RDKit (new 3D) descriptors “a study case”

Guillaume GODIN

19th September 2018

2018 RDKit UGM

Cambridge UK

http://cheminformatics.epfl.ch/workshop/20180914program.shtml

Contributors

- Gregory Landrum => RDKit (support on 3D descriptors)
- Igor Tetko => OCHEM (support RDkit & Firmenich Descriptors)
- Talia Kimber => CNN (Master thesis in progress at Firmenich)
- Arvind Jayaraman => Mathworks (support on DL toolbox enhancement)

- Firmenich IA Team:

  Eric

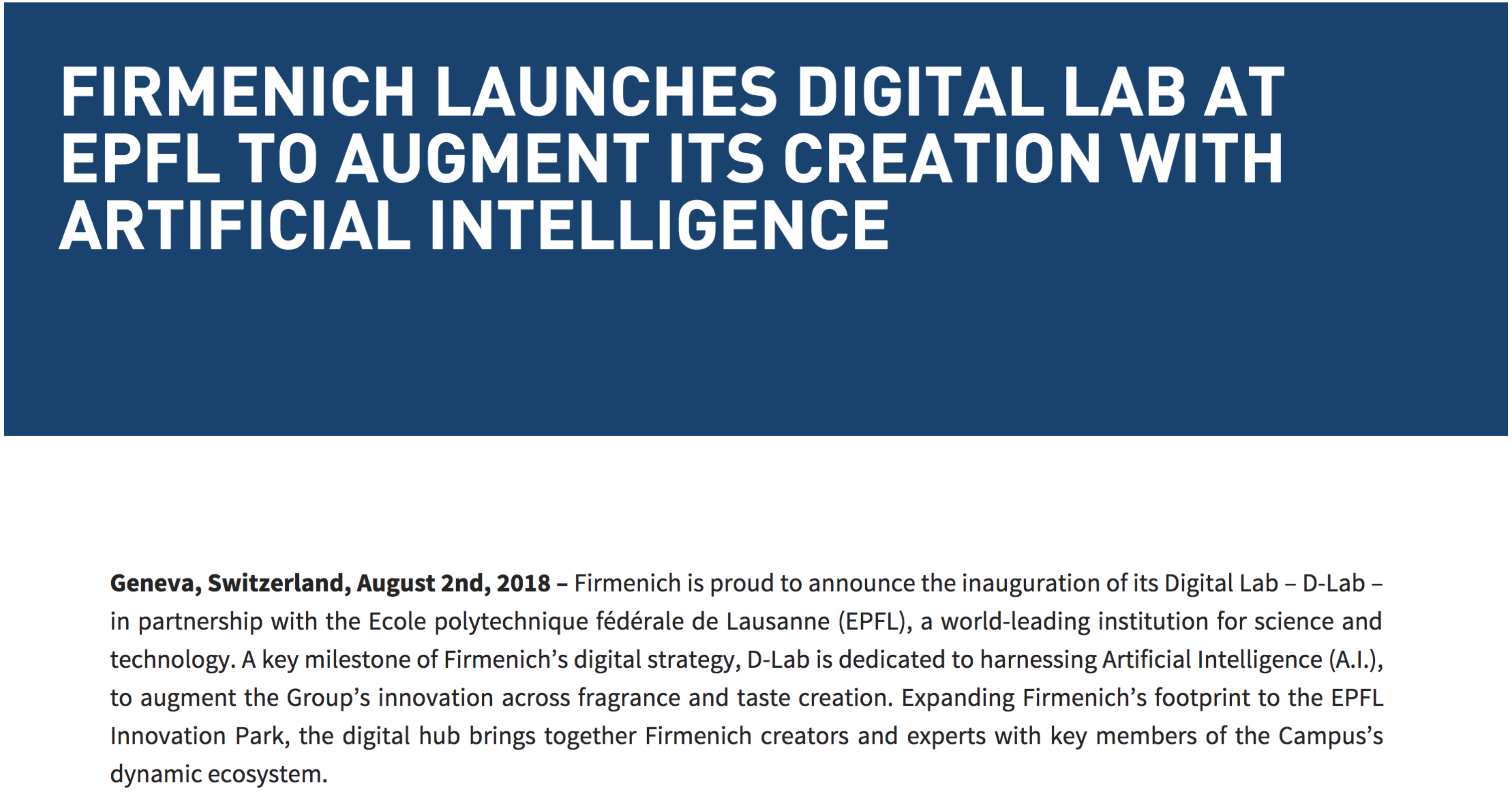
  Dario

  Addis

  Sven

http://cheminformatics.epfl.ch/workshop/20180914program.shtml

Firmenich D-Lab @ EPFL



FIRMENICH LAUNCHES DIGITAL LAB AT EPFL TO AUGMENT ITS CREATION WITH ARTIFICIAL INTELLIGENCE

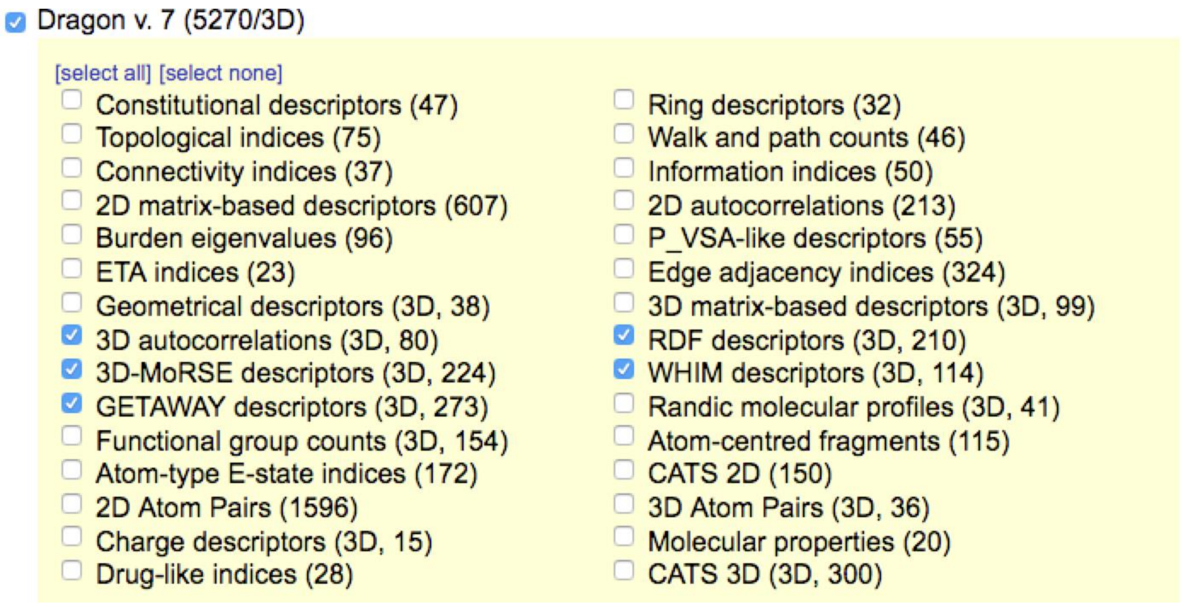
Geneva, Switzerland, August 2nd, 2018 – Firmenich is proud to announce the inauguration of its Digital Lab – D-Lab – in partnership with the Ecole polytechnique fédérale de Lausanne (EPFL), a world-leading institution for science and technology. A key milestone of Firmenich's digital strategy, D-Lab is dedicated to harnessing Artificial Intelligence (A.I.), to augment the Group's innovation across fragrance and taste creation. Expanding Firmenich's footprint to the EPFL Innovation Park, the digital hub brings together Firmenich creators and experts with key members of the Campus's dynamic ecosystem.

RDKit 3D descriptors since 2017.09

Autocorr3D	New in 2017.09 release. Todeschini and Consoni “Descriptors from Molecular Geometry” Handbook of Chemoinformaticshttp://dx.doi.org/10.1002/9783527618279.ch37	C++
RDF	same	C++
MORSE	same	C++
WHIM	same	C++
GETAWAY	same	C++
Autocorr2D	same	C++

Last Hackathon: Since Version 2018.09 custom atomic properties can be injected in those descriptors ”algorithms”

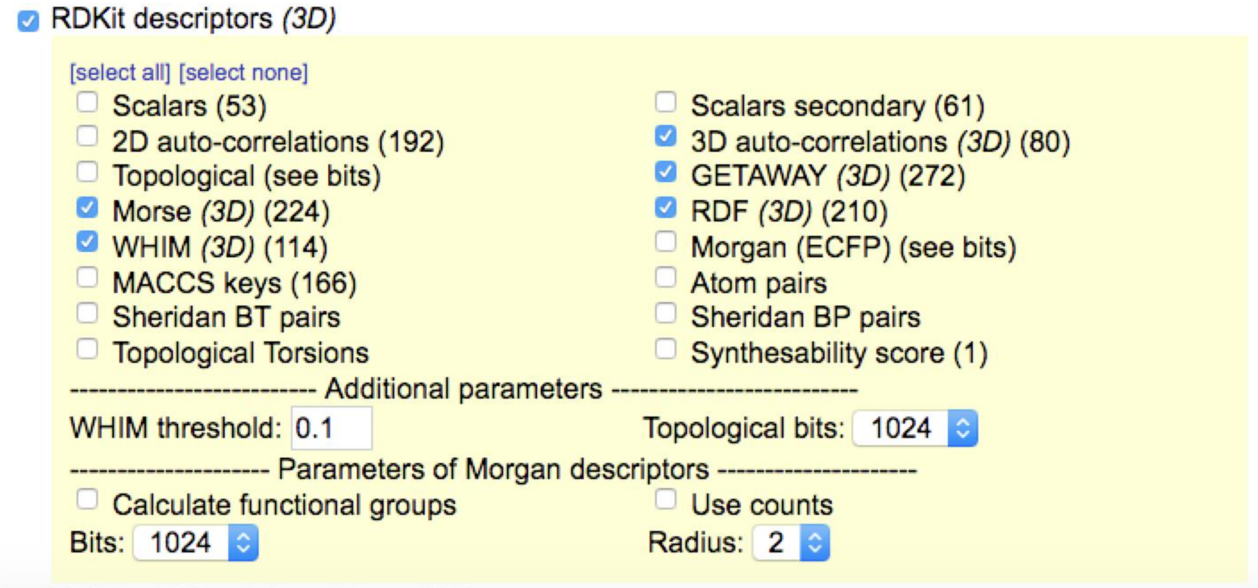
OCHEM* Dragon v7 & RDKit Descriptors



☑ Dragon v. 7 (5270/3D)

[select all] [select none]

☐ Constitutional descriptors (47)
☐ Topological indices (75)
☐ Connectivity indices (37)
☐ 2D matrix-based descriptors (607)
☐ Burden eigenvalues (96)
☐ ETA indices (23)
☐ Geometrical descriptors (3D, 38)
☑ 3D autocorrelations (3D, 80)
☑ 3D-MoRSE descriptors (3D, 224)
☑ GETAWAY descriptors (3D, 273)
☐ Functional group counts (3D, 154)
☐ Atom-type E-state indices (172)
☐ 2D Atom Pairs (1596)
☐ Charge descriptors (3D, 15)
☐ Drug-like indices (28)
☐ Ring descriptors (32)
☐ Walk and path counts (46)
☐ Information indices (50)
☐ 2D autocorrelations (213)
☐ P_VSA-like descriptors (55)
☐ Edge adjacency indices (324)
☐ 3D matrix-based descriptors (3D, 99)
☑ RDF descriptors (3D, 210)
☑ WHIM descriptors (3D, 114)
☐ Randic molecular profiles (3D, 41)
☐ Atom-centred fragments (115)
☐ CATS 2D (150)
☐ 3D Atom Pairs (3D, 36)
☐ Molecular properties (20)
☐ CATS 3D (3D, 300)



☑ RDKit descriptors *(3D)*

[select all] [select none]

☐ Scalars (53)
☐ 2D auto-correlations (192)
☐ Topological (see bits)
☑ Morse *(3D)* (224)
☑ WHIM *(3D)* (114)
☐ MACCS keys (166)
☐ Sheridan BT pairs
☐ Topological Torsions
☐ Scalars secondary (61)
☑ 3D auto-correlations *(3D)* (80)
☑ GETAWAY *(3D)* (272)
☑ RDF *(3D)* (210)
☐ Morgan (ECFP) (see bits)
☐ Atom pairs
☐ Sheridan BP pairs
☐ Synthesability score (1)

------------------------- Additional parameters -------------------------

WHIM threshold: 0.1
Topological bits: 1024

-------------------- Parameters of Morgan descriptors --------------------

☐ Calculate functional groups
☐ Use counts
Bits: 1024
Radius: 2

2 tests:

- only 3D in common (see check blue boxes)
- all 2D + 3D (without RDKit Sheridan pairs & Topological Torsions)

***All computation made using OCHEM**

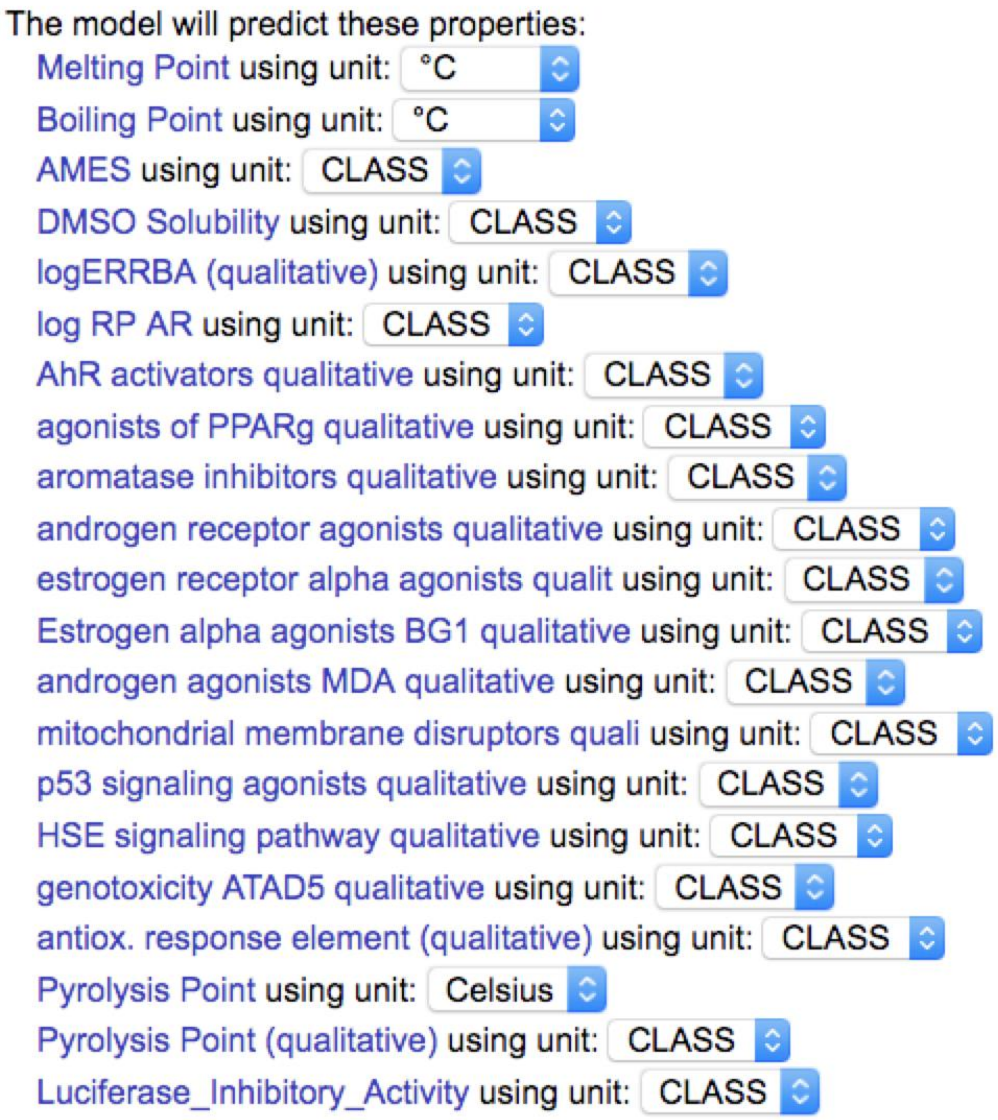
Study 1: Multi learning tasks

Target selects:

- Regression
  - MP
  - BP
  - Pyrolysis Point
- Classification 18
  - Toxicities
  - biological agonists
  - …

Dataset = 1’015’745 data points

We can learn targets with heterogenous chemical datasets



The model will predict these properties:

Melting Point using unit: °C

Boiling Point using unit: °C

AMES using unit: CLASS

DMSO Solubility using unit: CLASS

logERRBA (qualitative) using unit: CLASS

log RP AR using unit: CLASS

AhR activators qualitative using unit: CLASS

agonists of PPARg qualitative using unit: CLASS

aromatase inhibitors qualitative using unit: CLASS

androgen receptor agonists qualitative using unit: CLASS

estrogen receptor alpha agonists qualit using unit: CLASS

Estrogen alpha agonists BG1 qualitative using unit: CLASS

androgen agonists MDA qualitative using unit: CLASS

mitochondrial membrane disruptors quali using unit: CLASS

p53 signaling agonists qualitative using unit: CLASS

HSE signaling pathway qualitative using unit: CLASS

genotoxicity ATAD5 qualitative using unit: CLASS

antiox. response element (qualitative) using unit: CLASS

Pyrolysis Point using unit: Celsius

Pyrolysis Point (qualitative) using unit: CLASS

Luciferase_Inhibitory_Activity using unit: CLASS

Databases are Public available from OCHEM website (grab from original articles)

DNN architecture

We train **unique dense deep network** to learn all targets **simultaneity**

Benefits

- One global model
- Use targets synergy
- Faster inference

Input: dimension FP size

Hidden layer 1 : 512 neurons

- Dropout 0.5
- Relu

Hidden layer 2 : 256 neurons

- Dropout 0.5
- Relu

Hidden layer 3 : 128 neurons

- Dropout 0.5
- Relu

Hidden layer 4 : 64 neurons

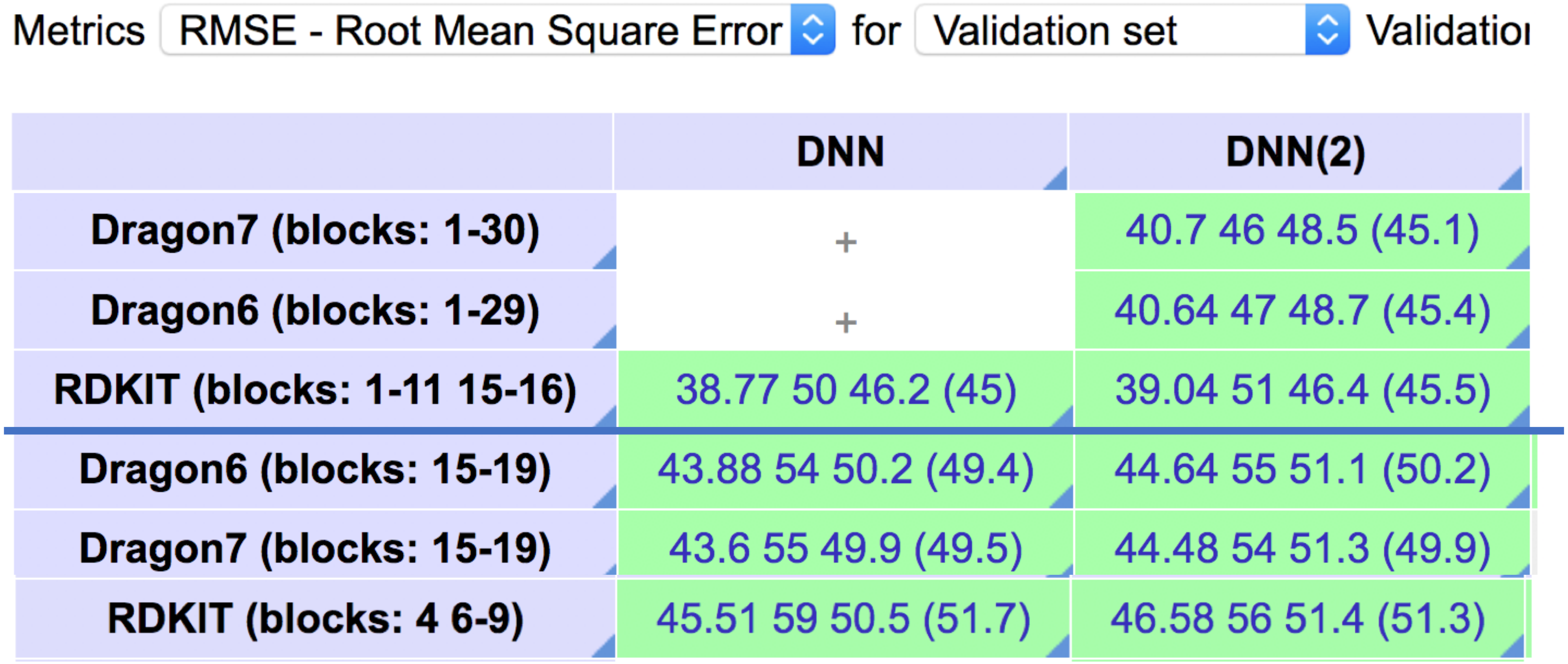
- Dropout 0.25
- Relu

Hidden layer 5 : 32 neurons

- Dropout 0.1
- Relu

output : dimension 21

Results for regression targets (RMSE)



Metrics RMSE - Root Mean Square Error for Validation set Validatio

	DNN	DNN(2)
Dragon7 (blocks: 1-30)	+	40.7 46 48.5 (45.1)
Dragon6 (blocks: 1-29)	+	40.64 47 48.7 (45.4)
RDKIT (blocks: 1-11 15-16)	38.77 50 46.2 (45)	39.04 51 46.4 (45.5)
Dragon6 (blocks: 15-19)	43.88 54 50.2 (49.4)	44.64 55 51.1 (50.2)
Dragon7 (blocks: 15-19)	43.6 55 49.9 (49.5)	44.48 54 51.3 (49.9)
RDKIT (blocks: 4 6-9)	45.51 59 50.5 (51.7)	46.58 56 51.4 (51.3)

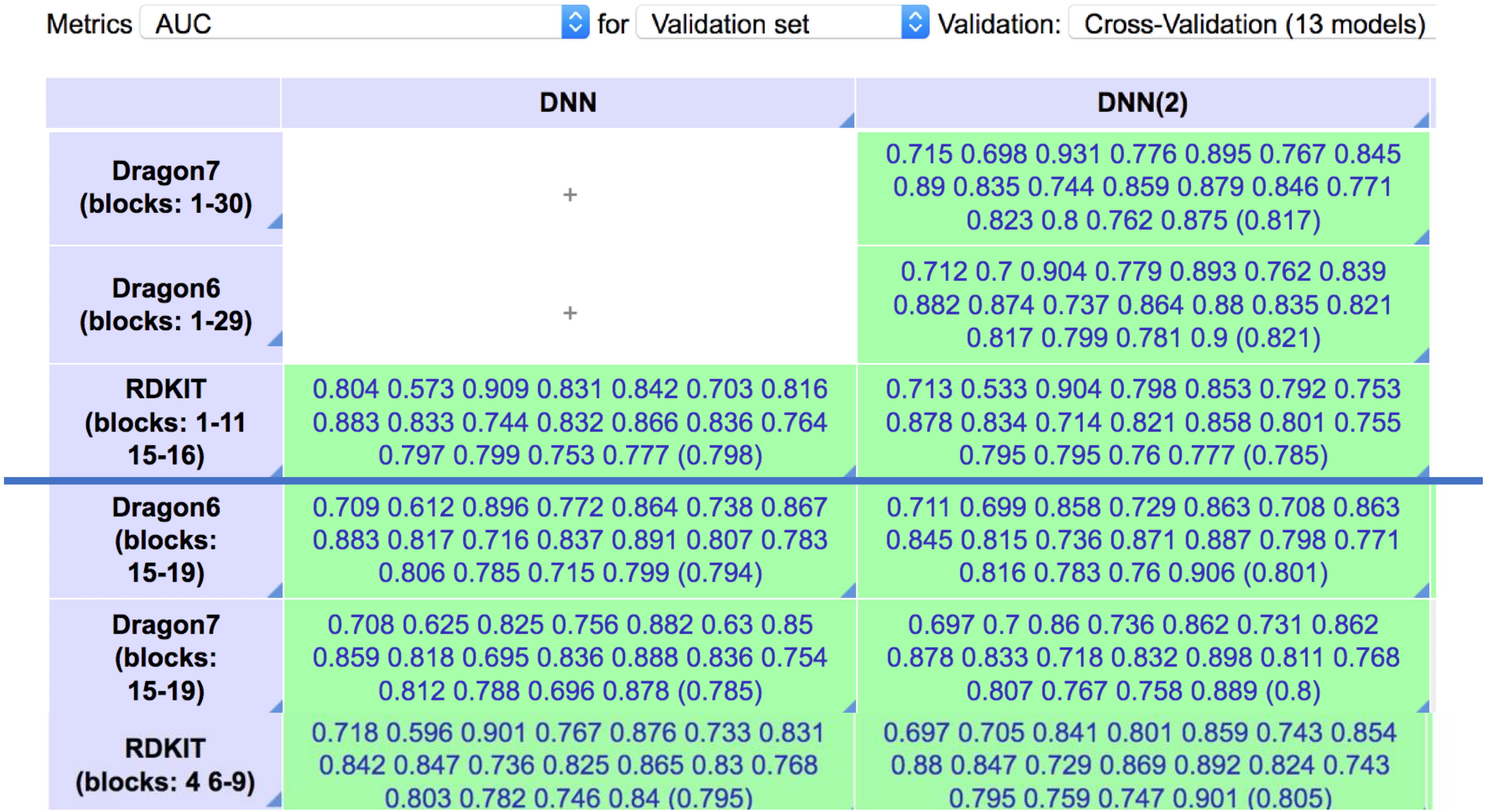
All descriptors

Only 3D

10000 epochs

2000 epochs

Results for classification targets (AUC)



Metrics AUC for Validation set Validation: Cross-Validation (13 models)

	DNN	DNN(2)
Dragon7 (blocks: 1-30)	+	0.715 0.698 0.931 0.776 0.895 0.767 0.845 0.89 0.835 0.744 0.859 0.879 0.846 0.771 0.823 0.8 0.762 0.875 (0.817)
Dragon6 (blocks: 1-29)	+	0.712 0.7 0.904 0.779 0.893 0.762 0.839 0.882 0.874 0.737 0.864 0.88 0.835 0.821 0.817 0.799 0.781 0.9 (0.821)
RDKIT (blocks: 1-11 15-16)	0.804 0.573 0.909 0.831 0.842 0.703 0.816 0.883 0.833 0.744 0.832 0.866 0.836 0.764 0.797 0.799 0.753 0.777 (0.798)	0.713 0.533 0.904 0.798 0.853 0.792 0.753 0.878 0.834 0.714 0.821 0.858 0.801 0.755 0.795 0.795 0.76 0.777 (0.785)
Dragon6 (blocks: 15-19)	0.709 0.612 0.896 0.772 0.864 0.738 0.867 0.883 0.817 0.716 0.837 0.891 0.807 0.783 0.806 0.785 0.715 0.799 (0.794)	0.711 0.699 0.858 0.729 0.863 0.708 0.863 0.845 0.815 0.736 0.871 0.887 0.798 0.771 0.816 0.783 0.76 0.906 (0.801)
Dragon7 (blocks: 15-19)	0.708 0.625 0.825 0.756 0.882 0.63 0.85 0.859 0.818 0.695 0.836 0.888 0.836 0.754 0.812 0.788 0.696 0.878 (0.785)	0.697 0.7 0.86 0.736 0.862 0.731 0.862 0.878 0.833 0.718 0.832 0.898 0.811 0.768 0.807 0.767 0.758 0.889 (0.8)
RDKIT (blocks: 4 6-9)	0.718 0.596 0.901 0.767 0.876 0.733 0.831 0.842 0.847 0.736 0.825 0.865 0.83 0.768 0.803 0.782 0.746 0.84 (0.795)	0.697 0.705 0.841 0.801 0.859 0.743 0.854 0.88 0.847 0.729 0.869 0.892 0.824 0.743 0.795 0.759 0.747 0.901 (0.805)

All descriptors

Only 3D

Study 1: Conclusion

- RDkit provide very similar accuracy than Dragon v6 or v7
- RDkit provide flexibility our own descriptors (3D custom atomic properties option available since 2018.09 version)
- RDkit is faster (x6) than Dragon

What we need now ?

- Faster way to enumerate multiple smiles from a given molecule
- Master Students @ EPFL (D-lab)
- PhD in chemoinformatic & Deep Learning @ EPFL or Geneva
- 2 Full positions: junior & senior Data scientists (with or without chemical background)
- ~~1 Full position: senior Chemoinformatic scientist (with deep learning experience)~~

Q&A

THANK YOU

Data sources "article"

- Ghosh, D.; Koch, U.; Hadian, K.; Sattler, M.; Tetko, I.V. Luciferase advisor: High-accuracy model to flag false positive hits in luciferase hts assays. *J. Chem. Inf. Model.* **2018**, *58*, 933-942.
- Tetko, I.V.; Novotarskyi, S.; Sushko, I.; Ivanov, V.; Petrenko, A.E.; Dieden, R.; Lebon, F.; Mathieu, B. Development of dimethyl sulfoxide solubility models using 163 000 molecules: Using a domain applicability metric to select more reliable predictions. *J. Chem. Inf. Model.* **2013**, *53*, 1990-2000.
- Tetko, I.V.; Sushko, Y.; Novotarskyi, S.; Patiny, L.; Kondratov, I.; Petrenko, A.E.; Charochkina, L.; Asiri, A.M. How accurately can we predict the melting points of drug-like compounds? *J. Chem. Inf. Model.* **2014**, *54*, 3320-3329.
- Tetko, I.V.; D, M.L.; Williams, A.J. The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from patents. *J. Cheminform.* **2016**, *8*, 2.
- Abdelaziz, A.; Spahn-Langguth, H.; Werner-Schramm, K.; Tetko, I.V. Consensus modeling for hts assays using in silico descriptors calculates the best balanced accuracy in tox21 challenge. *Frontiers Environ. Sci.* **2016**, *4*, 2.
- Rybacka, A.; Ruden, C.; Tetko, I.V.; Andersson, P.L. Identifying potential endocrine disruptors among industrial chemicals and their metabolites - development and evaluation of in silico tools. *Chemosphere* **2015**, *139*, 372-378.
- Brandmaier, S.; Sahlin, U.; Tetko, I.V.; Oberg, T. Pls-optimal: A stepwise d-optimal design based on latent variables. *J. Chem. Inf. Model.* **2012**, *52*, 975-983.
- Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A.K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Muller, K.R.*, et al.* Applicability domains for classification problems: Benchmarking of distance to models for ames mutagenicity set. *J. Chem. Inf. Model.* **2010**, *50*, 2094-2111.