

SOFTWARE

Open Access



Dimorphite-DL: an open-source program for enumerating the ionization states of drug-like small molecules

Patrick J. Ropp[†], Jesse C. Kaminsky[†], Sara Yablonski and Jacob D. Durrant^{*}

Abstract

Small-molecule protonation can promote or discourage protein binding by altering hydrogen-bond, electrostatic, and van-der-Waals interactions. To improve virtual-screen pose and affinity predictions, researchers must account for all major small-molecule ionization states. But existing programs for calculating these states have notable limitations such as high cost, restrictive licenses, slow execution times, and poor modularity. Here, we present dimorphite-DL 1.0, a fast, accurate, accessible, and modular open-source program for enumerating small-molecule ionization states. Dimorphite-DL uses a straightforward empirical algorithm that leverages substructure searching and draws on a database of experimentally characterized ionizable molecules. We have tested dimorphite-DL using several versions of Python and RDKit on all major operating systems. We release it under the terms of the Apache License, Version 2.0. A copy is available free of charge from <http://durrantlab.com/dimorphite-dl/>.

Keywords: Ionization, pH, Protonation, Modeling, Virtual screening, Drug discovery

Introduction

Structure-based virtual screening (VS) predicts the geometry of a small molecule bound to its receptor (i.e., the docked pose) and maps that geometry to a score that correlates with affinity. Ligand protonation can impact electrostatic, hydrogen-bond, and van-der-Waals interactions between the ligand and receptor [1], potentially affecting both VS steps. Many ligands adopt multiple protonation states, or protomers. Protomers encompass ionization forms, which involve the gain or loss of a proton, and tautomeric forms, which involve the intramolecular transfer of a proton from one ligand atom to another [1]. Transitions between protomers (e.g., via proton uptake or release [2]) often accompany binding [3]. As most small-molecule drugs are ionizable [4, 5], accurate VS must consider the protomer that best complements the binding pocket [6, 7].

Predicting acid ionization constants (pK_a) is a critical first step. Empirical approaches such as linear free-energy calculations [8], quantitative structure–property relationships, and database similarity searches perform this prediction quickly and so are well suited for processing large compound libraries [5]. In contrast, quantum mechanical methods are slower and not necessarily more accurate [5].

After using predicted pK_a values to identify all possible ionization forms, the next step is to discard those forms that are rare. Ligands interconvert between all ionization states in solution, but the pH determines which state is favored. For example, at physiological pH (7.4), 99.96% of 3-chloropropanoic acid ($pK_a = 4.0$ [9]) exists in the deprotonated form, 3-chloropropanoate. It is reasonable to ignore the rare protonated form when performing a VS with limited computational resources. In contrast, 44.27% of 2,2,2-trifluoroethane 1-thiol ($pK_a = 7.3$ [10]) exists in the deprotonated form at physiological pH. Proper small-molecule preparation should consider both the deprotonated and protonated forms of this compound.

*Correspondence: durrantj@pitt.edu

[†]Patrick J. Ropp and Jesse C. Kaminsky have contributed equally to this work

Department of Biological Sciences, University of Pittsburgh, 4249 Fifth Avenue, Pittsburgh, PA 15260, USA



Enumerating major small-molecule ionization states can improve virtual-screen predictivity, but available programs for performing this task are generally too expensive, have restrictive licenses, are too slow for use in high-throughput contexts, predict a single state rather than all major states, and/or cannot be easily incorporated into broader drug-discovery pipelines. There is a need for a fast, accurate, accessible, and modular open-source alternative. We have developed a computer program called dimorphite-DL to address this need. We have tested dimorphite-DL using several versions of Python (2.7.13, 3.6.3, 3.6.5, and 3.6.6) and RDKit (2016.09.2, 2018.03.1, and 2018.03.4) on macOS High Sierra 10.13.4, Ubuntu 18.04.1 LTS, and Windows 10 Home 1709. We release it under the terms of the Apache License, Version 2.0. A copy is available free of charge from <http://durranlab.com/dimorphite-dl/>.

Implementation

A set of compounds with experimental pK_a values

We assembled a set of 1938 small molecules with single, diverse ionizable sites and mostly experimentally determined pK_a values. To the extent possible, we limited our search to pK_a values measured in neutral aqueous solutions near room temperature (e.g., between 23 and 27 °C). Sources of experimental data included PubChem[®], a chemical database provided by the NIH's National Library of Medicine; iBonD 2.0, the Internet Bond-Energy Databank provided by Tsinghua and Nankai Universities [11]; Reaxys, a chemical database provided by Elsevier Life Sciences IP Limited; and a published work by Lee et al. [12] that describes monoprotic small molecules. We also separately considered a set of 78 phosphates and phosphonates, which can lose up to two protons.

We performed limited data filtering to improve applicability and accuracy. For example, we removed some molecules with multiple disconnected fragments (e.g., salts) and chiral centers. If a given molecule included multiple experimental pK_a values that spanned a range greater than 1.0, we assumed experimental uncertainty and discarded the molecule. Otherwise, we averaged the available pK_a values. Per previous studies [12], we generally only considered molecules with measured pK_a values between -1.74 (H₃O⁺) and 15.7 (H₂O). In total, 98.8% of the pK_a values in our database met this criterion. To ensure proper coverage, we included nine sulfonates and sulfates with pK_a values less than -1.74. We also included fourteen molecules with pK_a values greater than 15.7: four non-phenol alcohols, three amides, and seven molecules with protonated but uncharged aromatic nitrogen atoms. Finally, to ensure that nitro-group oxygen atoms

are always deprotonated, we assigned a very negative, arbitrary pK_a value (-1000.0) to this group.

We grouped these compounds by ionizable moiety and constructed pK_a histograms for each group. Although chemical features beyond the moiety itself (e.g., neighboring electronegative groups) do impact pK_a, our analysis provided a typical pK_a range for each ionizable site. In some cases, visual inspection of the histograms led us to reconsider some moiety definitions. For example, the distribution of amide pK_a values was initially bimodal. By separating amides from nitrogens bonded to electronegative atoms, we divided this group into two chemically distinct populations.

Ultimately, we settled on 38 ionizable substructures. In some cases, a given moiety could belong to two such categories. For example, every amide contains an amine group. To uniquely assign each moiety to a given categorization, we prioritized the substructures. Atoms belonging to high-priority substructures that cover more atoms (e.g., amides) were not considered when subsequently searching for lower-priority substructures (e.g., amines).

For each set of compounds matching one of these substructures, we calculated the mean (μ) and standard deviation (σ) of the associated pK_a values (Table 1). In the case of the phosphates and phosphonates, each moiety was associated with two separate pK_a means and standard deviations (μ_1 and σ_1 ; and μ_2 and σ_2), one for each ionizable proton. A range of reasonable pK_a values for each moiety, $\text{range}_{\text{pK}_a}$, is given by $[\mu - n\sigma, \mu + n\sigma]$, where n is a user-defined parameter we call the "pK_a precision factor."

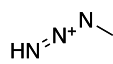

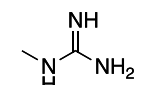

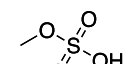
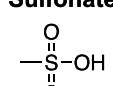

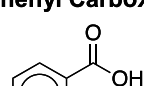

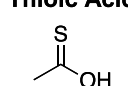
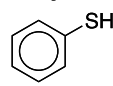

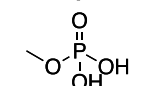

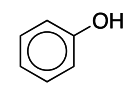
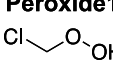

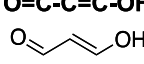

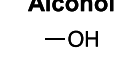
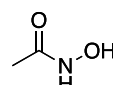

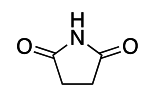

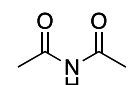
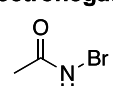

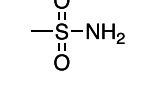

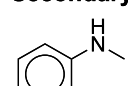
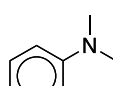

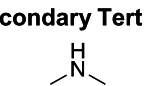

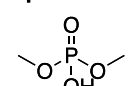
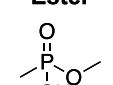

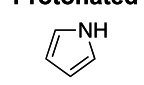
Predicting ionization states

Dimorphite-DL 1.0 uses the μ and σ values associated with each ionizable moiety to predict small-molecule ionization states for a given pH range. It accepts the following user inputs:

1. A small-molecule library in SMILES format [13], with each compound SMILES on its own line. Alternatively, the user can provide a single SMILES as a command-line parameter.
2. The pK_a precision factor to use when estimating moiety pK_a ranges (n , 1.0 by default).
3. The minimum pH to consider (pH_{min} , 6.4 by default).
4. The maximum pH to consider (pH_{max} , 8.4 by default).

For each molecule, dimorphite-DL uses RDKit [14], an open-source cheminformatics library, to search for the 38 ionizable substructures described above (Table 1 and Additional file 1: Table S1). The same prioritization scheme ensures that any given atom is assigned to at most only one category. The program outputs

Table 1 The 38 ionizable dimorphite-DL substructures in order of decreasing priority from left to right, with representative compounds

Azide  4.65 ± 0.07	Nitro  -1000 ± 0	AmidineGuanidine1  12.03 ± 1.59	AmidineGuanidine2  10.04 ± 2.13	Sulfate  -2.36 ± 1.30
Sulfonate  -1.82 ± 1.41	Sulfinic Acid  1.79 ± 0.44	Phenyl Carboxyl  3.46 ± 1.25	Carboxyl  3.46 ± 1.29	Thioic Acid  0.68 ± 1.50
Phenyl Thiol  4.98 ± 2.61	Thiol  9.12 ± 1.33	Phosphate  2.42 ± 1.11 6.51 ± 0.95	Phosphonate  1.88 ± 7.25 0.59 ± 0.85	Phenol  7.07 ± 3.28
Peroxide1  8.74 ± 0.76	Peroxide2  11.98 ± 0.87	O=C-C=C-OH  3.55 ± 0.80	Vinyl Alcohol  8.87 ± 1.66	Alcohol  14.78 ± 2.55
N-hydroxyamide  9.30 ± 1.22	Ringed Imide1  6.45 ± 0.56	Ringed Imide2  8.68 ± 1.87	Imide  2.47 ± 1.48	Imide2  10.23 ± 1.12
Amide Electronegative  3.49 ± 2.69	Amide  12.01 ± 4.51	Sulfonamide  7.92 ± 1.98	Anilines Primary  3.90 ± 2.07	Anilines Secondary  4.34 ± 2.18
Anilines Tertiary  4.17 ± 2.01	Aromatic Nitrogen Unprotonated  4.35 ± 2.07	Amines Primary Secondary Tertiary  8.16 ± 2.52	Phosphinic Acid  2.97 ± 0.69	Phosphate Diester  2.73 ± 2.54
Phosphonate Ester  2.09 ± 0.45	Primary Hydroxyl Amine  4.04 ± 0.85	Aromatic Nitrogen Protonated  12.08 ± 5.10		

Exact substructure definitions are given in Additional file 1: Table S1. The pK_a range is the average of all associated pK_a values in the database, plus or minus the standard deviation

protonated/deprotonated SMILES, as appropriate for the user-specified pH range.

Dimorphite-DL does not calculate pK_a values explicitly. Rather, for each categorized moiety, it takes one of three actions (Fig. 1) based on a moiety-specific pK_a range, $range_{pK_a} = [\mu - n\sigma, \mu + n\sigma]$, and a user-defined pH range, $range_{pH} = [pH_{min}, pH_{max}]$:

1. If $\mu + n\sigma < pH_{min}$ (i.e., the entirety of $range_{pK_a}$ is less than the entirety of $range_{pH}$), the moiety is deprotonated.
2. If $pH_{max} < \mu - n\sigma$ (i.e., the entirety of $range_{pK_a}$ is greater than the entirety of $range_{pH}$), the moiety is protonated.
3. If $range_{pK_a}$ and $range_{pH}$ overlap, two distinct small-molecule models are generated, with the moiety protonated and deprotonated, respectively.

Substructure identification using SMARTS

Dimorphite-DL uses SMARTS (SMILES arbitrary target specification [15]) to determine which atoms in a given molecule belong to one of the 38 ionizable substructures. SMARTS is a language for molecular subgraph isomorphism and pattern matching that is an extension of the popular SMILES format [13]. SMILES describes a molecule, but SMARTS describes a molecular pattern (e.g., a substructure). All SMILES strings are valid SMARTS strings, but SMARTS additionally allows for variable atom and bond specifications. A given SMARTS string can thus represent multiple related chemical structures.

SMARTS pattern recognition requires standardized input molecules. Dimorphite-DL attempts to standardize

all input SMILES strings automatically. For example, $N[N+] \# N$, $N=[N+] = N$, $NN \# N$, and $N=N=N$ are all recognized as azides. Both $[N+]([O-])=O$ and $N(=O)=O$ are recognized as nitro groups. And aromatic rings are recognized regardless of whether the input SMILES string describes aromatic bonds (e.g., both $Oc1ccccc1$ and $OC1=CC=CC=C1$ are recognized as phenols).

The Additional file 1 includes the SMARTS representations of the 38 substructures, as well as the calculated μ and σ values (Additional file 1: Table S1). The Additional file 1 also describes how dimorphite-DL independently handles phosphate and phosphonate groups.

Results and discussion

Dimorphite-DL 1.0 is a fast, accessible, open-source Python program for predicting small-molecule ionization states. As a simple illustration of the importance of accounting for alternate ionization states, consider the local anesthetic lidocaine. The pK_a of lidocaine is 8.01 [16], so both the charged protonated- and deprotonated-amine forms are prevalent at physiological pH. Computational evidence suggests that both forms bind the sodium channel NavPas, preventing cell depolarization [17]. But the two protomers have different poses [17], and charged lidocaine binds with higher affinity [17]. It is thus critical to account for both ionization states. Dimorphite-DL applied to lidocaine successfully predicted both forms.

The dimorphite-DL approach

Dimorphite-DL uses a substructure-based empirical algorithm to quickly prepare large compound libraries for virtual screening (VS). Importantly, it is not limited to identifying a single ionizable state per molecule. Rather,

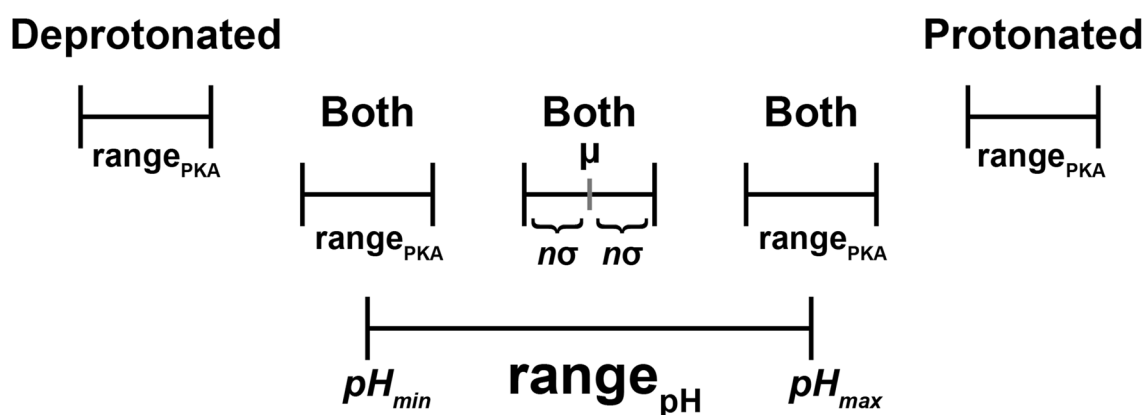


Fig. 1 A schematic representation of the dimorphite-DL approach. Each ionizable moiety is associated with a pK_a range ($range_{pK_a}$) defined by three parameters: μ , σ , and n . The user specifies a pH range ($range_{pH}$) and pK_a precision factor (n ; default: 1.0). The mean (μ) and standard deviation (σ) associated with each moiety are derived from the database of small molecules with experimentally characterized pK_a values. If $range_{pK_a}$ is entirely less than $range_{pH}$, dimorphite-DL outputs a deprotonated molecule. If $range_{pK_a}$ is entirely greater than $range_{pH}$, dimorphite-DL outputs a protonated molecule. If $range_{pK_a}$ and $range_{pH}$ overlap, dimorphite-DL outputs both deprotonated and protonated molecules

it can generate multiple states as appropriate for a given pH range, thus increasing the chance of identifying the most binding-compatible state.

Because distant chemical groups can impact proton dissociation, the ideal program for enumerating ionizable states would calculate pK_a values in the context of the whole molecule. But creating such a program is challenging. Methods that consider whole-molecule contexts, such as quantum mechanical approaches, are too slow for high-throughput use. Surprisingly, their pK_a calculations are not necessarily more accurate than those of simpler algorithms [5]. Empirical methods such as dimorphite-DL are faster, but they draw on chemical databases that cannot account for all possible chemical contexts. Dimorphite-DL predicts ionization states by considering only at most a few atoms adjacent to each ionizable moiety. In compounds with multiple ionizable sites, each site is considered independently. Our algorithm thus does not account for interactions between sites or other electronic effects.

To compensate for this limitation, we associate 38 ionizable moieties with pK_a ranges rather than point values. We derive ranges from the experimental pK_a values of 1938 small molecules (see Materials and Methods). For each moiety, $\text{range}_{pK_a} = [\mu - n\sigma, \mu + n\sigma]$, where μ and σ are the mean and standard deviation of the associated experimental pK_a values, respectively; and n is a user-defined “ pK_a precision factor”. Protonation is assigned based on the overlap between range_{pK_a} and the user-specified pH range (Fig. 1).

Dimorphite-DL accuracy: correct, excessive, and incorrect predictions

Defining terms will allow us to better describe the accuracy of Dimorphite-DL predictions. Consider a user-defined pH range, $\text{range}_{pH} = [pH_{min}, pH_{max}]$, and a compound with an experimentally determined pK_a value. Further assume that the compound can lose at most one proton. Applying dimorphite-DL to this compound can have one of three outcomes:

- Dimorphite-DL predicts the correct state
 - $pK_a < pH_{min}$, and dimorphite-DL deprotonates the compound
 - $pK_a > pH_{max}$, and dimorphite-DL protonates the compound
 - $pH_{min} \leq pK_a \leq pH_{max}$, and dimorphite-DL generates both deprotonated and protonated forms
- Dimorphite-DL predicts an excess state (i.e., two states when only one is appropriate)

- $pK_a < pH_{min}$ or $pK_a > pH_{max}$, but dimorphite-DL generates both deprotonated and protonated forms
- Dimorphite-DL predicts the incorrect (or incomplete) state
 - $pK_a < pH_{min}$, but dimorphite-DL protonates the compound
 - $pK_a > pH_{max}$, but dimorphite-DL deprotonates the compound
 - $pH_{min} \leq pK_a \leq pH_{max}$, and dimorphite-DL either deprotonates or protonates the compound (not both)

We distinguish between “excess-state” and “incorrect-state” outcomes because they differ in their consequences. If dimorphite-DL predicts an excess state, it needlessly expands the compound library and increases the computational expense of subsequent VS. But if it predicts an incorrect or incomplete state, VS accuracy may suffer because a relevant state is never generated.

The influence of the pK_a precision factor on accuracy

Recall that each moiety has an associated range_{pK_a} ($[\mu - n\sigma, \mu + n\sigma]$) determined in part by the user-specified pK_a precision factor, n . To assess the influence of this factor on accuracy, we evaluated the compounds in our database using different values of n (Table 2), always over the default $\text{range}_{pH} = [6.4, 8.4]$. As n increases, more compounds are assigned excess states, reducing the number of entirely correct and entirely incorrect assignments (Table 2, Additional file 1: Tables S2, S3, and S4). We select $n = 1.0$ as our default, as it strikes a good balance between the three outcomes.

Users can specify other values of n according to their needs. Tuning the pK_a precision factor is particularly

Table 2 Dimorphite-DL accuracy

pK_a precision factor, n (standard deviation)	Correct (%)	Excess (%)	Incorrect (%)
0.0	70.9	23.9	5.2
0.5	69.1	26.5	4.4
1.0	58.8	40.2	0.9
1.5	51.2	48.8	0.0
2.0	50.7	49.3	0.0
2.5	23.9	76.1	0.0
3.0	22.1	77.9	0.0

The percentage of molecules that are correctly/excessively/incorrectly protonated at different pK_a precision factors (n), at physiological pH (6.4–8.4). To generate these statistics, we considered all 1938 compounds in our primary set, as well as the 78 additional phosphate and phosphonate compounds described in the Additional file 1

useful when one needs to limit the size of the compound library. To illustrate, consider a given library compound with i distinct ionizable moieties. Dimorphite-DL will process each moiety separately. If no moiety has a $\text{range}_{\text{pK}_A}$ that overlaps with range_{pH} , dimorphite-DL will produce only one protomer. If every $\text{range}_{\text{pK}_A}$ overlaps with range_{pH} , dimorphite-DL will produce 2^i distinct protomers.

Selecting lower values of n protects against combinatorial explosions, as $\text{range}_{\text{pK}_A}$ and range_{pH} are less likely to overlap. Narrowing the difference between pH_{min} and pH_{max} can further reduce overlap. These measures limit the size of the resulting compound library, reducing the computational cost of any subsequent VS. But restrictive parameters may force dimorphite-DL to ignore binding-relevant states, reducing VS accuracy. Table 2, Additional file 1: Tables S2, S3, and S4 will help the user find a good balance between accuracy, generalizability, and performance.

Accuracy per ionizable moiety

Next, we evaluated how accurately dimorphite-DL predicts the ionization states of individual moieties (Table 3). To simplify analysis, we considered only $n=1.0$ and physiological pH ($\text{range}_{\text{pH}}=[6.4, 8.4]$). Here, we focus on the amine (1°, 2°, and 3°), carboxylic acid, phenol, benzoic acid, and sulfonamide moieties because they are drug like and are well represented in our 1938-member compound set (21%, 20%, 10%, 7%, and 2%, respectively). A similar analysis of the remaining moieties can be found in the Additional file 1: Tables S2, S3, and S4.

We evaluated each moiety using threefold cross validation. For each fold, we divided all the relevant molecules from our compound set into a training set (two thirds of all samples) and a testing set (the remaining one third of all samples). We calculated the pK_a mean (μ) and standard deviation (σ) of the compounds in the training set and defined $\text{range}_{\text{pK}_A}$ to be $[\mu - 1.0\sigma, \mu + 1.0\sigma]$. To evaluate accuracy, we calculated the percentage of testing-set

compounds with correct, excess, and incorrect predicted states. This cross-validation approach was used only to evaluate our model. The published program uses $\text{range}_{\text{pK}_A}$ values derived from all molecules.

Comparing dimorphite-DL to similar commercial programs

Several other programs can predict small-molecule ionization states. A complete review of these programs is beyond the scope of this work. We direct interested readers to refs. [5, 18–20]. But we do wish to mention a few advantages that dimorphite-DL has over other packages.

Dimorphite-DL is free and open source. Similar commercial programs can be expensive (e.g., Schrödinger's Epik [1, 4] and Jaguar; BIOVIA's Pipeline Pilot; etc.). Not all academic researchers can afford the subscription fees, and labs that focus primarily on experimental work cannot justify so large a computational investment.

Some programs (e.g., software by ChemAxon and OpenEye) have “free” academic licenses with concerning commercialization and intellectual-property (IP) restrictions. For example, OpenEye's free license requires researchers to give up any IP rights and to promptly release their work to the public domain. Eligibility is also regularly reevaluated, and access may be unexpectedly and suddenly withdrawn. Many researchers are reluctant to incorporate commercial tools into existing pipelines, as they limit dissemination.

Dimorphite-DL is well suited for preparing large compound libraries for use in VS. Unlike some other programs (e.g., ChemAxon's Marvin [5]), dimorphite-DL can process small molecules in batch. Its empirical approach also prepares large libraries quickly. The expensive quantum mechanical calculations used by some other programs (e.g., Schrödinger's Jaguar) cannot be easily applied at scale. ARChem's SPARC program, though capable of batch processing, also reportedly suffers from long runtimes [5].

Comparing dimorphite-DL to Open Babel

The popular cheminformatics program Open Babel [21] is arguably most similar to dimorphite-DL in terms of its license and features. Like dimorphite-DL, Open Babel allows users to ionize small-molecule models as appropriate for a given pH. For each input molecule, Open Babel produces a single output molecule. In contrast, dimorphite-DL can produce multiple outputs, each with different ionization states. Open Babel and dimorphite-DL are also similar in that both are released under open-source licenses. But Open Babel is licensed under the GNU General Public License, a so-called viral license. Many who wish to incorporate an ionization module into their existing software will find this license unacceptable.

Table 3 Dimorphite-DL accuracy at physiological pH (6.4–8.4) for five common moieties

	Correct (%)	Excess (%)	Incorrect (%)
Amine (1°, 2°, and 3°)	26.9 ± 3.0	73.1 ± 3.0	0.0 ± 0.0
Carboxylic acid	100.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Phenol	33.7 ± 3.8	66.3 ± 3.8	0.0 ± 0.0
Benzoic acid	100.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Sulfonamide	37.1 ± 11.6	62.9 ± 11.6	0.0 ± 0.0

Mean ± standard-deviation percentages were calculated using three-fold cross validation. The pK_a precision factor (n) is 1.0. Additional file 1: Tables S2, S3, and S4 report similar accuracy measures for additional moieties, range_{pH} , and n

In contrast, dimorphite-DL is released under the more permissive Apache License, Version 2.0.

Open Babel ionizes small molecules per a rule-based approach similar to that of dimorphite-DL. Excluding its substructure rules specific to amino acids, Open Babel considers 14 generally applicable ionizable substructures. In contrast, dimorphite-DL considers 38 such substructures. There is a one-to-one mapping between many Open-Babel and dimorphite-DL substructure rules. Both programs recognize the same amines, hydroxamic acids, phosphate diesters, phosphonate esters, sulfinic acids, guanidines/amidines, and azides. In other cases, a single Open-Babel substructure maps to multiple dimorphite-DL substructures. For example, unlike Open Babel, dimorphite-DL distinguishes between carboxylates and phenyl carboxylates; phosphates and phosphonates; sulfates and sulfonates; and vinyl alcohols with and without conjugated ketones. Open Babel and dimorphite-DL also handle aromatic nitrogen atoms differently. Whereas Open Babel considers imidazoles and tetrazoles specifically, dimorphite-DL takes a more generalizable approach. When predicting ionization, dimorphite-DL considers only whether an uncharged aromatic nitrogen atom is protonated (e.g., 1*H*-pyrrole) or unprotonated (e.g., pyridine).

Dimorphite-DL also better accounts for phosphate and phosphonate groups. These groups can exist in three ionization states (doubly deprotonated, singly deprotonated, and fully protonated). Dimorphite-DL can generate all three forms, but Open Babel generates only one of two (doubly deprotonated or fully protonated). Dimorphite-DL also generates both protonated and deprotonated azides ($pK_a = 4.65$ [22]); in contrast, Open Babel always protonates azides.

Limitations

As mentioned above, Dimorphite-DL uses an empirical approach with the advantages of speed and reasonable accuracy, but it does assign ionization states without regard for the larger intramolecular context. To understand why intramolecular effects are at times important, consider phenol (pK_a of 9.99 at 25° in water [23]). Adding hydrocarbon substituents to the phenyl ring increases the hydroxyl pK_a (e.g., *m*-cresol, 4-(*tert*-butyl)phenol, and 2,6-di-*tert*-butyl-4-methyl-phenol have pK_a values of 10.09 [24], 10.32 [25], and 12.55 [26], respectively). In contrast, halide substituents tend to decrease the hydroxyl pK_a (e.g., 2,3,4,5,6-pentachlorophenol, 2,4,6-trichlorophenol, 2,4-dichlorophenol, and 4-chlorophenol have pK_a values of 4.79 [27], 6.15 [28], 7.85 [29], and 9.59 [25], respectively). Dimorphite-DL considers only the phenol substructure when predicting ionization states. It knows only that phenols in all their forms tend

to have hydroxyl pK_a values that center around 7.07, with a standard deviation of 3.28 (Table 1).

The same limitation applies to additional moieties that are themselves ionizable. For example, when a second ionizable hydroxyl group is added to a phenol aromatic ring, the pK_a is slightly reduced (e.g., pyrocatechol, resorcinol, and hydroquinone have pK_a values of 9.25 [30], 9.44 [31], and 9.85 [32], respectively). Ionizable carboxylate groups also impact the pK_a (e.g., 4-hydroxybenzoic acid and salicylic acid have hydroxyl pK_a values of 9.23 [33] and 13.3 [34], respectively), as do ionizable sulfonate groups (e.g., 4-hydroxybenzenesulfonic acid and 3-hydroxybenzenesulfonic acid have hydroxyl pK_a values of 8.7 and 9.07 [35], respectively).

Applying dimorphite-DL to salicylic acid (i.e., 2-hydroxybenzoic acid) at physiological pH (pH 6.4–8.4, default settings) illustrates the occasional pitfalls of our limited-substructure approach. Dimorphite-DL correctly recognized that protonated carboxyl groups are rare at this pH. But it incorrectly predicted that the hydroxyl group exists in both protonated and unprotonated forms. In reality, the pK_a of the salicylic acid hydroxyl group is unusually high (13.3 [34]), such that only the protonated hydroxyl is truly prevalent. While salicylic acid presents an admittedly extraordinary use case, we nevertheless welcome future high-throughput methods that take a more whole-molecule approach to ionization-state prediction.

We note also that dimorphite-DL computes ionization states, but not prototropic tautomerization states [36]. To clarify, ionization involves the gain or loss of a proton. Prototropic tautomerization involves intramolecular proton transfer from one atom to another. Existing open-source tools (e.g., MolVS [37]) are available for modeling prototropic tautomerization. Using dimorphite-DL with these other programs will allow researchers to fully enumerate all protonation (i.e., ionization and tautomeric) states.

These limitations aside, we expect that dimorphite-DL will be a useful tool for researchers engaged in structure-based VS. This free and open-source program for predicting small-molecule ionization states will improve VS accuracy, helping to identify novel bioactive molecules.

Additional file

[Additional file 1.](#) Supplementary discussion and tables.

Authors' contributions

PJR and JDD wrote the software. JCK and SY evaluated the accuracy of the software. SY performed critical background research. JCK, SY, and JDD wrote the manuscript. All authors read and approved the final version.

Acknowledgements

We acknowledge Jacob O. Spiegel for helpful discussions and code reviews.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

Project name: Dimorphite-DL 1.0, Project home page: <http://durrantlab.com/dimorphite-dl/>, Operating systems: macOS High Sierra, Ubuntu 18.04, and Windows 10 Home, Programming language: Python 2.7 or 3.6, Other requirements: RDKit 2016.09.2 or higher, License: Apache License, Version 2.0.

Funding

Not applicable.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 20 November 2018 Accepted: 8 February 2019

Published online: 14 February 2019

References

- Shelley JC, Cholleti A, Frye LL, Greenwood JR, Timlin MR, Uchimaya M (2007) Epik: a software program for pKa prediction and protonation state generation for drug-like molecules. *J Comput Aided Mol Des* 21:681–691. <https://doi.org/10.1007/s10822-007-9133-z>
- Mitra R, Shyam R, Mitra I, Miteva MA, Alexov E (2008) Calculating the protonation states of proteins and small molecules: implications to ligand-receptor interactions. *Curr Comput Aided Drug Des* 4:169–179
- Petukh M, Stefl S, Alexov E (2013) The role of protonation states in ligand-receptor recognition and binding. *Curr Pharm Des* 19:4182–4190
- Greenwood JR, Calkins D, Sullivan AP, Shelley JC (2010) Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J Comput Aided Mol Des* 24:591–604. <https://doi.org/10.1007/s10822-010-9349-1>
- Liao C, Nicklaus MC (2009) Comparison of nine programs predicting pKa values of pharmaceutical substances. *J Chem Inf Model* 49:2801–2812. <https://doi.org/10.1021/ci900289x>
- Rapp CS, Schonbrun C, Jacobson MP, Kalyanaraman C, Huang N (2009) Automated site preparation in physics-based rescoring of receptor ligand complexes. *Proteins* 77:52–61. <https://doi.org/10.1002/prot.22415>
- Knox AJ, Meegan MJ, Carta G, Lloyd DG (2005) Considerations in compound database preparation—"hidden" impact on virtual screening results. *J Chem Inf Model* 45:1908–1919. <https://doi.org/10.1021/ci050185z>
- Perrin DD, Dempsey B, Serjeant EP (1981) pKa prediction for organic acids and bases, vol 1. Springer, Netherlands
- Lomas JS (2012) 1H NMR study of the hetero-association of unsaturated alcohols with pyridine. *J Phys Org Chem* 25:620–627
- DeCollo TV, Lees WJ (2001) Effects of aromatic thiols on thiol–disulfide interchange reactions that occur during protein folding. *J Org Chem* 66:4244–4249
- Internet Bond-Energy Databank (2018). Tsinghua and Nankai Universities. <http://ibond.nankai.edu.cn/>. Accessed 13 Nov 2018
- Lee AC, Yu JY, Crippen GM (2008) pKa prediction of monoprotic small molecules the SMARTS way. *J Chem Inf Model* 48:2042–2053. <https://doi.org/10.1021/ci8001815>
- Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 28:31–36
- Landrum G (2018) RDKit: open-source cheminformatics. <http://www.rdkit.org/>. Accessed 13 Nov 2018
- Daylight (2011) Daylight theory manual: SMARTS: a language for describing molecular patterns. Daylight Chemical Information Systems, Inc. <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>. Accessed 1 Jan 2019
- Wishart DS et al (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucl Acids Res* 34:D668–D672. <https://doi.org/10.1093/nar/gkj067>
- Buyan A, Sun D, Corry B (2018) Protonation state of inhibitors determines interaction sites within voltage-gated sodium channels. *Proc Natl Acad Sci U S A* 115:E3135–E3144. <https://doi.org/10.1073/pnas.1714131115>
- Dardonville C (2018) Automated techniques in pKa determination: low, medium and high-throughput screening methods. *Drug Discov Today Technol* 27:49–58. <https://doi.org/10.1016/j.ddtec.2018.04.001>
- Shields GC, Seybold PG (2013) Computational approaches for the prediction of pKa values. CRC Press, Boca Raton
- Balogh GT, Tarcsay A, Keseru GM (2012) Comparative evaluation of pKa prediction tools on a drug discovery dataset. *J Pharm Biomed Anal* 67–68:63–70. <https://doi.org/10.1016/j.jpba.2012.04.021>
- O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. *J Cheminform* 3:33. <https://doi.org/10.1186/1758-2946-3-33>
- Betterton EA, Craig D (1999) Kinetics and mechanism of the reaction of azide with ozone in aqueous solution. *J Air Waste Manag Assoc* 49:1347–1354
- Bouchard G, Carrupt PA, Testa B, Gobry V, Girault HH (2002) Lipophilicity and solvation of anionic drugs. *Chemistry* 8:3478–3484. [https://doi.org/10.1002/1521-3765\(20020802\)8:15%3c3478::AID-CHEM3478%3e3.0.CO;2-U](https://doi.org/10.1002/1521-3765(20020802)8:15%3c3478::AID-CHEM3478%3e3.0.CO;2-U)
- Biggs AI (1956) The ionization constants of phenol and of some substituted phenols. *Trans Faraday Soc* 52:35–39. <https://doi.org/10.1039/tf9565200035>
- Lugo-Gonzalez JC, Gomez-Tagle P, Huang XM, del Campo JM, Yatsimirsky AK (2017) Substrate specificity and leaving group effect in ester cleavage by metal complexes of an oximate nucleophile. *Inorg Chem* 56:2060–2069. <https://doi.org/10.1021/acs.inorgchem.6b02739>
- Ohmori H, Ueda C, Nakagawa T, Nishiguchi S, Jeong J, Masui M (1986) Unsymmetrical anodic C–C coupling of 2,6-Di-tert-butyl-4-methylphenol. *Chem Pharm Bull* 34:508–515
- Cevasco G, Thea S (1998) The dissociative route in the alkaline hydrolysis of aryl 4-hydroxy-beta-styrenesulfonates. *J Org Chem* 63:2125–2129. <https://doi.org/10.1021/jo971508w>
- Lente G, Espenson JH (2004) Unusual kinetic role of a water-soluble iron (III) porphyrin catalyst in the oxidation of 2, 4, 6-trichlorophenol by hydrogen peroxide. *Int J Chem Kinet* 36:449–455
- Mock WL, Morsch LA (2001) Low barrier hydrogen bonds within salicylate mono-anions. *Tetrahedron* 57:2957–2964. [https://doi.org/10.1016/S0040-4020\(01\)00158-2](https://doi.org/10.1016/S0040-4020(01)00158-2)
- Sever MJ, Wilker JJ (2006) Absorption spectroscopy and binding constants for first-row transition metal complexes of a DOPA-containing peptide. *Dalton Trans* 6:813–822. <https://doi.org/10.1039/b509586g>
- Sanyal SK, Mandal SK (1985) Diffusion of phenols and benzyl alcohol through porous (G-4) diaphragm. *Indian J Chem Sect A Inorg Bio-inorg Phys Theor Anal Chem* 24:603–604
- Baxendale JH, Hardy HR (1953) The ionization constants of some hydroquinones. *Trans Faraday Soc* 49:1140–1144
- Lugo-Gonzalez JC, Gomez-Tagle P, Huang X, del Campo JM, Yatsimirsky AK (2017) Substrate specificity and leaving group effect in ester cleavage by metal complexes of an oximate nucleophile. *Inorg Chem* 56:2060–2069. <https://doi.org/10.1021/acs.inorgchem.6b02739>
- Dhat CR, Jahagirdar DV (1982) Copper (II) chelates of substituted salicylic acids—a thermodynamic study. *Indian J Chem Sect A Inorg Bio-inorg Phys Theor Anal Chem* 21:792–795
- Zollinger H, Büchler W, Wittwer C (1953) Wirkung der Sulfosäuregruppe auf aromatische Systeme: Hammett's σ -Werte des SO_3^- -Substituenten. *Helv Chim Acta* 36:1711–1722
- McNaught AD, Wilkinson A (1997) Tautomerism. In: *Compendium of chemical terminology, gold book*, 2nd edn. International Union of Pure and Applied Chemistry, Research Triangle Park, North Carolina, pp 1513–1515
- Swain M (2018) MolVS: molecule validation and standardization. <https://github.com/mcs07/MolVS>. Accessed 31 Dec 2018