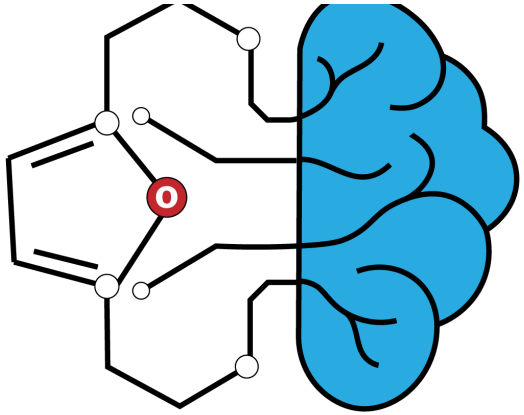


AiChemist

Institute of Structural Biology

Predicting HTS data using modern ML methods: what have we learned from recent EUOS competitions?

Igor V. Tetko
Helmholtz Munich and BigChem GmbH

7 September, CECEC2026, Prague, Czech Republic

HELMHOLTZ MUNICH

https://en.wikipedia.org/wiki/Helmholtz_Zentrum_M%C3%BCnchen

European RI for Chemical Biology and early Drug Discovery

EU-OPENSCREEN provides access to

- Technologies (e.g., screening platforms)
- Resources (e.g., compound collections)
- Expertise (e.g., in medicinal chemistry)
- Data (e.g., bioactivity data)
- Training

Collaborators in academia and industry

EU-OPENSCREEN PARTNER SITES 2024

- 8 member countries and 1 observer country
- 22 Screening partner sites
- 11 Chemistry partner sites
- 3 Chemoproteomics and spatial MS-based omics partner sites
- 1 database host

Screening partner site

Chemistry partner site

Chemoproteomics and spatial MS-based omics partner site

Database host

Member country

Observer country

EU-OPENSCREEN Libraries

European Chemical Biology Library (ECBL)

Diversity library

- 98,560 structurally highly diverse compounds
- Average MW=350 g/mol

Horvath D. *et al.*, ChemMedChem 2014, *9*, 2309

CHEMMEDCHEM FULL PAPERS

ChemPubSoc Europe

DOI: 10.1002/cmdc.201402126

Design of a General-Purpose European Compound Screening Library for EU-OPENSCREEN

Dragos Horvath,[a,b] Michael Lisurek,[b] Bernd Rupp,[b] Ronald Kühne,[b] Edgar Specker,[b] Jens von Kries,[b] Didier Rognan,[c] C. David Andersson,[d] Fredrik Almqvist,[d] Mikael Elofsson,[e] Per-Anders Enqvist,[e] Anna-Lena Gustavsson,[b] Nikita Remez,[f] Jordi Mestres,[f] Gilles Marcou,[g] Alexander Varnek,[g] Marcel Hibert,[g] Jordi Quintana,[h] and Ronald Frank[b]

Pilot library

Representative diversity set (part of ECBL)

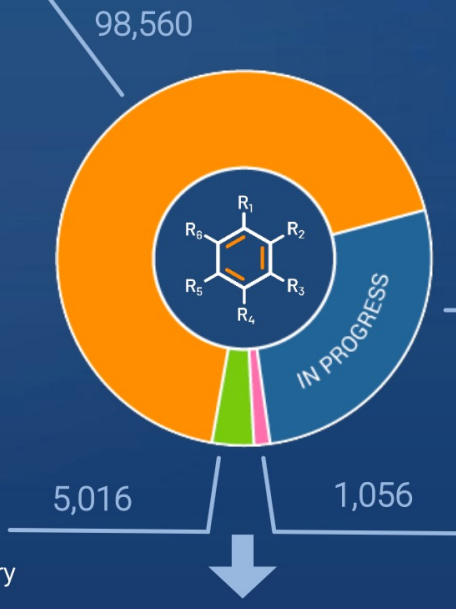
- 2,464 representative compounds of the diversity library

Bioactives

- 2,464 bioactives: active against 1039 different targets, contain 654 approved drugs and 368 highly selective probes

Assa interference set

- 88 assay interference compounds in 4 dilutions



European Chemical Biology Database
https://ecbd.eu/

The European Academic Compound Library (EACL)

Novel donated compounds from chemists worldwide

- ~ 6,000 compounds and growing
- Regulated and confidential access (e.g., MTA)
- IP stays with the chemist
- Embargo period up to 3 years

- Chemogenonics 1000
- TPD 600
- Covalent 500
- RNA binding 8,000
- Diversity 6,000

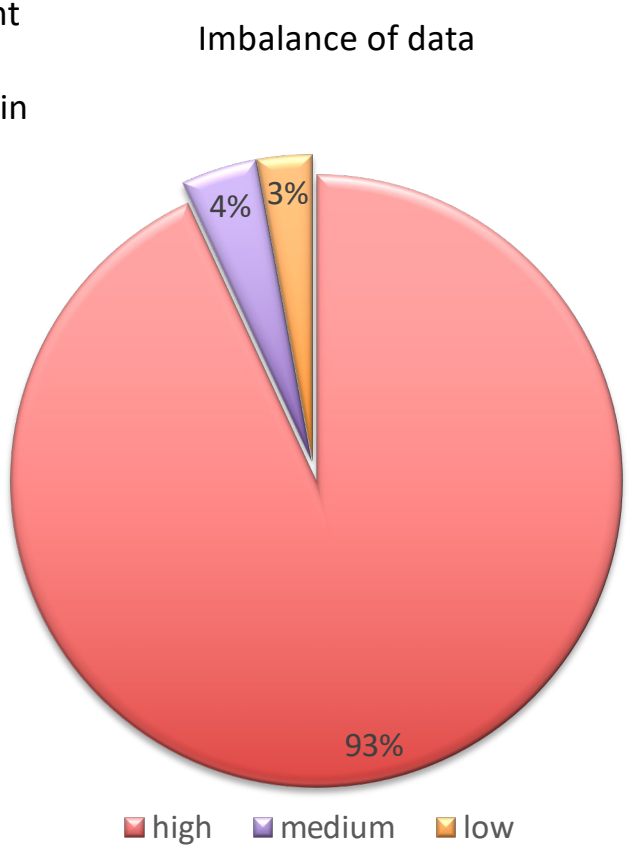
Fragment Library (EFSL)

Set of low MW and ultra-low MW fragments

- Collaboration with INSTRUCT/ iNEXT-Discovery sites
- 968 fragments with HAC > 8 in DMSO-d6 (c = 100 mM)
- 88 "minifrags with HAC < 8 in DMSO-d6 (c = 1M)

First EUOS Solubility 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



Kaggle SLAS Solubility prediction Winning model: OCHEM-generated consensus model

Andrea Hunklinger

SLAS Europe 2023

25.05.2023

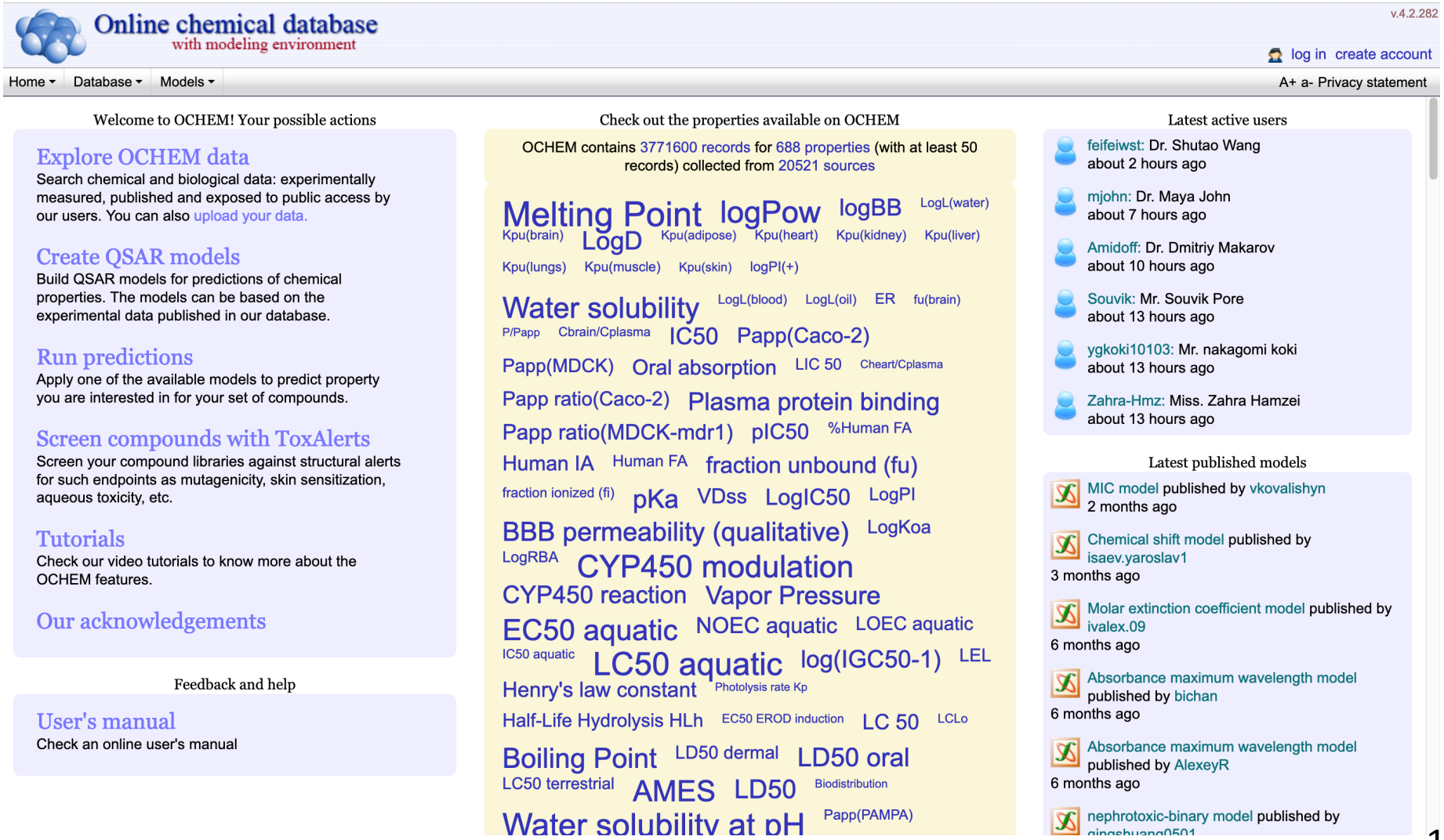
HELMHOLTZ MUNICH

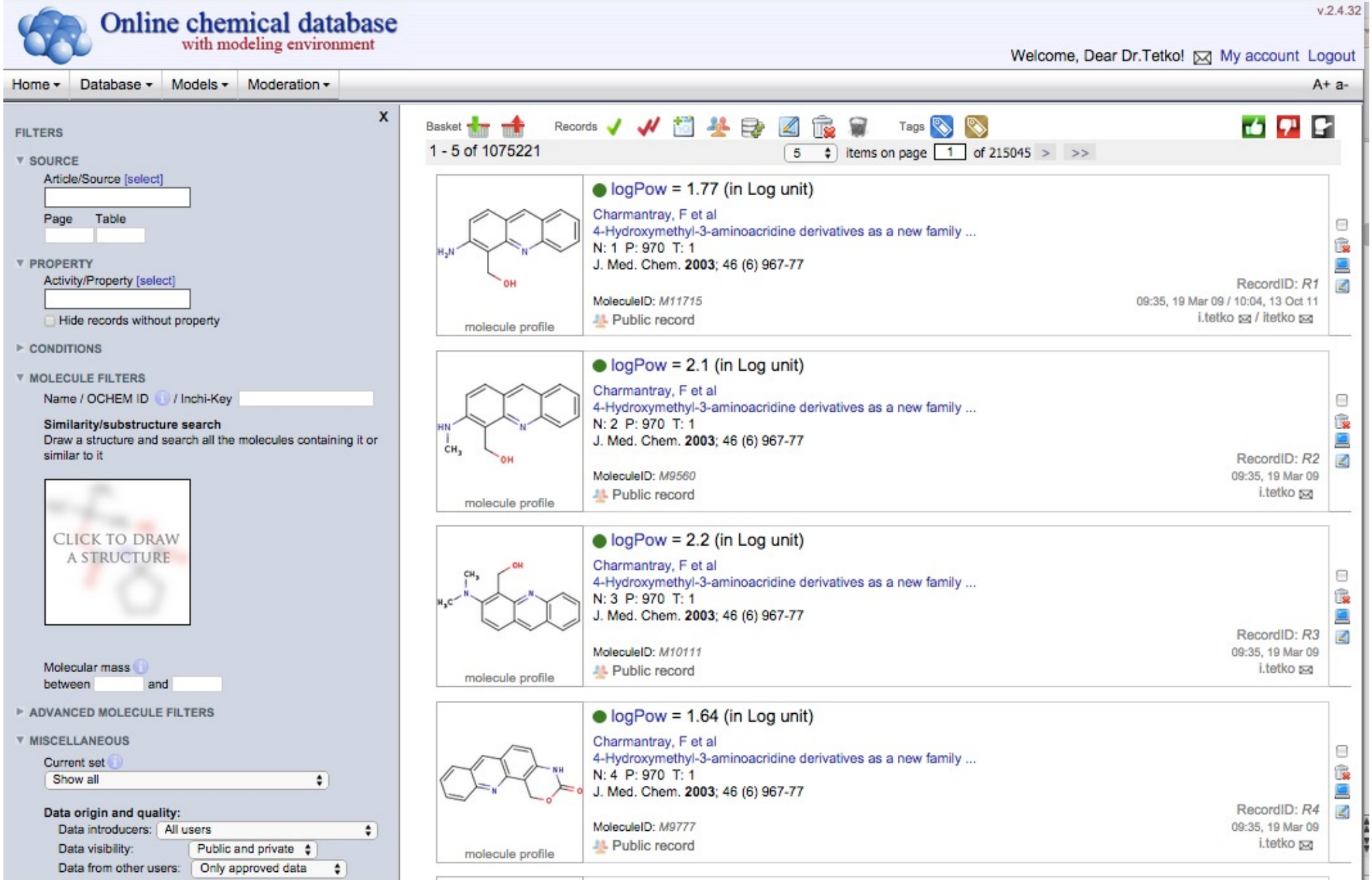
Peter Hartog, Martin Šícho , Guillaume Godin, Igor Tetko

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

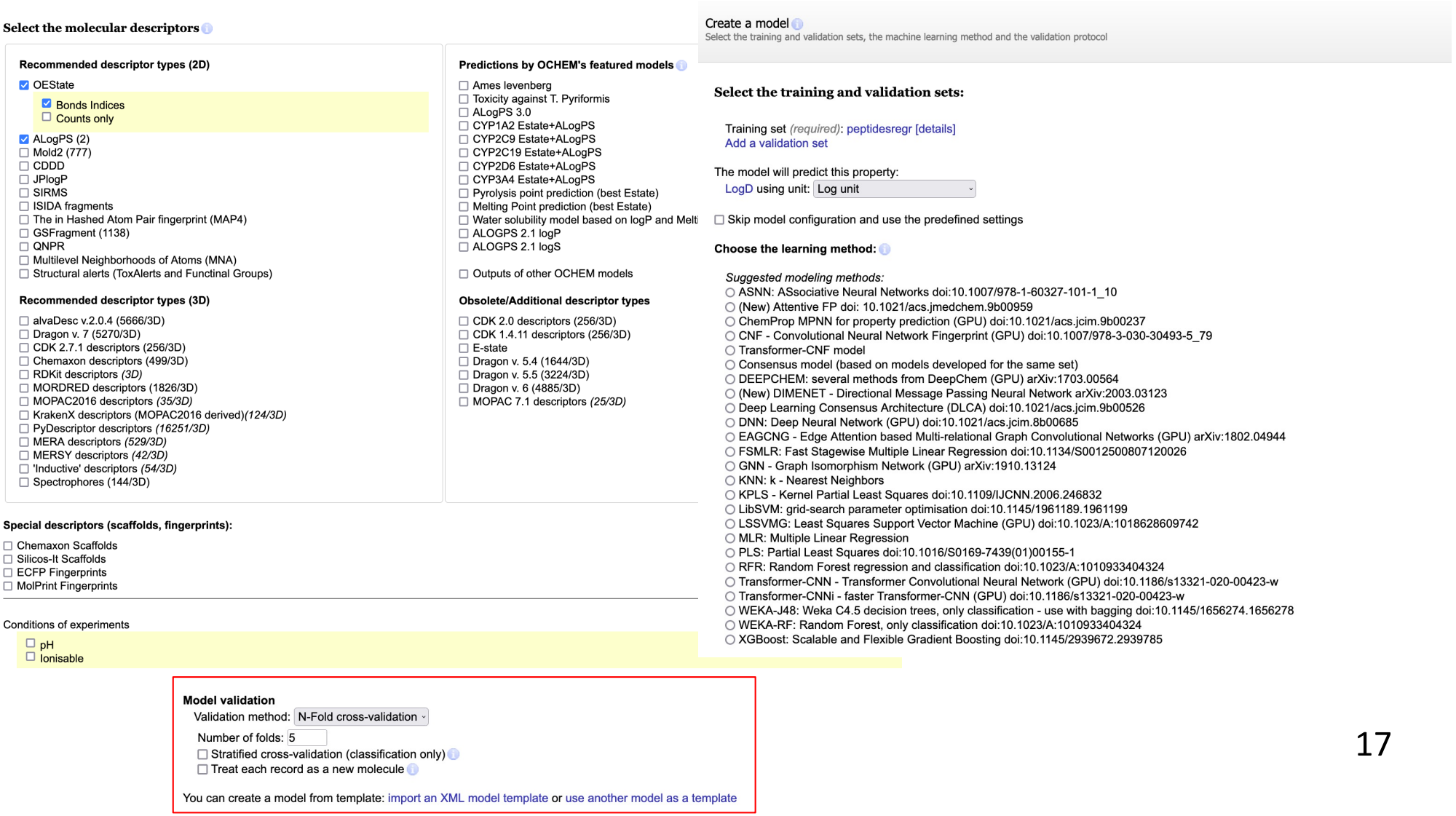
Hunklinger at al, DOI: 10.1016/j.slasd.2024.01.005

OCHEM https://ochem.eu

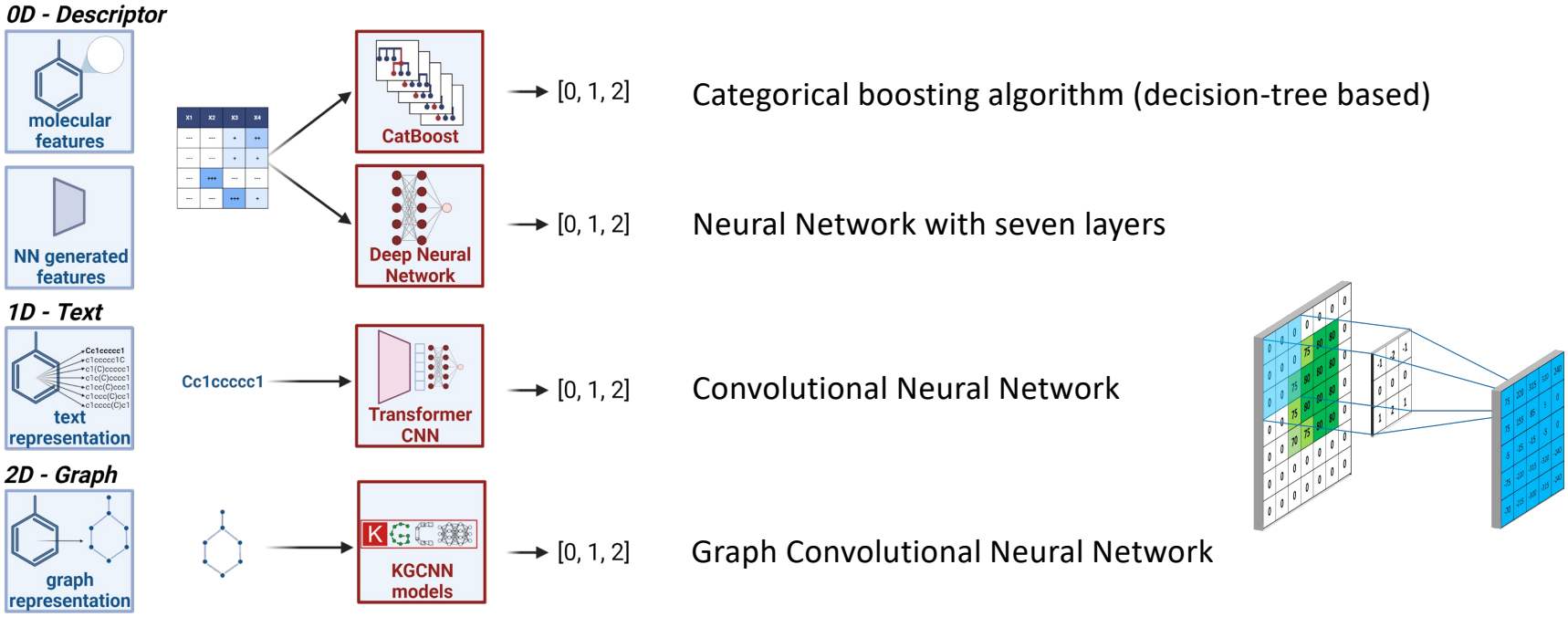



Online chemical database
with modeling environment
v.2.4.32
Welcome, Dear Dr.Tetko! My account Logout
Home Database Models Moderation
A+ a-
FILTERS
SOURCE
Article/Source [select]
Page Table
PROPERTY
Activity/Property [select]
Hide records without property
CONDITIONS
MOLECULE FILTERS
Name / OCHEM ID / Inchi-Key
Similarity/substructure search
Draw a structure and search all the molecules containing it or similar to it
CLICK TO DRAW A STRUCTURE
Molecular mass between and
ADVANCED MOLECULE FILTERS
MISCELLANEOUS
Current set
Show all
Data origin and quality:
Data introducers: All users
Data visibility: Public and private
Data from other users: Only approved data
Basket
Records
Tags
1 - 5 of 1075221
5 items on page 1 of 215045 > >>
molecule profile
logPow = 1.77 (in Log unit)
Charmantray, F et al
4-Hydroxymethyl-3-aminoacridine derivatives as a new family ...
N: 1 P: 970 T: 1
J. Med. Chem. 2003; 46 (6) 967-77
MoleculeID: M11715
Public record
RecordID: R1
09:35, 19 Mar 09 / 10:04, 13 Oct 11
i.tetko / itetko
molecule profile
logPow = 2.1 (in Log unit)
Charmantray, F et al
4-Hydroxymethyl-3-aminoacridine derivatives as a new family ...
N: 2 P: 970 T: 1
J. Med. Chem. 2003; 46 (6) 967-77
MoleculeID: M9560
Public record
RecordID: R2
09:35, 19 Mar 09
i.tetko
molecule profile
logPow = 2.2 (in Log unit)
Charmantray, F et al
4-Hydroxymethyl-3-aminoacridine derivatives as a new family ...
N: 3 P: 970 T: 1
J. Med. Chem. 2003; 46 (6) 967-77
MoleculeID: M10111
Public record
RecordID: R3
09:35, 19 Mar 09
i.tetko
molecule profile
logPow = 1.64 (in Log unit)
Charmantray, F et al
4-Hydroxymethyl-3-aminoacridine derivatives as a new family ...
N: 4 P: 970 T: 1
J. Med. Chem. 2003; 46 (6) 967-77
MoleculeID: M9777
Public record
RecordID: R4
09:35, 19 Mar 09
i.tetko

QSPR/QSAR modelling in OCHEM



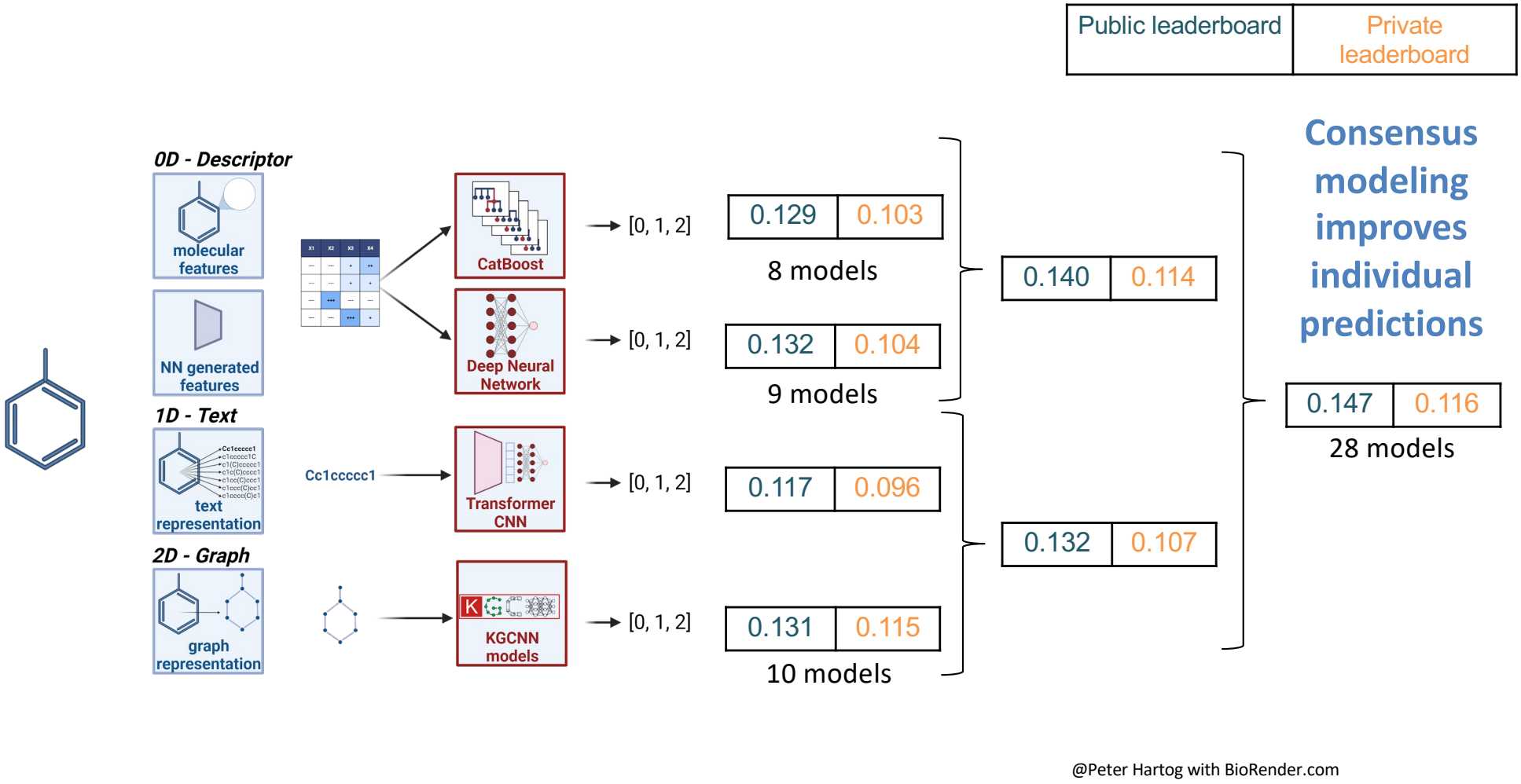
Molecular representation



@Peter Hartog with BioRender.com

Hunklinger at al, DOI: 10.1016/j.slasd.2024.01.005

Quadratic kappa metric scores



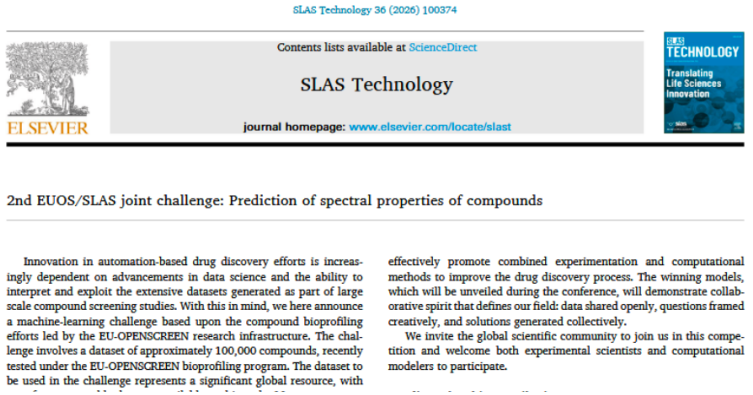
Hunklinger at al, DOI: 10.1016/j.slasd.2024.01.005

Each descriptor/method sees only part of a molecule

Blind monks examining an elephant, an ukiyo-e print by Hanabusa Itchō (1652–1724).
https://en.wikipedia.org/wiki/Blind_men_and_an_elephant

2ND EU-OPENSCREEN / SLAS MACHINE LEARNING CHALLENGE

- Focus: predicting optical properties of small molecules
- Based on ~100,000 compounds from EU-OPENSCREEN bioprofiling studies
- Builds on success of the 1st solubility prediction challenge
- Impact:
  - Supports assay design & compound triaging
  - Promotes open data and collaborative ML in drug discovery



SLAS Technology 36 (2026) 100374

Contents lists available at ScienceDirect

SLAS Technology

journal homepage: www.elsevier.com/locate/slast

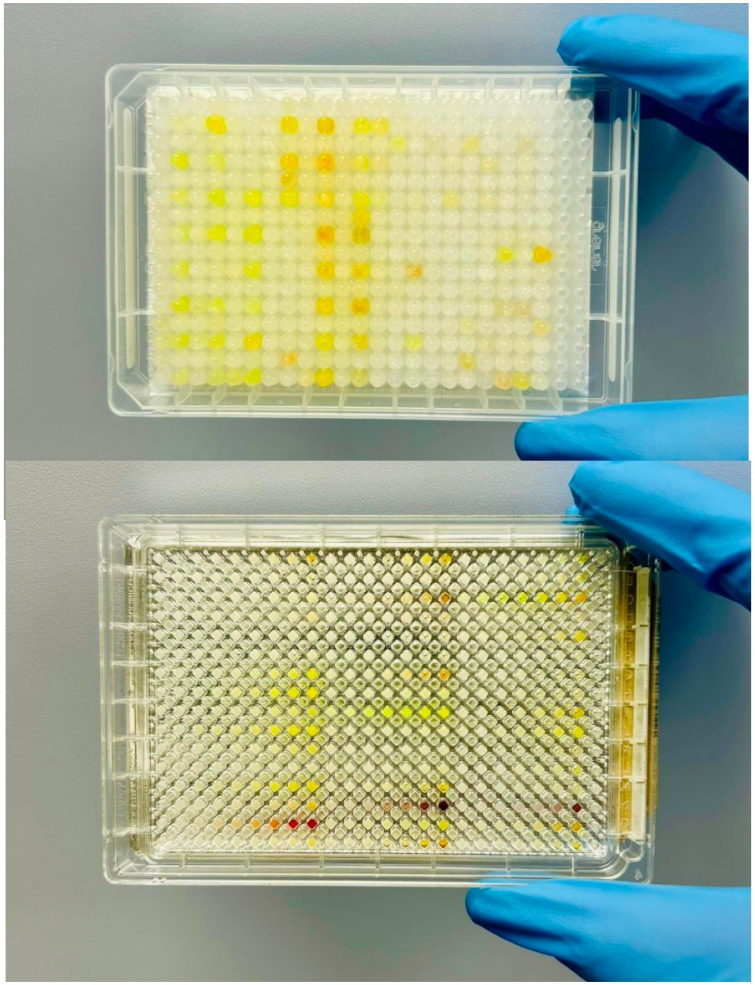
2nd EUOS/SLAS joint challenge: Prediction of spectral properties of compounds

Innovation in automation-based drug discovery efforts is increasingly dependent on advancements in data science and the ability to interpret and exploit the extensive datasets generated as part of large scale compound screening studies. With this in mind, we here announce a machine-learning challenge based upon the compound bioprofiling efforts led by the EU-OPENSCREEN research infrastructure. The challenge involves a dataset of approximately 100,000 compounds, recently tested under the EU-OPENSCREEN bioprofiling program. The dataset to be used in the challenge represents a significant global resource, with effectively promote combined experimentation and computational methods to improve the drug discovery process. The winning models, which will be unveiled during the conference, will demonstrate collaborative spirit that defines our field: data shared openly, questions framed creatively, and solutions generated collectively.

We invite the global scientific community to join us in this competition and welcome both experimental scientists and computational modelers to participate.

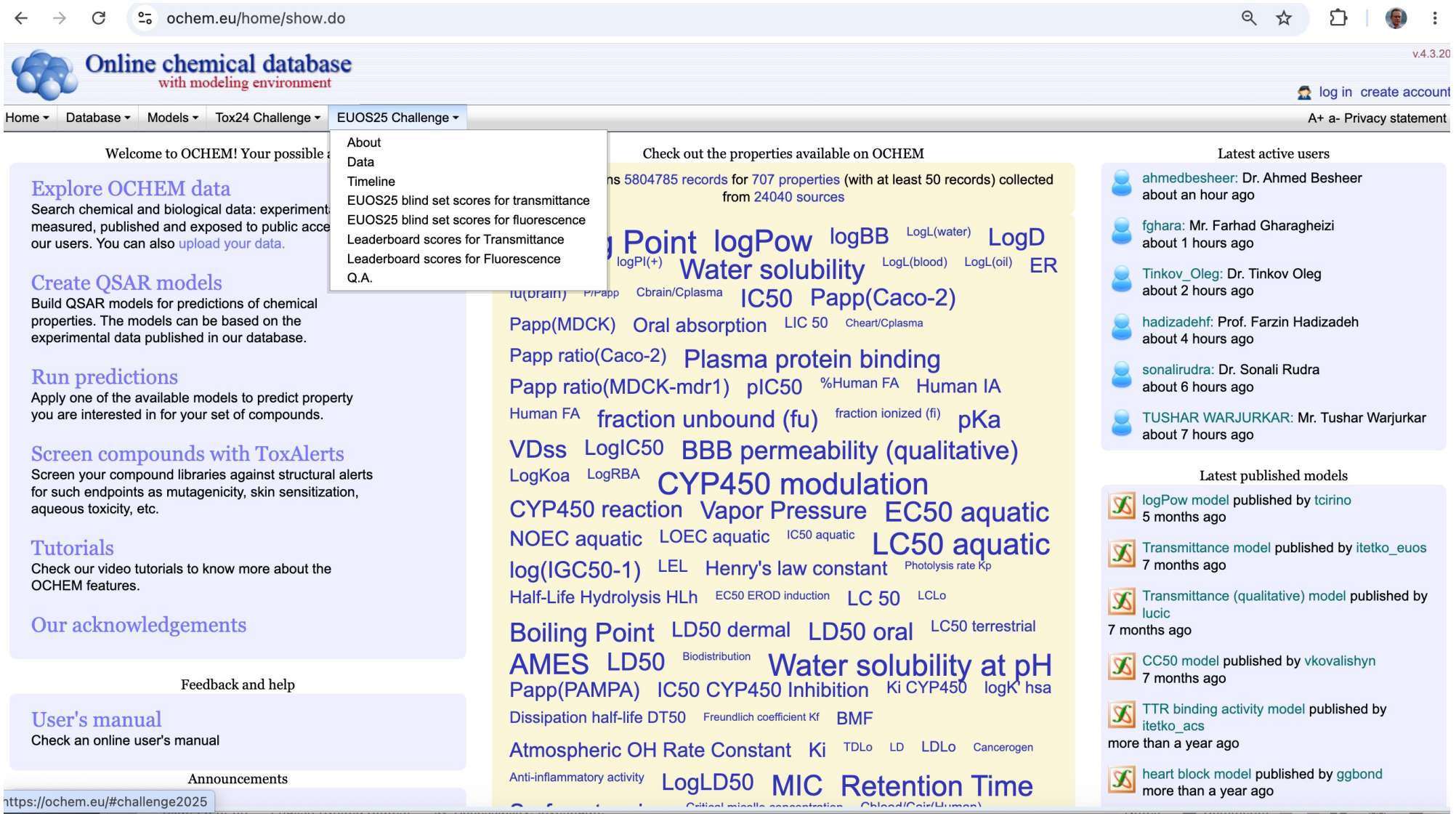
CHALLENGE OBJECTIVES & DATASET

- Objectives:
  - Predict optical transmittance (<70%) in UV and visible ranges
  - Predict compound fluorescence at assay-relevant wavelengths
- Dataset:
  - ~100k chemically diverse compounds
  - Measurements: absorbance, transmittance, fluorescence
  - Training (70%), leaderboard (15%), blind test (15%)



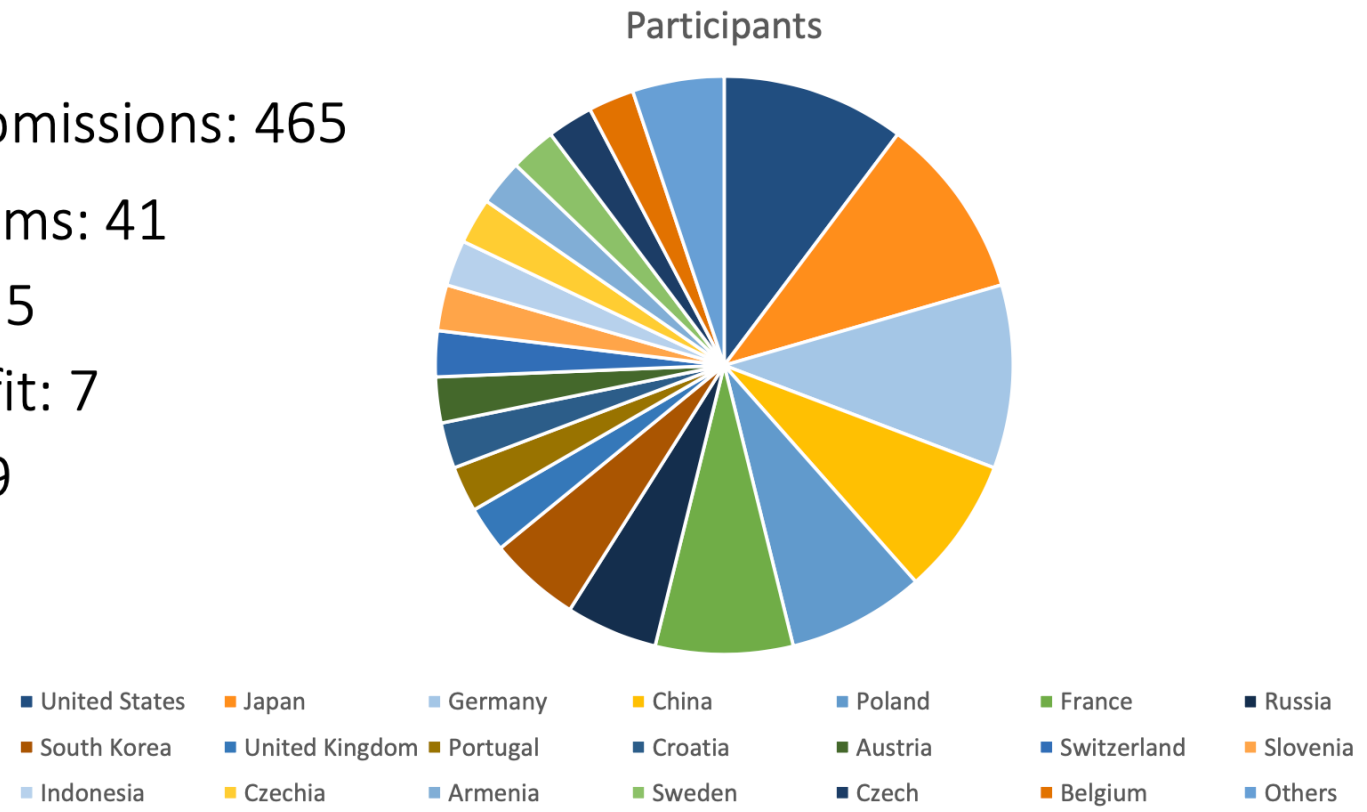
eu-openscreen

Highly imbalanced: 0.2% active for fluorescence at > 450 nm


ochem.eu/home/show.do
Online chemical database
with modeling environment
v.4.3.20
log in create account
Home
Database
Models
Tox24 Challenge
EUOS25 Challenge
About
Data
Timeline
EUOS25 blind set scores for transmittance
EUOS25 blind set scores for fluorescence
Leaderboard scores for Transmittance
Leaderboard scores for Fluorescence
Q.A.
A+ a- Privacy statement
Welcome to OCHEM! Your possible
Explore OCHEM data
Search chemical and biological data: experiment
measured, published and exposed to public acce
our users. You can also upload your data.
Create QSAR models
Build QSAR models for predictions of chemical properties. The models can be based on the experimental data published in our database.
Run predictions
Apply one of the available models to predict property you are interested in for your set of compounds.
Screen compounds with ToxAlerts
Screen your compound libraries against structural alerts for such endpoints as mutagenicity, skin sensitization, aqueous toxicity, etc.
Tutorials
Check our video tutorials to know more about the OCHEM features.
Our acknowledgements
Feedback and help
User's manual
Check an online user's manual
Announcements
https://ochem.eu/#challenge2025
Check out the properties available on OCHEM
ns 5804785 records for 707 properties (with at least 50 records) collected from 24040 sources
Point
logPow
logBB
LogL(water)
LogD
logPI(+)
Water solubility
LogL(blood)
LogL(oil)
ER
fu(brain)
P/Papp
Cbrain/Cplasma
IC50
Papp(Caco-2)
Papp(MDCK)
Oral absorption
LIC 50
Cheart/Cplasma
Papp ratio(Caco-2)
Plasma protein binding
Papp ratio(MDCK-mdr1)
pIC50
%Human FA
Human IA
Human FA
fraction unbound (fu)
fraction ionized (fi)
pKa
VDss
LogIC50
BBB permeability (qualitative)
LogKoa
LogRBA
CYP450 modulation
CYP450 reaction
Vapor Pressure
EC50 aquatic
NOEC aquatic
LOEC aquatic
IC50 aquatic
LC50 aquatic
log(IGC50-1)
LEL
Henry's law constant
Photolysis rate Kp
Half-Life Hydrolysis HLh
EC50 EROD induction
LC 50
LCLo
Boiling Point
LD50 dermal
LD50 oral
LC50 terrestrial
AMES
LD50
Biodistribution
Water solubility at pH
Papp(PAMPA)
IC50 CYP450 Inhibition
Ki CYP450
logK hsa
Dissipation half-life DT50
Freundlich coefficient Kf
BMF
Atmospheric OH Rate Constant
Ki
TDLo
LD
LDLo
Cancerogen
Anti-inflammatory activity
LogLD50
MIC
Retention Time
Latest active users
ahmedbesheer: Dr. Ahmed Besheer
about an hour ago
fghara: Mr. Farhad Gharagheizi
about 1 hours ago
Tinkov_Oleg: Dr. Tinkov Oleg
about 2 hours ago
hadizadehf: Prof. Farzin Hadizadeh
about 4 hours ago
sonalirudra: Dr. Sonali Rudra
about 6 hours ago
TUSHAR WARJURKAR: Mr. Tushar Warjurkar
about 7 hours ago
Latest published models
logPow model published by tcirino
5 months ago
Transmittance model published by itetko_euos
7 months ago
Transmittance (qualitative) model published by lucic
7 months ago
CC50 model published by vkovalishyn
7 months ago
TTR binding activity model published by itetko_acs
more than a year ago
heart block model published by ggbond
more than a year ago

APPLICANT OVERVIEW

- Number of submissions: 465
- Number of teams: 41
  - Commercial: 5
  - Non-For-Profit: 7
  - Academic: 29



EUOS25 Challenge: Transmittance Category
1st place Solution

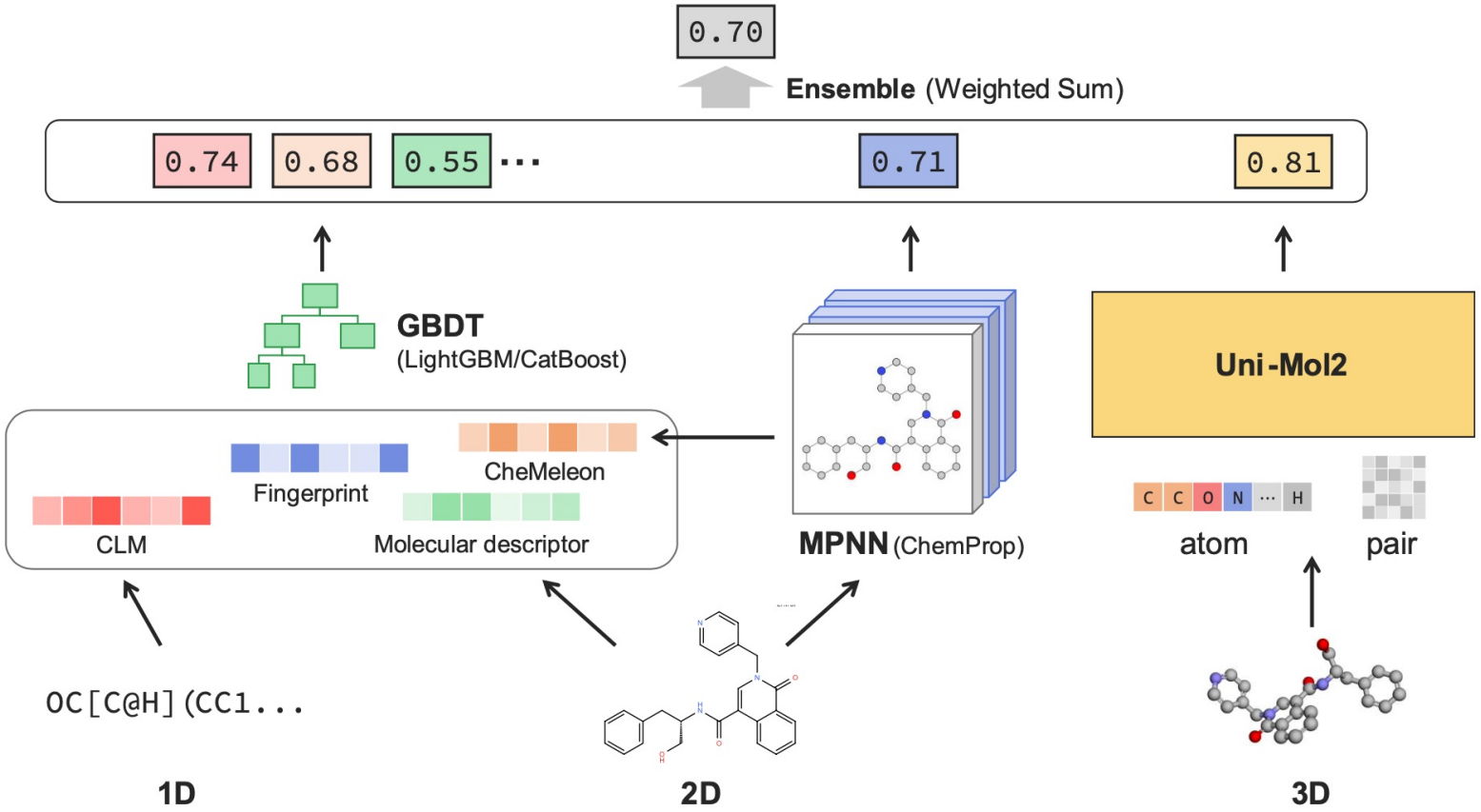
Team yumiz

Institute of SCIENCE TOKYO
School of Computing
Ohue Laboratory

Kairi Furui
Apakorn Kengkanna
Koh Sakano

furui@li.comp.isct.ac.jp

Overview of Winning Solution



- Multimodal molecular property prediction model integrating 1D+2D+3D information
- Ensemble of various prediction models through weighted sum

Training Techniques

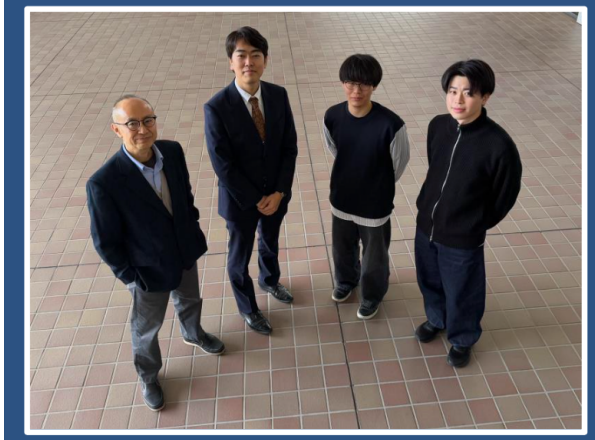
- **Data Preprocessing**
  - Standardization using RDKit
  - Salt/solvent removal and extraction of largest fragment
  - Molecular Sanitize, Normalize, Reionize, etc.

- **Data Splitting**
  - 5-fold CV based on Murcko scaffold
  - Parameter tuning for GBDT models via Optuna

- **Other Techniques**
  - **Focal Loss** to handle imbalanced data
  - **Weighted ensemble** of multiple models
    - Model selection and weighting based on CV performance

Meiji Pharmaceutical University

Team microsomes



MEIJI PHARMACEUTICAL UNIVERSITY

Multimodal Descriptors & Sequential Stacking

Winning Solution for Fluorescence Prediction in the EUOS25 Challenge

Members:
Mr. Wataru Miyahara, Mr. Kaito Inden, Mr. Yuma Iwashita, Prof. Yoshihiro Uesawa

Multimodal Strategy

MEIJI PHARMACEUTICAL UNIVERSITY

Precision QM Strategy

- **Combined 6 MOPAC Hamiltonians and MACE-xTB workflow**
- **Thermodynamic properties calculated in a simulated DMSO environment**
  - ML-optimized geometry
  - GFN2-xTB with ALPB solvation
  - RRHO thermodynamics

Massive Informatics

- **Leveraged 1,800+ Mordred descriptors**
- **Custom Extension:**

- PathCounts (up to Order 50) for long-range correlations.
- Conjugation Features (e.g., GraphEnergy, BLA) for π-system quality.

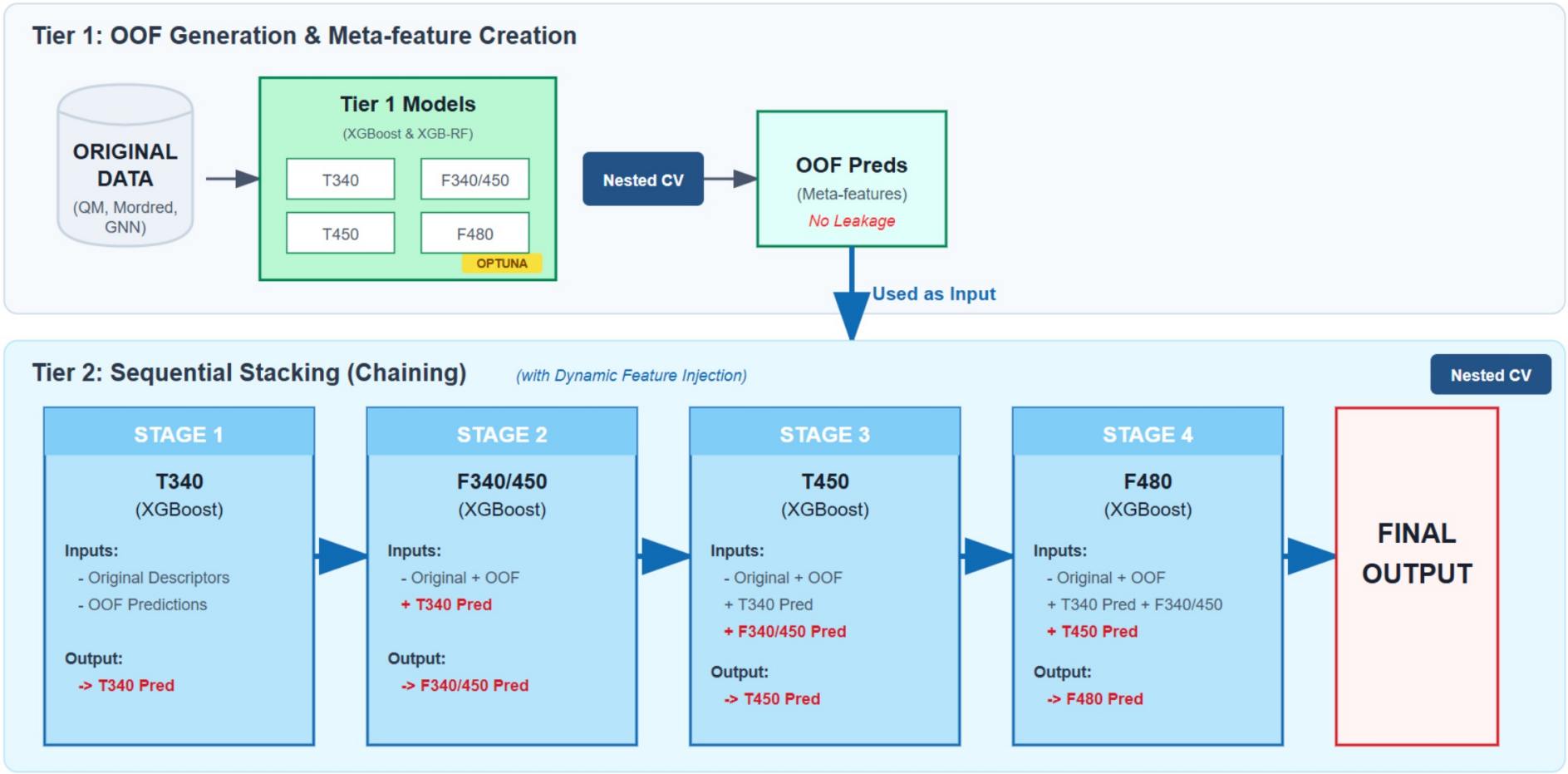
GNN Feature Generation

- **Utilized custom GINE-Net**

- Predictions as features
- Hidden layer embeddings
- Enriches feature space

Architecture: Sequential Stacking

MEIJI PHARMACEUTICAL UNIVERSITY



Online chemical database
with modeling environment

Home | Database | Models | Tox24 Challenge | EUOS25 Challenge | Retrosynthesis Submit | Retrosynthesis | Administration | Moder

SLAS2025 Blind Set AUC for prediction of Transmittance <= 70% at certain wavelength

Top scores of challenge participants

rank	User	Blind Test	340 nm	>450 nm	id	submissions	comment	Submission time (GMT-12)
1	yumiz	0.789	0.871	0.708	406	32	final sub	14-01-2026 15:56:36
2	filipsPL	0.788	0.88	0.697	400	14	20260114	13-01-2026 21:00:52
3	longhung25	0.784	0.874	0.695	439	5	no_plate_full	20-01-2026 00:31:07
4	Abdulfatai_Lawal	0.783	0.867	0.699	403	8	Fopt	14-01-2026 07:40:44
5	steffen.hirte	0.783	0.872	0.694	463	6	full ensemble	20-01-2026 20:48:34
	lucic	0.782	0.864	0.699	455	14	Team_5ZaMoSt	20-01-2026 14:01:09
7	rced	0.781	0.876	0.687	398	4	ensemble_CP	13-01-2026 07:01:48
	itetko_euos	0.781	0.876	0.685	465	6	organisers' own submission	22-01-2026 08:43:46
9	znavoyan	0.78	0.864	0.697	383	2	first_model	11-01-2026 22:55:38
10	Amidoff	0.779	0.87	0.688	408	20	finalmodel1.csv	14-01-2026 21:54:30
11	VSA	0.778	0.869	0.687	433	3	final_submission	19-01-2026 14:47:41
12	ArkaVUB	0.775	0.855	0.694	323	6	ArkaJ_v2	30-12-2025 03:48:15
13	uesawa	0.773	0.869	0.676	437	4	team_microsomes#3	19-01-2026 22:56:53
14	vchupakhin	0.773	0.854	0.692	459	7	subm5_aqaiopt.csv	20-01-2026 16:39:38
15	vfastovskii_unistra	0.769	0.864	0.673	466	1	submission_VFASTOVSKII.csv	21-01-2026 02:17:17
16	dodo.flp	0.765	0.853	0.677	436	46	submission_final.csv	19-01-2026 21:23:16
17	jwburns	0.764	0.856	0.672	372	26	CheMeleon (159022f)	09-01-2026 07:48:45
18	wangze09	0.763	0.857	0.668	294	12	overfitting_version	26-12-2025 20:12:43

Online chemical database
with modeling environment

Home | Database | Models | Tox24 Challenge | EUOS25 Challenge | Retrosynthesis Submit | Retrosynthesis | Administration | Moderation

SLAS2025 Blind Set AUC for Fluorescence

Top scores of challenge participants

rank	User	Blind Test	340/450 fluorescence	480/540, 525/598, or 560/610 fluorescence	id	submissions	comment	Submission time (GMT-12)
1	uesawa	0.801	0.894	0.707	437	4	team_microsomes#3	19-01-2026 22:56:53
2	vfastovskii_unistra	0.79	0.886	0.693	466	1	submission_VFASTOVSKII.csv	21-01-2026 02:17:17
3	ArkaVUB	0.789	0.88	0.698	323	6	ArkaJ_v2	30-12-2025 03:48:15
4	filipsPL	0.786	0.901	0.67	400	14	20260114	13-01-2026 21:00:52
	itetko_euos	0.784	0.896	0.673	465	6	organisers' own submission	22-01-2026 08:43:46
6	yumiz	0.783	0.897	0.668	406	32	final sub	14-01-2026 15:56:36
7	znavoyan	0.779	0.891	0.667	383	2	first_model	11-01-2026 22:55:38
8	longhung25	0.778	0.898	0.658	439	5	no_plate_full	20-01-2026 00:31:07
9	vchupakhin	0.778	0.873	0.683	459	7	subm5_aqaiopt.csv	20-01-2026 16:39:38
10	steffen.hirte	0.778	0.891	0.666	463	6	full ensemble	20-01-2026 20:48:34
11	wangze09	0.775	0.882	0.669	294	12	overfitting_version	26-12-2025 20:12:43
12	VSA	0.77	0.89	0.651	433	3	final_submission	19-01-2026 14:47:41
13	rced	0.768	0.902	0.634	398	4	ensemble_CP	13-01-2026 07:01:48
14	wim	0.768	0.892	0.644	424	2	ensembled_LGBM.csv	18-01-2026 05:51:36
15	Abdulfatai_Lawal	0.764	0.897	0.631	403	8	Fopt	14-01-2026 07:40:44
	naibaF	0.761	0.893	0.628	390	2	first try	12-01-2026 08:47:41

fillipsPL: 3 days (72 hours) per target using 384 AMD EPYC 9684X CPU cores

itetko_euos & naibaF: hours to ten of hours of GPU card NVIDIA Tesla V100 per one method

Conclusions

- Representation learning models demonstrate great power
- Consensus/Ensemble/multimodal models demonstrate the highest accuracy*
  - An access to HPC cluster is a great asset
- Select and combine best methods & descriptors
  - Combine diverse methods
  - Use correct validation metrics which match the task
- **Most importantly: have a good luck! ☺**

*Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena Pyriformis. J. Chem. Inf. Model. 2008, 48 (4), 766–784. https://doi.org/10.1021/ci700443v - one of the first published consensus model

Acknowledgements

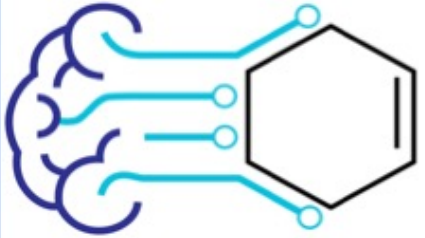
EU-OPENSCREEN

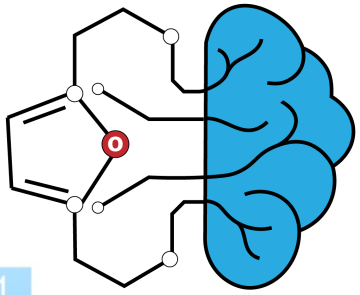
Katya Ahmad
Fabian Krüger
Dina Khasanova
Mark Embrechts
Michael Sattler

Peter Hartog
Thalita Cirino
Nesma Mousa
Massa Zahdeh
Andi Hunklinger
Alessandro Dipalma
Paula Torren-Peraire
Varvara Voinarovska

Guillaume Godin (OSMO)
Ruud van Deursen (DSM-Firmenich)

Alexander von Humboldt
Stiftung/Foundation





AiChemist