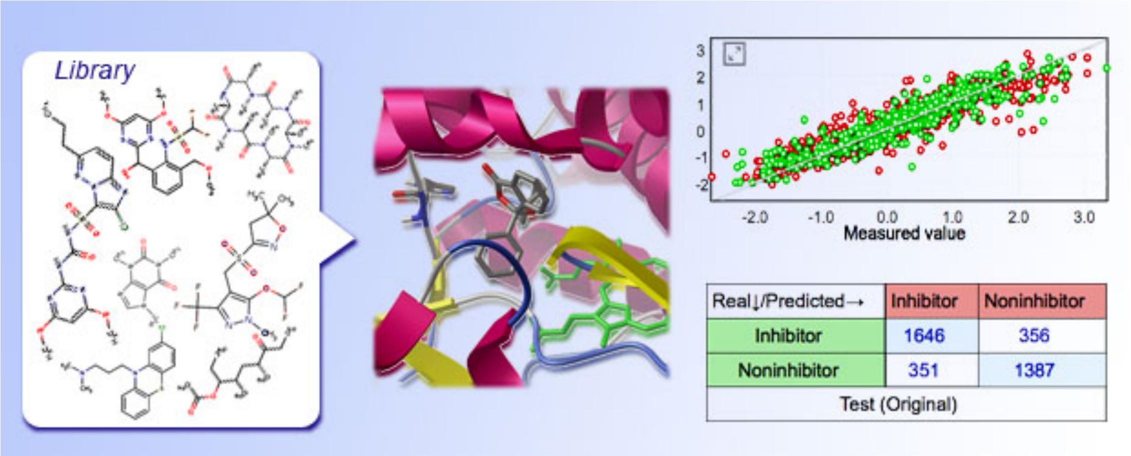


Real↓/Predicted→	Inhibitor	Noninhibitor
Inhibitor	1646	356
Noninhibitor	351	1387

Test (Original)

(Big) Data analysis using On-line Chemical database and Modelling platform

Dr. Igor V. Tetko

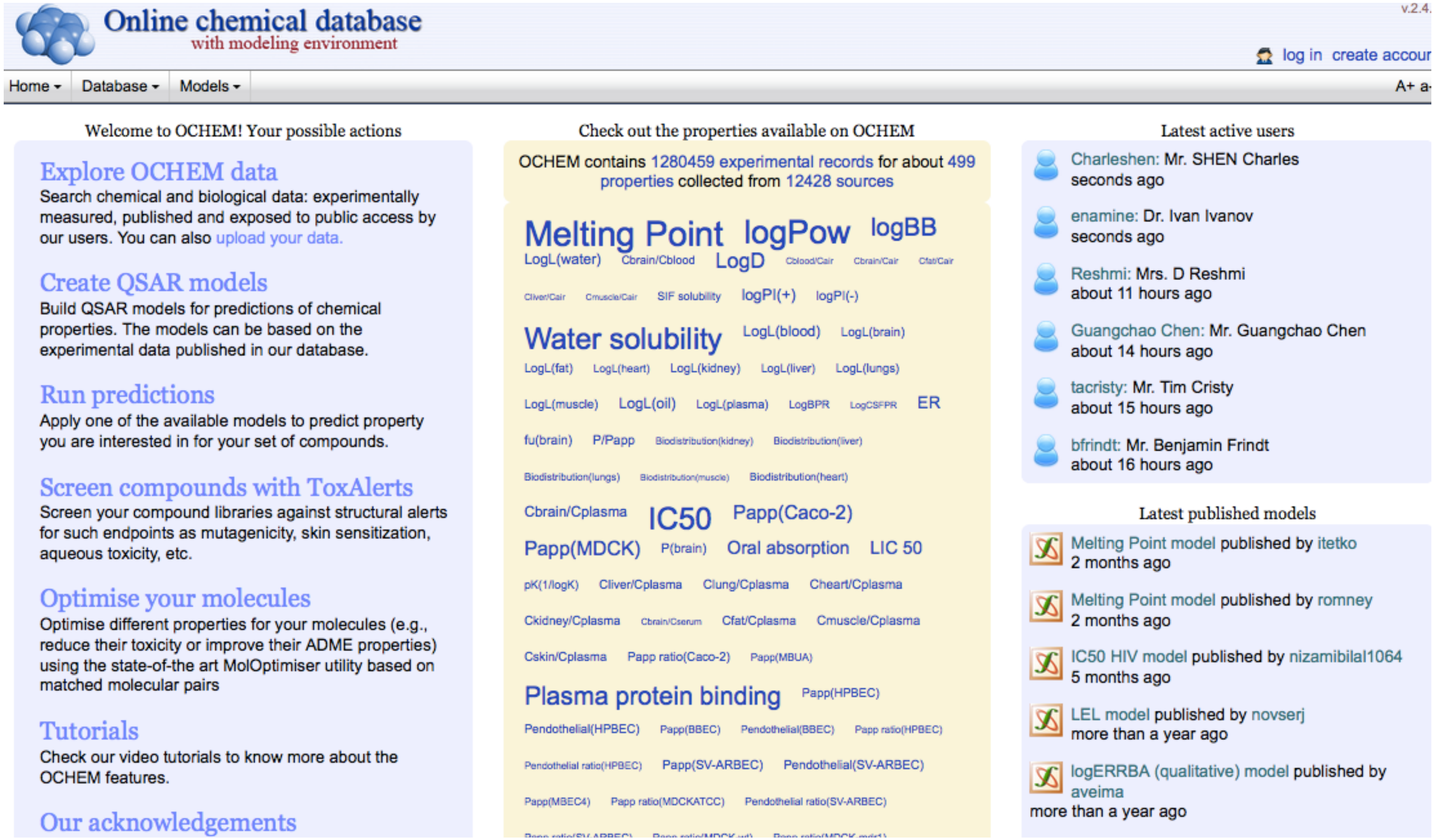
Institute of Structural Biology, Helmholtz Zentrum München & BIGCHEM GmbH

September 14, 2018, EPFL, Lausanne

HelmholtzZentrum münchen
German Research Center for Environmental Health

big
chem

Data storage and model development: http://ochem.eu



Welcome to OCHEM! Your possible actions

Explore OCHEM data
Search chemical and biological data: experimentally measured, published and exposed to public access by our users. You can also upload your data.

Create QSAR models
Build QSAR models for predictions of chemical properties. The models can be based on the experimental data published in our database.

Run predictions
Apply one of the available models to predict property you are interested in for your set of compounds.

Screen compounds with ToxAlerts
Screen your compound libraries against structural alerts for such endpoints as mutagenicity, skin sensitization, aqueous toxicity, etc.

Optimise your molecules
Optimise different properties for your molecules (e.g., reduce their toxicity or improve their ADME properties) using the state-of-the art MolOptimiser utility based on matched molecular pairs

Tutorials
Check our video tutorials to know more about the OCHEM features.

Our acknowledgements

Check out the properties available on OCHEM

OCHEM contains 1280459 experimental records for about 499 properties collected from 12428 sources

Melting Point logPow logBB LogL(water) Cbrain/Cblood LogD Cblood/Cair Cbrain/Cair Cfat/Cair Cliver/Cair Cmuscle/Cair SIF solubility logPI(+) logPI(-) Water solubility LogL(blood) LogL(brain) LogL(fat) LogL(heart) LogL(kidney) LogL(liver) LogL(lungs) LogL(muscle) LogL(oil) LogL(plasma) LogBPR LogCSFPR ER fu(brain) P/Papp Biodistribution(kidney) Biodistribution(liver) Biodistribution(lungs) Biodistribution(muscle) Biodistribution(heart) Cbrain/Cplasma IC50 Papp(Caco-2) Papp(MDCK) P(brain) Oral absorption LIC 50 pK(1/logK) Cliver/Cplasma Clung/Cplasma Cheart/Cplasma Ckidney/Cplasma Cbrain/Cserum Cfat/Cplasma Cmuscle/Cplasma Cskin/Cplasma Papp ratio(Caco-2) Papp(MBUA) Plasma protein binding Papp(HPBEC) Pendothelial(HPBEC) Papp(BBEC) Pendothelial(BBEC) Papp ratio(HPBEC) Pendothelial ratio(HPBEC) Papp(SV-ARBEC) Pendothelial(SV-ARBEC) Papp(MBEC4) Papp ratio(MDCKATCC) Pendothelial ratio(SV-ARBEC) Papp ratio(SV-ARBEC) Papp ratio(MDCK-wt) Papp ratio(MDCK-mdr1)

Latest active users

Charleshen: Mr. SHEN Charles
seconds ago

enamine: Dr. Ivan Ivanov
seconds ago

Reshmi: Mrs. D Reshmi
about 11 hours ago

Guangchao Chen: Mr. Guangchao Chen
about 14 hours ago

tacristy: Mr. Tim Cristy
about 15 hours ago

bfrindt: Mr. Benjamin Frindt
about 16 hours ago

Latest published models

Melting Point model published by itetko
2 months ago

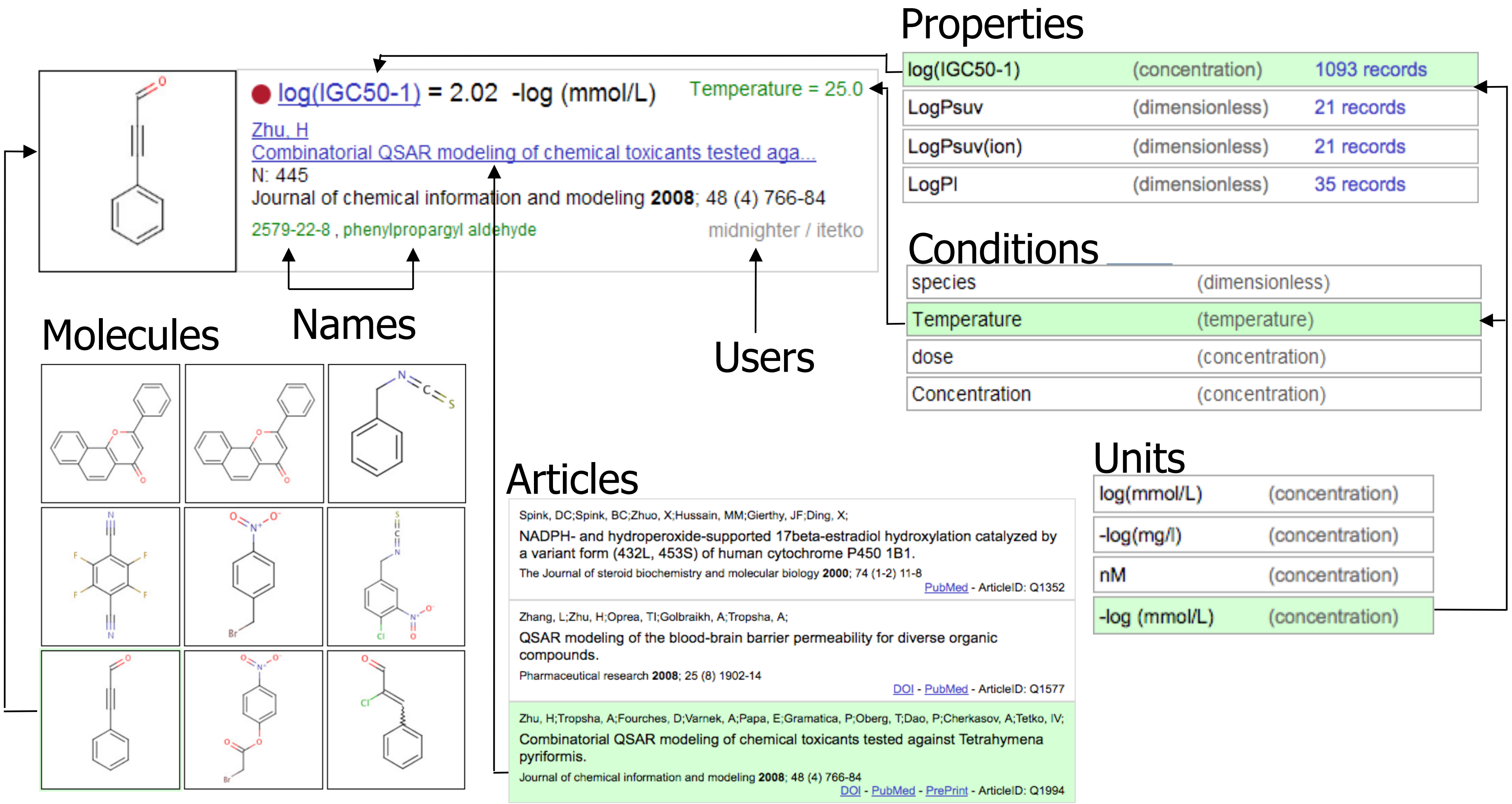
Melting Point model published by romney
2 months ago

IC50 HIV model published by nizamibilal1064
5 months ago

LEL model published by novserj
more than a year ago

logERRBA (qualitative) model published by aveima
more than a year ago

OCHEM Database schema



log(IGC50-1) = 2.02 -log (mmol/L) Temperature = 25.0

Zhu, H

Combinatorial QSAR modeling of chemical toxicants tested aga...

N: 445

Journal of chemical information and modeling **2008**; 48 (4) 766-84

2579-22-8 , phenylpropargyl aldehyde

midnighter / itetko

Properties

log(IGC50-1)	(concentration)	1093 records
LogPsuv	(dimensionless)	21 records
LogPsuv(ion)	(dimensionless)	21 records
LogPI	(dimensionless)	35 records

Conditions

species	(dimensionless)
Temperature	(temperature)
dose	(concentration)
Concentration	(concentration)

Articles

Spink, DC;Spink, BC;Zhuo, X;Hussain, MM;Gierthy, JF;Ding, X;

NADPH- and hydroperoxide-supported 17beta-estradiol hydroxylation catalyzed by a variant form (432L, 453S) of human cytochrome P450 1B1.

The Journal of steroid biochemistry and molecular biology **2000**; 74 (1-2) 11-8

PubMed - ArticleID: Q1352

Zhang, L;Zhu, H;Oprea, TI;Golbraikh, A;Tropsha, A;

QSAR modeling of the blood-brain barrier permeability for diverse organic compounds.

Pharmaceutical research **2008**; 25 (8) 1902-14

DOI - PubMed - ArticleID: Q1577

Zhu, H;Tropsha, A;Fourches, D;Varnek, A;Papa, E;Gramatica, P;Oberg, T;Dao, P;Cherkasov, A;Tetko, IV;

Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis.

Journal of chemical information and modeling **2008**; 48 (4) 766-84

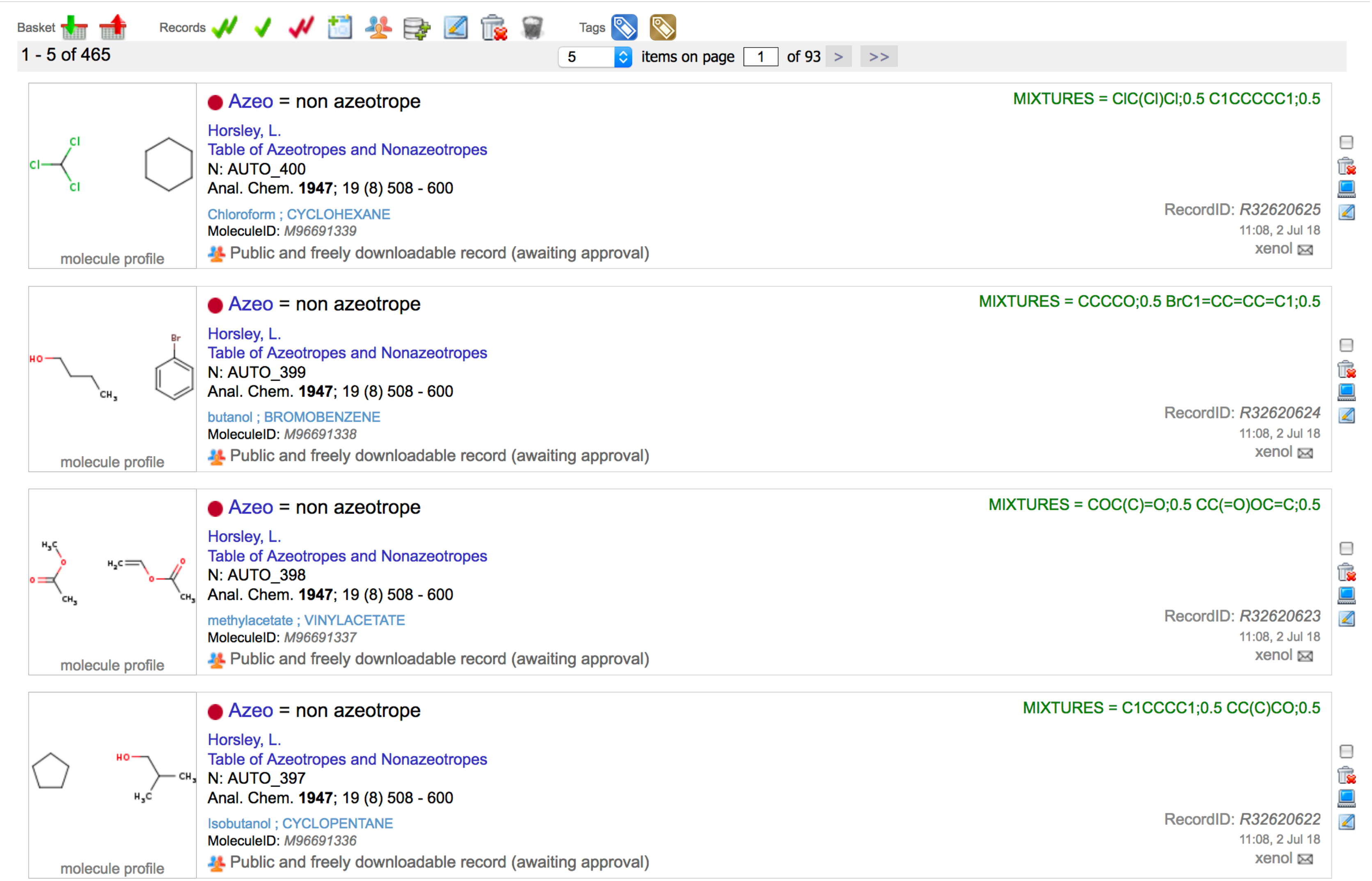
DOI - PubMed - PrePrint - ArticleID: Q1994

Units

log(mmol/L)	(concentration)
-log(mg/l)	(concentration)
nM	(concentration)
-log (mmol/L)	(concentration)

CC-BY license for the uploaded data

Support of mixtures



1 - 5 of 465 | 5 items on page 1 of 93 > >>

● Azeo = non azeotrope

MIXTURES = ClC(Cl)Cl;0.5 C1CCCCC1;0.5

Horsley, L.
Table of Azeotropes and Nonazeotropes
N: AUTO_400
Anal. Chem. **1947**; 19 (8) 508 - 600
Chloroform ; CYCLOHEXANE
MoleculeID: *M96691339*
Public and freely downloadable record (awaiting approval)

RecordID: *R32620625*
11:08, 2 Jul 18
xenol

molecule profile

● Azeo = non azeotrope

MIXTURES = CCCCO;0.5 BrC1=CC=CC=C1;0.5

Horsley, L.
Table of Azeotropes and Nonazeotropes
N: AUTO_399
Anal. Chem. **1947**; 19 (8) 508 - 600
butanol ; BROMOBENZENE
MoleculeID: *M96691338*
Public and freely downloadable record (awaiting approval)

RecordID: *R32620624*
11:08, 2 Jul 18
xenol

molecule profile

● Azeo = non azeotrope

MIXTURES = COC(C)=O;0.5 CC(=O)OC=C;0.5

Horsley, L.
Table of Azeotropes and Nonazeotropes
N: AUTO_398
Anal. Chem. **1947**; 19 (8) 508 - 600
methylacetate ; VINYLACETATE
MoleculeID: *M96691337*
Public and freely downloadable record (awaiting approval)

RecordID: *R32620623*
11:08, 2 Jul 18
xenol

molecule profile

● Azeo = non azeotrope

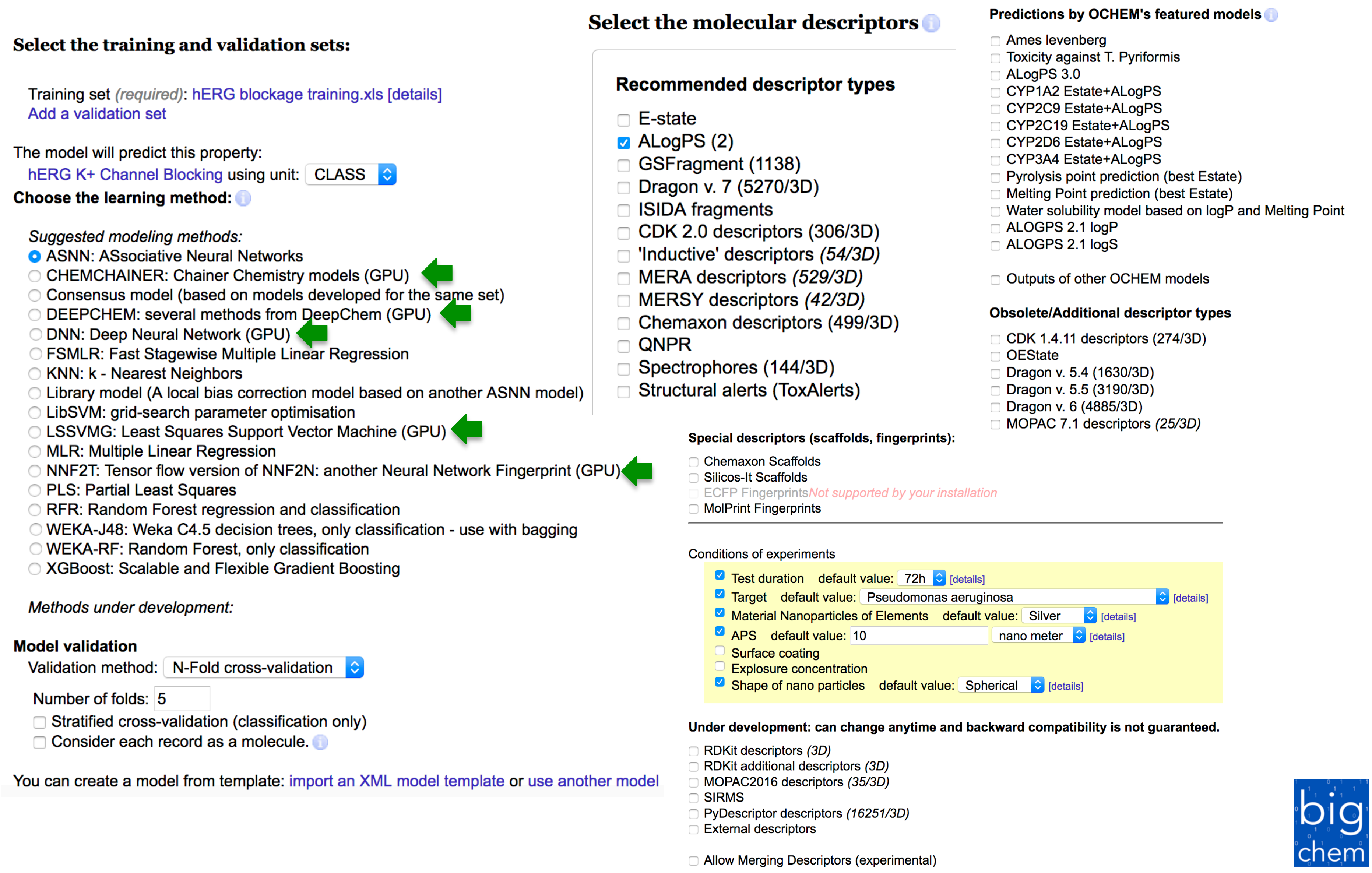
MIXTURES = C1CCCC1;0.5 CC(C)CO;0.5

Horsley, L.
Table of Azeotropes and Nonazeotropes
N: AUTO_397
Anal. Chem. **1947**; 19 (8) 508 - 600
Isobutanol ; CYCLOPENTANE
MoleculeID: *M96691336*
Public and freely downloadable record (awaiting approval)

RecordID: *R32620622*
11:08, 2 Jul 18
xenol

molecule profile

QSPR/QSAR modelling in OCHEM



Select the training and validation sets:

Training set *(required)*: hERG blockage training.xls [details]
Add a validation set

The model will predict this property:
hERG K+ Channel Blocking using unit: CLASS

Choose the learning method:

Suggested modeling methods:

- ● ASNN: ASsociative Neural Networks
- ○ CHEMCHAINER: Chainer Chemistry models (GPU)
- ○ Consensus model (based on models developed for the same set)
- ○ DEEPCHEM: several methods from DeepChem (GPU)
- ○ DNN: Deep Neural Network (GPU)
- ○ FSMLR: Fast Stagewise Multiple Linear Regression
- ○ KNN: k - Nearest Neighbors
- ○ Library model (A local bias correction model based on another ASNN model)
- ○ LibSVM: grid-search parameter optimisation
- ○ LSSVMG: Least Squares Support Vector Machine (GPU)
- ○ MLR: Multiple Linear Regression
- ○ NNF2T: Tensor flow version of NNF2N: another Neural Network Fingerprint (GPU)
- ○ PLS: Partial Least Squares
- ○ RFR: Random Forest regression and classification
- ○ WEKA-J48: Weka C4.5 decision trees, only classification - use with bagging
- ○ WEKA-RF: Random Forest, only classification
- ○ XGBoost: Scalable and Flexible Gradient Boosting

Methods under development:

Model validation

Validation method: N-Fold cross-validation

Number of folds: 5

□ Stratified cross-validation (classification only)
□ Consider each record as a molecule.

You can create a model from template: import an XML model template or use another model

Select the molecular descriptors

Recommended descriptor types

- □ E-state
- ☑ ALogPS (2)
- □ GSFragment (1138)
- □ Dragon v. 7 (5270/3D)
- □ ISIDA fragments
- □ CDK 2.0 descriptors (306/3D)
- □ 'Inductive' descriptors *(54/3D)*
- □ MERA descriptors *(529/3D)*
- □ MERSY descriptors *(42/3D)*
- □ Chemaxon descriptors (499/3D)
- □ QNPR
- □ Spectrophores (144/3D)
- □ Structural alerts (ToxAlerts)

Special descriptors (scaffolds, fingerprints):

- □ Chemaxon Scaffolds
- □ Silicos-It Scaffolds
- □ ECFP Fingerprints*Not supported by your installation*
- □ MolPrint Fingerprints

Conditions of experiments

- ☑ Test duration default value: 72h [details]
- ☑ Target default value: Pseudomonas aeruginosa [details]
- ☑ Material Nanoparticles of Elements default value: Silver [details]
- ☑ APS default value: 10 nano meter [details]
- □ Surface coating
- □ Explosure concentration
- ☑ Shape of nano particles default value: Spherical [details]

Under development: can change anytime and backward compatibility is not guaranteed.

- □ RDKit descriptors *(3D)*
- □ RDKit additional descriptors *(3D)*
- □ MOPAC2016 descriptors *(35/3D)*
- □ SIRMS
- □ PyDescriptor descriptors *(16251/3D)*
- □ External descriptors

□ Allow Merging Descriptors (experimental)

Predictions by OCHEM's featured models

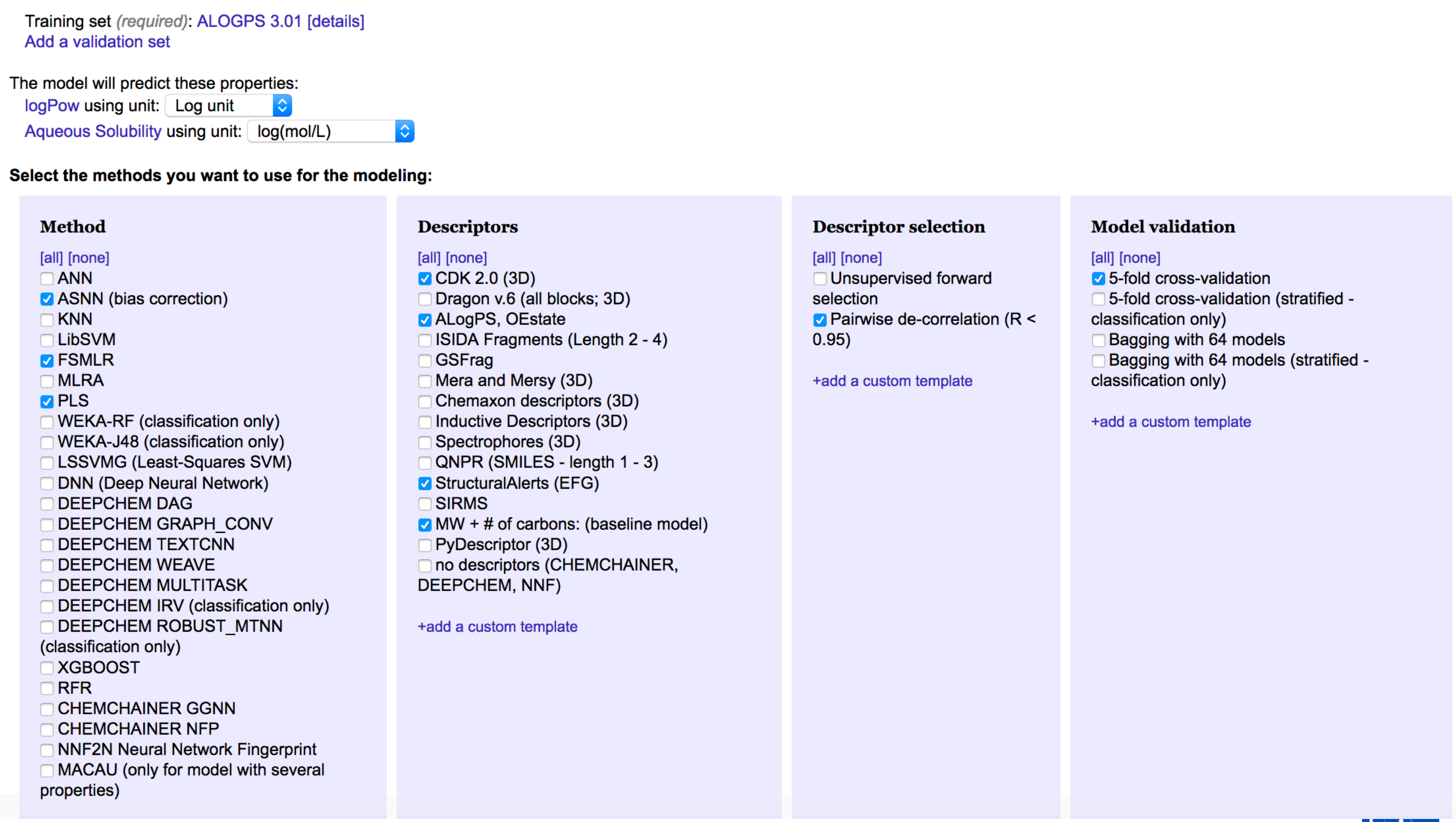
- □ Ames levenberg
- □ Toxicity against T. Pyriformis
- □ ALogPS 3.0
- □ CYP1A2 Estate+ALogPS
- □ CYP2C9 Estate+ALogPS
- □ CYP2C19 Estate+ALogPS
- □ CYP2D6 Estate+ALogPS
- □ CYP3A4 Estate+ALogPS
- □ Pyrolysis point prediction (best Estate)
- □ Melting Point prediction (best Estate)
- □ Water solubility model based on logP and Melting Point
- □ ALOGPS 2.1 logP
- □ ALOGPS 2.1 logS

□ Outputs of other OCHEM models

Obsolete/Additional descriptor types

- □ CDK 1.4.11 descriptors (274/3D)
- □ OEState
- □ Dragon v. 5.4 (1630/3D)
- □ Dragon v. 5.5 (3190/3D)
- □ Dragon v. 6 (4885/3D)
- □ MOPAC 7.1 descriptors *(25/3D)*

Comprehensive Modeling



Training set *(required)*: ALOGPS 3.01 [details]
Add a validation set

The model will predict these properties:
logPow using unit: Log unit
Aqueous Solubility using unit: log(mol/L)

Select the methods you want to use for the modeling:

Method

[all] [none]
- □ ANN
- ☑ ASNN (bias correction)
- □ KNN
- □ LibSVM
- ☑ FSMLR
- □ MLRA
- ☑ PLS
- □ WEKA-RF (classification only)
- □ WEKA-J48 (classification only)
- □ LSSVMG (Least-Squares SVM)
- □ DNN (Deep Neural Network)
- □ DEEPCHEM DAG
- □ DEEPCHEM GRAPH_CONV
- □ DEEPCHEM TEXTCNN
- □ DEEPCHEM WEAVE
- □ DEEPCHEM MULTITASK
- □ DEEPCHEM IRV (classification only)
- □ DEEPCHEM ROBUST_MTNN (classification only)
- □ XGBOOST
- □ RFR
- □ CHEMCHAINER GGNN
- □ CHEMCHAINER NFP
- □ NNF2N Neural Network Fingerprint
- □ MACAU (only for model with several properties)

Descriptors

[all] [none]
- ☑ CDK 2.0 (3D)
- □ Dragon v.6 (all blocks; 3D)
- ☑ ALogPS, OEstate
- □ ISIDA Fragments (Length 2 - 4)
- □ GSFrag
- □ Mera and Mersy (3D)
- □ Chemaxon descriptors (3D)
- □ Inductive Descriptors (3D)
- □ Spectrophores (3D)
- □ QNPR (SMILES - length 1 - 3)
- ☑ StructuralAlerts (EFG)
- □ SIRMS
- ☑ MW + # of carbons: (baseline model)
- □ PyDescriptor (3D)
- □ no descriptors (CHEMCHAINER, DEEPCHEM, NNF)

+add a custom template

Descriptor selection

[all] [none]
- □ Unsupervised forward selection
- ☑ Pairwise de-correlation (R < 0.95)

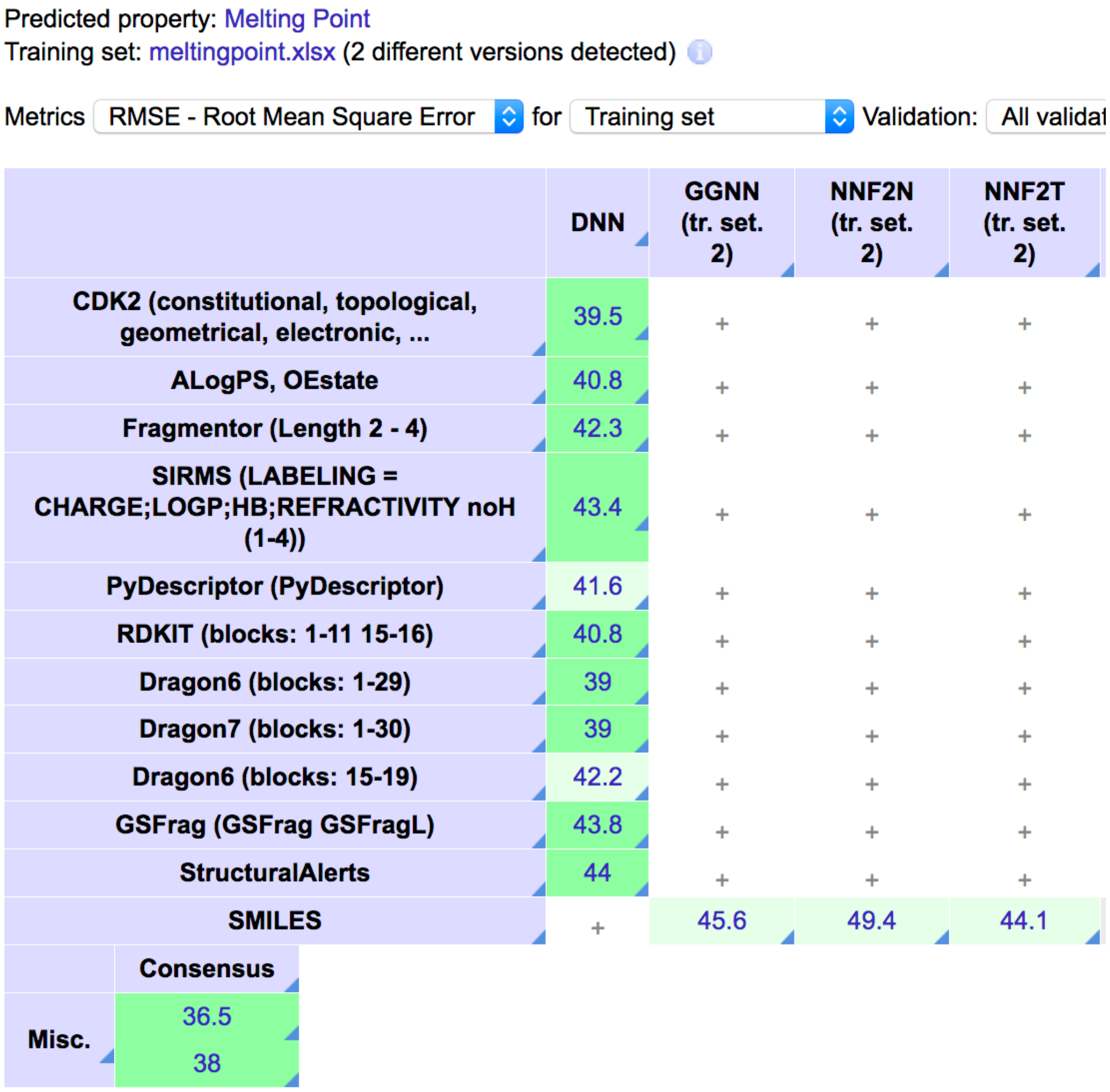
+add a custom template

Model validation

[all] [none]
- ☑ 5-fold cross-validation
- □ 5-fold cross-validation (stratified - classification only)
- □ Bagging with 64 models
- □ Bagging with 64 models (stratified - classification only)

+add a custom template

Comprehensive View



Predicted property: Melting Point

Training set: meltingpoint.xlsx (2 different versions detected)

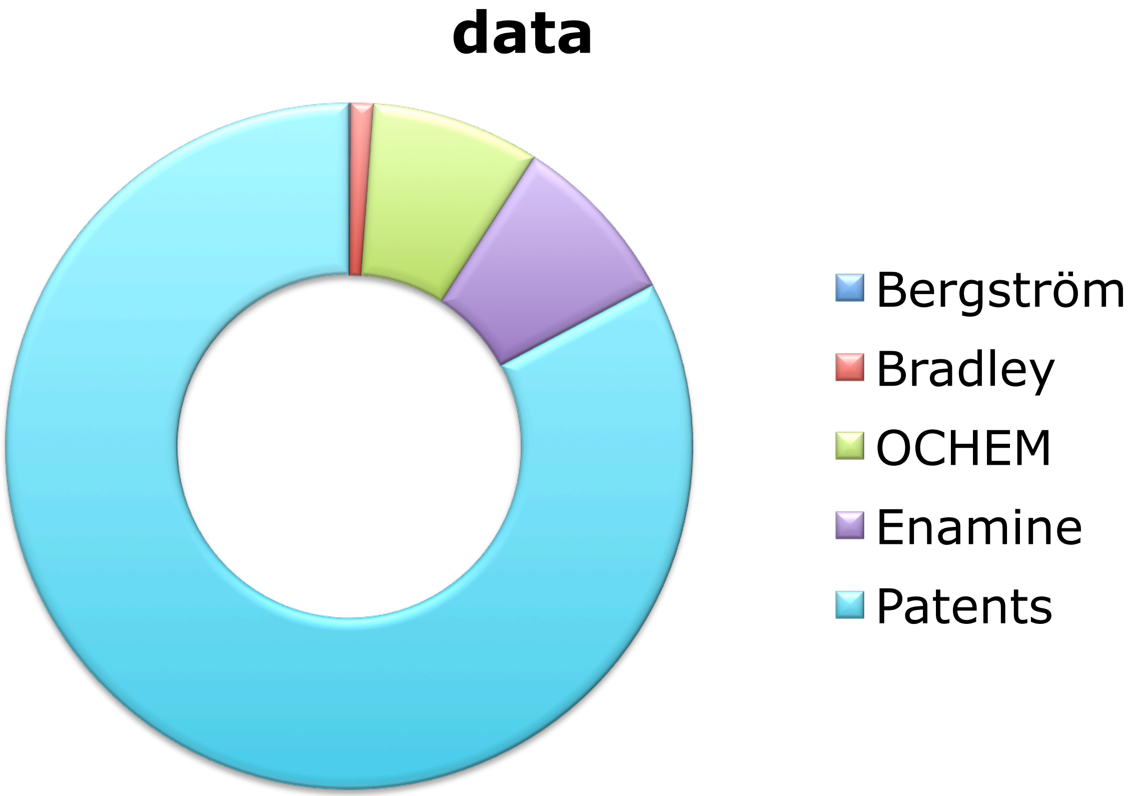
Metrics: RMSE - Root Mean Square Error for Training set Validation: All validat

	DNN	GGNN (tr. set. 2)	NNF2N (tr. set. 2)	NNF2T (tr. set. 2)
CDK2 (constitutional, topological, geometrical, electronic, ...	39.5	+	+	+
ALogPS, OEstate	40.8	+	+	+
Fragmentor (Length 2 - 4)	42.3	+	+	+
SIRMS (LABELING = CHARGE;LOGP;HB;REFRACTIVITY noH (1-4))	43.4	+	+	+
PyDescriptor (PyDescriptor)	41.6	+	+	+
RDKIT (blocks: 1-11 15-16)	40.8	+	+	+
Dragon6 (blocks: 1-29)	39	+	+	+
Dragon7 (blocks: 1-30)	39	+	+	+
Dragon6 (blocks: 15-19)	42.2	+	+	+
GSFrag (GSFrag GSFragL)	43.8	+	+	+
StructuralAlerts	44	+	+	+
SMILES	+	45.6	49.4	44.1

	Consensus
Misc.	36.5 38

275k Melting Point Datasets (Big Data)

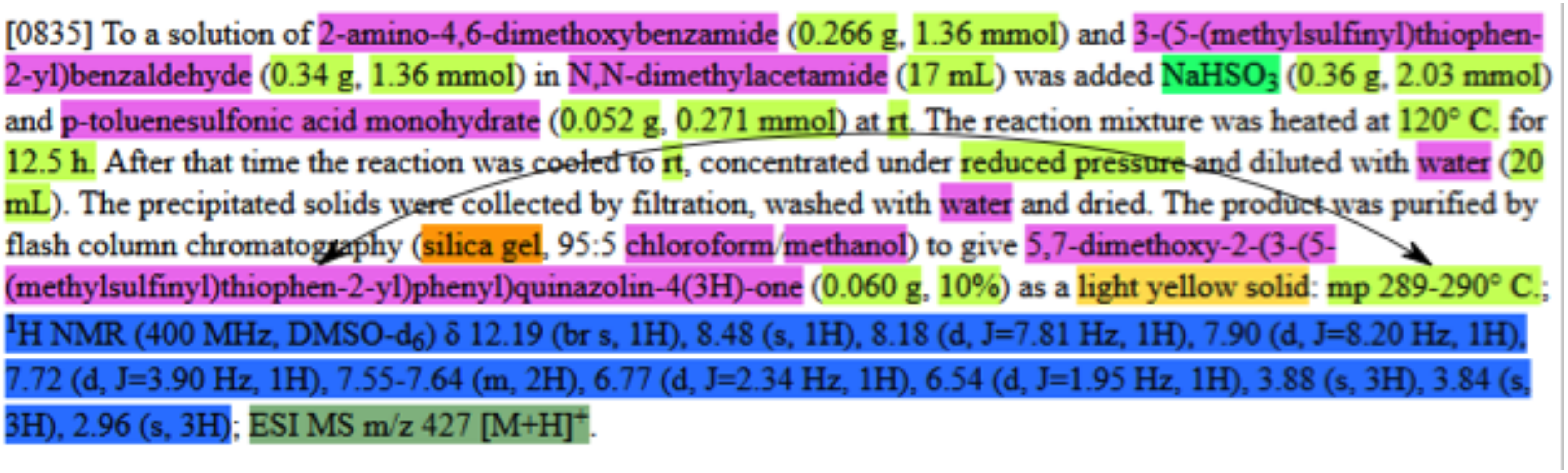
Bergström	277
Bradley	2886
OCHEM	22404
Enamine	21883
PATENTS	228079



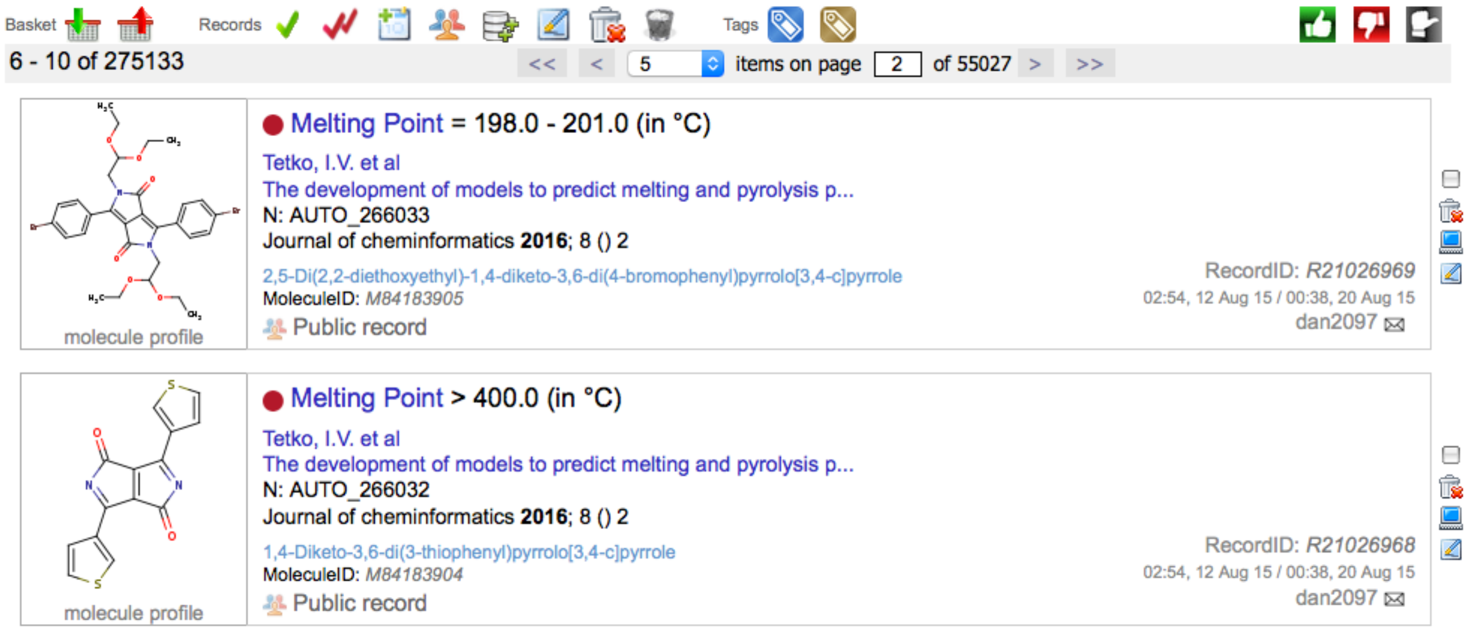
COMBINED: OCHEM + Enamine + Bradley + Bergström

Tetko et al *J. Chemoinformatics, 2016, 8, 2.*

Extraction of MP information from patents



[0835] To a solution of 2-amino-4,6-dimethoxybenzamide (0.266 g, 1.36 mmol) and 3-(5-(methylsulfinyl)thiophen-2-yl)benzaldehyde (0.34 g, 1.36 mmol) in N,N-dimethylacetamide (17 mL) was added $NaHSO_3$ (0.36 g, 2.03 mmol) and p-toluenesulfonic acid monohydrate (0.052 g, 0.271 mmol) at rt. The reaction mixture was heated at 120° C. for 12.5 h. After that time the reaction was cooled to rt, concentrated under reduced pressure and diluted with water (20 mL). The precipitated solids were collected by filtration, washed with water and dried. The product was purified by flash column chromatography (silica gel, 95:5 chloroform/methanol) to give 5,7-dimethoxy-2-(3-(5-(methylsulfinyl)thiophen-2-yl)phenyl)quinazolin-4(3H)-one (0.060 g, 10%) as a light yellow solid: mp 289-290° C.; ^{1}H NMR (400 MHz, DMSO-d_6) δ 12.19 (br s, 1H), 8.48 (s, 1H), 8.18 (d, J=7.81 Hz, 1H), 7.90 (d, J=8.20 Hz, 1H), 7.72 (d, J=3.90 Hz, 1H), 7.55-7.64 (m, 2H), 6.77 (d, J=2.34 Hz, 1H), 6.54 (d, J=1.95 Hz, 1H), 3.88 (s, 3H), 3.84 (s, 3H), 2.96 (s, 3H); ESI MS m/z 427 [M+H]$^+$.



6 - 10 of 275133

● Melting Point = 198.0 - 201.0 (in °C)

Tetko, I.V. et al
The development of models to predict melting and pyrolysis p...
N: AUTO_266033
Journal of cheminformatics **2016**; 8 () 2
2,5-Di(2,2-diethoxyethyl)-1,4-diketo-3,6-di(4-bromophenyl)pyrrolo[3,4-c]pyrrole
MoleculeID: *M84183905*
Public record

RecordID: *R21026969*
02:54, 12 Aug 15 / 00:38, 20 Aug 15
dan2097

molecule profile

● Melting Point > 400.0 (in °C)

Tetko, I.V. et al
The development of models to predict melting and pyrolysis p...
N: AUTO_266032
Journal of cheminformatics **2016**; 8 () 2
1,4-Diketo-3,6-di(3-thiophenyl)pyrrolo[3,4-c]pyrrole
MoleculeID: *M84183904*
Public record

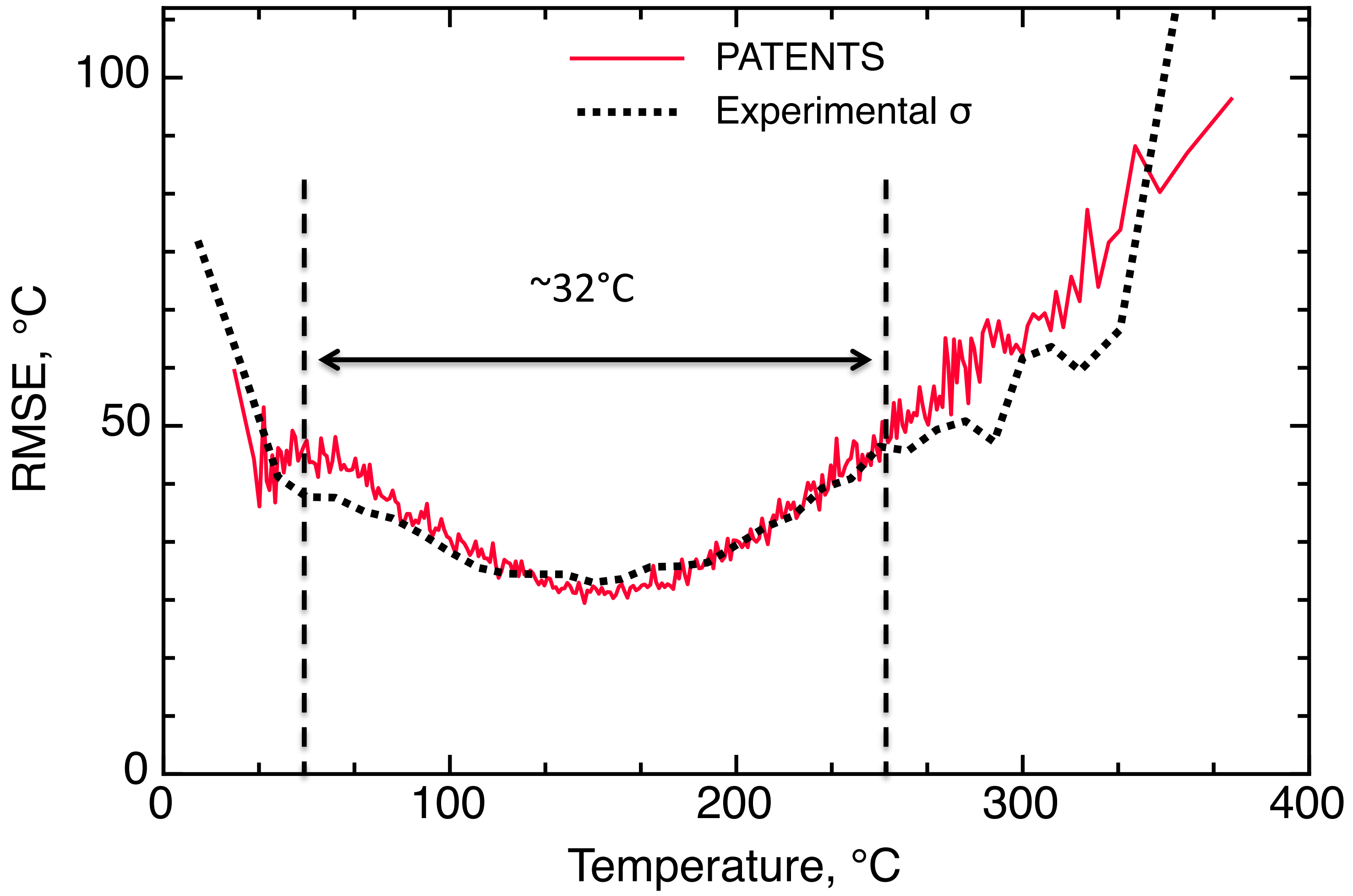
RecordID: *R21026968*
02:54, 12 Aug 15 / 00:38, 20 Aug 15
dan2097

molecule profile

Modeling of MP data

Package name	Type of descriptors	Number of descriptors	Matrix size, billions	Non zero values, millions	Sparseness
Functional Groups	integer	595	0.18	3.1	33
QNPR	integer	1502	0.45	6.3	49
MolPrint	binary	688634	205	8.1	7200
Estate count	float	631	0.19	10	14
Inductive	float	54	0.02	11	1
ECFP4	binary	1024	0.31	12	25
Isida	integer	5886	1.75	18	37
ChemAxon	float	498	0.15	23	1.5
GSFrag	integer	1138	0.34	24	5.7
CDK	float	239	0.07	27	2
Adriana	float	200	0.06	32	1.3
Mera, Mersy	float	571	0.17	61	1.1
Dragon	float	1647	0.49	183	1.5

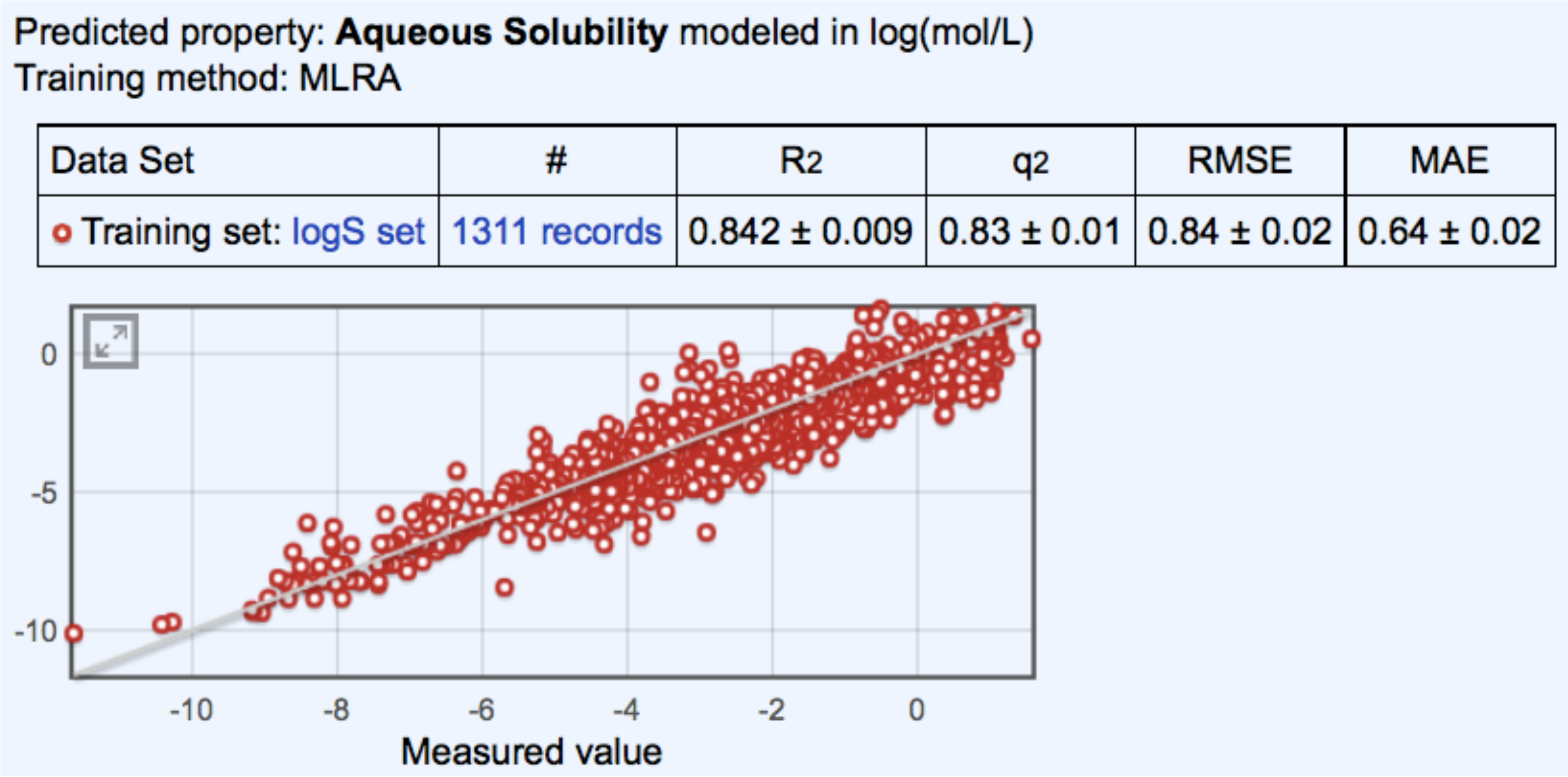
Prediction and experimental errors for consensus model based on the PATENTS set



Experimental accuracy was based on *N* = 18058 duplicated measurements

Prediction of Huuskonen set using ALOGPS logP and MP based on 230k measurements

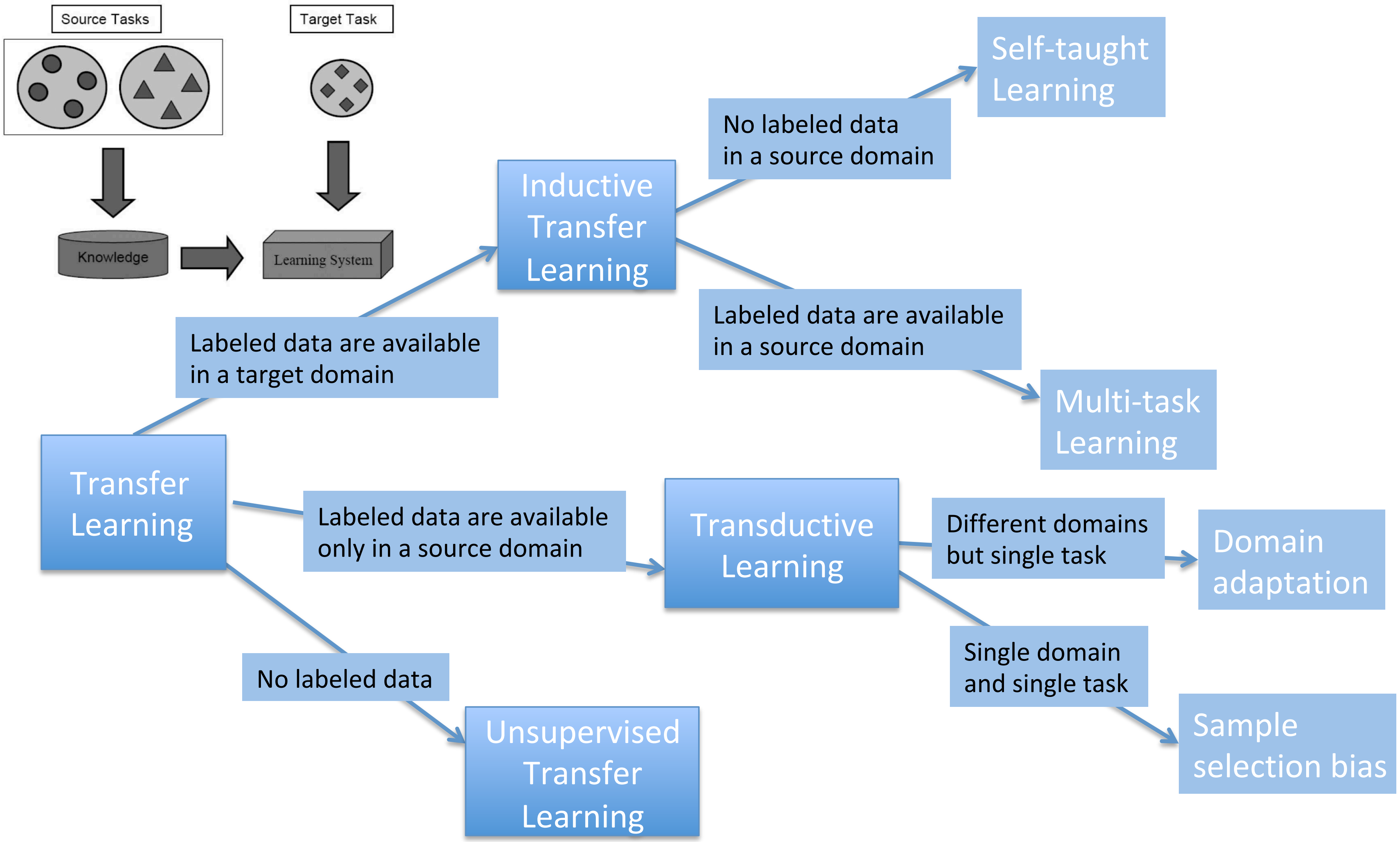
$$\log S = 0.5 - 0.01(MP-25) - \log K_{ow} \; ^{*}$$



Predicted property: **Aqueous Solubility** modeled in log(mol/L)
Training method: MLRA

Data Set	#	R2	q2	RMSE	MAE
○ Training set: logS set	1311 records	0.842 ± 0.009	0.83 ± 0.01	0.84 ± 0.02	0.64 ± 0.02

Feature net model: uses other models as descriptors



Transfer Learning

Labeled data are available in a target domain

Inductive Transfer Learning

No labeled data in a source domain

Self-taught Learning

Labeled data are available in a source domain

Multi-task Learning

Labeled data are available only in a source domain

Transductive Learning

Different domains but single task

Domain adaptation

Single domain and single task

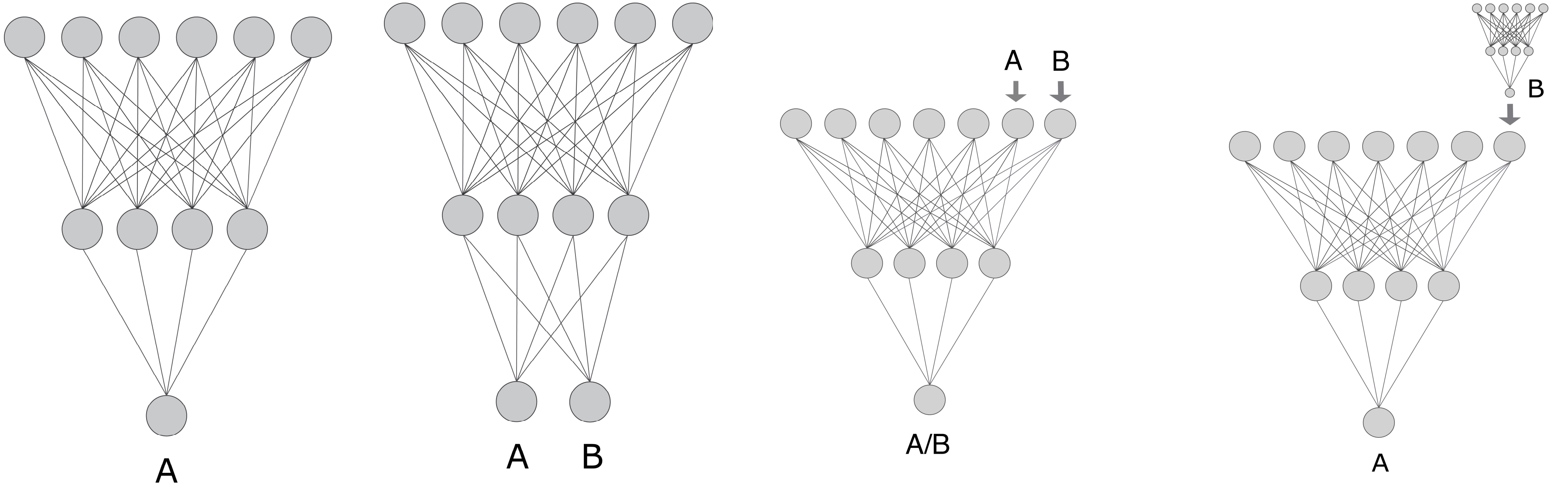
Sample selection bias

No labeled data

Unsupervised Transfer Learning

Adapted from: Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* **2010**, *22*, 1345-1359.

Multi-task learning



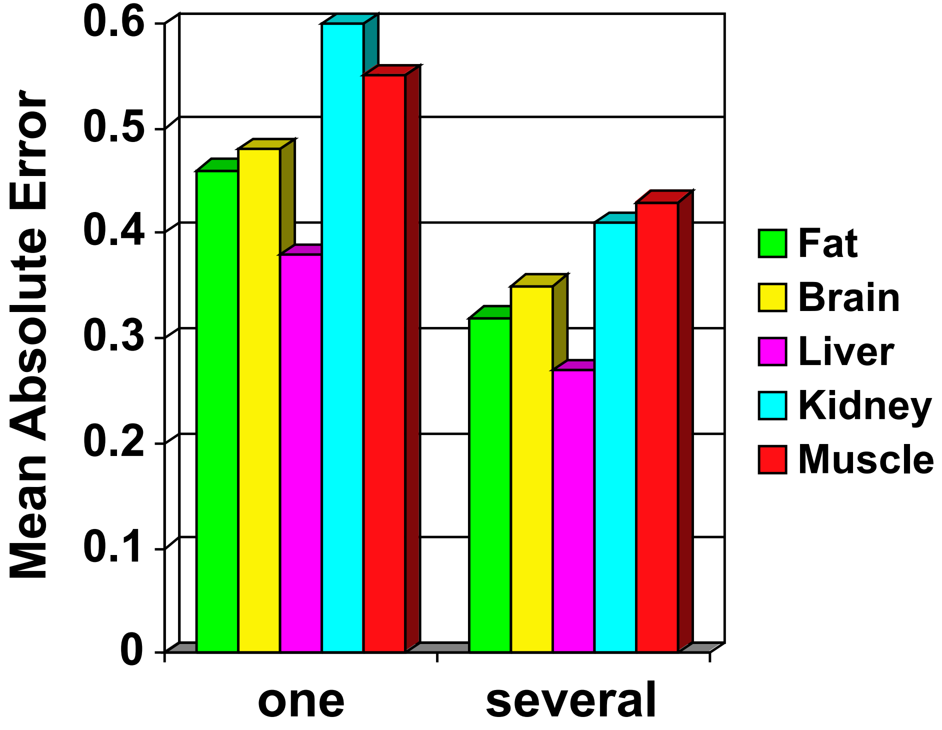
Multi-task learning

Problem:

- prediction of tissue-air partition coefficients
- small datasets 30-100 molecules (human & rat data)

Results:

simultaneous prediction of several properties increased the accuracy of models



Varnek, A. et al J. Chem. Inf. Model. 2009, 49, 133-44.

Prediction of toxicity of chemical compounds:

REGISTRY OF TOXIC EFFECTS OF CHEMICAL SUBSTANCES (RTECS®)

Different species

- Rat
- Mouse
- Rabbit
- …
- Human

~ 129k records
~ 87k compounds
29 properties

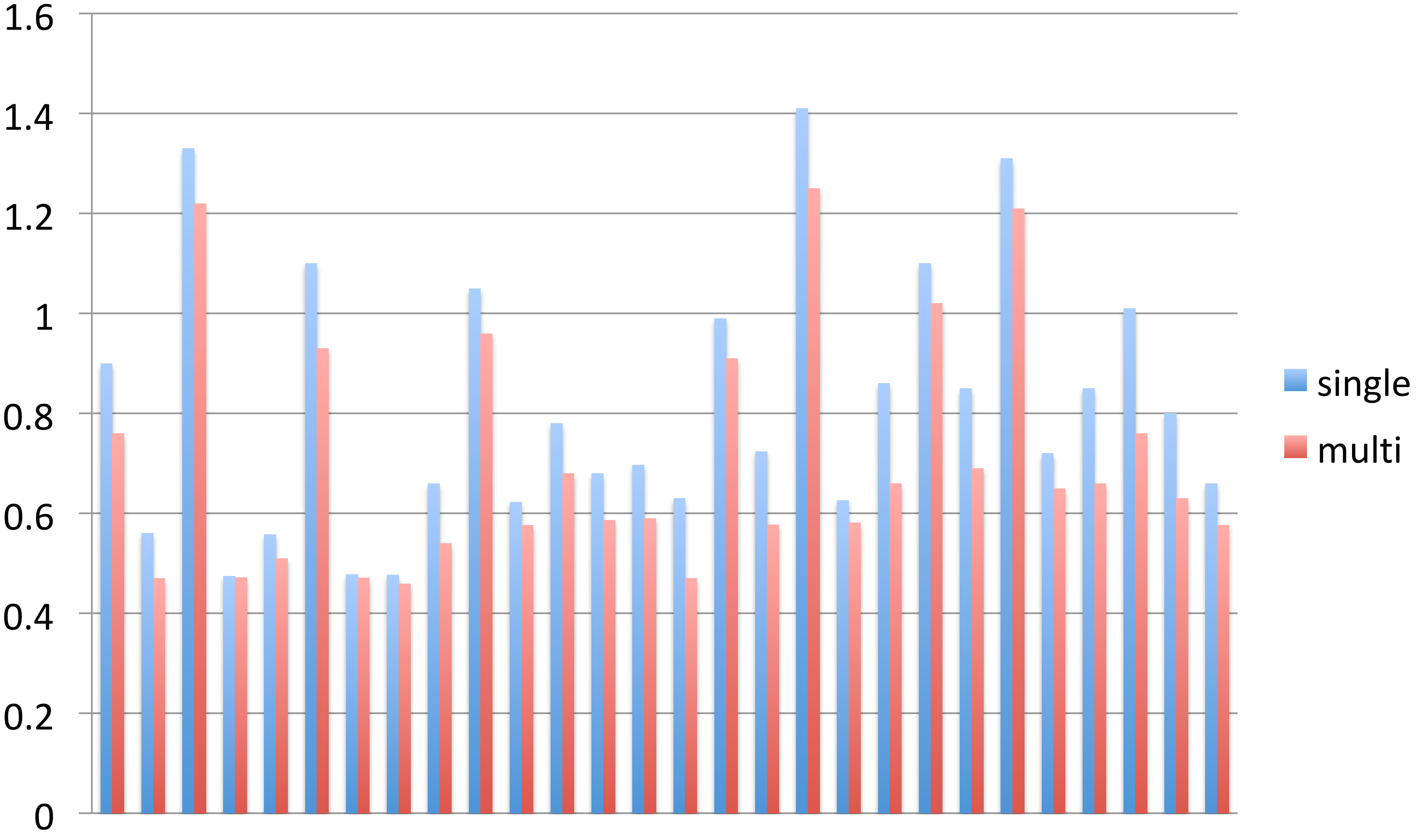
- Different toxicities
  - LD50
  - TDL
  - NOEL
  - LDLo

- Administartion
  - Oral
  - IPR (intraperitoneal)
  - IVR (intravenous)

Sosnin, S.; Karlov, D.; Tetko, I.V.; Fedorov, M.V. A comparative study of prediction of multi-target toxicity for a broad chemical space. *Chem. Res. Toxicol.* 2018, *in prep*.

big chem

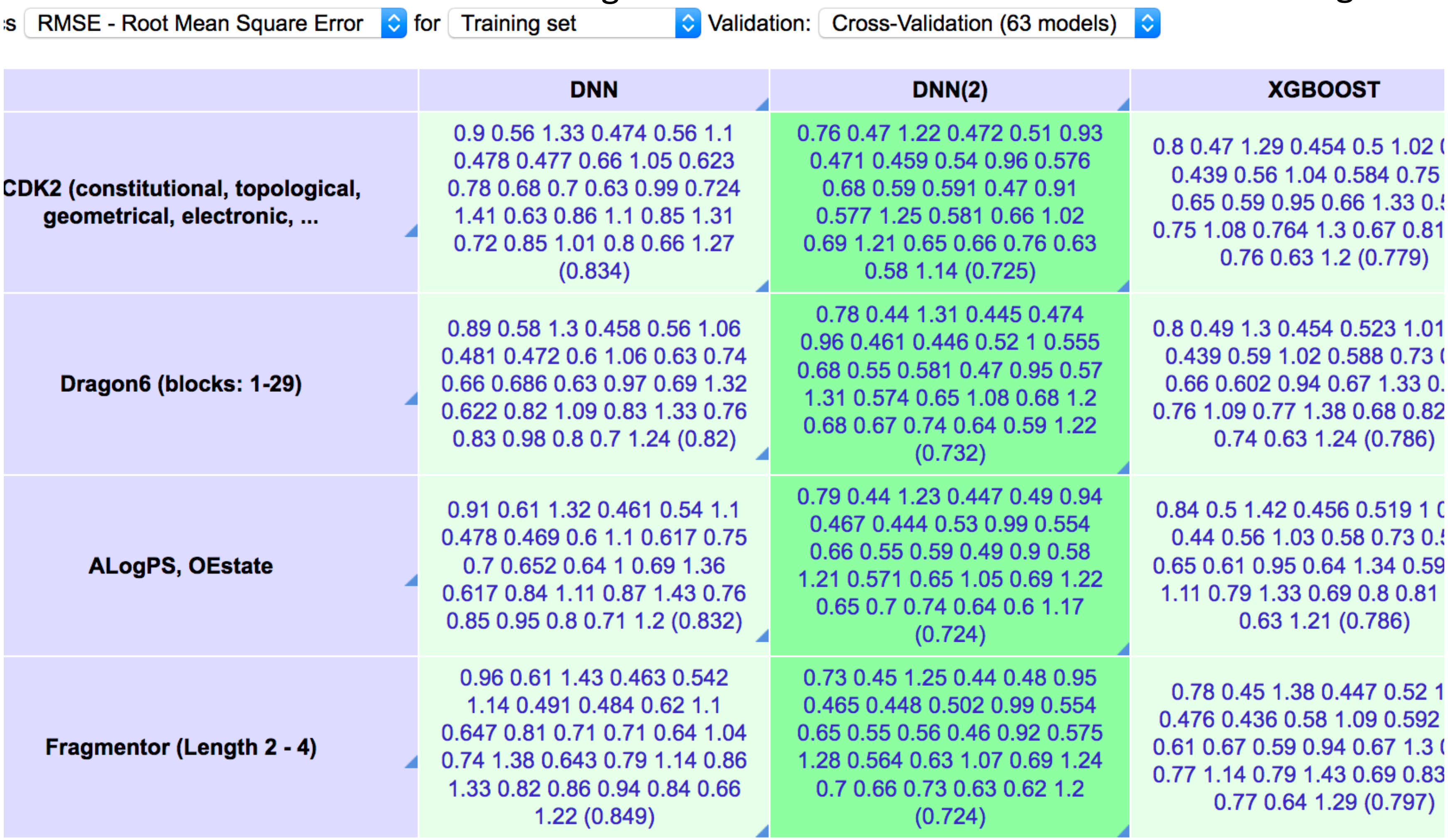
RMSE for different toxicities using CDK descriptors



Sosnin, S. et al. A comparative study of prediction of multi-target toxicity for a broad chemical space. *Chem. Res. Toxicol.* 2018, *in prep*.

Comparison of different models to predict toxicity (RMSE)

single multi single

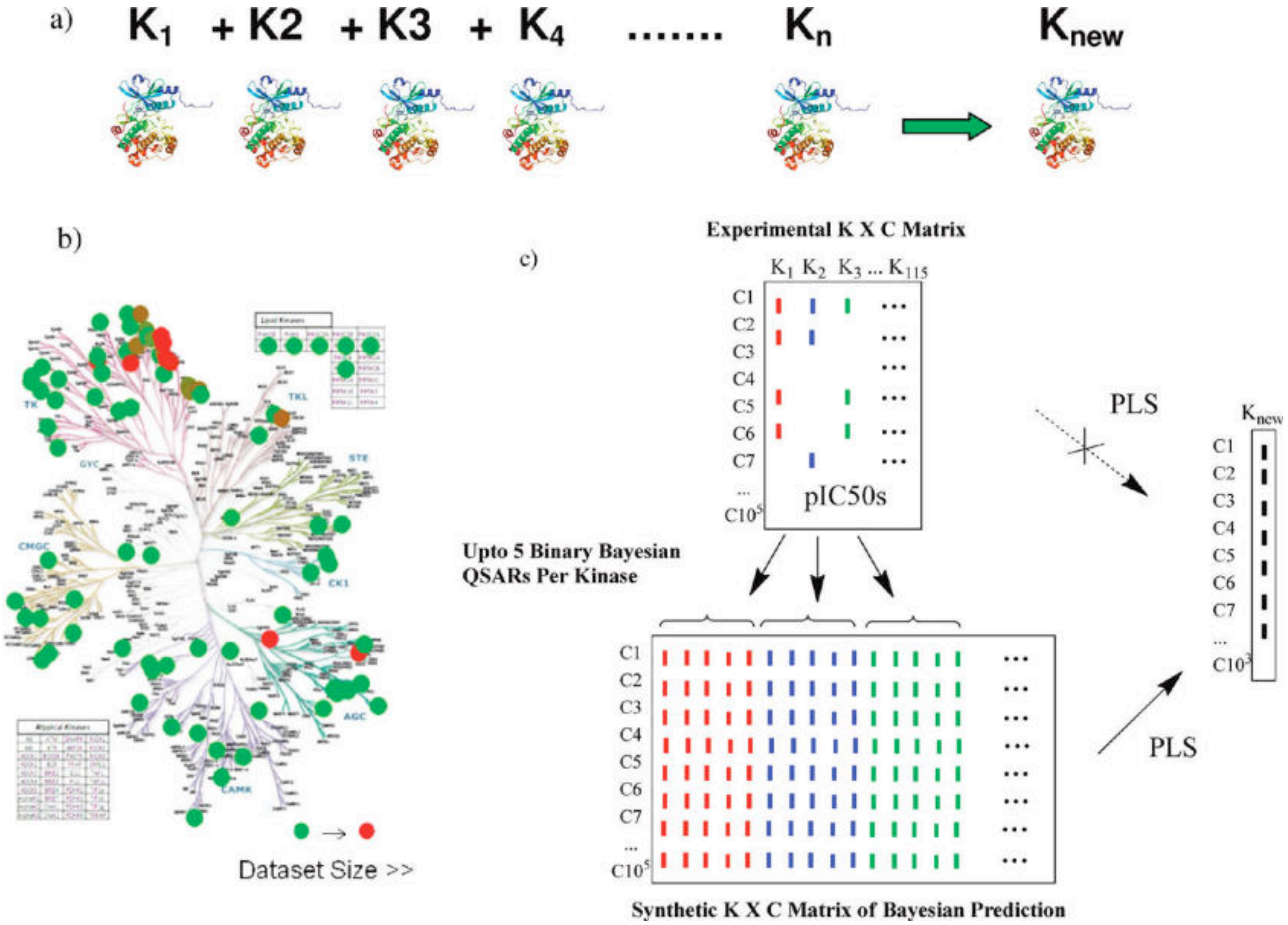


s RMSE - Root Mean Square Error for Training set Validation: Cross-Validation (63 models)

	DNN	DNN(2)	XGBOOST
CDK2 (constitutional, topological, geometrical, electronic, ...	0.9 0.56 1.33 0.474 0.56 1.1 0.478 0.477 0.66 1.05 0.623 0.78 0.68 0.7 0.63 0.99 0.724 1.41 0.63 0.86 1.1 0.85 1.31 0.72 0.85 1.01 0.8 0.66 1.27 (0.834)	0.76 0.47 1.22 0.472 0.51 0.93 0.471 0.459 0.54 0.96 0.576 0.68 0.59 0.591 0.47 0.91 0.577 1.25 0.581 0.66 1.02 0.69 1.21 0.65 0.66 0.76 0.63 0.58 1.14 (0.725)	0.8 0.47 1.29 0.454 0.5 1.02 0.439 0.56 1.04 0.584 0.75 0.65 0.59 0.95 0.66 1.33 0. 0.75 1.08 0.764 1.3 0.67 0.81 0.76 0.63 1.2 (0.779)
Dragon6 (blocks: 1-29)	0.89 0.58 1.3 0.458 0.56 1.06 0.481 0.472 0.6 1.06 0.63 0.74 0.66 0.686 0.63 0.97 0.69 1.32 0.622 0.82 1.09 0.83 1.33 0.76 0.83 0.98 0.8 0.7 1.24 (0.82)	0.78 0.44 1.31 0.445 0.474 0.96 0.461 0.446 0.52 1 0.555 0.68 0.55 0.581 0.47 0.95 0.57 1.31 0.574 0.65 1.08 0.68 1.2 0.68 0.67 0.74 0.64 0.59 1.22 (0.732)	0.8 0.49 1.3 0.454 0.523 1.01 0.439 0.59 1.02 0.588 0.73 0.66 0.602 0.94 0.67 1.33 0. 0.76 1.09 0.77 1.38 0.68 0.82 0.74 0.63 1.24 (0.786)
ALogPS, OEstate	0.91 0.61 1.32 0.461 0.54 1.1 0.478 0.469 0.6 1.1 0.617 0.75 0.7 0.652 0.64 1 0.69 1.36 0.617 0.84 1.11 0.87 1.43 0.76 0.85 0.95 0.8 0.71 1.2 (0.832)	0.79 0.44 1.23 0.447 0.49 0.94 0.467 0.444 0.53 0.99 0.554 0.66 0.55 0.59 0.49 0.9 0.58 1.21 0.571 0.65 1.05 0.69 1.22 0.65 0.7 0.74 0.64 0.6 1.17 (0.724)	0.84 0.5 1.42 0.456 0.519 1 0.44 0.56 1.03 0.58 0.73 0. 0.65 0.61 0.95 0.64 1.34 0.59 1.11 0.79 1.33 0.69 0.8 0.81 0.63 1.21 (0.786)
Fragmentor (Length 2 - 4)	0.96 0.61 1.43 0.463 0.542 1.14 0.491 0.484 0.62 1.1 0.647 0.81 0.71 0.71 0.64 1.04 0.74 1.38 0.643 0.79 1.14 0.86 1.33 0.82 0.86 0.94 0.84 0.66 1.22 (0.849)	0.73 0.45 1.25 0.44 0.48 0.95 0.465 0.448 0.502 0.99 0.554 0.65 0.55 0.56 0.46 0.92 0.575 1.28 0.564 0.63 1.07 0.69 1.24 0.7 0.66 0.73 0.63 0.62 1.2 (0.724)	0.78 0.45 1.38 0.447 0.52 1 0.476 0.436 0.58 1.09 0.592 0.61 0.67 0.59 0.94 0.67 1.3 0.77 1.14 0.79 1.43 0.69 0.83 0.77 0.64 1.29 (0.797)

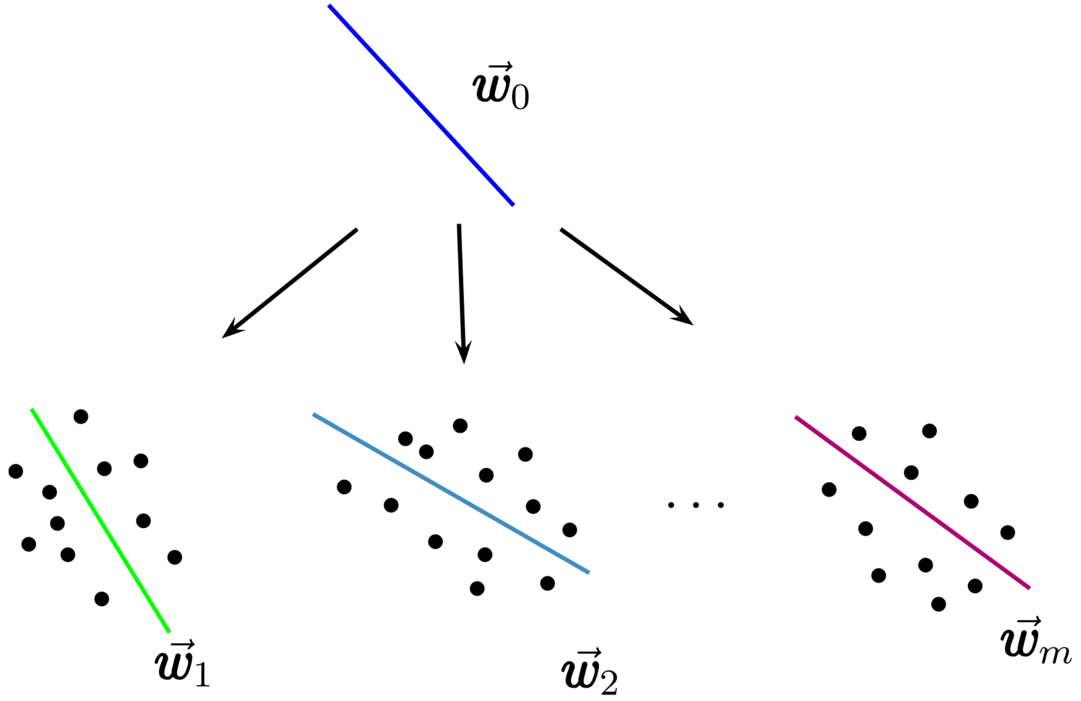
Sosnin, S. et al. A comparative study of prediction of multi-target toxicity for a broad chemical space. *Chem. Res. Toxicol.* 2018, *in prep*.

Profile-like QSAR



Martin, E.; Mukherjee, P.; Sullivan, D.; Jansen, J. Profile-QSAR: A novel meta-QSAR method that combines activities across the kinase family to accurately predict affinity, selectivity, and cellular activity. *J. Chem. Inf. Model.* **2011**, *51*, 1942-1956.

Non-neural network approaches to multi-learning: Least Squares Support Vector Regression (LSSVM)

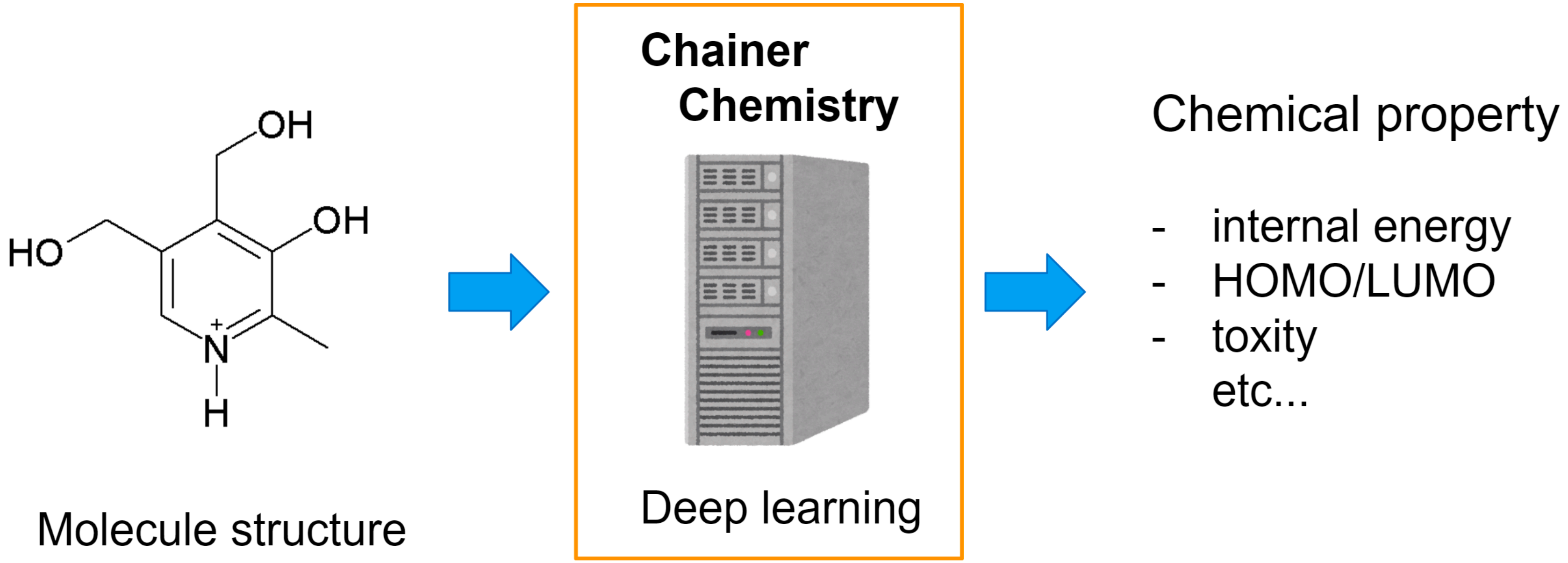


Suykens, J.A.K.; Vandewalle, J. Least squares support vector machine classifiers. *Neural Process. Lett.* 1999, *9*, 293-300.

Xu, S.; An, X.; Qiao, X.; Zhu, L.; Li, L. Multi-output least-squares support vector regression machines. Pattern Recognition Letters 2013, 34, 1078-1084.

Chainer Chemistry ("ChemChainer")

- Chainer – one of popular frameworks for Deep Learning
- Algorithms provided by Chainer developers
- Can be installed using Python tools
- https://github.com/pfnet-research/chainer-chemistry



Molecule structure

Chainer Chemistry

Deep learning

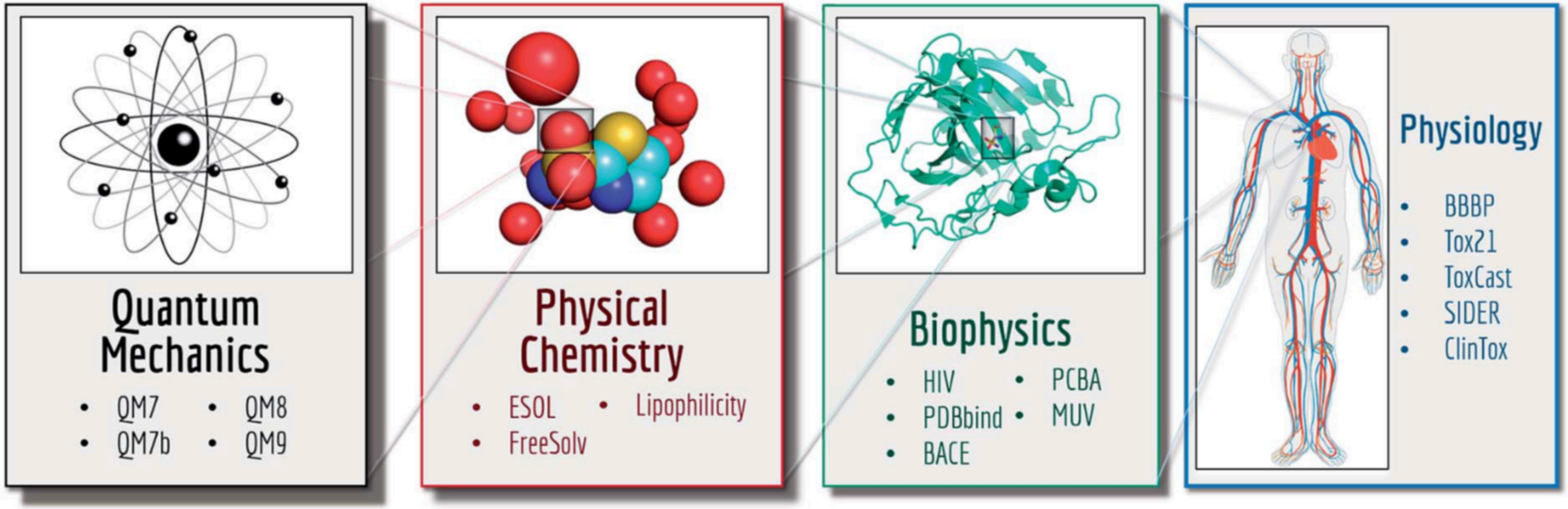
Chemical property

- internal energy
- HOMO/LUMO
- toxity

etc...

DEEPCHEM

- Based on TensorFlow (google)
- Available as part of Python (Anaconda) or as a Docker
- Supports multiple MTL and STL approaches
- https://github.com/deepchem/deepchem



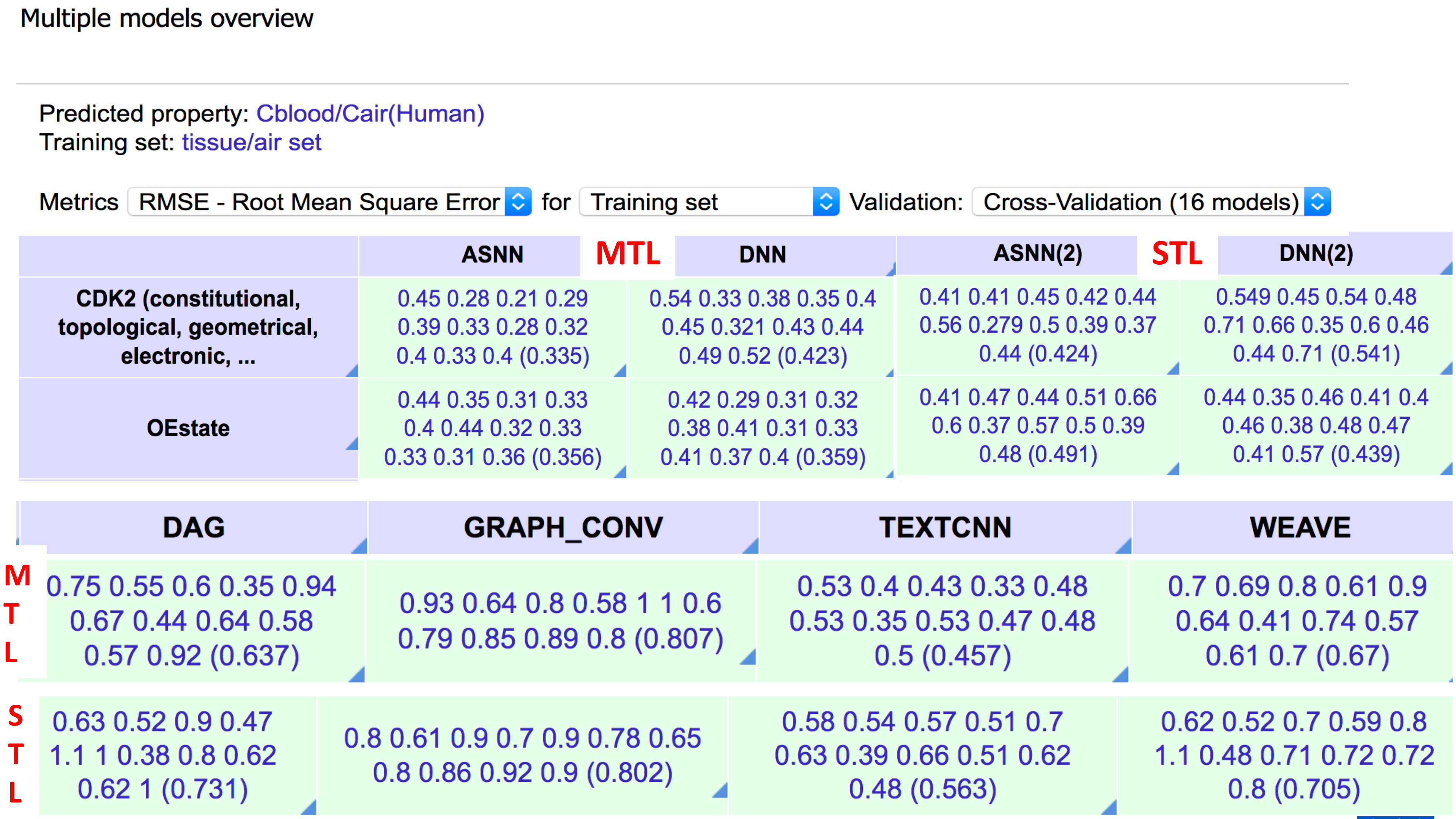
Wu, Z. et al Moleculenet: A benchmark for molecular machine learning. *Chem Sci* **2018**, *9*, 513-530.

Summary of “readily” available methods

Package	Examples of supported algorithms
Chainer Chemistry	NFP, GGNN, RSGCN, WeaveNet, SchNet
DeepChem	DAG, NNF, MPNN, TEXTCNN, WEAVE, IRV
OCHEM	Above methods + DNN, LSSVM, Macau, feature net as well as use of tasks classes as descriptors

NFP/NNF - Neural Fingreprint; GGNN - Gated Graph Neural Network; MPNN - Message Passing Neural Networks; SchNet - continuous-filter convolutional neural network for modeling quantum interactions; DAG - Directed Acyclic Graphs; IRV - Influence Relevance Voters ; LSSVM − Least Squares Support Vector Machines

Comparison of MTL and STL



Multiple models overview

Predicted property: Cblood/Cair(Human)
Training set: tissue/air set

Metrics RMSE - Root Mean Square Error for Training set Validation: Cross-Validation (16 models)

	MTL		STL	
	ASNN	DNN	ASNN(2)	DNN(2)
CDK2 (constitutional, topological, geometrical, electronic, ...	0.45 0.28 0.21 0.29 0.39 0.33 0.28 0.32 0.4 0.33 0.4 (0.335)	0.54 0.33 0.38 0.35 0.4 0.45 0.321 0.43 0.44 0.49 0.52 (0.423)	0.41 0.41 0.45 0.42 0.44 0.56 0.279 0.5 0.39 0.37 0.44 (0.424)	0.549 0.45 0.54 0.48 0.71 0.66 0.35 0.6 0.46 0.44 0.71 (0.541)
OEstate	0.44 0.35 0.31 0.33 0.4 0.44 0.32 0.33 0.33 0.31 0.36 (0.356)	0.42 0.29 0.31 0.32 0.38 0.41 0.31 0.33 0.41 0.37 0.4 (0.359)	0.41 0.47 0.44 0.51 0.66 0.6 0.37 0.57 0.5 0.39 0.48 (0.491)	0.44 0.35 0.46 0.41 0.4 0.46 0.38 0.48 0.47 0.41 0.57 (0.439)

	DAG	GRAPH_CONV	TEXTCNN	WEAVE
MTL	0.75 0.55 0.6 0.35 0.94 0.67 0.44 0.64 0.58 0.57 0.92 (0.637)	0.93 0.64 0.8 0.58 1 1 0.6 0.79 0.85 0.89 0.8 (0.807)	0.53 0.4 0.43 0.33 0.48 0.53 0.35 0.53 0.47 0.48 0.5 (0.457)	0.7 0.69 0.8 0.61 0.9 0.64 0.41 0.74 0.57 0.61 0.7 (0.67)
STL	0.63 0.52 0.9 0.47 1.1 1 0.38 0.8 0.62 0.62 1 (0.731)	0.8 0.61 0.9 0.7 0.9 0.78 0.65 0.8 0.86 0.92 0.9 (0.802)	0.58 0.54 0.57 0.51 0.7 0.63 0.39 0.66 0.51 0.62 0.48 (0.563)	0.62 0.52 0.7 0.59 0.8 1.1 0.48 0.71 0.72 0.72 0.8 (0.705)

big chem

http://bigchem.eu

big data in chemistry + informatics = chemoinformatics

The **increasing volume of biomedical data** in chemistry and life sciences requires development of **new methods and approaches for their analysis**.

The BIGCHEM project will provide **innovative education in large chemical data analysis**. The innovative research program will be implemented with the target users, **large pharma companies and SMEs**, which generate and analyze large chemical data as well as will promote technology transfer from academy to industrial applications.

Marie Skłodowska-Curie European Industrial Doctorate (EID)

Beneficiaries

HelmholtzZentrum münchen
German Research Center for Environmental Health

u^b
b
UNIVERSITÄT
BERN

Lead Discovery Center
LDC

AstraZeneca

ETH zürich

UNIVERSITÀ DEGLI STUDI
DI MODENA E REGGIO EMILIA

Boehringer
Ingelheim

universitätbonn
Rheinische
Friedrich-Wilhelms-
Universität Bonn

UNIVERSITÉ DE STRASBOURG

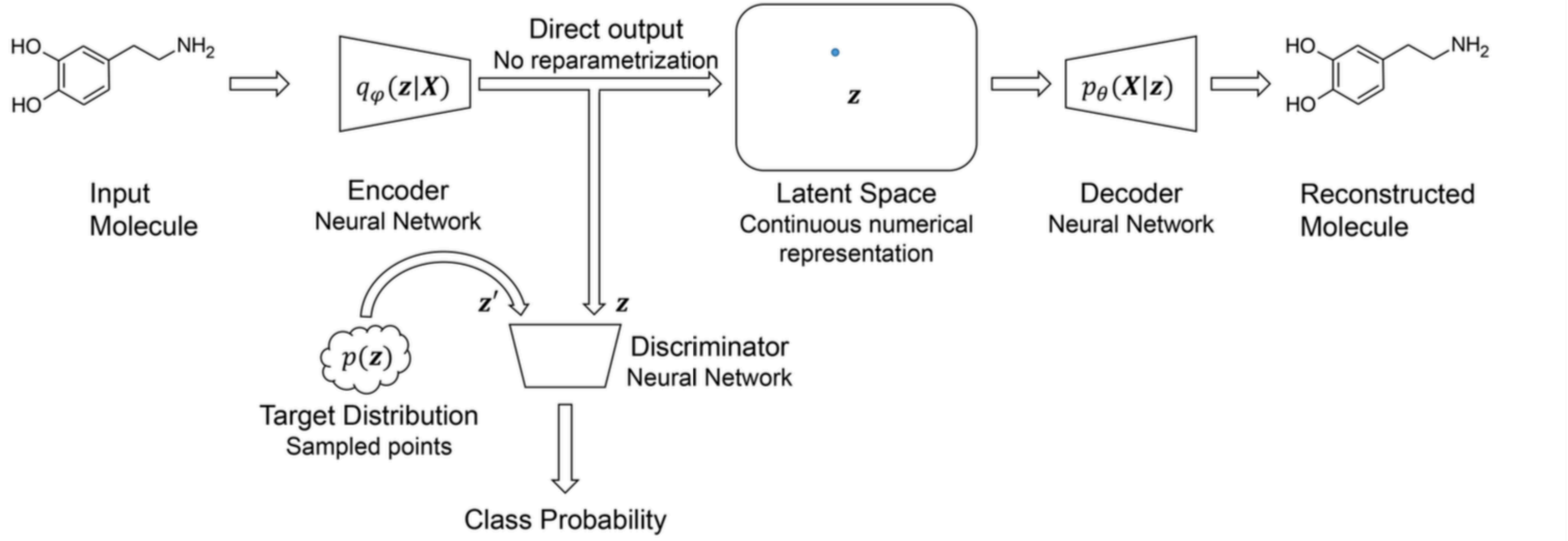
DOI: 10.1002/minf.201700123

Application of Generative Autoencoder in de Novo Molecular Design

Thomas Blaschke,*[a, b] Marcus Olivecrona,[a] Ola Engkvist,[a] Jürgen Bajorath,[b] and Hongming Chen*[a]

Abstract: A major challenge in computational chemistry is the generation of novel molecular structures with desirable pharmacological and physiochemical properties. In this work, we investigate the potential use of autoencoder, a deep learning methodology, for de novo molecular design. Various generative autoencoders were used to map molecule structures into a continuous latent space and vice versa and their performance as structure generator was assessed. Our results show that the latent space preserves chemical similarity principle and thus can be used for the generation of analogue structures. Furthermore, the latent space created by autoencoders were searched systematically to generate novel compounds with predicted activity against dopamine receptor type 2 and compounds similar to known active compounds not included in the trainings set were identified.

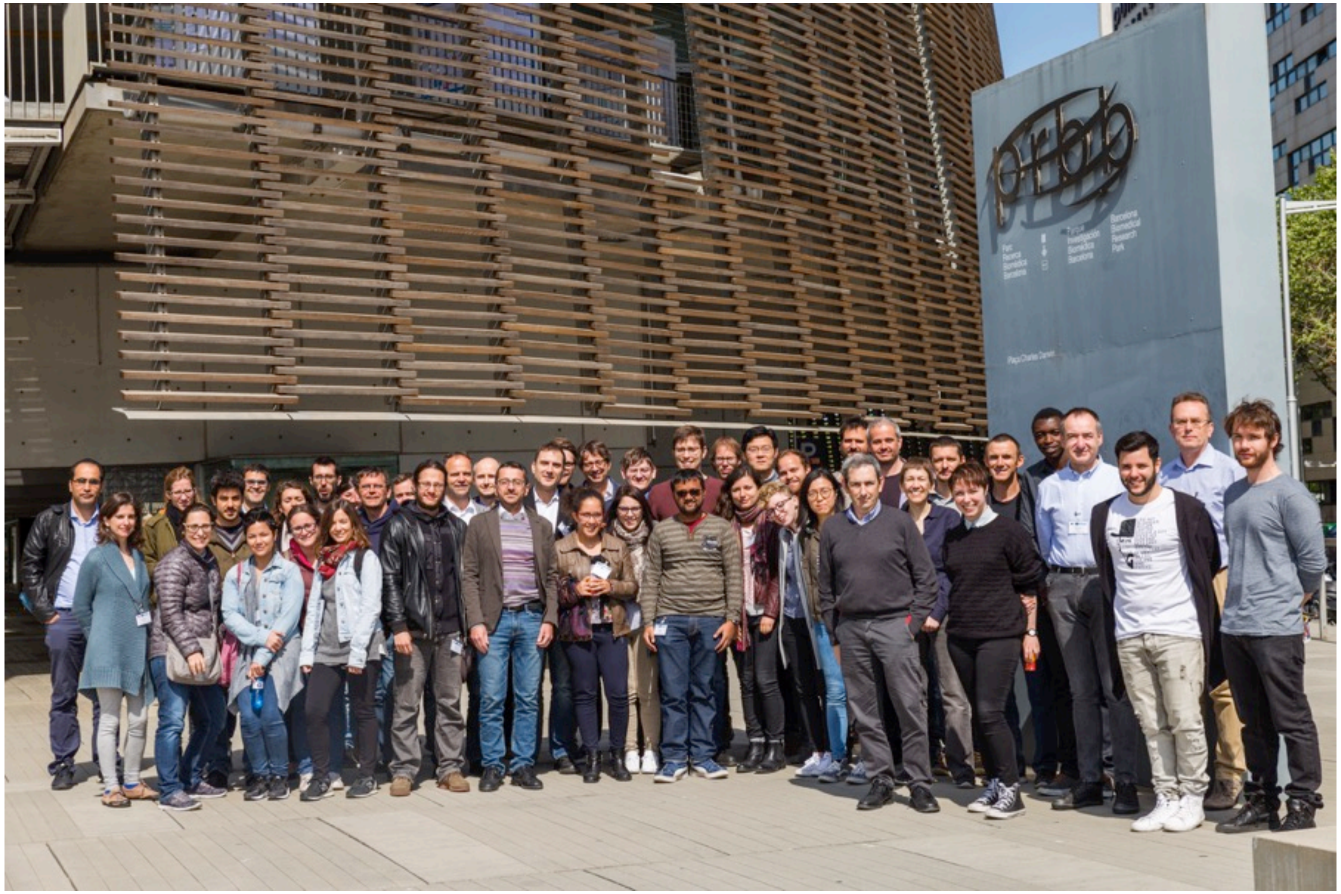
Keywords: Autoencoder · chemoinformatics · de novo molecular design · deep learning · inverse QSAR



Summary

- OCHEM is powerful extendable platform for data storage
- Works with millions of datapoints
- Provide an integrated support of various (multi-learning) algorithms
- Very useful for ADMETox and (Q)SAR studies

Acknowledgements



Pavel Karpov
Dipan Ghosh
Michael Withnall
Allison Keys
Zhonghua Xia
Barbara Gasset
Monica Campillos

Guillaume Godin
Sergey Sosnin
Daniel Lowe
Maxim Fedorov
Anthony Williams

Yura Sushko
Sergey Novotarskyi
Robert Körner

Michael Sattler

MARIE CURIE ACTIONS

BigChem

ERA CVD