Firmenich

for good, naturally

REPRODUCIBLE AI !?

GUILLAUME GODIN

SCIENTIFIC DIRECTOR

DIGITAL LAB

FIRMENICH SA

“ESOL” BENCHMARK FOR WATER SOLUBILITY

I found 3 data sources available … :

- Delaney article (2004): 1144
- RDKit: Christos Kannas (2013 2nd RDKit UGM meeting): 1143
- Deepchem / Moleculenet.ai (2017): 1128

Recent papers use modified benchmark => up to 16 molecules removed :

- Are the harder one ? Hopefully not ;-)

STATISTICAL LIMIT REPORTED

- We make 5 CV BUT:
  - We do not always report CV deviation of models in paper or on slides!
- So please add «Error bars» in presentations
- Correct judgement of errors in NN:
  - Smaller error = more «robust» model
  - => better compression & high generalisation

OPEN SOURCE CODE SHARING

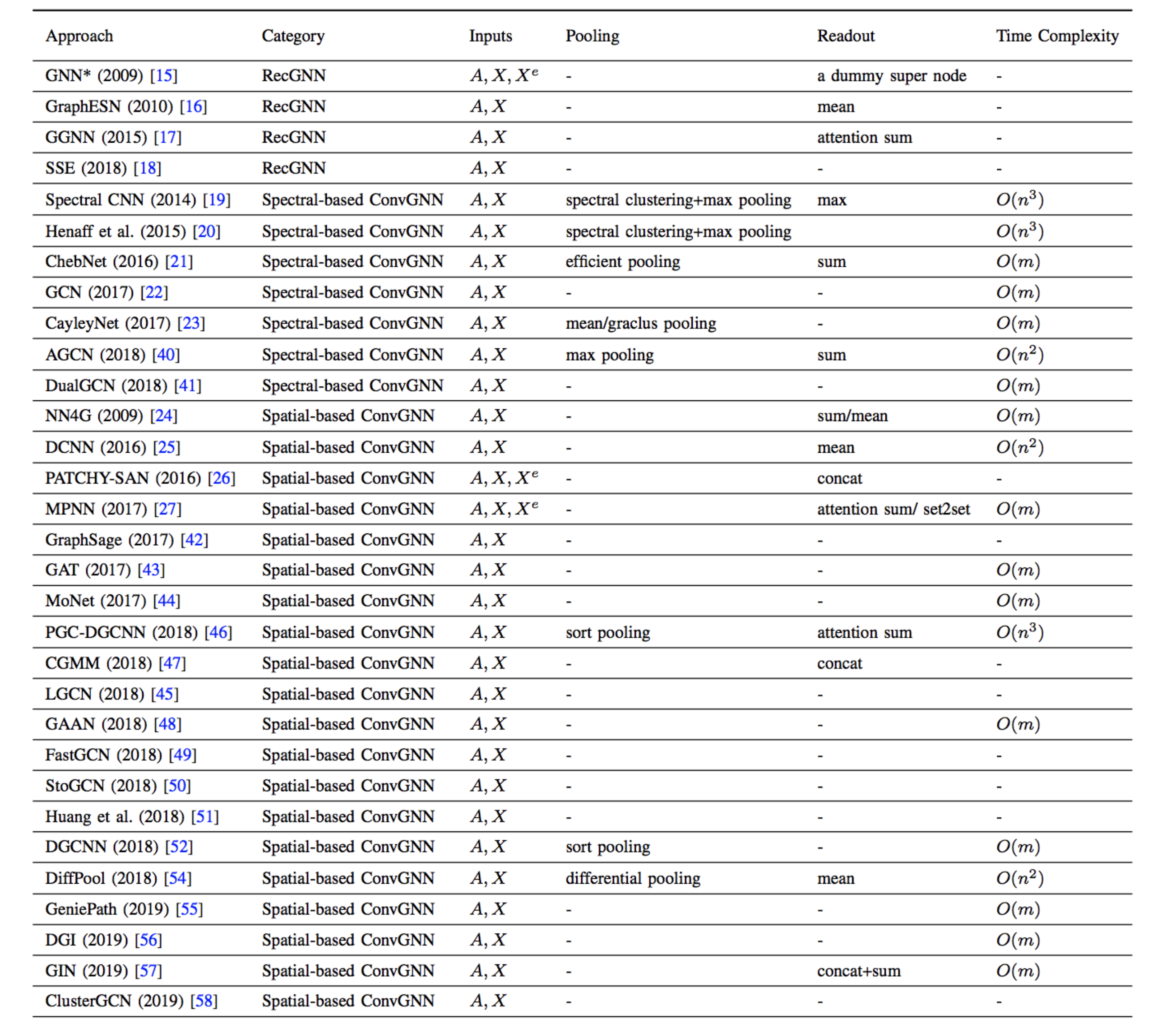
As you know, open source code is the best to validate paper: “Bryan Kelly”

If possible with unit tests ;-)

« CAUTION », simple « seed » changes may affect the performance of the model:

(aka) not robust and not general model

How to chose Graph conv => like FP all!



Approach	Category	Inputs	Pooling	Readout	Time Complexity
GNN* (2009) [15]	RecGNN	A, X, X^e	-	a dummy super node	-
GraphESN (2010) [16]	RecGNN	A, X	-	mean	-
GGNN (2015) [17]	RecGNN	A, X	-	attention sum	-
SSE (2018) [18]	RecGNN	A, X	-	-	-
Spectral CNN (2014) [19]	Spectral-based ConvGNN	A, X	spectral clustering+max pooling	max	$O(n^3)$
Henaff et al. (2015) [20]	Spectral-based ConvGNN	A, X	spectral clustering+max pooling		$O(n^3)$
ChebNet (2016) [21]	Spectral-based ConvGNN	A, X	efficient pooling	sum	$O(m)$
GCN (2017) [22]	Spectral-based ConvGNN	A, X	-	-	$O(m)$
CayleyNet (2017) [23]	Spectral-based ConvGNN	A, X	mean/graclus pooling	-	$O(m)$
AGCN (2018) [40]	Spectral-based ConvGNN	A, X	max pooling	sum	$O(n^2)$
DualGCN (2018) [41]	Spectral-based ConvGNN	A, X	-	-	$O(m)$
NN4G (2009) [24]	Spatial-based ConvGNN	A, X	-	sum/mean	$O(m)$
DCNN (2016) [25]	Spatial-based ConvGNN	A, X	-	mean	$O(n^2)$
PATCHY-SAN (2016) [26]	Spatial-based ConvGNN	A, X, X^e	-	concat	-
MPNN (2017) [27]	Spatial-based ConvGNN	A, X, X^e	-	attention sum/ set2set	$O(m)$
GraphSage (2017) [42]	Spatial-based ConvGNN	A, X	-	-	-
GAT (2017) [43]	Spatial-based ConvGNN	A, X	-	-	$O(m)$
MoNet (2017) [44]	Spatial-based ConvGNN	A, X	-	-	$O(m)$
PGC-DGCNN (2018) [46]	Spatial-based ConvGNN	A, X	sort pooling	attention sum	$O(n^3)$
CGMM (2018) [47]	Spatial-based ConvGNN	A, X	-	concat	-
LGCN (2018) [45]	Spatial-based ConvGNN	A, X	-	-	-
GAAN (2018) [48]	Spatial-based ConvGNN	A, X	-	-	$O(m)$
FastGCN (2018) [49]	Spatial-based ConvGNN	A, X	-	-	-
StoGCN (2018) [50]	Spatial-based ConvGNN	A, X	-	-	-
Huang et al. (2018) [51]	Spatial-based ConvGNN	A, X	-	-	-
DGCNN (2018) [52]	Spatial-based ConvGNN	A, X	sort pooling	-	$O(m)$
DiffPool (2018) [54]	Spatial-based ConvGNN	A, X	differential pooling	mean	$O(n^2)$
GeniePath (2019) [55]	Spatial-based ConvGNN	A, X	-	-	$O(m)$
DGI (2019) [56]	Spatial-based ConvGNN	A, X	-	-	$O(m)$
GIN (2019) [57]	Spatial-based ConvGNN	A, X	-	concat+sum	$O(m)$
ClusterGCN (2019) [58]	Spatial-based ConvGNN	A, X	-	-	-

But not all code available

https://arxiv.org/pdf/1901.00596.pdf

Firmenich

Measurement & Instrument anomaly detection

Minor[1]:

Outliers detection against internal machine reference performance

Major[2]:

Change detection overtime analysis fluctuation (maintenance)

1: https://link.springer.com/article/10.1007/s41060-019-00186-0
2: https://www.frontiersin.org/articles/10.3389/fphys.2018.00325/full

“ESOL” BENCHMARK FOR WATER SOLUBILITY

I found 3 datasources available … :

- Delaney article (2004): 1144
- RDKit: Christos Kannas (2013 2nd RDKit UGM meeting): 1143
- Deepchem / Moleculenet.ai (2017): 1128

Recent papers use modified benchmark => up to 16 molecules removed :

- Are the harder one ? Hopefully not ;-)

“Data cleaning” is now recognize as a science…

«AqSolDB» August 2019:	9,982 unique compounds curated
Article	https://www.nature.com/articles/s41597-019-0151-1
Code & raw data	https://codeocean.com/capsule/8848590/tree/v1
Curate DB	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/OVHAW8

Sorkun et al.

48 persons downloaded it since … including me!

OUR PUBLISHED WORKS

Human Knowledge comes almost exclusively from text

What is for AI in Chemistry ?

Augmentation Text “CNF”	Kimber, T et al, https://arxiv.org/abs/1812.04439
Augmentation TextCNF/CNN	Tetko, IV et al, “Augmentation Is What You Need!” In *Artificial Neural Networks and Machine Learning – ICANN 2019*
GEN SMILES	Van Deursen R. et al, https://arxiv.org/abs/1909.04825
GEN Graph *G(V,E)*	Van Deursen R. et al, https://arxiv.org/abs/1909.11472
Retrosynthesis Transformers	https://chemrxiv.org/articles/A_Transformer_Model_for_Retrosynthesis/8058464/1 Karpov P, et al, “A Transformer Model for Retrosynthesis.” In *Artificial Neural Networks and Machine Learning – ICANN 2019*

More in the pipeline…

https://github.com/RuudFirsa

All codes available

https://github.com/bigchem/retrosynthesis

textCNF / CNN augmentation including in OCHEM interface (code available soon)