Winning model: OCHEM-generated consensus model

Andrea Kopp

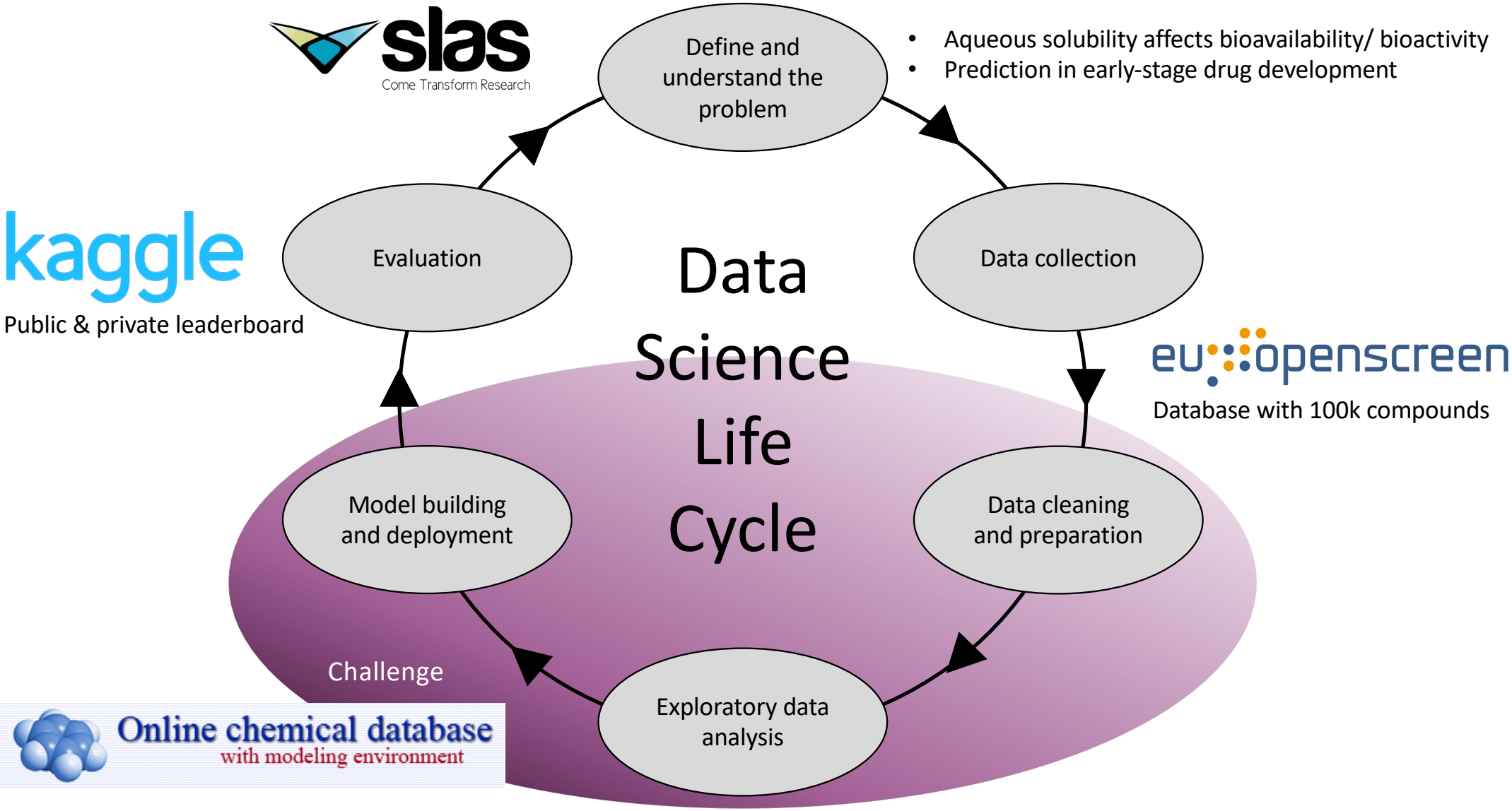
SLAS Europe 2023

25.05.2023

HELMHOLTZ MUNICH

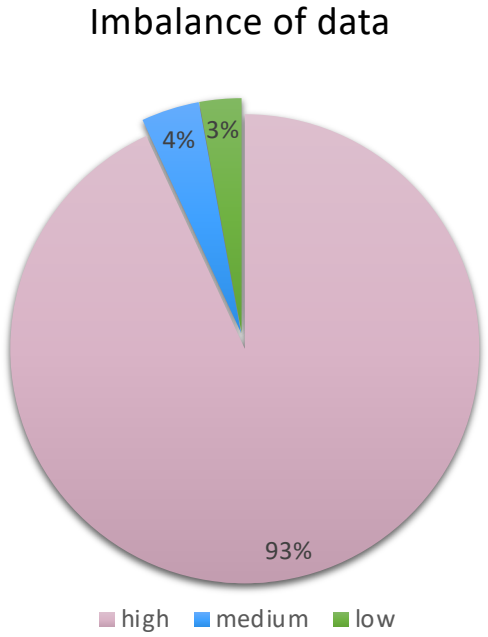
Team of Igor Tetko with Peter Hartog, Martin Šícho and Guillaume Godin

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN


slas
Come Transform Research
Define and understand the problem
Data collection
Data cleaning and preparation
Exploratory data analysis
Model building and deployment
Evaluation
Data Science Life Cycle
Challenge
• Aqueous solubility affects bioavailability/ bioactivity
• Prediction in early-stage drug development
kaggle
Public & private leaderboard
eu:openscreen
Database with 100k compounds
Online chemical database
with modeling environment

Challenge set-up

- Experimentally: Nephelometer measures undissolved sediment
- Classification into *low*, *medium* and *high* soluble with phenytoin and amiodarone as thresholds
- 70k training datapoints, 15k public leaderboard, 15k private leaderboard
- Stratified random sampling



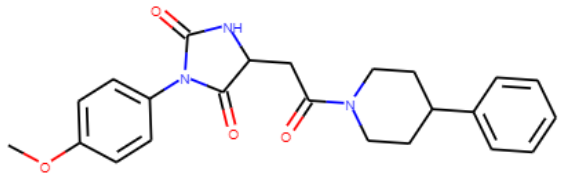
Challenge set-up

- Metric for prediction: quadratic kappa metric
  - good for imbalanced and ordered categorical data
  - Perfect alignment with $\kappa^2 = 1$
  - Random alignment with $\kappa^2 = 0$

$$\kappa^2 = 1 - \frac{\sum_{i,j} w_{i,j} * O_{i,j}}{\sum_{i,j} w_{i,j} * E_{i,j}}$$

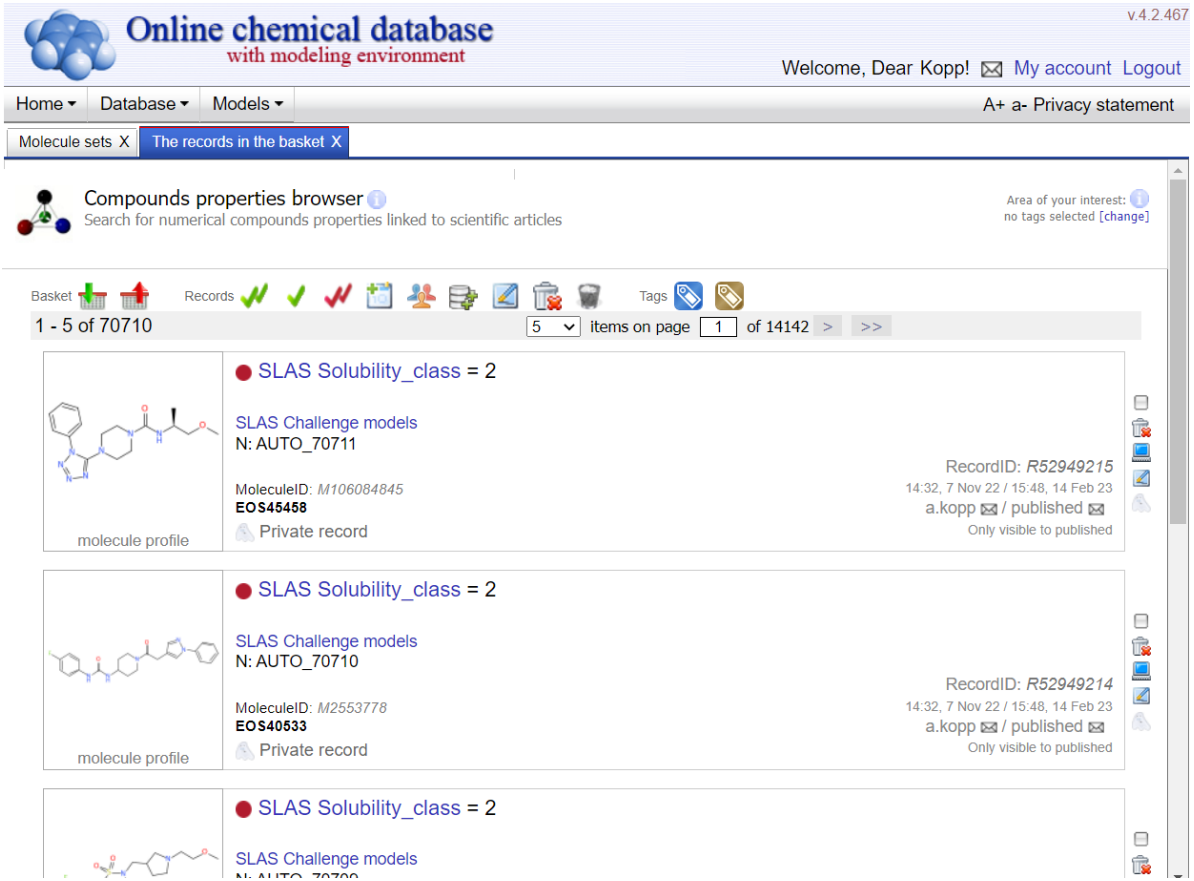
Molecules given as SMILES strings

`COc1ccc(N2C(=O)NC(CC(=O)N3CCC(c4ccccc4)CC3)C2=O)cc1`



→ Challenge: Classification model with molecular features

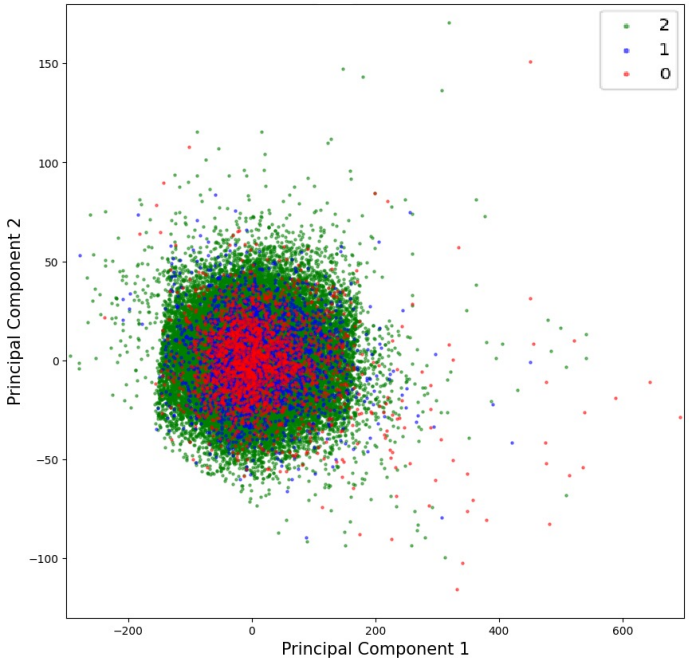
Molecular descriptors



- https://www.ochem.eu
- Free to participate
- User-contributed database with focus on quality and verifiability
- Easy calculated descriptors organised in >20 descriptors blocks
- Download licence-free descriptor packages

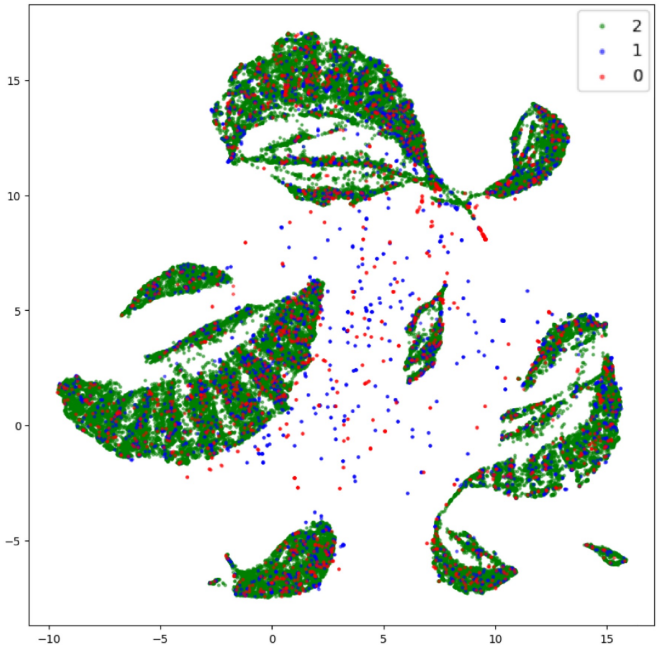
Data exploration

PCA



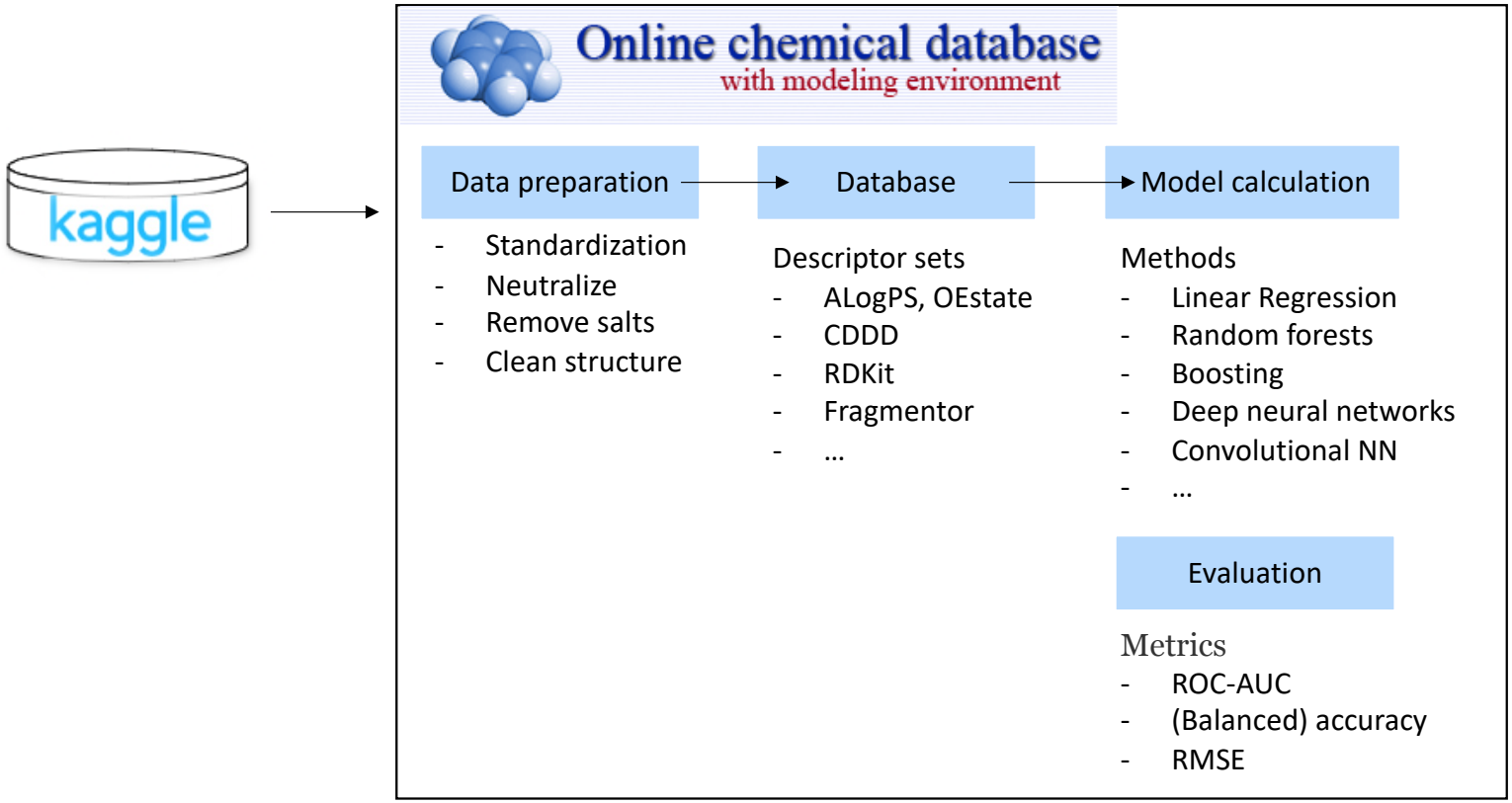
86% explained variance

UMAP



- ALogPS, OEstate descriptor set
- Classification with *k*-NN
- Optimisation of workflow
- Highest κ^2 = 0.00832

Workflow with OCHEM



OCHEM for modeling

- ~20 modern methods
- Estimation of Applicability and Accuracy
- Generate baskets for your modeling needs
- Stratified learning (bagging, cross validation) for imbalanced datasets

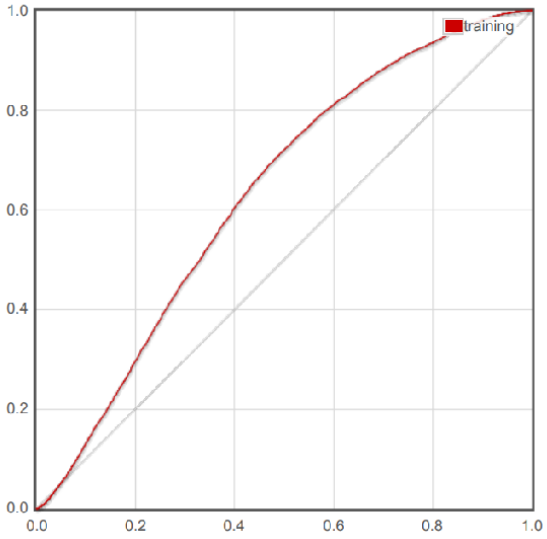
Open OCHEM:
https://github.com/openochem

- Open-source OCHEM
- Local and private modeling

Predicted property: **Solubility_class (qualitative)** modeled in CLASS
Training method: RFR

Data Set	#	Accuracy	Balanced Accuracy	AUC
Training set: train_set_solubility_challenge (discretized)_low	62445 records	64.5% ± 0.2	60.8% ± 0.3	0.632 ± 0.004
Excluded from training set	8265 records	17.6% ± 0.4	8.8% ± 0.2	NaN

ROC curve



[ALogPS, OEstate]
Correl. limit: 0.95 Variance threshold: 0.01
Maximum value: 999999,
trees=512
5-fold cross-validation (stratified)
-
260 pre-filtered descriptors
trees=512

Calculated in 3638 seconds
Size: 36813 Kb

Real↓/Predicted→	high	low+medium	Hit rate
high	37500	20069	0.651
low+medium	2124	2752	0.564
Precision	0.946	0.121	
Training (Original)			

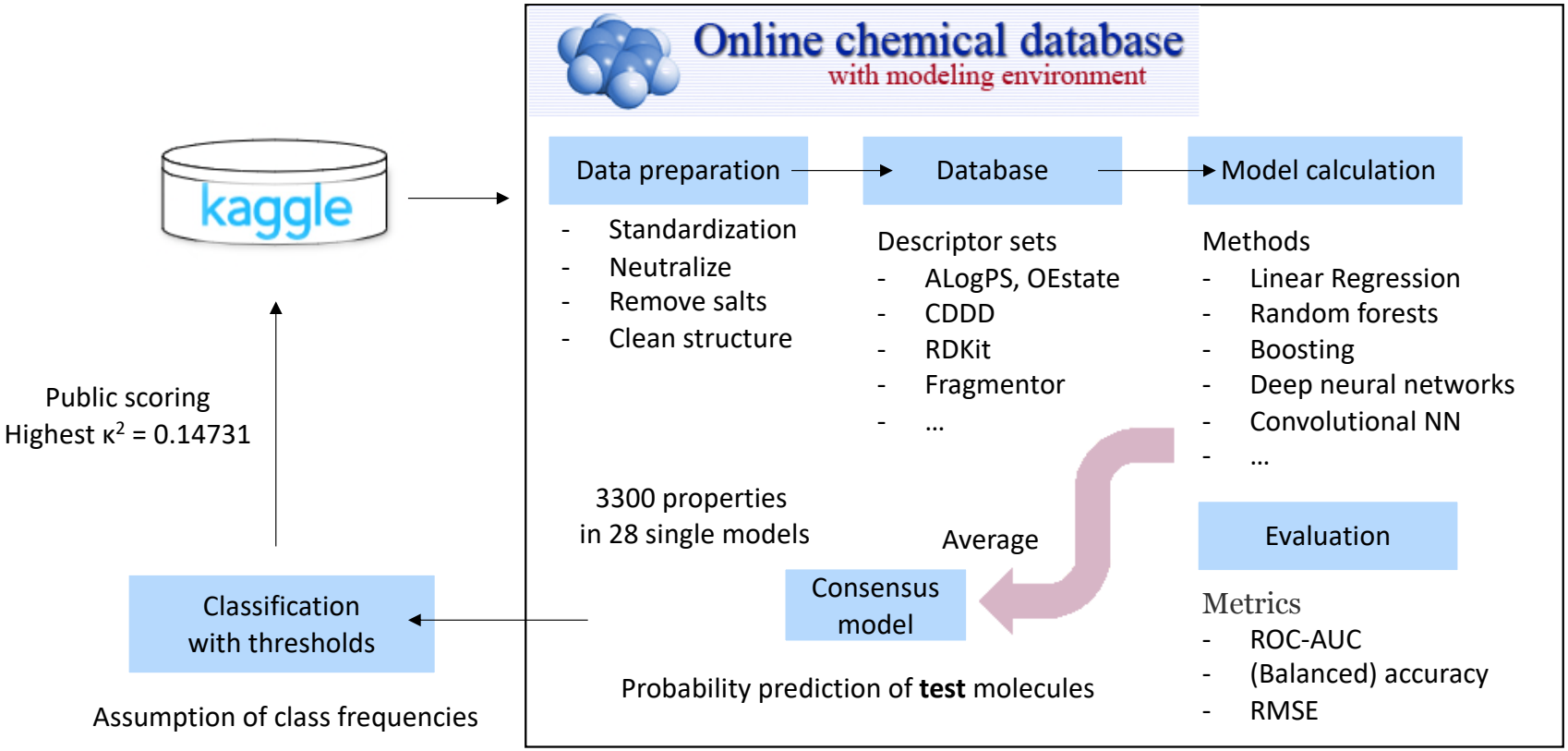
OCHEM for modeling

- Graphical interface allows comprehensive modeling without explicit coding
- Implementation for GPU and CPU use
- Consensus models:
  - Average over multiple models to improve prediction
  - Orthogonal models
  - Various descriptor sets/ molecular representations

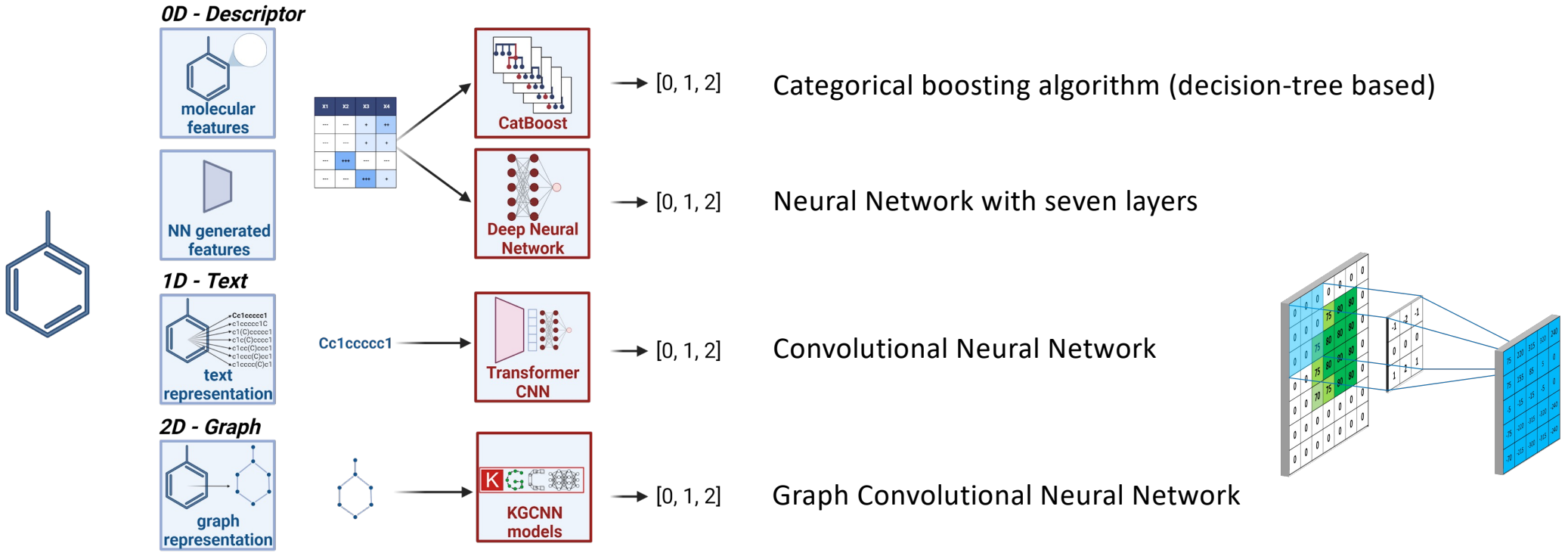
Metrics AUC for Training set Validation: Cross-Validation (84 models)

	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

Workflow with OCHEM

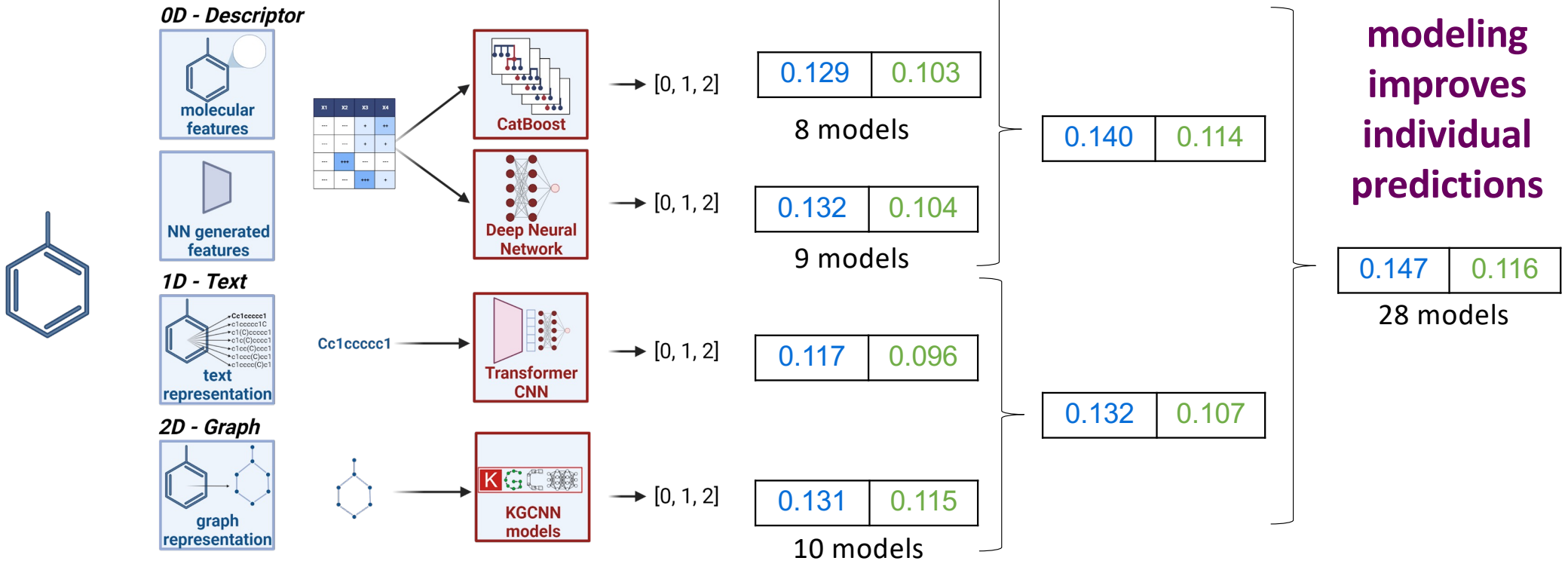


Molecular representation



Quadratic kappa metric scores →

Public leaderboard	Private leaderboard



Outcome

Solubility data

- Challenging task in general to predict solubility
- Very challenging dataset with an overall highest quadratic kappa score of 0.35 (0.6 is considered good)
- Systematic lower private score
- Three-classes and no numerical prediction

→ Limited use for experimental scientists

Private leaderboard

#	△	Team	Members	Score	Entries
1	—	olab		0.30785	35
2	—	Bernhard Rohde		0.21748	16
3	—	QP		0.15731	9
4	—	**a.kopp.chem**		0.11614	18
5	▲ 24	Emil Nichita		0.11562	1

→ Publication in SLAS Discovery coming soon

Acknowledgements

Igor Tetko
Peter Hartog
Martin Šícho (Uni Leiden)
Guillaume Godin (Firmenich)

Paula Torren-Peraire
Varvara Voinarovska
Marchela Pandelova
Nesma Mousa
Mark Embrechts
Michael Sattler

Eyke Hüllermeier (LMU)

HELMHOLTZ MUNICH

BigChem

CONCERTREACH
CONCERTING EXPERIMENTAL DATA AND IN SILICO MODELS FOR REACH
LIFE17 GIE/IT/000461

ERA CVD

eco
environmental cheminformatics
marie curie initial training network

Alexander von Humboldt Stiftung/Foundation

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

https://ai-dd.eu

AIDD – Advanced machine learning for Innovative Drug Discovery

Machine learning is changing our society, as exemplified by speech and image recognition applications. Also, the life sciences change rapidly through the use of artificial intelligence, and it is expected that **fields like drug development can take advantage of machine learning**. The main goal of the AIDD project is to **train and prepare the next generation of scientists** who need to have skills in both machine learning and drug discovery and will, after graduating, be able to contribute **to speed up the drug development process.**

Marie Skłodowska-Curie Innovative Training Network European Industrial Doctorate (2020-today)