pKa predictions for arsonic acid derivatives.

Miroslava Nedyalkova[‡,§], Diana Heredia[‖], Marco Lattuada[§] and Joaquín Barroso-Flores[†,¶]*

[‡]Swiss National Center for Competence in Research (NCCR) Bio-inspired Materials, University of Fribourg, Chemin des Verdiers 4, CH-1700 Fribourg, Switzerland

[§]Department of Chemistry, University of Fribourg, Chemin du Musée 9, Fribourg 1700, Switzerland

[‖]School of Chemical Sciences and Engineering, Yachay Tech University, 100119 Urcuquí, Ecuador

[†]Centro Conjunto de Investigación en Química Sustentable UAEM-UNAM, Carretera Toluca-Atlacomulco Km 14.5, Unidad San Cayetano, Toluca, Estado de México, 50200. México

[¶]Instituto de Química, Universidad Nacional Autónoma de México. Circuito Exterior S/N Ciudad Universitaria, Alcaldía Coyoacán, Ciudad de México, CP 05410 México

E-mail: jbarroso@unam.mx

Abstract

Food, water, air, and soil are regularly contaminated with natural and artificially occurring forms of arsenic, of which arsonic acid derivatives $RAsO(OH)_2$ are the major pentavalent compounds present in aqueous media. At a given pH, the resulting ionization state for these derivatives affects their lipophilicity, solubility, protein binding, and their ability to cross plasma membranes, potentially increasing their toxicity. Knowing their pKa values not only characterizes them but also helps design a strategy for bioremediation. Numerous challenges are associated with predicting pKa, and existing models are limited to specific chemical spaces. To leverage a pKa model for arsonic acids, we contrast machine learning (ML) methods based in Support Vector Machine and three DFT-based models: correlation to the maximum surface electrostatic potential ($V_{S,max}$) at the ωB97XD/cc-pVTZ level of theory; correlation to carboxylate atomic charges in conjunction with a density-based solvation model (SMD) at the level of M06L/6-311G(*d*,*p*); and the scaled solvent-accessible surface approach, which yielded high mean unsigned errors for predicted pKa, and therefore it is not an efficient method for calculating the pKas of arsenic acids, in contrast with reported data for carboxylic acids, aliphatic amines, and thiols. The highest agreement was obtained with the atomic charges calculation on the conjugated arsonate base. ML-

based and $V_{S,max}$ models rank second and third, respectively, in terms of prediction performance.

Keywords: arsenic, arsonic acid, pKa, theoretical calculations, data science

Synopsis

Predicting pKa values for arsonic acid derivatives improves the specific design of bioremediation strategies for their extraction from water sources. We contrast various computational methods for their fast and precise calculation.

Introduction

The toxicity and contamination potential of arsonic acids are influenced by their pKa values, which determine the relative concentrations of neutral and ionized species in the environment. The pKa, or acid dissociation constant, is a critical parameter that affects how these compounds behave under various pH conditions, impacting their mobility, bioavailability, and environmental persistence. [1] ; By predicting their pKa values and their degree of ionization is crucial for predicting their behavior in soil and water, and thus designing strategies for their bioremediation[2–4] . For instance, in the natural pH range of 5.5 to 8.5, arsenite is predominantly present in its protonated, more toxic form. However, as pH decreases, arsonic acid derivatives deprotonate, acquiring a net electrostatic charge, reducing its toxicity and enhancing removal efficiency [1,5,6] . For instance, the pKa values for arsonic acids like benzoic acid, *p*-chloroaniline, 2-chlorophenol, 2,4-dichlorophenol, 2,4,5-trichlorophenol, and 2,4,6-trichlorophenol fall within the range of environmentally significant pH conditions from 6.5 to 9.5, where the dominant arsenic species will be a mixture of $H_2AsO_4^-$ and $HAsO_4^{2-}$, meaning that mixed systems will prevail under such conditions, with both species contributing to total absorption. The charge state of arsonic acids affects their adsorption to soil particles. Generally, negatively charged species (at higher pH) are less likely to absorb negatively charged soil particles, increasing their mobility and bioavailability. In contrast, neutral species are more likely to passively diffuse through cell membranes of

microorganisms and plant roots compared to charged species. This affects the uptake and bioaccumulation of arsonic acids.[7] The speciation of arsonic acids also affects their uptake by plants. For example, arsenates (As(V)) are taken up by phosphate transporters, while arsenites (As(III)) are more likely to be taken up by aquaglyceroporins.[8] Different microbial species have evolved specific transport systems for various arsenic species. The ionic state of the arsonic acid, determined by its pKa and environmental pH, influences which uptake mechanism is most effective.[9]

Phenylarsonic acid compounds have been a primary component of animal feed additives for several decades, primarily used to promote growth and control bacterial and parasitic diseases in livestock. Nitarsone and Roxarsone (molecules **24** and **26** in the present study, respectively) are widely used as organo-arsenic-based bird feed additives, which are then excreted as inorganic arsenic, causing a steep rise in the soil concentration of arsenic which can in turn also be transported into groundwater [10–12] . These compounds undergo minimal metabolism in animal bodies and are largely expelled through manure and urine. Despite the low toxicity of these drugs, they can decompose into more harmful inorganic arsenic forms, specifically arsenite and arsenate, through both biotic and abiotic processes. Therefore, pKa calculations are essential for predicting and evaluating the environmental toxicity of different types and souses for such a contaminant.

The pKa helps predict the solubility and mobility of a compound in water. Compounds with different protonation states based on their pKa values can dissolve to different extents, affecting how they spread through water systems. Highly soluble contaminants can disperse widely, potentially affecting larger areas.

The biological impact of a compound often depends on its chemical form, which is influenced by its pKa . For example, the toxic effects of a compound might be more pronounced in its dissociated form, which is more likely to interact with biological molecules. pKa calculations can help predict which form predominates under environmental conditions. pKa values provide insights into the chemical stability and reactivity of contaminants. This is important for understanding how contaminants degrade or transform into more toxic or less toxic substances over time.

Environmental pH can vary widely, and the pKa of a contaminant determines its state at different pH levels. This interaction is vital for understanding how contaminants behave in

various environmental settings, such as acidic mine drainage or alkaline lake waters.

Understanding pKa values helps in predicting how compounds interact with the environment and living organisms, forming the basis for assessing potential risks and the design of interventions.

However, accurate and quick prediction models for pKa remain a challenge. Several factors affect the accuracy of pKa models, such as the conformational flexibility of the ionizable groups, structural symmetry, unusual heterocyclic structures, multiple ionization centers, charge transfer in conjugated systems, tautomerism, and intra- and intermolecular interactions.

Various computational methods have been developed for calculating pKa values, mainly divided into two major categories: macroscopic and microscopic. As one of the macroscopic methods, continuum electrostatic methods (CE) used in predicting protonation states of proteins rely on descriptions of electrostatic potentials and parametrization of known residue values [13–16]. Other successful macroscopic empirical methods include PROpKa, although the residues used for parameterization are those located near the protein surface and have pKa values that are close to those of the model compound [17,18]. Consequently, empirical methods may not be able to predict large pKa shifts and may only be effective when it comes to residues with small perturbations. The alchemical free energy calculations based on molecular dynamics (MD) simulations, as well as the nonequilibrium (NEQ) free energy methods, have shown promise in accurately estimating the pKa [19]. The key advantage of MD-based approaches like constant pH molecular dynamics (CpHMD) is that they can explicitly sample protonation events and consider the effects of the local environment on pKa values rather than relying on empirical models. CpHMD uses either Monte Carlo sampling or λ-dynamics to dynamically update the protonation states during the simulation, allowing the free energy gradient to drive the protonation changes[20–22].

A quantum mechanical (QM) method based on density functional theory can be used to model the free energy of proton dissociation with very high accuracy; however, these free energy relationships tend to be modeled on specific functional groups (carboxylic acids, anilines, phenols, etc.) and are applied to a limited number of data sets due to the computational cost of these high-accuracy methods [23–26]. From semi-empirical QM calculations to advanced DFT calculations, the level of theory used in computation methods can significantly impact the accuracy of pKa predictions. Higher-level *ab initio* approaches provide more reliable and transferable pKa values, particularly for complex systems. Using continuum solvation model based on the quantum

mechanical charge density of a solute molecule interacting with a continuum description of the solvent (SMD) was used by Sabuzi et al. [27] for accurately predict pKa for carboxylic acid derivatives using B3LYP and CAM-B3LYP. As a result of these findings, it was demonstrated that neither complex theory nor external factors are necessary for accurate prediction of carboxylic acid pKa. Coote et al.[28] have demonstrated that thermocycle-based approaches for pKa prediction have limitations when dealing with complex organic molecules where all molecular conformations must be considered. The computational costs associated with thermodynamic cycle-based approaches for predicting pKa values can be significant. These costs arise due to the complexity of the calculations necessary to determine a chemical reaction's solution-phase Gibbs free energy, which involves multiple steps, including two geometry optimizations and determining the change in Gibbs free energy, ΔG. This process can be computationally intensive, making them unattractive for systematic conformational searches.

We have proven that pKa values for carboxylic compounds can be accurately predicted to within half a pKa unit by calculating $V_{S,max}$ values over the acidic hydrogen atoms. Similarly, calculating $V_{S,min}$ over basic nitrogen atoms allows us to predict pKb values for amines, and when used in conjunction with the previous model, isoelectric point values can be accurately predicted for amino acids. Calculated atomic charges and experimental pKa values of carboxylate fragments in their anionic form showed a good correlation [29,30] The applied SVM model was used to predict pKa values; the prediction rate was lower than that of DFT but better than that of ESP. To the best of our knowledge, no systematic study has used the DFT, ML, and ESP approaches to calculate pKa values for arsonic acids.

A useful tool for studying the charge distribution around the molecule is calculating the electrostatic potential surface (EPS). On the EPS, areas with high electron density have minimum values, the lowest of which, on the confinements of a given atom, are called $V_{S,min}$. These minimum values mean that over these regions, the electrons pass more time on average, and as the electrons have a negative charge, this results in minimum values of EPS. On the contrary, the regions with low electron density have maximum values, and the larger homologous electrostatic values are called $V_{S,max}$.

In the context of an acid-base reaction, the ease with which the proton is released from the

acid molecule is important. As the bond maintains the atoms joined in a molecule, the weaker the bond, the easier the proton release and, therefore, the more acidic the molecule is. From these facts, the relationship between the EPS value over acidic hydrogens and the pKa value of arsonic acids is studied.

In recent years, machine learning (ML) techniques have been applied to many scientific topics, including predicting pKa values. Cai and co-workers reported a deep-learning-based pKa predictor, DeepKa, trained on data generated by constant-pH simulations[31]. Reis and co-workers also reported a deep-learning-based pKa predictor, pKaI, which was trained on pKa values calculated by a continuum electrostatics method[32]. Another protein pKa prediction paper from Gokcan and Isayev introduced a new empirical scheme based on deep representation learning trained on experimental pKa data[33]. The advantages of Support Vector Machine and Cascade Deep Forest are that they could perform well on small datasets[34,35]. This is the reason why we use the SVM for our case. To gain physical insights from the ML models, we evaluate feature importance and determine the features causing pKa shifts for the explored set of 35 arsonic acids.

Herein, we report pKa calculations for 35 arsonic acid derivatives (see Figure 1) obtained by four methods. One ML method, and three DFT-based methods, namely: direct thermodynamic cycle calculation with a solvent accessible surface corrected solvation model (SMDSAS) by Smith; a multivariate regression analysis based on calculations of atomic charges on the carboxylate by Monard; a linear correlation of maximum surface potential ($V_{S,max}$) on the acidic hydrogen atom by Caballero-García. The results for all four methods are provided and compared.

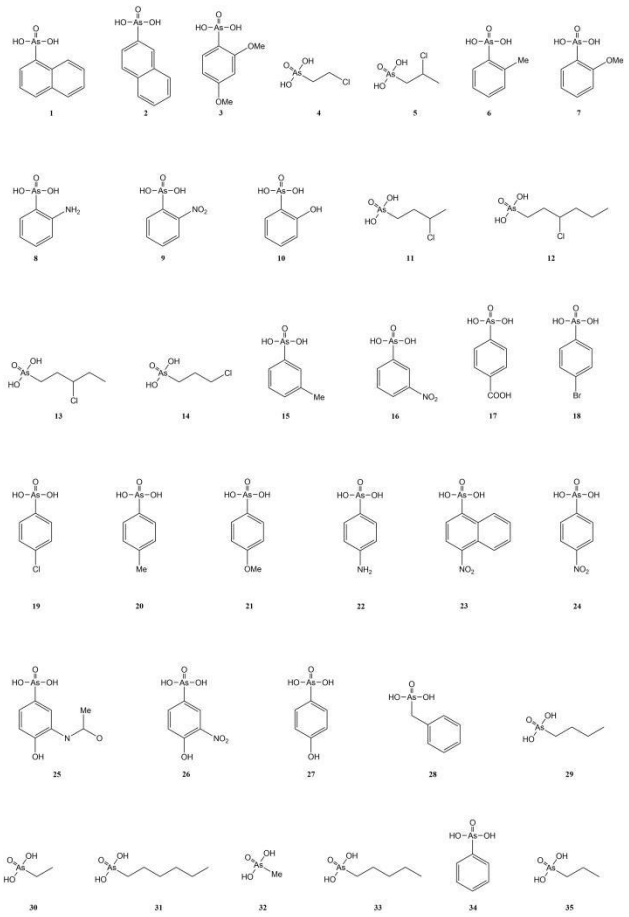

Figure 1 Chemical structures of all 35 arsonic acids under study.

Methods

To calculate $V_{S,max}$ values, the structure of the thirty-five arsonic acids in Figure 1 were optimized at the ωB97XD/cc-pvTZ level of theory using the Gaussian suite of programs[36]. To corroborate that the optimized structures corresponded to a minimum, all vibrational frequencies were calculated, too, and no negative frequencies were found. Their corresponding WFX files were then processed with MultiWFN [37] to calculate the electrostatic potential values around the molecules, thus obtaining Vs,max values for relevant atoms (vide infra), i.e., acidic hydrogen atoms in our case.

MultiWFN uses equations 1 and 2 to quantify both $\overline{V}_S^+$ and $\overline{V}_S^-$, which correspond to the average of positive and negative ESP over VdW surface[37], respectively:

$$\overline{V}_S^+ = \frac{1}{m}\sum_{i=1}^{m} V(r_i) \qquad (1)$$

$$\overline{V_S^-} = \frac{1}{m}\sum_{j=1}^{m} V(r_j) \qquad (2)$$

where *i* and *j* are index of the positive and negative regions of sampling points over the EPS. The arsonic acid group has two acidic hydrogen atoms and therefore also has two pKa values corresponding to the first and second deprotonation events. These two values are obtained from experimental measurements reported previously, and they are collected in Table 1 in which they are labeled pKa1 for the first deprotonation and pKa2 for the second deprotonation.

For the correlation with the atomic charges on the carboxylate method by Monard et al. [38] , all geometries were optimized at the M06-2X level of theory, using the SDD basis set for As and the 6-311G(*d*,*p*) basis for all remaining atoms (namely, C, H and O), using the aforementioned suite of programs. Once again, all molecules were checked to ensure that there were no imaginary frequencies. Natural Population Analysis (NPA) was performed on the resulting structures to obtain the formal charges on the atoms of interest. Using the highest, and average of the oxygen Natural Atomic Charges of the conjugate arsonate oxygen atoms fragment, were compared with the experimental pKa of the corresponding molecule. From these three, the average NPA oxygen charges yielded the best agreement with the experimental pKa values and were thus used throughout the study.

A linear equation is obtained by a least-square fit for the *Q* descriptor shown in equation 3, which is the average atomic charge of the arsonate oxygens. The predicted pKa's are computed using equation 3 (i.e., by reporting average{q(O_1), q(O_2), q(O_3),} of a given molecule into the parametrized equation).

$$Q = avg\{qO_1, qO_2, qO_3\} \qquad (3)$$

The third DFT approach developed by Lian et al.[39] used is based on the optimal description of the solute-solvent boundary, an essential component of continuum solvation models. To calculate the bulk electrostatic contribution for the default SMD model (described as SMDDefault), the solute-solvent boundary and cavity are optimized, and the scaled solvent-accessible surface (sSAS) is used to construct the cavity in the SMD continuum model. This is known as SMDsSAS. The SCRF section of SMDsSAS allowed simultaneous tuning of the surface type and scaling factor options. When SAS was chosen as the solute-solvent boundary, the solvent radius (1.385 x water) was added to the intrinsic Coulomb radii to construct the cavity. From 0.4 to 0.8, the scaling factor was

introduced to adjust the size of the SAS cavity, which can be used to calculate the solvent-accessible surface area (SASA) of individual molecules.

In the case of the ML methods selected, SVM is a supervised learning approach developed by Cortes and Vapnik [40] with sparse generalization applicability and widely used in drug and material design [41–43]. The SVM, over the other algorithms, has an advantage in that it can be used for small datasets. All SVM calculations in this work were conducted by using the AlvaModel[44]. The test and train sets are selected randomly with a test: train ratio equal to 1:4.

Results and Discussion

$V_{S,max}$ correlation

The two highest maximum values of the EPS around the acids are located just in front of the acidic hydrogen atoms. The highest maximum value was labeled as $V_{S,max1}$, and the second highest maximum value was labeled as $V_{S,max2}$, as shown in Table 1.

Table 1 $V_{S,max1}$ (eV), $V_{S,max2}$ (eV), experimental pKa1 and pKa2 values for all thirty-five arsonic acid derivatives used in this study.

N°	Compound name	$V_{S,max1}$ (eV)	$V_{S,max2}$ (eV)	pKa1	pKa2
1	1-Naphthyl arsonic acid	2.13	2.10	3.66	8.66
2	2-Naphthyl arsonic acid	2.20	2.07	4.2	8.46
3	2,4-Dimetoxyphenyl arsonic acid	1.86	1.86	4.35	9.55
4	2-Chloroethyl arsonic acid	2.38	2.36	3.68	8.37
5	2-Chloropropyl arsonic acid	2.28	2.06	3.76	8.39
6	2-Methylphenyl arsonic acid	2.11	2.11	3.82	8.85
7	2-Methoxyphenyl arsonic acid	1.94	1.94	4.08	9.40
8	2-Aminophenyl arsonic acid	2.32	2.04	3.79	8.93
9	2-Nitrophenyl arsonic acid	2.13	2.12	3.37	8.54
10	2-Hydroxyphenyl arsonic acid	1.97	1.97	4.00	7.92
11	3-Chlorobutyl arsonic acid	2.20	2.19	3.95	8.85
12	3-Chlorohexyl-1-arsonic acid	2.20	2.17	3.51	8.31
13	3-Chloropentyl-1-arsonic acid	2.20	2.17	3.71	8.77
14	3-Chloropropyl arsonic acid	2.29	2.25	3.63	8.53
15	3-Methylphenyl arsonic acid	2.17	2.06	3.82	8.60
16	3-Nitrophenyl arsonic acid	2.50	2.39	3.41	7.80
17	4-Arsenobenzoic acid	2.34	2.23	4.22	8.44
18	4-Bromophenyl arsonic acid	2.34	2.22	3.25	8.19
19	4-Chlorophenyl arsonic acid	2.33	2.21	3.33	8.25
20	4-Methylphenyl arsonic acid	2.16	2.04	3.70	8.68
21	4-Methoxyphenyl arsonic acid	2.14	2.00	3.79	8.93
22	4-Aminophenyl arsonic acid	2.06	1.92	4.13	9.19
23	4-nitronaphthalen-1-yl-1-arsonic acid	2.40	2.39	-	7.87
24	4-Nitrophenyl arsonic acid	2.52	2.42	2.90	7.80
25	3-acetylamino-4-hydroxyphenyl arsonic acid	2.19	2.31	3.78	7.9

26	4-Hydroxy-3-nitrophenyl arsonic acid	2.43	2.32	3.46	
27	4-Hydroxyphenyl arsonic acid	2.17	2.04	3.89	8.37
28	Benzyl arsonic acid	2.14	2.14	3.81	8.49
29	Butyl arsonic acid	2.10	2.10	4.23	8.91
30	Ethyl arsonic acid	2.12	2.12	3.89	8.35
31	Hexyl arsonic acid	2.09	2.09	4.16	9.19
32	Methyl arsonic acid	2.16	2.16	3.41	8.18
33	Pentyl arsonic acid	2.09	2.10	4.14	9.07
34	Phenyl arsonic acid	2.21	2.09	3.47	8.48
35	Propyl arsonic acid	2.11	2.11	4.21	9.09

All four correlations between $V_{S,max1}$, $V_{S,max2}$, and pKa1, pKa2 were evaluated after obtaining the fitted linear equation and their r^2 correlation coefficient for each case. Next, these equations were used to predict the pKa of the same acids. Then, compared to the predicted pKa and experimental pKa, the absolute error and the mean absolute error were calculated. Finally, a cross-validation test and graphical plot of residuals were carried out to prove the method's robustness.

To determine what pKa data correlated better with the maximum value of EPS, all possible correlations between them were explored, and the results are summarized in Table 2. The second column corresponds to the experimental pKa1 value. In contrast, the third column is the calculated pKa1 obtained by using the $V_{S,max1}$, and pKa1 correlation equation; the fourth column is the pKa1 calculated using the $V_{S,max2}$, and pKa1 correlation equation; the fifth column is the experimental pKa2 value; the sixth column is the pKa2 calculated using the $V_{S,max1}$ and pKa2 correlation equation; the seventh is the pKa2 calculated using the $V_{S,max2}$ and pKa2 correlation equation.

The absolute error between the experimental and calculated pKa has values between 0.00 to 1.04, indicating that in some acids, the predicted value is equal to the experimental value, and in other cases, the calculated values move away up to 1.04 pKa units with these

correlations. The mean absolute error (MAE) tells us, on average, how far our calculated values are from the experimental values. MAE for $V_{S,max1}$, and pKa1 relationship is 0.18, and MAE for $V_{S,max2}$, and pKa1 relationship is 0.20; therefore, the value of $V_{S,max1}$ correlated with pKa1 predicts better values for pKa1 than the $V_{S,max2}$ with pKa1 correlation. In the same way, MAE for $V_{S,max1}$, and $V_{S,max2}$ correlated with pKa2 is 0.25, meaning that both correlation equations predict the values of pKa2 with similar accuracy.

Table 2 Experimental pKa1 and pKa2 values and predicted pKa values pKa1-cal and pKa2-cal for each correlation.

N°	pKa1	$pKa1_{calc}$	$pKa1_{calc}$	pKa2	$pKa2_{calc}$	$pKa2_{calc}$
1	3.66	3.87	3.82	8.66	8.70	8.65
2	Will check	-	-	8.46	8.55	8.72
3	4.35	4.28	4.19	9.55	9.27	9.23
4	3.68	3.48	3.41	8.37	8.17	8.03
5	3.76	3.63	3.89	8.39	8.39	8.75
6	3.82	3.89	3.80	8.85	8.74	8.63
7	4.08	4.16	4.07	9.40	9.10	9.03
8	3.79	3.57	3.91	8.93	8.29	8.80
9	3.37	3.87	3.78	8.54	8.71	8.59
10	4.00	4.11	4.02	7.92	9.03	8.96
11	3.95	3.76	3.67	8.85	8.55	8.43
12	3.51	3.76	3.72	8.31	8.56	8.49
13	3.71	3.75	3.71	8.77	8.55	8.49
14	3.63	3.62	3.58	8.53	8.37	8.28
15	3.82	3.80	3.89	8.60	8.62	8.75
16	3.41	3.29	3.37	7.80	7.92	7.96
17	4.22	3.54	3.61	8.44	8.26	8.33
18	3.25	3.54	3.63	8.19	8.26	8.37

19	3.33	3.55	3.64	8.25	8.27	8.38
20	3.70	3.82	3.92	8.68	8.64	8.80
21	3.79	3.85	3.97	8.93	8.69	8.88
22	4.13	3.98	4.10	9.19	8.85	9.08
23	-	-	-	7.87	8.14	7.95
24	2.90	3.26	3.32	7.80	7.87	7.88
25	3.78	3.78	3.49	7.9	8.58	8.15
26	3.46	3.39	3.48			
27	3.89	3.80	3.91	8.37	8.62	8.78
28	3.81	3.85	3.76	8.49	8.68	8.56
29	4.23	3.91	3.82	8.91	8.76	8.65
30	3.89	3.88	3.79	8.35	8.72	8.60
31	4.16	3.92	3.83	9.19	8.78	8.67
32	3.41	3.81	3.72	8.18	8.63	8.50
33	4.14	3.92	3.82	9.07	8.78	8.66
34	3.47	3.74	3.83	8.48	8.53	8.67
35	4.21	3.90	3.81	9.09	8.75	8.64

Figure 2 shows all linear correlations graphically with their respective fitted equations and correlation coefficient (r^2) values. As previously mentioned, these equations are used to predict the pKa, as shown in Table 2. The predicted values have sufficiently high accuracy to be used. Based on the MAE and the r^2 values, the correlations between $V_{S,max1}$ and pKa1, and between $V_{S,max2}$ and pKa2 were chosen to continue the analysis, because they have the lowest MAE values and the highest r^2 values.

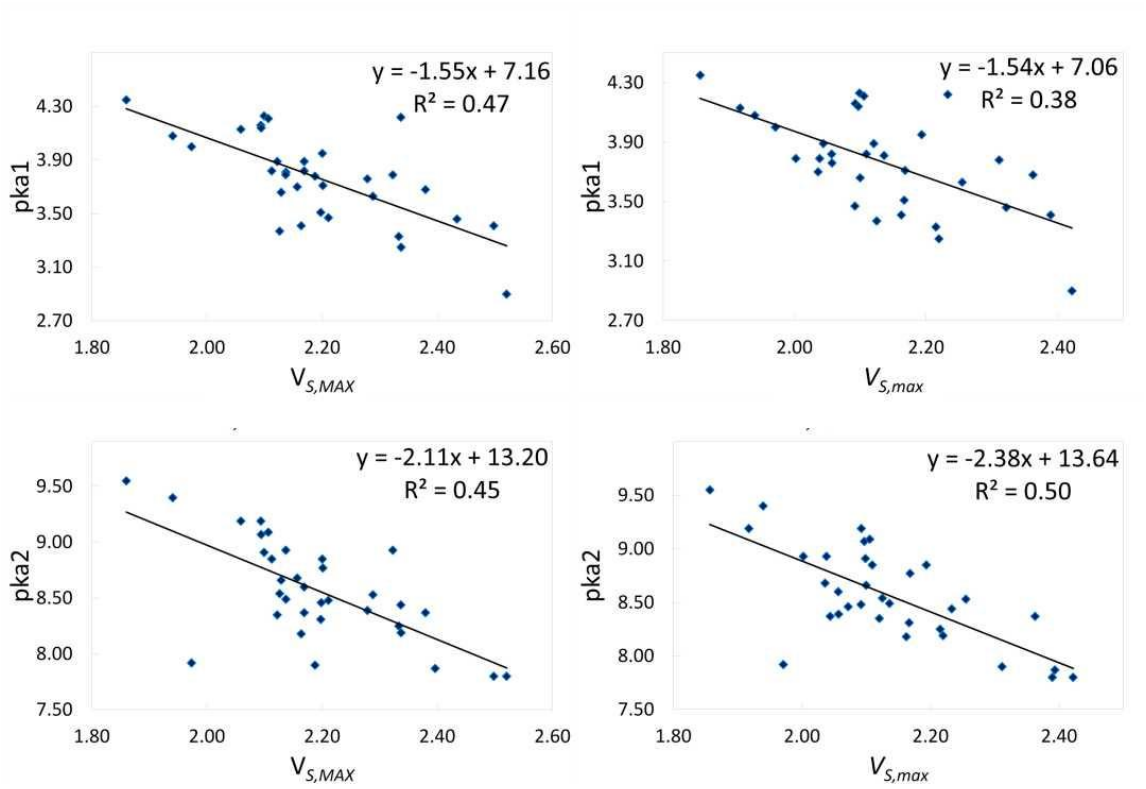

Figure 2 Correlation of $V_{S,max1}$ (top row) and $V_{S,max2}$ (bottom row) vs pKa1 and pka2.

The residual plots were studied to determine the possible reason and solution for the low value of R^2 (see Figure 3); in these graphs, the residuals or the error between the experimental and theoretical values were plotted. No trends or agglomerate data are needed in residual plots to support the hypothesis of a linear correlation for the data. Fortunately, the residual data are randomly dispersed in the plot, which means the linear model is adequate for this data. Moreover, in the residuals, it is possible to appreciate what values have the greatest error and, in some cases, treat that data as an outlier and, because of that, delete it from the whole set. An additional statistical test to ensure the data follow a linear correlation is the *F* test, which analyses the fit applied to a set of data under the acceptance or rejection of the null or alternative hypothesis; in these cases, the calculated *F* is higher than the critical one, supporting once again that the data have a linear correlation which is the alternative hypothesis.

However, it is insufficient to prove that the data follow a linear correlation to use them as a model to predict values. The cross-validation test is very helpful in creating a robust predictive

model (see Figure 3). The cross-validation method tells that if we subtract randomly a small subset of the whole set and do the same procedure to predict the values, the new predicted values must have the same accuracy as the predicted values with the correlation obtained from all data. This result helps us to ensure that the predicted results are independent of the partition between training and test data. Therefore, ten of the thirty-five values were randomly eliminated. Next, the data were plotted, obtaining new correlation equations used to predict the pKa1 and pKa2. Similar results were found using the complete set of data and the data selected on the cross-validations test. The R^2 value, in this case, improves, but the predicted values have the same accuracy.

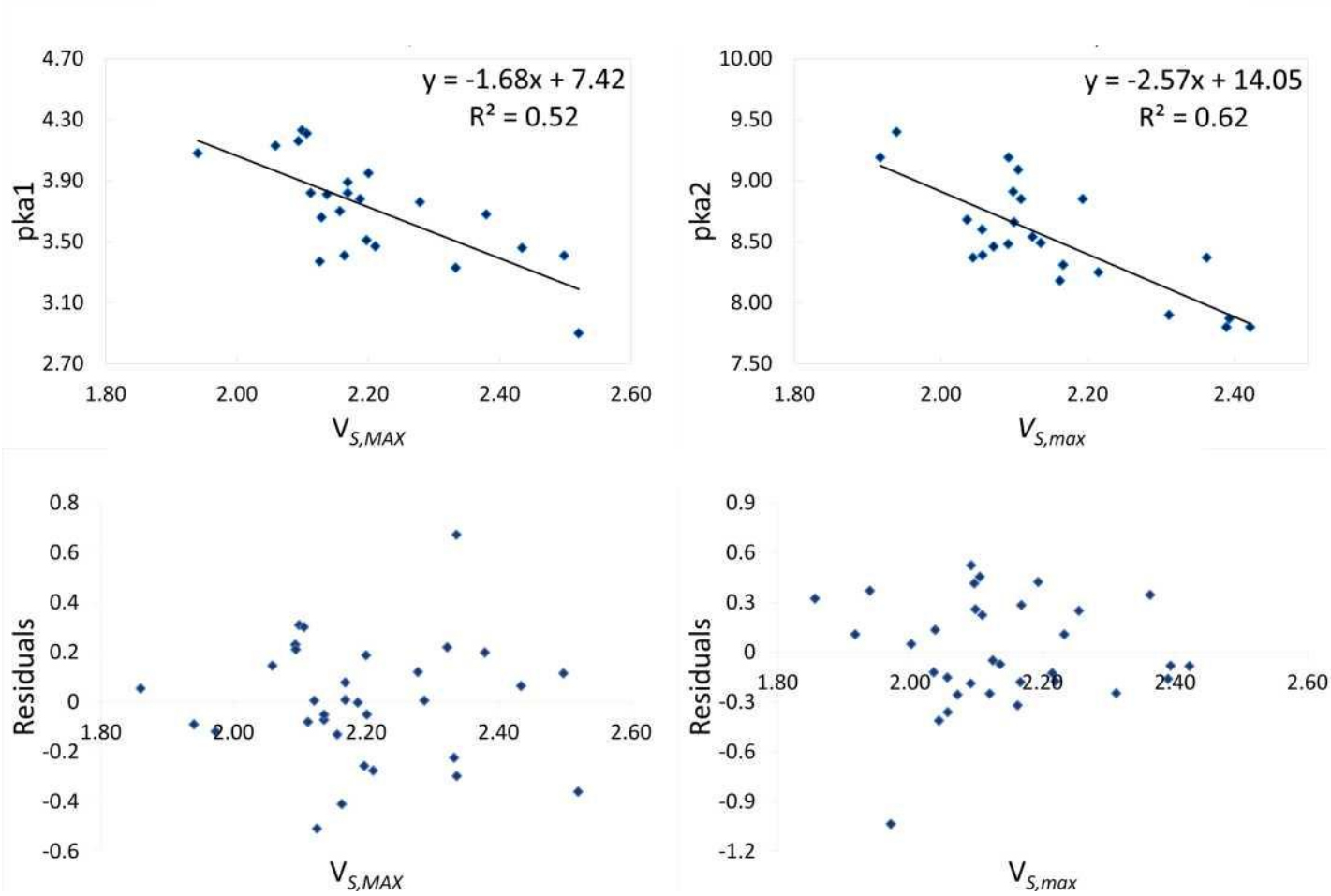

Figure 3 Residual and cross-validation plots.

Scaled solvent-accessible surface (sSAS) approach with SMD model

Using the default SMD model for the arsonic acids set, MAE was used to evaluate the accuracy of the prediction rate for both aug-cc-pvdz and M06-2X for the arsenic and compared with the obtained cavity values. In the work of Smith et al.[39] have shown that the accuracy of calculated pKas with DFT and the SMD model by tuning the scaled solvent-accessible surface (SAS) approach for constructing the solute-solvent boundary can improve the prediction for the case of carboxylic acids, amines, and thiols. We performed an optimization search to determine the optimal value of scaling SAS factor - α by minimizing the MAE for α (the range was 0.3 – 0.8). The pKa was calculated with the scaled value of alfa and as well without by the direct method based on SMD (default) one by the Arrhenius equation $\mathrm{pKa} = \frac{\Delta G_{aq}}{2.303RT}$

The prediction performance of the two SMD models (default and scaled based on solvent-accessible surface) on the data sets was assessed by MAE and was presented in Figure 4. Unfortunately, the prediction rate is unacceptable for neither of the chosen approaches based on SMD. The obtained pKa for the arsonic acid data vs experimentally measured pKa values are presented in the SI. The SMD variants without an effect of the scaling factor based only on SDD performance for this data set were better, but they were far from having a reasonable prediction rate.

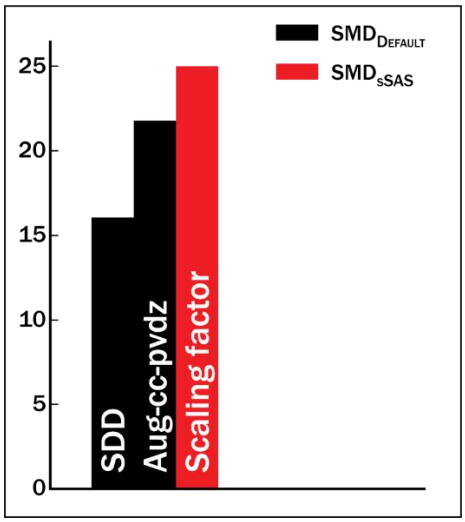

Figure 4 MAE and standard deviations for pKas of arsonic acids. The SDD basis set for Arsenic and 6-31+G(d,p) basis set for H, O, and C were used for all the calculations.

The pKas calculated with different SMD variants show no correlation with the corresponding experimental values, with R^2 values of 0.45 for the SMD default model with SDD basis set. However, for the set of arsonic acids, the obtained results are insufficient to use the two approaches compared to the published results for some other acids.

Atomic charges on the conjugated base

The next approach was linking partial atomic charges to pKa experimental data. Monard reported a benchmarking study based on NPA charges computed using the CPCM solvation model with the B3LYP/3-21G level. This resulted in the most accurate combination for reproducing experimental pKa values for alcohols, thiols, and amino acids. While other charge models, such as Mulliken,

Löwdin, and AIM charges, can also be used to predict pKa, the NPA charge scheme consistently outperforms these other methods[45].

Our study utilized the NPA charges on the oxygen to estimate pKa values, a method that demonstrated very high correlation coefficients (See Figure 5). The models for the obtained correlations for the average NPA distribution on the oxygen atoms were presented in the SI (Table S1). The NPA charges, a key aspect of our research, play a crucial role in understanding the acidity of the compounds and their potential environmental impact.

The NPA was used to calculate atomic charges. The methodology for revealing the linearity of the relationship between experimental pKa's and atomic charges was inspired by the work of Ugur et al.[38] and Monard et al.[25]

Using the SMD implicit solvent model, the average charge on the oxygen of each arsonic fragment was computed with NPA at the M06-2X/SDD (for the Arsenic atom).

According to Ugur et al.[38], the negative charge of carboxylate can be shared between two oxygen atoms and two carbon atoms, as opposed to alcohols and thiols. The atomic charges for this fragment can be extracted in various ways and then compared with experimental pKa values using Equation 3.

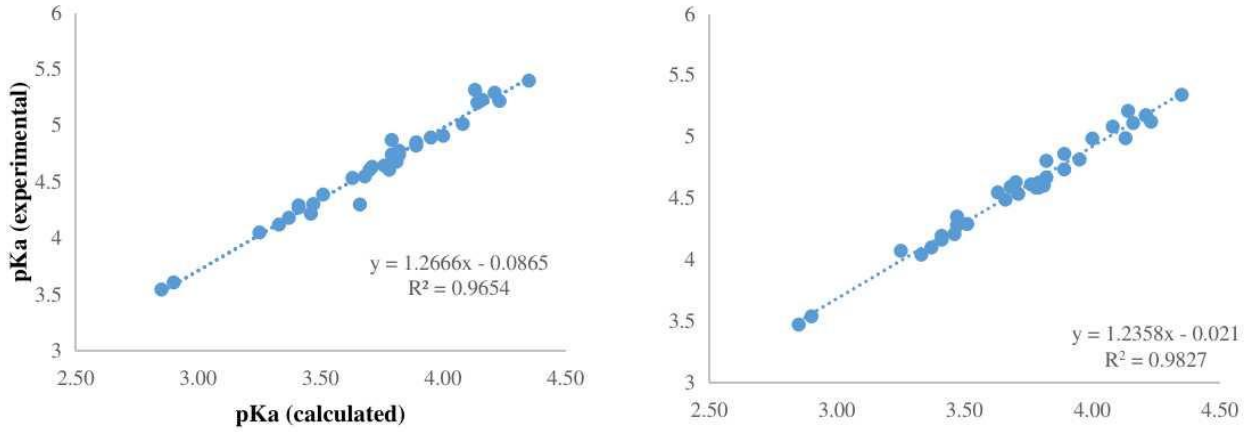

Figure 5 Linear correlation of experimental pKa's and values calculated using M06-2X and 6-311G(d,p) with SMD based on: a) average oxygen charges and b) on oxygen charges for the monoprotonated form (anion)

Our present protocol for obtaining accurate and fast pKa predictions for a limited set of arsonic acids unveils a new pattern in the charge extraction scheme, setting our study apart from another

available research. The linear regression was evaluated between the experimental pKa's and the average NPA atomic charges on the ionizable OH groups. The best combination of DFT functionals and basis sets is M06-2X/SDD with SMD default model. The next goal would be to transfer the suggested extracted scheme of charges protocol to a set of tri-carboxylic acids by calculating the average atomic charge of the carboxylate form into the tricarboxylic acids, such as hemimellitic acid (1,2,3-benzene tricarboxylic acid), trimellitic acid (1,2,4-benzene tricarboxylic acid), and trimesic acid (1,3,5-benzene tricarboxylic acid). These tricarboxylic acids can act as environmental pollutants due to their acidic nature and potential toxicity. The pKa values for these tricarboxylic acids vary depending on the positions of the carboxyl groups on the benzene ring. The other test group will be PFOA (perfluorooctanoic acid), which has a significantly different pKa value at the air-water surface compared to the reported bulk pKa values. Focusing our interest on accurate pKa values of PFAS is essential for understanding and modeling their environmental fate and transport, as the pKa determines the speciation and behavior of these persistent pollutants.

Machine Learning Approach for pKa Prediction

The datasets with the experimental values of the pKa1 values were modeled using the SVM algorithm, and the results are shown in Figure 6. For the deprotonated form, the set of descriptors for each compound was generated by Alvadesc [46,47] from a 3-dimensional conformation. A set of molecular descriptors has been calculated for the first ionization form of the arsonic acids. The initial descriptor space was based on 4,000 descriptors, which are categorized into 22 classes, including constitutional descriptors, topological descriptors, walk and path counts, connectivity indices, information indices, 2D autocorrelations, edge adjacency indices, Burden eigenvalues, topological charge indices, eigenvalue-based indices, Randic molecular profiles, geometrical descriptors, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, functional group counts, Ghose-Crippen atom-centered fragments, charge descriptors, molecular properties, 2D binary fingerprints and 2D frequency. One hundred forty-five descriptors were selected to be used in this QSPR modeling. We used the genetic algorithm-multi linear regression (GA-MLR) approach to select features. Table S2 in the Supporting information section presents the selected molecular descriptors with the highest obtained score with a maximum value for R^2_{adj}, based on various molecular descriptors ranging from one to twenty.

An SVM model was used to predict the pKa using the best combination of selected descriptors presented in Table S2 in SI. The predicted pKa values derived from this machine learning method are plotted vs. the experimental values in Figure 6. The applied selected set of descriptors can predict the pKa values. In the presented plots, outliers are highlighted with red dots for the molecules biased by the main trend. The presented R^2 values for each of the combinations of the descriptors are shown in Figure 6.

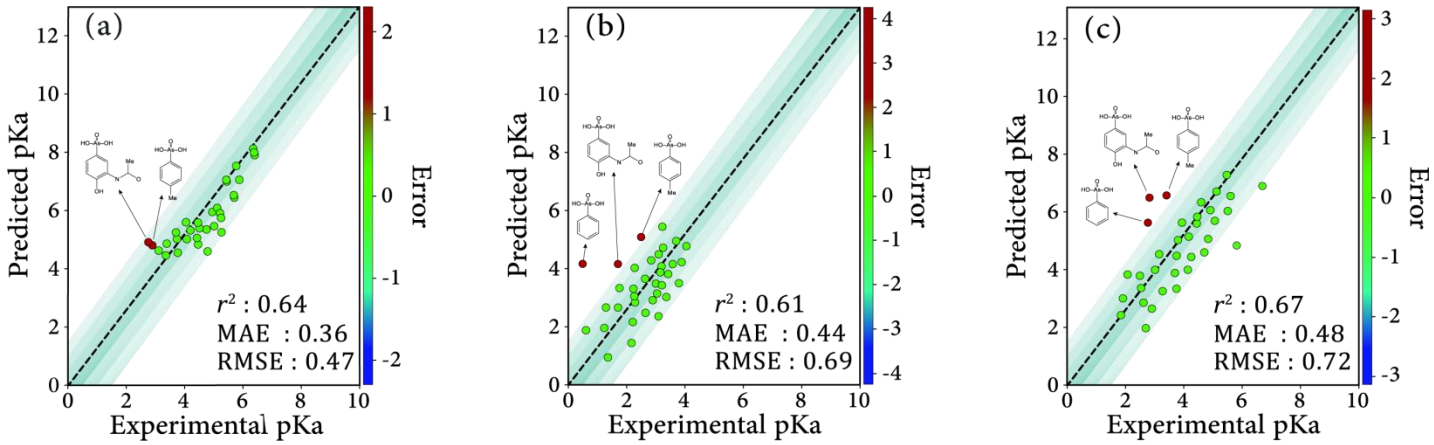

Figure 6 Comparison between predicted and experimental pKa values based on the SVM model for the selected combination of the descriptors.

These descriptors are related to the pKa in the following manner. The McGowan volume (Vx) is one of the parameters that can be used to evaluate and predict the pKa values of compounds. It is considered an Abraham descriptor, which is a type of molecular descriptor used in QSAR modeling [48,49] . For example, one study used the McGowan volume and other descriptors like log Kow to model the pKa values of chlorinated phenols. The McGowan volume appears to be a molecular descriptor related to and can be used to predict pKa values. Using the Vx descriptor can accurately predict pKa based on the leverage models. The McGowan volume provides a compact and easy-to-interpret representation of molecular size and shape, which can be helpful for pKa prediction [50] . Connectivity indices (X3Av), also known as branching indices, are a class of topological descriptors that capture information about the molecular structure and connectivity [51–53]. The X3AvX3 is a numerical descriptor that encapsulates information about the molecular graph

(the way atoms are connected) of a compound. This index helps in predicting the physicochemical properties, biological activities, and chemical reactivity of molecules based on their structure. These indices quantify the degree of branching and connectivity in a molecule, which can be relevant for predicting properties like pKa. The used 2D Autocorrelations (ATS1m) as molecular descriptors are one of the types of molecular descriptors that can capture information about the distribution of specific properties (e.g., atomic properties, bond properties) along the 2D molecular structure. This suggests that 2D autocorrelation descriptors can provide relevant information for predicting the pKa values of molecules, as they capture structural features that influence the acid-base behavior. The 2D autocorrelations and other molecular descriptors are commonly used as inputs for developing QSAR and QSPR models to predict pKa and other properties.

TPSA, or topological polar surface area of a molecule, is a critical measure defined as the total surface area of all polar atoms within a molecule [54] . The polar surface area has been used in medicinal chemistry to optimize a drug's potential to permeate cells[55,56], and it is considered a main descriptor to evaluate the blood-brain barrier penetration[57]. According to Lipinski's Rule of Five, an orally active drug typically has a TPSA less than 140 Å² and a pKa value that ensures the molecule is not too ionized at physiological pH. Adjusting TPSA and pKa helps achieve a balance between solubility and permeability, enhancing oral bioavailability. Compounds with a TPSA less than 90 Å² are more likely to cross the blood-brain barrier. Incorporating pKa values helps predict the ionization state at physiological pH, which affects

Furthermore, the MLOGP[58] descriptor plays a significant role in quantitative structure-activity relationship (QSAR) models, where it is frequently employed to predict the permeability of compounds across the blood-brain barrier. This is particularly important for developing drugs to treat neurological conditions, as it helps researchers predict how effectively a molecule can deliver therapeutic effects to the brain. By using MLOGP and other molecular descriptors, scientists can better understand the pharmacokinetic properties of new drug candidates, leading to more effective and targeted drug therapies. The other group of descriptors is Atom-Centred Fragments (C-002), which are a type of molecular descriptor that can capture information about the environment surrounding each atom in a molecule [59] . These descriptors can store details about a molecule's chemical environment, connectivity, and substructures, which are relevant for predicting properties

like pKa. The atom-centered fragments and other molecular descriptors in our model are well-suited for constructing QSAR and QSPR models, including those aimed at predicting pKa values. In summary, the combination of different types of molecular descriptors can be effectively utilized in developing predictive models for pKa values, leveraging the detailed information they can provide about the chemical environment surrounding each atom in a molecule.

Conclusions

Four computational models were contrasted to assess thirty-five arsonic acids pKa quickly; three were based on DFT calculations, and the fourth was based on an SVM. Accurate prediction of these values for arsonic acid derivatives is essential in planning their extraction strategies. However, this has proven to be a more elusive task than organic molecules such as carboxylic acids or thiols. Contrary to our initial expectations, neither ML nor correlation to $V_{S,max}$ calculations provided acceptable MAE values, and instead, the method proposed by Smith et al. for the scaled SAS SMD solvation model yields the best predictions for the present family of arsonic acids. While DFT models provide highly detailed and accurate predictions, the SVM model offers a potentially faster and more efficient alternative, provided it is well-trained on relevant data. The comparison aimed to identify the best method for reliable and expedient predictions.

Acknowledgments

ML and MN gratefully acknowledge financial support from the NCCR Bioinspired Materials. We thank DGTIC - UNAM for granting access to the supercomputer '*Miztli*' and to Ms. Citlalit Martínez-Soto for keeping local computing facilities running.

Supporting Information Available

The Supporting Information is available free of charge at

- Atomic charges on the deprotonated Arsonate and the value of pKa1 to which they were correlated (PDF)
- Best R^2_{adj} for GA algorithm at different numbers of molecular descriptors (PDF)
- XYZ coordinates for all optimized compounds in the $V_{S,max}$ study at the B97XD/cc-pVDZ level

of theory. First number on the header describes the number of atoms in each molecule. Energies in atomic units. (PDF)

- ML descriptors and obtained models (XLSX)

References

(1) Reid, M. S.; Hoy, K. S.; Schofield, J. R. M.; Uppal, J. S.; Lin, Y.; Lu, X.; Peng, H.; Le, X. C. Arsenic Speciation Analysis: A Review with an Emphasis on Chromatographic Separations. *TrAC Trends in Analytical Chemistry* **2020**, *123*, 115770.

(2) Kong, L.; Zhao, J.; Hu, X.; Zhu, F.; Peng, X. Reductive Removal and Recovery of As(V) and As(III) from Strongly Acidic Wastewater by a UV/Formic Acid Process. *Environ Sci Technol* **2022**, *56* (13), 9732–9743.

(3) Tian, C.; Zhao, J.; Ou, X.; Wan, J.; Cai, Y.; Lin, Z.; Dang, Z.; Xing, B. Enhanced Adsorption of *p* -Arsanilic Acid from Water by Amine-Modified UiO-67 as Examined Using Extended X-Ray Absorption Fine Structure, X-Ray Photoelectron Spectroscopy, and Density Functional Theory Calculations. *Environ Sci Technol* **2018**, *52* (6), 3466–3475.

(4) Hasan, M. H.; McCrum, I. T. pKa as a Predictive Descriptor for Electrochemical Anion Adsorption. *Angewandte Chemie International Edition* **2024**, *63* (13).

(5) Banerjee, C.; Singh, A.; Raman, R.; Mazumder, S. Calmodulin–CaMKII Mediated Alteration of Oxidative Stress: Interplay of the CAMP/PKA–ERK 1/2-NF-KB–NO Axis on Arsenic-Induced Head Kidney Macrophage Apoptosis. *Toxicol Res (Camb)* **2013**, *2* (6), 413.

(6) Fendorf, S.; Nico, P. S.; Kocar, B. D.; Masue, Y.; Tufano, K. J. *Arsenic Chemistry in Soils and Sediments*; Lawrence Berkeley National Laboratory: Berkeley, **2009**.

(7) Bolan, N.; Mahimairaja, S.; Kunhikrishnan, A.; Naidu, R. Sorption–Bioavailability Nexus of Arsenic and Cadmium in Variable-Charge Soils. *J Hazard Mater* **2013**, *261*, 725–732.

(8) Zhao, F. J.; Ma, J. F.; Meharg, A. A.; McGrath, S. P. Arsenic Uptake and Metabolism in Plants. *New Phytologist* **2009**, *181* (4), 777–794.

(9) De Francisco, P.; Martín-González, A.; Rodriguez-Martín, D.; Díaz, S. Interactions with Arsenic: Mechanisms of Toxicity and Cellular Resistance in Eukaryotic Microorganisms. *Int J Environ Res Public Health* **2021**, *18* (22), 12226.

(10) Fisher, D. J.; Yonkos, L. T.; Staver, K. W. Environmental Concerns of Roxarsone in Broiler Poultry Feed and Litter in Maryland, USA. *Environ Sci Technol* **2015**, *49* (4), 1999–2012.

(11) Chen, J.; Zhang, J.; Rosen, B. P. Role of ArsEFG in Roxarsone and Nitarsone Detoxification and Resistance. *Environ Sci Technol* **2019**, *53* (11), 6182–6191.

(12) Yang, Z.; Peng, H.; Lu, X.; Liu, Q.; Huang, R.; Hu, B.; Kachanoski, G.; Zuidhof, M. J.; Le, X. C. Arsenic Metabolites, Including *N*-Acetyl-4-Hydroxy-m-Arsanilic Acid, in Chicken Litter from a Roxarsone-Feeding Study Involving 1600 Chickens. *Environ Sci Technol* **2016**, *50* (13), 6737–6743.

(13) Rocchia, W.; Alexov, E.; Honig, B. Extending the Applicability of the Nonlinear Poisson−Boltzmann Equation: Multiple Dielectric Constants and Multivalent Ions. *J Phys Chem B* **2001**, *105* (28), 6507–6514.

(14) Holst, M.; Baker, N.; Wang, F. Adaptive Multilevel Finite Element Solution of the Poisson-Boltzmann Equation I. Algorithms and Examples. *J Comput Chem* **2000**, *21* (15), 1319–1342.

(15) Feig, M.; Brooks, C. L. Recent Advances in the Development and Application of Implicit Solvent Models in Biomolecule Simulations. *Curr Opin Struct Biol* **2004**, *14* (2), 217–224.

(16) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. Performance Comparison of Generalized Born and Poisson Methods in the Calculation of Electrostatic Solvation Energies for Protein Structures. *J Comput Chem* **2004**, *25* (2), 265–284.

(17) Alexov, E.; Mehler, E. L.; Baker, N.; M. Baptista, A.; Huang, Y.; Milletti, F.; Erik Nielsen, J.; Farrell, D.; Carstensen, T.; Olsson, M. H. M.; Shen, J. K.; Warwicker, J.; Williams, S.; Word, J. M. Progress in the Prediction of *pKa* Values in Proteins. *Proteins: Structure, Function, and Bioinformatics* **2011**, *79* (12), 3260–3275.

(18) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J Chem Theory Comput* **2011**, *7* (2), 525–537.

(19) Wilson, C. J.; Karttunen, M.; de Groot, B. L.; Gapsys, V. Accurately Predicting Protein pKa Values Using Nonequilibrium Alchemy. *J Chem Theory Comput* **2023**, *19* (21), 7833–7845.

(20) Vila-Viçosa, D.; Reis, P. B. P. S.; Baptista, A. M.; Oostenbrink, C.; Machuqueiro, M. A PH Replica Exchange Scheme in the Stochastic Titration Constant-PH MD Method. *J Chem Theory Comput* **2019**, *15* (5), 3108–3116.

(21) Hofer, F.; Kraml, J.; Kahler, U.; Kamenik, A. S.; Liedl, K. R. Catalytic Site pKa Values of Aspartic, Cysteine, and Serine Proteases: Constant PH MD Simulations. *J Chem Inf Model* **2020**, *60* (6), 3030–3042.

(22) Buslaev, P.; Aho, N.; Jansen, A.; Bauer, P.; Hess, B.; Groenhof, G. Best Practices in Constant PH MD Simulations: Accuracy and Sampling. *J Chem Theory Comput* **2022**, *18* (10), 6134–6147.

(23) Fujiki, R.; Matsui, T.; Shigeta, Y.; Nakano, H.; Yoshida, N. Recent Developments of Computational Methods for pKa Prediction Based on Electronic Structure Theory with Solvation Models. *J (Basel)* **2021**, *4* (4), 849–864.

(24) Pezzola, S.; Tarallo, S.; Iannini, A.; Venanzi, M.; Galloni, P.; Conte, V.; Sabuzi, F. An Accurate Approach for Computational pKa Determination of Phenolic Compounds. *Molecules* **2022**, *27* (23), 8590.

(25) Haslak, Z. P.; Zareb, S.; Dogan, I.; Aviyente, V.; Monard, G. Using Atomic Charges to Describe the pKa of Carboxylic Acids. *J Chem Inf Model* **2021**, *61* (6), 2733–2743.

(26) Khalili, B.; Rimaz, M. Theoretical Calculations of the Relative pKa Values of Some Selected Aromatic Arsonic Acids in Water Using Density Functional Theory. *Current Chemistry Letters* **2016**, 7–18.

(27) Pezzola, S.; Venanzi, M.; Galloni, P.; Conte, V.; Sabuzi, F. Easy to Use DFT Approach for Computational *p* Ka Determination of Carboxylic Acids. *Chemistry – A European Journal* **2024**, *30* (1).

(28) Haworth, N. L.; Wang, Q.; Coote, M. L. Modeling Flexible Molecules in Solution: A pKa Case Study. *J Phys Chem A* **2017**, *121* (27), 5217–5225.

(29) Caballero-García, G.; Mondragón-Solórzano, G.; Torres-Cadena, R.; Díaz-García, M.; Sandoval-Lira, J.; Barroso-Flores, J. Calculation of VS, Max and Its Use as a Descriptor for the Theoretical Calculation of pKa Values for Carboxylic Acids. *Molecules* **2018**, *24* (1), 79.

(30) Sandoval-Lira, J.; Mondragón-Solórzano, G.; Lugo-Fuentes, L. I.; Barroso-Flores, J. Accurate Estimation of pKb Values for Amino Groups from Surface Electrostatic Potential ($V_{S,min}$) Calculations: The Isoelectric Points of Amino Acids as a Case Study. *J Chem Inf Model* **2020**, *60* (3), 1445–1452.

(31) Cai, Z.; Luo, F.; Wang, Y.; Li, E.; Huang, Y. Protein pKa Prediction with Machine Learning. *ACS Omega* **2021**, *6* (50), 34823–34831.

(32) Reis, P. B. P. S.; Bertolini, M.; Montanari, F.; Rocchia, W.; Machuqueiro, M.; Clevert, D.-A. A Fast and Interpretable Deep Learning Approach for Accurate Electrostatics-Driven pKa Predictions in Proteins. *J Chem Theory Comput* **2022**, *18* (8), 5068–5078.

(33) Gokcan, H.; Isayev, O. Prediction of Protein pKa with Representation Learning. *Chem Sci* **2022**, *13* (8), 2462–2474.

(34) Xiong, Y.; Liu, J.; Wei, D. An Accurate Feature-based Method for Identifying DNA-binding Residues on Protein Surfaces. *Proteins: Structure, Function, and Bioinformatics* **2011**, *79* (2), 509–517.

(35) Mansouri, K.; Cariello, N. F.; Korotcov, A.; Tkachenko, V.; Grulke, C. M.; Sprankle, C. S.; Allen, D.; Casey, W. M.; Kleinstreuer, N. C.; Williams, A. J. Open-Source QSAR Models for pKa Prediction Using Multiple Machine Learning Approaches. *J Cheminform* **2019**, *11* (1), 60.

(36) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A. , Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16, Revision C.01. Gaussian, Inc.: Wallingford CT **2016**.

(37) Lu, T.; Chen, F. Multiwfn: A Multifunctional Wavefunction Analyzer. *J Comput Chem* **2012**, *33* (5), 580–592.

(38) Ugur, I.; Marion, A.; Parant, S.; Jensen, J. H.; Monard, G. Rationalization of the pKa Values of Alcohols and Thiols Using Atomic Charge Descriptors and Its Application to the Prediction of Amino Acid pKa's. *J Chem Inf Model* **2014**, *54* (8), 2200–2213.

(39) Lian, P.; Johnston, R. C.; Parks, J. M.; Smith, J. C. Quantum Chemical Calculation of pKa's of Environmentally Relevant Functional Groups: Carboxylic Acids, Amines, and Thiols in Aqueous Solution. *J Phys Chem A* **2018**, *122* (17), 4366–4374.

(40) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach Learn* **1995**, *20* (3), 273–297.

(41) Borislavov, L.; Nedyalkova, M.; Tadjer, A.; Aydemir, O.; Romanova, J. Machine Learning-Based Screening for Potential Singlet Fission Chromophores: The Challenge of Imbalanced Data Sets. *J Phys Chem Lett* **2023**, *14* (45), 10103–10112.

(42) Nedyalkova, M.; Vasighi, M.; Azmoon, A.; Naneva, L.; Simeonov, V. Sequence-Based Prediction of Plant Allergenic Proteins: Machine Learning Classification Approach. *ACS Omega* **2023**, *8* (4), 3698–3704.

(43) Nedyalkova, M.; Paluch, A. S.; Vecini, D. P.; Lattuada, M. Progress and Future of the Computational Design of Antimicrobial Peptides (AMPs): Bio-Inspired Functional Molecules. *Digital Discovery* **2024**, *3* (1), 9–22.

(44) Mauri, A.; Bertola, M. Alvascience: A New Software Suite for the QSAR Workflow Applied to the Blood–Brain Barrier Permeability. *Int J Mol Sci* **2022**, *23* (21), 12882.

(45) Gross, K. C.; Seybold, P. G.; Hadad, C. M. Comparison of Different Atomic Charge Schemes for Predicting pKa Variations in Substituted Anilines and Phenols. *Int J Quantum Chem* **2002**, *90* (1), 445–458.

(46) Mauri, A. AlvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints; **2020**; pp 801–820.

(47) Mauri, A.; Bertola, M. Alvascience: A New Software Suite for the QSAR Workflow Applied to the Blood–Brain Barrier Permeability. *Int J Mol Sci* **2022**, *23* (21), 12882.

(48) Soriano-Meseguer, S.; Fuguet, E.; Port, A.; Rosés, M. Suitability of Skin-PAMPA and Chromatographic Systems to Emulate Skin Permeation. Influence of PH on Skin-PAMPA Permeability. *Microchemical Journal* **2023**, *190*, 108567.

(49) Lončarski, M.; Tubić, A.; Apostolović, T.; Nikić, J.; Agbaba, J. *Modelling of the Adsorption of Chlorinated Phenols on Polyethylene and Polyethylene Terephthalate Microplastic* **2020**, *85*.

(50) Sun, N.; Avdeef, A. Biorelevant PKa (37°C) Predicted from the 2D Structure of the Molecule and Its pKa at 25°C. *J Pharm Biomed Anal* **2011**, *56* (2), 173–182.

(51) Pompe, M.; Randić, M. Variable Connectivity Model for Determination of PKa Values for Selected Organic Acids Variable Connectivity Model for Determination of pKa Values for Selected Organic Acids. *Acta Chimica Slovenica* **2007**, *54*.

(52) Miličević, A.; Šinko, G. Use of Connectivity Index and Simple Topological Parameters for Estimating the Inhibition Potency of Acetylcholinesterase. *Saudi Pharmaceutical Journal* **2022**, *30* (4), 369–376.

(53) Ščavničar, A.; Balaban, A. T.; Pompe, M. Application of Variable Anti-Connectivity Index to Active Sites. Modelling pKa Values of Aliphatic Monocarboxylic Acids. *SAR QSAR Environ Res* **2013**, *24* (7), 553–563.

(54) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J Med Chem* **2000**, *43* (20), 3714–3717.

(55) Pajouhesh, H.; Lenz, G. R. Medicinal Chemical Properties of Successful Central Nervous System Drugs. *NeuroRX* **2005**, *2* (4), 541–553.

(56) Hitchcock, S. A.; Pennington, L. D. Structure−Brain Exposure Relationships. *J Med Chem* **2006**, *49* (26), 7559–7583.

(57) Gupta, M.; Lee, H. J.; Barden, C. J.; Weaver, D. F. The Blood–Brain Barrier (BBB) Score. *J Med Chem* **2019**, *62* (21), 9824–9836.

(58) Shaker, B.; Yu, M.-S.; Song, J. S.; Ahn, S.; Ryu, J. Y.; Oh, K.-S.; Na, D. LightBBB: Computational Prediction Model of Blood–Brain-Barrier Penetration Based on LightGBM. *Bioinformatics* **2021**, *37* (8), 1135–1139.

(59) Carracedo-Reboredo, P.; Liñares-Blanco, J.; Rodríguez-Fernández, N.; Cedrón, F.; Novoa, F. J.; Carballal, A.; Maojo, V.; Pazos, A.; Fernandez-Lozano, C. A Review on Machine Learning Approaches and Trends in Drug Discovery. *Comput Struct Biotechnol J* **2021**, *19*, 4538–4558.

TOC Graphic

Some journals require a graphical entry for the Table of Contents. This should be laid out "print ready" so that the sizing of the text is correct. Inside the tocentry environment, the font used is Helvetica 8 pt, as required by *Journal of the American Chemical Society*.
The surrounding frame is 9 cm by 3.5 cm, which is the maximum permitted for *Journal of the American Chemical Society* graphical table of content entries. The box will not resize if the content is too big: instead it will overflow the edge of the box.

This box and the associated title will always be printed on a separate page at the end of the document.