OPERA, AN OPEN-SOURCE AND OPEN-DATA SUITE OF QSAR MODELS

NTP National Toxicology Program, U.S. Department of Health and Human Services

K. Mansouri[1], X. Chang[2], D. Allen[2], R. Judson[3], A.J. Williams[3], W. Casey[1], N. Kleinstreuer[1]

[1]NIH/NIEHS/DNTP/NICEATM, RTP, NC, USA; [2]ILS, RTP, NC, USA; [3]CCTE/EPA, RTP, NC, USA

Introduction

- OPERA is a free and open-source/open-data suite of QSAR models providing predictions for toxicity endpoints and physicochemical, environmental fate, and ADME properties.
- In addition to predictions, OPERA provides accuracy estimates, applicability domain assessment and experimental data when available.
- Recent additions to OPERA include models for estrogenic activity, androgenic activity, and acute oral systemic toxicity developed through international collaborative modeling projects, and updates to models predicting plasma protein binding and intrinsic hepatic clearance.
- OPERA predictions for ADME parameters (CL_{int} and F_U) as well as physicochemical parameters (logP, pKa, and logD) are used as inputs for the *in vitro* to *in vivo* extrapolation (IVIVE) workflow on the NTP's Integrated Chemical Environment (ICE: https://ice.ntp.niehs.nih.gov/).
- OPERA predictions are also available both via the user interface and for download from the EPA's CompTox Chemicals Dashboard. (https://comptox.epa.gov/dashboard).

OPERA application

General approach:

- OECD 5 principles for QSAR validation are employed during modeling
- Only high-quality curated data are used to build the models
- Chemical structures are processed using the QSAR-ready standardization workflow prior to modeling
- The QSAR-ready workflow is also implemented in the app for user input processing structures prior to prediction
- Works with different input and output formats
- Provides applicability domain and prediction accuracy assessment
- Provides experimental values when available
- Provides information about the nearest neighbors
- Provides molecular descriptor values for transparency
- OECD-compliant QSAR model reporting format (QMRF) reports available

Availability:

Predictions:

- EPA CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard)
- NTP's Integrated Chemical Environment (https://ice.ntp.niehs.nih.gov/)

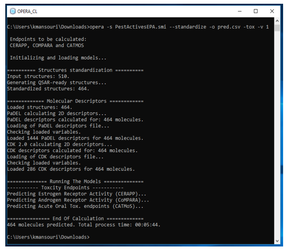
Standalone desktop application (current version 2.7):

- Github: https://github.com/NIEHS/OPERA
  - Windows and Linux packaged installers with dependencies.
  - Additional wrappers and libraries: Java, Python, C/C++
- NTP KNIME server: knime.niehs.nih.gov/knime/

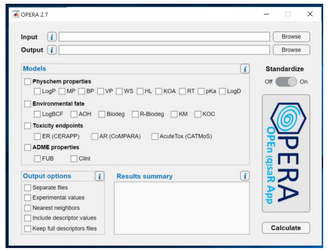
More info:

- https://ntp.niehs.nih.gov/go/opera

Interfaces:



Command line



Graphical user interface

Existing, recently updated and future models

All models:

Environmental fate	
AOH	Atmospheric Hydroxylation Rate
BCF	Bioconcentration Factor
BioHL	Biodegradation Half-life
RB	Ready Biodegradability
KM	Fish Biotransformation Half-life
KOC	Soil Adsorption Coefficient

ADME properties	
FUB	Atmospheric Hydroxylation Rate
Clint	Bioconcentration Factor

Physchem properties	
BP	Boiling Point
HL	Henry's Law Constant
KOA	Octanol/Air Partition Coefficient
LogP	Octanol-water Partition Coefficient
MP	Melting Point
KOC	Soil Adsorption Coefficient
VP	Vapor Pressure
WS	Water Solubility
RT	HPLC Retention Time
pKa	Acid Dissociation Constant
logD	Distribution Coefficient

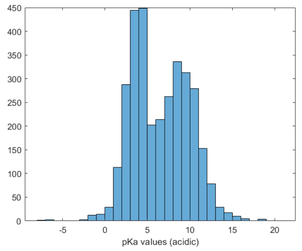
Toxicity endpoints	
ER	Estrogen Receptor Activity
AR	Androgen Receptor Activity
AcuteTox	Acute Oral Systemic Toxicity

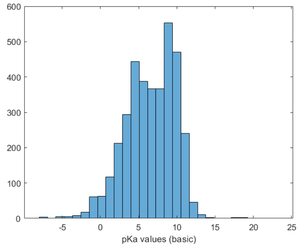
Future models	
CACO2	Caco-2 permeability
Inhalation	Acute Inhalation Systemic Toxicity
SixPack	Acute Toxicity Six-Pack Endpoints
UGT	Glucuronidation: substrate selectivity
SULT	Sulfation: substrate selectivity

Recent model updates:

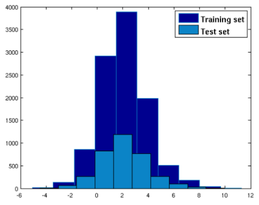
pKa: acid dissociation constant

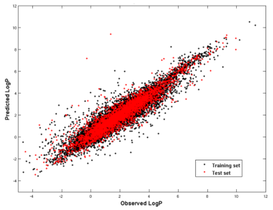
- The OPERA pKa model was built on a curated version of the DataWarrior dataset.
- The acidic (3260 chemicals) and basic (3680 chemicals) datasets were modeled separately
- First, a weighted-kNN classification model predicts whether a chemical is acidic, basic or both. Then a SVM model predicts the strongest acidic and basic pKa values
- The acidic and basic pKa models reached an R^2 of 0.72 and 0.78 and RMSE of 1.80 and 1.53, respectively.





LogP: octanol-water partition coefficient





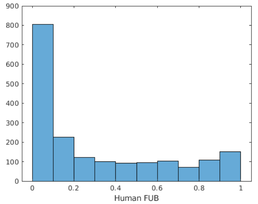
- The OPERA logP model was initially built using a curated dataset from the PHYSPROP database.
- The overall statistics of the model reached an R^2 of 0.86 and an RMSE of 0.78 for the test set.
- The logP model as well as other OPERA models (water solubility, and vapor pressure) have been updated to account for highly investigated groups of chemicals such as polyfluorinated substances (PFAS).

LogD: distribution coefficient

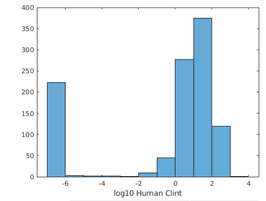
- LogD is the distribution coefficient that takes into account pH-dependence and is used to estimate the different relative concentrations of the ionized and non-ionized forms of a chemical at a given pH.
- OPERA uses both pKa and logP predictions to provide logD estimates for ionizable chemicals at pH 5.5 and pH 7.4.
- LogD is estimated using the following formula: $logD_{(pH)} = logP - \log(1 + 10^{(pH - pKa)})$

PK parameters: F_U and CL_{int}

- Both CL_{int} and F_U OPERA models were built using datasets combined from different sources.
- Most of the data entries are also available in the EPA's high-throughput toxicokinetic (*httk*) R package.
- After several rounds of automated and manual curation to reduce errors, variability and outliers, the CL_{int} and F_U datasets consisted of 1056 and 1873 chemicals, respectively.
- The CL_{int} dataset was modeled in two steps:
  - First a classification model to separate the cleared from non-cleared chemicals
  - Then, a regression model is applied to predict the CL_{int} value for the cleared chemicals.

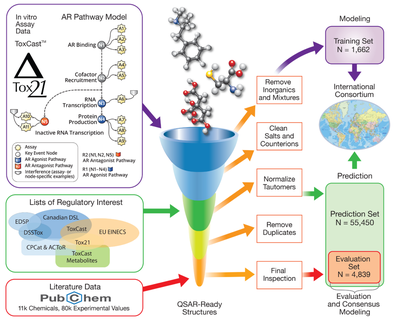


	Cross-validation	Test
R^2	0.63	0.65
RMSE	0.20	0.19



	Cross-validation	Test
BA	0.70	0.57
R^2	0.40	0.39
RMSE	0.73	0.79

Consensus of international collaborations



- The toxicity endpoints included in OPERA are the estrogen and androgen pathway activities and the acute oral toxicity
- The models were the result of three international collaborations including over a hundred scientists from a total of 35 research groups covering governmental institutions, industry and academia
- Multiple models were combined into a unique consensus as show in the diagram.

CERAPP: Collaborative Estrogen Receptor Activity Prediction Project

	Binding		Agonist		Antagonist	
	Training	Validation	Training	Validation	Training	Validation
Sn	0.93	0.58	0.85	0.94	0.67	0.18
Sp	0.97	0.92	0.98	0.94	0.94	0.90
BA	0.95	0.75	0.92	0.94	0.80	0.54

CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity

	Binding		Agonist		Antagonist	
	Training	Validation	Training	Validation	Training	Validation
Sn	0.99	0.69	0.95	0.74	1.00	0.61
Sp	0.91	0.87	0.98	0.97	0.95	0.87
BA	0.95	0.78	0.97	0.86	0.97	0.74

CATMoS: Collaborative Acute Toxicity Modeling Suite

- CATMoS consisted of five different endpoints and the final consensus model was a combination of all predictions using a weight of evidence approach.
- CATMoS is currently being evaluated for regulatory use by the US EPA.

	Very-Toxic		Non-Toxic	
	Training	Evaluation	Train	Evaluation
BA	0.93	0.84	0.92	0.78
Sn	0.87	0.70	0.88	0.67
Sp	0.99	0.97	0.97	0.90

GHS categories										
	Training					Evaluation				
	Cat 1	Cat 2	Cat 3	Cat 4	Cat 5	Cat 1	Cat 2	Cat 3	Cat 4	Cat 5
BA	0.88					0.74				
Sn	0.73	0.75	0.84	0.80	0.88	0.50	0.53	0.56	0.66	0.67
Sp	0.99	0.99	0.92	0.89	0.96	0.99	0.97	0.89	0.74	0.90

LD50		
	Training	Evaluation
R^2	0.85	0.65
RMSE	0.30	0.49

EPA categories								
	Training				Evaluation			
	Cat 1	Cat 2	Cat 3	Cat 4	Cat 1	Cat 2	Cat 3	Cat 4
BA	0.87				0.74			
Sn	0.87	0.83	0.91	0.63	0.70	0.56	0.81	0.40
Sp	0.99	0.95	0.75	0.98	0.97	0.88	0.62	0.97

References

[1] Mansouri K. et al. J Cheminform (2018) https://doi.org/10.1186/s13321-018-0263-1
[2] Mansouri, K. et al. SAR & QSAR in Env. Res. (2016) https://doi.org/10.1080/1062936X.2016.1253611
[3] Williams A. J. et al. J Cheminform (2017) https://doi.org/10.1186/s13321-017-0247-6
[4] JRC QSAR Model Database https://qsardb.jrc.ec.europa.eu/qmrf/endpoint
[5] Mansouri, K. et al. EHP (2016) https://doi.org/10.1289/ehp.1510267
[6] Mansouri, K. et al. J Cheminform (2019) https://doi.org/10.1186/s13321-019-0384-1
[7] Mansouri, K. et al. EHP (2020) https://doi.org/10.1289/EHP5580
[8] Kleinstreuer et al. Comp Tox (2018) https://doi.org/10.1016/j.comtox.2018.08.002
[9] Mansouri, K et al. EHP "CATMoS manuscript" (2021) In Press