Supporting Information

Kostal et al. 10.1073/pnas.1314991111

Computational Approaches for Identifying Chemicals with Minimal Acute Toxicity to the Fathead Minnow (*Pimephales promelas*)

Categorization of Chemicals in Categories of Concern for Acute Aquatic Toxicity Based on Molar LC_{50} Values. The established US EPA thresholds for categories of concern for acute aquatic toxicity are as follows: <1 mg/L, high concern for acute aquatic toxicity; 1–100 mg/L, moderate concern for acute aquatic toxicity; and >100 mg/L, low concern for acute aquatic toxicity. We added a fourth category at >500 mg/L, which designates chemicals of no concern.

The average molecular weight of the 555 compounds in the fathead minnow dataset is 149.05 g/mol. Thus, the thresholds for the categories were as follows: <0.0067 mmol/L, high concern for acute aquatic toxicity; 0.0067–0.67 mg/L, moderate concern for acute aquatic toxicity; 0.67–3.3 mg/L, low concern for acute aquatic toxicity; and >3.3 mg/L, no concern for acute aquatic toxicity.

Distinctions of the Approach Presented to Categorical Quantitative Structure–Activity Relationships. The approach presented in this study shares a similarity with categorical quantitative structure–activity relationship (QSAR) models in that both attempt to identify a statistical relationship between a response variable (a set of categories of biological activity) and a set of independent variables (molecular descriptors or properties of chemicals). However, it is distinct from categorical QSAR models in three ways. First, whereas categorical QSARs aim to predict the class of biological activity (e.g., active, nonactive, sensitizer, nonsensitizer), our model aims to primarily identify one class of chemicals: those least likely to exert the unintended biological activity. Second, categorical QSAR/QSPRs models often use complex statistical approaches, such as random forest, neural network, or machine learning, to identify the classification algorithm, which implies that the relationship between the descriptors (often numerous) and response is not obvious to the user. By contrast, our approach uses only three descriptors and a simple, transparent statistical approach to derive the relationship between the independent and response variables. This leads to the third difference, which is that our approach aims to inform the design of new chemicals by providing simple and easily interpretable design guidelines, whereas categorical QSARs cannot provide such an understanding and rely on applying the model to obtain a prediction.

Because most categorical QSAR models rely on a multitude of variables, they must use efficient algorithms to perform variable selection. Among these is the stochastic QSAR sampler (SQS) method, which offers an alternative to the more traditional, stepwise regression approach of reducing the initial descriptor set to a tractable size. Because our approach uses only three variables and strives for mechanistic underpinning of these variables, it was not necessary to use SQS to identify the descriptors used in the model. Instead, variables were selected primarily based on their relevance to acute aquatic toxicity, and an exhaustive search of combinations of two or three variables was carried out to confirm the most relevant variables in the final model.

Correlations of Electronic Parameters with Acute Aquatic Toxicity by Mode of Action. To understand the reaction mechanism, apart from global properties like frontier orbital energies, we can also apply local quantities as a measure of reactivity of different sites in a molecule. Fukui indices (FIs), functional derivatives of electron density with respect to the number of electrons at constant potential, can provide such chemical information (1, 2). Broadly, FIs represent the response of the chemical potential of a system to a change in external potential and thus can be regarded as important measures of reactivity. Condensed to atoms, they are useful predictors of local susceptibility to nucleophilic, electrophilic, or radical attack. FIs for atom A in a molecule with N electrons can be computed as

$$\begin{aligned} f_A^+ &= p_A(N+1) - p_A(N) \\ f_A^- &= p_A(N) - p_A(N-1) \\ f_A^0 &= \frac{1}{2}\left[\left(p_A(N+1) - p_A(N-1)\right)\right]. \end{aligned} \qquad \textbf{[S1]}$$

From Eq. **S1**, f_A^+, f_A^-, and f_A^0 denote FIs for attacks by a nucleophile, electrophile, and radical, respectively, where p_A is the electronic population of atom A. From Pearson's hard/soft acid-base (HSAB) principle, maxima in FIs represent chemically softer regions, where electron density change is most favorable (3). These regions can be interpreted as most favorable for attack by a nucleophile, an electrophile, or a radical. Conversely, minima in FIs correspond to chemically harder regions that favor ionic interactions. Thus, one may establish the behavior of different sites in a molecule with respect to hard or soft reagents. In our calculation of FIs, orbital relaxation effects were taken into account by performing separate calculations for the ground state systems, the systems with added electron, and the systems with electron removed, all at the ground state geometry. Unphysical negative indices were disregarded.

Noncovalent interactions between a chemical and a biological target (e.g., protein–ligand binding interactions) can be described by a molecule's electrostatic potential (ESP) or $V(r)$ as

$$V(r) = \sum_A^{N_A} \frac{Z_A}{|r - R_A|} - \int \frac{\rho(r')dr'}{|r - r'|}. \qquad \textbf{[S2]}$$

From Eq. **S2**, ESP can be computed exactly for any position r from nuclei N_A with Z_A nuclear charge. Electron attraction depends on electron charge distribution $\rho(r)$, which may be obtained from the exact normalized solution for the electron Schrödinger equation of the system. For polar molecules, ESP maps are excellent predictors of charge–dipole, dipole–dipole, and quadrupole–dipole interactions. The ESP surface fit to atom-centered charges can be used to identify nucleophilic/electrophilic atoms in the molecule and estimate the strength of their Coulomb interactions. In this study, partial atomic charges were derived from quantum chemical electrostatic potentials using the Merz-Singh-Kollman (MKS) scheme (4). MKS charges are noted to be less sensitive to the choice of density functional and basis set than charges obtained from the popular density-based Mulliken, Löwdin, or natural bond orbital population analyses (5). Further, dipole moments calculated using MKS charges are on average more accurate than those derived from the aforementioned density-based models, especially for small molecules (5). Where allowed by parameterization, MKS charges were validated with ChElPG (CHG) electrostatic potential scheme (6); a strong correlation ($R^2 > 0.96$) was observed between the two charge schemes.

Thus, in addition to the property-based filters, chemical reactivity, modeled by different electronic parameters, is correlated to different toxic modes of action (MOAs) depending on the type.

Fig. S4 shows a heat map of the univariate correlations of each of the electronic parameters to $\log LC_{50}$ values. The correlation coefficients are listed in Table S2. As expected, the octanol-water distribution coefficient ($\log D_{o/w}$) has a high absolute correlation across all MOAs, ranging from 0.59 for acetylcholinesterase inhibitors to 0.88 for chemicals that act by narcosis. Similarly, ΔE has moderate but consistent absolute correlations across all MOA categories, with the highest value for acetylcholinesterase inhibition (0.71). However, when looking for information relevant to particular MOAs, there are clearly relationships between certain electronic parameters and specific MOAs. For example, partial atomic charges derived from ESP were highly correlated with $\log LC_{50}$ for chemicals that cause central nervous system (CNS) seizures, −0.85 for the maximum ESP charge, and 0.68 for the minimum ESP charge. Further, neurodepressants showed very high correlations with FIs relating to electrophilic (f^-_{max}) and radical attack (f^0_{max}): 0.90 and 0.86. (It should be noted that all of the neurodepressants in the data set are barbiturates and are thus structurally similar.) As relevant as the high correlations with $\log LC_{50}$ are the electronic parameters and MOAs where low correlations were observed. CNS seizure agents, for example, showed negligible correlation with ΔE (0.08). Acetylcholinesterase inhibitors, on the other hand, had low correlations with all FIs and ESP charges but had the highest correlation with ΔE (0.713). Narcotic compounds (excluding polar narcotics) have low correlations with all of the electronic parameters, consistent with their nonreactive mechanism of toxicity. The above correlation analyses can provide guidance to QSAR developers by identifying electronic parameters with very high univariate correlations to toxicity thresholds within specific MOA classes. Further, it substantiates the use of ΔE in the global model by showing consistently good predictive ability of this descriptor across all MOAs.

1. Parr RG, Yang WT (1984) Density functional approach to the frontier electron theory of chemical reactivity. *J Am Chem Soc* 106(14):4049–4050.
2. Yang W, Parr RG, Pucci R (1984) Electron density, Kohn-Sham frontier orbitals, and Fukui functions. *J Chem Phys* 81(6):2862–2863.
3. Mendez F, Gazquez JL (1994) Chemical reactivity of enolate ions: The local hard and soft acids and bases principle viewpoint. *J Am Chem Soc* 116(20):9298–9301.
4. Singh UC, Kollman PA (1984) An approach to computing electrostatic charges for molecules. *J Comput Chem* 5(2):129–145.
5. Marenich AV, Jerome SV, Cramer CJ, Truhlar DG (2012) *J Chem Theory Comput* 8: 527–541.
6. Breneman CM, Wiberg KB (1990) *J Comput Chem* 11:361–373.

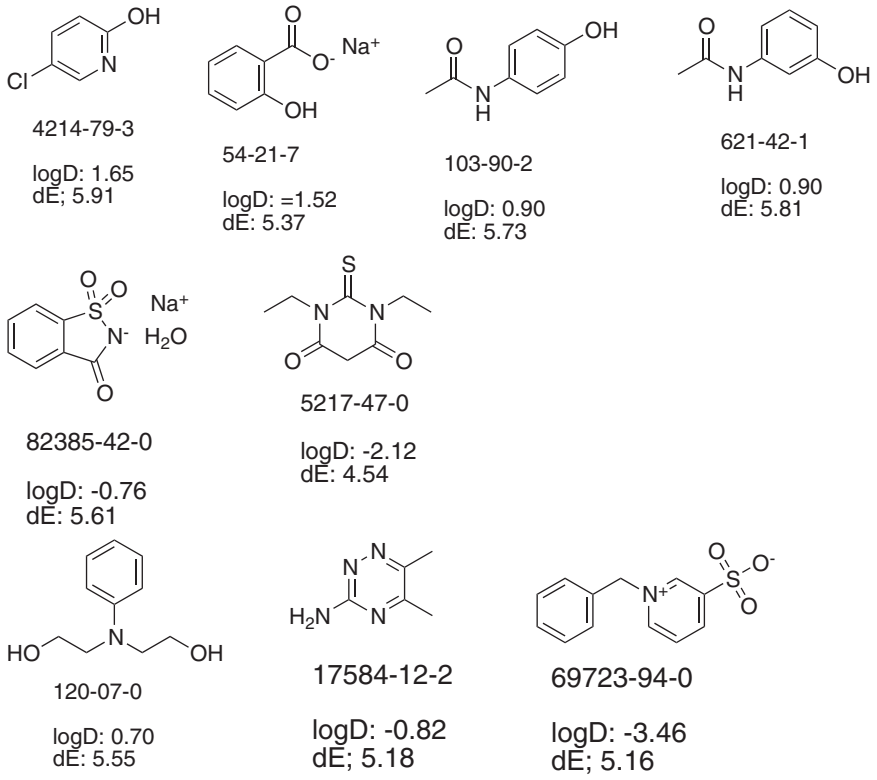

Fig. S1. List of false positives, i.e., compounds in the no-concern category that do not meet the $\log D_{o/w} < 1.7$, $\Delta E > 6$ eV, and $V < 620$ Å^3 criteria. Listed under each are the CAS numbers, $\log D_{o/w}$, and ΔE.

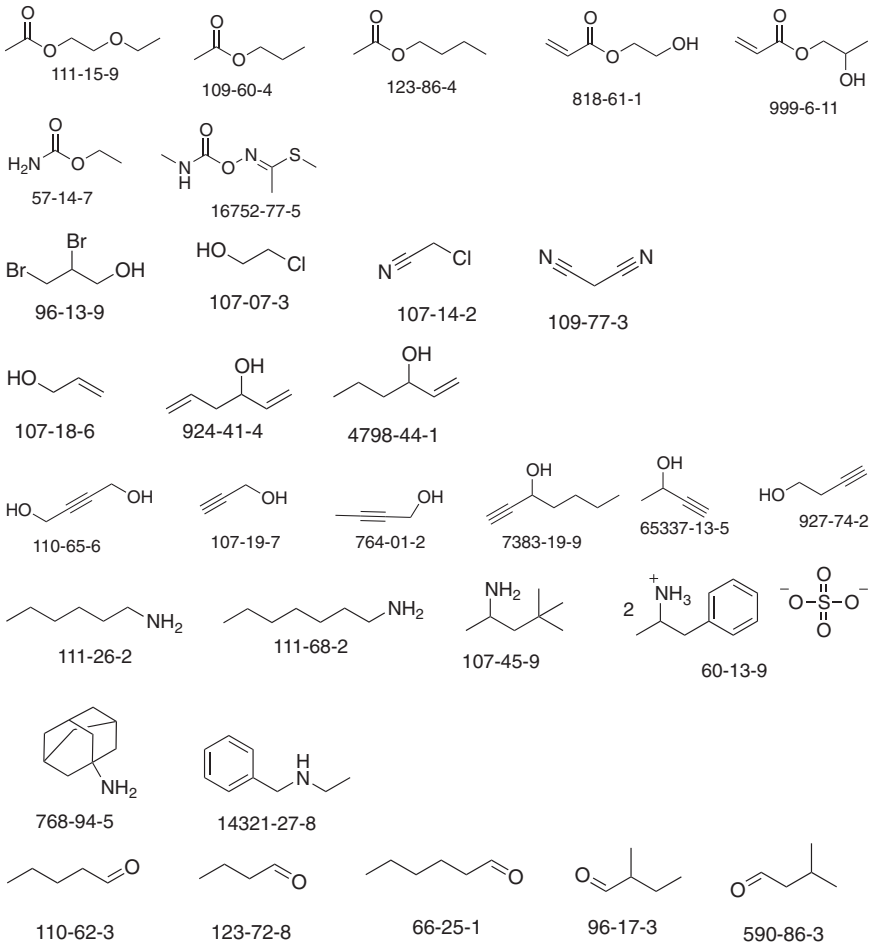

Fig. S2. List of false negatives, i.e., compounds in the high and moderate concern categories that meet the log$D_{o/w}$ < 1.7, ΔE > 6 eV, and V < 620 Å^3 criteria. Only compound 107-18-6 is in the high concern category; the rest are in the moderate concern category. Listed under each are the CAS numbers, log$D_{o/w}$, and ΔE.

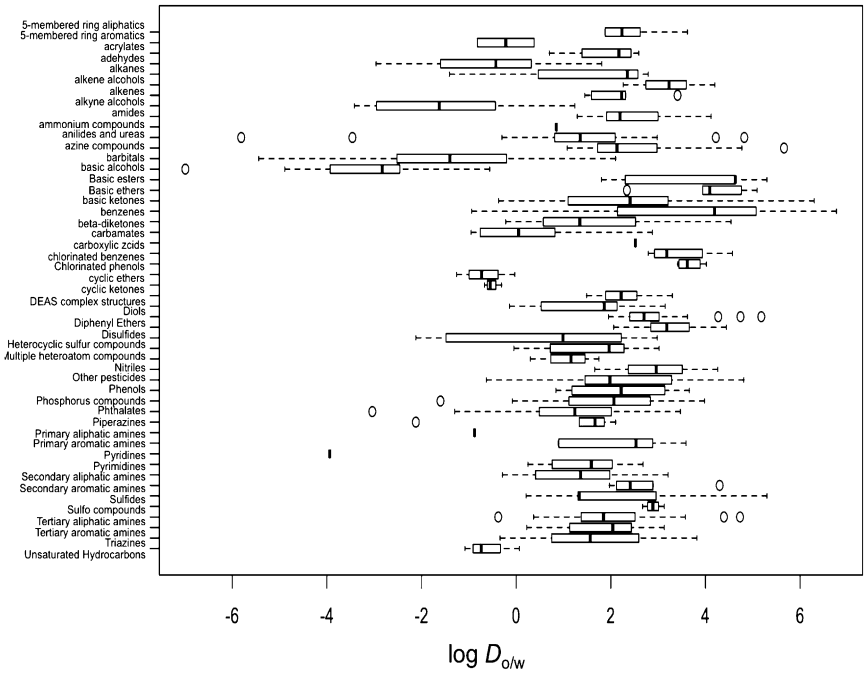

Fig. S3. Box plot of octanol-water distribution coefficients (log$D_{o/w}$) of compounds in dataset by chemical class.

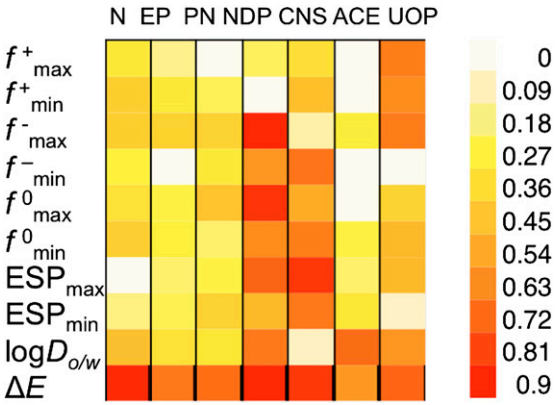

Fig. S4. Heat map of univariate correlations of electronic parameters and the octanol-water distribution coefficient at pH = 7.4, log$D_{o/w}$; the highest occupied and lowest unoccupied frontier orbitals, ΔE (eV); and the log of the mean lethal concentration, logLC_{50}, for the 555 compounds (Table S1) in the fathead minnow acute aquatic toxicity dataset and designated by MOA. *f*, Fukui index; +, −, and 0, nucleophilic, electrophilic, and radical attack, respectively; max and min, local maxima and minima in the Fukui function, respectively; ESP, most positive (max) and negative (min) atomic charge derived from the molecular electrostatic potential. MOA categories are as follows: N, narcosis; EP, electrophiles; PN, polar narcosis; NDP, neurodepressant; CNS, central nervous system seizure or stimulant; ACE, acetylcholinesterase inhibition; UOP, uncoupler of oxidative phosphorylation.

Table S1. Functional classes represented in the dataset

Chemical class	Compound no.
5-Membered ring aliphatics	3
5 Membered ring aromatics	6
Acrylates	11
Aldehydes	42
Alkanes	23
Alkene alcohols	7
Alkenes	5
Alkyne alcohols	14
Amides	13
Ammonium compounds	1
Anilides and ureas	7
Azine compounds	1
Barbitals	6
Basic alcohols	35
Basic esters	28
Basic ethers	6
Basic ketones	31
Benzenes	17
Beta-diketones	3
Carbamates	10
Carboxylic acids	9
Chlorinated benzenes	15
Chlorinated phenols	17
Cyclic ethers	13
Cyclic ketones	5
DEAS complex structures	3
Diols	3
Diphenyl ethers	4
Disulfides	4
Heterocyclic sulfur compounds	1
Multiple heteroatom compounds	9
Nitriles	16
Other pesticides	17
Phenols	30
Phosphorus compounds	5
Phthalates	7
Piperazines	7
Primary aliphatic amines	24
Primary aromatic amines	24
Pyridines	28
Pyrimidines	1
Secondary aliphatic amines	6
Secondary aromatic amines	6
Sulfides	8
Sulfo compounds	3
Tertiary aliphatic amines	9
Tertiary aromatic amines	4
Triazines	2
Unsaturated hydrocarbons	6

The chemical structures of the compounds are available through the EPA website (www.epa.gov/med/Prods_Pubs/fathead_minnow.htm).

Table S2. Univariate correlations of electronic parameters and logD with the $\log LC_{50}$ (mmol/L) by MOA

MOA correlation	N	EP	PN	NDP	CNS	ACE	UOP	All
f^{+}_{max}	0.315	0.164	—	0.229	0.351	—	0.666	0.212
f^{+}_{min}	0.412	0.309	0.241	—	0.460	—	0.626	0.361
f^{-}_{max}	0.411	0.404	0.403	−0.899	0.123	−0.293	0.671	−0.312
f^{-}_{min}	0.269	—	-0.316	0.591	0.696	0.667	—	0.276
f^{0}_{max}	0.347	0.263	0.448	−0.861	0.523	—	0.391	0.298
f^{0}_{min}	0.409	0.289	0.211	−0.625	0.668	0.264	−0.489	0.345
ESP_{max}	—	0.211	0.263	−0.735	−0.849	0.216	0.49	—
ESP_{min}	-0.180	−0.245	−0.400	0.485	0.683	0.323	0.07	—
ΔE	0.457	0.348	0.331	−0.679	−0.08	0.713	−0.60	—
$\log D_{o/w}$	−0.880	−0.680	−0.705	−0.89	−0.848	0.583	−0.729	−0.681

The correlations not listed were <0.10 (absolute value).