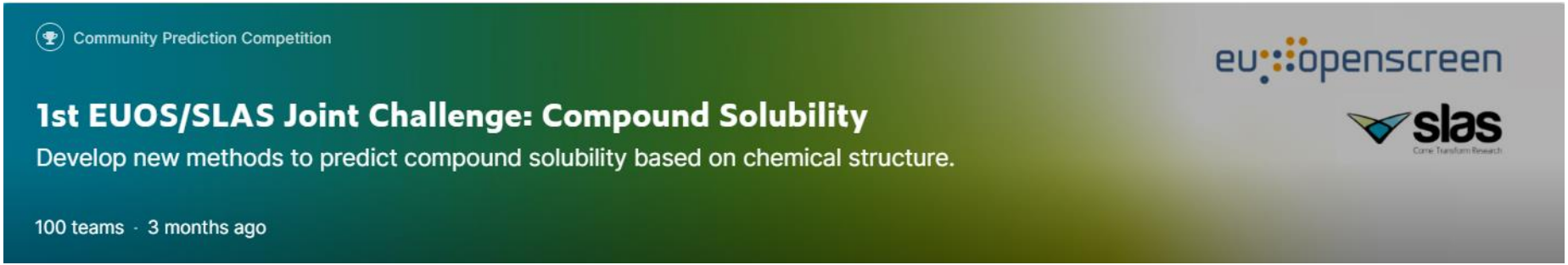


Challenge organizers:

Robert Harmel, Andrea Zaliani, **Wenyu Wang#,** Julio Martin, Christian Parker, Jing Tang#

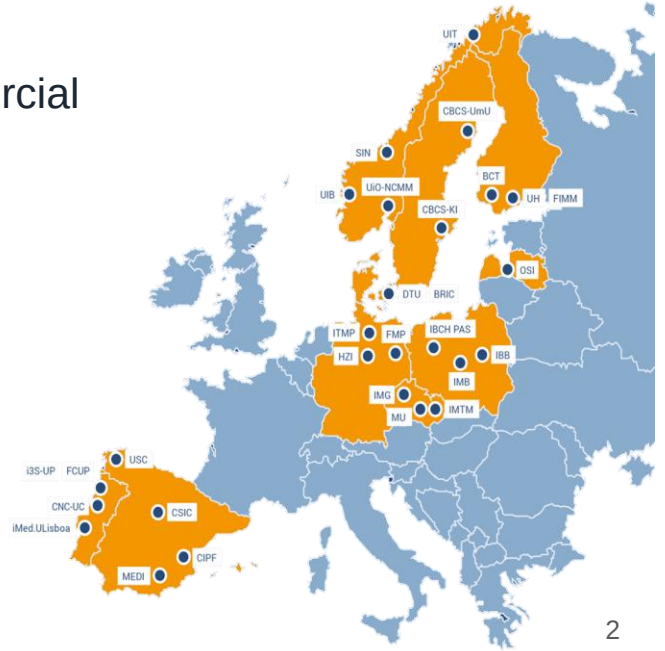
#Network Pharmacology for Precision Medicine (**NetPhar**), Faculty of Medicine, University of Helsinki

Date: 2023/03/23

EU-OPENSCREEN

- 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-OPENSCREEN laboratory
Cell painting	Ongoing assay validation at four sites

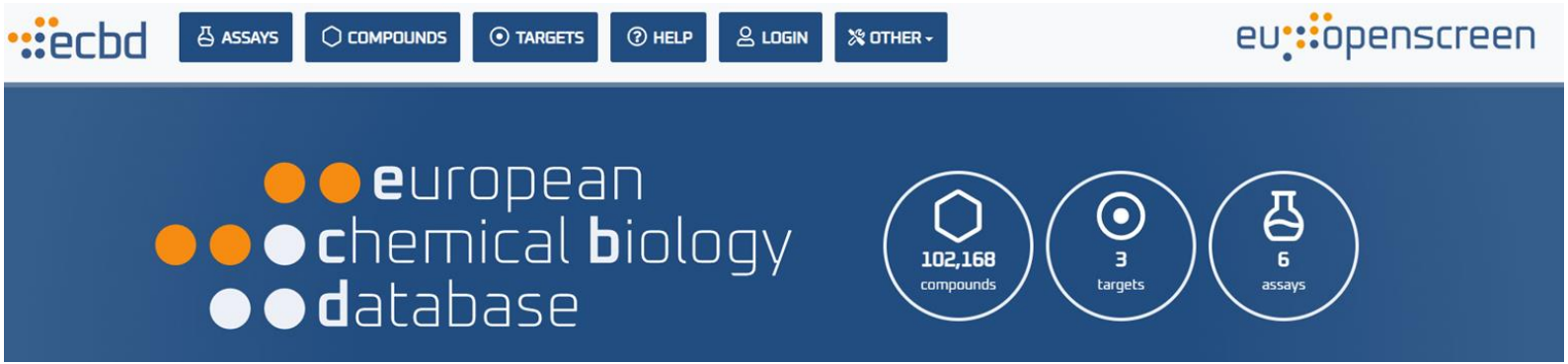


EU-OPENSCREEN

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

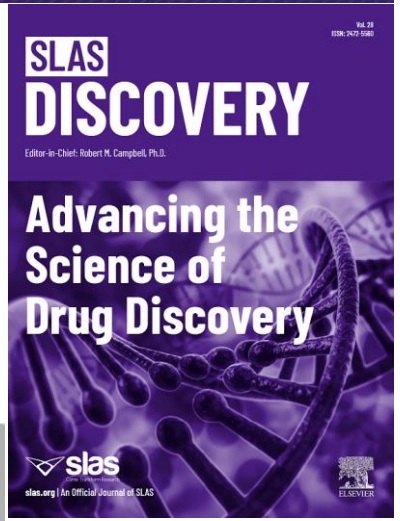
- FAIR data principles
- Optional embargo period up to 36 months
- Annotations + links to other databases (e.g. ChEMBL)

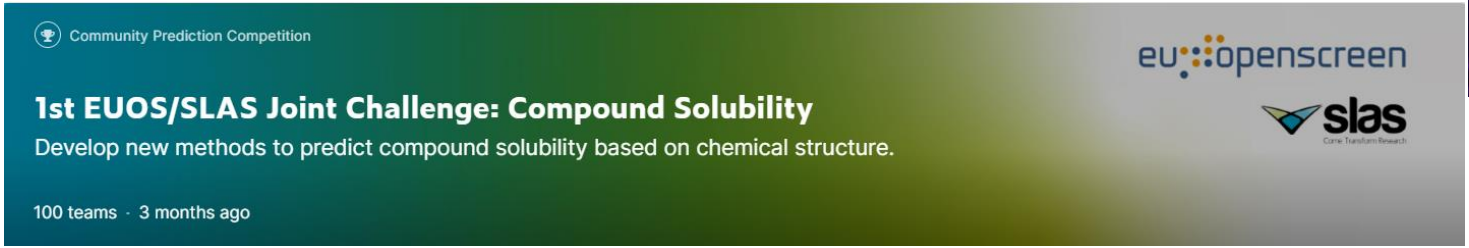
https://ecbd.eu/



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 **S**ociety of **L**ab **A**utomation & **S**creening
- Solubility: an essential feature of all biologically active compounds

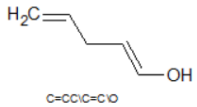


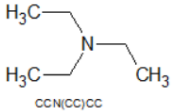


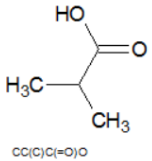
Challenge data

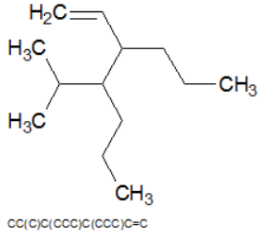
- Experimentally measured aqueous solubility of 100k small molecules
- 70K Training / 15K Public leaderboard / 15K Final evaluation
- Structure based predictors generated by participants (SMILE / InChIKey)
- Labels are solubility classes:

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









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

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

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

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

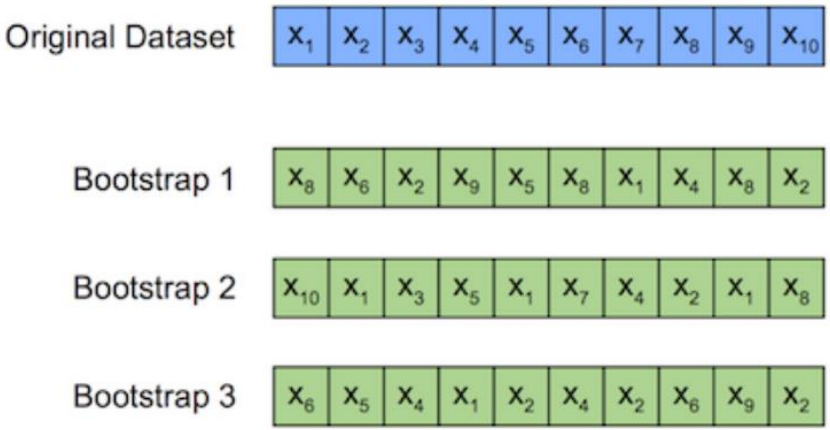
Kaggle default evaluation

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

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



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
	QR	0.1844	0.1573	Inf	Inf	Inf
1	akoppchem	0.1473	0.1161	NA	8.794174e+77	1.555826e+94
1	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
2	ZR	0.1377	0.1051	1.509180e+82	3.568496e-02	1.851443e+01
3	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