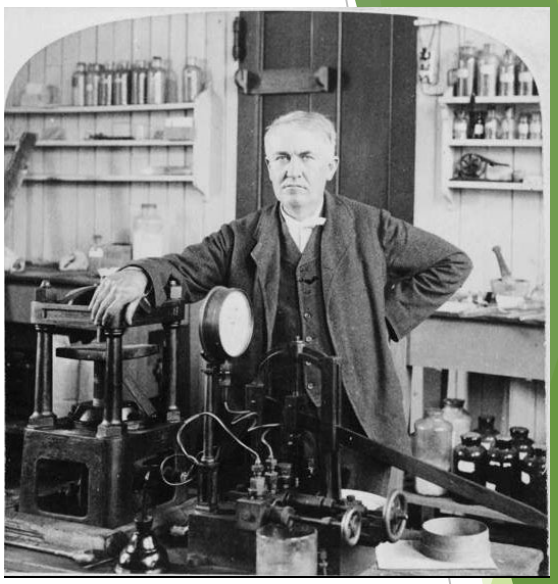
Machine Learning in Toxicology: Fundamentals of Application and Interpretation

Sean Ekins, Ph.D., D.Sc.
Email collaborationspharma@gmail.com
Phone 215-687-1320

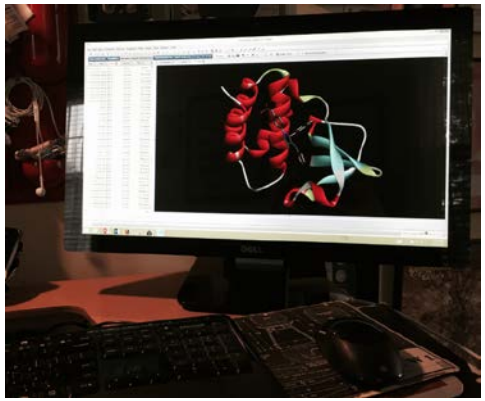
Collaborations Pharmaceuticals, Inc.

Laboratories past and present

Lavoisier's lab 18th C



Edison's lab 20th C



Author's lab 21th C

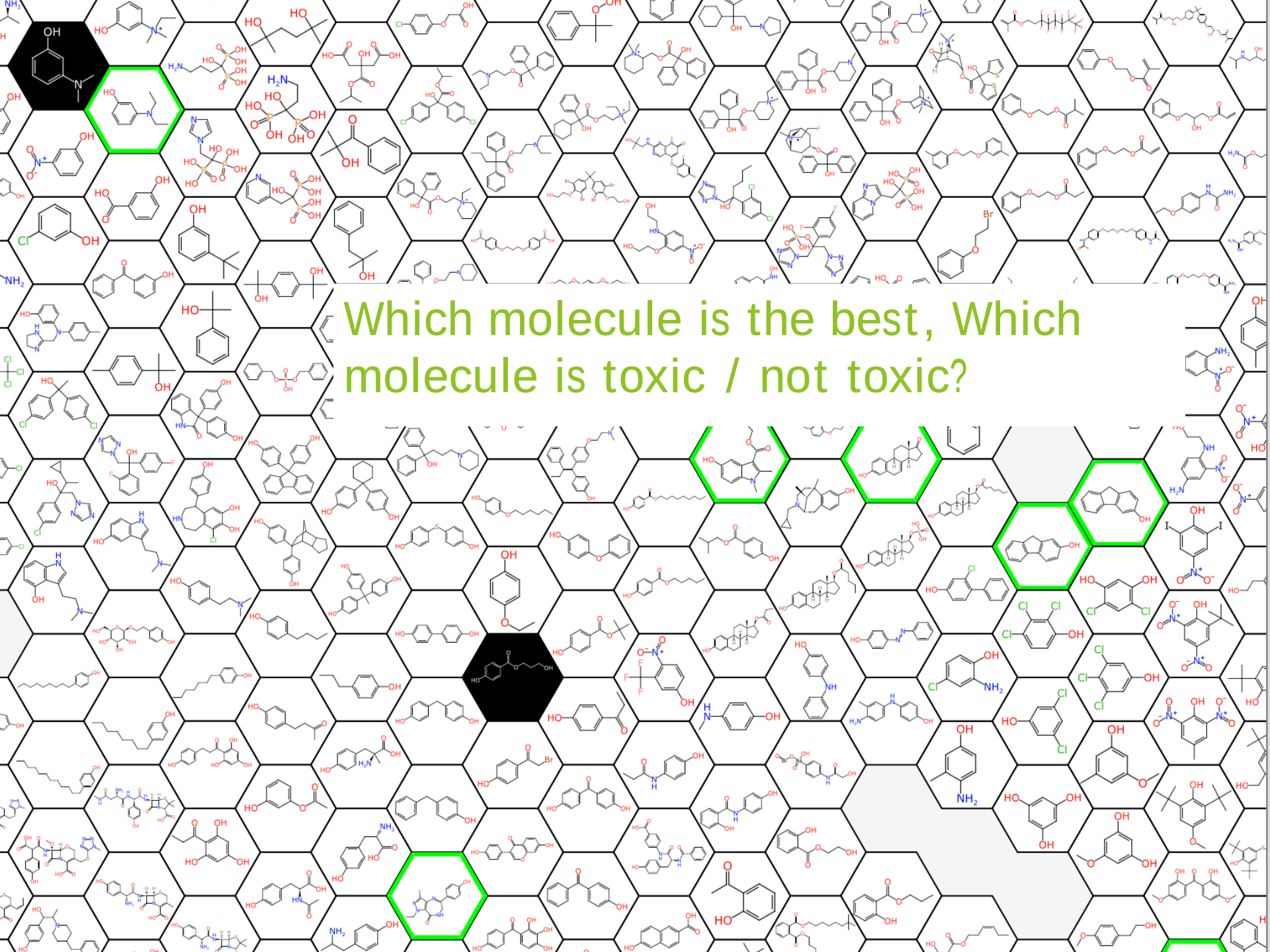
+ Network of global collaborators

What is Cheminformatics?

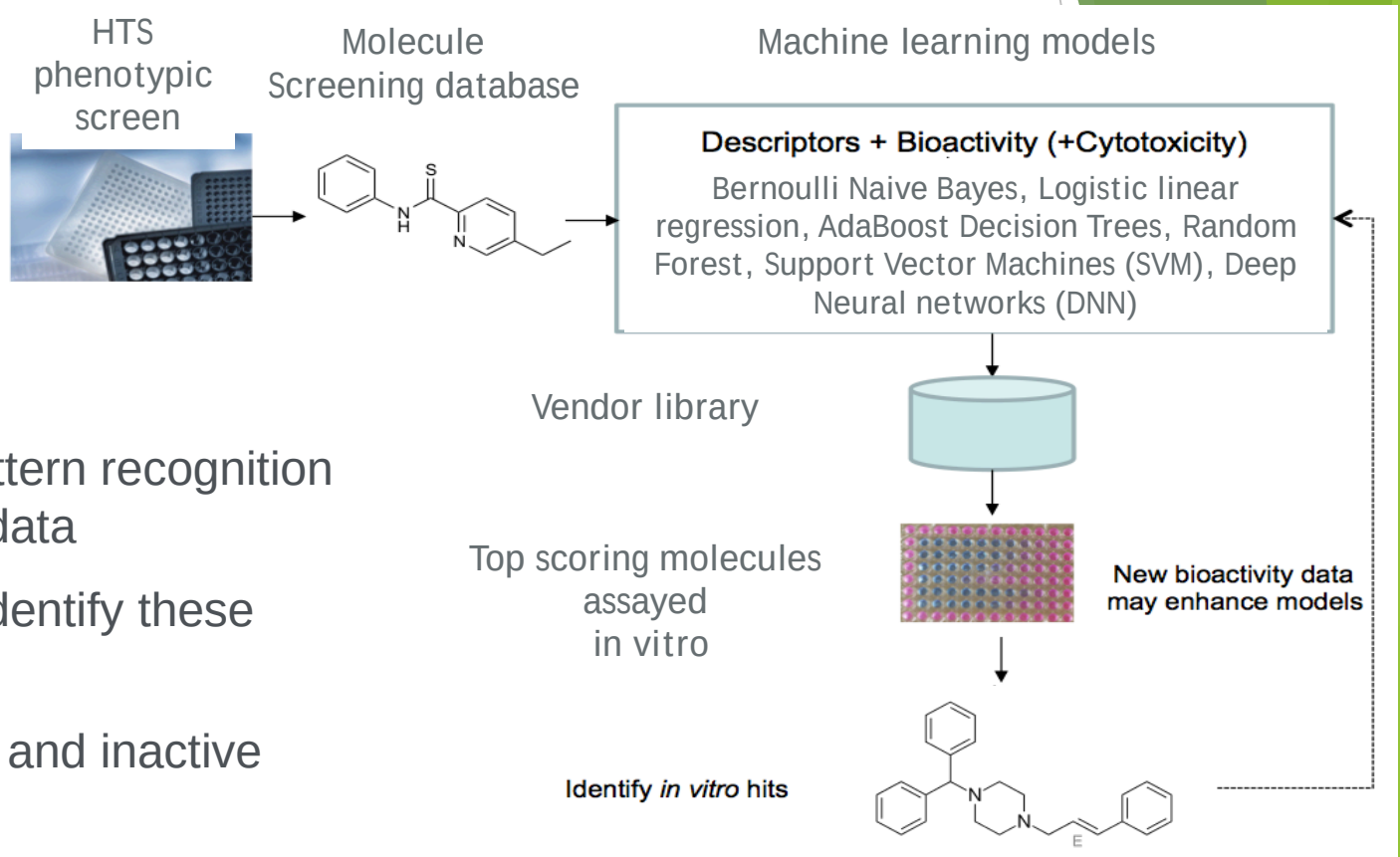
- Cheminformatics combines the scientific working fields of chemistry, computer science and information science

How and why do we use it for drug discovery?

- Learn from data to suggest compounds to make or avoid
- Can increase efficiency / cost effectiveness
- Minimize use of animals and costly materials
- Predict failure


Which molecule is the best, Which molecule is toxic / not toxic?

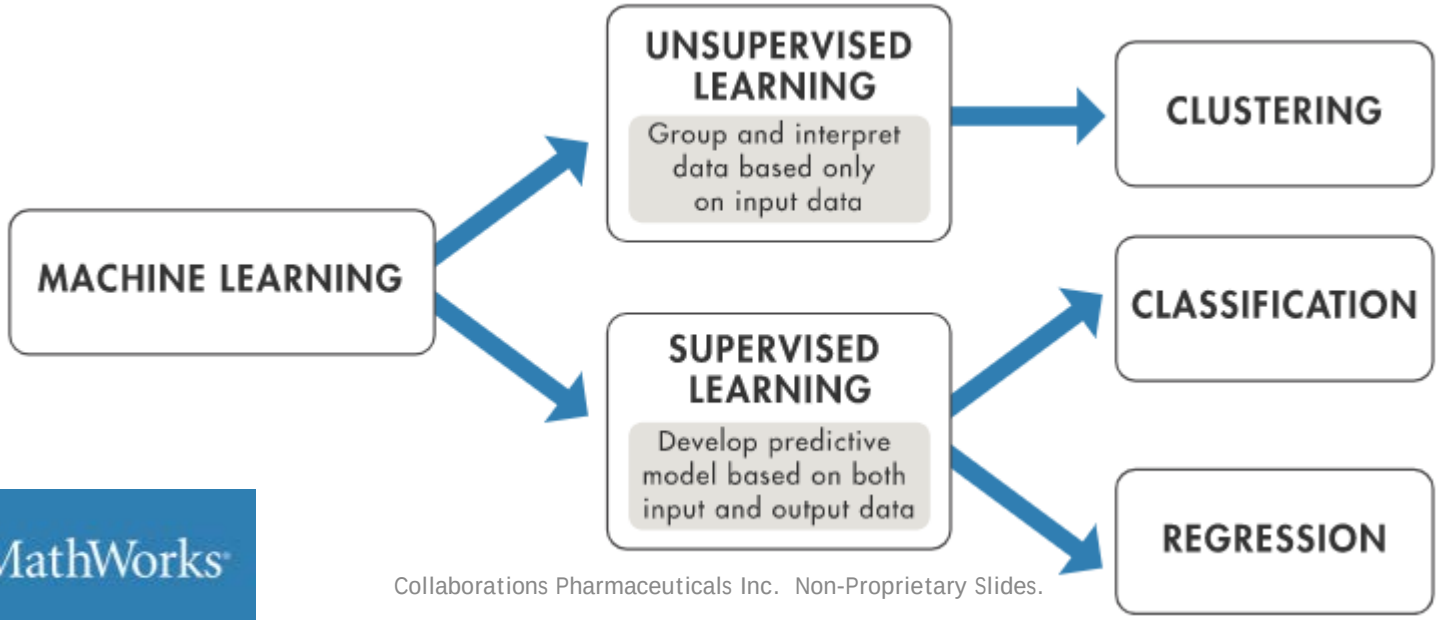
Speeding drug discovery with machine learning



- Molecular pattern recognition of biological data
- Descriptors identify these patterns
- Define active and inactive features
- Used to generate predictions for drug activity at a certain target (organism, protein of interest)

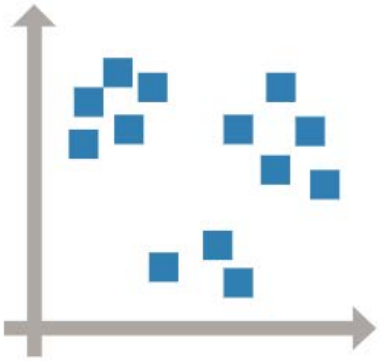
What is Machine Learning?

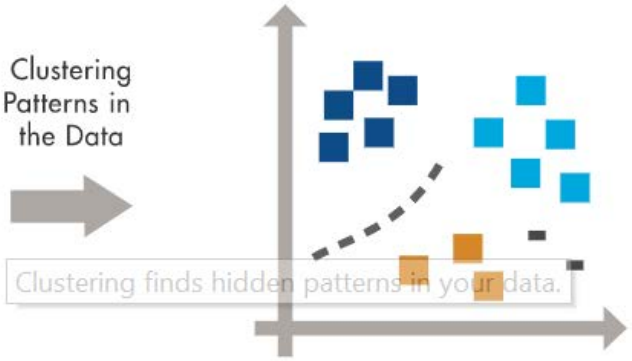
- Find patterns in data to create insights –
- We use examples of the correct output for a given input
- The algorithm learns from this input data
- The program created by the algorithm recognizes the correct response
- It works on objects similar to what it was trained on as well as new examples



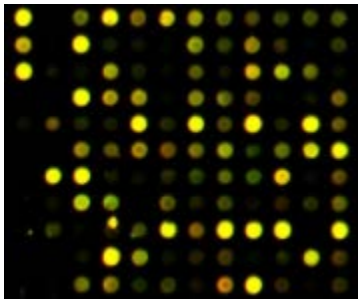
What can we model?

- Data with responses
- Image recognition
- Voice recognition
- Text recognition

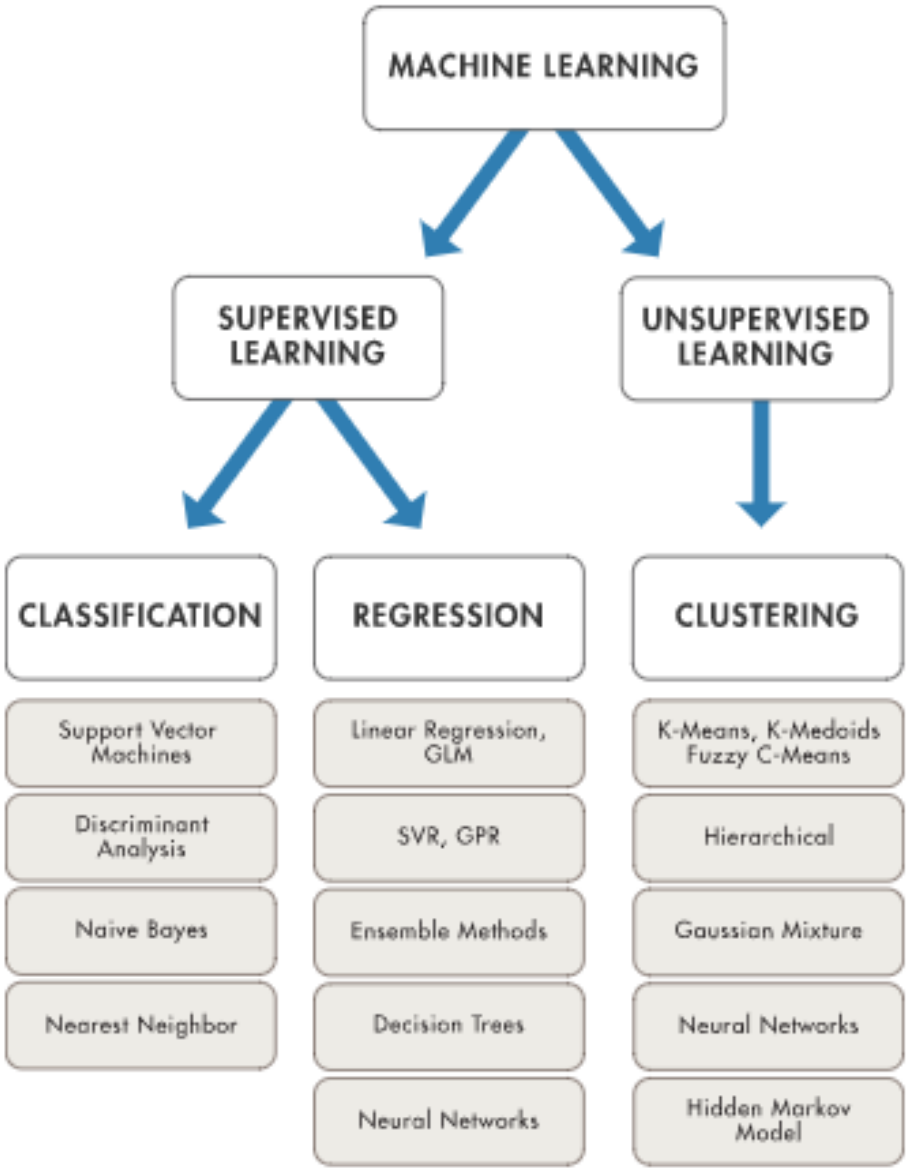




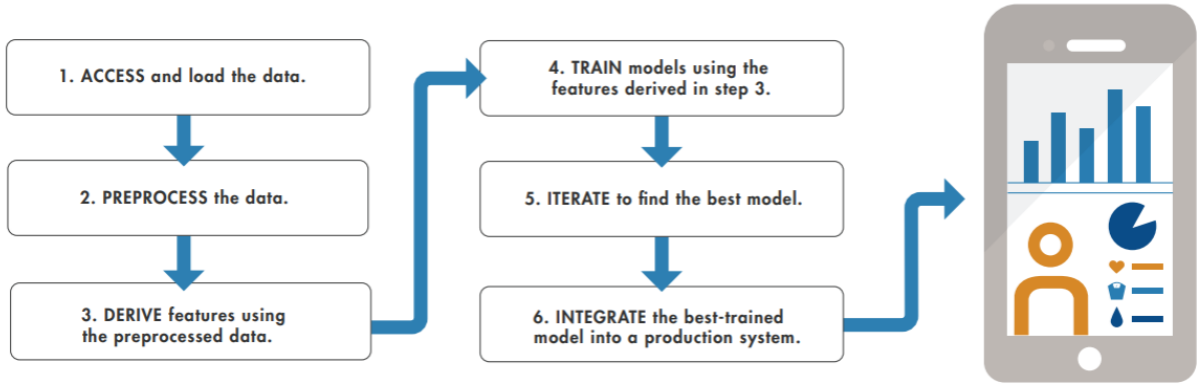
- e.g – molecules tested against a target / disease
- Complex data from genomics, proteomics, metabolomics
- e.g. microarray data for 100s, 1000s of compounds

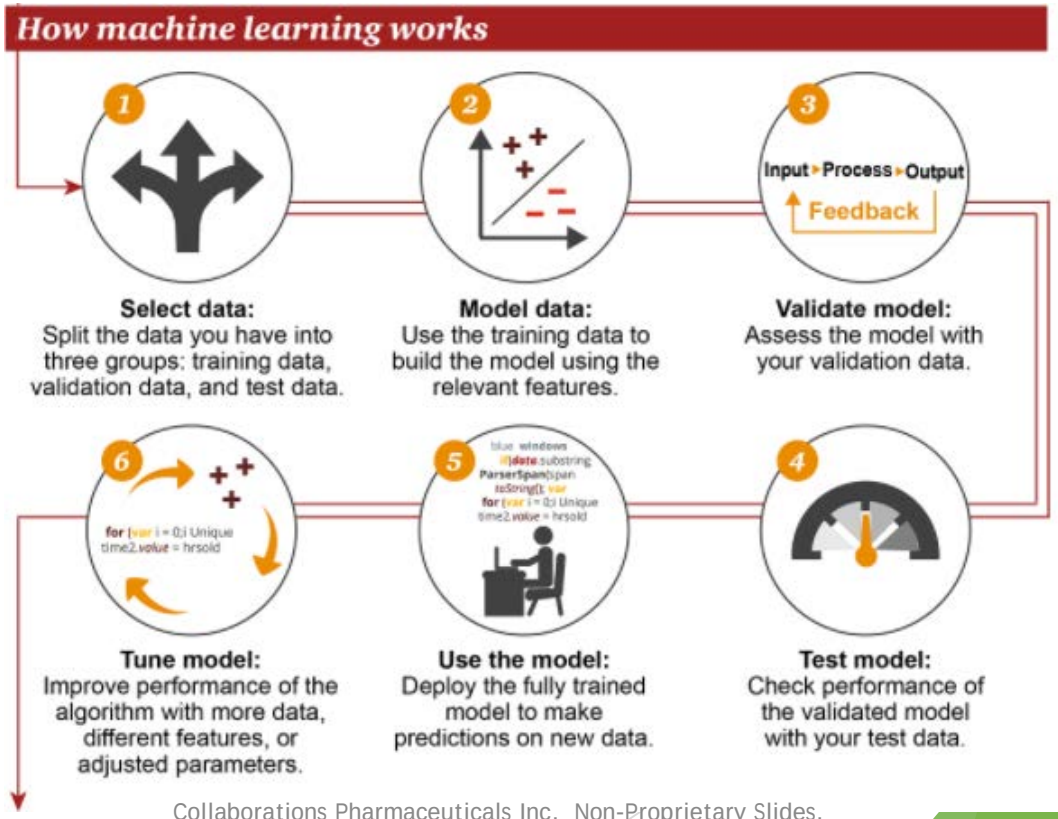


What algorithms do we use?



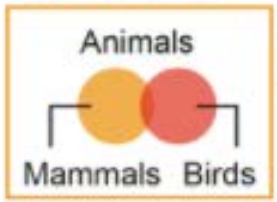
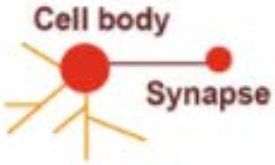
The Workflow





Which Machine Learning Tribe are you?

What are the five tribes?

Symbolists	Bayesians	Connectionists	Evolutionaries	Analogizers
				
Use symbols, rules, and logic to represent knowledge and draw logical inference	Assess the likelihood of occurrence for probabilistic inference	Recognize and generalize patterns dynamically with matrices of probabilistic, weighted neurons	Generate variations and then assess the fitness of each for a given purpose	Optimize a function in light of constraints ("going as high as you can while staying on the road")
Favored algorithm Rules and decision trees	**Favored algorithm** Naive Bayes or Markov	**Favored algorithm** Neural networks	**Favored algorithm** Genetic programs	**Favored algorithm** Support vectors

Source: Pedro Domingos, *The Master Algorithm*, 2015

Bayesian Machine Learning

Bayesian classification is a simple probabilistic classification model. It is based on Bayes′ theorem

$$p(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

h is the hypothesis or model
d is the observed data
p(*h*) is the prior belief (probability of hypothesis *h* before observing any data)
p(*d*) is the data evidence (marginal probability of the data)
p(*d*|*h*) is the likelihood (probability of data *d* if hypothesis *h* is true)
p(*h*|*d*) is the posterior probability (probability of hypothesis *h* being true given the observed data *d*)

A weight is calculated for each feature using a Laplacian-adjusted probability estimate to account for the different sampling frequencies of different features.

The weights are summed to provide a probability estimate

Ekins, Williams and Xu, Drug Metab Dispos 38: 2302-2308, 2010

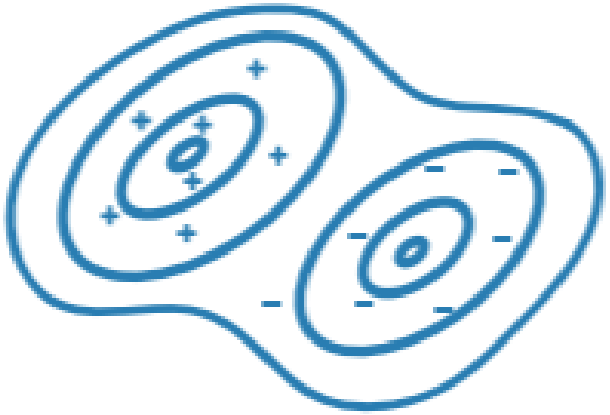
Naïve Bayes

How It Works

A naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. It classifies new data based on the highest probability of its belonging to a particular class.

Best Used...

- For a small dataset containing many parameters
- When you need a classifier that's easy to interpret
- When the model will encounter scenarios that weren't in the training data, as is the case with many financial and medical applications



Bayesian Models - Examples

PXR
Human sodium taurocholate co-transporting polypeptide (NTCP),
Human Multidrug And Toxin Extrusion Proteins, MATE1 and MATE-2K
Human apical sodium-dependent bile acid transporter
Cytochrome P450 3A4 Time-Dependent Inhibition
Human organic cation/carnitine transporter
Volume of distribution
hERG
TB
DILI
BBB, nephrotox, malaria, S Aureus, microsomal stability, cytotoxicity etc

Comparison of SVM and Bayesian ADME/Tox classification models generated with the same molecular descriptors.

Model	Reference	SVM 5 fold cross validation ROC	Bayesian 5 fold cross validation ROC	Bayesian Cross validation ROC
DILI (N = 532)	Ekins, Williams et al. (2010)	0.88	0.63	0.74
PXR (N = 312)	Kortagere et al. (2009)	0.81	0.78	0.84
$5HT_{2B}$ (N = 238)	Chekmarev et al. (2008)	0.83	0.82	0.87
hERG (N = 134)	Chekmarev et al. (2008)	0.82	0.71	0.74
hERG (N = 806)	Wang et al. (2012)	0.88	0.84	0.87
hERG (N = 305,616)[a]	Du et al. (2011)	0.83	0.85	0.86
BBB (N = 1968)	Martins et al. (2012)	0.90	0.91	0.92
AMES (N = 6512)	Hansen, Mika et al. (2009)	0.86	0.84	0.84
Nephrotoxicity (N = 104)	Lin and Will (2012)	0.53	0.64	0.65
Clearance iv (N = 512)[b]	Gombar and Hall (2013)	0.73	0.74	0.76
VDss iv (N = 556)[c]	Gombar and Hall (2013)	0.84	0.81	0.87

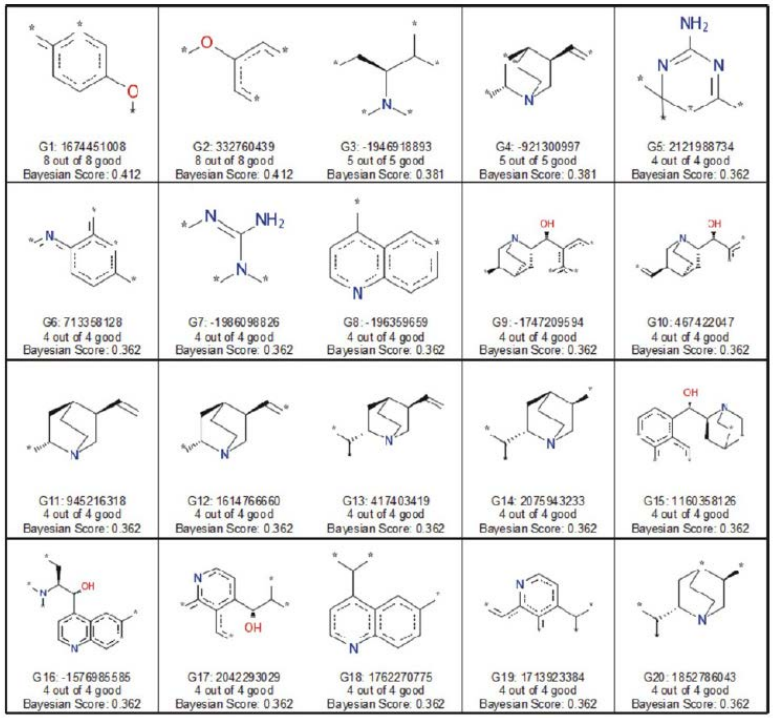
NT not tested. Note DILI model used ECFC_6 fingerprint descriptors instead of FCFP_6.

[a] Random forest Tree Out-of-bag training data ROC score: 0.78.
[b] SVM (regression)—5 fold cross validation $q^2 = 0.17$.
[c] SVM (regression)—5 fold cross validation $q^2 = 0.47$.

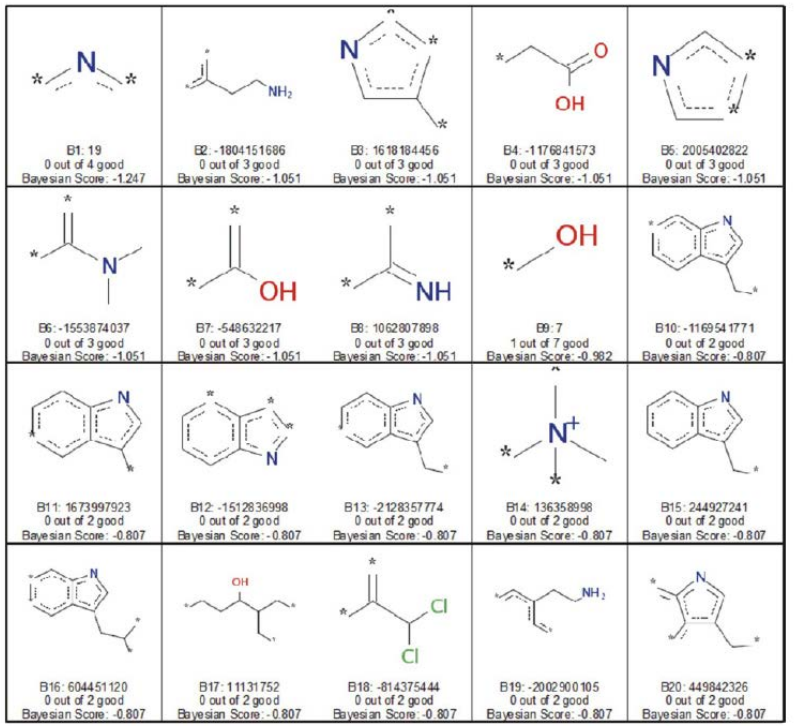
J Pharmacol Toxicol Methods 69: 115-140 (2014)

hMATE1 Bayesian Model Features

+ve



-ve



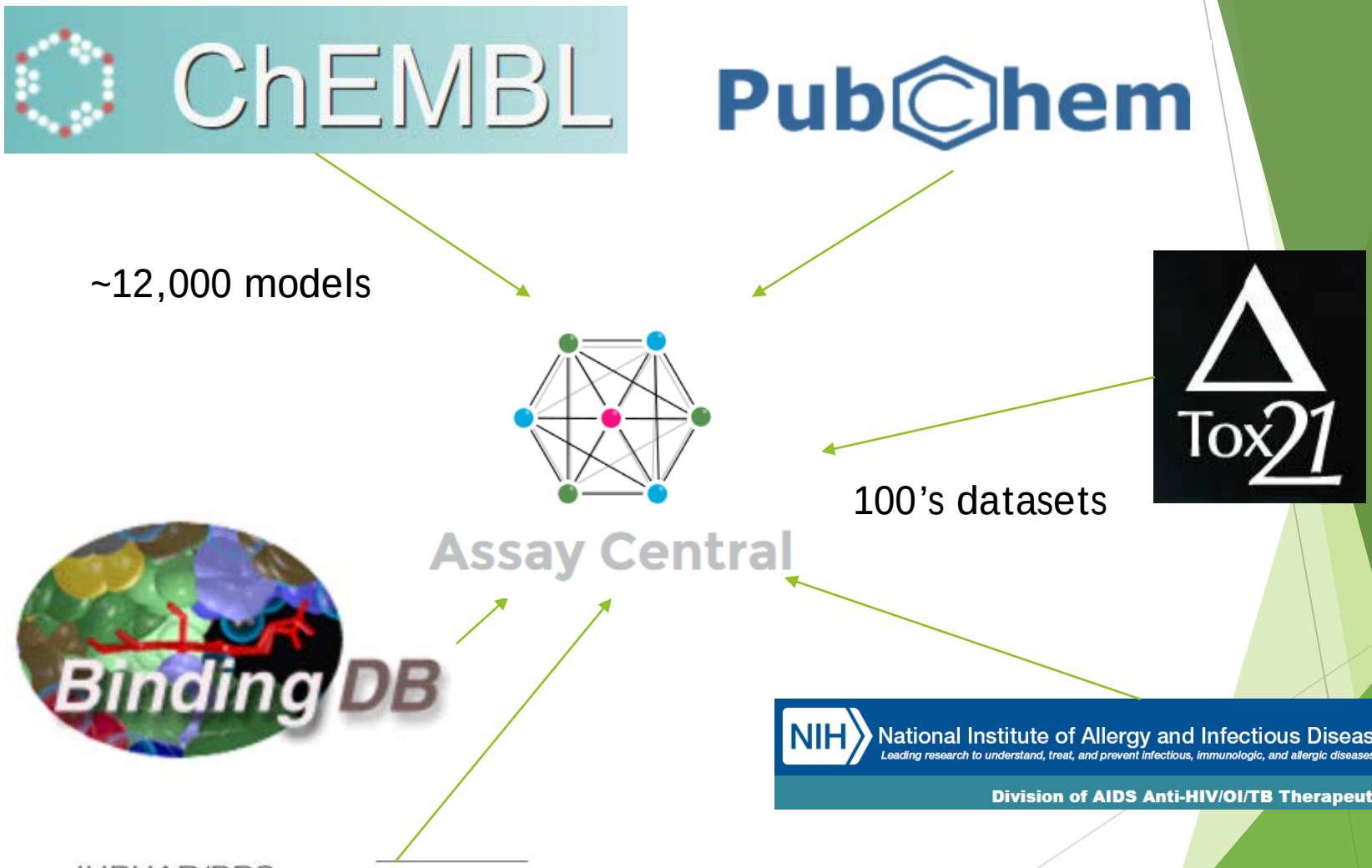
ROC = 0.88, leave out 50% x 100 ROC = 0.82

Bad features pyrole -low basicity

Charge important for increasing interaction with transporter

Astorga et al., JPET 341: 743-755 (2012)

Creating value out of public HTS data



Model resources for ADME/Tox

QSAR TOOLBOX

Bioclipse

admetSAR

CHEM BENCH

eADMET
THE REFERENCE IN CHEMOINFORMATICS

ToxPredict

OpenTox

ToxCreate

ADME SARfari

eTOXlab, an open source modeling framework for implementing predictive models in production environments

Pau Carrió, **Oriol López**, **Ferran Sanz** and **Manuel Pastor***

* Corresponding author: Manuel Pastor manuel.pastor@upf.edu ▾ Author Affiliations

Research Programme on Biomedical Informatics (GRIB), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, IMIM (Hospital del Mar Medical Research Institute), Dr. Aiguader 88, Barcelona, E-08003, Spain

For all author emails, please log on.

Journal of Cheminformatics 2015, **7**:8 doi:10.1186/s13321-015-0058-6

The electronic version of this article is the complete one and can be found online at: http://www.jcheminf.com/content/7/1/8

This article is part of the series Jean-Claude Bradley Memorial Series.

Database Highly accessed Open Access

QSAR DataBank repository: open and linked qualitative and quantitative structure–activity relationship models

V Ruusmann, **S Sild** and **U Maran***

* Corresponding author: U Maran uko.maran@ut.ee ▾ Author Affiliations

Institute of Chemistry, University of Tartu, Ravila 14a, Tartu, 50411, Estonia

For all author emails, please log on.

Journal of Cheminformatics 2015, **7**:32 doi:10.1186/s13321-015-0082-6

The electronic version of this article is the complete one and can be found online at: http://www.jcheminf.com/content/7/1/32

What's stopping us?

- Plenty of data available today… incorrectly formatted
- Vague details of experiments
- Minor & major errors in supplied SMILES/structures
- How do we know this structure is correct?
- How do we share results?
- How can the average scientist use this technology?

Quality of data

- Free from structure errors
- Free from data errors
- Free from experimental errors
- Structure Validation and Standardization
- Curation
- Annotation
- Structure filters
  - Incorrect valency, atom labels, aromatic bonds, stereochemistry, salts, duplication
- Structure standardization guidelines
  - Provided by the FDA (Substance Registration System Unique Ingredient Identifier (UNII): http://www.fda.gov/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/default.htm)
- Need a record of molecule provenance

Drug Discov Today. 2012 Jul;17(13-14):685-701. doi: 10.1016/j.drudis.2012.02.013. Epub 2012 Mar 8.

Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation.

Williams AJ[1], **Ekins** S, Tkachenko V.

MoDELS RESIDE IN PAPERS

NOT ACCESSIBLE...THIS IS UNDESIRABLE

How do we share them?

How do we use Them?

Enhancing Hit Identification in *Mycobacterium tuberculosis* Drug Discovery Using Validated Dual-Event Bayesian Models

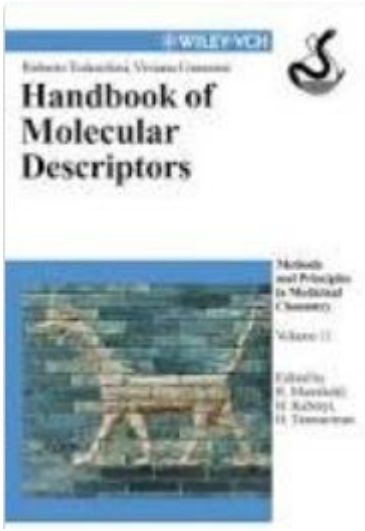
A collaborative database and computational models for tuberculosis drug discovery†

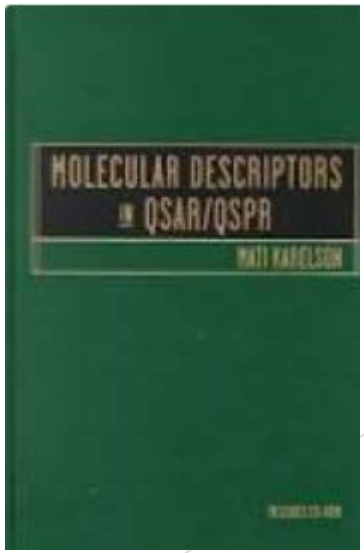
Combining Cheminformatics Methods and Pathway Analysis to Identify Molecules with Whole-Cell Activity Against *Mycobacterium Tuberculosis*

Bayesian Models Leveraging Bioactivity and Cytotoxicity Information for Drug Discovery

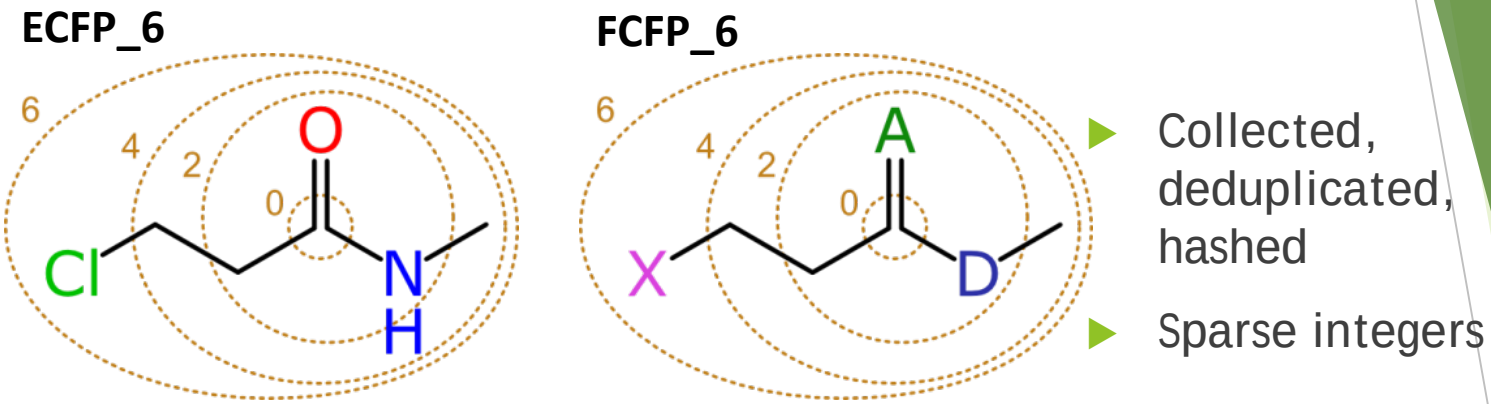
How do we describe Molecules?

- As fingerprints
- As properties
- as experimental measurement/s e.g. logP
- 0D, 1D, 2D, 3D, 4D descriptors





Open Extended Connectivity Fingerprints



- Collected, deduplicated, hashed
- Sparse integers

- Invented for Pipeline Pilot: public method, proprietary details
- Often used with Bayesian models: many published papers
- Built a new implementation: open source, Java, CDK
  - stable: fingerprints don't change with each new toolkit release
  - well defined: easy to document precise steps
  - easy to port: already migrated to iOS (Objective-C) for *TB Mobile* app

Clark et al., J Cheminform 6:38 2014

ChEMBL 20

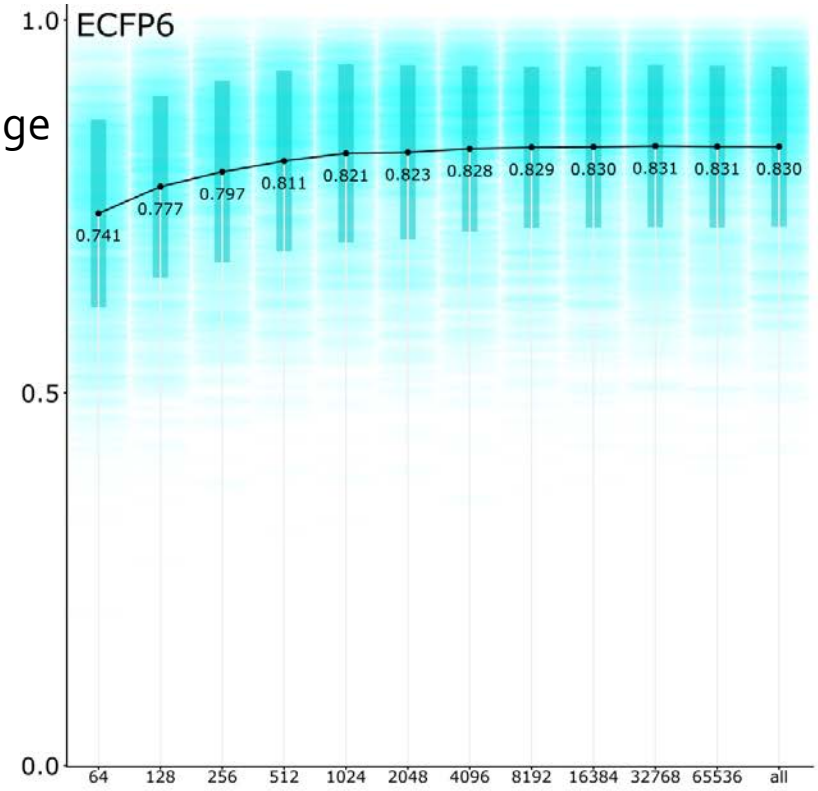
- Skipped targets with > 100,000 assays and sets with < 100 measurements

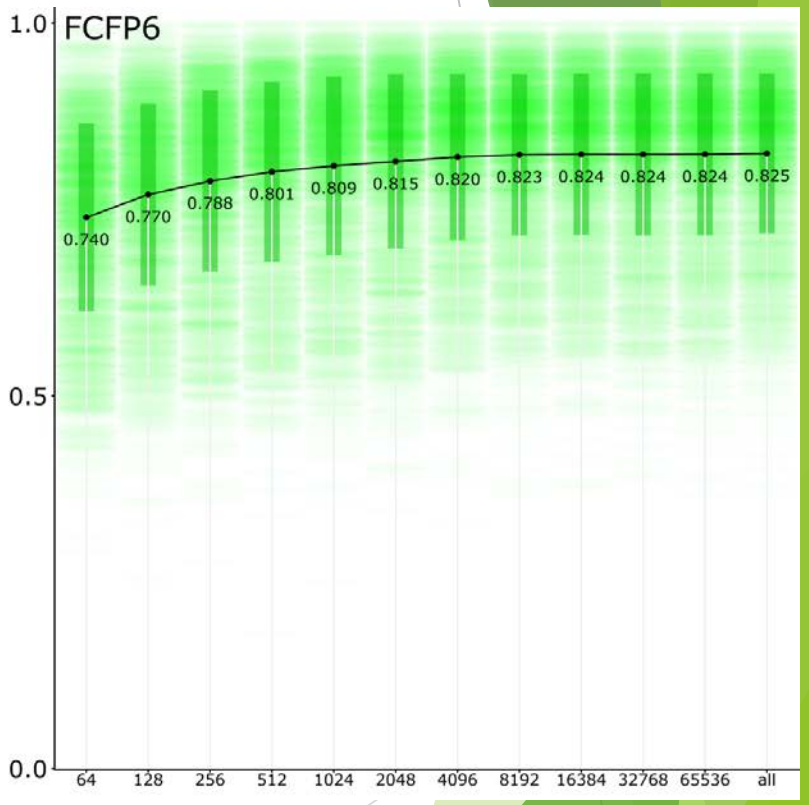
- Converted data to –log
- Dealt with duplicates
- 2152 datasets
- Cutoff determination
- Balance active/ inactive ratio
- Favor structural diversity and activity distribution

http://molsync.com/bayesian2

Clark and Ekins, J Chem Inf Model. 2015 Jun 22;55(6):1246-60

What do 2000 ChEMBL models look like





Folding bit size

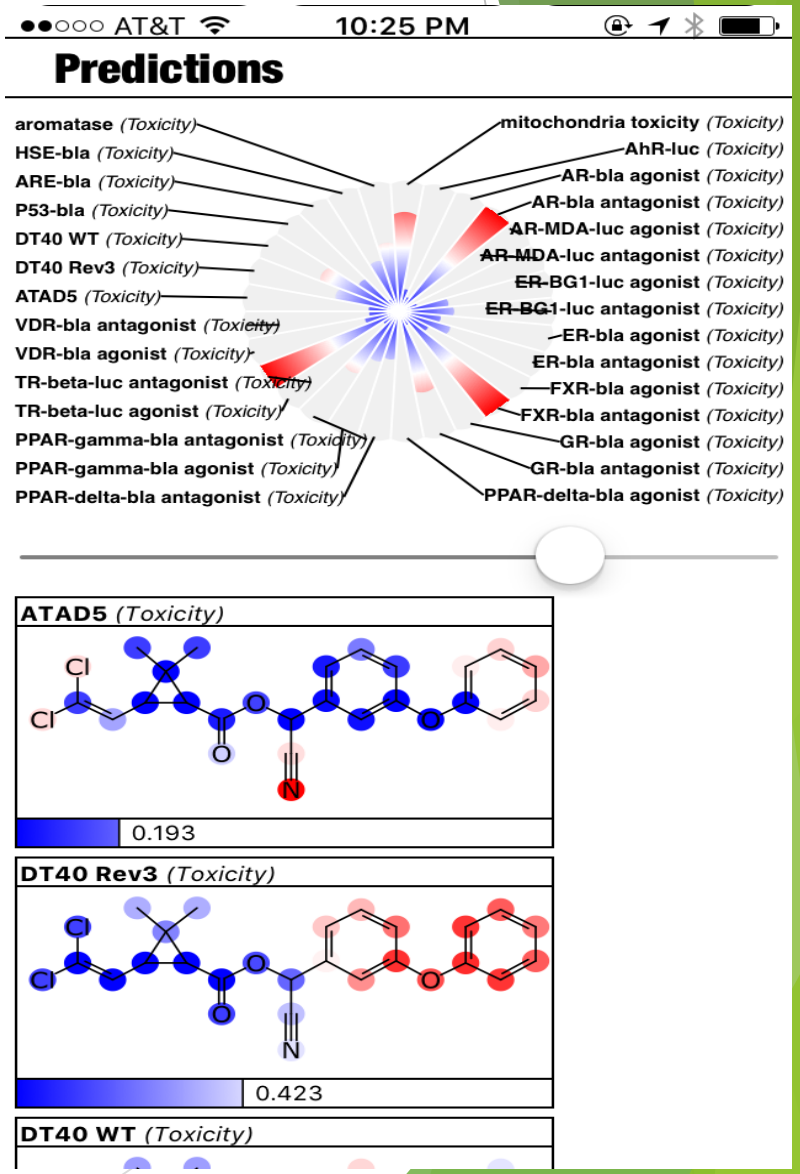
http://molsync.com/bayesian2

Clark and Ekins, J Chem Inf Model. 2015 Jun 22;55(6):1246-60

PolyPharma mobile app (iOS)

- Uses Tox21 data
- Enables prediction
- Visual output from Bayesian models
- Atom contributions – coloring
- free~!

- Alex Clark

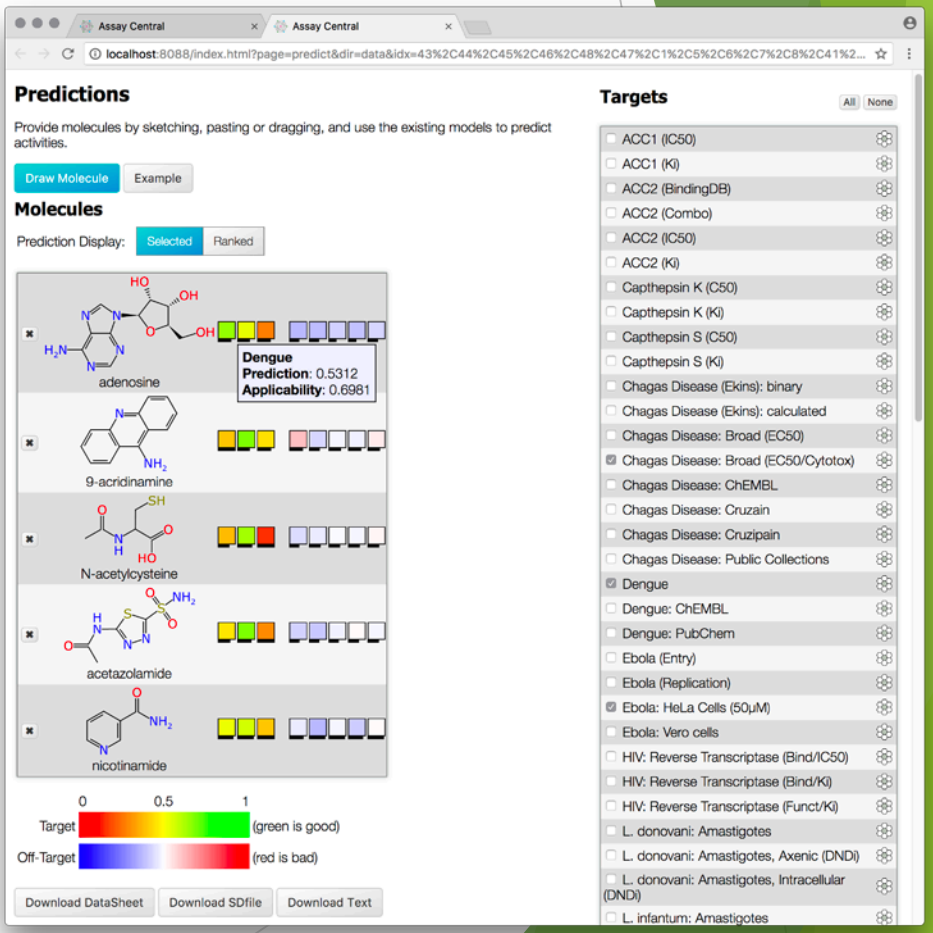


Assay Central

- A tool for building and sharing **Bayesian models** built with biological data from screens
- Assay Central can be used to generate predictions for new molecules (ADME/ Off targets etc)
- Provides model statistics and information on features contributing to activity

www.assaycentral.org

Support by NIH grant 1R43GM122196



ER data

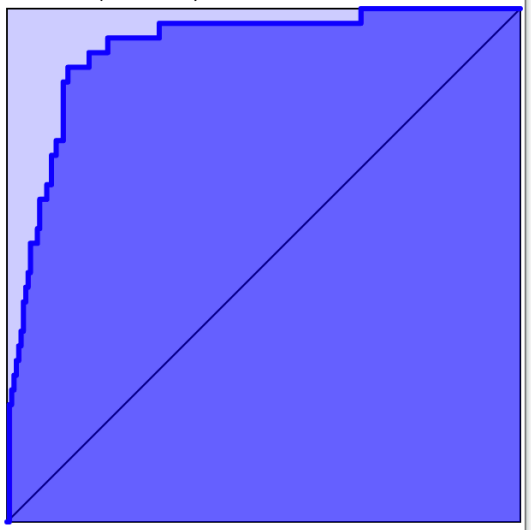
	Title ↑	Target	Organism	Size	ROC	F1	Kappa	MCC	Domain	Invalid
Data Model	ER (Funct/Ki)	Estrogen Receptor	Human	327	0.8774	0.8118	0.6082	0.6089	0.1649	
Data Model	ER: CERAPP Agonists	Estrogen Receptor	Human	1677	0.7691	0.4251	0.3042	0.3319	0.2830	
Data Model	ER: CERAPP Antagonists	Estrogen Receptor	Human	1677	0.6299	0.0911	0.0487	0.0956	0.2830	
Data Model	ER: CERAPP Binding	Estrogen Receptor	Human	1677	0.7745	0.4360	0.3078	0.3330	0.2830	
Data Model	ER: METI	Estrogen Receptor	Human	254	0.9202	0.6739	0.6068	0.6337	0.1811	
Data Model	ER: Tox21 Agonists	Estrogen Receptor alpha	Human	7351	0.8366	0.2973	0.2436	0.3083	0.3764	
Data Model	ER: Tox21 Antagonists	Estrogen Receptor alpha	Human	7351	0.7628	0.1726	0.1143	0.1848	0.3764	
Data Model	ER: alpha (All Binding)	Estrogen Receptor alpha	Human	2347	0.9428	0.8737	0.7426	0.7426	0.2932	
Data Model	ER: alpha (Bind/IC50)	Estrogen Receptor alpha	Human	1127	0.9709	0.9395	0.8698	0.8703	0.2627	
Data Model	ER: alpha (Bind/Ki)	Estrogen Receptor alpha	Human	2347	0.9426	0.8755	0.7512	0.7512	0.2932	
Data Model	ER: alpha (Funct/Ki)	Estrogen Receptor alpha	Human	488	0.8843	0.8169	0.6273	0.6285	0.1934	
Data Model	ER: beta (All Binding)	Estrogen Receptor beta	Human	1806	0.8840	0.8427	0.6645	0.6656	0.2615	
Data Model	ER: beta (Bind/IC50)	Estrogen Receptor beta	Human	969	0.9621	0.9357	0.8635	0.8656	0.2574	
Data Model	ER: beta (Bind/Ki)	Estrogen Receptor beta	Human	1806	0.8822	0.8478	0.6690	0.6690	0.2615	
Data Model	ER: beta (Funct/Ki)	Estrogen Receptor beta	Human	337	0.8907	0.8055	0.6559	0.6559	0.1713	

Collaborations Pharmaceuticals Inc. Non-Proprietary Slides.

An Example of ER model external testing

ER: METI

Origin: Assay Central
Field: ER/METI_v2
Comments: Domain Compatibility: 0.1811452
Training Actives: 35 / 254
ROC: 0.9202 (five-fold)
Curve:



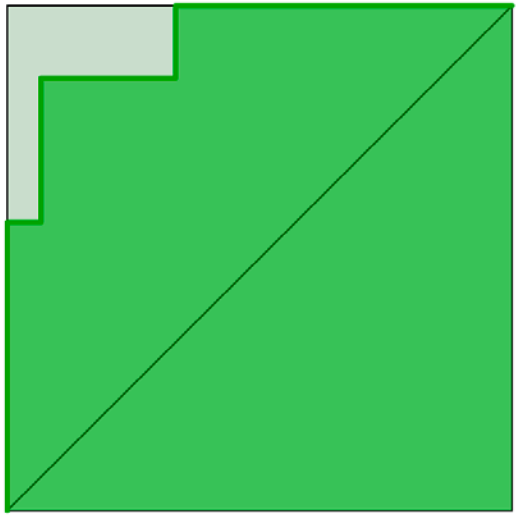
Truth Table:

		Predicted Yes	Predicted No
Actual	Yes	31	4
Actual	No	26	193

Precision: 0.5439
Recall: 0.8857
Specificity: 0.8813
F1 score: 0.6739
Kappa: 0.6068
MCC: 0.6337

subvalidation

Comments: External Cross Validation: actives=7 inactives=15
Training Actives: 28 / 232
ROC: 0.9333 (five-fold)
Curve:



Truth Table:

		Predicted Yes	Predicted No
Actual	Yes	5	2
Actual	No	1	14

Precision: 0.8333
Recall: 0.7143
Specificity: 0.9333
F1 score: 0.7692
Kappa: 0.6733
MCC: 0.6773

MegaTox

- Focused on providing a suite of ADME/Toxicity models
- Tox21 data, hepatotoxicity, cytotoxicity, mutagenicity, cardiotoxicity, drug-drug interactions, microsomal stability, Pregnane X receptor (PXR) and likelihood of causing drug-induced liver injury (DILI)

	Title ↑	Target	Organism	Actives	Size	ROC	F1	Kappa	MCC	Domain	Invalid
Data Model	CytochromeP450: 1A2	CytochromeP450 - 1A2	Human	310	810	0.9064	0.7945	0.6551	0.6580	0.2985	
Data Model	CytochromeP450: 2C19	CytochromeP450 - 2C19	Human	246	789	0.8835	0.7410	0.6203	0.6206	0.2972	
Data Model	CytochromeP450: 2C9	CytochromeP450 - 2C9	Human	458	1203	0.8594	0.7394	0.5589	0.5634	0.3056	
Data Model	CytochromeP450: 2D6	CytochromeP450 - 2D6	Human	732	1668	0.8965	0.8049	0.6379	0.6408	0.3024	
Data Model	CytochromeP450: 3A4	CytochromeP450 - 3A4	Human	1106	2135	0.8919	0.8113	0.6129	0.6131	0.3224	

Deep Learning uses

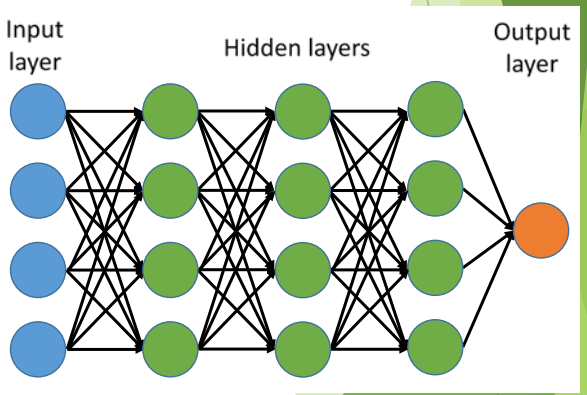
- facial recognition algorithms
  - Facebook tagging photos
- self-driving cars
- robot assistants
- Speech recognition
- Stock markets
- Fraud detection

http://tinyurl.com/hak4lcv

http://tinyurl.com/y8vjv8lp

Deep Learning in Pharmaceutical Research

- Bioinformatics
  - Protein disorder
  - Refine docking complexes
  - Model CLIP-seq data
  - High content image analysis data
  - Biomarkers
  - Protein contacts
  - Cancer diagnosis
- Pharmaceutical
  - Solubility
  - Gene expression data
  - Formulation
  - QSAR – Merck DL out performed random forests in 11 /15 and 13/15 datasets
  - Tox21



Pharm Res (2016) 33:2594–2603
DOI 10.1007/s11095-016-2029-7

PERSPECTIVE

The Next Era: Deep Learning in Pharmaceutical Research

Sean Ekins [1,2]

molecular pharmaceutics

Article

pubs.acs.org/molecularpharmaceutics

Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data sets

Alexandru Korotcov,[†] Valery Tkachenko,[*,†] Daniel P. Russo,[‡,§] and Sean Ekins[*,‡]

[†]Science Data Software, LLC, 14914 Bradwill Court, Rockville, Maryland 20850, United States

[‡]Collaborations Pharmaceuticals, Inc., 840 Main Campus Drive, Lab 3510, Raleigh, North Carolina 27606, United States

[§]The Rutgers Center for Computational and Integrative Biology, Camden, New Jersey 08102, United States

model	data sets used and references	cutoff for active	number of molecules and ratio
solubility	119	Log solubility = −5	1144 active, 155 inactive, ratio 7.38
probe-like	120	described in ref 120	253 active, 69 inactive, ratio 3.67
hERG	121	described in ref 121	373 active, 433 inactive, ratio 0.86
KCNQ1	PubChem BioAssay: AID 2642[122]	using actives assigned in PubChem	301,737 active, 3878 inactive, ratio 77.81
bubonic plague (*Yersina pestis*)	PubChem single-point screen BioAssay: AID 898	active when inhibition ≥50%	223 active, 139 710 inactive, ratio 0.0016
Chagas disease (*Typanosoma cruzi*)	Pubchem BioAssay: AID 2044	with EC_{50} < 1 μM, > 10-fold difference in cytotoxicity as active as described in[88]	1692 active, 2363 inactive, ratio 0.72
TB (*Mycobacterium tuberculosis*)	*in vitro* bioactivity and cytotoxicity data from MLSMR, CB2, kinase, and ARRA data sets[22]	*Mtb* activity and acceptable Vero cell cytotoxicity selectivity index = (MIC or IC_{90})/CC_{50} ≥ 10	1434 active, 5789 inactive, ratio 0.25
malaria (*Plasmodium falciparum*)	CDD Public data sets (MMV, St. Jude, Novartis, and TCAMS)[123−125]	3D7 EC_{50} < 10 nM	175 active, 19 604 inactive, ratio 0.0089

Datasets preparation:

- Datasets were split into training (80%) and test (20%) datasets (default settings)
- Spit datasets maintain equal proportions of active to inactive class ratios (stratified splitting)
- 4-fold cross validation (default settings) on training data for better model generalization

Korotcov et al., Molecular Pharmaceutics 2017

AUC for all tested datasets (FCFP6, 1024 bits)

AUC values	BNB	LLR	ABDT	RF	SVM	DNN-2	DNN-3	DNN-4	DNN-5	Clark et al.
solubility train	0.959	0.991	0.996	0.934	0.983	1.000	1.000	1.000	1.000	0.866
solubility test	0.862	0.938	0.932	0.874	0.927	0.935	0.934	0.934	0.933	
probe-like train	0.989	0.932	1.000	0.984	0.995	1.000	1.000	1.000	1.000	0.757
probe-like test	0.636	0.662	0.658	0.571	0.665	0.559	0.563	0.565	0.563	
hERG train	0.930	0.916	0.992	0.922	0.960	1.000	1.000	1.000	1.000	0.849
hERG test	0.842	0.853	0.844	0.834	0.864	0.840	0.841	0.841	0.840	
KCNQ train	0.795	0.864	0.809	0.764	0.864	1.000	1.000	1.000	1.000	0.842
KCNQ test	0.786	0.826	0.801	0.732	0.832	0.861	0.856	0.852	0.848	
Bubonic plague train	0.956	0.946	0.985	0.895	0.992	1.000	1.000	1.000	1.000	0.810
Bubonic plague test	0.681	0.767	0.643	0.706	0.758	0.754	0.752	0.753	0.753	
Chagas disease train	0.812	0.847	0.865	0.815	0.926	1.000	1.000	1.000	1.000	0.800
Chagas disease test	0.731	0.763	0.768	0.732	0.789	0.790	0.791	0.790	0.789	
Tuberculosis train	0.721	0.737	0.760	0.735	0.800	1.000	1.000	1.000	1.000	0.727
Tuberculosis test	0.671	0.681	0.676	0.679	0.695	0.687	0.684	0.688	0.685	
Malaria train	0.994	0.993	0.999	0.979	0.998	1.000	1.000	1.000	1.000	0.977
Malaria test	0.984	0.982	0.966	0.953	0.975	0.975	0.975	0.974	0.974	

Clark *et al. J Chem Inf Model* 2015

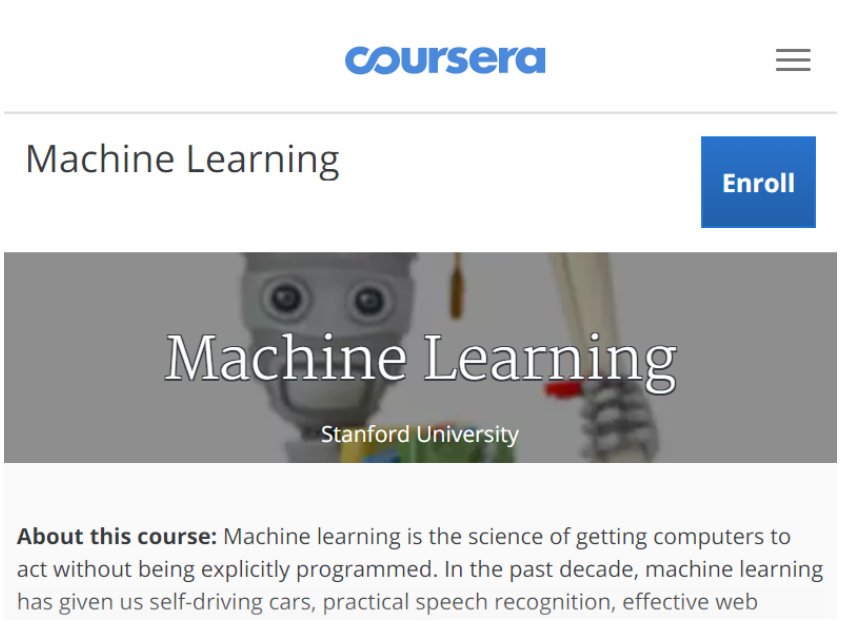
- **Deep learning wins using rank normalized score by metric or by dataset**

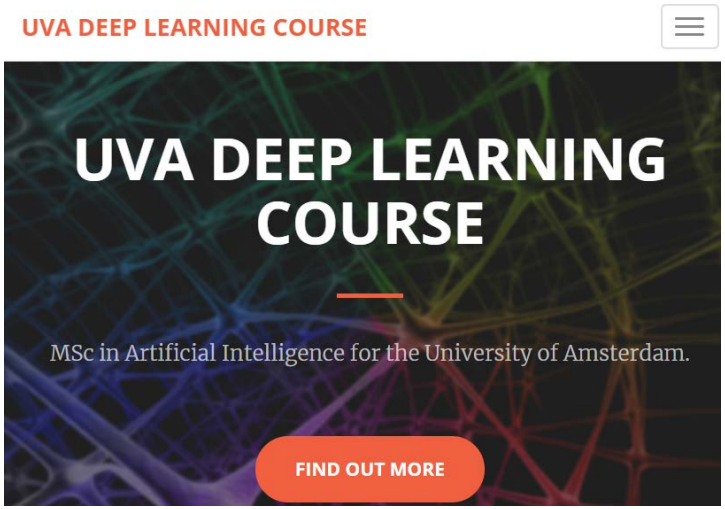
Korotcov et al., Molecular Pharmaceutics 2017

Summary

- Machine learning can be used over a broad array of:
  - projects, diseases, targets, whole cell assays, Tox endpoints etc..
- **Bayesian algorithm demonstrates wide utility**
- **Plenty of scope to pursue other machine learning methods with toxicology data**
- **Spillover of machine learning to new areas**

Where to learn more..










Thanks!

Collaborations Pharmaceuticals, Inc.
Kim Zorn
Dr. Maggie Hupcey
Dr. Tom Lane
[soon to be Dr.] Dan Russo

Consultants
Dr. Alex Clark (Assay Central)
Valery Tkachenko (Deep Learning)
Dr. Alex Korotcov (Deep Learning)

Joel Freundlich
Alex Perryman
Steve Wright
And many colleagues

NIH NIAID R41AI108003-01,
NIH NIGMS R43GM122196,
NIH NCATS R21TR001718,
NIH NINDS, 1R01NS102164-01,
NIH NCATS 1UH2TR002084-01,
One NC Small Business program grant

collaborationspharma@gmail.com

NVIDIA INCEPTION PROGRAM

BIOVIA