RELAY
THERAPEUTICS

Some (Hopefully) Useful Open Source Programs Built on the RDKit

Pat Walters

September 19, 2018

Journal of Molecular Graphics
Volume 11, Issue 2, June 1993, Pages 106-111

Papers

MOUSE: A teachable program for learning in conformational analysis

Daniel P. Dolata, W.Patrick Walters

Journal of Molecular Graphics
Volume 11, Issue 2, June 1993, Pages 112-117

Papers

Short-term learning in conformational analysis

Daniel P. Dolata, W.Patrick Walters

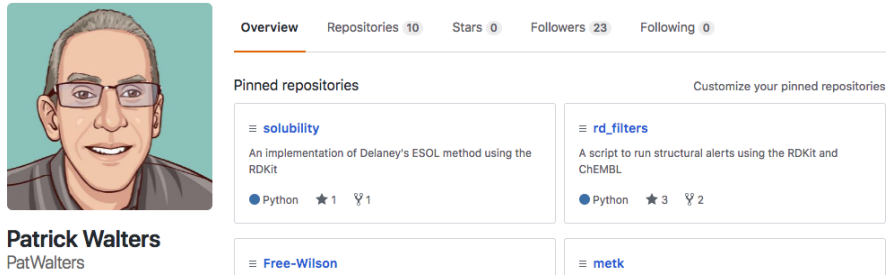
Overview

Free-Wilson Analysis

Filtering Chemical Libraries

Predicting (sort of) Aqueous Solubility

https://github.com/PatWalters



https://practicalcheminformatics.blogspot.com/

Have You Ever Been in This Situation?

RELAY THERAPEUTICS

Your project has synthesized several hundred compounds

You wonder what you might have missed

Is there any easy way to

- Evaluate contributions of different substituents
- Identify promising compounds which have yet to be synthesized

Have You Ever Been in This Situation?

RELAY THERAPEUTICS

Your project has synthesized several hundred compounds

You wonder what you might have missed

Is there any easy way to

- Evaluate contributions of different substituents
- Identify promising compounds which have yet to be synthesized

Journal of Medicinal Chemistry

VOLUME 7, NUMBER 4 JULY 6, 1964

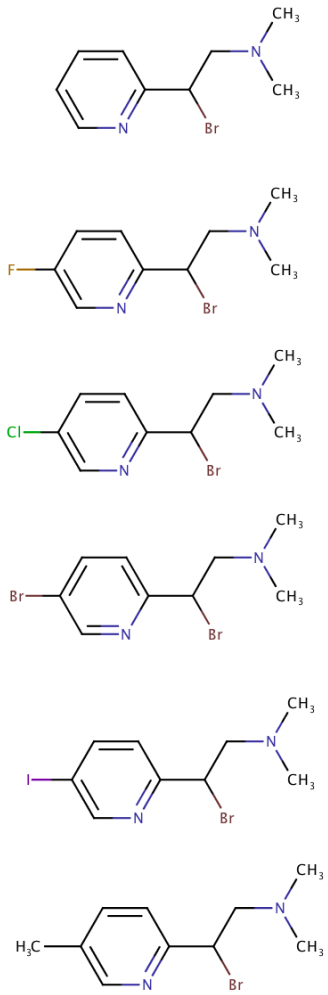
A Mathematical Contribution to Structure–Activity Studies

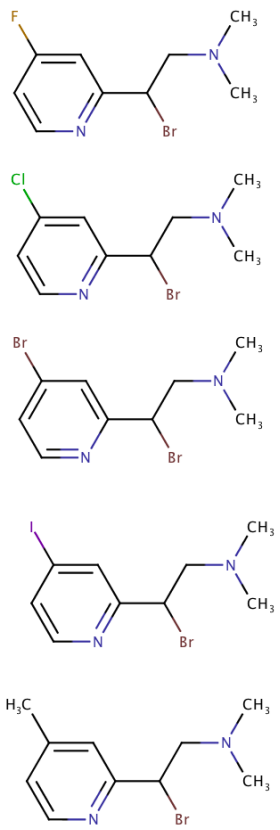
SPENCER M. FREE, JR., AND JAMES W. WILSON

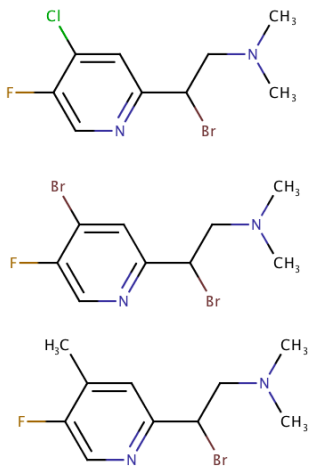
Research and Development Division, Smith Kline and French Laboratories, Philadelphia, Pennsylvania

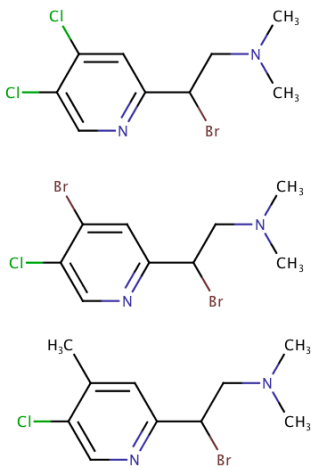
Received February 4, 1964

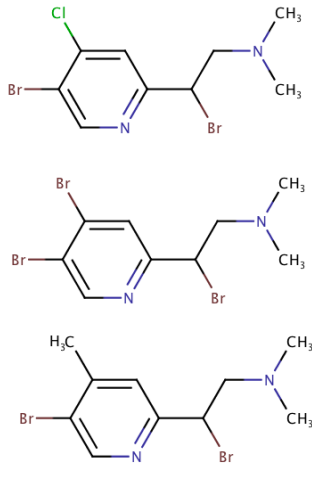
Here’s What We’ve Synthesized

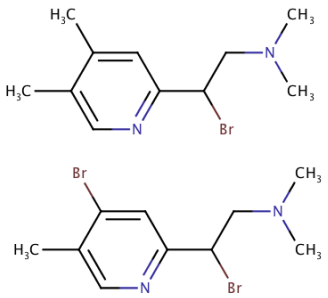




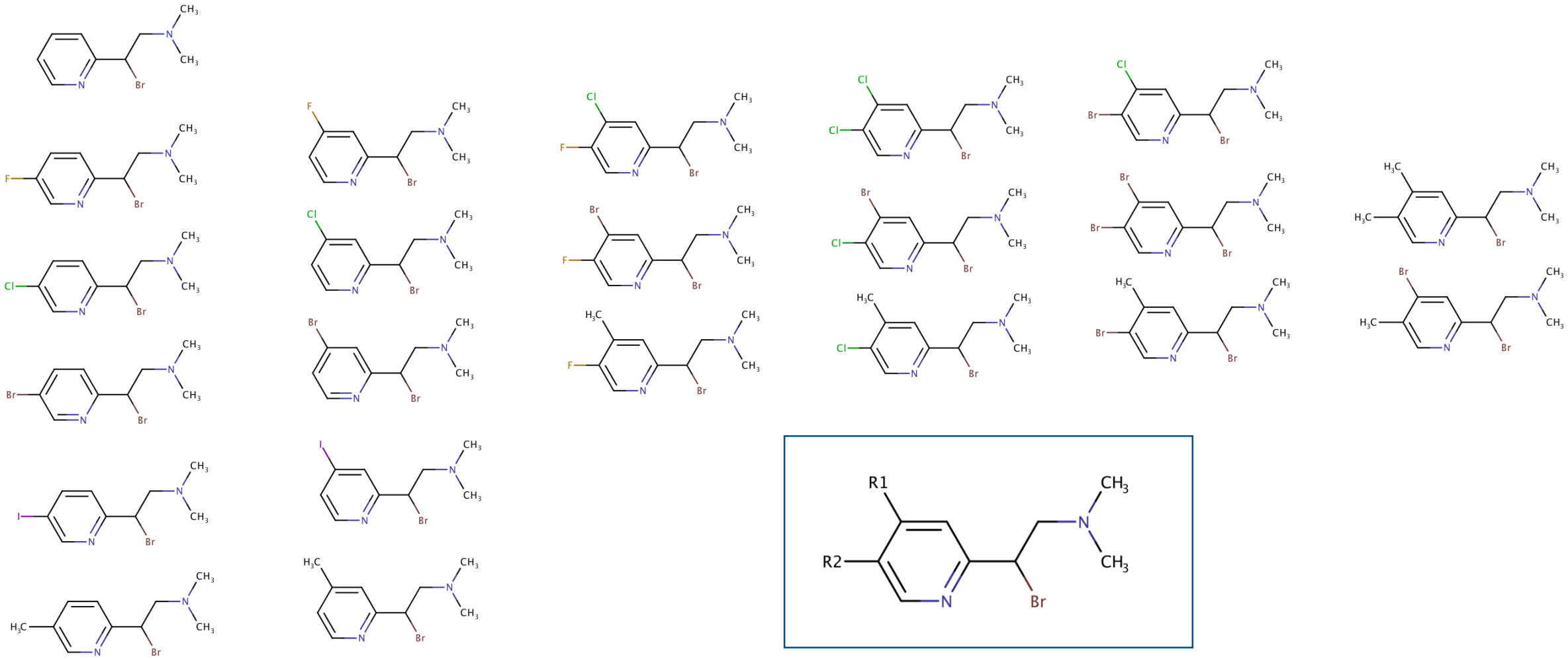




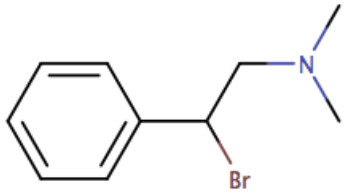
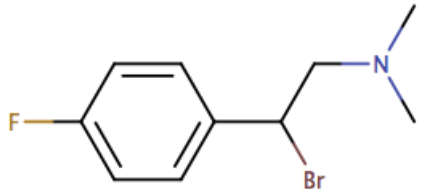
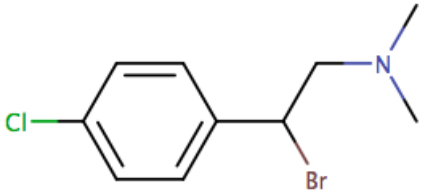
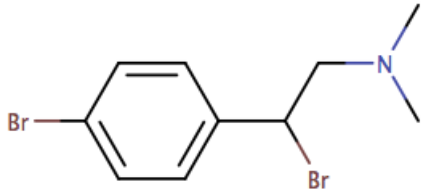


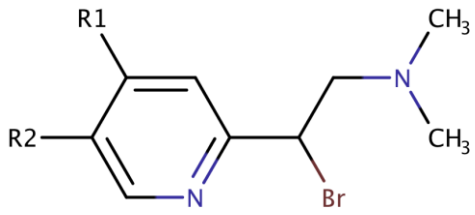


What Have We Missed?



Step 1 – Decompose the Molecules into R-Groups

	R1	R2
	H	H
	H	F
	H	Cl
	H	Br



Step 2 – Create a Matrix Containing Presence and Absence of R-Groups

	R1						R2					
[1]▲Name	H	F	Cl	Br	I	CH_3	H	F	Cl	Br	I	CH_3
MOL0001	1	0	0	0	0	0	1	0	0	0	0	0
MOL0002	1	0	0	0	0	0	0	1	0	0	0	0
MOL0003	1	0	0	0	0	0	0	0	1	0	0	0
MOL0004	1	0	0	0	0	0	0	0	0	1	0	0
MOL0005	1	0	0	0	0	0	0	0	0	0	1	0
MOL0006	1	0	0	0	0	0	0	0	0	0	0	1
MOL0007	0	1	0	0	0	0	1	0	0	0	0	0
MOL0008	0	0	1	0	0	0	1	0	0	0	0	0
MOL0009	0	0	0	1	0	0	1	0	0	0	0	0
MOL0010	0	0	0	0	1	0	1	0	0	0	0	0
MOL0011	0	0	0	0	0	1	1	0	0	0	0	0
MOL0012	0	0	1	0	0	0	0	1	0	0	0	0
MOL0013	0	0	0	1	0	0	0	1	0	0	0	0
MOL0014	0	0	0	0	0	1	0	1	0	0	0	0
MOL0015	0	0	1	0	0	0	0	0	1	0	0	0
MOL0016	0	0	0	1	0	0	0	0	1	0	0	0
MOL0017	0	0	0	0	0	1	0	0	1	0	0	0
MOL0018	0	0	1	0	0	0	0	0	0	1	0	0
MOL0019	0	0	0	1	0	0	0	0	0	1	0	0
MOL0020	0	0	0	0	0	1	0	0	0	1	0	0
MOL0021	0	0	0	0	0	1	0	0	0	0	0	1
MOL0022	0	0	0	1	0	0	0	0	0	0	0	1

Step 3– Regress R-Group Vectors vs pIC50

	X												Y
[1]▲Name	H	F	Cl	Br	I	CH_3	H	F	Cl	Br	I	CH_3	pIC_{50}
MOL0001	1	0	0	0	0	0	1	0	0	0	0	0	7.5
MOL0002	1	0	0	0	0	0	0	1	0	0	0	0	8.2
MOL0003	1	0	0	0	0	0	0	0	1	0	0	0	8.7
MOL0004	1	0	0	0	0	0	0	0	0	1	0	0	8.9
MOL0005	1	0	0	0	0	0	0	0	0	0	1	0	9.2
MOL0006	1	0	0	0	0	0	0	0	0	0	0	1	9.3
MOL0007	0	1	0	0	0	0	1	0	0	0	0	0	7.5
MOL0008	0	0	1	0	0	0	1	0	0	0	0	0	8.2
MOL0009	0	0	0	1	0	0	1	0	0	0	0	0	8.3
MOL0010	0	0	0	0	1	0	1	0	0	0	0	0	8.4
MOL0011	0	0	0	0	0	1	1	0	0	0	0	0	8.5
MOL0012	0	0	1	0	0	0	0	1	0	0	0	0	8.2
MOL0013	0	0	0	1	0	0	0	1	0	0	0	0	8.6
MOL0014	0	0	0	0	0	1	0	1	0	0	0	0	8.8
MOL0015	0	0	1	0	0	0	0	0	1	0	0	0	8.9
MOL0016	0	0	0	1	0	0	0	0	1	0	0	0	8.9
MOL0017	0	0	0	0	0	1	0	0	1	0	0	0	9.0
MOL0018	0	0	1	0	0	0	0	0	0	1	0	0	9.0
MOL0019	0	0	0	1	0	0	0	0	0	1	0	0	9.4
MOL0020	0	0	0	0	0	1	0	0	0	1	0	0	9.2
MOL0021	0	0	0	0	0	1	0	0	0	0	0	1	9.3
MOL0022	0	0	0	1	0	0	0	0	0	0	0	1	9.5

When is Linear Regression Poorly Behaved?

Number of characteristics (x-values) exceeds the number of samples (y-values)

Characteristics are colinear

Linear Regression

$$\text{Loss} = \Sigma(\hat{Y}_i - Y_i)^2$$

Ridge Regression

$$\text{Loss} = \Sigma(\hat{Y}_i - Y_i)^2 + \lambda\Sigma\beta^2$$

Step 5 - Examine Coefficients to Evaluate Substituent Contributions

R_1

R-Group SMILES	[1]▼Coefficient	R-group	Count
1-	0.193	1	5
*1-Br	0.176	1	5
*1-I	0.123	1	1
*1-Cl	-0.039	1	4
*1-H	-0.135	1	6
*1-F	-0.317	1	1

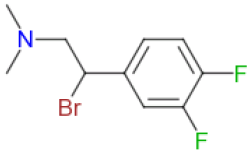
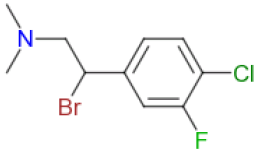
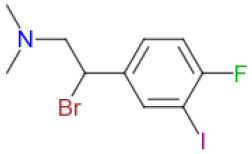
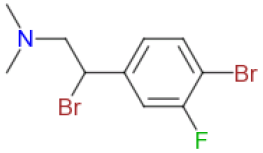
R_2

R-Group SMILES	[1]▼Coefficient	R-group	Count
*2-	0.386	2	3
*2-I	0.302	2	1
*2-Br	0.228	2	4
*2-Cl	0.026	2	4
*2-F	-0.316	2	4
*2-H	-0.627	2	6

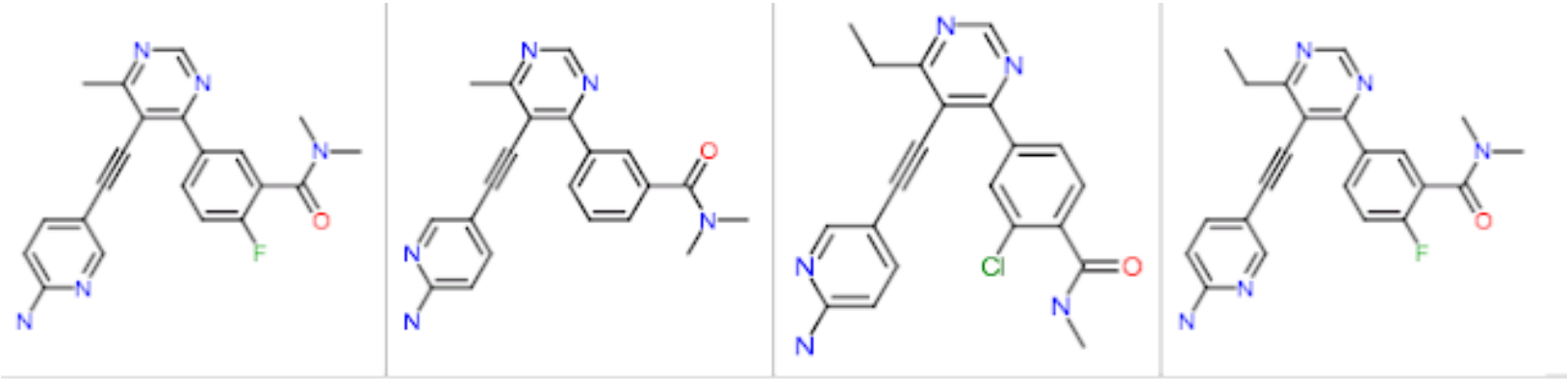
Examine Multivariate Contributions

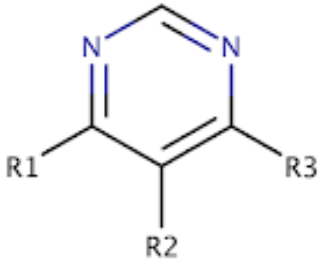
Bioavailability	Cellular IC_{50}	hERG IC_{50}
106.0	0.0004	0.0708
134.7	0.0006	0.316
75.2	0.2090	21.9
76.8	0.0117	4.07
40.3	0.4370	28.2
40.1	0.3720	28.2
87.7	0.0079	6.31

Examine Promising Combinations That Have Yet to be Synthesized

Molecule	R_1	R_2	Predicted pIC_{50}
			8.15
			8.49
			8.59
			8.69

What About Symmetric Scaffolds

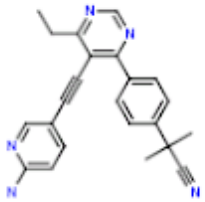
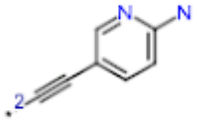
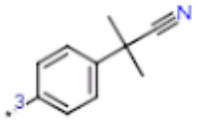
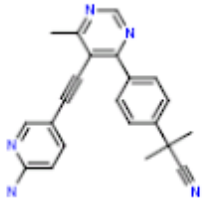
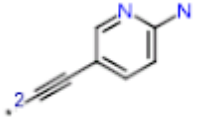
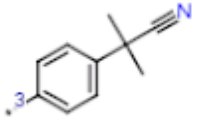
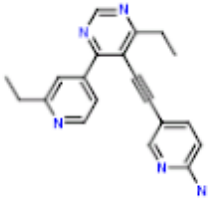
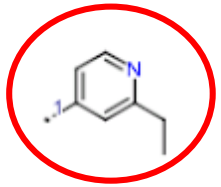
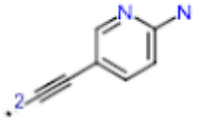
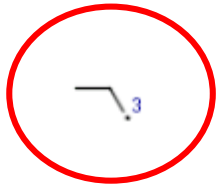
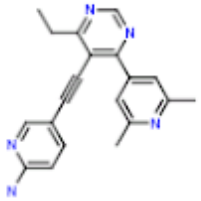
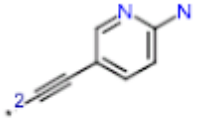
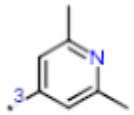




Where will the aromatic group end up?

CHEMBL3638592

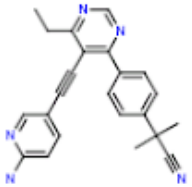
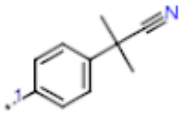
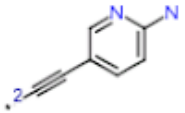
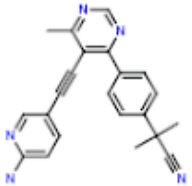
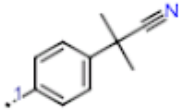
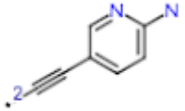
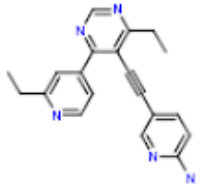
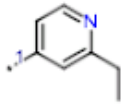
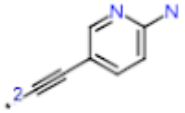
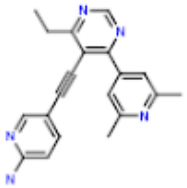
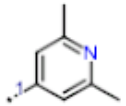
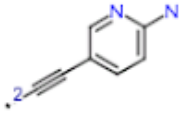
This is Not the Desired Result

SMILES	Name	R1_SMILES	R2_SMILES	R3_SMILES
	1973628			
	1973629			
	1973630			
	1973631			

Mixture of alkyl and aryl at R1 and R3

The “—smarts” Option “Pins” an R-group based on SMARTS

```
free_wilson.py rgroup --scaffold CHEMBL3638592_scaffold.mol --in
CHEMBL3638592.smi --prefix CHEMBL3638592 --smarts "3|c"
```

SMILES	Name	R1_SMILES	R2_SMILES	R3_SMILES
	1973628			
	1973629			
	1973630			
	1973631			

We Can Also Use Recursive SMARTS to Pin the Alkyl Group

anchor

ethyl

[#0;$([#0][CH3]),$([#0][CH2][CH3])]

methyl

or

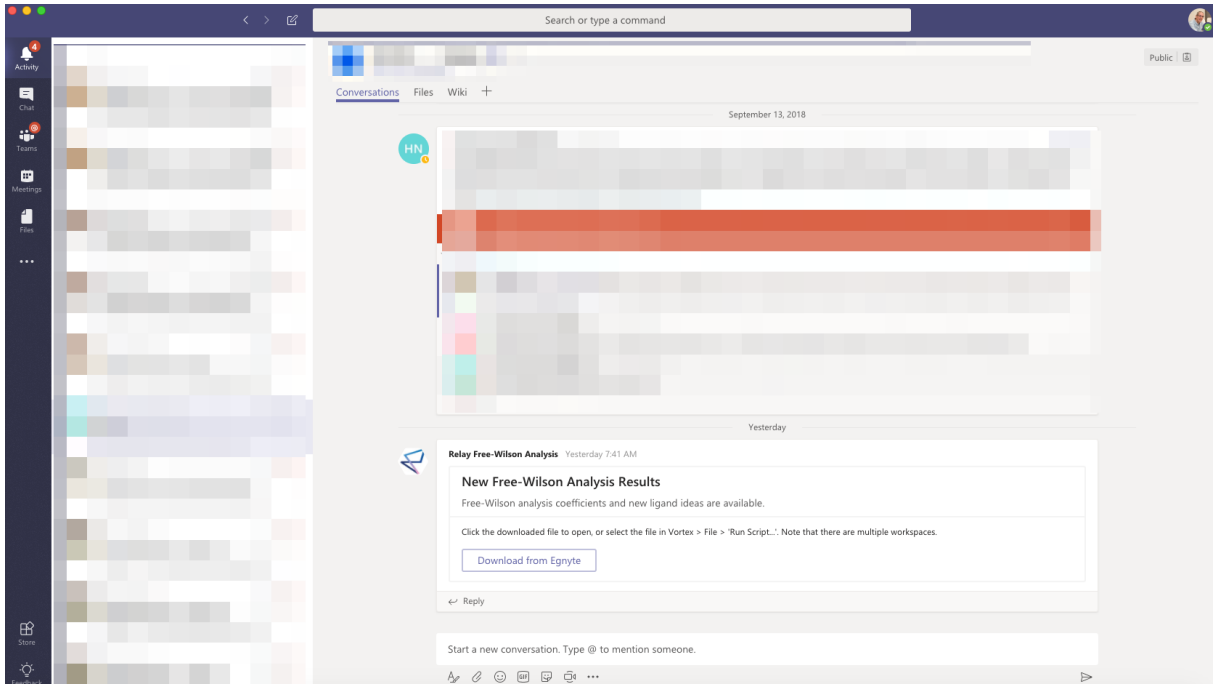
```
free_wilson.py rgroup --scaffold CHEMBL3638592_scaffold.mol --in
CHEMBL3638592.smi --prefix CHEMBL3638592 --smarts "3|[#0;$([#0][CH3]),$([#0][CH2][CH3])]"
```

Using a Chat Platform as the Center of an Informatics Infrastructure

RELAY THERAPEUTICS

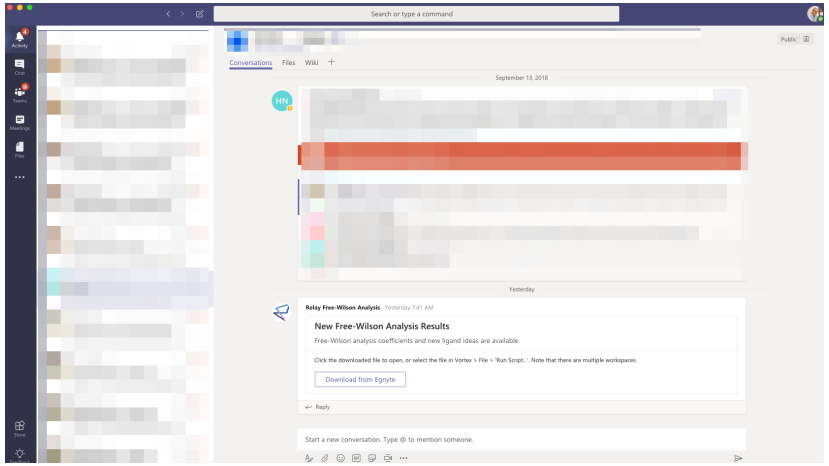
Dedicated channels for each drug discovery project

- Literature
- Assay Results
- ADME and PK
- Computation
- Biology
- Chemistry
- Structural Biology
- SAR Analysis

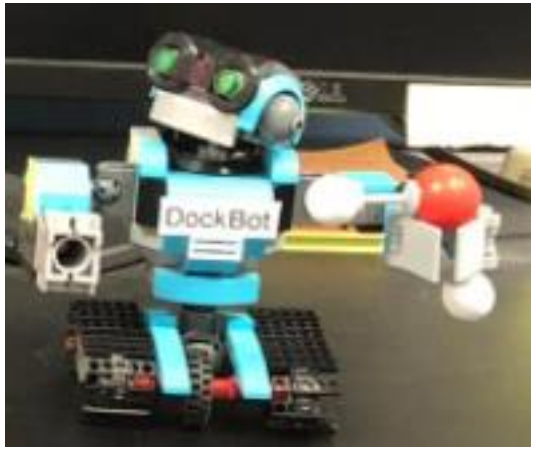


Microsoft Teams

Free-Wilson and the Relay Bot Infrastructure



Microsoft Teams



Relay Bots

Patents

Free-Wilson

Docking

ML Models

Similarity Calcs

Molly Schmidt and Jon Weiss

Filtering Chemical Libraries

Journal of Computer-Aided Molecular Design, **16:** 311–323, 2002.
KLUWER/ESCOM

Filtering databases and chemical libraries

Paul S. Charifson and W. Patrick Walters
Vertex Pharmaceuticals, 130 Waverly St, Cambridge, MA 02139, USA

molecular informatics
models – molecules – systems

Full Paper

Compound Selection and Filtering in Library Design

James A. Lumley

First published: 22 November 2005 | https://doi.org/10.1002/qsar.200520136 | Cited by: 14

ELSEVIER
Drug Discovery Today
Volume 2, Issue 9, September 1997, Pages 382-384

Review

Reactive compounds and *in vitro* false positives in HTS

Gilbert M. Rishton

Show more

https://doi.org/10.1016/S1359-6446(97)01083-0

SpringerOpen | Journal of Cheminformatics

J Cheminform. 2016; 8: 29.
Published online 2016 May 28. doi: 10.1186/s13321-016-0137-3
PMCID: PMC4884375
PMID: 27239230

Badapple: promiscuity patterns from noisy evidence

Jeremy J. Yang, Oleg Ursu, Christopher A. Lipinski, Larry A. Sklar, Tudor I. Oprea, and Cristian G. Bologa

Author information ▶ Article notes ▶ Copyright and License information ▶ Disclaimer

Journal of Medicinal Chemistry
Article
pubs.acs.org/jmc

Rules for Identifying Potentially Reactive or Promiscuous Compounds

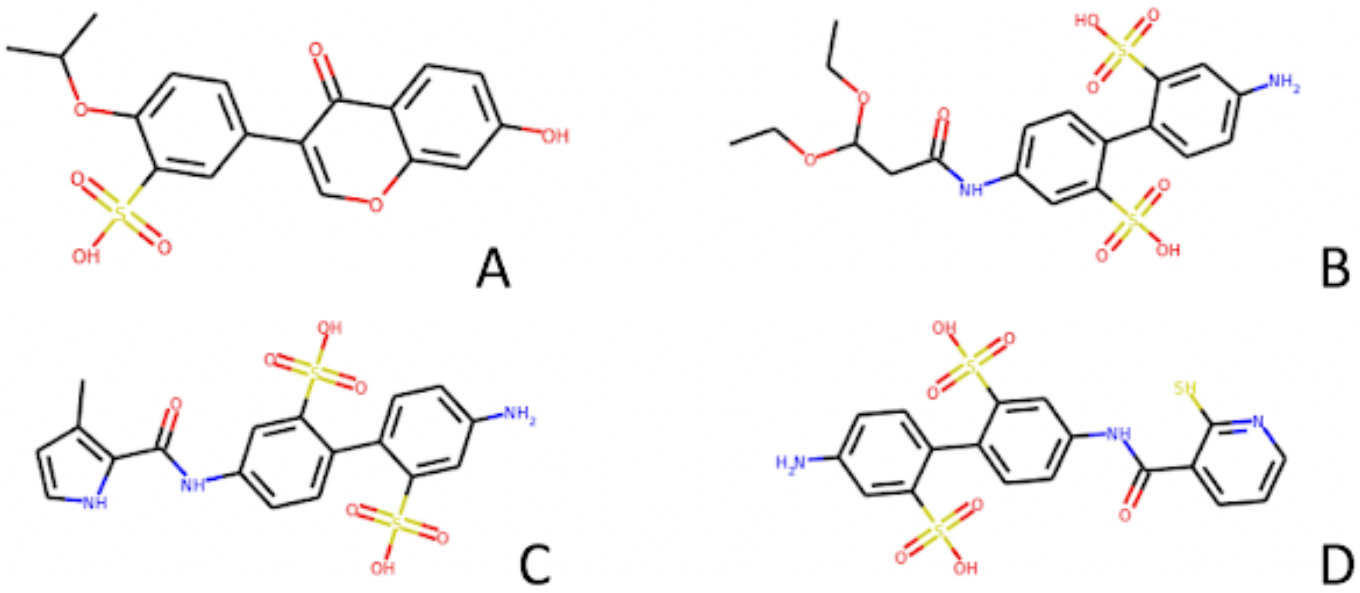
Robert F. Bruns* and Ian A. Watson
Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, Indiana 46285, United States

Journal of Medicinal Chemistry
Article
J. Med. Chem. **2010,** *53,* 2719–2740 **2719**
DOI: 10.1021/jm901137j

New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays

Jonathan B. Baell*[,†,‡] and Georgina A. Holloway[†,‡]
[†]*The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia and* [‡]*Cancer Therapeutics-CRC P/L, 4 Research Avenue, La Trobe R&D Park, Bundoora, Victoria 3086, Australia*

Filtering Chemical Libraries - My Motivation



https://github.com/lilleswing/deepchem/blob/large-scale-chemical-screens/examples/notebooks/Large_Scale_Chemical_Screens.ipynb

Structural Alerts Available in ChEMBL 20

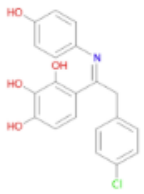
Structural Alerts

We have compiled a number of sets of publicly-available structural alerts where SMARTS were readily available and useable; these include Pfizer LINT filters, Glaxo Wellcome Hard Filters, Bristol-Myers Squibb HTS Deck Filters, NIH MLSMR Excluded Functionality Filters, University of Dundee NTD Screening Library Filters and Pan Assay Interference Compounds (PAINS) Filters. These sets of filters aim to identify compounds that **could** be problematic in a drug-discovery setting for various different reasons (*e.g.*, substructural/functional group features that might be associated with toxicity or instability in *in vivo* info settings, compounds that might interfere with assays and for example, appear to be 'frequent hitters' in HTS).

It should be noted however that some alerts/alert sets are more permissive than others and may flag a large number of compounds. Results should therefore be interpreted with care, depending on the use-case, and not treated as a blanket filter (*e.g.*, around 50% of approved drugs have 1 or more alerts from these pooled sets). The compound report card page now provides a summary count of the number of structural alerts hits picked up by a given molecule:

Compound Name and Classification

Compound ID	CHEMBL1082532
Compound Name	
ChEMBL Synonyms	
Max Phase	0
Trade Names	
Molecular Formula	C20H16ClNO4



CHEMBL1082532

Additional synonyms for CHEMBL1082532 found using NCI Chemical Identifier Resolver

Compound Representations

Molfile	Download MolFile
Canonical SMILES	Oc1ccc(cc1)\N=C(\Cc2ccc(Cl)cc2)/c3ccc(O)c(O)c3O
Standard InChI	InChI=1S/C20H16ClNO4/c21-13-3-1-12(2-4-13)11-17(22-14-5-7-15 ... Download InChI
Standard InChI Key	SIPOGQRDQZRLJC-XLNRJJMWSA-N

New **Structural Alerts**

There are 9 structural alerts for CHEMBL1082532. To view alerts please click here.

http://chembl.blogspot.com/2015/02/chembl-20-released.html

Structural Alert Sets in the ChEMBL Database

Pfizer LINT filters

Glaxo Wellcome Hard Filters

BMS HTS Deck Filters

NIH MLSMR Excluded Functionality Fitlers

University of Dundee NTD Screening Library Filters

Pan Assay Interference Compounds (PAINS) Filters

Inpharmatica Filters

SureChEMBL Filters

How can I apply these rules to my compound set?

rdfilters Applies Functional Group and Property Filters to Compound Sets

Usage:

rd_filters filter --in INPUT_FILE --prefix PREFIX [--rules RULES_FILE_NAME] [--alerts ALERT_FILE_NAME][--np NUM_CORES]

rd_filters template --out TEMPLATE_FILE [--rules RULES_FILE_NAME]

Options:

--in INPUT_FILE input file name

--prefix PREFIX prefix for output file names

--rules RULES_FILE_NAME name of the rules JSON file

--alerts ALERTS_FILE_NAME name of the structural alerts file

--np NUM_CORES the number of cpu cores to use (default is all)

--out TEMPLATE_FILE parameter template file name

Runs in parallel using pool.map()

rdfilters Template Files Control Operation

```
more tmplt.json
{
    "HBA": [
        0,
        10
    ],
    "HBD": [
        0,
        5
    ],
    "LogP": [
        -5,
        5
    ],
    "MW": [
        0,
        500
    ],
    "Rule_BMS": false,
    "Rule_Dundee": false,
    "Rule_Glaxo": false,
    "Rule_Inpharmatica": true,
    "Rule_LINT": false,
    "Rule_MLSMR": false,
    "Rule_PAINS": false,
    "Rule_SureChEMBL": false,
    "TPSA": [
        0,
        200
    ]
}
```

Predicting Aqueous Solubility

1000 *J. Chem. Inf. Comput. Sci.* **2004,** *44,* 1000−1005

ESOL: Estimating Aqueous Solubility Directly from Molecular Structure

John S. Delaney*

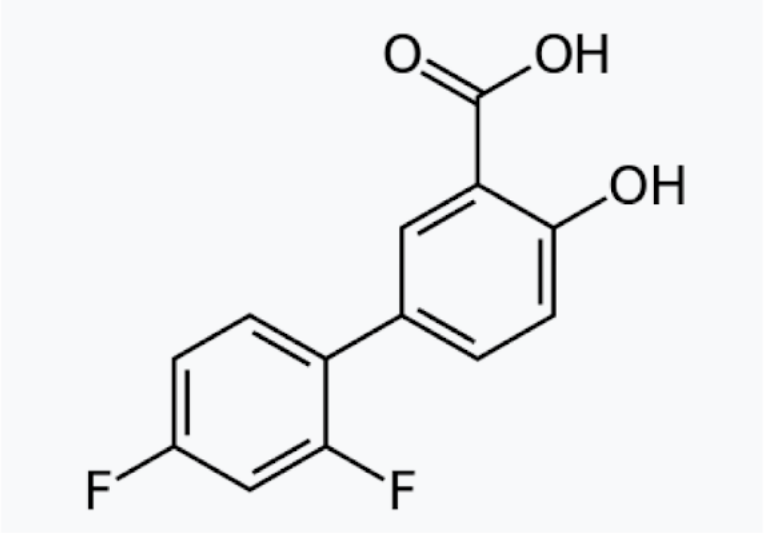
Syngenta, Jealott's Hill International Research Centre, Bracknell, Berkshire, RG42 6EY, United Kingdom

Received October 29, 2003

$$\mathrm{LogS} = 0.16 - 0.63\ \mathrm{cLogP} - 0.0062\ \mathrm{MW} + 0.066\ \mathrm{RB} - 0.74\ \mathrm{AP}$$

Useful method published in 2004
Data set has become a standard QSPR benchmark

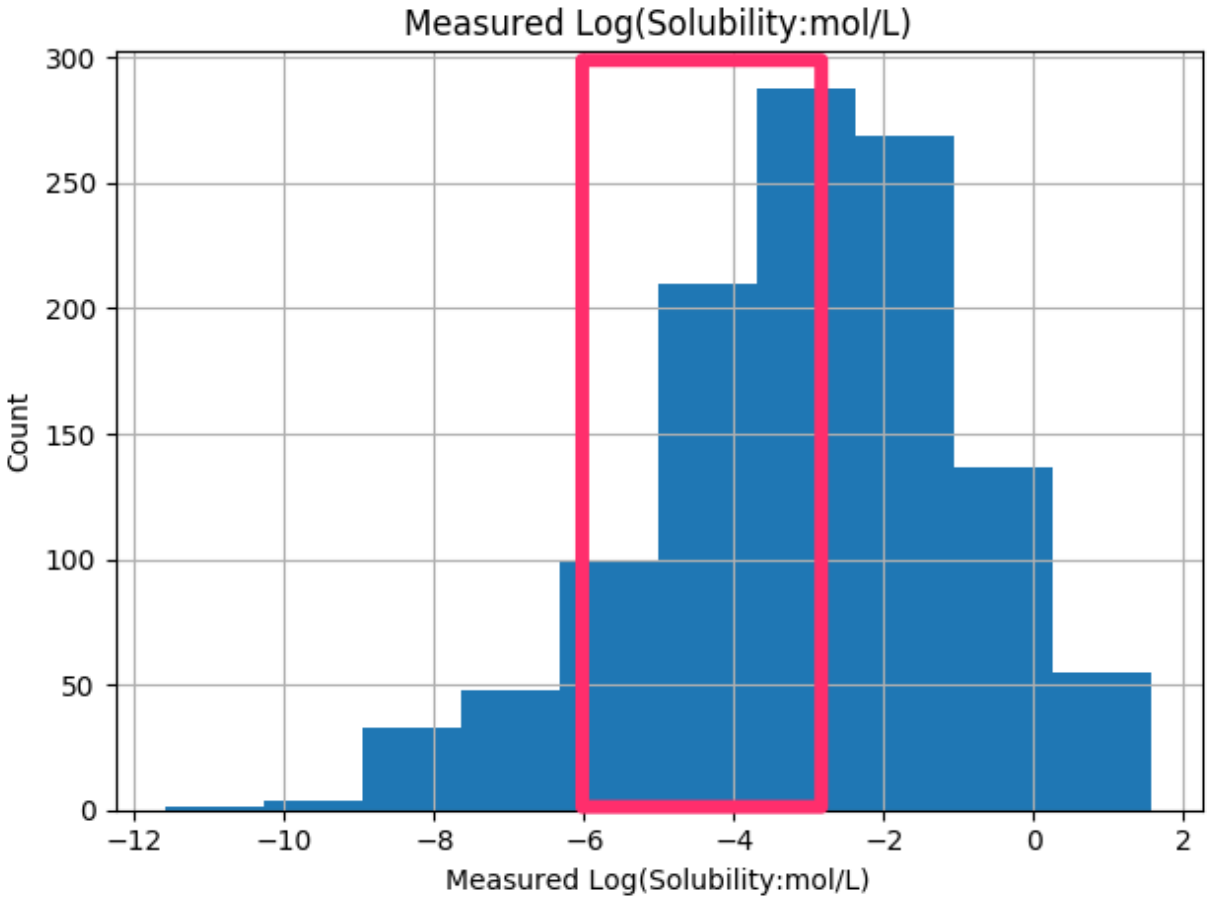
Even Experimental Solubility Measurements are Tricky



Diflunisal

Form	Solubility µg/ml	LogS mol/L
1	26	-3.9
2	7.6	-4.5
3	0.93	-5.4
4	0.29	-5.9

Most Solubility Datasets Have an Unrealistic Dynamic Range



Btw, most activity datasets have an equally unrealistic dynamic range

Testing the ESOL Implementation

Train on Delaney dataset

Test on 56 compounds from the University of St Andrews DLS-100 dataset

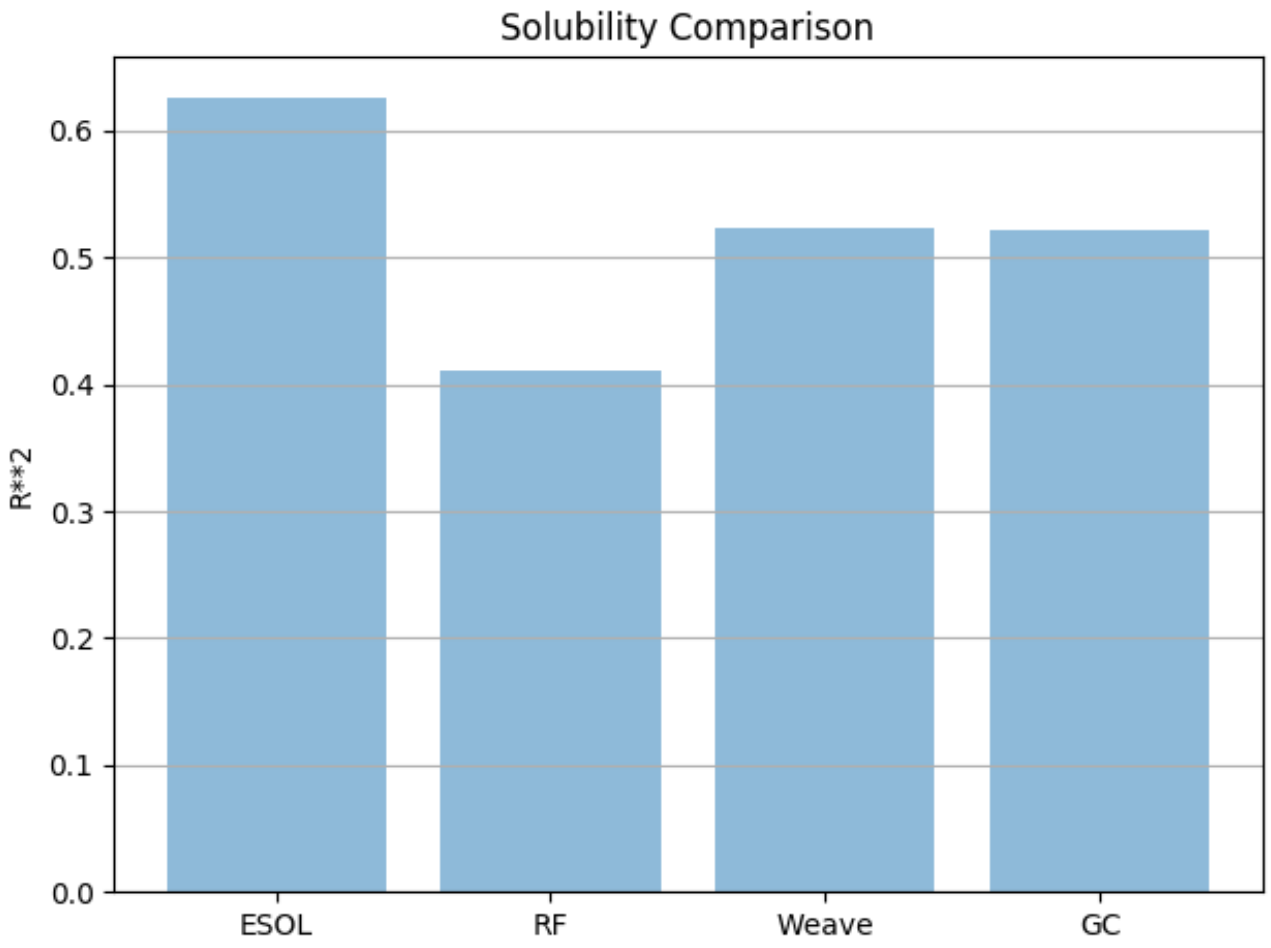
- John B. O. Mitchell

https://risweb.st-andrews.ac.uk/portal/en/datasets/dls100-solubility-dataset(3a3a5abc-8458-4924-8e6c-b804347605e8).html

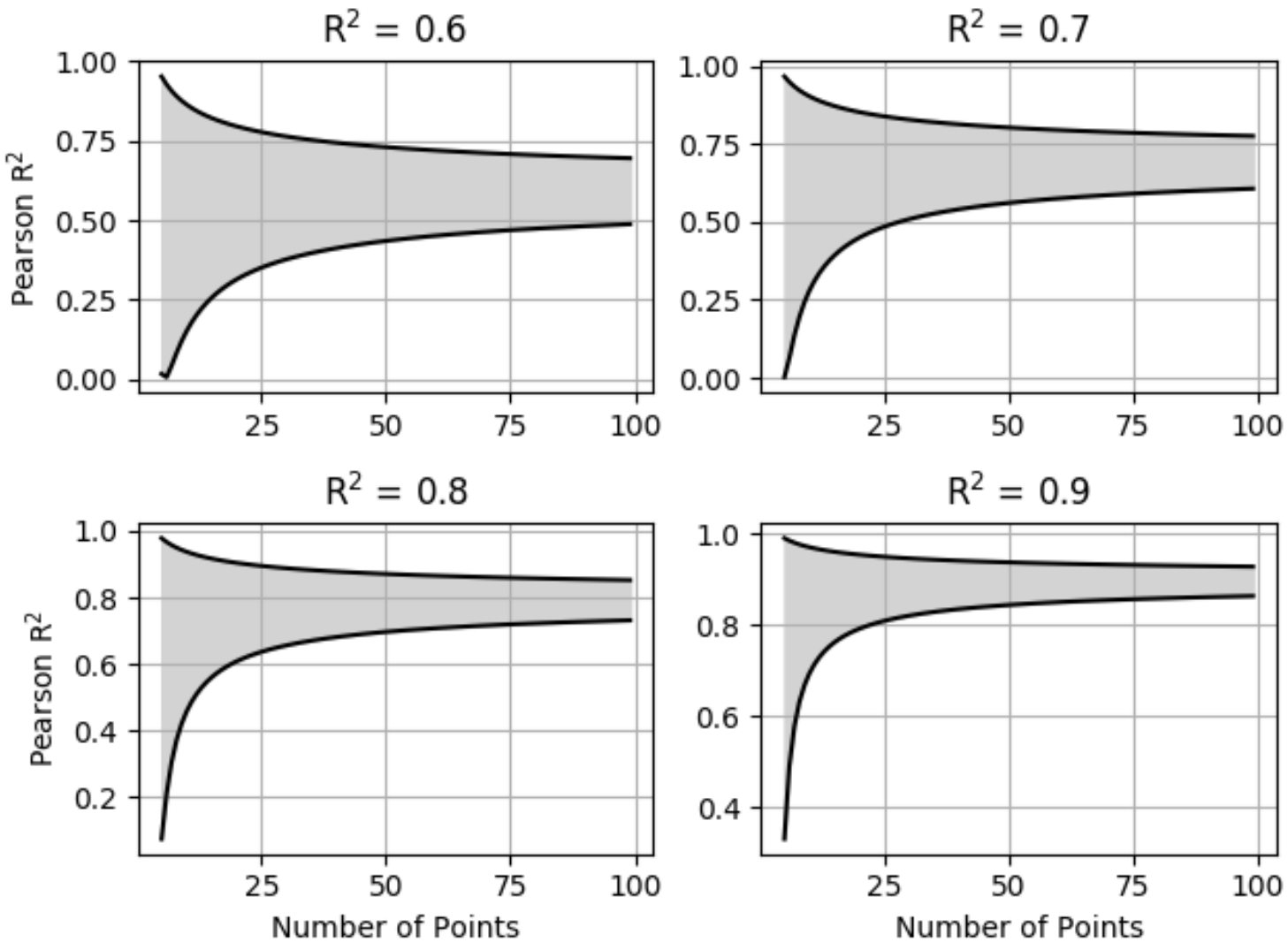
Compare with 3 methods from DeepChem

- Random Forest
- Weave
- Graph Convolutions

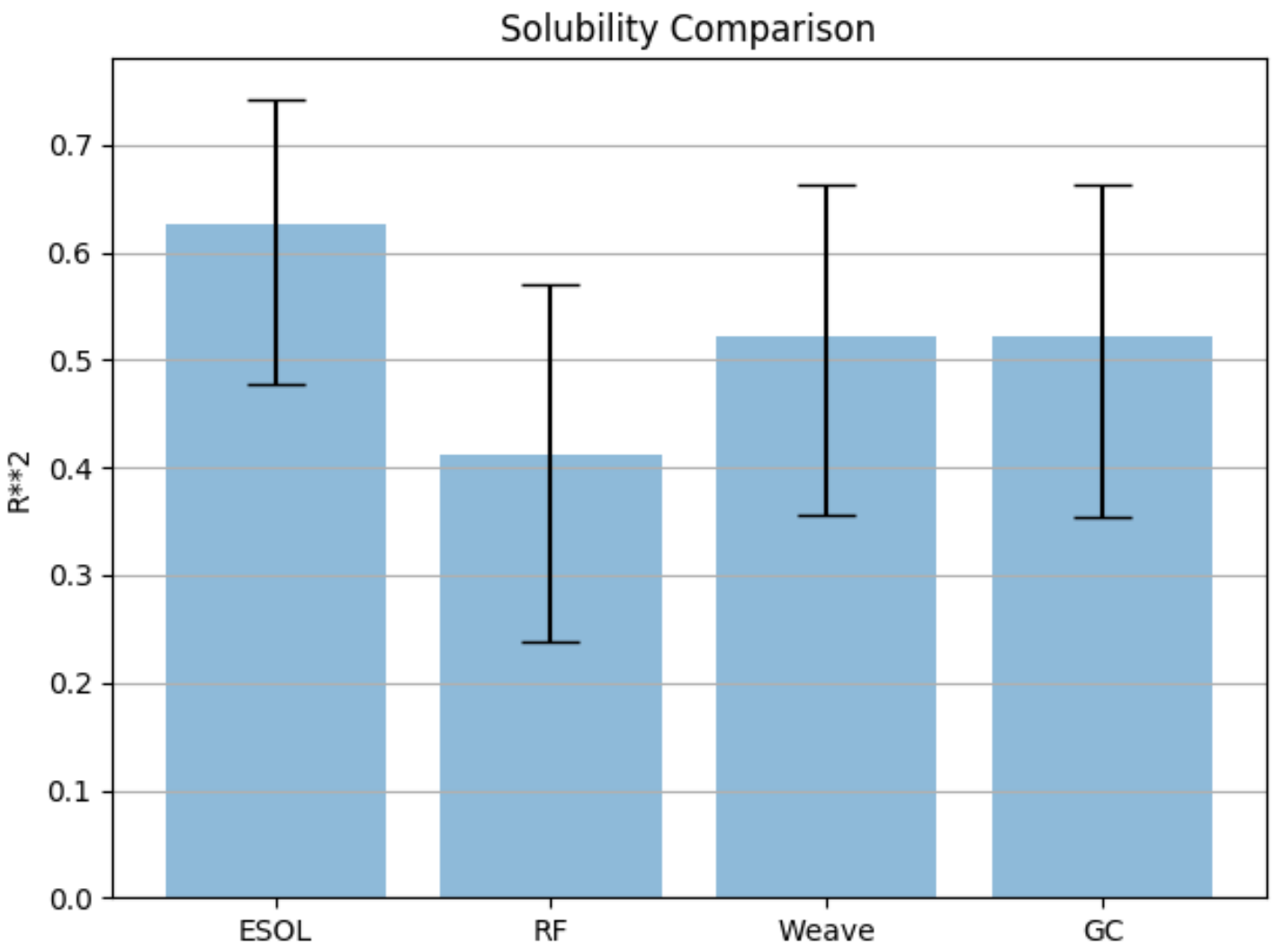
Does ESOL Outperform Deep Learning



Correlation Have Confidence Limits!



No Difference Within the 95% Confidence Limits



Technological Underpinnings

RDKit

Open-Source Cheminformatics and Machine Learning

Pandas

Acknowledgements

Hakan Gunaydin

Brandi Hudson

Demetri Moustakas

Mark Murcko

Nick Pabon

Levi Pierce

Molly Schmidt

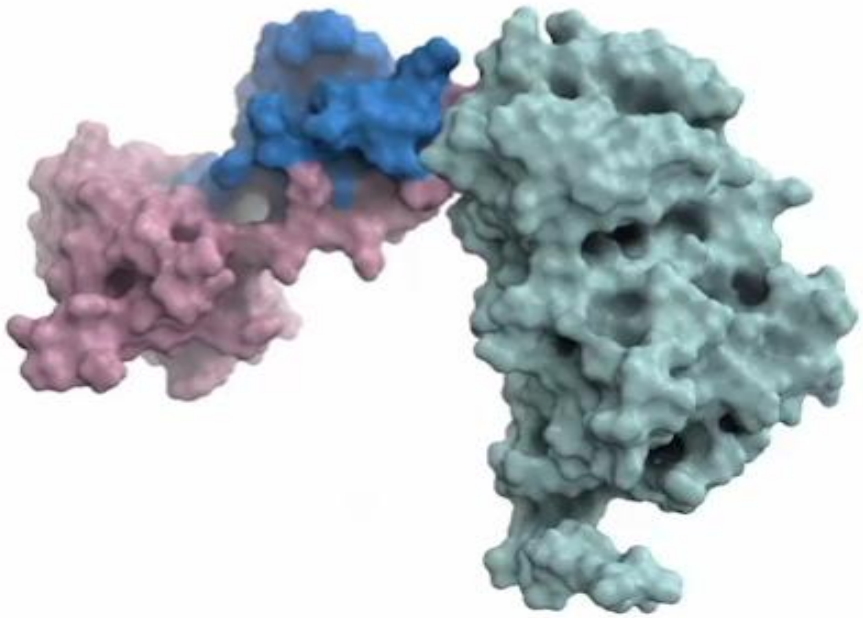
Jon Weiss

Paul Charifson

Emanuele Perola

Greg Landrum

The RDKit Community



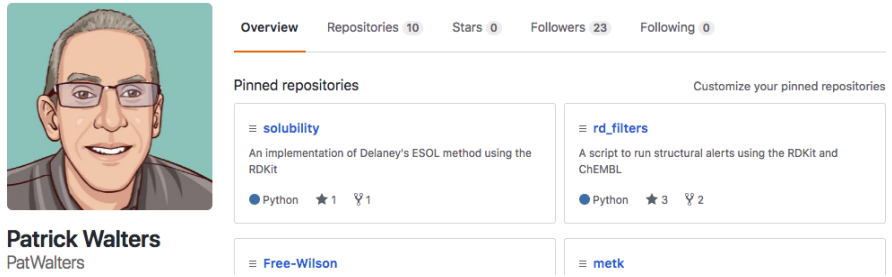
That's It – More to Come (Hopefully) Soon – Stay Tuned!

Free-Wilson Analysis

Filtering Chemical Libraries

Predicting (sort of) Aqueous Solubility

https://github.com/PatWalters



@wpwalters

https://practicalcheminformatics.blogspot.com/

RELAY

THERAPEUTICS