

 Community Prediction Competition

1st EUOS/SLAS Joint Challenge: Compound Solubility

Develop new methods to predict compound solubility based on chemical structure.

100 teams · 3 months ago

eu::openscreen

 **slas**
Care. Transform. Research.

Challenge organizers:

Robert Harmel, Andrea Zaliani, **Wenyu Wang#**, Julio Martin,

Christian Parker, Jing Tang#

#Network Pharmacology for Precision Medicine (**NetPhar**),

Faculty of Medicine, University of Helsinki

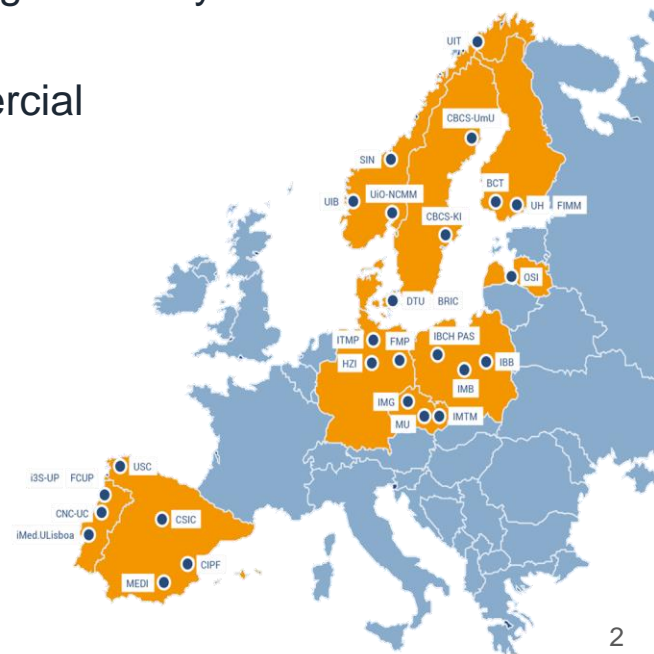
Date: 2023/03/23

EU-OPENSREEN



- European Research Infrastructure Consortium (ERIC)
- Non-profit organization for chemical biology and early drug discovery
- EU-wide integration of high-capacity screening platforms
- ECBL (European Chemical Biology Library): 100k commercial compounds from EU-OS partners

Solubility	University of Santiago de Compostela (USC)
Interference with bioluminescence reporters	Polish Academy of Sciences, Institute of Bioorganic Chemistry (IBCH PAS)
ROS (Reactive Oxygen Species)	Polish Academy of Sciences, Institute of Bioorganic Chemistry (IBCH PAS)
Cell viability	Institute for Molecular Medicine Finland (FIMM)
Antibacterial & antifungal assays	Fundación MEDINA (MEDI) Helmholtz-Centre for Infection Research (HZI)
Absorbance / autofluorescence	EU-OPENSREEN laboratory
Cell painting	Ongoing assay validation at four sites



EU-OPENSREEN



ECBD (European chemical biology database): a collaborative data sharing environment

- FAIR data principles
- Optional embargo period up to 36 months
- Annotations + links to other databases (e.g. ChEMBL)

<https://ecbd.eu/>

A screenshot of the ECBD website. The top navigation bar includes the "ecbd" logo and buttons for "ASSAYS", "COMPOUNDS", "TARGETS", "HELP", "LOGIN", and "OTHER". The main content area features the "eu:openscreen" logo and a large blue banner with the text "european chemical biology database" and three circular statistics: 102,168 compounds, 3 targets, and 6 assays.

ecbd ASSAYS COMPOUNDS TARGETS HELP LOGIN OTHER

eu:openscreen

european
chemical biology
database

102,168
compounds

3
targets

6
assays

Challenge motivation

- Impossible to probe million of compounds experimentally
- Chemical properties impact compound behavior in the environment.
- Computational predictions accelerate the research
- 1st Kaggle challenge: with **Society of Lab Automation & Screening**
- Solubility: an essential feature of all biologically active compounds



 Community Prediction Competition

1st EUOS/SLAS Joint Challenge: Compound Solubility

Develop new methods to predict compound solubility based on chemical structure.

100 teams · 3 months ago

eu:openscreen

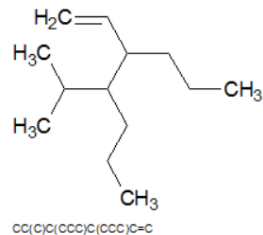
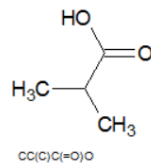
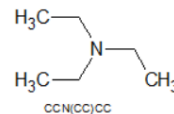
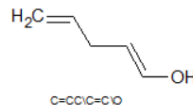
 slas
Come Transform Research

Challenge data



- Experimentally measured aqueous solubility of 100k small molecules
- 70K Training / 15K Public leaderboard / 15K Final evaluation
- Structure based predictors generated by participants (SMILE / InChIKey)
- Labels are solubility classes:

High (93%), Medium (4%) or Low (3%)



eos	smiles	inchi	inchikey	formula	mw
EOS102046	<chem>NS(N)(=O)=O</chem>	InChI=1S/H4N2O2S...	NVBFBHJWHLNUMCV...	H4N2O2S	96.111
EOS100468	<chem>O=P(O)(O)CP(=O)(...</chem>	InChI=1S/CH6O6P2...	MBKDYNNUVRNNR...	CH6O6P2	176.001
EOS100593	<chem>ClN=C(N)NN</chem>	InChI=1S/CH6N4.Cl...	UBDZFAGVPPMTIT...	CH7ClN4	110.548
EOS102045	<chem>NC(N)=S</chem>	InChI=1S/CH4N2S/c...	UMGDCJDMYOKAJ...	CH4N2S	76.124
EOS102399	<chem>O=C([O-])O.[Na+]</chem>	InChI=1S/CH2O3.Na...	UIIMBOGNXHQVGW...	CHNaO3	84.006
EOS13582	<chem>Cc1ccc2oc(=O)n(S(...</chem>	InChI=1S/C9H9NO4...	HAJTYZHIIIDXHX-U...	C9H9NO4S	227.241
EOS36981	<chem>CCN1C(=O)c2ccccc...</chem>	InChI=1S/C9H9NO3...	DQKSIWDBRCCINU...	C9H9NO3S	211.242
EOS102672	<chem>CCOc1cc(C#N)ccc1O</chem>	InChI=1S/C9H9NO2...	NBUPJWDUINJHFZ-...	C9H9NO2	163.176

Evaluation metric

- Accuracy is not an informative metric:
 - > 0.9 for a prediction based on class distribution of training data
- Kappa for imbalanced classification
- Quadratic weighted kappa for ordinal labels

$$\text{Cohen's Kappa} = 1 - \frac{\text{Error}}{\text{Baseline error}}$$

$$\text{Weighted Kappa} = 1 - \frac{\text{Weighted error}}{\text{Weighted baseline error}}$$

		Model Predicted		
		0	1	2
Ground Truth	$n_{i,j}$			
	0	4	0	1
	1	2	1	0
2	0	1	1	

		Weight matrix		
		0	1	2
0	0	1	4	
1	1	0	1	
2	4	1	0	

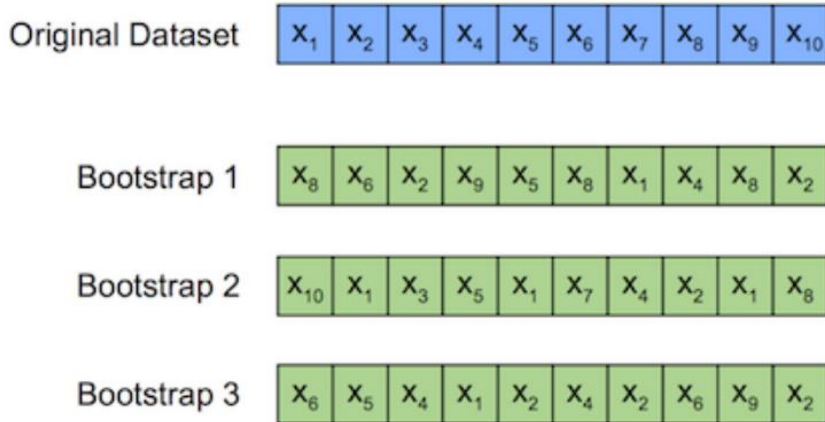
Kaggle default evaluation

Team name	prediction accuracy
a	0.9
b	0.8
c	0.7
d	0.6
..	..

- Only one value per team
- No uncertainty estimation
- No significant test
- Hard to make persuasive conclusion for teams with close performance

Bootstrap sampling based evaluation

$$K_m = \frac{\sum_{n=1}^{1000} \mathbb{1}_{Kappa_{n,m} \leq Kappa_{n,ref}}}{\sum_{n=1}^{1000} \mathbb{1}_{Kappa_{n,m} \geq Kappa_{n,ref}}}$$



Team name	resample id	ACC	mean	sd	Significance (Bayes factor)
a	1	0.95			
a	2	0.92			
a	3	0.69	0.85	0.14	Reference
b	1	0.73			
b	2	0.98			
b	3	0.82	0.84	0.13	0.33
c	1	0.76			
c	2	0.52			
c	3	0.35	0.54	0.21	12.86

Winner selection

- Rank: private leaderboard kappa
- Statistically tied teams: BF value < 5
- Solution has to be shared
- **Interpretable** (model transferable to future EUOS data):

Do not use EOS ID

Model accuracy stable upon shuffling the test set

Winner selection

- Rank: private leaderboard kappa
- Statistically tied teams: BF value < 5
- Solution has to be shared
- **Interpretable** (model transferable to future EUOS data):

Do not use EOS ID

Model accuracy stable upon shuffling the test set

team_name	public_kappa	private_kappa	bf_value1	bf_value2	bf_value3
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
olab	0.3452	0.3079	Inf	Inf	Inf
Bernhard Rohde	0.1978	0.2175	Inf	Inf	Inf
GP	0.1844	0.1573	Inf	Inf	Inf
1 (akoppchem	0.1473	0.1161	NA	8.794174e+77	1.555826e+94
Emil Nichita	0.1063	0.1156	1.186872e-01	1.900861e+84	8.021698e+102
David Huang	0.1202	0.1154	5.221648e+00	2.738914e+72	7.604751e+95
2 (Beardy Polonium	0.1372	0.1053	8.794174e+77	NA	1.574526e+01
ZR	0.1377	0.1051	1.509180e+82	3.568496e-02	1.851443e+01
Mr Maniac	0.1404	0.1021	1.555826e+94	1.574526e+01	NA
3 (Jack Dawe	0.0970	0.1015	1.214301e+129	5.908099e+07	1.691333e+00

Confirmed winning team

Top1. a.koppchem (Led by Prof. Igor Tetko)

Top2. Beardy Polonium

Top3. Mr Maniac

Upcoming session in SLAS EUROPE 2023



The banner features a dark blue background with a fine, light-colored grid pattern. On the left, a light blue vertical bar contains the text 'slas europe 2023' in white, with '2023' in a larger font. To the right of this bar, the text 'CONFERENCE & EXHIBITION' is written in white. In the center, the phrase 'SCIENCE SET IN MOTION' is displayed in large, white, sans-serif capital letters. To the right of the text, there is a graphic of four stylized human figures in silhouette, each holding a large, colorful circular object (yellow, red, green, and blue) that resembles a molecular or scientific structure. The date '22-26 MAY' is written in large, bold, yellow and blue letters, with 'Brussels, Belgium' in white below it. At the bottom left, the website 'SLAS.ORG/EUROPE2023' and the hashtag '#SLASEUROPE2023' are written in white. At the bottom right, the SLAS logo is shown, consisting of a white triangle and the word 'slas' in white lowercase letters. The word 'BRUSSELS' is written vertically in large, light blue letters on the right side of the banner.

slas
europe
2023 CONFERENCE & EXHIBITION

SCIENCE
SET IN
MOTION

22-26 MAY
Brussels, Belgium

SLAS.ORG/EUROPE2023
#SLASEUROPE2023

slas