OK GOOGLE
Find new drugs!

Large-Scale Machine Learning for Drug Discovery

Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships

Junshui Ma , Robert P. Sheridan , Andy Liaw , George E. Dahl , and Vladimir Svetnik

Massively Multitask Networks for Drug Discovery

Bharath Ramsundar[*,†,°] RBHARATH@STANFORD.EDU
Steven Kearnes[*,†] KEARNES@STANFORD.EDU
Patrick Riley[°] PFR@GOOGLE.COM
Dale Webster[°] DRW@GOOGLE.COM
David Konerding[°] DEK@GOOGLE.COM
Vijay Pande[†] PANDE@STANFORD.EDU
([*]Equal contribution, [†]Stanford University, [°]Google Inc.)

Journal of chemical information and modeling

Arxiv.org

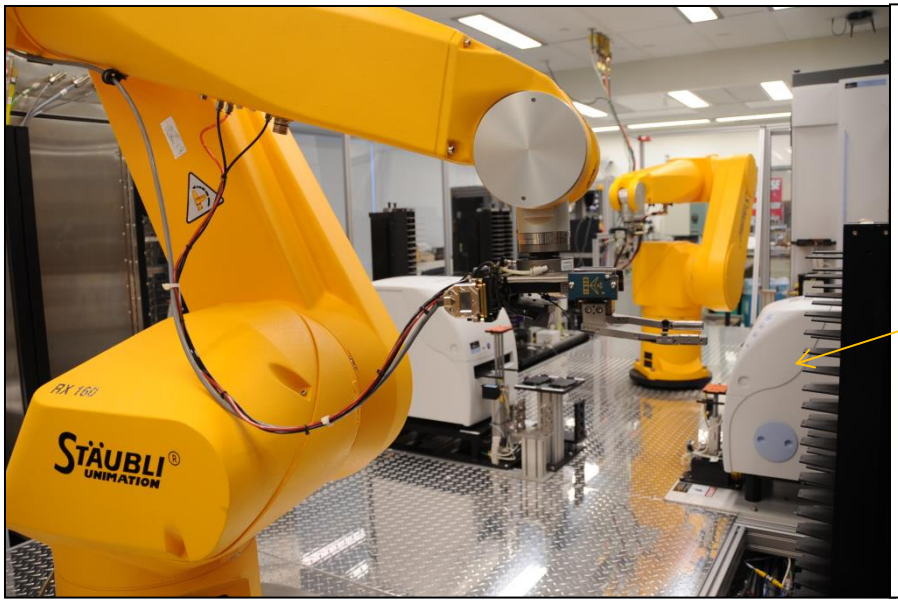
Feb, 2015

Sergey Yakushev, 24.3.15

Drug discovery and high-throughput screening

Compound stores

Robots

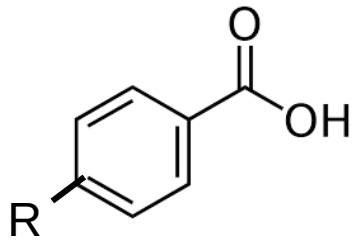


Automation

Drug discovery and high-throughput screening

Motivation for in silico drug discovery:

Known effective compound:

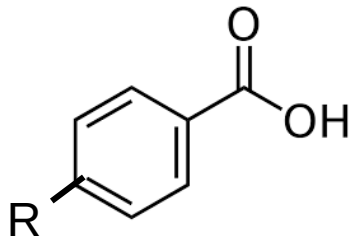


HTS strategy to optimize efficacy:

1. Select ~10 functional groups
2. Place their combination at 4 positions
3. Screen 10000 combination

Time and $

Instead of trial and error – use computer to help



1. Synthesize a small number of modified formulas (x10 than descriptors)
2. Measure various descriptors (properties of the molecules)
3. Assess biological activity
4. Find mathematical correlation between descriptors values and biological effect
5. Predict properties of the rest formulas
6. Synthesize and screen the molecules with desired predicted properties for biological activity

in silico drug discovery approach

Quantitative structure–activity relationships (QSAR)

Technique in the pharmaceutical industry for predicting on-target and off-target activities

Aim: Mathematical relationship between the structure and the biological effect

First application was in 1969
C. Hansch
Equation to calculate the biological activity: minimum effective dose

$$\log (1/C)=k_1 \log P - k_2 (\log P)^2 + k_3 \sigma + k_4$$

C = minimum effective dose
P = octanol-water partition coefficient
σ= Hammett substituent constant (reaction rates of specific group, dependent on the Gibbs free energy)
k_x = constants derived from regression analysis (statistical process estimating relationship among variables)

$$\log (1/C)=\mathbf{k_1} \log P - \mathbf{k_2} (\log P)^2 + \mathbf{k_3} \sigma + \mathbf{k_4}$$

Example:
Binding of drug to serum albumin. It can be determined by their hydrophobicity, in study of 40 compounds they resulted in the following equation:
$\log (1/C) = 0.75 \log P + 2.30$

B, Jhanwar, Pharmacologyonline 1: 306-344 (2011)

Quantitative structure–activity relationships (QSAR)

Descriptors (translate the value of the parameter in to numbers)

Parameters that explain properties in a group of related compounds.

Constitutional:
MW, number of H, functional groups, etc.

Quantum descriptors:
atomic charge, orbital densities, etc.

Geometrical:
volume, shape, surface area

Electrostatic:
dipole moment, polarizability

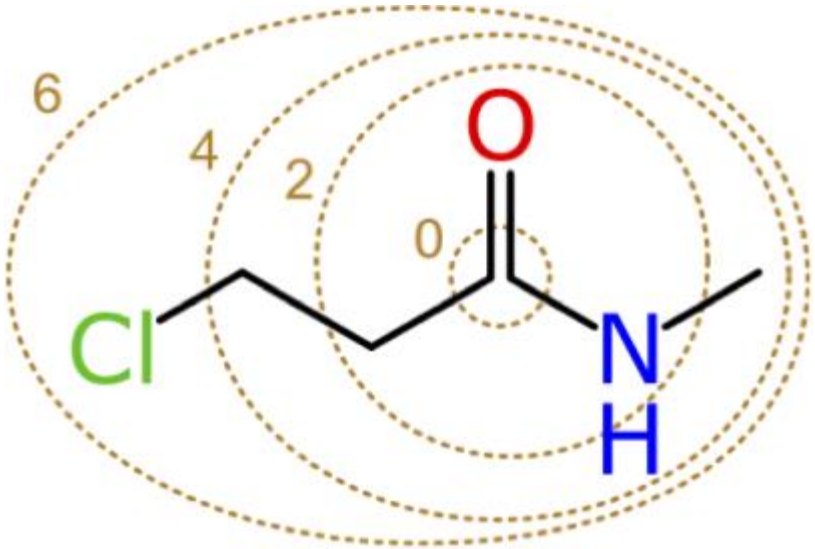
Experimental descriptors:
Come from empirical observations and can not be calcualted in advance. For example set of molecules (aminoacids) interacting with binding site. Can be delivered by site directed mutagenesis.

Quantitative structure–activity relationships (QSAR)

Descriptors

Extended-Connectivity Fingerprints

The molecule is decomposed into a set of fragments—each centered at a non-hydrogen atom—where each fragment extends radially along bonds to neighboring atoms. Each fragment is assigned a unique identifier, and the collection of identifiers for a molecule is hashed into a fixed-length bit vector to construct the molecular “fingerprint”. ECFP4 and other fingerprints are commonly used in cheminformatics applications, especially to measure similarity between compounds



> <ECFP_0>	> <ECFP_2>	> <ECFP_4>	> <ECFP_6>
734603939	734603939	734603939	734603939
1559650422	1559650422	1559650422	1559650422
-1100000244	-1100000244	-1100000244	-1100000244
1572579716	1572579716	1572579716	1572579716
-1074141656	-1074141656	-1074141656	-1074141656
	863188371	863188371	863188371
	-1793471910	-1793471910	-1793471910
	-1789102870	-1789102870	-1789102870
	-1708545601	-1708545601	-1708545601
	-932108170	-932108170	-932108170
	2099970318	2099970318	2099970318
		-87618679	-87618679
		1112638790	1112638790
		-627599602	-627599602

Extended-Connectivity Fingerprints
David Rogers, 2010

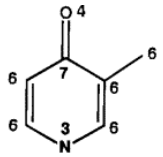
Quantitative structure–activity relationships (QSAR)

Descriptors

Atom Pairs and Donor-Aceptor Pairs

Atom Pairs

(atom 1 description)-(separation)-(atom 2 description)



	unique ap	frequency
1	CX3.-(2)-CX2.	3
2	CX3.-(1)-CX2.	2
3	OX1.-(3)-CX2.	2
4	NX2 -(1)-CX2.	2
5	CX3.-(3)-CX2.	1
6	CX2.-(2)-CX2.	1
7	CX2.-(1)-CX2.	1
8	CX2.-(3)-CX2.	1
9	CX3.-(1)-CX3.	1
10	NX2 -(3)-CX3.	1
11	NX2 -(2)-CX3.	1
12	OX1.-(2)-CX3.	1
13	OX1.-(4)-NX2	1
14	NX2 -(2)-CX2.	1
15	OX1.-(1)-CX3.	1
16	OX1.-(2)-CX2.	1
17	CX3.-(1)-CX1	1
18	CX2.-(4)-CX1	1
19	CX3.-(2)-CX1	1
20	OX1.-(3)-CX1	1
21	CX2.-(3)-CX1	1
22	CX2.-(2)-CX1	1
23	NX2 -(3)-CX1	1
		28 total

	unique bp	frequency
1	4-(3)-6	3
2	6-(1)-6	3
3	6-(2)-6	3
4	6-(3)-6	3
5	6-(2)-7	3
6	3-(1)-6	2
7	3-(2)-6	2
8	4-(2)-6	2
9	6-(1)-7	2
10	3-(4)-4	1
11	3-(3)-6	1
12	3-(3)-7	1
13	4-(1)-7	1
14	6-(4)-6	1
		28 total

	unique tt	frequency
1	OX1.-CX3.-CX3.-CX2.	1
2	OX1.-CX3.-CX3.-CX1	1
3	CX3.-CX3.-CX2.-CX2.	1
4	CX2.-CX3.-CX3.-CX2.	1
5	CX2.-CX3.-CX3.-CX1	1
6	NX2 -CX2.-CX3.-CX3.	1
7	OX1.-CX3.-CX2.-CX2.	1
8	CX3.-CX2.-NX2 -CX2.	1
9	NX2 -CX2.-CX2.-CX3.	1
10	NX2 -CX2.-CX3.-CX1	1
11	CX2.-NX2 -CX2.-CX2.	1
		11 total

	unique bt	frequency
1	4-7-6-6	3
2	6-6-7-6	3
3	3-6-6-7	2
4	6-3-6-6	2
5	3-6-6-6	1
		11 total

Donor-Aceptor Pairs

Add-on to AP, that describes the atom

1 = cations, 2 = anions, 3 = neutral hydrogen bond donors, 4 = neutral hydrogen bond acceptors, 5 = polar atoms (both donor and acceptor, e.g., hydroxy oxygen), 6 = hydrophobic atoms, 7 = other

Chemical Similarity Using Physiochemical Property Descriptors, Simon K. Kearsley, January 24, 1996

Quantitative structure–activity relationships (QSAR)

Quantitative (mathematical) relationship between structure and observed activity

Find correlation coefficients between observed activity and the descriptors

Curve fit – find an equation

No understanding of chemistry or mechanism is required!

Statistical correlation

Quantitative structure–activity relationships (QSAR)

Modeling

At present time there is a data across different molecules, different pathologies, different pathways and different biological effects

How one can teach the computer to classify the data and make the model?

Two major modeling system in QSAR:

1. Random Forest
2. Neural Networks

Modeling in QSAR

Decision tree

Modeling in QSAR

Decision tree

Predict if John will play tennis

Training examples: **9 yes / 5 no**

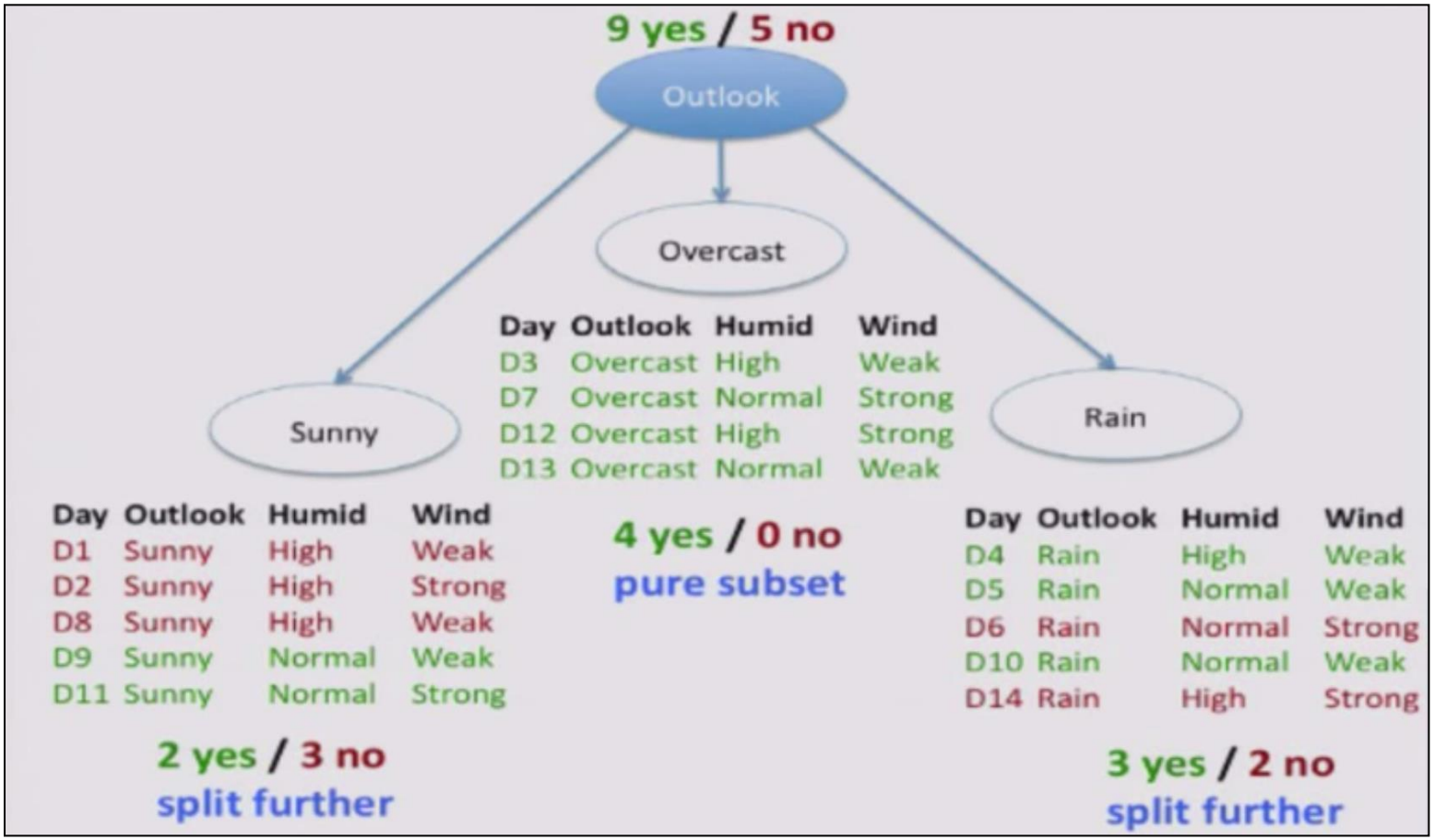
Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

New data:

D15	Rain	High	Weak	?

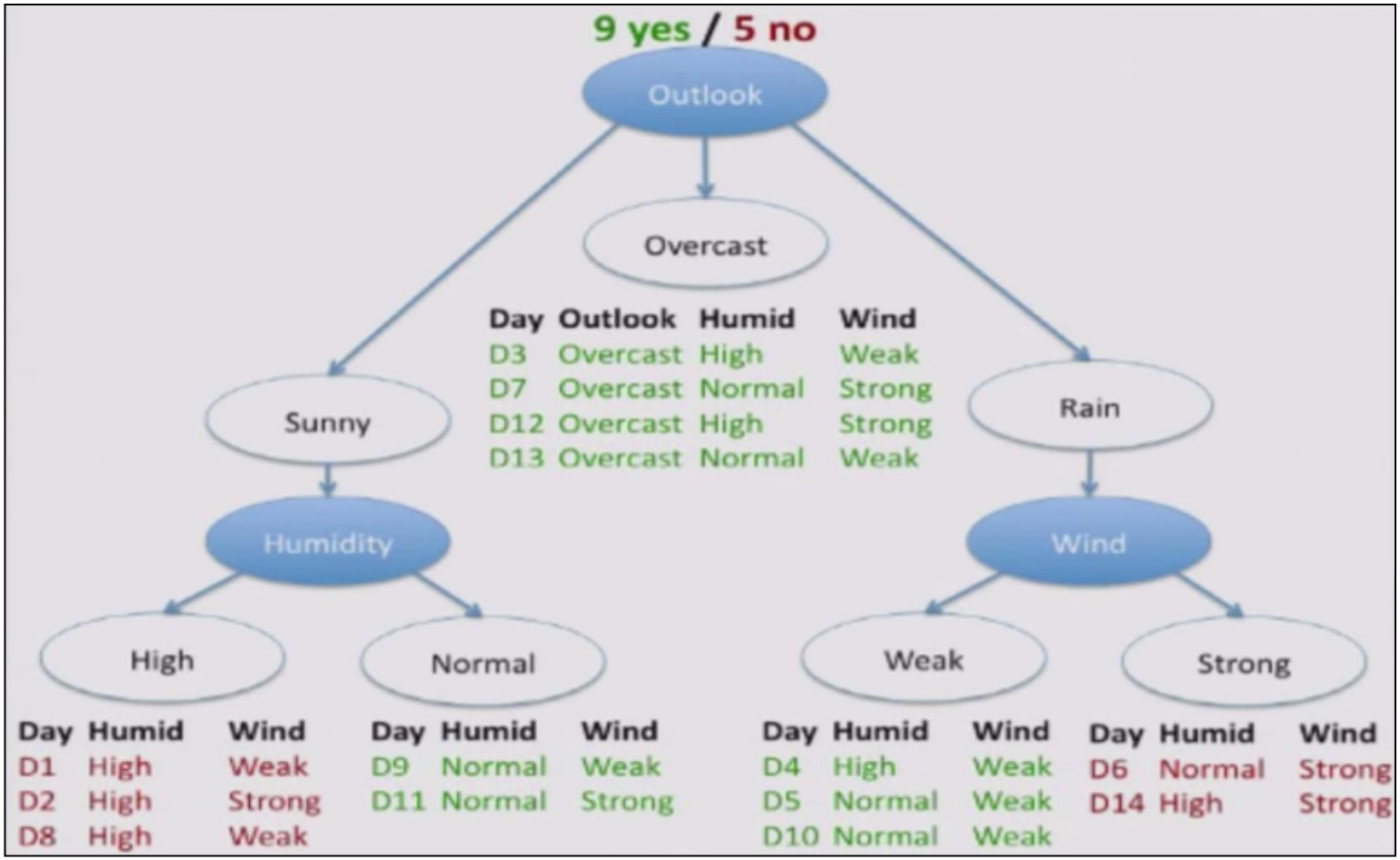
Modeling in QSAR

Decision tree



Modeling in QSAR

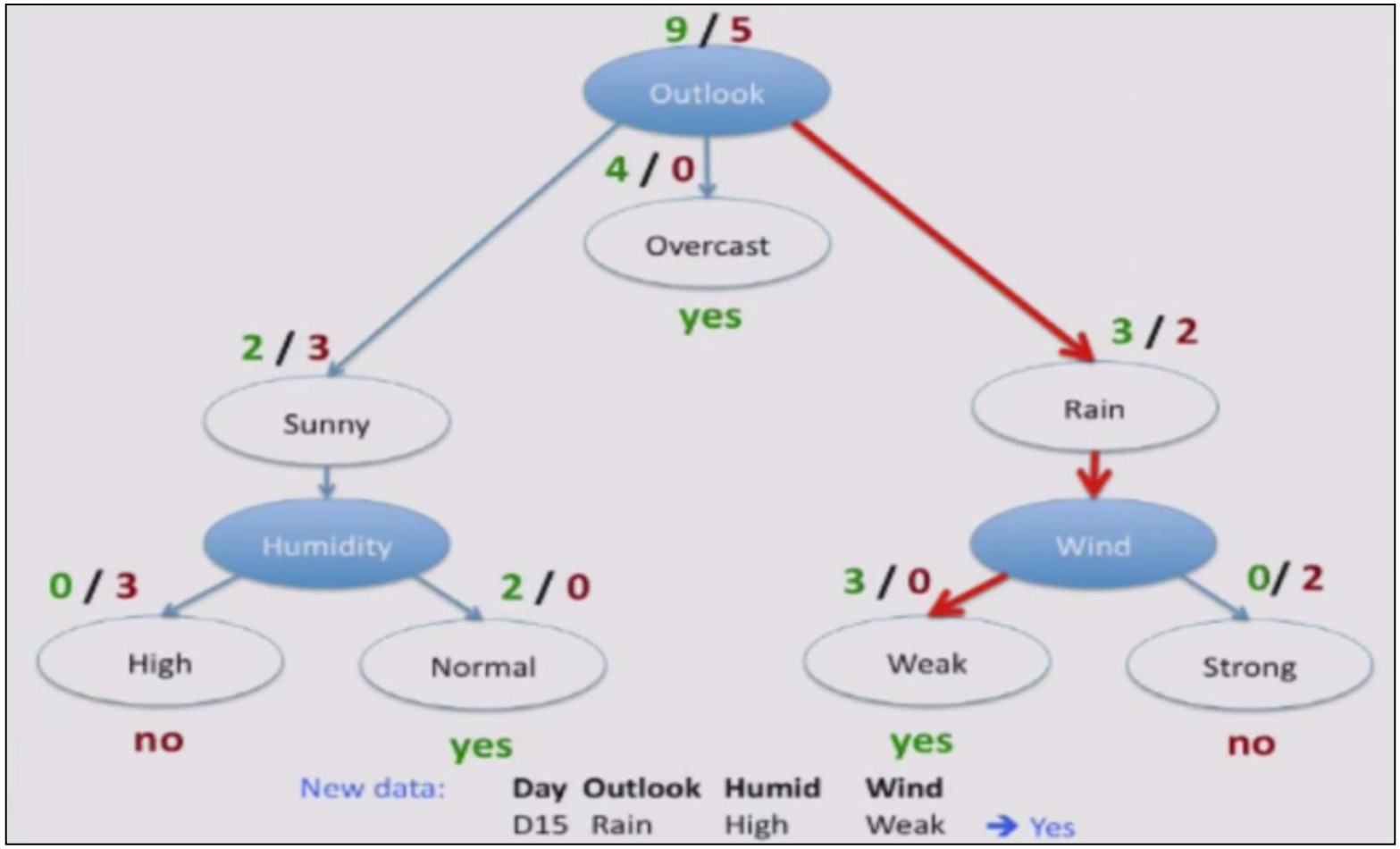
Decision tree



Pure subsets of descriptors

Modeling in QSAR

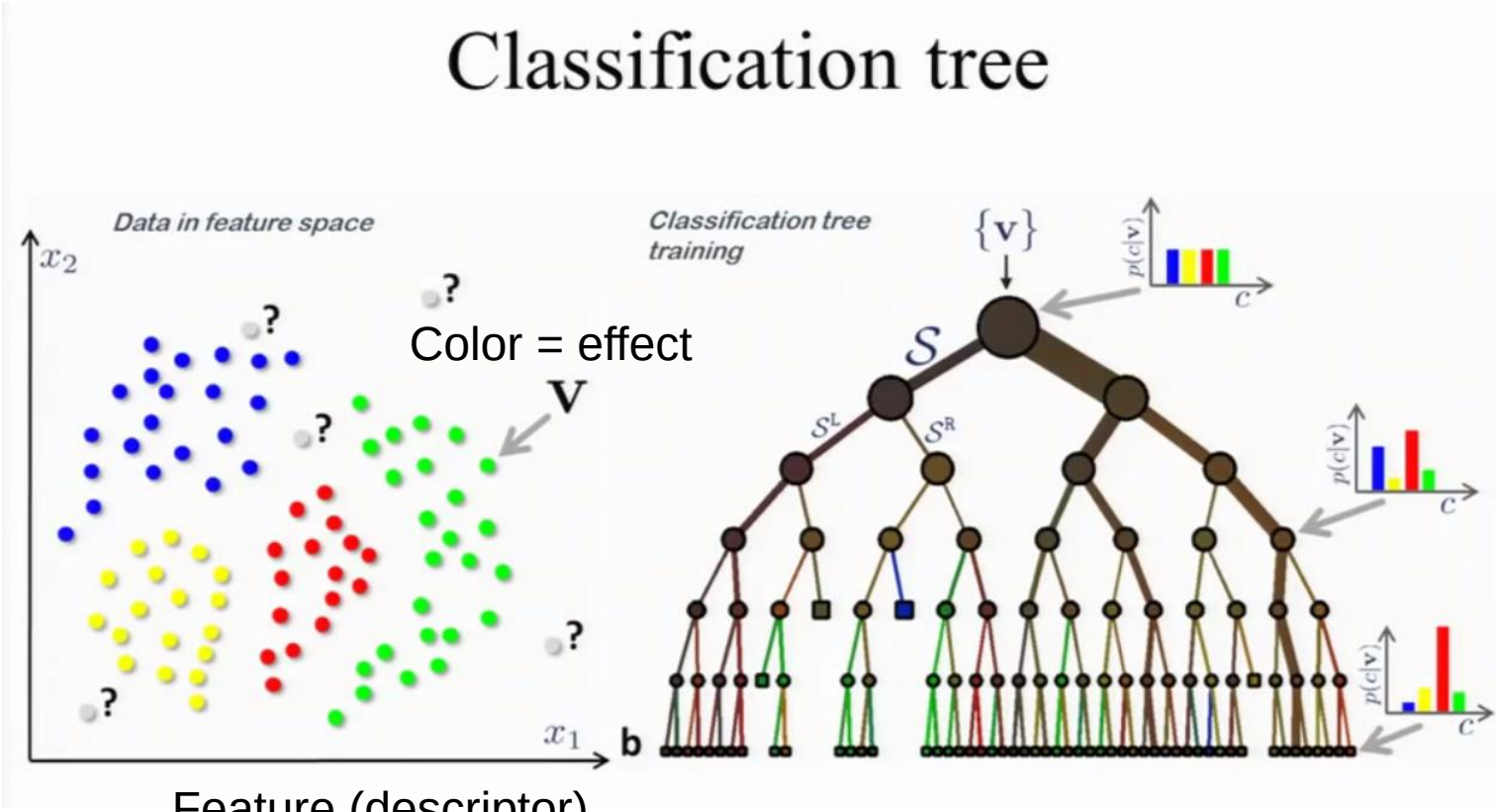
Decision tree



Modeling in QSAR

Decision tree

Data (value)



Modeling in QSAR

Decision tree

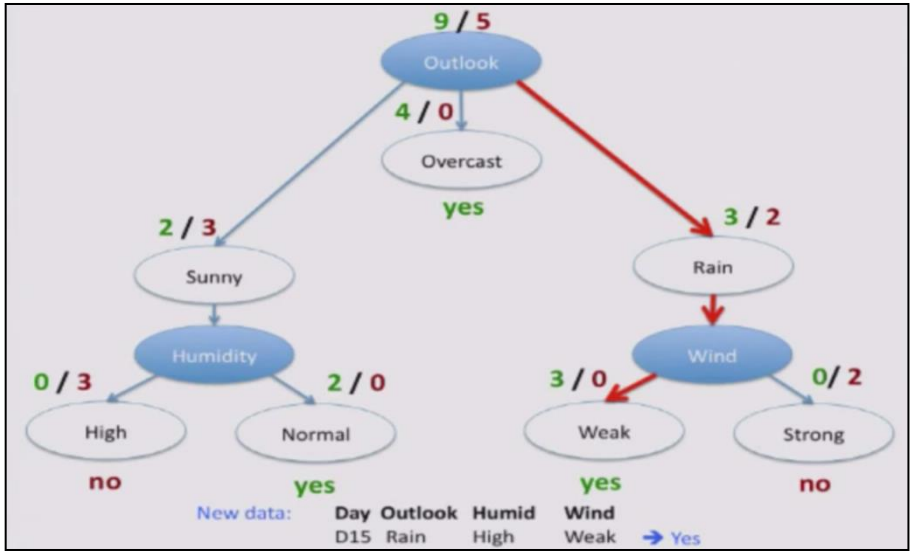
Many descriptors
Many values
Many effects

Matrix

descriptors

$$S = \begin{pmatrix} A0\ B0\ C0\ D0\ E0\ F0 \\ A1\ B1\ C1\ D1\ E1\ F1 \\ A2\ B2\ C2\ D2\ E2\ F2 \end{pmatrix}$$

values



We need a learning method to find correlations: classification of the descriptors

Modeling in QSAR

Random Forest (of decision trees)

From the matrix create many random subsets

$$S = \begin{pmatrix} A0\ B0\ C0\ D0\ E0\ F0 \\ A1\ B1\ C1\ D1\ E1\ F1 \\ A2\ B2\ C2\ D2\ E2\ F2 \end{pmatrix}$$

$$S1 = \begin{pmatrix} A3\ B8\ C1\ D9\ E1\ F0 \\ A8\ B5\ C9\ D7\ E4\ F1 \\ A2\ B2\ C2\ D2\ E2\ F2 \end{pmatrix}$$

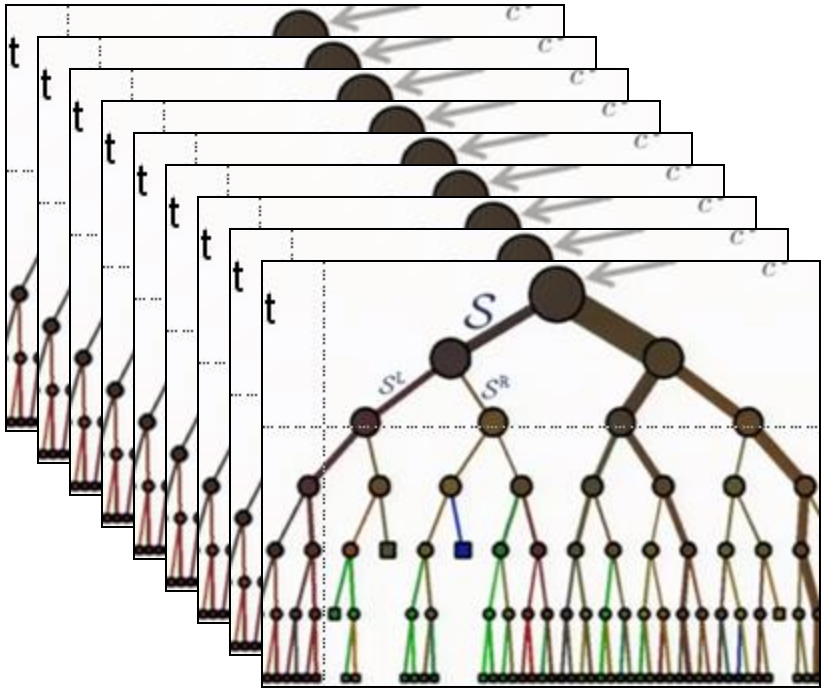
$$S2 = \begin{pmatrix} A9\ B4\ C2\ D1\ E9\ F6 \\ A8\ B5\ C9\ D7\ E4\ F1 \\ A2\ B2\ C2\ D2\ E2\ F2 \end{pmatrix}$$

…

Modeling in QSAR

Random Forest (of decision trees)

From the matrix create many random subsets
And for each of them make a decision tree



Ask each tree in the forest for prediction of the biological effect.

Vote for the most frequent one

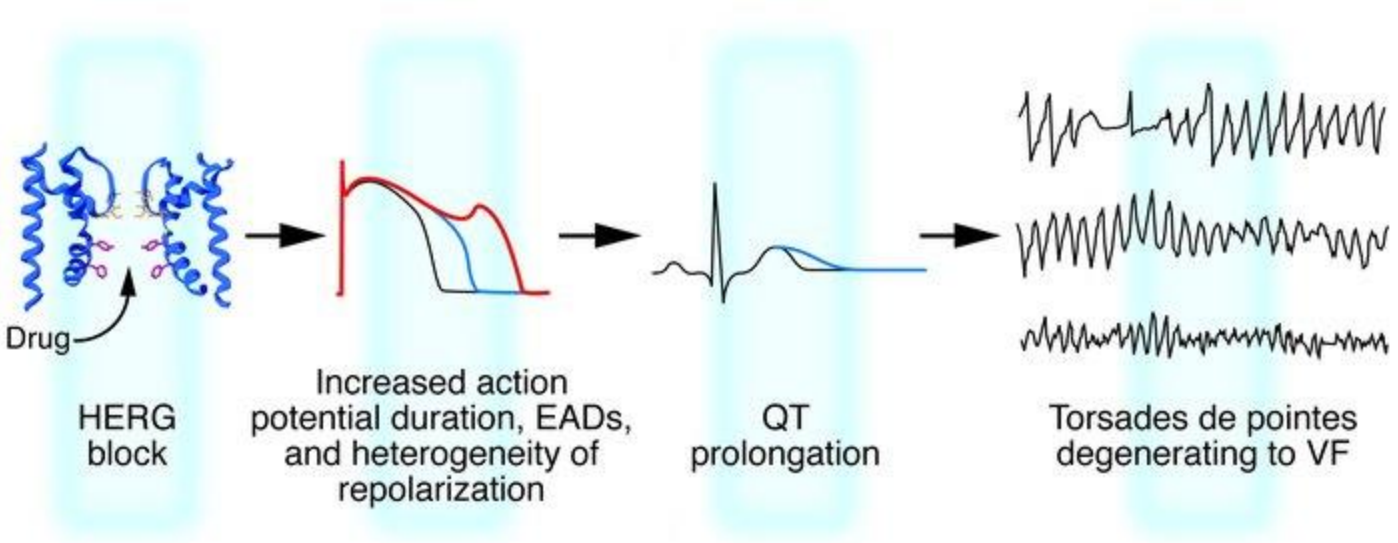
Modeling in QSAR

Random Forest (of decision trees)

Example of application

The human ether-a-go-go related gene (hERG) channel is a cardiac ion channel that is involved in regulation of cardiac action potential.

Blockage of this potassium channel extends the repolarization phase, leading to a prolonged QT interval.



Therefore, the hERG channel is a general anti-target

Genetics of acquired long QT syndrome
Dan M. Roden,
CJ Clin Invest. 2005;115(8):2025-2032.

Modeling in QSAR

Random Forest (of decision trees)

Example of application

Dataset: 280 compounds

in vitro hERG inhibition data were collected from the literature, Prous Science Integrity,19 and an inhouse experiment

Data (biological effect):

whole-cell patch-clamp, experimental IC50 or pIC50 (−log IC50) values

IC50 ≥ 10 μM or pIC50(−logIC50) ≤ 5 were assigned to class 0 (weak inhibitors)

the others were assigned to class 1 (strong inhibitors)

Descriptor:

Extended-Connectivity Fingerprints (ECFP4)

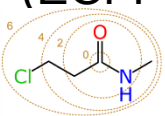


Table 1. The number of compounds for data set

Class	Learning	Validation A	Validation B
0	88	16	12
1	136	18	10
Total	224	34	22

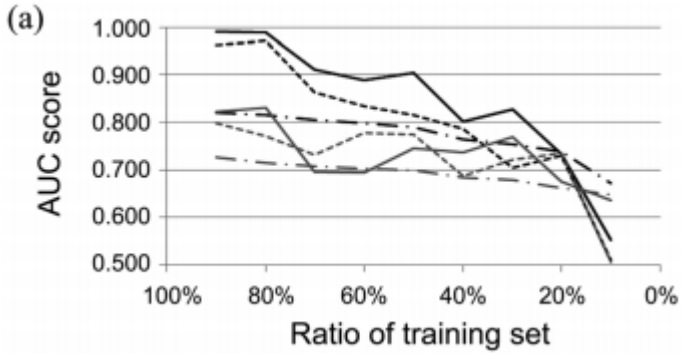
Variable % training set and test set

The Predictive QSAR Model for hERG Inhibitors Using Bayesian and Random Forest Classification Method, Jun Hyoung Kim, Bull. Korean Chem. Soc. 2011, Vol. 32, No. 4 1237

Modeling in QSAR

Random Forest (of decision trees)

Test of the model

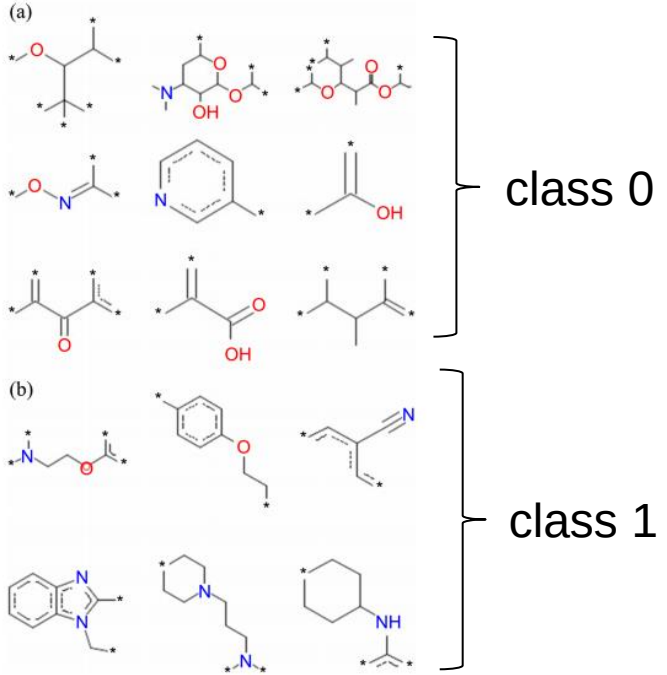


AUC = area under the curve (integral) of receiver operating characteristic.

The AUC provides a simple quality assessment for a classification model. The closer the AUC score is to 1.0, the better the model is at distinguishing samples in one of the classes

Identifying the important descriptor value

To identify the important features that contributed to class 0 or class 1, the ECFP4 descriptors were extracted from each learned model of which the normalized probability values are less than −0.70 (class 0) or more than 0.35 (class 1). The normalized probability is the final contribution of the descriptor value to the total relative estimate.



The Predictive QSAR Model for hERG Inhibitors Using Bayesian and Random Forest Classification Method, Jun Hyoung Kim, Bull. Korean Chem. Soc. 2011, Vol. 32, No. 4 1237

Modeling in QSAR

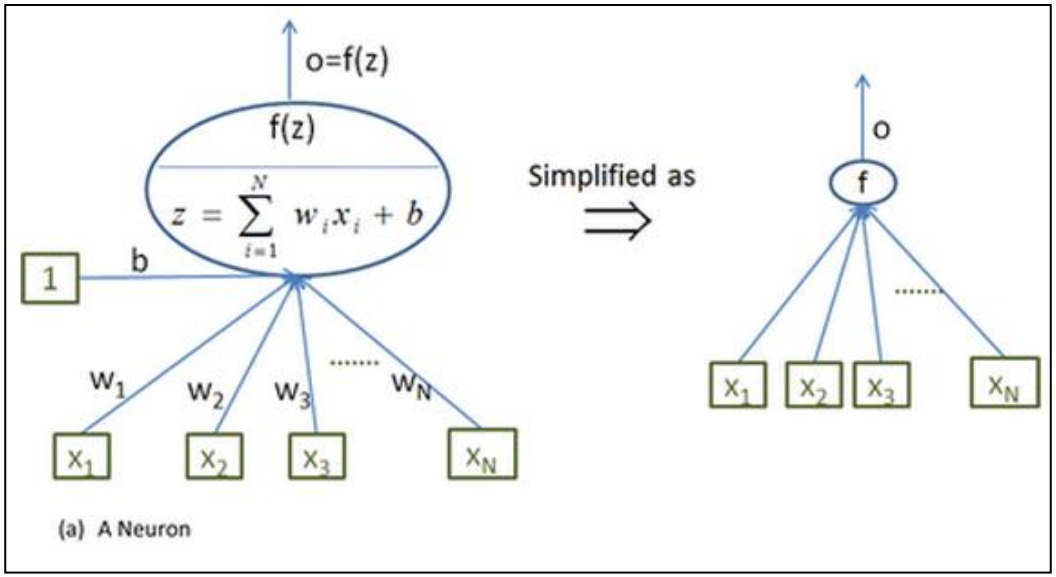
Neural Networks

effect

Activation function:

Default bias:

Input:



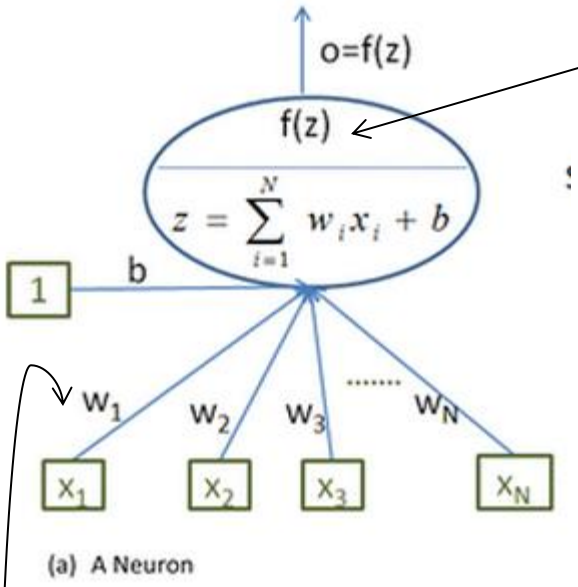
As many as descriptors and value X of each

Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships, Junshui Ma

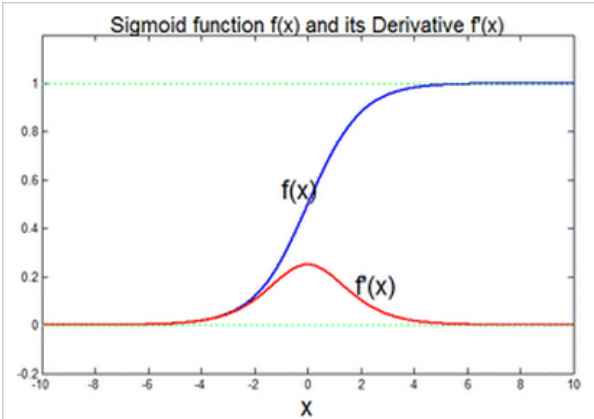
Modeling in QSAR

Neural Networks

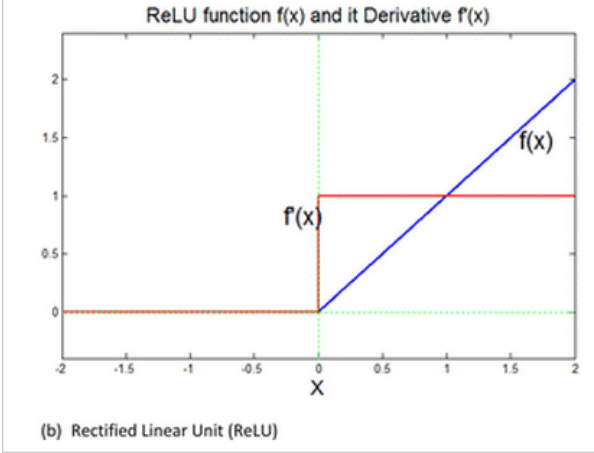
Objective function – result of learning



Activation function (from biology of neurons)



Red – derivative of the function



No saturation regime

Synaptic weight

(This is what we need for learning)

Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships, Junshui Ma

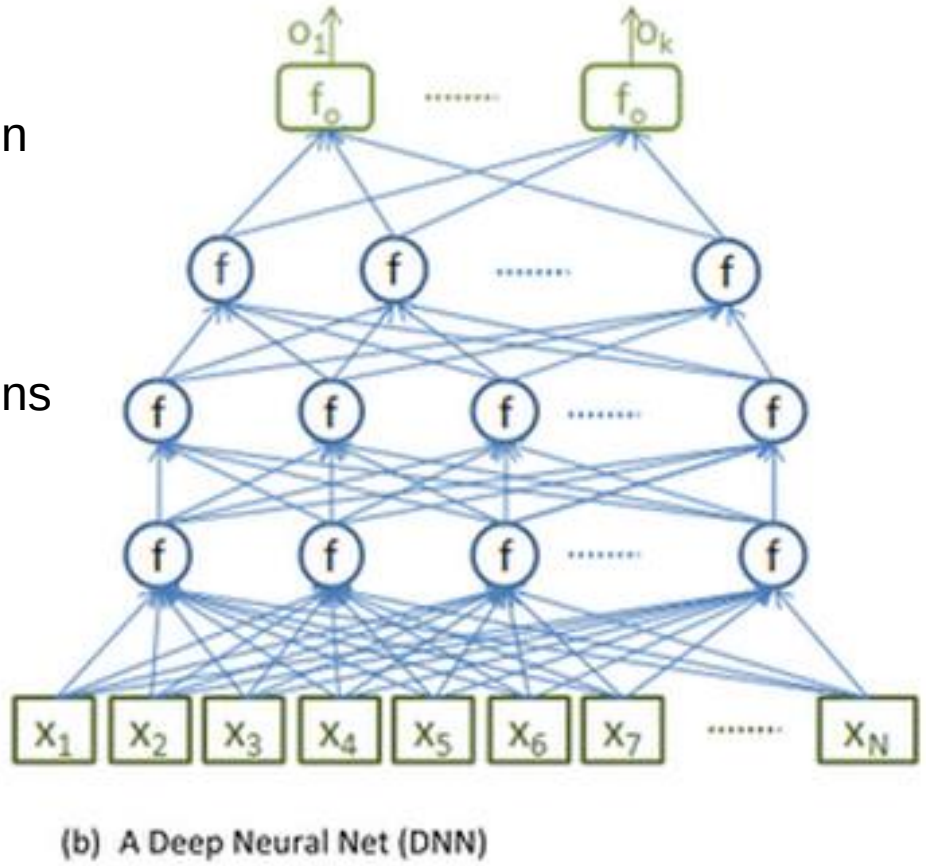
Modeling in QSAR

Deep Neural Networks

Output layer,
Generated prediction

Deep layers of neurons

Input layer,
Enter of descriptor

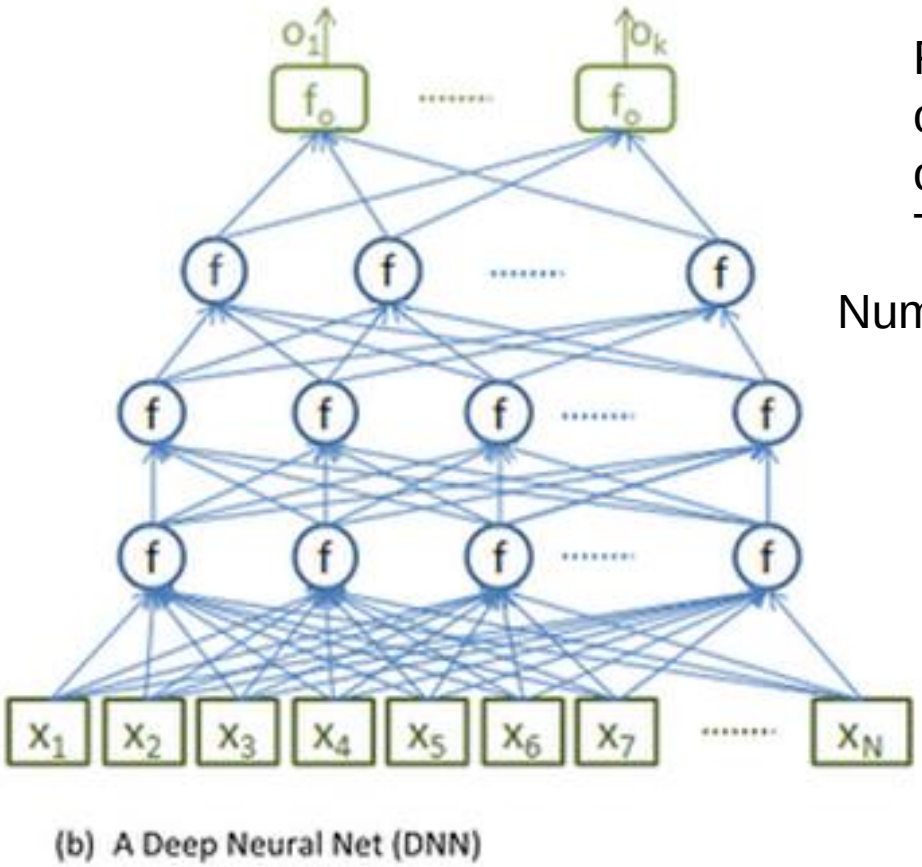


(b) A Deep Neural Net (DNN)

Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships, Junshui Ma

Quantitative structure–activity relationships (QSAR)

Deep Neural Networks



(b) A Deep Neural Net (DNN)

For the QSAR there should be multiple output neurons, each of them gives specific objecive function.
This is called multitask DNN.

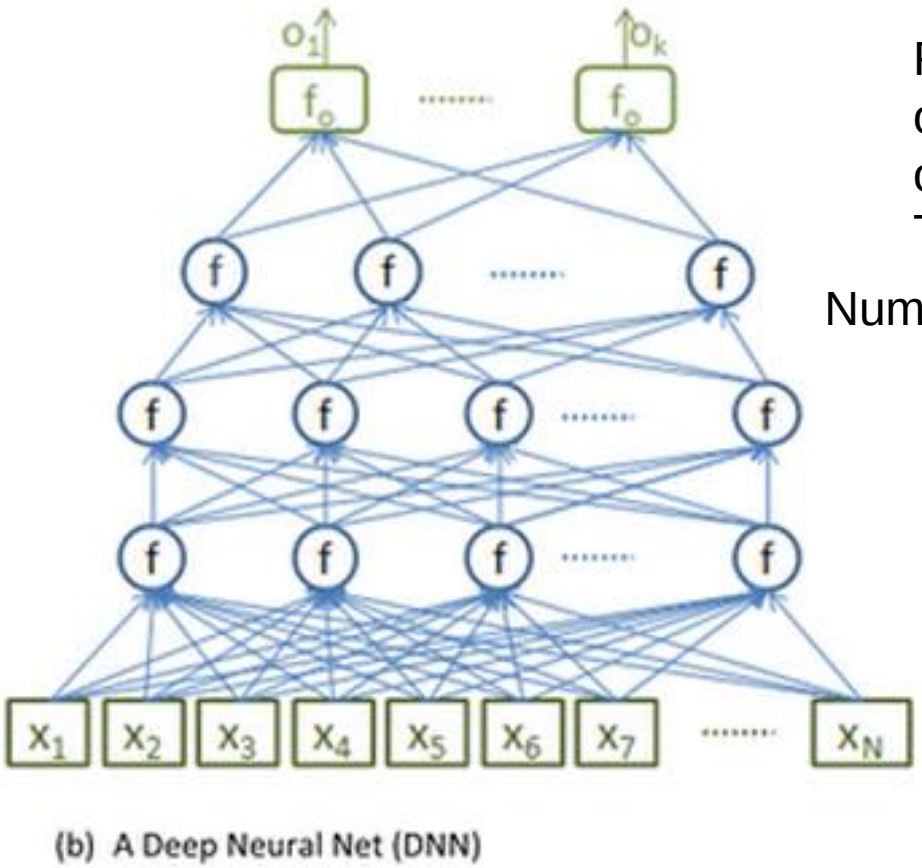
Number of tasks = number of objective functions

Why multitasking?
Different compounds might share several features. Even if data on compounds (compound classes) is poorly related, they are still governed by the laws of chemistry and it might still be important to learn more broadly useful higher-level features from the initial descriptors.
(example see previous paper: compound efficacy on one pathway and low inhibition).

Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships, Junshui Ma

Quantitative structure–activity relationships (QSAR)

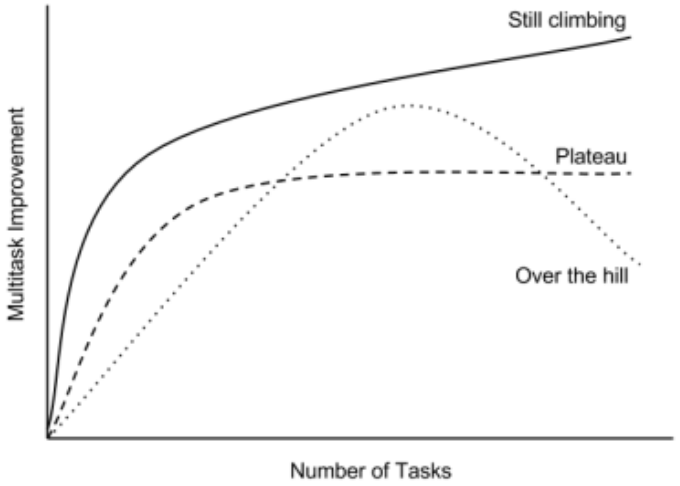
Deep Neural Networks



(b) A Deep Neural Net (DNN)

For the QSAR there should be multiple output neurons, each of them gives specific objecive function.
This is called multitask DNN.

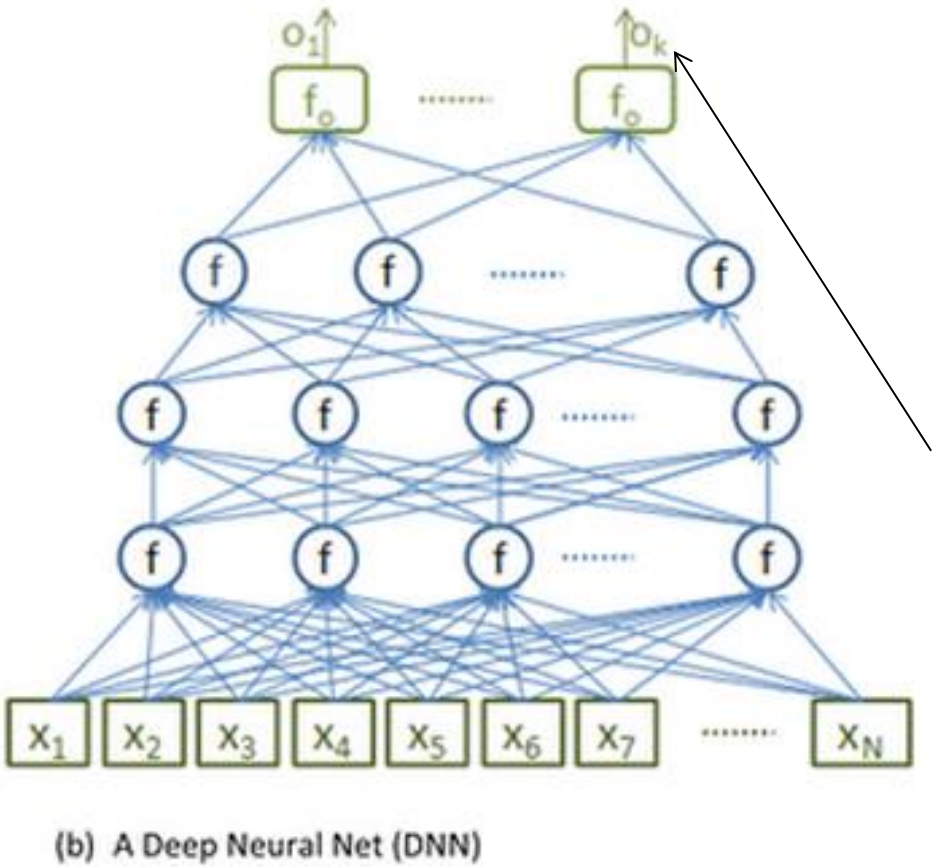
Number of tasks = number of objective functions



Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships, Junshui Ma

Quantitative structure–activity relationships (QSAR)

Deep Neural Networks



(b) A Deep Neural Net (DNN)

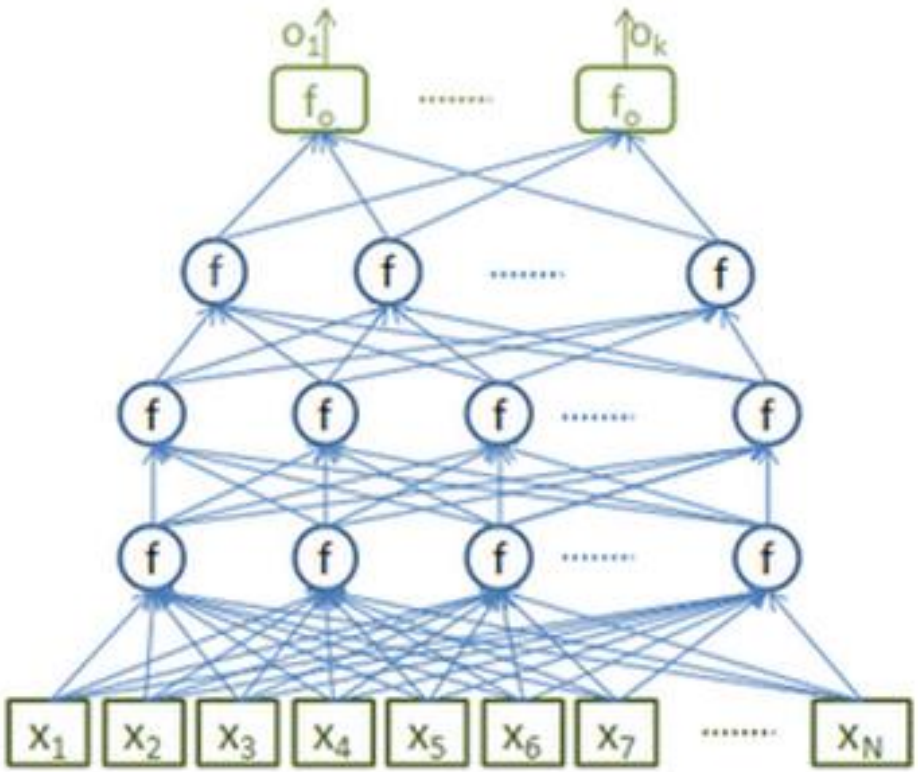
Training procedure:
Backward propagation

The objective function can have several millions of individual values (coef.).
This is dangerous for over fitting.
So initially the DNN is trained unsupervised with the set of compounds descriptors, without specifying their effect.

Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships, Junshui Ma

Quantitative structure–activity relationships (QSAR)

Deep Neural Networks



(b) A Deep Neural Net (DNN)

Training procedure:
Backward propagation

The large list of compounds is randomly split into sets and the DNN is trained on this sets. By the end of training the objective function is adjusted.

A learning of a all sets is an “Epoch”.

It is needed to have many epochs to train the DNN.

Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships, Junshui Ma

Data:

259 datasets gathered from publicly available data
contained 37.8M experimental data points for 1.6M compounds

Group	Datasets	Data Points / ea.	% Active
PCBA	128	282K (122K)	1.8 (3.8)
DUD-E	102	14K (11K)	1.6 (0.2)
MUV	17	15K (1)	0.2 (0)
Tox21	12	6K (500)	7.8 (4.7)

Massively Multitask Networks for Drug Discovery

Group	Datasets	Data Points / ea.	% Active
PCBA	128	282K (122K)	1.8 (3.8)
DUD-E	102	14K (11K)	1.6 (0.2)
MUV	17	15K (1)	0.2 (0)
Tox21	12	6K (500)	7.8 (4.7)

The PubChem BioAssay database currently contains 500,000 descriptions of assay protocols, covering 5000 protein targets, 30,000 gene targets and providing over 130 million bioactivity outcomes. PubChem's bioassay data are integrated into the NCBI Entrez information retrieval system, thus making PubChem data searchable and accessible by Entrez queries.

Massively Multitask Networks for Drug Discovery

Group	Datasets	Data Points / ea.	% Active
PCBA	128	282K (122K)	1.8 (3.8)
DUD-E	102	14K (11K)	1.6 (0.2)
MUV	17	15K (1)	0.2 (0)
Tox21	12	6K (500)	7.8 (4.7)

The DUD-E group contained 102 datasets that were designed for the evaluation of methods to predict interactions between proteins and small molecules

Massively Multitask Networks for Drug Discovery

Group	Datasets	Data Points / ea.	% Active
PCBA	128	282K (122K)	1.8 (3.8)
DUD-E	102	14K (11K)	1.6 (0.2)
MUV	17	15K (1)	0.2 (0)
Tox21	12	6K (500)	7.8 (4.7)

The MUV group contained 17 challenging datasets specifically designed to avoid common pitfalls in virtual screening

Massively Multitask Networks for Drug Discovery

Group	Datasets	Data Points / ea.	% Active
PCBA	128	282K (122K)	1.8 (3.8)
DUD-E	102	14K (11K)	1.6 (0.2)
MUV	17	15K (1)	0.2 (0)
Tox21	12	6K (500)	7.8 (4.7)

U.S. Department of Health & Human Services | National Institutes of Health

NIH National Center for Advancing Translational Sciences

Tox21 Data Challenge 2014

estrogen receptor alpha, LBD (ER, LBD)
estrogen receptor alpha, full (ER, full)
aromatase
aryl hydrocarbon receptor (AhR)
androgen receptor, full (AR, full)
androgen receptor, LBD (AR, LBD)
peroxisome proliferator-activated receptor gamma (PPAR-gamma)
nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element (Nrf2/ARE)
heat shock factor response element (HSE)
ATAD5
mitochondrial membrane potential (MMP)
p53

data set	type	description	number of molecules	number of unique descriptors
Training data set (ADME=absorption, distribution, metabolism, and excretion activities)				
3A4	ADME	CYP P450 3A4 inhibition –log(IC50) M	50000	9491
CB1	target	binding to cannabinoid receptor 1 –log(IC50) M	11640	5877
DPP4	target	inhibition of dipeptidyl peptidase 4 –log(IC50) M	8327	5203
HIVINT	target	inhibition of HIV integrase in a cell based assay –log(IC50) M	2421	4306
HIVPROT	target	inhibition of HIV protease –log(IC50) M	4311	6274
LOGD	ADME	logD measured by HPLC method	50000	8921
METAB	ADME	percent remaining after 30 min microsomal incubation	2092	4595
NK1	target	inhibition of neurokinin1 (substance P) receptor binding –log(IC50) M	13482	5803
OX1	target	inhibition of orexin 1 receptor –log(K_i) M	7135	4730
OX2	target	inhibition of orexin 2 receptor –log(K_i) M	14875	5790
PGP	ADME	transport by *p*-glycoprotein log(BA/AB)	8603	5135
PPB	ADME	human plasma protein binding log(bound/unbound)	11622	5470
RAT_F	ADME	log(rat bioavailability) at 2 mg/kg	7821	5698
TDI	ADME	time dependent 3A4 inhibitions log(IC50 without NADPH/IC50 with NADPH)	5559	5945
THROMBIN	target	human thrombin inhibition –log(IC50) M	6924	5552
Validation Data Sets				
2C8	ADME	CYP P450 2C8 inhibition –log(IC50) M	29958	8217
2C9	ADME	CYP P450 2C9 inhibition –log(IC50) M	189670	11730
2D6	ADME	CYP P450 2D6 inhibition –log(IC50) M	50000	9729
A-II	target	binding to Angiotensin-II receptor –log(IC50) M	2763	5242
BACE	target	inhibition of beta-secretase –log(IC50) M	17469	6200
CAV	ADME	inhibition of Cav1.2 ion channel	50000	8959
CLINT	ADME	clearance by human microsome log(clearance) μL/min·mg	23292	6782
ERK2	target	inhibition of ERK2 kinase –log(IC50) M	12843	6596
FACTORXIA	target	inhibition of factor Xla –log(IC50) M	9536	6136
FASSIF	ADME	solubility in simulated gut conditions log(solubility) mol/L	89531	9541
HERG	ADME	inhibition of hERG channel –log(IC50) M	50000	9388
HERG (full data set)	ADME	inhibition of hERG ion channel –log(IC50) M	318795	12508
NAV	ADME	inhibition of Nav1.5 ion channel –log(IC50) M	50000	8302
PAPP	ADME	apparent passive permeability in PK1 cells log(permeability) cm/s	30938	7713
PXR	ADME	induction of 3A4 by pregnane X receptor; percentage relative to rifampicin	50000	9282

Massively Multitask Networks for Drug Discovery

The DNN

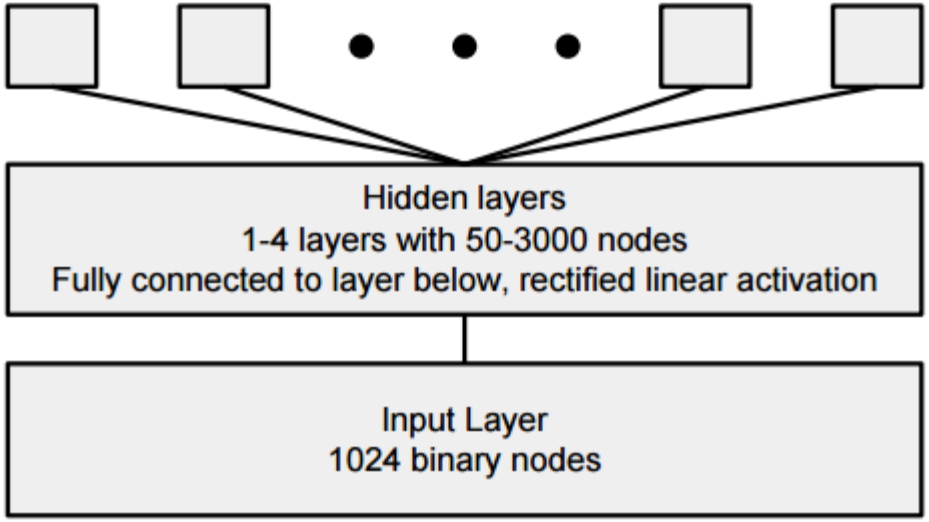


Figure 1. Multitask neural network.

Massively Multitask Networks for Drug Discovery

Metrics:

The metric to evaluate prediction performance is R^2, which is the squared Pearson correlation coefficient between predicted and observed activities in the test set

AUC The closer the AUC score is to 1.0, the better the model is at distinguishing samples in one of the classes

Massively Multitask Networks for Drug Discovery

What were they aiming in their research:

1. Do massively multitask networks provide a performance boost over simple machine learning methods?
2. How does the performance of a multitask network depend on the number of tasks? What are the crucial parameters?
3. Do massively multitask networks extract generalizable information about chemical space?
4. When do datasets benefit from multitask training?

Massively Multitask Networks for Drug Discovery

1. Model performance. Compare Metrics of Random Forest (and other less popular model and DNN

Paper 1 (newest)

Model	PCBA ($n = 128$)	MUV ($n = 17$)	Tox21 ($n = 12$)
Logistic Regression (LR)	.801	.752	.738
→ Random Forest (RF)	.800	.774	.790
Single-Task Neural Net (STNN)	.795	.732	.714
Pyramidal (2000, 100) STNN (PSTNN)	.809	.745	.740
Max{LR, RF, STNN, PSTNN}	.824	.781	.790
1-Hidden (1200) Layer Multitask Neural Net (MTNN)	.842	.797	.785
→ Pyramidal (2000, 100) Multitask Neural Net (PMTNN)	**.873**	**.841**	**.818**

Paper 2

Table 3. Comparing RF with DNN Trained Using Recommended Parameter Settings on 15 Additional Datasets

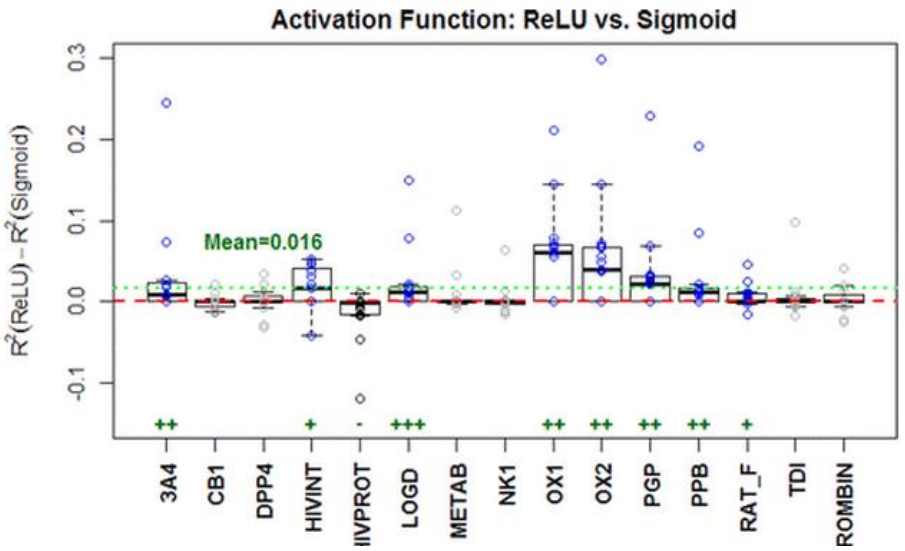
data set	random forest (R^2)	individual DNN (R^2)
2C8	0.158	**0.255**
2C9BIG	0.279	**0.363**
2D6	0.130	**0.195**
A-II	0.805	**0.812**
BACE	0.629	**0.644**
CAV	0.399	**0.463**
CLINT	0.393	**0.554**
ERK2	**0.257**	0.198
FACTORXIA	0.241	**0.244**
FASSIF	**0.294**	0.271
HERG	0.305	**0.352**
HERGfull	0.294	**0.367**
NAV	0.277	**0.347**
PAPP	0.621	**0.678**
PXR	0.333	**0.416**
mean	0.361	***0.411***

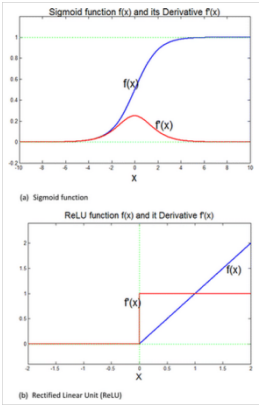
Both papers report significant improvement of the Metrics in DNN

Massively Multitask Networks for Drug Discovery

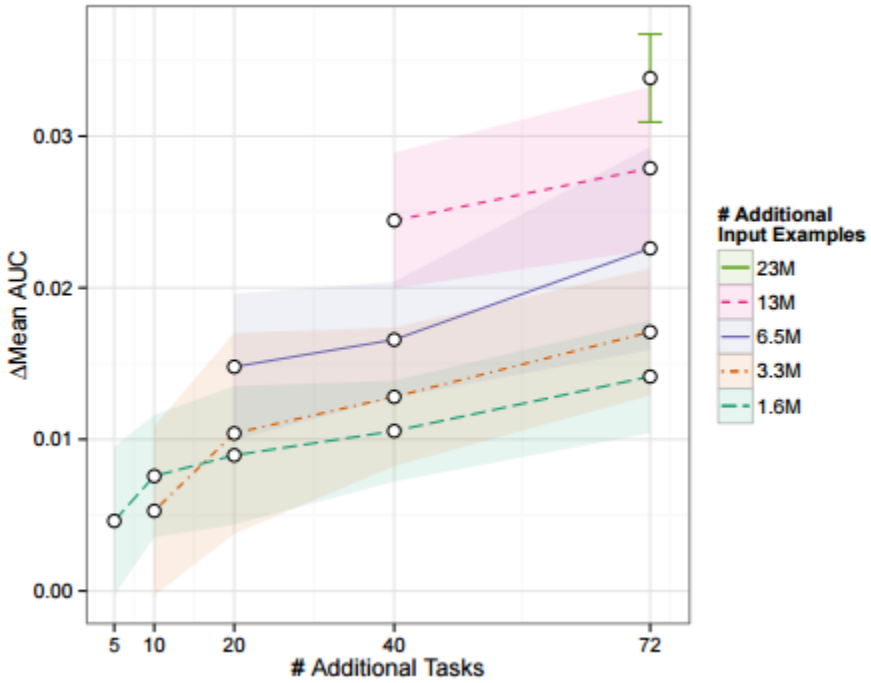
2. Model performance. Crucial parameters

Activation function



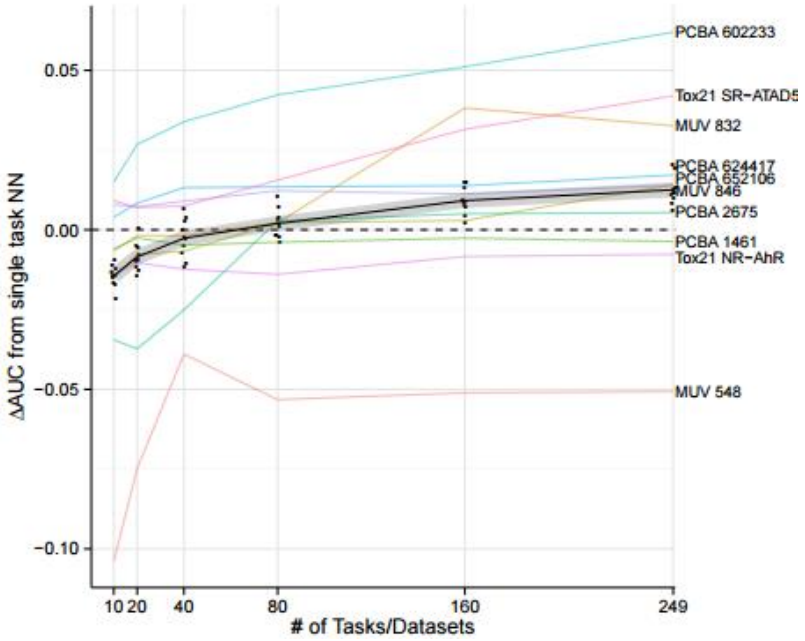


Number of tasks

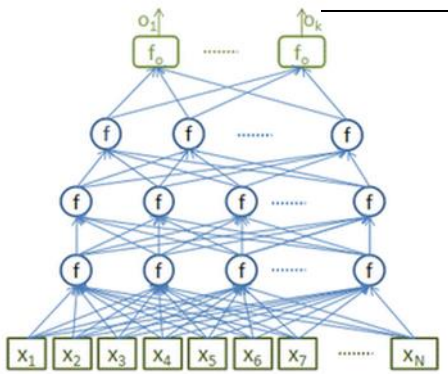


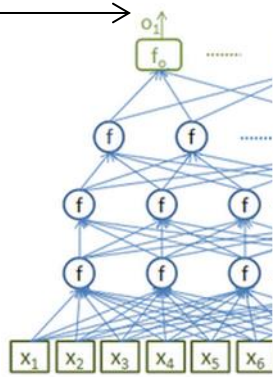
Massively Multitask Networks for Drug Discovery

3. Is the data generalizable



This plot represents the efficiency of the features of the resulting objective function to the dataset excluded from training in single task NN.

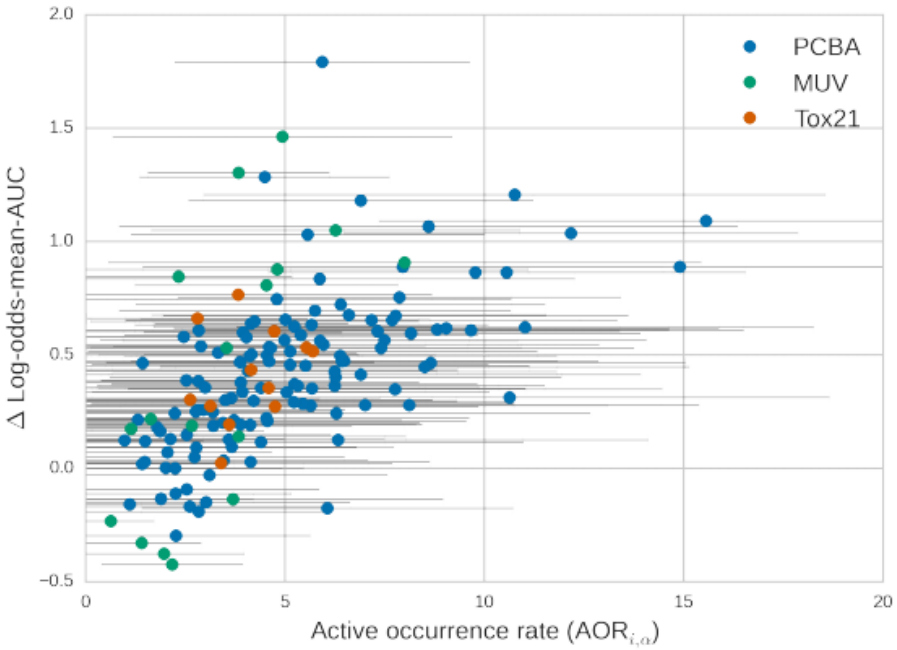




Massively Multitask Networks for Drug Discovery

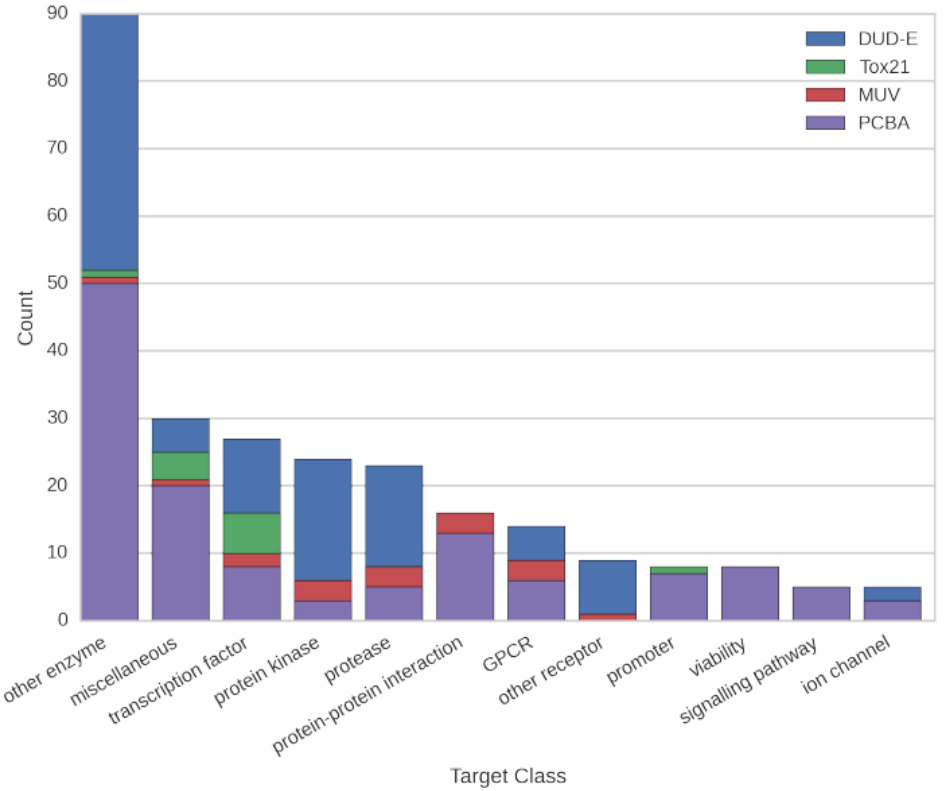
4. “Cross-talk” between databases more visible in multitask DNN

Active compounds are better distinguished from not-active



For each active compound α in dataset D_i, $\mathrm{AOR}_{i,\alpha}$ is defined as the number of additional datasets in which this compound is also active:

$$\mathrm{AOR}_{i,\alpha} = \sum_{d \neq i} \mathbb{1}(\alpha \in \text{Actives}\,(D_d)).$$



PBCA The PubChem BioAssay database currently contains 500,000 descriptions of assay protocols, covering 5000 protein targets, 30,000 gene targets and providing over 130 million bioactivity outcomes. PubChem's bioassay data are integrated into the NCBI Entrez information retrieval system, thus making PubChem data searchable and accessible by Entrez queries.

DUD-E group contained 102 datasets that were designed for the evaluation of methods to predict interactions between proteins and small molecules

MUV group contained 17 challenging datasets specifically designed to avoid common pitfalls in virtual screening

Massively Multitask Networks for Drug Discovery

Deep Neural Networks

DNNs have proven efficacy and already implemented in our daily life, ask Siri or OK, Google (natural language processing) about face recognition and movement recognition (Kinect)

Currently DNNs are in the development stage for the QSAR.
Their use will allow to screen the novel drugs for efficacy and toxicity across the complete knowledge base.

The hardware (i.e., a computer with graphic processing unit capability) used in this study costs about USD $4,000, and the supporting software is free.

OK GOOGLE
Find new drugs!