Prediction of Aqueous $pK_a$ Values for Guanidine-Containing Compounds Using Ab Initio Gas-Phase Equilibrium Bond Lengths

Beth A. Caine,†‡ Christophe Dardonville,§ and Paul L. A. Popelier*†‡

†Manchester Institute of Biotechnology (MIB), 131 Princess Street, Manchester M1 7DN, Great Britain
‡School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, Great Britain
§Instituto de Química Médica, IQM−CSIC, C/Juan de la Cierva 3, 28006 Madrid, Spain

ABSTRACT: In this work, we demonstrate the existence of linear relationships between gas-phase equilibrium bond lengths of the guanidine skeleton of 2-(arylamino)imidazolines and their aqueous $pK_a$ value. For a training set of 22 compounds, in the most stable conformation of their lowest energy tautomeric form, three bonds were found to exhibit $r^2$ and $q^2$ values >0.95 and root-mean-squared-error of estimation values $\leq$0.25 when regressed individually against $pK_a$. The equations describing these one-bond-length linear relationships, in addition to a multiple linear regression model using all three bond lengths, were then used to predict the experimental $pK_a$ values of an external test set of further 27 derivatives. The optimal protocol we derive here shows an overall mean absolute error (MAE) of 0.20 and standard deviation of errors of 0.18 for the test set. Predictions for a second test set of diphenyl-based bis(2-iminoimidazolidines) yielded an MAE of 0.27 and a standard deviation of 0.10. The predictive power of the optimal model is further demonstrated by its ability to correct erroneously reported experimental values. Finally, a previously established guanidine model is recalibrated at a new level of theory, and predictions are made for novel phenylguanidine derivatives, showing an MAE of just 0.29. The protocols established and tested here pass both of Roy’s modern and stringent MAE-based criteria for a “good” quantitative structure−activity relationship/quantitative structure−property relationship model predictivity. Notably, the ab initio bond length high correlation subset protocol developed in this work demonstrates lower MAE values than the Marvin program by ChemAxon for all test sets.

1. INTRODUCTION

Having a measure for the propensity of a molecule to lose or gain a proton, under given conditions, (e.g., pH, solvent, and temperature) has applications in numerous fields. For synthetic chemists, knowledge of $pK_a$ values provides means to manipulate mechanistic routes. In a physiological context, biochemical reactions may be rationalized with insight into the protonation state of the reacting species. The absorption, diffusion, metabolism, excretion, and toxicity profiles of a drug candidate strongly depends on its ionization state under physiological conditions. Furthermore, the Trapp model, used in the agrochemical industry to predict the ease with which a compound can travel through the phloem of a plant to its target site, uses $pK_a$ in addition to the log $P$ value of a lead compound.$^1$

The measurement of $pK_a$ values of small, weakly acidic molecules ($pK_a$ range 2–15) may be achieved accurately via a variety of methods, including ultraviolet−visible spectrometry, potentiometry, or nuclear magnetic resonance spectroscopy, to name a few. An accurate method for the determination of $pK_a$ without the need for synthesis offers great advantage in the context of early-stage drug and agrochemical discovery, when thousands of potential compounds must be assessed for target specificity and selectivity. The pursuit of a fast, accurate, and all-encompassing $pK_a$ prediction method is still a key challenge in chemistry, and in particular, the prediction accuracy of bases is of much interest.$^2$–$^4$

Empirical $pK_a$ predictors, such as ACD/$pK_a$ DB$^5$ of ACD/Labs and Epik,$^6$ utilize extensive databases of experimental values to formulate a set of rules to predict ionization feasibility,
in line with known substituent effects. To do so, they often invoke Hammett–Taft equations, with additional molecule-specific details such as mesomer standardization and charge cancellation. Marvin’s is the pK_a prediction package from ChemAxon, which uses quantitative structure property relationships (QSPRs) based on various chemical descriptors, including partial charge and polarizability. First principles-based approaches have also been widely explored but are yet to be widely implemented in popular chemical software packages, with the exception of Jaguar pK_a.2,8 by Schrödinger. These protocols calculate the change in Gibbs energy for dissociation, accessing pK_a values via the relation of this calculated quantity to the temperature, gas constant, and common logarithm of the dissociation constant. This approach to pK_a prediction prescribes implementation of a thermodynamic cycle of some sort, whereby both gas-phase and solution-phase energies are calculated for deprotonated and protonated species. Composite methods such as G3SX,9 CBS-QB3, and CBS-APNO10 have been shown to work favorably in the calculation of relevant thermodynamic quantities, but the high computational cost accrued by such high levels of theory remains a barrier to their utility. Recently, Ho11 suggested that the use of thermodynamic cycles can be avoided completely via high-level electronic structure calculations [MP2/GTMP2Large and G3(MP2)-RAD(+)], computed directly in the solution phase with the SMD solvation model.12 Other recent work13 details the use of an isodesmic reaction scheme, whereby relevant energetic quantities are computed using semiempirical methods. The authors implement AM1/SMD and PM3/SMD for a structurally diverse set of amines, yielding root-mean-squared-error values of 1.0 and 1.1 pK_a units, respectively. An error of 5.69 kJ mol\(^{-1}\) in the Gibbs free energy will alter the predicted pK_a value by one unit. Hence, accuracy in the determination of the intermediate free energy values for first principles calculations is imperative.

In the pursuit of prediction accuracy in the first principles-based methods, the solvation scheme must also be considered, in addition to the choice of the thermodynamic cycle variant and the level of theory. Implicit–explicit models, whereby at least one water molecule is included to represent hydrogen bonding interactions, have been shown to offer significant improvement in the solvation energies over the use of an implicit model alone.14–16 However, modeling the water molecules quantum mechanically further add to computational expense. The Jaguar pK_a program2,8 by Schrödinger works by applying empirical corrections to “raw” pK_a values obtained using density functional theory (DFT) and a direct thermodynamic cycle. Prediction accuracy and a low dependence on the initial input conformation are achieved by thorough consideration of conformational effects on pK_a and excellent mean absolute error (MAE) values are obtained (0.2 pK_a units for a set of cyclic amidines). Because of the treatment of compounds as three-dimensional (3D) structures, their protocol is proven capable of discerning different pK_a values for two stereoisomers as well as being sensitive to steric effects on pK_a which their purely empirical counterpart, Epik, is not.17 QSPR-based protocols often implement a lower level of theory than first principles methods and are generally reliant on a number of chemical descriptors. These descriptors are regressed against experimental values to assemble predictive models using multiple linear regression (MLR) or principal component regression.16–18 Univariate regression models that rely on only one descriptor may be labelled linear free energy relationships (LFERs). The method we propose here falls into this latter category. Other successful attempts to predict pK_a values from one calculated descriptor include the use of the quantum theory of atoms in molecules (QTAIM) atomic energy of the dissociating proton.19 Partial atomic charges of the atom bonded to the acidic proton have also been shown to have a linear relationship with pK_a.20

Here, we reintroduce our protocol, which reveals the linear relationships that exist between our quantum mechanically derived molecular descriptor (bond length) and an experimental property (pK_a). The ab initio bond length high correlation subset (AIBLHiCoS) method works on the basis that chemical space may be partitioned into a series of highly correlated local linear relationships between pK_a in the solvent and a relevant bond length in the gas phase. Membership of a compound to a subset is governed by structural and chemical commonality, defined by structural motifs such as tautomeric form and conformation, or the presence or absence of substituents at a defined proximity to the acidic proton. The existence of these relationships and simple implementation of the linear equations describing them to predict pK_a values have been demonstrated for phenols,21 naphthols,22 guanidines,23 benzoic acids and anilines,24 aryl guanidines and 2-(arylimino)-imidazolidines,25 benzohydroxamic, phenylacetic, and phenoxy-acetic acids,26 bicyclo[2.2.2]octane and cubane carboxylic acids,27 and pyridines, thiols, and barbituric acids. We use a descriptor from a 3D structure (in the gas phase), and so consideration of steric, intramolecular hydrogen bonding (IHb) and π–π interactions is inherent to the protocol within the constraints of our gas-phase model. These structural features are known to affect the measured pK_a value, and their treatment is not inherent to “two-dimensional methods,” such as the empirical protocol employed by Marvin, unless explicitly included in the parameterization.

The 2014 review of nine pK_a prediction packages by Settimo et al.3 found that prediction accuracy was harder to achieve for bases than for other weakly acidic compounds. This prompted our lab to work on the notoriously difficult guanidine functional group.25 The work on guanidines is revisited here later at a new theory level to predict the pK_a of 12 α,2-adrenoceptor antagonist phenylguanidine derivatives. A recent study25 detailed the synthesis and reported the pK_a values for this set of aryl guanidines in addition to 23 2-(arylimino)imidazolidines, using both ultraviolet (UV)-metric and pH-metric techniques. AIBLHiCoS predictions were also reported using a single-bond-length model, but the implementation of the MLR model was not included. The pK_a values for further five small 2-(arylimino)imidazolidines, two of which are the marketed drugs tizanidine and apraclonidine, were also obtained from the literature to further test the model. We also compared our predictions to the literature values for both sites of ionization present within a set of 10 diphenyl-based bis(2-iminoimidazolidines), further revealing the efficacy of the method in discerning the relative magnitude of the two microscopic dissociation constants. The 2-(arylimino)imidazolidine derivatives studied here also contain the Y-shaped guanidine functional group and thus may also adopt three tautomeric forms in their neutral state. The approach to high correlation subset (HiCoS) formation for these compounds is therefore similar to the approach described for guanidines,25 in that the tautomeric forms are used as a primary classifier of structural commonality.
Figure 1. (a) Five conformer/tautomers considered, labelled A to E, where A&B and D&E are rotamers of the same tautomer while the active bond is shown in green; (b) three tautomeric forms of 2-aminoimidazoline R-substituted derivatives (T1, T2, and T3), shown here for 2-(aminophenyl)imidazoline, where the bond lengths under investigation are labelled a to g. The active bonds used to form predictive models in the following sections are again shown in bold/green.

2. METHODOLOGY

2.1. HiCoS Formation. The first stage in the AIBLHiCoS protocol is to deduce where a so-called “active bond” lies and in which tautomeric form and conformation it is located. The active bond is identified by it returning superior \( r^2 \), \( q^2 \), and root-mean-squared-error of estimation (RMSEE) values when regressed against \( pK_a \) compared to those of other bonds. As a descriptor, the active bond length must reflect the implications on geometry that a diverse array of substituent groups may impose.

Aromatic compounds have been shown to provide a simplistic and chemically intuitive example of how chemical space may be partitioned using the bond length as a descriptor. Several of our studies have shown that the relationship between the bond length and \( pK_a \) for meta- and para-substituted compounds is such that they may form their own subset. In the case of naphthols, phenols, benzoic acids, and anilines, the relationships established between \( pK_a \) and the bond length for simple meta/para-substituted molecules provide linear equations capable of predictions within 0.5 \( pK_a \) units for more complex biomolecules. Compounds containing ortho-substituents often fall into their own subset, where membership to a set may be defined by the specific chemical characteristic of the substituent group. The partitioning of ortho-substituted species according to the chemical nature of the substituent is justified by the propensity for IHBs between ortho-substituents and the atoms of the ionizable group. In the presence of bulky ortho-substituents, steric effects can impose constraints on bonding and dihedral angles adopted by the ionizable group at the ipso carbon. This has a knock-on effect on bond lengths, causing data points to appear as outliers in plots of \( pK_a \) versus bond length. IHBs cause rearrangement of electron density in the region of interest, meaning that bond lengths adjacent to where the proton is lost often deviate from the expected trend, relative to a set of compounds absent in the IHB interaction. This effect leads to the phenomenon of self-selection, an expression that describes how compounds exhibiting these respective geometric features fall into their own separate HiCoSs, at differing ranges of bond lengths or occasionally with differing active bonds.

The identification of an optimally predictive model is guided by a number of internal and external validation metrics. The mathematical formalism of the statistics used, that is, the squared correlation coefficient \( r^2 \), RMSEE, and the external validation statistic denoted \( q^2 \) (in the form of leave-one-seventh-out), are illustrated in a previous work.\(^{21}\) A high \( r^2 \) value (0.9 and above) describes a strong relationship between our descriptor and the experimental observable, but this may be misleading. Indeed, a line of best fit connecting clusters of data points may return a high \( r^2 \) value generating a model of poor predictivity. The latter property is best captured by the cross-validated correlation coefficient \( q^2 \) and the MAE. A high \( q^2 \) value is indicative of a highly predictive model, and the leave-one-seventh-out protocol used here typically overcomes the known pitfalls of the widely used leave-one-out \( q^2 \). The need to overcome such deficiencies has been demonstrated by Troshpa et al.\(^{15}\) Furthermore, Roy recommends\(^{28}\) employing an external dataset of at least 10 compounds as a true test of a predictive model and advocates MAE evaluation, taking into consideration the standard deviation of the absolute errors (AE) as well as the range of the training set. For a model to be classed as “good” according to Roy, it must pass the following criteria:

(i) MAE \( \leq 10\% \) of the training set range
(ii) MAE + 3 \( \times \) \( \sigma \) \( \leq 20\% \) of the training set range (where \( \sigma \) is the standard deviation of the AE).

The above validation criteria will be used to assess the predictive ability of the most promising models we have developed.

It must be noted that a caveat of the current methodology (and of all methods that use experimental values in their calibration protocol) is its dependence on the availability of experimental data. On this occasion, the experimental values used to calibrate the imidazolines model were measured at 20 °C rather than at a standard state of 25 °C. Because of lack of other available data in aqueous solution at 25 °C, the values from this paper\(^{29}\) by Timmermans and van Zwieten were used, with the possible detriment to prediction accuracy in mind. The general trend of \( pK_a \) variance with temperature for the ionizable group in question is that there is a small decrease (0.1–0.2 \( pK_a \) units) for every 5 °C increment in the temperature. This effect is shown in Table S1 in the Supporting Information, which details the change in the measured \( pK_a \) for seven compounds for the temperature range of 20–40 °C. For the diphenyl-based bis(2-iminomimidazoline) test compounds, the \( pK_a \) data are taken from one and the same publication\(^{30}\) and were recorded at 30 °C. As a temperature increase of 10 °C is expected to have a non-negligible effect on \( pK_a \), a correction was applied to the training set \( pK_a \) values. The linear equation describing the
pKᵢ change between 20 and 30 °C is used to correct the values of the original training set. To enhance the model, the test set of 2-(arylilino)imidazolidines first introduced in section 4.1.3 is also added to the initial training set, after a correction for the temperature change from 25 to 30 °C.

The experimental values used in this work correspond to the pKᵢ values for dissociation from the guanidinium-type ion. However, as in the protocol described by Griffiths et al. for guanidine derivatives, the bond lengths we use as descriptors in this work correspond to the neutral, conjugate base form.

2.2. Tautomers and Bond Lengths. Individual HiCoSs were constructed by Griffiths et al. for five bond lengths within the guanidine fragment for a training set of compounds with chemically diverse R groups. The C within the guanidine fragment for a training set of compounds previously were consistently poor. The analysis on the basis that correlation coefficients obtained were constructed by Gri et al. for guanidine derivatives, the bond lengths we use as descriptors in this work correspond to the neutral, conjugate base form.

To identify the dominant conformations of each compound in the gas phase for each of the three tautomers, we used the “conformers plug-in” within the Marvin Sketch program to generate a number of starting geometries. For each type of mono-ortho-substituted species of the imidazolidine training set, 15 conformations were taken as input structures for optimization and vibrational analysis using the Gaussian 09 program. The most stable equilibrium geometries within the ensembles of each compound, in each tautomeric form, were then identified by a comparison of total ab initio energies. As the ortho-groups are identical for each of the 2,6-di-ortho-compounds of the training set, degenerate conformations are expected because of the inherent symmetry of the system. These compounds also exhibit a lower degree of conformational flexibility than their mono-ortho-substituted analogue because of restricted rotation around C=N and N–C (illustrated for T3 in Figure 2). Noting these inherent geometric features, all 2,6-di-ortho compounds were optimized from an input geometry suspected to be close to their most stable form, whereby the two rings are orthogonal with respect to each other. For each mono-ortho-substituted species, optimization from an input geometry with orthogonal orientation of the two rings was also carried out. This separate optimization was performed to assess the effect of conformational commonality on the LFERs produced from our training set, which, as will be discussed, has limited variation with respect to substitution patterns.

HF/6-31G(d,p) optimized structures of guanidine training set compounds from the previous study were reoptimized at the B3LYP/6-311G(d,p) level of theory. This reoptimization was carried out for each compound in each of the five conformer/tautomer forms defined in the previous paper and which are shown in Figure 1a. The pKᵢ data for further six small guanidine derivatives were obtained and optimized in each tautomer/conformer combination. Frequency calculations were carried out using the same level of theory for optimization, and inspection of the Hessian confirmed the absence of imaginary frequencies and the location of energy minima.
3.2. Construction of Models. Relevant bond lengths were extracted (seven bonds a–g for imidazolidines and five labelled i–v for guanidines) from all optimized structures of the tautomERICally distinct subsets. The values of $\tilde{r}$, RMSEE, and leave-one-seventh-out $q^2$ were obtained using the program SIMCA-P 10.0 following linear regression of each type of bond length versus aqueous $pK_a$ values. The coefficients of the three-bond-length MLR model were also obtained using the SIMCA-P program. For all test set compounds, the most stable geometries were ascertained for each of the tautomeric forms using the Marvin conformer search protocol described above, where for the diphenyl-based bis(2-iminoimidazolidines), the number of conformers generated was increased to 25. The appropriate bond length was then extracted and inserted into the HiCoS equation $[pK_a = m \times (bond \ length) + c]$ or MLR equation for the models deemed to be optimally predictive according to internal and external validation statistics.

3.3. Interatomic Exchange–Correlation Energy Calculations. Bond lengths by themselves do not reveal their degree of covalency by their mere value. Covalency is essentially a measure of electronic delocalization, which can be calculated within the context of a topological energy partitioning method according to internal and external validation statistics.

4. RESULTS AND DISCUSSION

4.1. Imidazolidines. 4.1.1. Conformation and Commonality. In a previous work, we have demonstrated the existence of conformation-specific linear relationships between the bond length and $pK_a$. For instance, we illustrated that two subsets are formed using the same interatomic bonding distance for a series of ortho-substituted naphthols, when the O–H group is either syn or anti with respect to the ortho-substituent. Consequently, bond lengths later used to predict unknown $pK_a$ values must be extracted from the compound in the specific tautomer/conformer combination used to form the predictive equation. Therefore, insertion of an O–H bond length from a syn conformer into the anti HiCoS equation, or vice versa, would likely return a poor prediction. Figure 2 illustrates the two torsional angles that define the orientation of the imidazolidine group relative to the phenyl ring: $D_1$ ($C^1$–$C^2$–$C^3$) and $D_2$ ($C^1$–$C^4$–$C^7$–$C^2$). The latter term, $V_{xc}$, is associated with the Coulomb hole and the electrostatic repulsion between the electrons. The absolute value of $V_{xc}$ evaluated between two atoms can be taken as the extent of delocalization of electrons between them, a factor that also determines the bond distance. The values for the exchange–correlation energy between C and N (denoted $V_{CN}$) were obtained by the AIMAll program (version 14.11.23), using the first-ever $^{17}$ DFT-compatible IQA partitioning and using default parameters on wave functions obtained at the B3LYP/6-311G(d,p) level. Note that the $V_{CN}$ values are always negative, and the more negative they are, the more covalent is the interaction between C and N.

3.4. Experimental Data. Compounds of the training set are labelled 1–23 and can be found in Table S2 of the Supporting Information. The structural diversity of these compounds in terms of substitution pattern is fairly low. There are 14 compounds of the set that possess two ortho-substituents, 9 of which have two identical o-halogen groups (both are either F, Cl, or Br), whereas the other 5 have two identical o-alkyl groups (either Me or Et). Eight compounds of the training set have only one ortho-substituent, three of which have an o-methyl group and five have an o-chloro group. Meta- and para-substituents are also present on the phenyl group of 15 o-substituted compounds of the set. The only compound of the set lacking any ortho-substituents is the unsubstituted form (shown in Figure 2 as tautomer T3), which will be referred to as compound 23.

The test compounds discussed in sections 4.1.3 and 4.1.4 were synthesized previously, as reported by Rozas and co-workers ($ph_1$–$ph_{10}$ and $py_1$–$py_{7}$) and Dardonville and co-workers ($ph_{11}$–$ph_{15}$, 1, 2a,b–5a,b, and 6h). The substitution patterns of the test set include meta- and para-substituted compounds. The $pK_a$ values were measured by both the potentiometric (pH-metric) and spectrophotometric (UV-metric) methods of $pK_a$ determination. It is pH-metric values that are used for comparison to theoretical predictions for $ph_1$–$ph_{15}$ and $py_1$–$py_7$ in section 4.1.3, whereas both UV- and pH-metric values are used for comparison in section 4.1.4. Further details on the methods used for $pK_a$ determination can be found in section S2 of the Supporting Information.
is one in which the imidazoline group is orientated away from the ortho-chloro group. If the same class of compound is instead represented as T2, a preference for almost absolute coplanar geometry is observed, illustrated by averaged D1 and D2 angles of around 178° and −1°, respectively. In contrast to the observations made for T1, the most stable arrangement for T3 is one in which the N–H of the imidazoline group is orientated toward the chlorine atom. An IQA analysis of compounds 5 and 7 (the 2,4-chloro and 2,5-chloro derivatives) in their most stable conformation of tautomer T3 reveals no BCP for N–H···Cl. Therefore, the propensity to adopt this conformation cannot be ascribed to the presence of an IHB.

The commonly visited geometries of compound 23 [2-(phenylimino)imidazolidine] will be quite similar to those identified for any other compounds substituted at the meta- and/or para-positions for small substituent groups. This may be asserted because of the infeasibility of significant through-space interactions between imidazole and these groups at the more distant meta- and para-positions. Consequently, inclusion of this compound within the training set of a model will ensure that the applicability domain encompasses compounds without ortho-substituents. For T2, one dominant conformation is identified, within which the two rings lie almost exactly coplanar, akin to all mono-ortho-substituted derivatives in this form. An IQA analysis was performed to assess the reasons for the enhanced stability of the coplanar conformation for T2 and can be found in section S4 of the Supporting Information. For the T1 tautomer, there is also one dominant conformation that corresponds to a structure with a D1 angle of 160° and a D2 dihedral of −12°. For T3, there are two conformations that have similar D1 and D2 angles but differ by the orientations of the hydrogen atoms of the N–H groups, with a difference in stability of only ~1 kJ mol⁻¹. In the analysis below, only the bond lengths of the most stable T3 conformers are considered.

So far, we have identified preferred conformations for the compounds of our training set in each tautomeric form, which has revealed a degree of commonality for subsets of compounds possessing similar substituent patterns. We also observed an anomalous geometry due to an IHB for the 2,6-F₂ derivative (compound 15), which is likely to be a common feature among other 2-F-substituted compounds. We are now in a position to construct plots from the constituent bond lengths of each compound and assess how LFERs emerge for each bond. In the following analysis, we include bond lengths from the 2-substituted species in their most stable conformations of T1 and T2 and also from a geometry that more closely resembles the orthogonal orientation of the 2,6-substituted compounds. This is performed to assess whether conformational congruence is indeed the guiding factor to HiCoS formation or simply whether the most stable, and therefore, the most commonly visited geometries provide bond lengths that have the strongest linear relationships with pKₐ values.

### 4.1.2. Identification of HiCoSs

In the search for a predictive model for 2-(phenylimino)imidazolidines, our initial motivation was to formulate a model using bond lengths from the most commonly visited forms of each compound of our training set. According to chemical intuition, within the constraints of our gas phase and static model, an active bond should be located within the most commonly visited state, that is, the lowest energy conformation of the most stable tautomeric form. Furthermore, in accordance with the results for other aromatic species, we may then expect to find that subsets form, where membership to a given subset is defined by the location of a substituent at the ortho- or meta- and para-positions on the phenyl group.

In the most stable conformation of each of their tautomeric forms, the “imino” tautomer (T3), where the C=N double bond is present between the imidazoline and the phenyl rings, is consistently more stable than either of the T1 and T2 rotamers (internal energy rankings can be found in Table S2). This energy ranking is consistent for a comparison of Gibbs energies. Plotting bond lengths a–d versus pKₐ, with the 23 training set compounds in their most stable conformation of the T3 tautomer, reveals strong, individual linear relationships for all bonds (r² = 0.946, 0.944, 0.830, and 0.960 for a–d, respectively). However, for bonds a and b (the endocyclic C–N...
bonds of the guanidine fragment), one obvious outlier is observed in each case. These anomalous bond lengths are identified as belonging to compound 15, 2-(2,6-difluorophenylimino)imidazolidine (blue circles in Figure 3). We may explain the occurrence of this outlier by the presence of the IHB we described above, which perturbs the electron density distribution between atoms and causes a more extreme distortion of the geometry than that observed for the 2-chloro-substituted analogues. We may thus remove the outlier from the set because compound 15 lacks conformational commonality with the rest of the set. If enough data for other ortho-fluoro-compounds were available, then we may surmise based on a previous work, that a separate HiCoS would be identified for bonds a and b, consisting of compounds that have this intramolecular interaction in common. There is only one other compound with pK_s data available: the 2-F,6-Cl-phenyl analogue. As expected, this second 2-F derivative exhibits a conformation very similar to compound 15 (D1 = 51°) and also appears below the trend line for bonds a and b, next to compound 15 (orange triangles in Figure 3). However, deriving a new HiCoS equation from two data points is unlikely to provide a meaningful relationship, thus rendering an all-encompassing, single-bond-length model using a or b as unfeasible. Removal of compound 15 reduces the training set to 22 congeners and reveals r^2 values of 0.973 for bond a, 0.969 for b, 0.836 for c, and 0.959 for d, as shown in Table 1, thus making bond a the active bond for all compounds contained in the training set, except for 2-F derivatives.

| set       | bond     | r^2     | q^2     | RMSEE  |
|-----------|----------|---------|---------|--------|
| full      | MLR(a,b,d) | 0.977   | 0.970   | 0.204  |
|           | a        | 0.946   | 0.943   | 0.286  |
|           | b        | 0.944   | 0.943   | 0.290  |
|           | c        | 0.830   | 0.816   | 0.507  |
|           | d        | 0.960   | 0.951   | 0.246  |
| w/o 15    | MLR(a,b,d) | 0.983   | 0.982   | 0.157  |
|           | a        | 0.973   | 0.969   | 0.205  |
|           | b        | 0.969   | 0.967   | 0.218  |
|           | c        | 0.836   | 0.827   | 0.504  |
|           | d        | 0.959   | 0.954   | 0.252  |

*Bond lengths and D1 and D2 angles can be found in Table S4 in the Supporting Information.*

In light of the existence of more than one “active bond,” MLR analysis was performed for bonds a, b, and d versus pK_s on the grounds that their individual r^2 and q^2 values are in excess of 0.95. The c bonds were excluded, as a detrimental effect on MLR validation metrics can be expected in accordance with the inferior (r^2 and q^2 < 0.9) single-bond-length values shown in Table 1. This new model, where pK_s is now defined as a function of three bonds, exhibits superior validation statistics when compared to any single-bond-length model (Table 1). The removal of the anomalously long bonds of compound 15 from the training set revealed a further increase in r^2 and notably an increase in q^2 of 0.12 in addition to a 0.05 reduction in RMSEE.

For the higher energy tautomeric forms, T1 and T2, we observe no other superior set of r^2, q^2, and RMSEE values for the training set in their most stable conformations, compared to those observed for T3 a (Table 2). In the case of T2, the bond length versus pK_s plots reveal that for a highly correlated model to be constructed using all compounds, each mono-ortho-substituted compound must occupy a higher energy conformation, whereby the two rings are arranged orthogonally. Separate HiCoSs are formed in the case of each bond, when the bond lengths of the more stable, more coplanar conformers of mono-ortho-compounds are plotted against their pK_s values. Therefore, for T2, inducing conformational commonality is the guiding factor for a strong LFER; indeed the r^2, q^2, and RMSEE values for bonds a, c, and d in the same orthogonal conformation are indicative of an almost perfect linear relationship. The new subsets constructed from the bond lengths of the coplanar conformers are found to run almost parallel to the full set containing the orthogonally arranged
Table 2. Model Statistics for Bond Lengths $a$—$d$, Where Each Compound is Represented in the T1 and T2 Forms$^a$

| bond | geometry | $r^2$ | $q^2$ | RMSE (Å) |
|------|----------|------|------|---------|
| T2   | $a$      | 0.863| 0.800| 0.456   |
|      | same conf (excl. 23) | 0.988| 0.987| 0.137   |
|      | $b$      | 0.603| 0.553| 0.792   |
|      | same conf (excl. 23) | 0.924| 0.922| 0.342   |
|      | $c$      | 0.583| 0.540| 0.812   |
|      | same conf (excl. 23) | 0.989| 0.989| 0.130   |
|      | $d$      | 0.423| 0.424| 0.920   |
|      | same conf (excl. 23) | 0.974| 0.973| 0.199   |
| T1   | $a$      | 0.906| 0.898| 0.377   |
|      | same conf | 0.952| 0.944| 0.270   |
|      | $b$      | 0.608| 0.599| 0.769   |
|      | same conf | 0.885| 0.879| 0.416   |
|      | $c$      | 0.803| 0.781| 0.545   |
|      | same conf | 0.745| 0.729| 0.620   |
|      | $d$      | 0.463| 0.456| 0.900   |
|      | same conf | 0.819| 0.814| 0.522   |

$^a$“Most stable” refers to the models formed with 2,6-di-ortho-compounds plus the more stable coplanar conformers of the mono-ortho-compounds. “Same conf” refers to models where mono-ortho-compounds are optimized from an orthogonal orientation of the two rings, so as to match that of the 2,6-di-ortho-substituted compounds (i.e., all compounds optimized from the same conformation). Compound 23 falls within the “most stable” set as it did not optimize to a local minimum with an orthogonal geometry. Corresponding plots are shown in Figures S1 and S2, whereas Tables S5 and S6 show bond lengths and angles D1 and D2.

The MAE value of the $a$ model was found to be the lowest at 0.26, with a standard deviation of AE ($\sigma$) of 0.23. Whereas a reasonable performance is observed using the $d$ model equation (MAE = 0.57, $\sigma$ = 0.53), the $b$ model performed very badly (MAE = 1.54, $\sigma$ = 2.47). The poor performance of the $b$ and $d$ models can be mainly attributed to large prediction errors for compounds py4—py7, which correspond to four 2-pyridyl derivatives (structures are shown in Table 3). For example, the errors for bond $b$ for these four compounds are −7.49, −7.24, −7.28, and −7.31. These very low predictions can be attributed to a significant shortening of the $b$ bond, which equates to a −0.03 Å difference relative to those of 3-pyridyl analogues. The $d$ bond lengths also experience this shortening effect but to a lesser degree, as illustrated by the lower prediction errors of −1.79, −1.65, −1.76, and −1.66. However, the $a$ bond lengths show no substantial deviation from the bond length range observed for the training set compounds, and consequently, the errors for py4—py7 are dramatically lower (−0.27, +0.01, −0.07, and −0.46). Inspection of the most stable equilibrium geometries identified for these compounds reveals the presence of an IHB interaction, which lies between the N−H of imidazolidine and the heteroatom at the 2-position of the six-membered ring. The existence of this IHB may be evidenced first (tenuously) by the presence of a (3,−1) bond critical point between the imidazolidine H and the N atom of pyridine in the most stable geometries of py4—py7 and second by the analysis of $V_{\text{HN}}$ and $V_{\text{NN}}$ values. The above terms correspond to the HN and NN exchange-correlation energies calculated for py4—py7, where $V_{\text{HN}}$ denotes the (N)−H−N interaction and $V_{\text{NN}}$ denotes the N(−H)···N interaction. The average magnitude of these values is even larger than that calculated for the constituent atoms of the IHB found in the most stable conformer of compound 15 (T3), equating to ∼−47 kJ mol$^{-1}$ for $V_{\text{HN}}$ and ∼37 kJ mol$^{-1}$ for $V_{\text{NN}}$. When these four 2-pyridyl compounds are excluded, the MAE and $\sigma$ values are significantly reduced to 0.53 and 0.25 for the $b$ model, and 0.37 and 0.23 for $d$. However, as the superior validation statistics derived in the previous section suggest, the $a$ model predictions are consistently more accurate, with an MAE of 0.26 and a standard deviation of 0.24 for the remaining 23 compounds.

Although the averaged error for the test set predictions using the $a$ model is impressively low, there are some compounds of the set that show significant errors, namely, ph11 (−0.70), ph13 (−0.82), and to a lesser extent, ph15 (−0.64). These errors correspond to the compounds 2-(4-nitrophenylimino), 2-(3-chloro,4-nitrophenylimino), and 2-(3-fluoro,4-nitrophenylimino)imidazolidine, and therefore, all three con-
tain a nitro group at the para position. The strong, resonance-induced electron-withdrawing effect of p-nitro groups is evidenced by their measured $pK_a$ values, which indicate a reduction in the basicity of $\sim 2pK_a$ units for all nitro compounds (ph1–ph15), relative to the unsubstituted form ($23$). The prediction errors for the three problematic nitro compounds are somewhat reduced by use of the $b$ model ($+0.58$, $+0.40$, and $+0.70$), but significant improvement is found by use of the $d$ model, for which the predicted values are found to be mere $−0.23$, $−0.05$, and $+0.10$ $pK_a$ units outside of the experiment. It is interesting to note that compounds ph12 and ph14 also contain a p-nitro group, but show consistently low errors for all three single-bond-length models (Table S7). The reason for the lower errors for ph12 and ph14 may lie in the fact that these compounds also have o-halogen substituents in addition to the p-nitro group, and thus share more structural commonality with the compounds used to calibrate the models.

With the introduction of our test set, we notice that certain phenyl substituents, like the ortho-fluoro groups of compound 15, can perturb bond lengths $a$, $b$, and $d$ to varying degrees, meaning that outliers are found to be inconsistent between single-bond-length models. This is not a surprising observation; but subsequently, to derive a $pK_a$ prediction protocol that is both highly accurate and consistent in its accuracy, we must go from one single-bond-model to another to minimize overall prediction errors.

The resultant predictive equation for the MLR of bonds $a$, $b$, and $d$ against $pK_a$ is expressed as follows: $pK_a = 92.511 \times r(a) + 105.33 \times r(b) + 33.596 \times r(d) − 311.79$. Implementation of this equation to predict for all 27 test set compounds revealed it to be inferior in terms of overall accuracy relative to the single-bond-length model (MAE = 0.63, $\sigma$ = 1.08). However, this high mean error can again be solely attributed to poor predictions for the 2-pyridyl compounds, which show errors of $−3.24$, $−3.01$, $−3.09$, and $−3.21$ for py4–py7. Excluding the errors of the 2-pyridyl derivatives reveals an improvement in the prediction accuracy over the use of the $a$ model for the remaining compounds, as the MAE falls to 0.20 and notably, the standard deviation of errors is reduced to just 0.18. In noting the sensitivity of the $b$ and $d$ bonds to the presence of the IHB of 2-pyridyl compounds, it appears that the optimal protocol for the $pK_a$ prediction of this class of derivatives should be via implementation of the $a$ model equation. For compounds that lack the heteroatom at the 2-position of the aryl group, implementation of the MLR three-bond-length model will likely provide the most consistently accurate predictions. Because of a lack of $pK_a$ data at this time, it is not possible to say how this protocol will fare when the phenyl

| ID | Exp $[25^\circ C]$ | AIBLHiCoS T3 | AE | ChemAxon T3 | AE |
|----|-------------------|--------------|----|-------------|----|
| ph1 | 10.49             | 10.46        | 0.03 | 8.88        | 1.61 | 10.10 | 0.39 |
| ph2 | 10.29             | 10.26        | 0.03 | 8.61        | 1.68 | 9.80  | 0.49 |
| ph3 | 10.42             | 10.38        | 0.04 | 9.88        | 0.54 | 11.04 | 0.62 |
| ph4 | 10.50             | 10.46        | 0.04 | 9.89        | 0.61 | 11.05 | 0.55 |
| ph5 | 10.44             | 10.51        | 0.07 | 9.74        | 0.70 | 10.91 | 0.47 |
| ph6 | 10.78             | 10.29        | 0.49 | 9.90        | 0.88 | 10.96 | 0.18 |
| ph7 | 10.62             | 10.73        | 0.11 | 9.78        | 0.84 | 10.96 | 0.34 |
| ph8 | 10.17             | 10.16        | 0.01 | 7.77        | 2.40 | 8.97  | 1.20 |
| ph9 | 9.11              | 9.06         | 0.05 | 9.15        | 0.04 | 10.32 | 1.21 |
| ph10| 9.08              | 9.35         | 0.27 | 7.74        | 1.34 | 10.03 | 0.95 |
| ph11| 8.52              | 8.39         | 0.13 | 9.75*       | 1.23 | 9.75  | 1.23 |
| ph12| 7.29              | 7.38         | 0.09 | 8.93*       | 1.64 | 8.93  | 1.64 |
| ph13| 8.07              | 7.89         | 0.18 | 9.25*       | 1.18 | 9.25  | 1.18 |
| ph14| 7.41              | 7.37         | 0.04 | 8.64*       | 1.23 | 8.64  | 1.23 |
| ph15| 7.80              | 7.83         | 0.03 | 9.21*       | 1.41 | 9.21  | 1.41 |
| Tiz | 8.20              | 7.78         | 0.42 | 9.72        | 0.23 | 7.49  | 0.71 |
| Apra| 9.22              | 8.75         | 0.47 | 8.48        | 0.74 | 8.48  | 0.74 |
| AG3 | 11.10             | 10.83        | 0.27 | 9.71        | 1.39 | 10.94 | 0.16 |
| AG4 | 9.40              | 9.70         | 0.30 | 9.02        | 0.38 | 10.30 | 0.90 |
| AG5 | 11.30             | 11.14        | 0.16 | 10.11       | 1.19 | 11.36 | 0.06 |
| py1 | 8.99              | 9.53         | 0.54 | 9.34        | 0.35 | 10.45 | 1.46 |
| py3 | 10.33             | 10.58        | 0.25 | 9.44        | 0.89 | 10.63 | 0.30 |
| py4 | 9.45              | 9.18         | 0.27 | 8.74        | 0.71 | 9.83  | 0.38 |
| py5 | 9.47              | 9.48         | 0.01 | 8.68        | 0.78 | 9.82  | 0.35 |
| py6 | 9.68              | 9.61         | 0.07 | 8.49        | 1.19 | 9.63  | 0.05 |
| py7 | 8.97              | 8.51         | 0.16 | 8.16        | 0.81 | 9.30  | 0.33 |

"Predictions and errors are also shown for further five literature compounds: tizanidine (Tiz), apraclonidine (Apra), and AG3–AG5. Predictions for compounds py4–py7 were made by implementation of the $a$ model, that is, $pK_a = 272.76 \times r(a) − 368.08$, and are denoted by (a). Predictions and AE obtained using ChemAxon using both T1 and T3 as input structures are also shown for comparison, where asterisks denote predictions where the Marvin program recognized the T3 form as the dominant tautomer."
Figure 4. (a–d) Plots of $V_{\text{xc}}$ vs $pK_a$ for bonds a–d, showing the outliers to be consistent with the analogous bond length vs $pK_a$ LFERs. The $r^2$ values correspond to training set compounds only. Values are listed in Table S8.

A high degree of prediction accuracy in modern terms may be considered as within 1 $pK_a$ unit per prediction, with a MAE for a test set of less than 0.5 units. According to these conditions, the error statistics ($\text{MAE} = 0.20$, $\sigma = 0.18$) indicate that ABLHiCoS performs very well and does so consistently. For predictions made using the $pK_a$ predictor plug-in within the Marvin Suite by ChemAxon, the imino tautomer (T3) as the input structure, eight prediction errors exceed 1.0 $pK_a$ units ($\text{ph8, ph9, ph11 to ph15, and py1}$), and the MAE value is in excess of 0.5 units (0.72). This value deteriorates further when T1 is used as an input structure, for which 11 errors exceed 1 $pK_a$ unit. The MAE of 0.97 for T1 shown in Table 3 is inclusive of the error values for $\text{ph11–ph15}$, for which the program implicitly recognizes the dominant tautomeric form as T3. Therefore, on the basis of our energy rankings (Table S3), optimal prediction accuracy is dependent on the user having prior knowledge about the dominant tautomeric form. The ABLHiCoS protocol we have developed for these compounds finds its place here as a useful tool for predicting subtle substituent effects on $pK_a$ with a high accuracy. Conversely, the wider applicability radius of the Marvin program means it demonstrates lower prediction accuracy in the case of these compounds.

According to Roy’s stringent assessment of a QSRR/quantitative structure–activity relationship model, the predictions made using the MLR model may be classed as “good” in his own parlance, that is, the MAE of 0.20 is smaller than 10% of the training set range (0.39 units) and the MAE $+ 3 \times \sigma$ (0.20 + 3 × 0.18) is 0.74 and therefore lower than 20% of the training set range (0.78).

We now discuss the exchange–correlation measure $V_{\text{xc}}$ introduced in the Computational Methods section 3. Figure 4a–d shows the relationship between the $V_{\text{xc}}$ values corresponding to the CN interactions for each of the bond lengths $a$–$d$ and the $pK_a$ values. Referring to the $r^2$ values on plots (a–d), which are for the training set compounds only (blue diamonds), remarkably strong linear relationships are observed in the case of all four CN bonds. Taking the $V_{\text{xc}}$ value as a more physically meaningful metric, the greater the exchange–correlation energy between the constituent atoms of the $C==N$ bond $\epsilon$, along with a lesser interaction energy between the atoms of the three $C==N$ bonds $a$, $b$, and $d$ adjacent to it, the more basic the guanidine group. Furthermore, plots of how the $V_{\text{xc}}$ values of the training set correlate with their corresponding bond lengths reveal $r^2$ values of 0.990, 0.871, 0.869, and 0.989 for $a$–$d$. Thus, although the bond length is not necessarily a measure of covalency, it is a measure in the case of the 23 training set compounds and most strongly for bonds $a$ and $d$. This observation therefore heightens the significance of our bond length versus $pK_a$ LFERs, that is, the most active bond lengths $a$ and $d$ (which exhibit the superior validation statistics $r^2$, $q^2$, and RMSEE) most strongly reflect the extent of delocalization between the two bonded atoms.

As the $V_{\text{xc}}$ values of $a$ and $d$ are very accurately reflected in the corresponding calculated bond distances, compounds $\text{ph11, ph13,}$ and $\text{ph15}$ (purple triangles in Figure 4) once again appear as outliers for $a$ but not for $d$. For the 2-pyridyl compounds $\text{py4–py7}$ (red circles), the presence of the IHB is seen to cause divergence of the $V_{\text{xc}}$ values of $b$, $c$, and $d$ from the linear trends that reflect the most substituent effects with impressive accuracy, leading to the formation of new HiCoSs. Accurate $pK_a$ prediction would be therefore be possible by the use of the linear equations describing the relationships shown in Figure 4a–d, where calculated $V_{\text{xc}}$ values are used in place
of bond lengths. The latter alternative prediction method could follow a procedure similar to the one described above for bond lengths. Specifically, the MLR model could be used to predict for substituted phenyl or 3-pyridyl compounds, but the \( V_{sc}^{C-N} \) model must again be used for 2-pyridyl compounds, as they belong to separate HiCoSs.

As the internal and external validation metrics for the bond length models are already impressive, it is likely that there is little to be gained from a protocol that uses \( V_{sc}^{C-N} \) as molecular descriptors, when weighed against the added computational cost involved in calculating them. However, it is interesting to note that the highest \( r^2 \) value for \( V_{sc}^{C-N} \) values versus \( pK_a \) is now found for bond c, which is the C=N double bond. Figure 4 shows that the \( V_{sc}^{C-N} \) values for both CN bonds adjacent to the site of ionization (c and d) are both highly correlated with the ionization feasibility. As the extent of delocalization increases between atoms of bond c and decreases for the atoms of bond d, the ability of the conjugate guanidinium-type cation to stabilize the positive charge increases. This finding is perhaps more in line with chemical intuition than what we observe for the bond lengths, where the C=N bond gives the worst \( r^2 \), \( q^2 \), and RMSEE values when regressed against \( pK_a \).

Notwithstanding the dramatic increase in \( r^2 \) from 0.83 to 0.98 for bond c by use of the corresponding \( V_{sc}^{C-N} \) values, Figure 4 also shows no substantial deterioration for other bonds. In fact, all \( V_{sc}^{C-N} \) versus \( pK_a \) plots show \( r^2 \) values above 0.9. The interatomic property, \( V_{sc}^{AB} \), as a measure of delocalization of electrons between two atoms, is related to the delocalization index, a QTAIM-derived metric that describes the number of shared or exchanged electrons between two QTAIM atoms. Work44–46 by Matta and co-workers has implemented delocalization indices, along with localization indices (a measure of electrons localized within a QTAIM atom), to define a series of compounds as what they call a \( \zeta \)-matrix. The authors showed47 that the Frobenius distances between the elements of the matrix representations of a set of \( p \)-benzoic acids containing only delocalization indices and that of a reference species with the lowest \( pK_a \) value show a high \( r^2 \) value when regressed against their \( pK_a \) values. Significantly, the distances between the submatrices representing just the [COOH] molecular fragment were found to give a higher

| Index | Y | R¹ | R² | R³ | R⁴ |
|-------|---|----|----|----|----|
| 1     | C |     |    |    | UV-metric | AIBLHiCoS | ChemAxon | UV-metric | AIBLHiCoS | ChemAxon |
|       |   |    |    |    | \( pK_{a(1)} \) | \( pK_{a(2)} \) | \( AE \) | \( pK_{a(1)} \) | \( pK_{a(2)} \) | \( AE \) |
| 2a    | C | Cl |    |    | 8.30 | 8.34 | 0.04 | 7.48 | 0.82 | 9.60 | 10.00 | 0.40 | 8.51 | 1.09 |
| 2b    | C | F  |    |    | 8.30 | 8.40 | 0.10 | 7.20 | 1.10 | 9.70 | 10.00 | 0.30 | 8.48 | 1.22 |
| 3a    | C | Cl |    |    | 8.90 | 8.75 | 0.05 | 7.71 | 1.09 | 10.10 | 9.99 | 0.11 | 8.42 | 1.68 |
| 3b    | C | F  |    |    | 8.90 | 8.77 | 0.03 | 7.69 | 1.11 | 10.20 | 10.04 | 0.16 | 8.46 | 1.74 |
| 4a    | C | Cl |    |    | 8.50 | 9.07 | 0.57 | 7.75 | 0.75 | 10.10 | 9.61 | 0.49 | 8.38 | 1.72 |
| 4b    | C | F  |    |    | 8.50 | 9.10 | 0.60 | 7.76 | 0.74 | 9.70 | 9.67 | 0.03 | 8.38 | 1.32 |
| 5a    | C | Cl |    |    | 9.00 | 9.02 | 0.02 | 8.32 | 0.68 | 10.10 | 9.07 | 1.03 | 7.66 | 2.44 |
| 5b    | C | F  |    |    | 8.60 | 9.04 | 0.44 | 8.23 | 0.37 | 9.60 | 9.44 | 0.16 | 7.44 | 2.16 |
| 6h    | N |    |    |    | 10.00 | 9.12 | 0.88 | 8.45 | 1.55 | 7.20 | 9.76 | 2.56 | 7.81 | 0.61 |

"UV-metric values were recorded at 30 °C, and both the AIBLHiCoS and ChemAxon predictions were adjusted accordingly, where the AIBLHiCoS equation implemented was \( K_{a(1)} = 86.34 \times r(a) + 90.40 \times r(b) + 34.08 \times r(c) - 283.43 \). The pH-metric experimental values correspond to a temperature of 25 °C, and the predicted values are made using the equation \( pK_{a(2)} = 92.31 \times r(a) + 96.00 \times r(b) + 36.77 \times r(c) - 303.02 \). The pH-metric values for 4b and 5b are missing because of insufficient sample quantities."
correlation with pKₐ than the [OH] submatrices. This is rationalized by the fact that the pKₐ value reflects the ability of the COOH fragment to accommodate ionization. Comparisons may therefore be drawn between this study and the current work, where using Vₛₑₐₜ, we see that each bonding interaction of the guanidine fragment correlates with pKₐ. This observation may be rationalized by the fact that the pKₐ values for our 2-iminoimidazolindines reflect the thermodynamic feasibility of the whole guanidine fragment to accommodate the positive charge.

4.1.4. External Test Set: Bis(2-iminoimidazolindines). To test the robustness of the AIBLHiCoS protocol described above for larger, more conformationally flexible species, the pKₐ values for a series of diphenyl-based bis(2-iminoimidazolindine) compounds were taken as a new test set (Table 4). These compounds contain two 2-(phenylimino)imidazolindine moieties linked by an amide bond at the para-positions, for which the two guanidine-type sites of ionization exhibit distinct pKₐ values. As the values obtained by Rios Martinez et al.⁴⁰ were measured at a nonstandard temperature of 30 °C, the values of the training set were adjusted to account for the expected deviation in pKₐ. Initial training set values (1–23) were adjusted using a correction from 20 to 30 °C, and for consistency, the former test set (ph1 through py7) values were adjusted using the appropriate equation for an increase from 25 to 30 °C. The dataset of guanidine-containing compounds used to approximate the pKₐ change is shown in Table S1 of the Supporting Information, and the new values for the enhanced training set are shown in Table S9. Both the a model and the three-bond-length MLR model formed using a, b, and d bond lengths were implemented to assess their relative accuracy for this new test set. The bond lengths of the 2-pyridyl compounds (py4–py7) and those of the three other largest outliers identified in the previous section (ph11, ph13, and ph15) were removed from the new training set on the basis that they belong to a separate HiCoS for these specific bonds.

All results shown in Table 4 correspond to predictions made using the MLRs for both temperatures, as they were both found to provide lower overall errors and standard deviation of errors when compared to the a models (Table S10). The upper part of Table 4 (corresponding to the UV-metric values of Rios Martinez et al.⁴⁰) demonstrates a poor performance for the AIBLHiCoS model, which shows an MAE of 0.41 and a standard deviation of 0.59 for the 20 sites of ionization. Notably, the prediction errors for compound 6h are huge (pKₐ(1) = +0.88 and pKₐ(2) = +2.56). These large errors cannot be attributed to the lack of training for species containing pyridines, as there are three 3-pyridyl species (py1–py3) present within the enhanced training set. Furthermore, as the predicted values for the majority of the congeneric species in Table 4 show errors below the 0.5 log unit threshold, we cannot rationalize the highly inaccurate predictions for both pHₐ(1) and pHₐ(2) of 6h and 5a pKₐ(2) by the approximate nature of the temperature corrections applied to the training set data.

On the basis of the success of the MLR model for the test compounds described in the previous section, the reliability of the UV-metric values was questioned. Subsequently, new pH-metric measurements were obtained for all compounds for which adequate sample quantity was available. These new pH-metric values were recorded at the standard 25 °C. Unfortunately, new measurements for 4b and 5b were not possible because of insufficient sample quantities.

The new predictions and pH-metric values are shown in the lower part of Table 4, and a comparison of error evaluation statistics is summarized in Table 5. Comparing the new predictions at 25 °C to the pH-metric values, the unsigned errors for 6h are now only 0.37 and 0.33 pKₐ units away from the experiment, proving the previously reported experimental values to be erroneous. The prediction error for 5a pKₐ(2) is now also below 0.5 units and closer in magnitude to that of the chemically similar pKₐ(1) site of 2a. All prediction errors are now found to be below 0.5.

Implementation of Roy’s criteria to assess prediction errors against the pH-metric values indicates that the model may be classified as “good,” that is, the MAE of 0.27 is smaller than 10% of the training set range (0.37) and the MAE + 3 × σ (0.27 + 3 × 0.10) is 0.56 and therefore lower than 20% of the training set range (0.74). The MAE values produced utilizing Marvin exceed 1 log unit for both the UV-metric values at 30 °C and pH-metric values at 25 °C. Notably, 19 out of 20 values exhibit errors of more than 0.5 units for the UV-metric values, and 14 out of 18 for the pH-metric values. Further, the order of microdissociation steps for sites (1) and (2), described by the relative magnitude of the two measured pKₐ values, is correct on every occasion for AIBLHiCoS but is incorrect for 5a and 6h for the predictions made by Marvin’s pKₐ predictor (of ChemAxon).

4.2. Guanidines. 4.2.1. HiCoSs from Tautomers. It has already been established that for guanidines, distinct HiCoSs may be formed by representing compounds in specific conformations of each tautomeric form. We have also proven that this approach to HiCoS formation is successful for 2-iminoimidazolindines. In the case of the imidazolindines discussed in the previous section, the most highly predictive model was formed using a bond distance within the most stable tautomer. For guanidines where substitution only occurs at one amino/imino group, previous work details the formation of a model using the C==N bond lengths of equilibrium geometries of compounds in a higher energy tautomer state. The purpose of the current section is to prove that the model already established in the previous work still functions as a predictive model for a new test set, as we shift the level of theory from our previous CPU-triggered HF/6-31G(d) level to the current B3LYP/6-311G(d,p) level. Because this section of the work on guanidines has already been discussed in the paper by Griffiths et al., further details can be found in section 2.3 of that paper.²³

Several changes have been made to the previous training set to form the model used in this work. First, eight new guanidine derivatives were added during recalibration to further corroborate the location of the active bond. Second, some pKₐ values used here differ from those used previously. For
guanidine itself, we use a value of 13.6 as per the Foye’s Principles of Medicinal Chemistry. A value of 8.23 also replaces that previously used for acetylguanidine. Both of these new values are reported to have been measured at 25 °C and are lower by 0.1 units than the values used in the guanidine model previously, which were taken from a study by Albert et al. and reported to have been measured at 20 °C.

Five candidates for the active bond were chosen as N−H, C≡N, C−N, C−N, and N−H, which correspond to those marked respectively as i−v in Figure 1a. Two remaining N−H bonds were excluded from the analysis, the first of which is one of the primary amino N−H bonds because including both primary N−H bonds was deemed unnecessary because of the fact they are very similar in magnitude. The N−H bond distance corresponding to the nitrogen atom, which is also bonded to the R group, was also omitted from the analysis as it is not present in any of the training or test set structures. The N−R bond was also excluded because the large variety of R groups causes too substantial a variation in the bond length. Using a total of 17 compounds from 10 different sources, bond length versus pKₐ plot using the C≡N bond of tautomer A is found to return the highest r² value of 0.925 (see Table 6). The removal of five outliers, that is, methylguanidine, acetylguanidine, metformin, amiloride, and biguanide, reveals an r² value of 0.97. These outliers are not wildly anomalous, and their inclusion still allows for a correlation of >0.90. We may therefore only rationalize their removal from the training set by the possibility of small deviations in experimental values, which are inherent to the varied experimental techniques and the conditions in which they were measured. The final set of 13 compounds used to form the predictive model implemented in the next section are shown in Table 7 and in the Supporting Information Tables S11−S15, respectively, for tautomers A to E.

4.2.2. External Test Set. The pKₐ prediction software by ChemAxon is again implemented to assess the quality of our predictions relative to a popular choice of a predictor. Our chosen model, as discussed in the previous section, is formed from tautomer A. For a fair comparison, the same tautomer used for the AIBLHiCoS model is also used as input for ChemAxon, in addition to the most stable form, C. As the r² value and other model validation statistics for the bonds of tautomer C are inferior to tautomer A, it can be confidently assumed that they would give inferior predictions to our model using A and are therefore not included.

### Table 6. r² Values for Bond Length vs pKₐ Plots for the Five Candidate Tautomer/Conformer Forms

| Tautomer/Conf | N≡N−H (i) | N≡C (ii) | C−N(sec) (iii) | C−N(pr) (iv) | (pr)N−H (v) | C−N(pr) (iv') |
|---------------|-----------|----------|----------------|-------------|------------|--------------|
| Full set      |           |          |                |             |            |              |
| A             | 0.324     | 0.925    | 0.843          | 0.018       | 0.027      |              |
| B             | 0.064     | 0.650    | 0.649          | 0.083       | 0.010      |              |
| C             |           | 0.257    | 0.149          | 0.014       | 0.237      |              |
| D             | 0.062     | 0.172    | 0.848          | 0.798       | 0.677      |              |
| E             | 0.005     | 0.469    | 0.563          | 0.380       | 0.445      |              |

| w/o Outliers  |           |          |                |             |            |              |
| A             | 0.360     | 0.977    | 0.823          | 0.077       | 0.042      |              |
| B             | 0.037     | 0.644    | 0.524          | 0.088       | 0.302      |              |
| C             |           | 0.408    | 0.501          | 0.077       | 0.244      |              |
| D             | 0.064     | 0.328    | 0.852          | 0.887       | 0.813      |              |
| E             | 0.002     | 0.469    | 0.563          | 0.380       | 0.445      |              |

As the bond order changes throughout structures A−E, (im) is an imino-type nitrogen; (sec) denotes a secondary amino nitrogen; (pr) denotes a primary amino nitrogen. C lacks a single-bonded secondary nitrogen and has an additional primary amino-type, denoted (iv'). The r² and RMSEE values are also listed in Tables S11 to S15 of the Supporting Information.

### Table 7. Compounds and Corresponding Aqueous pKₐ Values (with Citations to Their Sources) Used To Form the Initial Guanidine Training Set

| ID  | Compound                        | pKₐ  |
|-----|--------------------------------|------|
| 1b  | guanidine                       | 13.60|
| 2b  | phenylbiguanide                 | 10.71|
| 3b  | ethylenebis(biguanide)          | 11.76|
| 4b  | famotidine                      | 6.78 |
| 5b  | arginine                        | 12.50|
| 6b  | N,N-dimethylguanidine           | 13.40|
| 7b  | N,N',N''-trimethylguanidine     | 13.90|
| 8b  | hydroxyguanidine                | 7.96 |
| 9b  | N'-benzoylcarbonohydrionic diamide | 7.94 |
| 10b | N'-phenylcarbonohydrionic diamide | 8.26 |
| 11b | N-(diaminomethylene)benzamide   | 6.98 |
| 12b | ethoxycarbonylguanidine         | 7.03 |
| 13b | methoxyguanidine                | 7.46 |

The ChemAxon prediction error for the C tautomer is very respectable (0.54, Table 8), although the MAE does fall short of the <0.5 units threshold by a small margin. AIBLHiCoS offers only a minor advantage, with a slightly lower MAE (0.29) and 3 out of 12 errors in excess of 0.5 units. However, because of the large range of our models used to predict for the guanidines (7.2), both of Roy’s criteria are satisfied: (1) the MAE of 0.29 is smaller than 10% of the training set range or 0.71 and (2) MAE + (3 × σ) = 1.23 is smaller than 20% of the training set range, which is 1.42. ChemAxon shows 5 out of 12 errors above the 0.5 threshold using the C tautomer, but this rises to 9 out of 12 when tautomer A is used as the input structure. Once again, the prediction accuracy is highly dependent on user input with respect to the tautomeric form.
prediction errors observed for two 3-pyridyls, PG1 and PG2 (+0.55 and −0.24), for which the pyridine’s nitrogen is too far away for H-bonding to be feasible. There are \( pK_a \) data available for 1-(2-pyridyl)guanidine and 1-(6-methyl-2-pyridyl)-guanidine, and when the bond lengths of their optimized structures are included in the C≡N versus \( pK_a \) plot using tautomer A, they also appear below the line of best fit, near to the data point for PG3, giving errors of +1.53 and +1.22, respectively. A new HiCoS cannot be constructed using only two data points, and so a predictive equation that better describes the LFER for these compounds will be constructed when adequate data can be procured.

5. CONCLUSIONS

We have demonstrated the existence of novel, highly correlated linear relationships between gas-phase equilibrium bond lengths within the guanidine fragment of 2-(phenylimino)-imidazolidine derivatives and their aqueous \( pK_a \) values. It has been established that the relationship between bond lengths and \( pK_a \) is strengthened for the amino tautomers, T1 and T2, when conformational congruence is ensured. However, three single-bond-length models with excellent validation statistics (\( r^2 > 0.95 \)) emerged for compounds in their most stable conformations of the imino tautomer, T3. For this tautomer, shorter calculated \( a \), \( b \), and \( d \) bond lengths and a longer C≡N \( c \) bond length than that observed for the unsubstituted species can be collectively considered, indicative of a compound with higher acidity.

Predictions made using the three \( a \), \( b \), and \( d \) CN single-bond-length models were then compared to those obtained via implementation of an MLR model built from all three bond lengths \( a \), \( b \), and \( d \) versus \( pK_a \). The MLR model was found to exhibit superior validation statistics over any single-bond-length model. Prediction errors using the MLR model were found to be less than 0.5 for most compounds, but four anomalously large errors were observed for four 2-pyridyl compounds. These larger errors can be attributed to anomalously short \( b \) and \( d \) bonds, which are explained in terms of the presence of an IHB between the N–H of imidazolidine and the heteroatom of 2-pyridine. A robust protocol is then established, whereby the single-bond-length model \( a \) should be used to predict for compounds with a heteroatom at the 2-position, but the MLR model should be implemented for all other classes of compounds shown in the training and test sets. This protocol allowed for an MAE and standard deviation for all 27 compounds of 0.20 and 0.18, respectively.

It must be noted that the power of this relationship especially revealed itself on a few occasions during this part of the work. First, as mentioned in a recent study, the initial experimental value for compound \( \text{ph}12 \) (8.13) meant that our \( a \) model prediction gave an error of 0.85 log units, which was puzzling when compared to the high accuracy of the prediction for the fluor analogue, \( \text{ph}14 \). However, remeasurements revealed the initial value to be erroneous, and the real value (7.29) fitted our initial prediction to within the 0.5 unit accuracy threshold. Furthermore, implementation of the T3 \( a \) model using an enhanced training set has revealed that some of the 20 UV-metric values reported for diphenyl-based bis(2-iminimidazolines) by Rios Martinez et al. are inaccurate, most dramatically in the case of compound \( \text{6h} \), for which the prediction errors were reduced from +0.88 and −2.56 to +0.37 and +0.33, upon remeasurement using the pH-metric technique. The MAE for the 16 remeasured pH-metric values was found to be 0.27 with a standard deviation of 0.10. The MAE, standard deviation of \( AE \) of both external test sets, and training set ranges mean that both models may be considered “good” in terms of their predictive ability according to Roy’s criteria. The AIBLHiCoS method provides average prediction errors that are superior to ChemAxon’s \( pK_a \) predictor in all cases tested here. Further, the order of microdissociation steps

| ID | \( r(C\equiv N) \) | Exp (25°C) | AIBLHiCoS \( pK_a \) | AE | ChemAxon \( pK_a \) | AE | ChemAxon \( pK_a \) | AE |
|----|-----------------|------------|----------------------|----|------------------|----|------------------|----|
| G1 | 1.27459         | 10.90      | 11.01                | 0.11 | 10.89            | 0.01 | 11.45            | 0.55 |
| G2 | 1.27506         | 11.36      | 11.22                | 0.14 | 9.73             | 1.63 | 10.40            | 0.96 |
| G3 | 1.27478         | 10.98      | 11.10                | 0.12 | 9.46             | 1.52 | 10.10            | 0.88 |
| G4 | 1.27490         | 10.98      | 11.15                | 0.17 | 10.55            | 0.43 | 11.18            | 0.20 |
| G5 | 1.27511         | 11.35      | 11.24                | 0.11 | 10.71            | 0.64 | 11.22            | 0.13 |
| G6 | 1.27524         | 12.01      | 11.31                | 0.70 | 11.12            | 0.89 | 11.76            | 0.25 |
| G7 | 1.27632         | 11.68      | 11.79                | 0.11 | 9.35             | 2.33 | 10.02            | 1.66 |
| G8 | 1.27432         | 11.03      | 10.89                | 0.14 | 10.21            | 0.82 | 10.84            | 0.19 |
| G9 | 1.27462         | 11.01      | 11.03                | 0.02 | 10.21            | 0.80 | 10.85            | 0.16 |
| PG1| 1.27366         | 10.05      | 10.60                | 0.55 | 9.92             | 0.13 | 10.54            | 0.49 |
| PG2| 1.27405         | 11.01      | 10.77                | 0.24 | 10.26            | 0.75 | 10.91            | 0.10 |
| PG3| 1.27571         | 10.47      | 11.52                | 1.05 | 8.98             | 1.49 | 9.58             | 0.89 |

| MAE | 0.29 | 0.92 | 0.54 |
| \( \sigma \) | 0.31 | 0.67 | 0.43 |
predicted by Marvin was incorrect in the case of three compounds, whereas AIBLHiCoS is found to be correct in every case.

Revisiting the previously established model for guanidine derivatives provides corroboration of our previous work at a new level of theory, which suggests once again that, counterintuitively, it is not the most stable gas-phase tautomer that provides the active bond(s). The MAE for the test set of phenylguanidine analogues employed here was 0.29, which is both lower than that obtained by the ChemAxon program and in accordance with Roy’s stringent MAE evaluation criteria and is evidence of a “good” model.

We also present our findings from the first IQA-based study in the context of these LFERs, where in our efforts to rationalize outliers, it was found that the V^ab values for bonding interactions were also found to correlate highly with pK_a. Furthermore, the active equilibrium bond lengths (those which correlate most highly with pK_a) also correlate highly to the extent of delocalization between the two bonded atoms. Further work will explore the meaning of this interesting finding in the context of our protocol.

**ASSOCIATED CONTENT**

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.8b00142.

Detailed discussion related to the choice of level of theory, experimental details of pK_a measurements, IQA Analysis to confirm the presence of an HBI in compound 15, and IQA Analysis to analyze the enhanced stability of the coplanar conformer of compound 23 (PDF)

**AUTHOR INFORMATION**

Corresponding Author

*E-mail: pla@manchester.ac.uk. Phone: +44 161 3064511 (P.L.A.)*

ORCID

Christophe Dardonville: 0000-0001-3935-1932

Paul L. A. Popelier: 0000-0001-9053-1363

Notes

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

P.L.A. thanks the EPSRC for Fellowship funding (EP/K005472), and B.A.C. thanks the BBSRC "iCASE" award BB/L016788/1 and Syngenta Ltd for funding her PhD studentship. C.D. thanks the Spanish Ministerio de Economía y Competitividad (grant SAF2015-66690-R).

**REFERENCES**

(1) Trapp, S. Plant Uptake and Transport models for Neutral and Ionic Chemicals. *Environ. Sci. Pollut. Res.* 2004, 11, 33–39.

(2) Bochevarov, A. D.; Watson, M. A.; Greenwood, J. R. Multiconformation, Density Functional Theory-Based pKa prediction in Application to Large, Flexible Organic Molecules with Diverse Functional Groups. *J. Chem. Theory Comput.* 2016, 12, 6001–6019.

(3) Settimo, L.; Bellman, K.; Knecht, R. M. A. Comparison of the Accuracy of Experimental and Predicted pKa Values of Basic and Acidic Compounds. *Pharm. Res.* 2014, 31, 1082–1095.

(4) Jensen, J. H.; Swain, C. J.; Olsen, L. Prediction of pKa Values for Druglike Molecules Using Semiempirical Quantum Chemical Methods. *J. Phys. Chem. A* 2017, 121, 699–707.

(5) ACD/Labs. ACD/pKa: Predict accurate acid/base dissociation constants from structure. http://www.acdlabs.com/products/percepta/predictors/pka/

(6) Eipi, 3.6; Schrödinger LLC, 2016.

(7) Marvin, 16.1.4.0; ChemAxon, 2016.

(8) Yu, H. S.; Watson, M. A.; Bochevarov, A. D. Weighted Averaging Scheme and Local Atomic Descriptor for pKa Prediction Based on Density Functional Theory. *J. Chem. Inf. Model.* 2018, 58, 271–286.

(9) Sutton, C. C. R.; Franks, G. V.; da Silva, G. First principles pKa calculations on carboxylic acids using the SMD solvation model: Effect of thermodynamic cycle, model chemistry, and explicit solvent molecules. *J. Phys. Chem. B* 2012, 116, 11990–12006.

(10) Liptak, M. D.; Shields, G. C. Accurate pKa Calculations for Carboxylic Acids Using Complete Basis Set and Gaussian-n Models Combined with PCPCM Continuum Solvation Methods. *J. Am. Chem. Soc.* 2001, 123, 7314–7319.

(11) Ho, J. Are thermodynamic cycles necessary for continuum solvent calculation of pKas and reduction potentials? *Phys. Chem. Chem. Phys.* 2015, 17, 2859–2868.

(12) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* 2009, 113, 6378–6396.

(13) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. Adding explicit solvent molecules to continuum solvent calculations for the calculation of aqueous acid dissociation constants. *J. Phys. Chem. A* 2006, 110, 2493–2499.

(14) Thapa, B.; Schlegel, H. B. Theoretical Calculation of pKas of Selenols in Aqueous Solution Using an Implicit Solvation Model and Explicit Water Molecules. *J. Phys. Chem. A* 2016, 120, 8916–8922.

(15) Pliego, J. R., Jr.; Riveros, J. M. Theoretical Calculation of pK_a Using the Cluster-Continuum Model. *J. Phys. Chem. A* 2002, 106, 7434–7439.

(16) Goodarzi, N.; Goodarzi, M. Prediction of the acidic dissociation constant (pKa) of some organic compounds using linear and nonlinear QSPR methods. *Mol. Phys.* 2009, 107, 1495–1503.

(17) Klopman, G.; Fercu, D. Application of the Multiple Computer Automated Structure Evaluation Methodology to a Quantitative Structure-Activity Relationship Study of Acidity. *J. Comput. Chem.* 1994, 15, 1041–1050.

(18) Yu, D.; Du, R.; Xiao, J.-C. pKa prediction for acidic phosphorus-containing compounds using multiple linear regression with computational descriptors. *J. Comput. Chem.* 2016, 37, 1668–1671.

(19) Adam, K. R. New Density Functional and Atoms in Molecules Method of Computing Relative pKa Values in Solution. *J. Phys. Chem. A* 2002, 106, 11963–11972.

(20) Uğur, İ.; 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, 2200–2213.

(21) Harding, A. P.; Popelier, P. L. A. pKa prediction from an ab initio bond length: part 2—phenols. *Phys. Chem. Chem. Phys.* 2011, 13, 11264–11282.

(22) Anstötter, C.; Caine, B. A.; Popelier, P. L. A. The AIBLHiCoS Method : Predicting Aqueous pKa Values from Gas-Phase Equilibrium Bond Lengths. *J. Chem. Inf. Model.* 2016, 56, 471–483.

(23) Griffiths, M. Z.; Alkorta, I.; Popelier, P. L. A. Predicting pK_a Values in Aqueous Solution for the Guanidine Functional Group from Gas Phase Ab Initio Bond Lengths. *Mol. Inf.* 2013, 32, 363–376.

(24) Harding, A. P.; Popelier, P. L. A. pKa prediction from an ab initio bond length: Part 3—benzoic acids and anilines. *Phys. Chem. Chem. Phys.* 2011, 13, 11283–11297.

(25) Dardonville, C.; Caine, B. A.; Navarro de la Fuente, M.; Martín Herranz, G.; Corrales Mariblanca, B.; Popelier, P. L. A. Substituent effects on the basicity (pKa) of aryI guanidines and 2-(arylimino)-imidazolides: correlations of pH-metric and UV-metric values with predictions from gas-phase ab initio bond lengths. *New J. Chem.* 2017, 41, 11016–11028.
(26) Alkorta, I.; Popelier, P. L. A. Linear free-energy relationships between a single gas-phase ab initio equilibrium bond length and experimental pKa values in aqueous solution. ChemPhysChem 2015, 16, 465–469.

(27) Alkorta, I.; Griffiths, M. Z.; Popelier, P. L. A. Relationship between experimental pK_a values in aqueous solution and a gas phase bond length in bicyclo[2.2.2]octane and cubane carboxylic acids. J. Phys. Org. Chem. 2013, 26, 791–796.

(28) Roy, K.; Das, R. N.; Ambure, P.; Aher, A. B. Be aware of error measures. Further studies on validation of predictive QSAR models. Chemosom. Intell. Lab. Syst. 2016, 152, 18–33.

(29) Timmermans, P. B.; van Zwieten, P. A. Dissociation constants of clonidine and structurally related imidazolidines. Arzneimittelforschung 1978, 28, 1676–1681.

(30) Harding, A. P.; Popelier, P. L. A. pKa prediction from an ab initio bond length: part 3–benzoic acids and anilines. Phys. Chem. Chem. Phys. 2011, 13, 11283–11293.

(31) Anstöter, C.; Caine, B. A.; Popelier, P. L. A. The AIBLHiCoS Method: Predicting Aqueous pKa Values from Gas-Phase Equilibrium Bond Lengths. J. Chem. Inf. Model. 2016, 56, 471–483.

(32) SIMICA-P: UMETRICS: Umea, Sweden, 2002.

(33) Blanco, M. A.; Martin Pendas, Á.; Francisco, E. Interacting Quantum Atoms: A Correlated Energy Decomposition Scheme Based on the Quantum Theory of Atoms in Molecules. J. Chem. Theory Comput. 2005, 1, 1096–1109.

(34) Popelier, P. L. A. Atoms in Molecules: An Introduction; Pearson Education: London, Great Britain, 2000.

(35) Bader, R. F. W. Atoms in Molecules: A Quantum Theory; Oxford University Press: Oxford, Great Britain, 1990.

(36) Keith, T. A. AIMAll, TL Gristmill Software, Overland Park, KS, USA, 2014.

(37) Maxwell, P.; Martin Pendas, Á.; Popelier, P. L. A. Extension of the interacting quantum atoms (IQA) approach to B3LYP level density functional theory (DFT). Phys. Chem. Chem. Phys. 2016, 18, 20986–21000.

(38) Rodríguez, F.; Rozas, I.; Kaiser, M.; Brun, R.; Nguyen, B.; Wilson, W. D.; García, R. N.; Dardonville, C. New bis(2-amino-imidazoline) and bisguanidine DNA minor groove binders with potent in vivo antitrypanosomal and antiplasmoidal activity. J. Med. Chem. 2008, 51, 909–923.

(39) Kelly, B.; McMullan, M.; Muguruza, C.; Ortega, J. E.; Meana, J. J.; Callado, L. F.; Rozas, I. α2-adrenoceptor antagonists: synthesis, pharmacological evaluation, and molecular modeling investigation of pyridinoguanidine, pyridino-2-aminoimidazoline and their derivatives. J. Med. Chem. 2015, 58, 963–977.

(40) Rios Martinez, C. H.; Nué Martinez, J. J.; Ebiloma, G. U.; de Koning, H. P.; Alkorta, I.; Dardonville, C. Lowering the pKa of a bisimidazoline lead with halogen atoms results in improved activity and selectivity against Trypanosoma brucei. Eur. J. Med. Chem. 2015, 101, 806–817.

(41) Hanif, M.; Zaman, M. Thiolation of arabinoyxylan and its application in the fabrication of controlled release mucoadhesive oral films. Daru. J. Pharm. Sci. 2017, 25, 6.

(42) Williams, D. A.; Lemke, T. L. Foye’s Principles of Medicinal Chemistry, 7th ed.; Lippincott Williams & Wilkins: Philadelphia, PA, 2012; p 352.

(43) Munk, S. A.; Harcourt, D.; Ambrus, G.; Denys, L.; Głuchowski, C.; Burke, J. A.; Kharlamb, A. B.; Marlapaz, C. A.; Padillo, E. U.; Runde, E.; Williams, L.; Wheeler, L. A.; Garst, M. E. Synthesis and Evaluation of 2-[(5-Methylbenz-1-ox-4-szin-6-yl)iminom]imidazoline, a Potent, Peripherally Acting α2 Adrenoceptor Agonist. J. Med. Chem. 1996, 39, 3533–3538.

(44) Sumar, I.; Ayers, P. W.; Matta, C. F. Electron localization-delocalization matrices in the prediction of pKa’s and UV-wavelengths of maximum absorbance of p-benzoic acids and the definition of superatoms in molecules. Chem. Phys. Lett. 2014, 612, 190–197.

(45) Matta, C. F. Modeling biophysical and biological properties from the characteristics of the molecular electron density, electron localization and delocalization matrices, and the electrostatic potential. J. Comput. Chem. 2014, 35, 1165–1198.

(46) Matta, C. F. Molecules as networks: A localization-delocalization matrices approach. Comput. Theor. Chem. 2018, 1124, 1–14.

(47) Foye, I. Foye’s Principles of Medicinal Chemistry, 7th ed.; Lippincott Williams & Wilkins: Baltimore, 2013; p 1476.

(48) Verdolino, V.; Cammi, R.; Munk, B. H.; Schlegel, H. B. Calculation of pKa values of nucleobases and the guanine oxidation products guanidinohydantoin and spiroiminodihydantoin using density functional theory and a polarizable continuum model. J. Phys. Chem. B 2008, 112, 16860–16873.

(49) Albert, A.; Goldacre, R.; Phillips, J. The strength of heterocyclic bases. J. Chem. Soc. 1948, 2240–2249.

(50) Shaw, J. W.; Grayson, D. H.; Rosaz, I. Synthesis of Guanidines and Some of Their Biological Applications. In Guanidines as Reagents and Catalysts, 50th ed.; Selig, P., Ed.; Springer: Switzerland, 2017; pp 1–51.

(51) Bjerrum, J.; Schwarzenbach, G.; Sillen, L. G. Stability Constants; The Chemical Society: London, 1958; p 8.

(52) Degim, T.; Zaimoglu, V.; Akay, C.; Degim, Z. pH-Metric log K calculations of famotidine, naproxen, nizatidine, ranitidine and salicylic acid. I Farmaco 2001, 56, 659–663.

(53) Carrick, L. M.; Agelli, A.; Boden, N.; Fisher, J.; Ingham, E.; Waigh, T. A. Effect of ionic strength on the self-assembly, morphology and gelation of pH responsive β-sheet tape-forming peptides. Tetrahedron 2007, 63, 7457–7467.

(54) Davis, T. L.; Elderfield, R. C. The Determination Of The Ionization Constants Of Guanidine And Some Of Its Alkylated Derivatives. J. Am. Chem. Soc. 1932, 54, 1499–1503.

(55) Perrin, D. D. IUPAC: Dissociation Constants of Organic Acids and Bases In Aqueous Solution; Butterworths: London, 1965; p 446.

(56) Taylor, P. J.; Wait, A. R. Of Values for heterocycles. J. Chem. Soc., Perkin Trans. 2 1986, 1765–1770.

(57) Borek, E.; Clarke, H. T. Compounds Related to Calanine and Canavanine. J. Biol. Chem. 1938, 125, 479–494.