

The EPA Online Database of Experimental and Predicted Data to Support Environmental Scientists

1. U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology (NCCT), Research Triangle Park, NC 2. Oak Ridge Institute for Science and Education (ORISE) Participant, Research Triangle Park, NC

Problem Definition and Goals

Problem: There is limited access to freely available high quality experimental and predicted chemical data online.

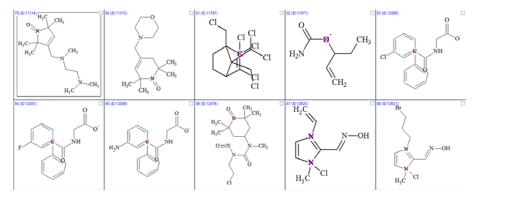
Goals: To deliver access to curated datasets of experimental and physicochemical data associated with chemicals of interest to environmental scientists. Make the data available as downloadable Open data. To develop predictive models from the curated data and use to predict properties for >700,000 chemicals and make available online. To provide details regarding the performance of the models.

Abstract

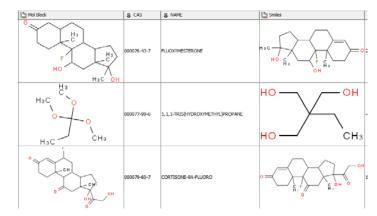
As part of our efforts to develop a public platform to provide access to experimental data and predictive models for physicochemical data we have delivered a web-based database, the EPA Chemistry Dashboard (https://comptox.epa.gov). In order to develop the prediction models we used publicly available data and a combination of manual and automated review processes to produce highly curated datasets. The automated curation processes for the validation of the data used a KNIME workflow and included approaches to validate different chemical structure representations (e.g. molfile and SMILES), identifiers (chemical names and registry numbers), and methods to standardize the data into QSAR-consumable formats for modeling. Machine-learning approaches were applied to create a series of models that have been used to generate predicted physicochemical and environmental parameters for over 700,000 chemicals. The data have been made available online via the EPA's iCSS Chemistry Dashboard and includes detailed model reports regarding whether or not a chemical is contained within a particular applicability domain, provides access to nearest neighbors within the training set as well as detailed QMRF (QSAR Modeling Report Format) reports for each of the models.

Source Data and Example Errors

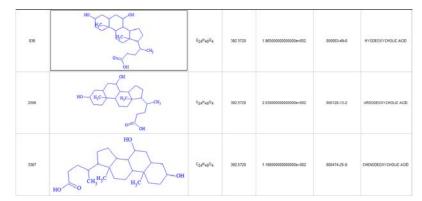
- The PHYSPROP data were sourced online¹ as SDF files. A total of 13 endpoints were represented, including LogKow, water solubility, melting point, boiling point, and others. The largest dataset, LogKow, contained over 15,800 individual chemicals. Each data point included the Mol-Block, SMILES, CASRN, Name, LogKow value and, where available, a reference. This dataset was chosen as representative for examining the quality of data.
- A manual examination of the data revealed a number of issues: e.g., SMILES and Mol-Block did not agree; CASRN did not match the correct structure; SMILES could not be converted; a single chemical structure would be listed multiple times with different property values. Example errors across various PHYSPROP datasets are shown below.



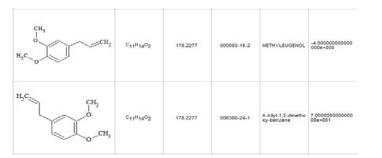
Examples of hypervalency in LogKow dataset



Different structures in Mol-Block and SMILES



Equivalent structures, different CASRN, names, and values in MP dataset



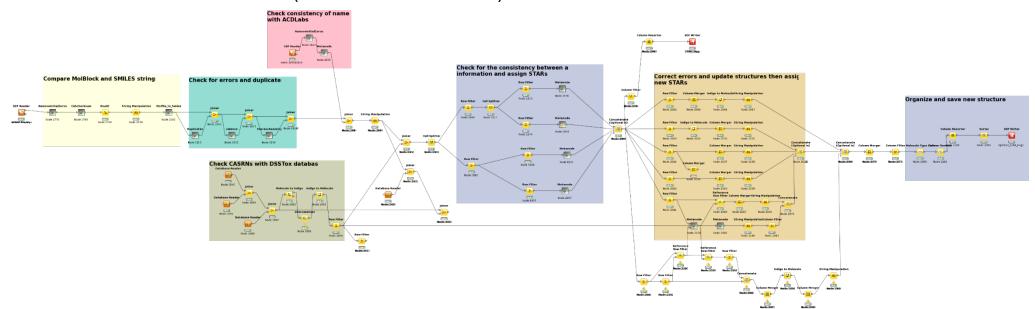
Equivalent structures but with different CASRN, names and values in MP dataset: -4 and + 70°C

Antony Williams^{1,*}, Christopher Grulke¹, Kamel Mansouri², Jennifer Smith¹, Jordan Foster¹, Michelle Krzyzanowski and Jeff Edwards¹

Automated Analysis Using KNIME

The manual investigation of the data allowed us to develop a KNIME² workflow for automated processing. This workflow was derived from earlier work by Mansouri et. al.³ and is represented in the figure below as a series of blocks representing, for example:

- Compare Mol-Block and SMILES (2268 different)
- Check names against dictionary (555 invalid) Assign Quality flags based on consistency
- Check for duplicates (657 structures, 531 names) Check CASRN Numbers (3646 invalid CASRN)
- among data fields



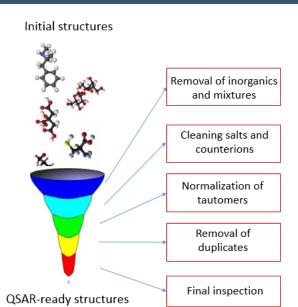
The KNIME workflow for automated processing of PHYSPROP data

The KNIME workflow was used to insert various levels of Quality Flags indicating consistency between chemical structure formats and identifiers. The consistency flag definitions and distribution are summarized below for the >15k chemicals.

4 STAR ENHANCED:	Name/CASRN/Mol/SMILES added Stereo:	550
4 STAR:	3 of 4 Name/CASRN/Mol/SMILES:	5967
3 STAR ENHANCED:	3 of 4 Name/CASRN/Mol/SMILES added Stereo:	177
3 STAR:	3 of 4 Name/CASRN/Mol/SMILES:	7910
2 STAR PLUS:	2 of 4 Name/CASRN/Mol/SMILES/Tautomer:	133
2 STAR:	2 of 4 Name/CASRN/Mol/SMILES:	1003
1 STAR:	No two fields consistent	379

Predictive models were developed using only the 3 and 4 star curated data.

Preparing data for QSAR Modeling



For the purposes of QSAR modeling, the 3 and 4 STAR datasets were processed through a KNIME workflow. This processing removed inorganics and mixtures, processed salts into neutral forms (except for melting point data), normalized tautomers, and removed duplicates. The resulting "QSAR-ready" file(s) were modeled using Genetic Algorithm-Partial Least Squares with 5-fold cross validation and utilizing 2D PaDEL⁴ molecular descriptors. Multiple modeling runs (100) produced the best models using a minimum number of descriptors. The models for all 13 endpoints are available as both Windows and Linux executable binaries and as a C++ library that can be called by a separate application.



Innovative Research for a Sustainable Future

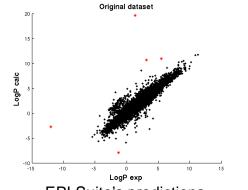
2627/P125 American Chemical Society Fall Meeting Philadelphia, PA August 21-25, 2016

ORCID: 0000-0002-2668-4821

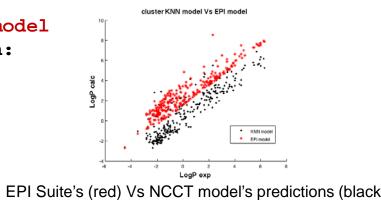
Antony Williams I williams.antony@epa.gov I 919-541-1033

Model Performance

The curation of the available data, utilization of large datasets for training and validation, and application of innovative machine-learning approaches produced high performance, yet simple models with a minimum number of descriptors. The logP model was built on a dataset of more than 14k chemicals using only 10 descriptors. With a wider applicability domain this model outperformed EPI Suite's predictions (left side plot) especially for the over-estimated cluster extracted and plotted together with the new model's predictions (right side plot).



Statistics of the new model 5-fold cross-validation: Q2: 0.87 RMSE: 0.67 Fitting 11247: R2: 0.87 RMSE: 0.66 Test set prediction : R2: 0.84 RMSE: 0.65



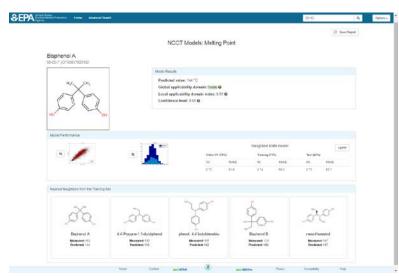
EPI Suite's predictions

Accessing ~1 Million Predicted Properties Online

The iCSS Chemistry Dashboard (ICD) is a public EPA-hosted web application providing access to over 700,000 chemicals from EPA's DSSTox database⁵. It integrates experimental and predicted data and is a hub to other NCCT apps and web-based resources. The 3 and 4 STAR curated experimental data are accessible via the ICD application. All chemicals were also passed through the prediction models and detailed model reports and results are freely available at http://comptox.epa.gov/dashboard

							Take -	Intervention	Shame.	-Cre
Bisphenol A										
80-05-7 DTXSID7020	(0) +									
000011011/001020	IDE O									
O Searched by Approved Name	e: Found 1 result for bisphenol &									
Q in 1 4. Q.										
a le le le le le l	Inter	naic Properties								
		decular Formula						1		
HyC, CHy							Q. Proj. AP.D			
X		verage Maasi 228.								
NE		inneisotopie Mass	1 228 114C8 gi	moi						
126	//									
	380	dura taethen								
196.5	Rec	ord information								
Charries Properties Enternal L	Lana Dynamyria Perskut Campoilian	To-Cast is Vite Date	Eronan	Analytical Pul	bOlum Comme	-				
Chamical Proportion Enternal L		To-Carl II Vite Date	Espenne	Analytical Pul	BOlum Conve	4				
	Deventeed as: Cov Court SCP Noperty						d) Range (Fred.)	Result Uni		
Summary Oxforce water Partition	Download as: Cov Dust SCP	Average (Esp					d.) Range (Fred.) 320 to 3.64	Result Uni		
Summary Octants water Partition Coefficient (rop)	Download as: cov cuel sor Property Octarol-water Partition Coefficient (or Vater Solubility	Average (Exp gpt 3.30 (2) 5.25e-04 (1)) Median (Ex	p.) Renge (Cap 3.43 5.25e-04	 Average (Prec 3.42 (2) 2.22e-03 (2) 	.) Median (Fre		+ nort		
Serenary Octandi valter Partition Coefficient (Ingo) Trates Solubity Metting Part	Download as: cov cover scr Property Octand-water Parition Coefficient (o Valet Solubility Valet Solubility	Average (Exp 3:30 (2) 5:26e-04 (1) 155 (7)	3 43 5 20+04 156	p.) Range (Exp 3.43 5.26e-34 153.10.158	 Average (Prec 3.42 (2) 2.22e-03 (2) 135 (2) 	 5.42 2.229-03 5.61 	3 20 to 3.64 7 56e-04 to 3 68e-03 132 to 144			
Summery Charal webs Partition Coefficient (tripp) Robert Solutiony Metting Front Bolling Partit	Download as: cov cuel sor Property Octarol-water Partition Coefficient (or Vater Solubility	Average (Exp gpt 3.30 (2) 5.25e-04 (1)) Median (Cx 3.43 5.250-04	p.) Renge (Cap 3.43 5.25e-04	 Average (Prec 3.42 (2) 2.22e-03 (2) 	 I.) Median (Fra 3.42 2.22+03 	3 20 to 3.64 7 56e-04 to 3 68e-03	+ nort		
Serenary Octandi valter Partition Coefficient (Ingo) Trates Solubity Metting Part	Deventised as: COV Event SEP Property Octanol-water Partition Doefficient (or Water Soluble) Moting Priori Babling Priori Mapor Priorites	Average (Exp 3:30 (2) 5:26e-04 (1) 155 (7)	3 43 5 20+04 156	p.) Range (Exp 3.43 5.26e-34 153.10.158	 Average (Prec 3.42 (2) 2.22e-33 (2) 136 (2) 349 (2) 7.05e-08 (1) 	 Median (Pre 3.42 2.25e-03 136 5e1 7.05e-08 	3 20 to 3.64 7 56e-04 to 3 68e-03 132 to 144 334 to 36a			
Summery Charal webs Partition Coefficient (tripp) Robert Solutiony Metting Front Bolling Partit	Devendend as: COV Deven SDP Property Octanor-water Particles Doefficient (or Valent Solubity Wettrap Prent Boding Forst	Average (Exp apt 3.30 (2) 5.25e-04 (1) 155 (7) 200 (1)	J Median (Da 3.43 5.20e-04 156 200	p.) Range (Exp 3.43 5.25e-04 153.10.158 200	 Average (Prec 3.42 (2) 2.22e-03 (2) 135 (2) 349 (2) 	 L) Median (Free 3.42 2.22e-03 538 549 	3 20 to 3.64 7 56e-04 to 3.68e-03 132 to 144 334 to 364	,c ,c		
Serverary Octorial water Partition Conflicted Partition Conflicted Partition Research Society Memory Part Delays Part Viscor Pressure	Drawieed as: COV Evel SDP Property Outerst-outer Particle Darficient (or Vater Soundly Warry Parts Beiling Frant Vapor Presson Sol Adorption Coefficient Octavio and Particle Coefficient	Average (Exp apt 3.30 (2) 5.26e-04 (1) 155 (7) 200 (1) -	J Median (Da 3.43 5.200-04 156 200 -	p.) Range (Dap 3.43 5.25e-34 153 to 158 200	 Average (Prec 3.42 (2) 2.22e-33 (2) 136 (2) 349 (2) 7.05e-08 (1) 	 Median (Fre 3.42 2.25×03 138 341 7.05±08 2.52 8.39 	3 20 to 3.64 7 56e-04 to 3 68e-03 132 to 144 334 to 36a	rion Jion C génne		
Senser Ordersi ander Partition Conference (regio) Reletti Sololiky Mettry Part Soliky Part Soliky Part Solik Astroption Conflicent Ordersi An Prettion Conflicent	Deventioned as: COV Event SDP Nospenty Ocurrin-water Particles Coefficient (or Youter Sociality Metrog Parts Metrog Parts Sol Advergon Coefficient	Average (Exp apt 3.30 (2) 5.26e-84 (1) 155 (7) 200 (1) -	J Median (Da 3.43 5.200-04 196 200 -	p.) Range (Exp 3.43 5.25e-34 153.10.158 200 -	 Average (Pee 3.42 (2) 2.22e-33 (2) 135 (2) 349 (2) 7.06e-08 (1) 2.92 (2) 	 I. Median (Pre- 3.42 2.22+03 38 341 7.06±08 2.92 8.39 10.4 	3 20 to 3.64 7 56e-04 to 3 66e-02 132 to 144 334 to 36a - 2 74 to 3.60	- - - - - - - -		
Semmery Octorial water Partition Confinent (page) Natern Sochilly Meterng Part Delary Part Vagor Pressure Soli Adorphico Coofficient	Devented as: Corr Duet 307 Preperty Octansi-earls Partices Duefficient (or Youte Soulding Manry Parts Manry Parts Manry Parts Manry Parts Said Asseption Coefficient Octanoi a Profession Coefficient Annuelphili Many John Rate Biodegoadation Mid Unit	Average (Exp api 3.30 (2) 5.26e-64 (1) 155 (7) 200 (1) -	 Median (Cx 3.43 5.20+04 156 200 - - - 	p.) Range (Exp 3.43 5.25e-34 153.19.158 200 -	 Average (Pee 3.42 (2) 2.22e-33 (2) 136 (2) 349 (2) 7.06e-08 (1) 2.92 (2) 8.39 (1) 	 Median (Fre 3.42 2.25×03 138 341 7.05±08 2.52 8.39 	3 20 to 3.64 7 56e-04 to 3 60e-03 132 to 144 334 to 364 - 2 74 to 3.63	- - - - - - - -		
Semmery Chards where Partian Cartiever (rigg) Nate: Soliday Nate: Soliday Part Nate: Part Soliday Part Soliday Part Contras & Partian Conflicted Adversion Conflicted Chards & Partian Conflicted Adversion Partian Conflicted Adversion Partian Conflicted Adversion Partian Conflicted	Deventeed as: 007 Event 307 Neperty Outant-values Partition Coefficient (or Water double): Metrop Parel Beiling Forst Sid Adverption Coefficient Organia a Rodrig Coefficient Designa in Phylory Advin Rate Disclamations Factor	Average (Exp api 3.50 (2) 5.26e-84 (1) 185 (7) 200 (1) - -) Median (Da 3.43 5.20e-04 156 200 - - - - -	p.) Range (Dap 3.43 5.26e-04 153 to 158 200 - - - -	 Average (Prec 3.42 (2) 2.22e-03 (2) 135 (2) 349 (2) 7.066-08 (1) 2.92 (2) 8.39 (1) 10.4 (1) 55.1 (1) 173 (1) 	 Median (Pre 3.42 2.22+03 348 348 7.06+08 9.92 8.39 10.4 15.1 173 	3 20 to 3.64 7 50e-04 to 3 60e-03 132 to 144 334 to 364 - 2 74 to 3 10 -	+ Juan 2' - - -		
Secondary Orthon's scales Partition Conference (Party) Verter Society Marting Party Society Party Society Party Society Party Society Party Orthon Society Conference Amongonales (Independent Refer	Devented as: Corr Duet 307 Preperty Octansi-earls Partices Duefficient (or Youte Soulding Manry Parts Manry Parts Manry Parts Manry Parts Said Asseption Coefficient Octanoi a Profession Coefficient Annuelphili Many John Rate Biodegoadation Mid Unit	Average (Exp apt 3.30 (2) 5.26e-64 (1) 155 (7) 200 (1) - - -	 Median (Da 3.43 5.259-04 196 200 - <	p.) Range (Exp 3.43 5.25e-34 153 to 158 200 - - -	 L) Average (Perc 3.42 (2) 2.22e-03 (2) 136 (2) 340 (2) 7.06e-08 (1) 2.92 (2) 8.39 (1) 10.4 (1) 15.1 (1) 	 I. Median (Pre- 3.42 2.22x-03 38 341 7.06x-08 2.52 8.39 40.4 15.4 	3 20 to 3.64 7 56e-04 to 3 60e-03 132 to 144 334 to 36a - 2 74 to 3 50 -	+ moit, iC embig -		
Semmery Chards where Partian Cartiever (rigg) Nate: Soliday Nate: Soliday Part Nate: Part Soliday Part Soliday Part Contras & Partian Conflicted Adversion Conflicted Chards & Partian Conflicted Adversion Partian Conflicted Adversion Partian Conflicted Adversion Partian Conflicted	Deventeed as: 007 Event 307 Neperty Outant-values Partition Coefficient (or Water double): Metrop Parel Beiling Forst Sid Adverption Coefficient Organia a Rodrig Coefficient Designa in Phylory Advin Rate Disclamations Factor	Average (Exp api 3.50 (2) 5.26e-84 (1) 185 (7) 200 (1) - -) Median (Da 3.43 5.20e-04 156 200 - - - - -	p.) Range (Dap 3.43 5.26e-04 153 to 158 200 - - - -	 Average (Prec 3.42 (2) 2.22e-03 (2) 135 (2) 349 (2) 7.066-08 (1) 2.92 (2) 8.39 (1) 10.4 (1) 55.1 (1) 173 (1) 	 Median (Pre 3.42 2.22+03 348 348 7.06+08 9.92 8.39 10.4 15.1 173 	3 20 to 3.64 7 50e-04 to 3 60e-03 132 to 144 334 to 364 - 2 74 to 3 10 -	+ moit, ic rc embig -		

A summary of available chemical properties for Bisphenol A



The model report for the melting point prediction for Bisphenol A – including nearest neighbors.

Future Work

• Release NCCT models as interactive online prediction tools in the near future via the ICD. • Integrate the suite of EPA T.E.S.T⁶ physchem and toxicity prediction models to expand the collection of available models.

• Source additional data to expand the training sets underpinning the prediction algorithms.

References

PHYSPROP Data: http://esc.syrres.com/interkow/EpiSuiteData_ISIS_SDF.htm

KNIME: https://www.knime.org/

Mansouri et al. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project, Environ Health Perspect; DOI:10.1289/ehp.1510267

PaDEL descriptors, http://padel.nus.edu.sg/software/padeldescriptor/

EPA Distributed Structure-Searchable Toxicity (DSSTox) Database, http://www.epa.gov/chemicalresearch/distributed-structure-searchable-toxicity-dsstox-database

EPA Toxicity Estimation Software Tool (T.E.S.T.) software http://www.epa.gov/chemical-research/toxicityestimation-software-tool-test