Supplemental Information:

**Rapid Experimental Measurements of Physicochemical Properties to Inform Models and Testing**

Chantel I. Nicolas1,2,5, Kamel Mansouri1,2,5, Katherine A. Phillips3, Christopher M. Grulke2, Ann M. Richard2, Antony J. Williams2, James Rabinowitz2, Kristin K. Isaacs3, Alice Yau4, John F. Wambaugh2\*

1ScitoVation, LLC

6 Davis Drive, Durham, NC 27703

2National Center for Computational Toxicology

Office of Research and Development

US EPA, Research Triangle Park, NC 27711

3National Exposure Research Laboratory

4Southwest Research Institute

San Antonio, TX 78238

5Oak Ridge Institute for Science and Education,

Oak Ridge, TN 37831

**\*Corresponding Author:**

109 T.W Alexander Dr., NC 27711, USA

Wambaugh.john@epa.gov

Phone: (919) 541-7641; fax: (919) 541-1194

**Disclaimer:** The views expressed in this publication are those of the authors and do not necessarily represent the views for policies of the U.S. Environmental Protection Agency. Reference to commercial products or services does not constitute endorsement.

# SI Methods

## Determination of Ionization Potential

In order to determine pKa via ultraviolet-visible (UV-Vis) spectroscopy, absorbance measurements for pH values ranging from acidic (pH = 3) to basic (pH = 10) were collected at various wavelengths (nm) depending on the compound1. In order to bypass the need to look up extinction coefficients for hundreds of compounds as is done in Tomsho (2011)2, a simplified algebraic method was used per Patterson (1999)3

= p Eq. S1

where is the absorbance of the basic pH solution (10 in most cases) and is the absorbance of the acidic pH solution (3 in most cases), and A is the absorbance at various intermediate pH solutions (4,5,7,8 and 9 in most cases). If an absorbance was reported at multiple wavelengths, the method of Salgado and Hernández4 was used to select the optimal wavelength for pKa determination . In this method, the wavelength that corresponds to the greatest difference between absorbance curves of the acidic and basic solutions is selected to determine pKa. An example of this is where a chemical (4-Nitrophenol) had absorbance values collected at three different wavelengths (230, 320, and 400 nm). The absorbance data from the 400 nm wavelength measurement were used to parameterize the Patterson3 equation and thus pKa was determined from a linear model fitted to that data (Figure S10).

## Structural Similarity of Measured Chemicals

In an effort to understand the diversity of the chemicals analyzed experimentally, as well as explain why some chemicals were able to produce experimental estimate values whereas others were not, the structural similarity of the 200 chemicals was compared. This was done by computing a pair-wise similarity matrix (i.e., a 200-by-200 matrix) where each element represents the Tanimoto5 similarity between a pair of chemicals. ToxPrint fingerprints6 were used as bits in the computation of the similarity matrix. Fingerprints were removed if all values were constant across the 200 chemicals (e.g., if all chemicals had a value of 0 for a given fingerprint, it was removed). Tanimoto5 similarity is calculated according to the equation:

Eq. S2

where S is the pair-wise Tanimoto5 similarity between two chemicals, nTF is the number of bits in the fingerprint that are true (present) in the first chemicals but are false (absent) in the second chemical, nFT are the number of bits that are false in the first chemical, but are true in the second chemical, and nTT are the number of bits that are true in both the first chemical and the second chemical. In essence, this metric is the ratio of the number of bits that are present in both chemicals to the total number of bits that are present either chemicals of the pair. The QSAR-ready SMILES string for each chemical, as well as the ToxPrint fingerprints of all 200 chemicals, are provided in Table S12.

# SI CODE

Code S1: KNIME workflow for test chemical selection and associated data (S1\_Chemical\_Selection\_Data.zip)

# SI FIGURES



Figure S1. Structural similarity of chemicals with octanol-water partition coefficient values for both previous measurements (i.e., PHYSPROP measurements) and new estimated experiments.



Figure S2. Structural similarity of chemicals with vapor pressure values for both previous measurements (i.e., PHYSPROP measurements) and new estimated experiments.



Figure S3. Structural similarity of chemicals with water solubility values for both previous measurements (i.e., PHYSPROP measurements) and new estimated experiments.



Figure S4. Structural similarity of chemicals with Henry's Law constant values for both previous measurements (i.e., PHYSPROP measurements) and new estimated experiments.

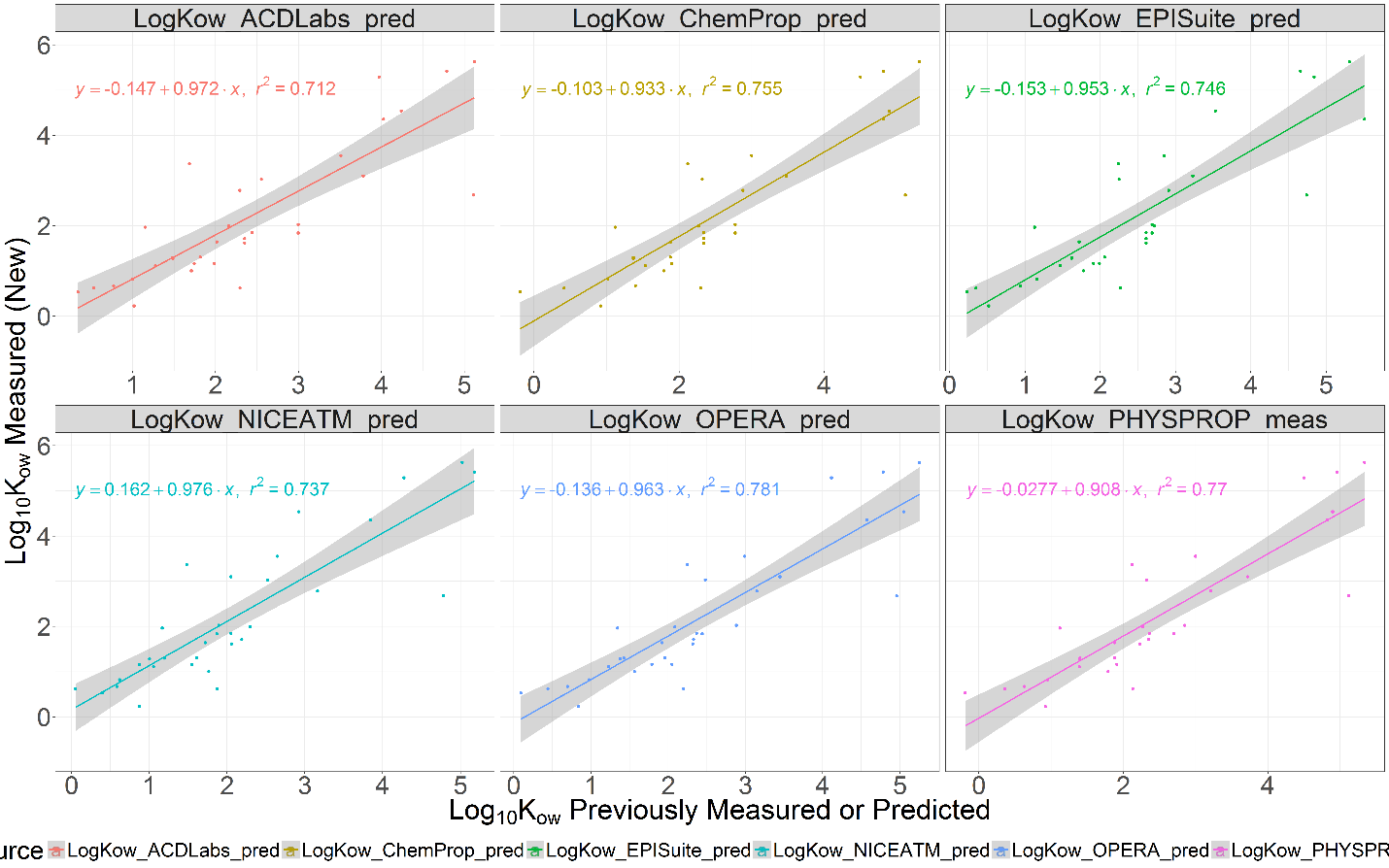


Figure S5. Plot of new experimental estimates vs predicted values along with linear regression equation, r2, and 95th percent confidence interval from A) ACD/Labs, B) ChemProp, C) EPI Suite, D) NICEATM, E) OPERA, and previous measurements from F) PHYSPROP for octanol-water partitioning coefficient (Log10(Kow)).

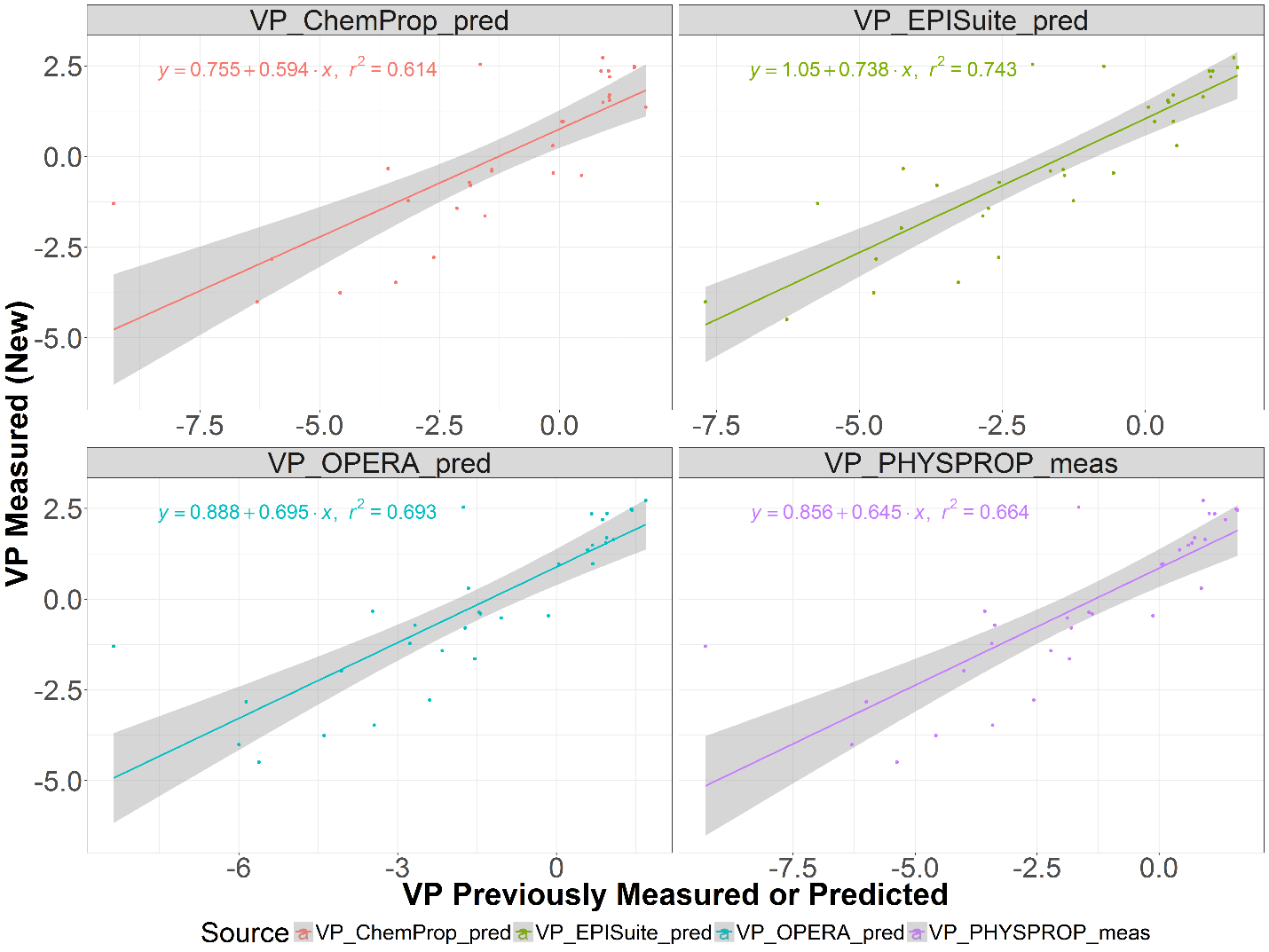


Figure S6. Plot of new experimental estimates vs predicted values along with linear regression equation, r2, and 95th percent confidence interval from A) ChemProp, B) EPI Suite, C) OPERA, and previous measurements from D) PHYSPROP for vapor pressure (VP).

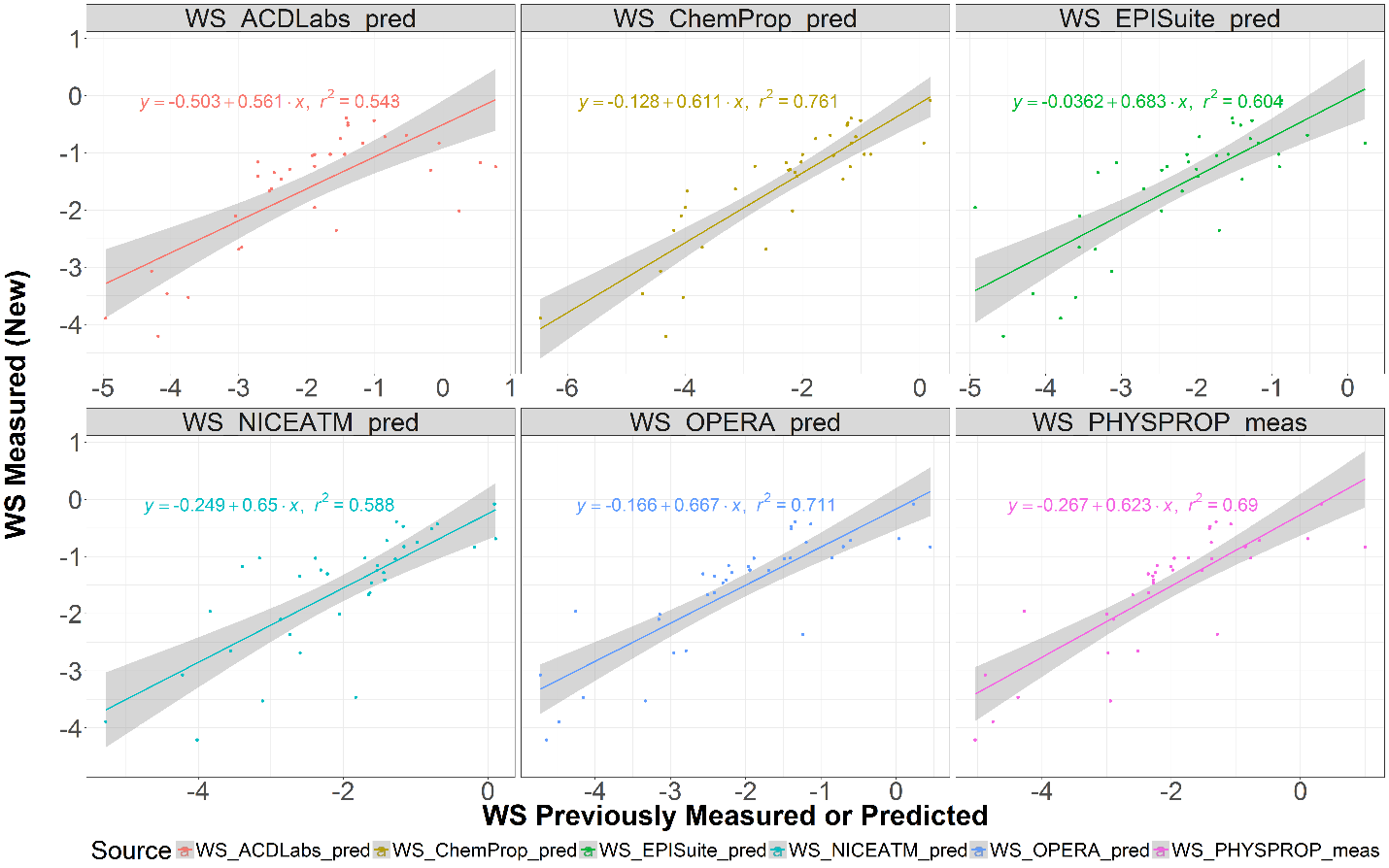


Figure S7. Plot of new experimental estimates vs predicted values along with linear regression equation, r2, and 95th percent confidence interval from A) ACD/Labs (sequoia dots), B) ChemProp, C) EPI Suite, D) NICEATM, E) OPERA, and previous measurements from F) PHYSPROP, and previous measurements from F) PHYSPROP for water solubility (WS).

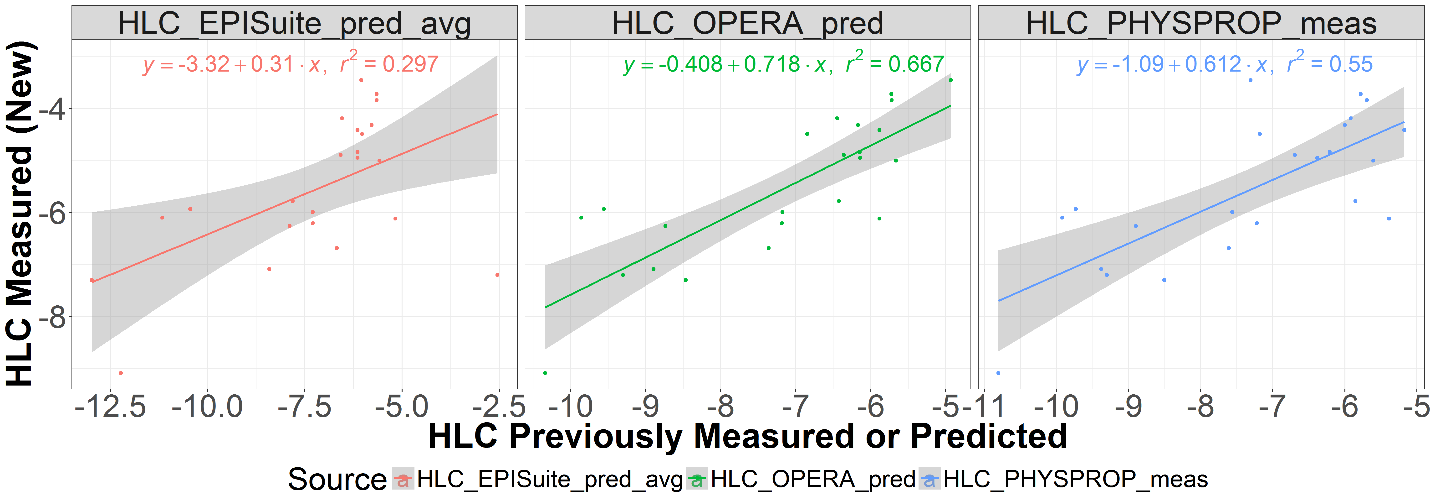


Figure S8. Plot of new experimental estimates vs predicted values along with linear regression equation, r2, and 95th percent confidence interval from A) EPI Suite, B) OPERA, and previous measurements from C) PHYSPROP for Henry’s Law Constant (HLC).



Figure S9: Heatmap representing the Tanimoto similarity matrix of all 200 chemicals investigated in this study. The color bars to the top and left side of the matrix indicate the types of experimental estimates that failed (if any). Within each type of failure, the similarity matrix has been sorted via hierarchical clustering of similarity values throughout the matrix. High similarity values (i.e., pairs of chemical structures that are very similar to one another) are shown in black, while low similarity values (i.e., pairs of chemical structures that are very dissimilar to one another) are shown in white according to the color bar at the bottom of the heatmap. The numbers of chemicals that had a specific combination of failures are shown in parentheses in the legend.

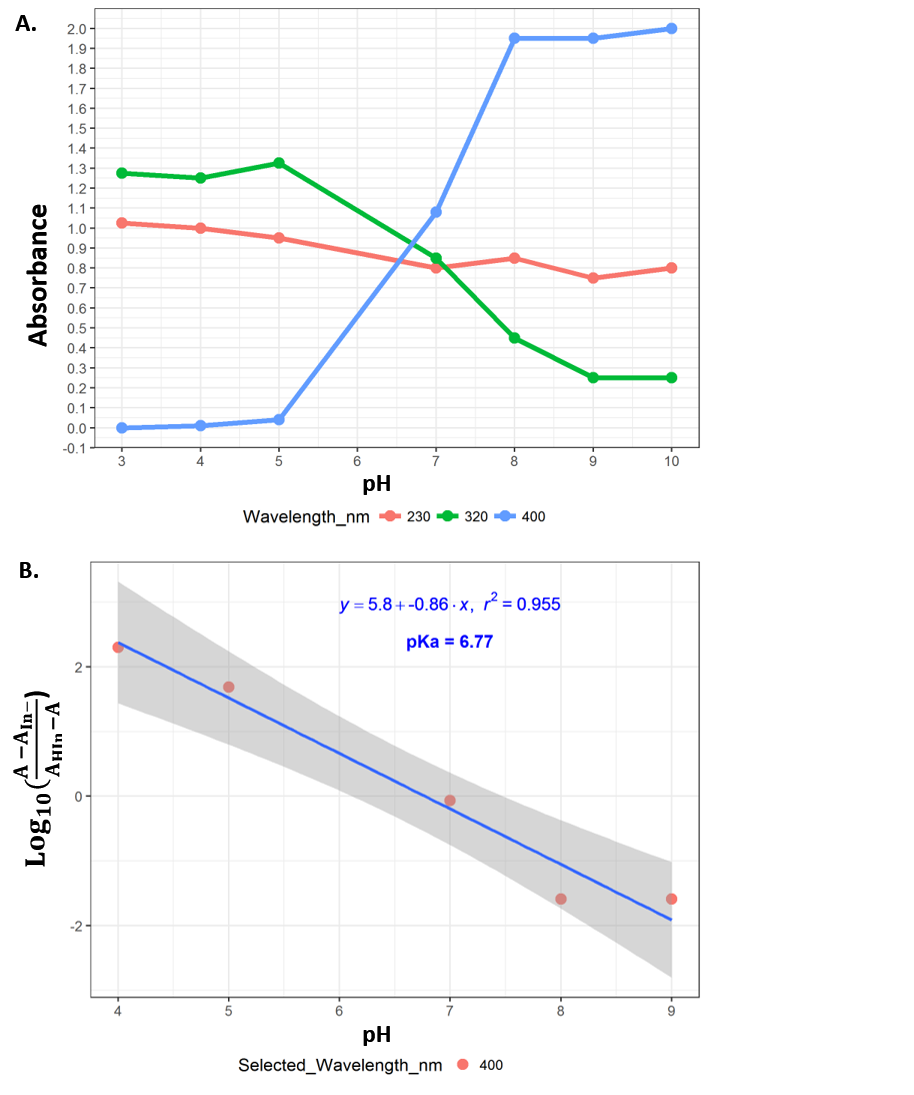


Figure S10. Illustration of wavelength selection for the determination of the acid dissociation constant (pKa): A. various wavelengths (nm) and absorbance measurements for a range of pH values and B. wavelength selection used in the determination of pKa values using a linear equation, where A in the absorbance at intermediate pH values, AHIn is the absorbance at pH = 3, and AIn- is the absorbance at pH = 10.

# SI TABLES

Table S1. Experimental estimates of log10(Kow), VP, HLC, WS, and pKa experimental estimates for 200 submitted compounds (attached spreadsheet).

Table S2. OPERA predictions for log10(Kow), VP, WS, and HLC along with confidence levels and their global and local applicability domain (attached spreadsheet).

Table S3. Physicochemical property units given by each of the seven sources where applicable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| **Data Source** | **Kow Units** | **VP Units** | **WS Units** | **HLC Units** | **pKa Units** |
| **ACD/Labs** | log(Kow) | - | - | - | † |
| **ChemProp** | log(Kow) | log(Pa) | log(mol/L) | - | - |
| **EPI Suite** | log(Kow) | mmHg | mol/L | atm-m3/mol | - |
| **New Measurements** | log(Kow) | Pa | g/mL | atm-m3/mol | † |
| **NICEATM** | log(Kow) | - | mol/L | - | - |
| **OPERA** | log(Kow) | log(mmHg) | log(mol/L) | log(atm-m3/mol) | - |
| **PHYSPROP** | log(Kow) | log(mmHg) | mol/L | atm-m3/mol | - |
| †pKavalues were derived indirectly from reported pH absorbance measurements and are always in logarithmic form. | | | | | |

Table S4. Summary of compounds that have previously measured values as per PHYSPROP along with their respective predictions for log10(Kow) in ACD/Labs, ChemProp, EPI Suite, NICEATM, and OPERA. (Attached Spreadsheet)

Table S5. Summary of compounds that have previously measured values as per PHYSPROP along with their respective predictions for VP in ChemProp, EPI Suite, and OPERA. (Attached Spreadsheet)

Table S6. Summary of compounds that have previously measured values as per PHYSPROP along with their respective predictions for HLC in EPI Suite, and OPERA. (Attached Spreadsheet)

Table S7. Summary of compounds that have previously measured values as per PHYSPROP along with their respective predictions for WS in ACD/Labs, ChemProp, EPI Suite, NICEATM, and OPERA. (Attached Spreadsheet)

Table S8. Summary of compounds that have values derived from new absorbance measurements along with their predicted pKa values in ACD/Labs. (Attached Spreadsheet)

Table S9. Acid/base determinations for the 200 pilot chemicals using ACD/Labs predictions. (Attached Spreadsheet)

Table S10. Absorbance measurements for pKa determination. (Attached Spreadsheet)

Table S11. ToxPrint fingerprints utilized for the 200 pilot chemicals based on their QSAR-ready SMILES.

Table S12. Odds ratios for experimental failures based on ToxPrint fingerprints (attached spreadsheet).

# SI REFERENCES

1. Martínez, C. H. R.; Dardonville, C., Rapid Determination of Ionization Constants (p K a) by UV Spectroscopy Using 96-Well Microtiter Plates. *ACS medicinal chemistry letters* **2012,** *4*, (1), 142-145.

2. Tomsho, J. W.; Pal, A.; Hall, D. G.; Benkovic, S. J., Ring Structure and Aromatic Substituent Effects on the p K a of the Benzoxaborole Pharmacophore. *ACS medicinal chemistry letters* **2011,** *3*, (1), 48-52.

3. Patterson, G. S., A simplified method for finding the pKa of an acid-base indicator by spectrophotometry. *J. Chem. Educ* **1999,** *76*, (3), 395.

4. Salgado, L. V.; Vargas-Hernández, C., Spectrophotometric determination of the pKa, isosbestic point and equation of absorbance vs. pH for a universal pH indicator. *American Journal of Analytical Chemistry* **2014,** *5*, (17), 1290.

5. Tanimoto, T. T., Elementary mathematical theory of classification and prediction. **1958**.

6. Yang, C.; Arvidson, K.; Richard, A.; Worth, A.; Tarkhov, A.; Ringeissen, S.; Marusczyk, J.; Gasteiger, J.; Rathman, J.; Schwab, C. In *CHEMOTYPES AND CHEMOTYPER: A new structure representation standard to include atomic/bond properties into structural alerts for toxicity effects and mechanisms*, 52nd Annual Meeting of the Society of Toxicology, 2013; 2013; pp 10-14.

7. Schmitt, W., General approach for the calculation of tissue to plasma partition coefficients. *Toxicology in Vitro* **2008,** *22*, (2), 457-467.

8. Archer, K. J. K., R. V., , Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis* **2008,** *52*, (4), 2249-2260.

9. Advanced Chemistry Development Inc. (ACD/Labs) [www.acdlabs.com](http://www.acdlabs.com) (August 11),

10. Pearce, R. G.; Setzer, R. W.; Strope, C. L.; Sipes, N. S.; Wambaugh, J. F., httk: R Package for High-Throughput Toxicokinetics. *2017* **2017,** *79*, (4), 26.

11. Wambaugh, J.; Pearce, R.; Davis, J.; Setzer, R. W.; Sipes, N., httk: High-Throughput Toxicokinetics. R package version 1.3. **2015**.

12. Ingle, B. L.; Veber, B. C.; Nichols, J. W.; Tornero-Velez, R., Informing the Human Plasma Protein Binding of Environmental Chemicals by Machine Learning in the Pharmaceutical Space: Applicability Domain and Limits of Predictability. *Journal of Chemical Information and Modeling* **2016,** *56*, (11), 2243-2252.

13. USEPA Chemistry Dashboard. <https://comptox.epa.gov/dashboard/> (August 10),

14. Pearce, R.; Setzer, W.; Davis, J.; Wambaugh, J., Evaluation and Calibration of High-Throughput Predictions of Chemical Distribution to Tissues. *in preparation* **2017**.