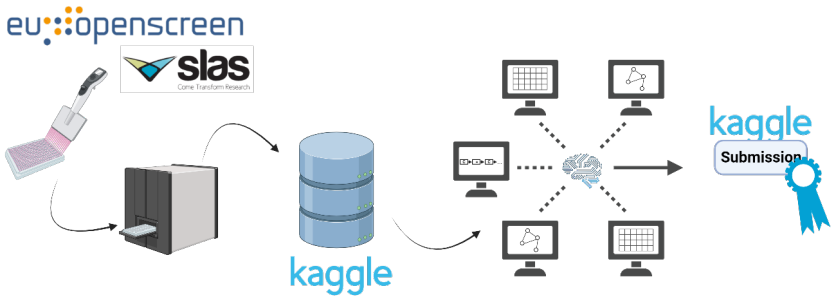
MOLECULAR TARGETS & THERAPEUTICS CENTER

OCHEM consensus model wins Kaggle solubility challenge

Andrea Kopp,[1] Peter Hartog,[1] Martin Šícho,[2,3] Guillaume Godin,[4] and Igor V. Tetko,[1,5,*]

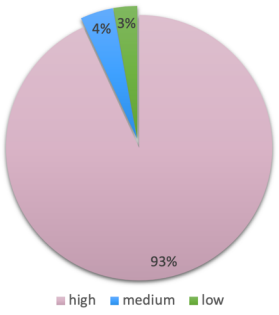
[1]Institute of Structural Biology, Molecular Targets and Therapeutics Center, Helmholtz Munich-Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), DE-85764 Neuherberg, Germany; [2]Leiden Academic Centre for Drug Research, Leiden University, 55 Einsteinweg, 2333 CC Leiden, The Netherlands; [3]CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technická 5, 166 28, Prague, Czech Republic; [4]DSM-Firmenich SA, Rue de la Bergère 7, CH-1242 Satigny, Switzerland; [5]BIGCHEM GmbH, Valerystr. 59, DE-85716 Unterschleißheim, Germany

We outline our success in the EU-OPENSCREEN - a not-for-profit European Research Infrastructure Consortium (ERIC) - and The Society for Laboratory Automation and Screening (SLAS) solubility challenge. The challenge was established to identify the state-of-the-art computational methods for reliable predictions of threshold solubility of compounds. Here, we present our consensus model which was the winning solution amid 100 contributing teams.



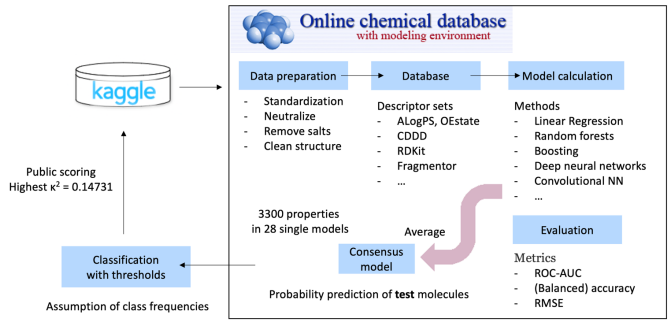
https://ochem.eu

Imbalance of data



Challenge setup: classification of highly imbalanced data into three ordered classes

Workflow with OCHEM



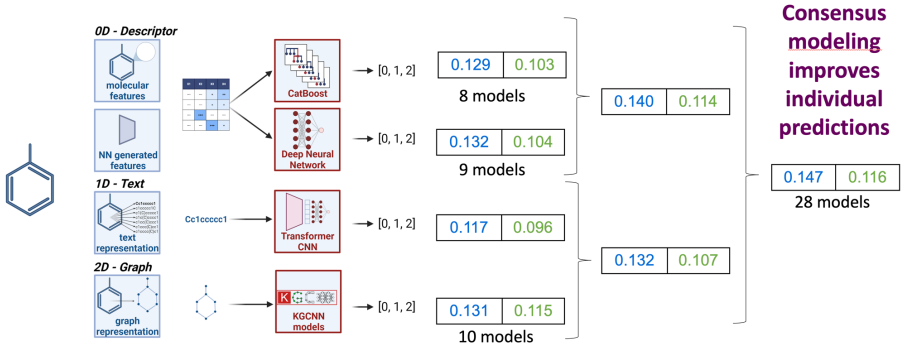
Workflow used for model development

Metrics AUC for Training set Validation: Cross-Validation (84 models)

	LSSVMG	ASNN	PLS	KNN
ALogPS, OEstate	0.74	0.68	0.61	0.64
CDDD	0.8	0.74	0.75	0.71
CDK2 (cons,topol,geom,elec,hybrid) 3D:corina	0.75	0.71	0.56	0.71
ChemaxonDescriptors (pH 0 - 14:1) 3D:corina	0.76	0.7	0.59	0.68
Dragon6 (2D blocks: 1 28)	0.64	0.66	0.59	0.65
Dragon6 (3D blocks: 1-29) 3D:corina	0.76	0.72	0.57	0.65
Fragmentor (length:2 - 4)	0.72	0.7	0.59	0.63
GSFrag (F + L)	0.69	0.69	0.61	0.61
InductiveDescriptors 3D:corina	0.69	0.71	0.57	0.67
JPlogP	0.73	0.74	0.59	0.67
MAP4	0.71	0.65	0.59	0.67
MORDRED (All) 3D:corina	0.77	0.73	0.57	0.68
Mera, Mersy 3D:corina	0.73	0.69	0.55	0.67
OEstate	0.74	0.67	0.63	0.68
PyDescriptor 3D:corina	0.71	0.71	0.7	0.67
QNPR (length:1 - 3)	0.68	0.62	0.58	0.58
RDKIT (3D blocks: 1-11 15-16) 3D:corina	0.77	0.72	0.56	0.65
SIRMS (labels:charge+logp+hb+refractivity)	0.76	0.73	0.59	0.67
Spectrophores (accuracy=20) 3D:corina	0.68	0.6	0.52	0.6
StructuralAlerts	0.67	0.64	0.58	0.51
alvaDesc (3D blocks: (only) 1-30) 3D:corina	0.75	0.71	0.57	0.68

Example of models developed using OCHEM

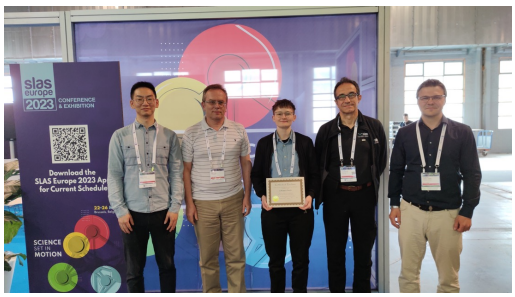
Quadratic kappa metric scores → Public leaderboard | Private leaderboard



Overview of molecular representations and models. Molecules were represented by descriptors, SMILES text, and graph representations. Subsequently, descriptor-based models including CatBoost and DNNs, TransformerCNN and KGCNN models were generated and tested. Combinations of models improved performance.

Conclusions:

- Concensus modellling provided the best accuracy
- Different representations enhance the performance
- Reliable protocol is important to get best results
- Do not give up!



Challenge winners (Ms. A. Kopp is in the centre) with the challenge organizers during 2023 SLAS conference in Brussels

See pre-print at https://doi.org/10.26434/chemrxiv-2023-p8qcv

We acknowledge the OPENSCREEN consortium and thank them for organizing the 1st EUOS/SLAS Joint Kaggle Challenge: Compound Solubility. This study was partially funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions, grant agreement "Advanced machine learning for Innovative Drug Discovery (AIDD)" No. 956832.

HELMHOLTZ MUNICH