Supplemental Figures

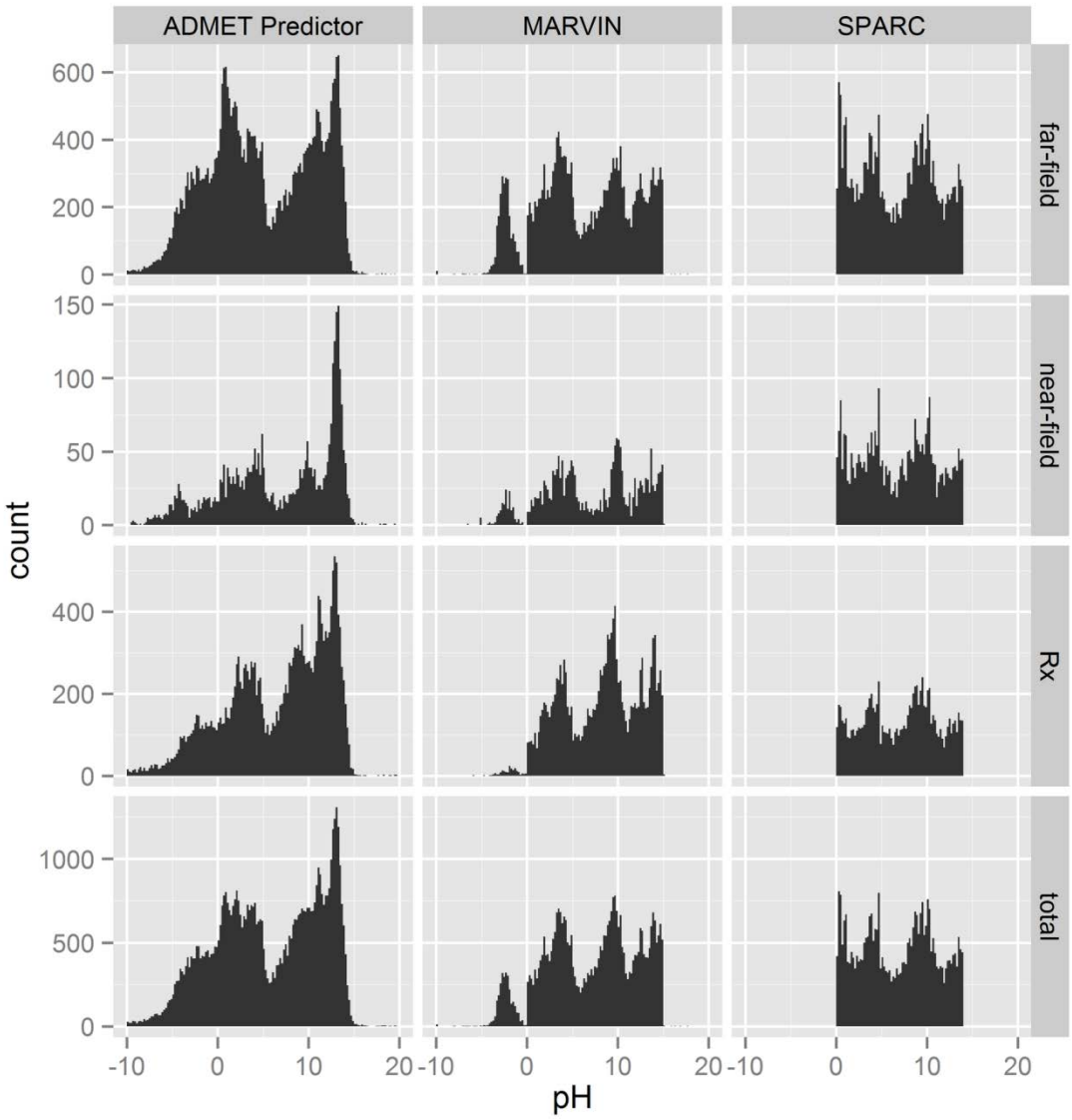
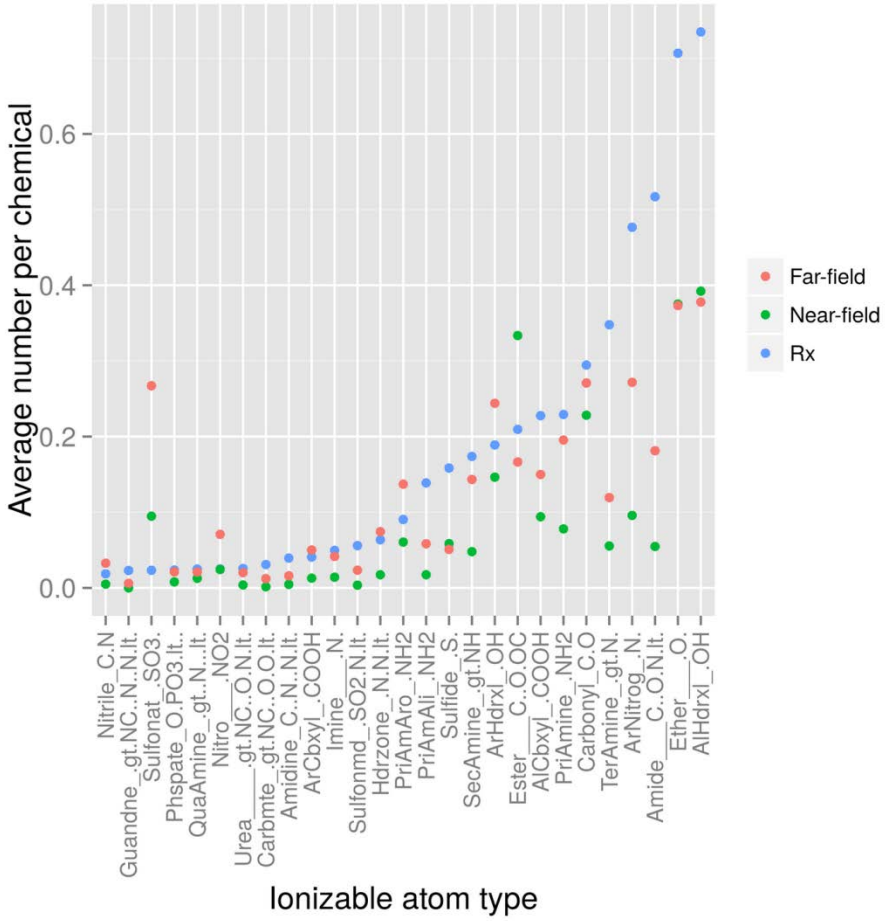


Fig. S1. Comparison of pK_a prediction distributions made by ADMET Predictor, MARVIN, and SPARC for near-field, far-field, and pharmaceutical compounds, and the entire data set.



795

796 **Fig. S2.** Average number of occurrences of the most prevalent atom types per compound for

797 pharmaceutical near- and far-field environmental chemicals, sorted by the average number of

798 each atom type that is in pharmaceutical compounds.

799

800 **Table S1.** Number of predicted neutral chemicals per chemical class for ChemAxon pK_a plug-in.

Chemical Class	Non-ionizable Chemicals	Total Chemicals	Percent Neutral
Pharmaceutical	1015	7766	13%
Near-field	2575	3888	66%
Far-field	9051	20759	47%

801

802

Appendix

Several pK_a prediction programs exist (Liao and Nicklaus, 2009). Commercial predictors span a range of mechanisms to predict the protonation state of particular atoms, including linear free energy relationships (LFER) that use a dictionary of chemical substructures (Lee et al., 2007), quantitative structure-property relationships (QSPR) (Jover et al., 2008; Palaz et al., 2012), and quantum chemical and *ab initio* methods (Bochevarov et al., 2013; Eckert and Klamt, 2006; Eckert et al., 2009; Klamt et al., 2010; Klamt et al., 2003; Vareková et al., 2011). Semi-empirical models calculate descriptors for each ionizable chemical functional group, after which pK_a values are predicted using machine learning or tree-based models (Jelfs et al., 2007; Xing et al., 2003). These semi-empirical models are limited by the number of chemicals used (Xing et al., 2003) and the usage of a proprietary, non-releasable training set (Jelfs et al., 2007).

Empirical methods employ substructure databases and use LFER to predict pK_a values based on the prior assignments for the atomic groups stored in a database. As such, their prediction accuracy is limited to the substructures contained in their database. If additional training data are available, many of these tools can be recalibrated to apply to new chemical structures. Unfortunately, such data are not available for many environmental chemicals. The data limitations of these methods will improve with the addition of more pK_a data and could be aided by efforts to contribute pK_a data that are currently underway (https://gist.github.com/egonw/5aa53abe480a8625fe81). Such is also the case with predictors using QSPR. These prediction methods have been developed using machine learning algorithms along with structural and chemical descriptors to make predictions of pK_a values (Fraczkiewicz et al., 2014; Szegezdi and Csizmadia, 2007; Szegezdi and Czismadia, 2004).

825 Quantum chemical methods and *ab initio* methods offer great promise, but currently both are
826 computationally intensive and generally do not perform as well as LFER and QSPR methods
827 (Elyashberg et al., 2010). Due to their computational inefficiency, these methods are
828 incompatible with high-throughput methodologies.

829 The majority of pK_a prediction programs inspect a particular chemical, including the
830 interplay between ionizable sites, to predict the pK_a value. Calculating the interactions between
831 sites, however, exponentially increases the computation time. In SPARC (Lee et al., 2007),
832 chemicals with complex atomic interactions can result in calculations that last weeks to months
833 for a single chemical, for which SPARC will return an incomplete calculation error (Lee et al.,
834 2007).

835