor 2C	P28335	68	4992

ExCAPE-DB

"ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics", Sun et al. *J. Cheminform.* (2017) 9:17



ExCAPE-DB Data Curation

Chemical structure standardization

- fragment splitting
- isotope removal
- stereochemistry
- tautomer generation
- neutralization

Bioactivity data standardization

- single protein target
- active if concentration < 10 μM
- organic compounds (without metals)
- molecular weight < 1000 Da
- number of heavy atoms > 12 (remove small and inorganic compounds)
- multiple activity data for the same chemical-target pair are aggregated with maximal potency as final value

Studied Proteins

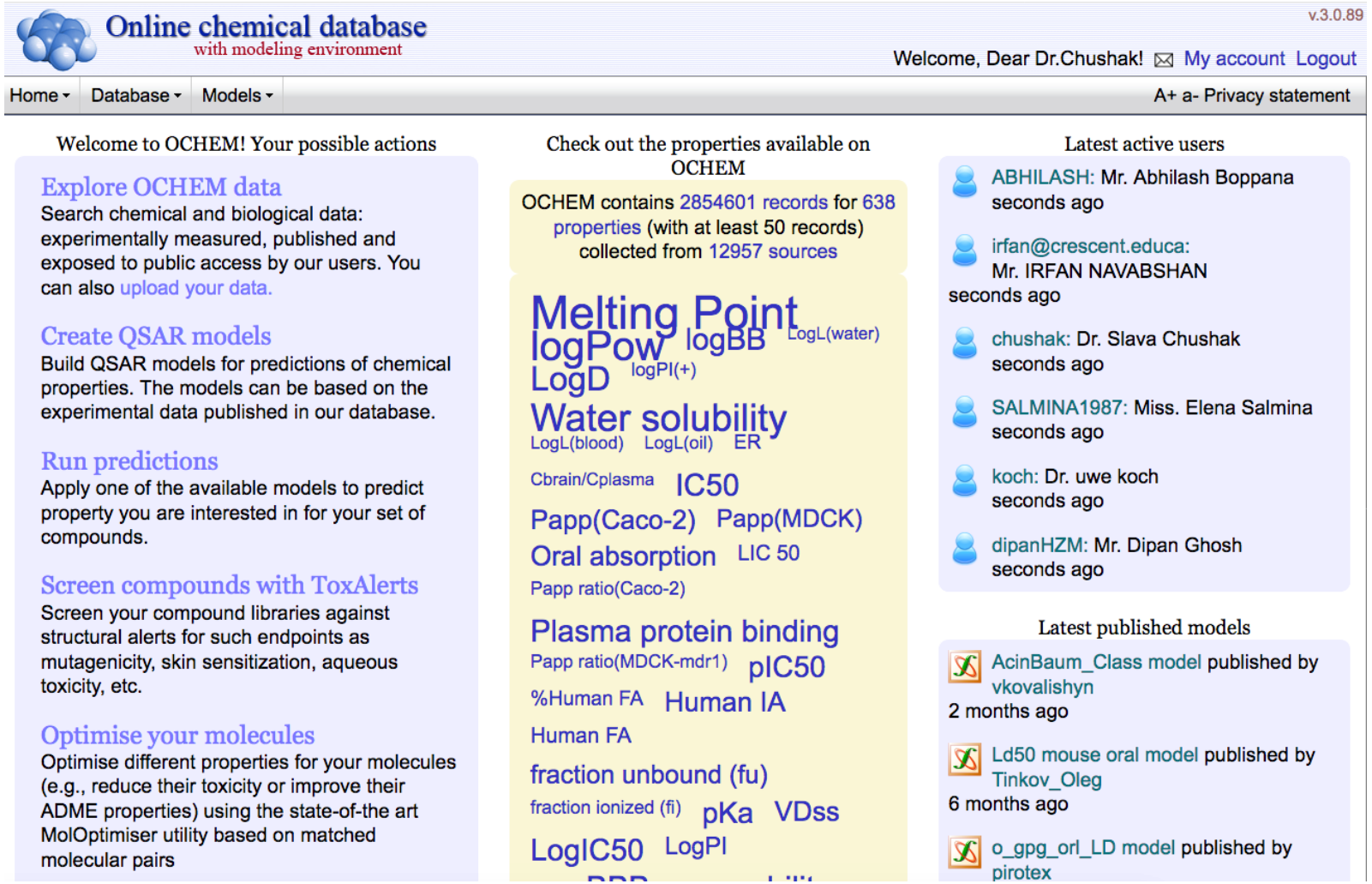
	Gene	Name	Protein Classification	Actives	Inactives
2	ACHE	Acetylcholinesterase	Esterase	2698	3014
3	ADORA1	Adenosine A1 receptor	GPCR	2955	2010
4	ADORA2A	Adenosine A2a receptor	GPCR	3523	3640
5	ADRA1A	Alpha-1a adrenergic receptor	GPCR	1220	1237
6	ADRB2	Beta-2 adrenergic receptor	GPCR	1075	1114
7	CACNA1B	Voltage-gated N-type calcium channel alpha-1B	Ion_Channel	677	699
8	CHRM1	Muscarinic acetylcholine receptor M1	GPCR	1578	1750
9	DRD1	Dopamine D1 receptor	GPCR	1733	2080
10	DRD2	Dopamine D2 receptor	GPCR	4612	6678
11	GRM4	Metabotropic glutamate receptor 4	GPCR	630	619
12	HTR1A	Serotonin 1a (5-HT1a) receptor	GPCR	2967	2722
13	OPRD1	Deltha-type opioid receptor	GPCR	2215	2783
14	OPRK1	Kappa opioid receptor	GPCR	3279	3407
15	OX1R	Orexin receptor 1	GPCR	2553	3117
16	SCN9A	Sodium channel protein type IX alpha subunit	Ion_Channel	3388	3213
17	SLC6A3	Sodium-dependent dopamine transporter	Transporter	2357	2667
18	SLC6A4	Sodium-dependent serotonin transporter	Transporter	1395	1039

Two datasets were generated:

- approximately equal number of Active and Inactive compounds
- activity values for active compounds

Online Chemical Modeling Environment

QSAR modeling was performed using OCHEM web-server

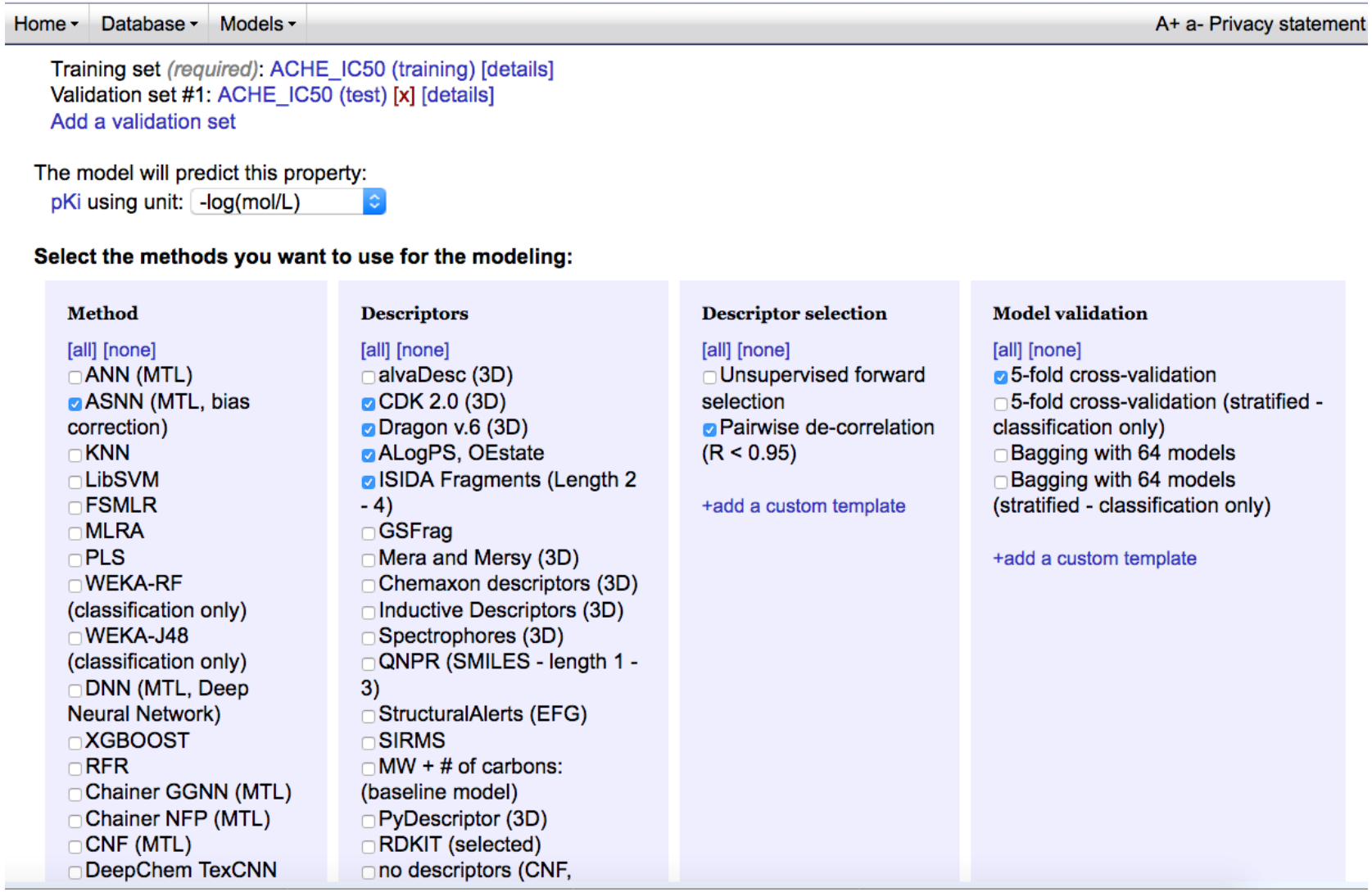


Datasets for QSAR Modeling

Home | Database | Models | A+ a- Privacy statement

adora2a_Class_training.csv	6354 records	8 models
ADORA2A_pKi (test)	352 records	
ADORA2A_pKi (training)	3171 records	4 models
ADORA2A_pKi	3523 records	
ADRB2_pKi (test)	108 records	
ADRB2_pKi (training)	967 records	5 models
ADRB2_pKi	1075 records	
ADRA1A_pKI (test)	122 records	
ADRA1A_pKI (training)	1098 records	5 models
scn9a_Class_test.csv	695 records	
scn9a_Class_training.csv	6261 records	8 models
ADRA1A_pKI	1220 records	
SBP_Phase1_SMILES.csv	68 records	
OPRD1_pKi (test)	213 records	
OPRD1_pKi (training)	1913 records	4 models
OPRD1_pKi	2126 records	
SBP_Phase1.csv	68 records	
DRD1_pKi (test)	157 records	
DRD1_pKi (training)	1414 records	5 models
DRD1_excape_Act.csv	1725 records	1 models

Methods for QSAR Modeling



QSAR Statistical Parameters

Regression models

In the formulae below, $\tilde{y}_i$ and y_i denote predicted and real values of the predicted property for *i*-th compound in the set. $E(\tilde{y})$ and $E(y)$ are the means of the predicted and real property values; s (y) denotes the standard deviation.

Root Mean Square Error

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\tilde{y}_i - y_i)^2}{N}}$$

R^2 (Pearson correlation coefficient)

$$r^2 = \frac{\sum_{i=1}^{N} (\tilde{y}_i - E(\tilde{y}))(y_i - E(y))}{\sigma(\tilde{y}) \cdot \sigma(y)}$$

Classification models

TP – true positive
FP – false positive
TN – true negative
FN – false negative

Accuracy

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

Balanced accuracy

$$BA = \frac{1}{2}\left(\frac{TP}{TP + FP} + \frac{TN}{TN + FN}\right)$$

Classification Models for ACHE

Predicted property: Classifier
Training set: ache_Class_training.csv

Metrics Balanced accuracy for Validation set Validation: Cross-Validation (8 models)

	ASNN	WEKA-RF
ALogPS, OEstate	95	97
CDK2 (constitutional, topological, geometrical, electronic, ...	93	92
ChemaxonDescriptors (pH 0 - 14:1)	93	94
Fragmentor (Length 2 - 4)	95	96

Overview

Model name: Classifier_WEKA-RF_[Fragmentor (Length 2 - 4)] [rename]
Temporal Public ID: 32886279 - use this link to share the model

Predicted property: **Classifier** modeled in CLASS
Training method: WEKA-RF

[Fragmentor (Length 2 - 4)]
Correl. limit: 0.95 Variance threshold: 0.01,
Maximum value: 999999,
trees=128, features=0, depth=0
5-fold cross-validation
-
866 pre-filtered descriptors
trees=128, features=0, depth=0

Calculated in 244 seconds
Size: 2466 Kb

Data Set	#	Accuracy	Balanced Accuracy	MCC	AUC
Training set: ache_Class_training.csv	4599 records	94.9% ± 0.3	94.9% ± 0.3	0.898 ± 0.006	0.949 ± 0.003
Test set: ache_Class_test.csv [x]	559 records	96.4% ± 0.7	96.3% ± 0.8	0.93 ± 0.01	0.963 ± 0.008

Show ROC curves

Real↓/Predicted→	Inactive	Active	Hit rate
Inactive	2258	89	0.962
Active	146	2106	0.935
Precision	0.939	0.959	

Training (Original)

Real↓/Predicted→	Inactive	Active	Hit rate
Inactive	285	4	0.986
Active	16	254	0.94
Precision	0.95	0.984	

Test (Original)

Regression Models for ACHE

Overview | Applicability domain

Model name: Consensus pKi - 322666 [rename]
Temporal Public ID: 5026460 - use this link to share the model

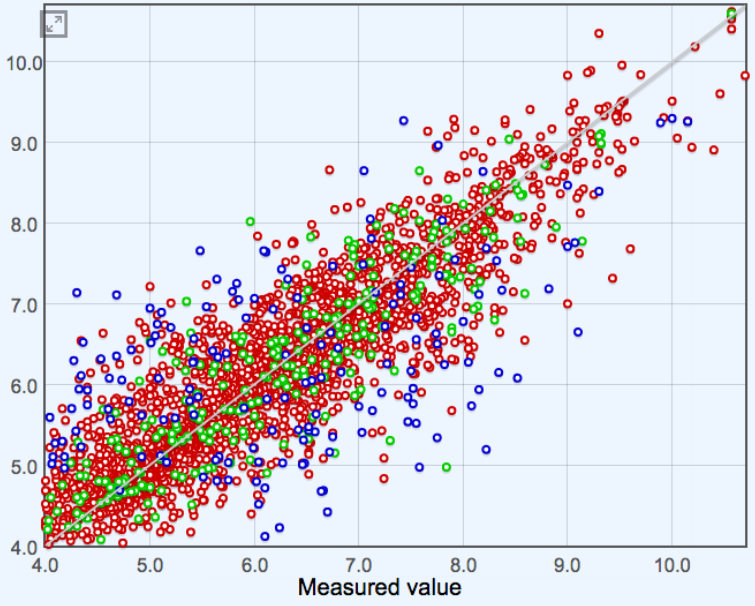
Predicted property: **pKi** modeled in -log(mol/L)
Training method: Consensus

3 individual models:
ModelID: 49875453
ModelID: 22704829
ModelID: 28859546
RMSE_WEIGHTED

No validation

Size: 1 Kb

Data Set	#	R_2	q_2	RMSE	MAE
Training set: ACHE_IC50 (training)	2269 records	0.81 ± 0.008	0.808 ± 0.009	0.57 ± 0.01	0.422 ± 0.008
Test set: ACHE_IC50 (test) [x]	270 records	0.8 ± 0.03	0.8 ± 0.03	0.61 ± 0.04	0.44 ± 0.03
Excluded from training set	159 records	0.23 ± 0.06	0.1 ± 0.1	1.3 ± 0.05	1.14 ± 0.05



Summary Table of QSAR Models

Validation set

Classification | Regression

Target	Actives	Inactives	Acc train	BAcc train	Acc valid	BAcc valid	R2 train	RMSE train	R2 valid	RMSE valid
ACHE	280	289	95	95	96	96	0.81	0.57	0.8	0.61
ADORA1	296	111	99	98	95	91	0.73	0.48	0.67	0.53
ADORA2A	352	373	97	97	98	98	0.7	0.6	0.76	0.56
ADRA1A	122	126	96	96	97	97	0.72	0.67	0.63	0.67
ADRB2	108	104	98	98	99.5	99.5	0.75	0.51	0.82	0.49
CACNA1B	123	226	86	86	88	87	0.74	0.32	0.78	0.35
CHRM1	158	227	96.3	96.2	96.6	96	0.71	0.7	0.73	0.67
DRD1	162	170	95	94	95	95	0.71	0.66	0.65	0.65
DRD2	420	561	98.1	98.1	92.9	94.3	0.65	0.6	0.68	0.56
GRM4	63	45	96.9	96.9	98	99	0.62	0.45	0.73	0.37
HTR1A	297	332	98	98	99	99	0.68	0.63	0.64	0.65
OPRD1	386	332	98.7	98.7	95.1	95.3	0.74	0.68	0.76	0.63
OPRK1	328	401	96.8	96.8	97	96.9	0.75	0.7	0.72	0.78
OX1R	139	168	93	92	93	93	0.86	0.39	0.8	0.43
SCN9A	356	339	99	99	99	99	0.8	0.42	0.8	0.43
SLC6A3	236	264	97.1	97	97.8	97.8	0.72	0.57	0.58	0.73
SLC6A4	140	49	98	98	95	90	0.8	0.53	0.7	0.62

Summary

- Developed and validated classification and regression QSAR models for 17 neurological proteins
- Accuracy for classification models >= 90%
- R2 for regression models (training sets) > 0.62

Acknowledgements

Collaborators:

- Dr. Jeffery Gearhart
- Dr. Heather Pangburn
- Dr. Dirk Yamamoto
- Dr. Darrin Ott

Funding:

- Defense Health Agency under the contract RSAAC 18-089
- 711 Human Performance Wing

Questions?