TOWARDS A GENERAL DESCRIPTOR FOR THE PREDICTION OF PKA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AYŞE GÖÇER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
CHEMISTRY

AUGUST 2022

Approval of the thesis:

TOWARDS A GENERAL DESCRIPTOR FOR THE PREDICTION OF PKA

submitted by **AYŞE GÖÇER** in partial fulfillment of the requirements for the degree of **Master of Science in Chemistry Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** ________________

Prof. Dr. Özdemir Doğan
Head of Department, **Chemistry** ________________

Assist. Prof. Dr. Antoine Marion
Supervisor, **Chemistry Department, METU** ________________

Examining Committee Members:

Assist. Prof. Dr. Burcu Dedeoğlu
Chemistry Department, Gebze Technical University ________________

Assist. Prof. Dr. Antoine Marion
Chemistry Department, METU ________________

Prof. Dr. Ali Çırpan
Chemistry Department, METU ________________

Assoc. Prof. Dr. E. Görkem Günbaş
Chemistry Department, METU ________________

Assist. Prof. Dr. Erol Yıldırım
Chemistry Department, METU ________________

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Ayşe Göçer

Signature :

ABSTRACT

TOWARDS A GENERAL DESCRIPTOR FOR THE PREDICTION OF PKA

Göçer, Ayşe
M.S., Department of Chemistry
Supervisor: Assist. Prof. Dr. Antoine Marion

August 2022, 58 pages

Estimating pKa of molecules with accurate, general, and efficient protocols is essential for many applications, notably for proteins modelling. For example, although individual amino acids pKa is known in aqueous solutions, it might be changed in the protein environment and in general in a different environment. Desolvation, hydrogen bonding, or charge charge interactions can cause such pKa shift, and it is often crucial to explain the bonding preferences of molecules during the chemical reaction. There are various ways to predict pKa values with approaches such as QM based, MM based, continuum based, and knowledge based methods. These approaches, however, tend to lack generality and are usually tailored to predict the pKa of certain classes of molecules only. Our aim is to identify a general electronic descriptor that would grasp the essence of what makes proton affinity vary from one molecule to another, and from one environment to another. Our protocol is based on descriptors initially designed to characterize topological changes in molecular electron density upon electronic excitation. By repurposing these tools, we aim at exploring the reorganization of the electron density upon de-protonation to eventually rationalize the concept of pKa with clear-cut electronic descriptors that can be transferred to various molecular environments, i.e., from the gas phase to a protein interior.

Keywords: pKa, quantum chemistry, ab-initio, acidity

ÖZ

PKA TAHMİNİ İÇİN GENEL BİR TANIMLAYICIYA DOĞRU

Göçer, Ayşe
Yüksek Lisans, Kimya Bölümü
Tez Yöneticisi: Dr. Öğr. Üyesi. Antoine Marion

Ağustos 2022 , 58 sayfa

Moleküllerin pKa'sını doğru, genel ve verimli protokollerle tahmin etmek, birçok uygulama için, özellikle protein modellemesi için gereklidir. Örneğin, tek tek pKa amino asitleri sulu çözeltilerde bilinmesine rağmen, protein ortamında ve genel olarak farklı bir ortamda değişebilir. Desolvasyon, hidrojen bağı veya yük yükü etkileşimleri bu tür pKa kaymasına neden olabilir ve kimyasal reaksiyon sırasında moleküllerin bağlanma tercihlerini açıklamak genellikle çok önemlidir. QM tabanlı, MM tabanlı, süreklilik tabanlı ve bilgi tabanlı yöntemler gibi yaklaşımlarla pKa değerlerini tahmin etmenin çeşitli yolları vardır. Bununla birlikte, bu yaklaşımlar genellikten yoksun olma eğilimindedir ve genellikle yalnızca belirli molekül sınıflarının pKa'sını tahmin etmek için uyarlanmıştır. Amacımız, proton afinitesinin bir molekülden diğerine ve bir ortamdan diğerine değişmesini sağlayan şeyin özünü kavrayacak genel bir elektronik tanımlayıcı belirlemektir. Protokolümüz, başlangıçta elektronik uyarma üzerine moleküler elektron yoğunluğundaki topolojik değişiklikleri karakterize etmek için tasarlanmış tanımlayıcılara dayanmaktadır. Bu araçları başka bir amaca uygun hale getirerek, pKa kavramını çeşitli moleküler ortamlara, yani gaz fazından bir protein içine aktarılabilen kesin elektronik tanımlayıcılarla rasyonalize etmek için de-protonasyon üzerine elektron yoğunluğunun yeniden düzenlenmesini keşfetmeyi amaçlıyoruz.

Anahtar Kelimeler: pKa, kuantum kimya, ab-initio, asitlik

This study is dedicated to humanity so that, maybe, one day, they can learn how to live in peace.

ACKNOWLEDGMENTS

I would like to thank my supervisor Assist. Prof. Dr. Antoine Marion for his time, support, patience, and understanding about me for both my undergraduate project and masters study. Being one of his research group members is really awesome. I am very grateful for his guidance throughout the research.

I would like to express my thankfulness to Assoc. Prof. Dr. Thibaud Etienne from the Université de Lorraine, Nancy, France for his very helpful and informative discussions.

I also would like to thank Prof. Dr. Paul W. Ayers from McMaster University for giving me a great opportunity to study with him. I really appreciate his time and resources throughout the project.

I would like to give my special thanks to my lovely family as well. I really appreciate all their support and love towards me.

I find myself lucky to have Tuğba Sönmez, Meltem Melisa Çetin, Heba and Menna Mohamed who are my lovely close friends. They are great people and I am very thankful for all their friendships.

The numerical calculations reported in this study were performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

QM	Quantum Mechanics
SE	Schrodinger Equation
H	Hamiltonian operator
T	Kinetic Energy
V	Potential Energy
BO	Born-Oppenheimer
E	Energy
HF	Hartree-Fock
SCF	Self Consistent Field
CI	Configuration Interaction
CC	Coupled Cluster
CASSCF	Complete Active Space Self Consistent Field
DFT	Density Functional Theory
D/A	Detachment/Attachment
SP	Single Point

CHAPTER 1

INTRODUCTION

1.1 Motivation and Problem Definition

pKa is a fundamental quantity to determine the acid strength and dissociation constant of molecules, and it determines relative reactivity and stability. It is critical for several branches of science, such as chemistry, biochemistry, and drug design.[1]

The strength of an acid is relative, and so, pKa (also known as acid dissociation constant) is also a relative quantity. It is always to be compared with the other acids/bases in the direct environment.

In solution, an acid AH dissociates into its conjugate base A^- and a proton H^+ as follows:

$$AH^q_{soln} \xrightarrow{\Delta G_q} A^{q-1} + H^+_{soln} \tag{1.1}$$

q is the formal charge of AH, and $q - 1$ is the formal charge of A. Acid dissociation is often studied in aqueous solutions, but equation 1.1 is a general expression for all solvents.

The acidity constant K_a for infinitely diluted solutions is:

$$K_a = \frac{[A]\,[H]}{[AH]} \tag{1.2}$$

and pKa is calculated as follows:

$$pK_a = -log(K_a) \tag{1.3}$$

Experimentally, potentiometry[2], conductometry[3], electrophoresis[4], NMR [5, 6, 7], UV[8], HPLC[9], and fluorometry[10] could be performed to determine pKa values.[11] However, for very strong or very weak acids, experimental pKa determination is hard and has more considerable uncertainties. Because there are multiple tautomeric species in the protonation/de-protonation reaction or when the species' concentration in solution approaches the limits of quantification, building a good model for pKa helps to understand what really makes one molecule more acidic than the other and can trigger the rational design of new molecules or help predict the behavior of a known molecule in a new environment.[12]

In general, for computational chemistry methods, pKa is used rather than Ka constant.

$$K_a(calc) = e^{-\frac{\Delta G_a(calc)}{RT}} \tag{1.4}$$

and an error in prediction of Ka thus, δK_a is:

$$\delta K_a(calc) \approx e^{-\frac{\Delta G_a}{RT}} \delta G_a(calc) \tag{1.5}$$

Because error of predictions (δK_a) depends exponentially on free energies ΔG_a, this exponential ΔG_a dependence increases the error on K_a predictions. When ΔG_a is low, δK_a (the error in the Ka) increases exponentially. [12]

However, in pKa calculation:

$$pK_a(calc) = \frac{\Delta G_a(calc)}{RTIn(10)} \tag{1.6}$$

and an error in prediction of pKa is:

$$\delta pK_a(calc) \approx \frac{\delta \Delta G_a(calc)}{RTIn(10)} \tag{1.7}$$

the pKa values are proportional to the dissociation free energies ΔG_a but, the errors on the predictions of pKa do not depend on the magnitude of the free energies as in K_a. Thus, the error δpKa does not depend on how acidic is the studied species AH.[12]

Theoretical pKa determinations could be classified as QM based methods[13, 14, 15, 16, 17, 18], MM based methods[19, 20, 21, 22, 23], continuum solvent-based methods[24, 25, 26, 27, 28, 29], and knowledge-based methods.[30, 31, 32]

Gibbs free energy method is one of the commonly used methods for pKa determination. However, with Gibbs free energy method, due to the high charge of the hydronium ion, the change in free energy of solvation of hydronium ion is hard to calculate. Also, adding water molecules to the equation requires more calculations, which is very challenging. [33] To overcome previous methods' limitations, several methods have been tried to build up. The methods based on de-protonation energy [34][35] and *ab initio* bond length [36] are used to make pKa prediction as well. Another study[37] uses the maximum surface electrostatic potential ($V_{S,max}$) on the acidic hydrogen atoms of carboxylic acids to make a pKa prediction.

Ugur et al. investigated the possibility of using atomic partial charges to predict pKa of alcohols and thiols[38]. The study was later extended by Haslak et al. [39] for carboxylic acids. In their approach, the authors calculated the charge on the oxygen (or sulphur) atom after de-protonation. The idea follows our general chemistry understanding of what makes a molecule to have a lower pKa, i.e., if the charge in the de-protonated form is delocalized in the molecule, the stability of the basic form increases.

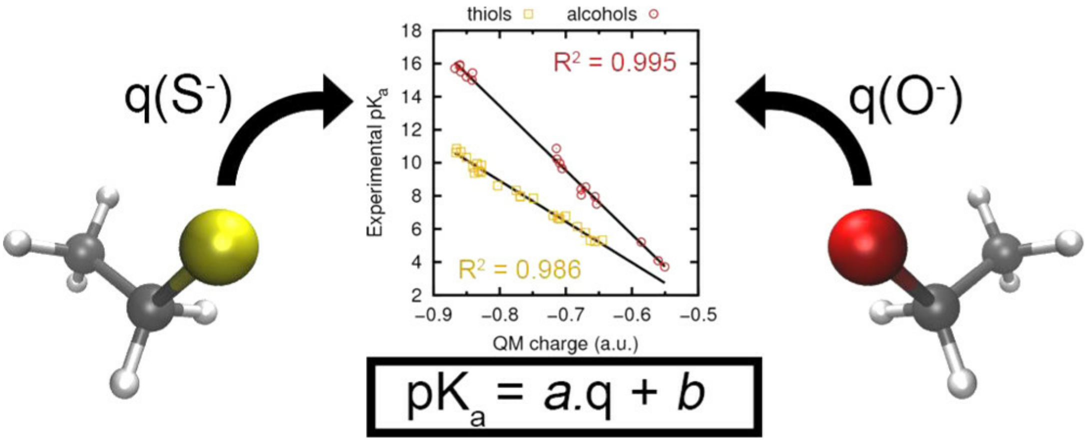


Figure 1.1: Correlation of experimental pKa vs. QM charge[38]

Figure from Ugur et al.1.1 demonstrates the linear relationship between experimental pKa and quantum mechanical (QM) charge for a series of alcohols and thiols [38]. The study investigated 9 DFT functionals with 16 different basis sets, semiempirical Hamiltonians, five charge models, three solvent models, and gas-phase calculations. As can be seen, the linear fit is remarkable with correlation coefficients of $R^2 = 0.995$ and $R^2 = 0.986$ for alcohols and thiols, respectively. Although the method is very powerful to predict pKas, it is limited to a single family of molecules at a time. Also, since the partial charge is not a physically observable, there is no unique way to determine it (and the equation to be used in the fit is different for each model). Thus, there can be several methods to predict it, so the results come from different theories (DFT, HF, MP, etc.) or basis sets give different values.

1.1.1 Aim

The limitation observed in the method by Ugur et al. is common to many parameterized methods for the prediction of pKa, i.e., there is no single general equation that can predict the pKa of different chemical functions. Nevertheless, this study highlighted a promising result showing that the electron density contains information about the relative reactivity of the molecules. Our motivation in this study is to find a more general electronic descriptor that can be used for the pKa prediction of any chemical function. A well-designed and well-understood descriptor will ultimately give us the key to designing molecules with tailored acidity.

CHAPTER 2

METHODOLOGY

2.1 Quantum Chemistry

In quantum chemistry[40], non-relativistic and time-independent, Hamiltonian operator $(\hat{H})$ for N electron and M nuclei system, in atomic units, can be written as follows:

$$\hat{H} = -\sum_{i=1}^{N}\frac{1}{2}\nabla_i{}^2 - \sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A{}^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{A=1}^{M}\sum_{B>A}^{M}\frac{(Z_A)(Z_B)}{R_{AB}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} \tag{2.1}$$

The first two terms in the above equation are kinetic energy operators $(\hat{T})$ of electrons and nuclei, respectively, and the last three are potential energies $(\hat{V})$. The third term is electron-nucleus interaction, and the fourth one is nucleus-nucleus interaction. The last term is electron-electron repulsion energy. The r and R letters refer to the distance between electrons and nucleus, respectively. Z_A, and Z_B are the charges of nuclei A and B.

The above equation is solvable for only few systems, such as particle in a box, harmonic oscillator, rigid rotor, and H atom. Because in other, more realistic cases, the Hamiltonian is not separable, the equation cannot be solved analytically and requires several approximations.

Born-Oppenheimer's (BO) approximation is fundamental for quantum chemistry calculations. The approximation says that due to the mass of a nucleus being significantly heavier than a mass of an electron, we can assume that nuclei are fixed at a particular position, and electrons are moving around them. In other words, the masses of nuclei and electrons are very different, and their motion can be decoupled. That is, when solving the electronic structure, we can take the position of the nuclei as fixed. By decoupling the electrons and nuclei's motion, the nucleus-nucleus term is now a constant for a given position of the nuclei, and the electron-nucleus attraction term is fairly simple. It is just the interaction of one electron with a field of fixed positive charges. Now the problematic term becomes the electron-electron repulsion term because the motion of electrons is coupled.

Thus, by applying BO approximation, just the following terms are left: the kinetic energy of electrons, electron-nucleus attraction, and electron-electron repulsion. After these simplifications, we have an electronic Hamiltonian, which depends just on electrons' degree of freedom.

$$\hat{H}_{elec} = -\sum_{i=1}^{N} \frac{1}{2}\nabla_i{}^2 - \sum_{i=1}^{N}\sum_{A=1}^{M} \frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}} \tag{2.2}$$

This electronic Hamiltonian $\hat{H}_{elec}$, which consists of three operators, can be separated as one electron and two-electron operators. The operator which contains just one electron degree of freedom is called the one-electron operator (h_{core} in the equations below). The kinetic energy, $\hat{T}$, of electrons and electron- nucleus potential energy, $\hat{V}_{en}$, are one-electron operators. Similarly, the two-electron operator includes the degrees of freedom of two electrons. Thus, the operator which corresponds to electron-electron attraction is called the two-electron operator.

$$\hat{H}_{elec} = \hat{h}_{core} + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}} \tag{2.3}$$

$$\hat{h}_{core} = -\sum_{i=1}^{N} \frac{1}{2}\nabla_i{}^2 - \sum_{i=1}^{N}\sum_{A=1}^{M} \frac{Z_A}{r_{iA}} \tag{2.4}$$

Due to the non-separable nature of the electron-electron interaction, the N-electron Hamiltonian cannot be solved analytically, and numerical solutions are needed. There are ample approximations to treat this electron-electron interaction, which can also be named an electron correlation issue. They can be divided into wavefunction-based and density-based methods.

For wavefunction-based methods, Hartree-Fock (HF) theory is central.

2.1.1 Hartree-Fock (HF) Theory

In HF theory[40], electron correlation is treated in an average manner. It is said that each electron feels an average potential due to other electrons.

Fock operator, $\hat{f}_1$, is;

$$\hat{f}_1 = \hat{h}_{core} + \hat{v}^{HF}(1) \tag{2.5}$$

$\hat{v}^{HF}(1)$ is the Hartree-Fock potential to define electron-electron interaction. It consists of two terms, Coulomb and Exchange terms, $\hat{\mathcal{F}}_b(1)$ and $\hat{\kappa}_b(1)$, respectively.

$$\hat{v}^{HF}(1) = \sum_b \hat{\mathcal{F}}_b(1) + \hat{\kappa}_b(1) \tag{2.6}$$

$$\hat{f}_1 = \hat{h}_{core} + \sum_b \hat{\mathcal{F}}_b(1) + \hat{\kappa}_b(1) \tag{2.7}$$

Hartree-Fock potential, $\hat{v}^{HF}(1)$, which consists of Coulomb and Exchange parts, is an average potential. Because the interaction is treated in an average manner, there is no explicit correlation between the motion of the electrons.

$\hat{\mathcal{F}}_b(1)$ is Coulomb operator;

$$\hat{\mathcal{J}}_b(1) = \int dx_2 |\chi_b(2)|^2 r_{12}^{-1} \tag{2.8}$$

$$\hat{\mathcal{J}}_b(1)\chi_a(1) = \left[\int dx_2 \chi_b^*(2) r_{12}^{-1} \chi_b(2)\right] \chi_a(1) \tag{2.9}$$

The possible interpretation of Coulomb operator when $\hat{\mathcal{J}}_b(1)$ acts on $\chi_a(1)$ is the potential electron 1 feels in spin-orbital χ_a due to existence of electron 2 in spin-orbital χ_b.

$\hat{\kappa}_b(1)$ is Exchange operator which is;

$$\hat{\kappa}_b(1)\chi_a(1) = \left[\int dx_2 \chi_b^*(2) r_{12}^{-1} \chi_a(2)\right] \chi_b(1) \tag{2.10}$$

The exchange term, however, does not have a simple classical interpretation. It comes from the anti-symmetry nature of a single determinant. The name exchange is due to the fact that electron 1 and electron 2 are changed relative to the Coulomb operator. ($\hat{\kappa}_b(1)\chi_a(1) = \left[\int dx_2 \chi_b^*(2) r_{12}^{-1} \boldsymbol{\chi_a(2)}\right] \boldsymbol{\chi_b(1)}$, but

$\hat{\mathcal{J}}_b(1)\chi_a(1) = \left[\int dx_2 \chi_b^*(2) r_{12}^{-1} \boldsymbol{\chi_b(2)}\right] \boldsymbol{\chi_a(1)}$).

The power of this approximation is that it converts many-body problems into one-body problems. However, it is a crude approximation due to no explicit electron-electron interaction in HF. Also, due to this average potential energy for electron-electron interaction, which depends on the positions of other electrons, HF theory is non-linear and must be solved iteratively.

2.1.2 Self Consistent Field (SCF) procedure

SCF[40] is an iterative procedure to obtain optimal orbitals. HF, DFT, Configuration Interaction (CI), Coupled Cluster (CC) theory, and Complete Active Space SCF (CASSCF) are examples of SCF methods.

The procedure is started with the determination of a trial (guess) wave function. Charge density is calculated using these functions, and potential energies are calculated. Then Schrodinger equation, SE, is solved, and the charge density is calculated again. If the calculated density is the same as the previous guess one, the SCF procedure is stopped. However, if not, the procedure continues using this different charge density, and potential energies will be calculated again. By solving SE, a new charge density will be calculated. The procedure stops if the resulting density is the same as the previous one up to a given convergence threshold. If not, these steps are followed until reaching a convergence.

2.1.3 Roothaan-Hall equation

In chemistry, we deal with molecules and use molecular orbitals (MO), so their calculations are important. The calculation of molecular orbitals is equal to the solution of the spatial integro-differential equation 2.11 below. For only atomic calculations, numerical solutions are available and practical for this equation 2.11, but there are no practical calculations available for molecules. Roothaan and Hall showed independently how the differential equation 2.11 could be converted by using known spatial basis functions (ϕ_μ here) into a linear algebra problem that provides standard matrix techniques for a solution.[41, 40]

$$f(r_1)\Psi_a(r_1) = \epsilon_a \Psi_a(r_1) \tag{2.11}$$

If we expand K unknown Ψ molecular orbitals in terms of known basis functions ϕ_μ with the expansion coefficients C as follows:

$$\Psi_a = \sum_{\mu=1}^{K} C_{\mu a}\phi_\mu \qquad a = 1, 2, 3, ..., K \tag{2.12}$$

and by putting this equation into the above equation 2.11, we get the equation 2.13. By the expansion of Ψ in terms of basis functions ϕs, HF MO calculations become "finding the coefficients C".

$$f(1)\sum_{\nu} C_{\nu a}\phi_\nu(1) = \epsilon_a \sum_{\nu} C_{\nu a}\phi_\nu(1) \tag{2.13}$$

By multiplying the above equation by $\phi_\mu^*(1)$:

$$\sum_\nu C_{\nu a} \int \phi_\mu^*(1) f(1) \phi_\nu(1) dr_1 = \epsilon_a \sum_\nu C_{\nu a} \int \phi_\mu^*(1) \phi_\nu(1) dr_1 \tag{2.14}$$

we have two new quantities which are $\int \boldsymbol{\phi_i^*(1)\phi_j(1)dr_1}$ and $\int \boldsymbol{\phi_\mu^*(1)f(1)\phi_\nu(1)dr_1}$. In quantum chemistry calculations, basis sets are not orthogonal, so we have overlap matrix S. Elements of overlap matrix S can be calculated as follows:

$$S_{ij} = \int \phi_i^*(1)\phi_j(1)dr_1 \tag{2.15}$$

Matrix elements of Fock matrix F :

$$F_{\mu\nu} = \int \phi_\mu^*(1) f(1) \phi_\nu(1) dr_1 \tag{2.16}$$

Finally the equation 2.14 becomes :

$$\sum_\nu F_{\mu\nu} C_{\nu a} = \epsilon_a \sum_\nu S_{\mu\nu} C_{\nu a} \qquad a = 1, 2, 3, ..., K \tag{2.17}$$

We can write the above equation 2.17 in a more compact form, and this is the Roothaan-Hall equation :

$$\boldsymbol{FC = \epsilon SC} \tag{2.18}$$

F is the Fock matrix, C is eigenfunctions of F (coefficients matrix), ϵ is the vector of eigenvalues of F, and S is the overlap matrix between basis functions ϕ_μ. The columns of matrix C describe the molecular orbitals. The first column of C consists of the coefficients of Ψ_1; likewise, the second column of C consists of the coefficients of Ψ_2 and so on.

C is the KxK square matrix of MO coefficients:

$$\text{C} = \begin{pmatrix} C_{11} & C_{12} & C_{13} & \dots & C_{1k} \\ C_{21} & C_{22} & C_{23} & \dots & C_{2k} \\ C_{31} & C_{32} & C_{33} & \dots & C_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{k1} & C_{k2} & C_{k3} & \dots & C_{kk} \end{pmatrix}$$

ϵ is a diagonal matrix that consists of K eigenvalues of F (orbital energies) on a diagonal plane. The first eigenvalue (ϵ_1) is for the first MO, and similarly, ϵ_2 is the second MO in the same order with eigenfunctions.

$$\epsilon = \begin{pmatrix} \epsilon_1 & & 0 \\ & \ddots & \\ 0 & & \epsilon_k \end{pmatrix}$$

2.1.4 Density Functional Theory (DFT)

DFT differs from wave function-based methods. It does not require finding the wave functions, DFT calculates the energy from electron density.[42]

Energy and other physical properties of a system can be derived from the electron density of the system.

Electron density, ρ, can be calculated as follows;

$$\rho = N \int ... \int |\psi(x_1, x_2 ... x_N)|^2 dx_1 dx_2 ... dx_N \tag{2.19}$$

N is the number of electrons. If we integrate the electron density ρ overall space, we get the total number of electrons N.

$$N = \int \rho(r) dr \tag{2.20}$$

Energy in DFT is calculated as:

$$E[\rho(r)] = T[\rho(r)] + E_{ne}[\rho(r)] + E_{ee}[\rho(r)] \tag{2.21}$$

$$E_{ee}[\rho(r)] = J[\rho(r)] + K[\rho(r)] \tag{2.22}$$

$$E^{DFT}[\rho(r)] = T_s[\rho(r)] + E_{ne}[\rho(r)] + J[\rho(r)] + E_{xc}[\rho(r)] \qquad (2.23)$$

Because the energy depends on ρ, which depends on coordinates, it is said that energy is density functional, which means the function of a function.

If the correct $E_{xc}[\rho(r)]$ was known, $E[\rho(r)]$ would be exact. Although we do not know the correct form of $E_{xc}[\rho(r)]$, there are ample approximations for it, such as Local Density Approximation (LDA), Local Spin Density Approximation (LSDA), Generalized Gradient Approximation (GGA), and hybrid functionals. All of them have different strengths and weaknesses.[43][44]

One of the strengths of DFT is no need to solve Schrodinger's equation, and no direct knowledge of wave functions is needed. Due to computational cost and efficiency, DFT is a commonly used method for quantum chemistry calculations for large molecules (approximately $20-100?$ atoms) or for the dynamics of small molecules.[45]

2.2 Analysis of electron density reorganization - MESRA

The analysis of electron density reorganization is of particular importance in the field of excited states. [46][47] Gaining a clearlcut understanding of the phenomenon at play during the interaction of light with matter helps in the design of new molecules with tailored properties. Among the numerous approaches, the concept of detachment and attachment densities is particularly appealing.

One of the recent application of D/A density is activity-based photosensitizers. In the paper, Kilic et al. investigated impact of iodine atom and its position on the electronic transition nature of two dicyanomethylene-4H-chromene (DCM) cores (which are $DCM_O - I$ and $I - DCM_O - Cl$). D/A density concept and related descriptors were discussed.

Molecular Electronic Structure Reorganization Analysis, MESRA [46], was created by Thibaud Etienne. Mesra is capable of calculating detachment/attachment density matrices, construction of TDDFT auxiliary many-body wave function, construction of three types of transition orbitals, quantitative electronic transition analysis, analysis of the post-linear response detachment/attachment relaxation, adiabatic connection of the Z-vector, computation of a relaxation-adapted transition locality index and manipulation of Cartesian grids for numerical analysis. However, we only use detachment/attachment density matrices for our purposes.

Mesra calculates detachment/attachment (D/A) densities from excitation process. Detachment density is the portion of electron density that is removed from the ground state during the electronic excitation. Attachment density is the rearranged portion of density in the excited state. D/A densities can be seen in the Figure 2.1 below from reference [46].

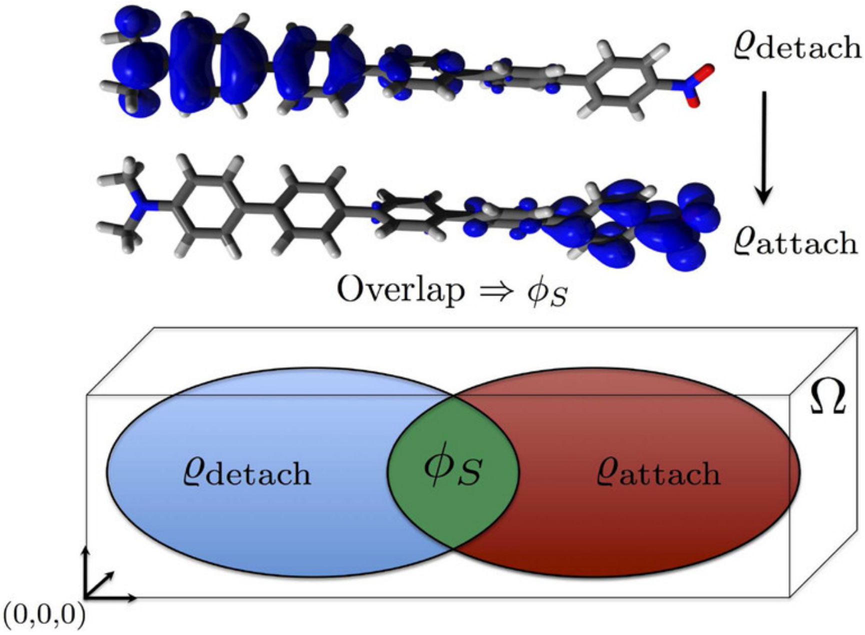

$\varrho_{\mathrm{detach}}$
$\varrho_{\mathrm{attach}}$
Overlap $\Rightarrow \phi_S$
Ω
$\varrho_{\mathrm{detach}}$
ϕ_S
$\varrho_{\mathrm{attach}}$
(0,0,0)

Figure 2.1: Graphical representation of Detachment density(blue region)/Attachment density(red region) overlap (green region)[46]

As can be seen from Figure 2.1, during the electronic transition, some portion of electron density moves away from the ground state (from ϱ_{detach} in the figure) and reorganizes at the excited state (from ϱ_{attach} in the figure). We can visualize their overlap region ϕ_S as the green region in Figure 2.1.

Mesra produces the terms overlap between detachment attachment densities (ϕ_S), Hole-particle overlap, Integral of D/A density, and Fraction of D/A contributing to the net displaced charge.

We have modified Mesra to achieve the differences in density between protonated and de-protonated states; thus, the modified Mesra works by taking the protonated state density as an originally ground state density and taking the de-protonated state density from the initially excited state density. Eventually the code can work now with any two different densities coming from two different calculations. The only requirement is that the basis set used for both calculations should be the exact same (in size and actual position of the functions' centers).

2.3 Mathematical details

In this section, we give some fundamental details about the mathematics used to derive D/A densities. Those require matrix diagonalization and linear algebra, which we briefly summarize hereafter.[46]

The diagonalization is a linear algebra operation related with the eigenvalue problem for matrices.

If γ is a L x L matrix

$$\gamma = \begin{pmatrix} \gamma_{1,1} & \gamma_{1,2} & \cdots & \gamma_{1,L} \\ \gamma_{2,1} & \gamma_{2,2} & \cdots & \gamma_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{L,1} & \gamma_{L,2} & \cdots & \gamma_{L,L} \end{pmatrix} \tag{2.24}$$

To diagonalize the γ matrix, finding the approximate U matrix so that $U^{\dagger}\gamma U$is diagonal require.

$$\gamma = \begin{pmatrix} \gamma_{1,1} & \gamma_{1,2} & \cdots & \gamma_{1,L} \\ \gamma_{2,1} & \gamma_{2,2} & \cdots & \gamma_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{L,1} & \gamma_{L,2} & \cdots & \gamma_{L,L} \end{pmatrix} \quad \underrightarrow{U} \quad \gamma^{diag} = \begin{pmatrix} \gamma_{1,1} & 0 & \cdots & 0 \\ 0 & \gamma_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_{L,L} \end{pmatrix} \tag{2.25}$$

thus,

$$\gamma^{diag} = U^{\dagger}\gamma U \tag{2.26}$$

and U is an eigenfunctions of γ matrix and $U^{\dagger}$ is a complex conjugate and transpose of U matrix. The problem of the diagonalization of a matrix is equal to the issue of finding the unitary matrix U that converts a non-diagonal matrix into a diagonal matrix. There are L eigenvalues of γ if it is L x L matrix. If γ is a matrix with both positive and negative eigenvalues, we can define the two new matrices, which consists of only negative or non-negative elements. After giving some fundamentals related to this study, the calculations of the descriptors can be seen as follows:

In a system that consists of N electrons and a K-sized basis set with S represents the overlap matrix, P represents the density matrix, P_0 represents the ground state density matrix, and P_x represents the excited state density matrix. The difference between excited-state electron density matrix and ground-state electron density matrix is Δ and is defined as:

$$\Delta = P_x - P_0 \Rightarrow \sum_{k=1}^{K} \Delta S_{kk} = 0 \tag{2.27}$$

The trace i.e., the sum of all the diagonal elements of a matrix, of ΔS_{kk} is zero because there is no electron gain or loss by the system.

By making the Δ matrix a diagonal matrix δ with unitary similarity transformation, we get eigenvalues of Δ

$$\exists\ U \mid \delta = U^{\dagger}\Delta U;\ \ \delta ij = 0 \qquad \forall i \neq j \tag{2.28}$$

By diagonalizing the Δ matrix, one can get detachment and attachment matrices. Δ has positive and negative eigenvalues. We can define two new diagonal matrices, which are σ_+ and σ_-, whose components are functions of those from δ.

$$(\sigma_\pm)_{kj} = \frac{1}{2}(\sqrt{(\delta_{kk}^2}) \pm \delta_{kk}) \times \delta_{kj} \quad \forall \delta = \sum_{\omega=+,-} \omega \sigma_\omega \tag{2.29}$$

where δ_{kj} is the Kronecker delta function and $(\delta)_{kk}$ is a component of the δ matrix. The σ function splits Δ's eigenvalues into two arrays: σ keeps the absolute value of negative δ values and sets non-negative ones to zero. Conversely, σ_+ contains only positive values from the δ diagonal and zeros. Backtransforming δ, σ_- and σ_+ gives

$$\Delta = U\delta U^\dagger \tag{2.30}$$

$$\Gamma = U\sigma_- U^\dagger \tag{2.31}$$

$$\Lambda = U\sigma_+ U^\dagger \tag{2.32}$$

and

$$\Delta = \Lambda - \Gamma \tag{2.33}$$

Λ and Γ are respectively the attachment and detachment density matrices, and they are expressed in the space of K atomic orbitals Ψ_μ . Due to the fact that no electron is lost during the electronic excitation;

$\sum_{\mu=1}^{K} (\Gamma S)_{\mu\mu} = \sum_{\mu=1}^{K} (\Lambda S)_{\mu\mu}$

It is also possible to define the related detachment and attachment densities in the real 3D space (ζ_1, ζ_2, and ζ_3 being the three spatial coordinates)

$$\varrho_\tau(\zeta_1, \zeta_2, \zeta_3) = \sum_{\mu=1}^{K} \sum_{\nu=1}^{K} \tau_{\mu\nu} \Psi_\mu(\zeta_1, \zeta_2, \zeta_3) \Psi_\nu^*(\zeta_1, \zeta_2, \zeta_3) \tag{2.34}$$

$$\tau \equiv \Gamma, \Lambda \tag{2.35}$$

We might define detachment/attachment charge ϑ_τ:

$$\vartheta_\tau = \int\limits_R d_{\zeta 1} \int\limits_R d_{\zeta 2} \int\limits_R d_{\zeta 3} \varrho_\tau(\zeta_1, \zeta_2, \zeta_3) \equiv \int\limits_{R^3} d_\zeta^3 \varrho_\tau(\zeta) \tag{2.36}$$

$$\tau \equiv \Gamma, \Lambda \tag{2.37}$$

ϑ_τ is the integral of the density (D or A), which gives a charge. This is the charge that is involved in the reorganization of the density during the process of interest. ϑ is the average of ϑ_Γ and ϑ_Λ. Detachment charge and attachment charge (ϑ_Γ and ϑ_Λ) should be exactly the same. However, for numerical reasons, they are not exactly the same in practical calculations.

Among many other ways to characterize the D and A densities, one can define the spacial overlap between the two densities. This is given by the descriptor ϕ_S [46] and is defined as follows:

$$\phi_S = \vartheta^{-1} \int\limits_{R^3} \sqrt{\varrho_\Gamma \varrho_\Lambda}\, d_\zeta^3 \quad \phi_S \in [0; 1] \tag{2.38}$$

$$\vartheta \equiv \frac{1}{2} \int\limits_{R^3} d_\zeta^3 \sum_{\tau=\Gamma,\Lambda} \varrho_\tau(\zeta) \tag{2.39}$$

The ϕ_S index ranges from 0 to 1, depending on the electronic transition's charge-transfer character. When ϕ_S is "0", this means no overlap between detachment and attachment densities. Contrarily, "1" corresponds to the extreme case where a zero electronic density changes between the state of interest and the reference state (e.g., excited and ground state, in excited states calculations).

From charge density variation ϱ_Δ, one can obtain the transferred charge between final and initial state electron density ϱ_x and ϱ_0 respectively. By defining ϱ_Δ, ϱ_+ and ϱ_- functions as follows:

$$\varrho_\Delta(\zeta) = \varrho_x(\zeta) - \varrho_0(\zeta) \tag{2.40}$$

$$\varrho_\pm(\zeta) = \frac{1}{2}(\sqrt{\varrho^2_\Delta(\zeta)} \pm \varrho_\Delta(\zeta)) \tag{2.41}$$

Finally, the amount of transferred charge χ (we later name it as a qCT) can be calculates as fallows:

$$\chi_\omega = \int\limits_{R^3} \varrho_\omega(\zeta)\, d^3_\zeta \tag{2.42}$$

$$\chi = \frac{1}{2} \sum_{\omega=+,-} \varrho_\omega \tag{2.43}$$

ϑ is total charge involved in the transition. qCT (χ in the equations to make the formula consistent with the source article) is the amount of charge effectively moved from initial to final state during the process. Both ϑ and qCT are also investigated in recent study.

To summarize, qCT tells you how much charge is actually relocated during the transition, and ϕ_S tells you how local is this relocation. Large values of qCT tell that an important amount of charge was involved in transition. Low value of ϕ_S indicates that this charge was relocated far from where it was taken. If ϕ_S is large, then the charge was rearranged in the same space.

2.4 Computational Details

The calculations during the study were done by Gaussian 09 software because the tool we re-purpose, which calculates D/A overlap, runs just for Gaussian outputs.

As a first step, geometry optimization is done for the protonated form of the molecule by adding the following line in the input file;

#P ωb97xd/6-311+G* SCRF=CPCM opt=tight freq

After having optimized geometry, single point (SP) calculations are done for the protonated and de-protonated forms of the molecule. The example line for SP calculations for protonated and de-protonated forms of the molecules can be seen below;

#P ωb97xd/6-311+G* SCRF=CPCM

The only differences between protonated and de-protonated forms are that the charge in the input file of the de-protonated form should be negative 1 (when if the molecule is neutral in the protonated form, if the molecule has already charged, then the new charge should be "initial charge - 1"). Also, in de-protonated form, the Hydrogen atom, which binds to the Oxygen atom, acidic H, should be replaced by a ghost atom by adding "H-Bq" instead of H. The purpose of using a ghost atom is to have the exact same basis functions for the protonated and the de-protonated forms to compare their densities and calculate the difference density matrix. If we removed H atom directly, the number of basis functions for the protonated/de-protonated forms would not be equal, so comparing their densities would not be possible.

A few files are necessary as an input of MESRA and those can be generated as follow. During the SP calculation, one can add the following lines that should be placed in input files ("name.chk" file will be converted "name.fchk" file by using formchk utility of Gaussian.):

```
chk=name.chk
rwf=name.rwf
```

The "name.fchk" and "name.rwf" are generated by using formchk and rwfdump utilities of Gaussian for both protonated and de-protonated forms as it can be seen below, and they are necessary files for Mesra[48] to calculate D/A overlap;

```
formchk name.chk
```

```
rwfdump name.rwf overlap 514r
```

```
rwfdump name.rwf density 633r
```

To run Mesra, one should have "protonated-name.fchk", "de-protonated-name.fchk", "overlap", and "density" files that come from the formchk and rwfdump utilities of Gaussian in the directory.

Mesra allows making several different calculations for distinct purposes. When having the necessary files to we can run the Mesra, the following line is used;

```
mesra dau2
```

"dau2" keyword is for computation and storage of the detachment/attachment density matrices, and the evaluation of density-based descriptors using Löwdin-like detachment/attachment population. Löwdin population analysis is one of the atomic orbital-based population analyses. It uses a transformation of all the atomic orbitals to an orthogonal basis.[49] "dau2" is our new keyword to take densities coming from two different ground state calculations. It is a modified version of "dau", which takes ground state and excited states densities from a single file.

After having D/A densities from "dau" keyword, we can start to analyze them by using "qmni" keyword (e.q. to get ϕ_S). "qmni" keyword is for computing density overlap, normalized charge displacement, and centroids from two cube files. After the analysis is performed, one can generate 3D representations of the D and A densities as cube files with the following command:

```
cubegen 0 fdensity=scf name.fchk name.cube -X h
```

The cube files can be analyzed with the qmni module of mesra in order to calculate the descriptors described in the previous section (i.e., ϕ_S, ϑ, qCT) :

```
mesra qmni cube1 cube2
```

These steps were done for all molecules in our training sets.

CHAPTER 3

RESULTS

3.1 General observations and trends

pKa is associated with de-protonation reaction and it is related to relative reactivity and stability of acidic chemical functions. In de-protonation reaction, a Bronsted acid gives its proton and becomes negatively charged ion.

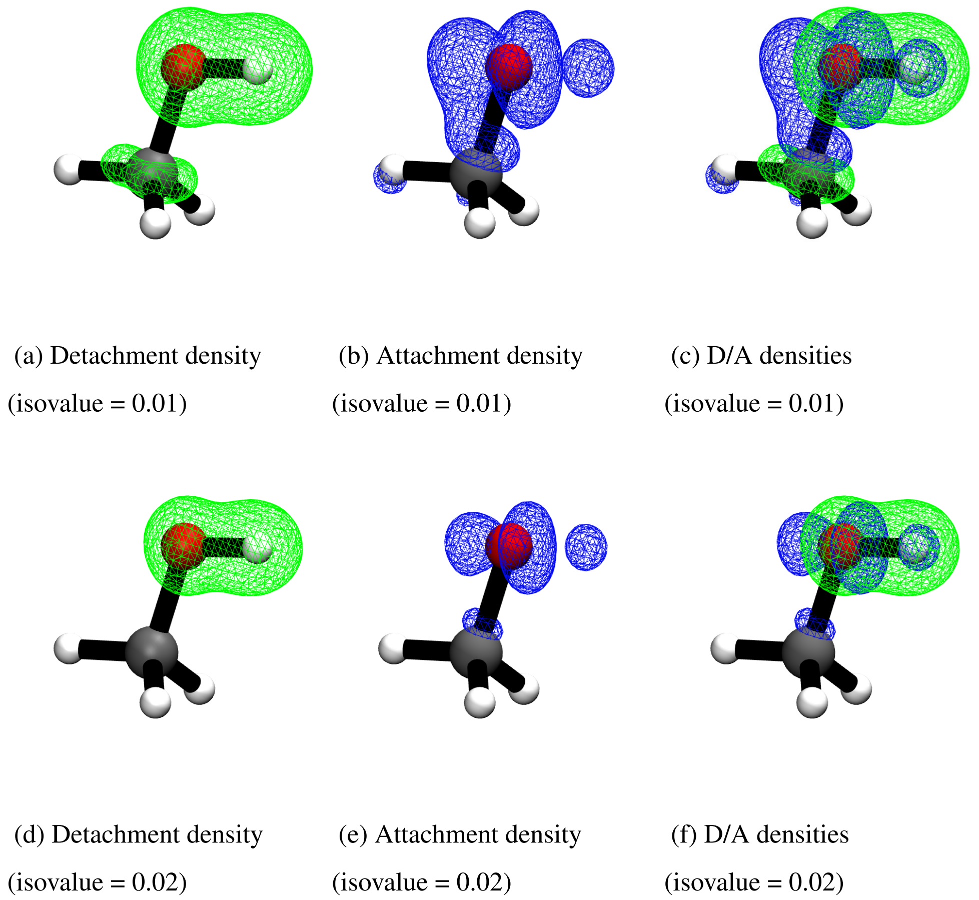
Upon the de-protonation, the number of electron does not change; only the electron density is affected. This conservation of electron number is the same for the excitation process. The number of electrons is not affected during the excitation process, just electron density changes. One method to determine this change in density is detachment attachment density analysis.

Detachment (D) density is a part of the ground-state density, which is removed from the molecule's ground state during the excitation process. Attachment (A) density is the rearranged detachment density in the excited state. The remaining electron density is common between the two states. The spacial overlap between detachment and attachment densities is ϕ_S. Our idea in the present work is to re-purpose the tool to calculate D/A densities, originally for the excitation process, for de-protonation reaction to make pKa prediction. Here, we define initial and final states as protonated and de-protonated states, respectively. The D/A densities and related descriptors are thus expected to capture the capability of different molecules to adapt to the loss of a proton, which we hypothesize to be related with their pKa.

In this study, geometry optimization and single point (SP) calculation were performed. During the SP calculation of the de-protonated form, a ghost atom with neither any nuclear charge nor electrons was used to replace the acidic Hydrogen atom, removed from the protonated form in the de-protonation reaction. The purpose of this ghost atoms is to have the exact same basis functions for the protonated and the de-protonated forms to compare their densities and calculate the difference density matrix. If the SP calculation were performed by removing the H atom directly, the number of basis functions for the protonated/de-protonated forms would not be equal, so comparing their densities would not be possible.

Then, density matrices were calculated for both the protonated/de-protonated forms. The detachment density matrix is the sum of all the natural orbitals of the difference density matrix, which corresponds to negative eigenvalues of the difference density matrix. Similarly, the attachment density matrix is the sum of all natural orbitals corresponding to positive eigenvalues of the difference matrix.[50] Finally, the overlap between these two densities, ϕ_S, was calculated based on the cube files. D/A density analysis reduces the complexity of difference density analysis because the difference density is a complicated function and it is not always possible to characterize every electronic transition with difference density. [50]

Detachment and attachment densities and their overlap density can be visualized with VDM or other visualisation program by using cube files. In Figure 3.1, a comparison of different isovalues is shown. The more spread density can be seen by lower isovalue (0.01). Generally, lower isovalues show electron density reorganization in detail. When isovalue is higher, some parts of the electron delocalization are lost, and very low isovalues would not be good either because electron density would cover all over the molecule, and we would not get the information. From chemical intuition, 0.01 is the chosen isovalue in this study but it should be noted that this choice is arbitrary.



(a) Detachment density (isovalue = 0.01)

(b) Attachment density (isovalue = 0.01)

(c) D/A densities (isovalue = 0.01)

(d) Detachment density (isovalue = 0.02)

(e) Attachment density (isovalue = 0.02)

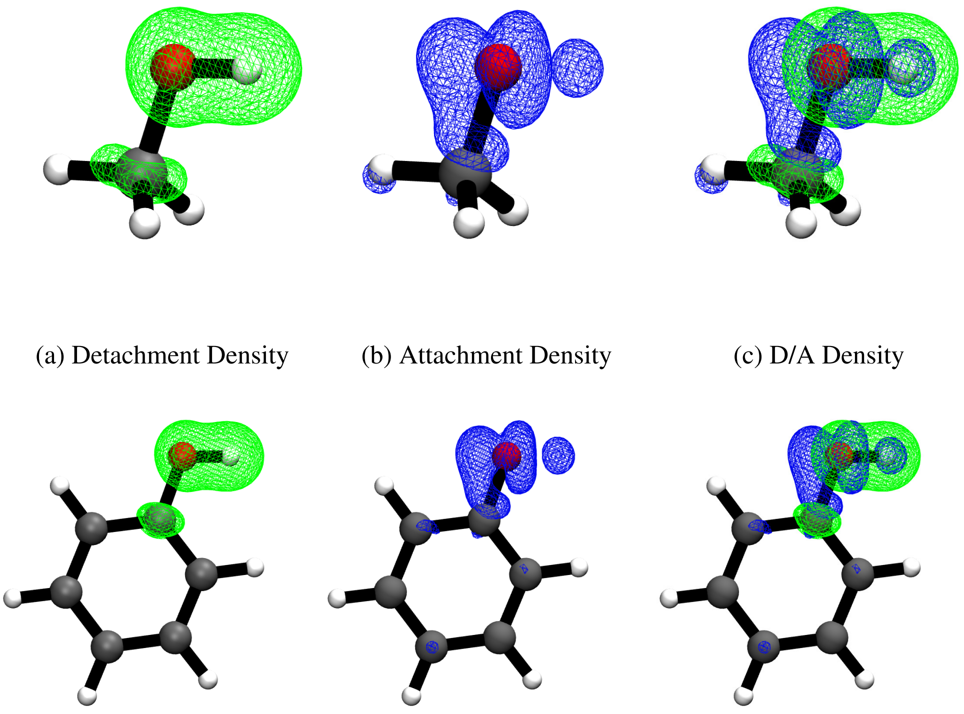
(f) D/A densities (isovalue = 0.02)

Figure 3.1: Representation of detachment, attachment, and detachment/attachment (D/A) densities of methanol by ωb97xd/6-311+G*/CPCM.

An analysis of the densities in Figure 3.1 shows that D density was placed on the sigma bond. This was somewhat expected since this is the main part of the disturbed electron density. It can be seen from Figure 3.1 that the green region (Detachment Density) is removed from the protonated form of the methanol during the de-protonation process and become the blue region (Attachment Density) in the deprotonated form. After removing H^+, electron density rearranges itself and spreads across the molecule.

Figure 3.2 compares the D/A densities on methanol and phenol. When H^+ is removed from phenol, electron density reorganizes over the molecule, and this reorganization is significantly different from methanol. We see that the ring is involved and that density in ortho and para positions is also involved. As we expect from general chemistry knowledge, ortho and para positions receive electron density in the de-protonated state. Our general understanding of the lower pKa of phenols over methanol is due to the delocalization of the negative charge in the phenolate in ortho and para positions. Our analysis goes in the same direction.

The densities are distinct from each other. The rearrangement in density is different for methanol and phenol, which lets us speculate that the D/A analysis holds information about the molecule's reactivity.



(a) Detachment Density (b) Attachment Density (c) D/A Density

(d) Detachment Density (e) Attachment Density (f) D/A Density

Figure 3.2: Representation of detachment (the green region) and attachment (the blue region) densities on methane and phenol by ωb97xd/6-311+G*/CPCM (isovalue = 0.01)

3.2 Aliphatic alcohols

First, the relation between ϕ_S and the number of carbon atoms in simple primary alcohols (methanol to nonanol) was investigated.

From general chemistry knowledge, it is known that acidity decreases with the increasing number of carbon atoms in simple primary alcohols (methanol to nonanol). Figure 3.3 comes from a study in which the authors investigated the acidity trend in alcohols by the increasing number of carbon atoms by using pentanol, hexanol, heptanol, octanol, and nonanol.[51] It demonstrates that acidity decreases and tends to converge to a given a value when increasing the number of carbons.

Figure 3.4 shows the variation of ϕ_S as a function of the number of carbons in a series of aliphatic primary alcohols. We see that ϕ_S increases with the number of carbons and tends to reach a plateau. The key takeaway is that there is a variation between ϕ_S and the C number, which goes in a direction similar to experimental observations.

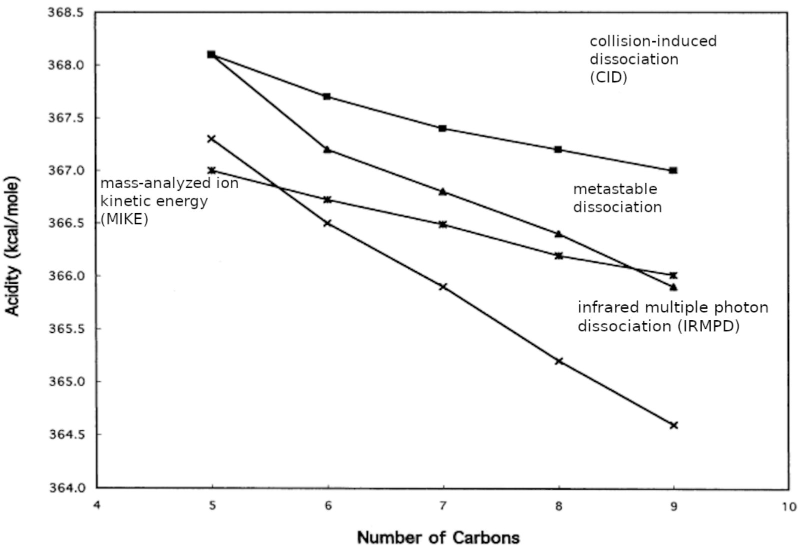


Figure 3.3: Acidity trend with respect to increase in C number (this graph is modified from the original paper by adding explicit names of the lines)[51]

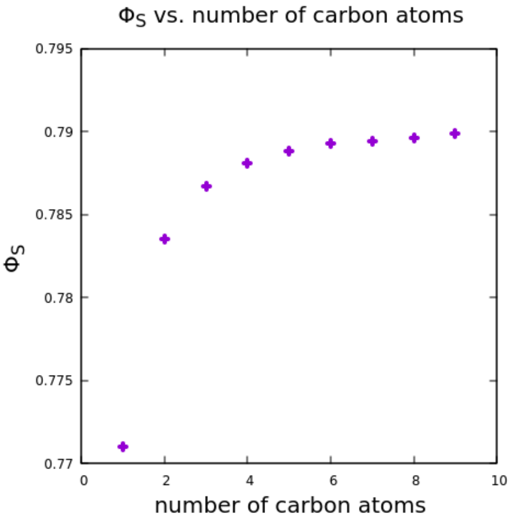


Figure 3.4: Variation of the D/A overlap ϕ_S as a function of the number of carbon in primary aliphatic alcohols.

3.3 Training set of alcohols and thiols

This study includes 25 alcohols and 26 thiols to investigate the correlation between pKa and ϕ_S and with other descriptors towards the end of this thesis. The molecules in this study are listed in Table 3.1 and Table 3.2 below. The Table 3.1 and Table 3.2 give the name of molecules, their experimental pKas, ϕ_S values, predicted pKas (by using pka vs. ϕ_S graphs which come from ωb97xd/6-311+G*/cpcm calculations) and the difference between predicted and true pKa values. The molecules with diverse pKa ranges for both alcohols and thiols were chosen to see whether D/A density overlap can be a general descriptor. To compare the predicted and measured pKa values, mean absolute error (MAE) and standard error (SE) was calculated. Among the 51 molecules selected in the present study, the error on the pKa prediction was fairly low (less than 2 pKa units). For six of them, however, the deviation was more significant. The molecules marked with a * in the Tables were considered as outliers. The origin of the failure of the prediction for those molecules may be an error on the experimental measurement or an intrinsic issue with our method and descriptor (see section 3.5). They were not included in the R^2 calculations or mean absolute error (MAE) and standard error (SE) calculations and they were also omitted from the method dependence analysis in section 3.4.

Table 3.1: Alcohols Training Set: IUPAC Nomenclature, Experimental pKa, ϕ_S, Predicted pKa, predicted pKa - experimental pKa (Δ(pKa)). Molecules with * were considered as outliers for this analysis and MAE and SE are calculated without outlier molecules.

$Alcohols$	**pKa**$_{exp}$**[38][52]**	ϕ_S	**pKa**$_{pred}$	$\Delta(pKa)$
2,6-dinitrophenol	3.71	0.8409	5.12	+1.41
2,4-dinitrophenol*	4.09	0.8373	5.89	+1.80
2,5-dinitrophenol	5.21	0.8392	5.48	+0.27
2,5-dichlorophenol	7.51	0.8273	8.01	+0.50
4-cyanophenol	7.96	0.8243	8.64	+0.68
3-hydroxyquinoline	8.06	0.8281	7.84	-0.22
3-methylsulfonylphenol	8.40	0.8240	8.71	+0.31
5-hydroxyquinolin	8.54	0.8271	8.05	-0.49
3-metoxyphenol	9.65	0.8226	9.01	-0.64
phenol	9.97	0.8187	9.83	-0.14
4-tertbutylphenol	10.23	0.8225	9.03	-1.20
2-4-6-trimethylphenol*	10.87	0.8250	8.50	-2.37
2-Trifluoromethyl-2-propanol	11.60	0.8124	11.17	-0.43
2,2,2-trichloroethanol	12.02	0.8089	11.91	-0.11
2,2,2-trifluoroethanol	12.43	0.7989	14.04	+1.61
2-propyn-1-ol	13.55	0.7944	14.99	+1.44
2-methoxyethanol	15.00	0.7957	14.72	-0.29
methanol	15.20	0.7821	17.60	+2.40
phenylmethanol*	15.44	0.8039	12.98	-2.47
ethanol	15.50	0.7922	15.46	-0.04
2-buten-1-ol	15.52	0.7959	14.67	-0.85
2-propanol	15.70	0.7986	14.10	-1.60
1-propanol	15.87	0.7957	14.72	-1.16
1-buthanol	15.92	0.7969	14.46	-1.46
tert-butanol*	16.00	0.8054	12.66	-3.34
	Mean Absolute Error (MAE) =			0.8212
	Standard Error (SE) =			0.7910

Table 3.2: Thiols Training Set: IUPAC Nomenclature, Experimental pKa, ϕ_S, Predicted pKa, predicted pKa - experimental pKa (Δ(pKa)). Molecules with * were considered as outliers for this analysis and MAE and SE are calculated without outlier molecules.

$Thiols$	**pKa**$_{exp}$**[38][52]**	ϕ_S	**pKa**$_{pred}$	$\Delta(pKa)$
3-nitrobenzenethiol	5.24	0.7649	5.67	+0.43
5-mercaptouracil	5.30	0.7569	6.81	+1.51
1-4-mercaptophenyl ethane-1-one	5.33	0.7674	5.31	-0.02
3-chlorobenzenthiol	5.78	0.7623	6.04	+0.26
4-bromobenzene-thiol	6.02	0.7620	6.08	+0.06
4-chlorobenzenthiol	6.14	0.7551	7.07	+0.93
3-methoxy-benzene-thiol	6.39	0.7621	6.07	-0.32
benzenethiol	6.61	0.7583	6.61	+0.00
2-methylbenzenethiol	6.64	0.7623	6.04	-0.60
3-methylbenzenethiol	6.66	0.7602	6.34	-0.32
4-methoxybenzenethiol	6.78	0.7555	7.01	+0.23
4-methylbenzenethiol	6.82	0.7599	6.38	-0.44
2-2-2-trifluoro ethane 1-thiol	7.30	0.7396	9.29	+1.99
prop-1-ene-2-thiol	7.86	0.7420	8.94	+1.08
2-ethoxyethanethiol	9.38	0.7380	9.52	+0.14
phenylmethanethiol	9.43	0.7456	8.43	-1.00
3-mercaptopropane-1-2-diol	9.51	0.7452	8.49	-1.02
2-mercaptoethanol	9.72	0.7365	9.73	+0.01
prop-2-ene-1-thiol	9.96	0.7360	9.80	-0.16
methanethiol	10.33	0.7248	11.41	+1.08
ethanethiol	10.61	0.7344	10.03	-0.58
butane-1-thiol	10.67	0.7401	9.22	-1.45
2-propanethiol	10.86	0.7411	9.07	-1.79
2-methylpropane-2-thiol*	11.05	0.7489	7.96	-3.09
2-methyl-2-butanethiol*	11.22	0.7517	7.56	-3.67
Mean Absolute Error (MAE) =				0.7749
Standard Error (SE) =				0.3494

Four representative molecules of the alcohols set are depicted on a scale in Figure 3.5. The Figure shows electron density reorganization in 2,6-dinitrophenol, 4-cyanophenol, phenol, and propanol with their pKa and corresponding ϕ_S values. pKa decreases from propanol to 2,6-dinitrophenol (propanol < phenol < 4-cyanophenol < 2,6-dinitrophenol). The trend (for both alcohol and thiols) that we observe from calculations is that the ϕ_S increases with decreasing pKa, and we get different ϕ_S values for different molecules with specific pKa values. The trend for alcohols and thiols is not perfect, but the general trend for alcohols and thiols is clear.

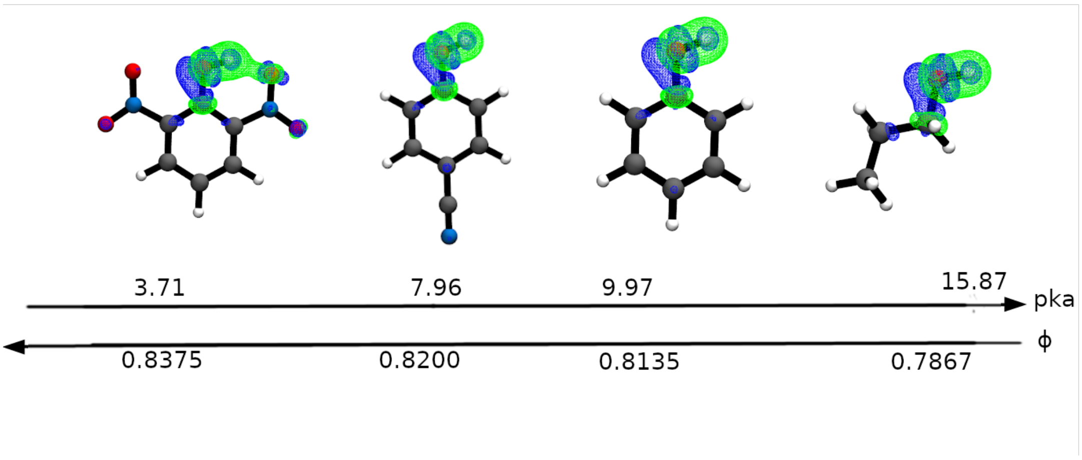
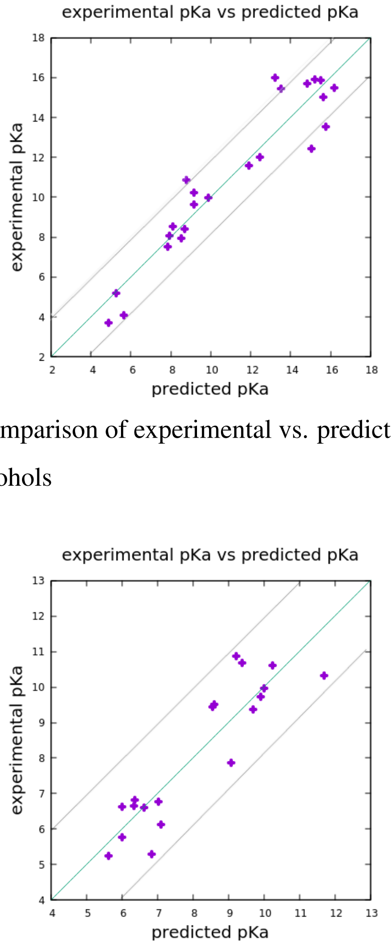


Figure 3.5: Representative molecules from the alcohols set with their pKa and corresponding ϕ_S values.

By using pKa vs. ϕ_S results from Tables 3.1 and 3.2, pKa predictions were done and shown in Figure 3.6. Figure 3.6 demonstrates experimental vs. predicted pKa values for alcohols and thiols. ωb97xd/6-311+G*/CPCM was used for both of them. In both graphs, predicted values are in the range of ± 2 with respect to measured values. For alcohols and thiols individually, the ϕ_S descriptors gives a fair prediction of pKa values.



(a) Comparison of experimental vs. predicted pKa for alcohols

(b) Comparison of experimental vs. predicted pKa for thiols

Figure 3.6: Comparison of experimental vs. predicted pKa for alcohols and thiols by ωb97xd/6-311+G**/CPCM

Figure 3.7 demonstrates the correlation of pKa of alcohols and thiols with ϕ_S by using ωb97xd/6-31++G** and water as a solvent. As shown, both alcohols and thiols correlate with pKa very well. Correlation constants, without outlier molecules, are 0.93 and 0.87 respectively (with outlier moleculesR^2 = 0.88 for alcohols and R^2 = 0.58 for thiols). However, it is not possible to predict the pKa of alcohols and thiols with a unique ϕ_S descriptor as the two families of molecules fall in different regions of the plot. This means ϕ_S alone is not enough for a general pKa descriptor. A second (or more) descriptor is needed to capture the correlation.

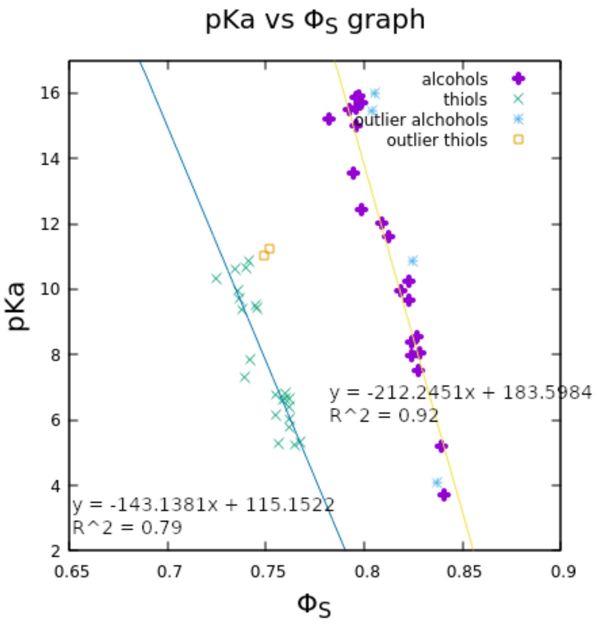
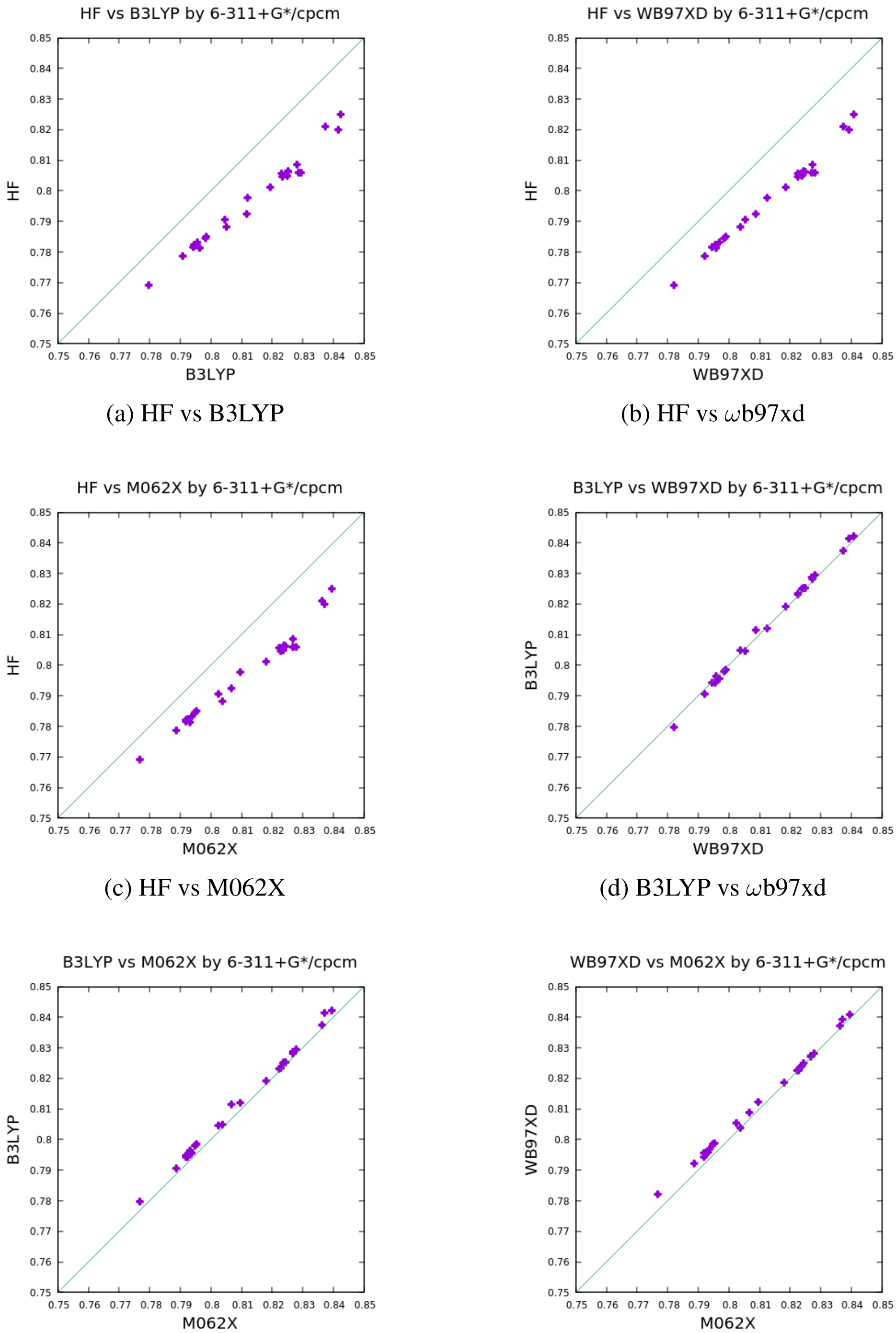


Figure 3.7: Correlation between pKa and ϕ_S for alcohols and thiols by ωb97xd/6-311+G*/CPCM

3.4 Methodological aspects affecting the value of ϕ_S

Before exploring other descriptors from the D/A density analysis, we analyzed the method dependency of our approach. The calculations were performed with several methods and basis sets to investigate whether ϕ_S has any dependency. The findings of the method dependency investigation can be shown in Figure 3.8.

HF, B3LYP, ωb97xd, and M062X were used to determine whether ϕ_S changes from method to method. A fixed basis set was used (6-31++G**) with the CPCM solvation method. The findings from Figure 3.8d, Figure3.8e, and Figure3.8f of Figure 3.8 suggest that the descriptor is not density functional dependent. However, from Figure3.8a, Figure3.8b and Figure3.8c, the results from HF to DFT matter. We observe that there is a linear relation between HF and DFT functionals yet, the ϕ_S which comes from HF is always lower than the density functionals.



(a) HF vs B3LYP

(b) HF vs ωb97xd

(c) HF vs M062X

(d) B3LYP vs ωb97xd

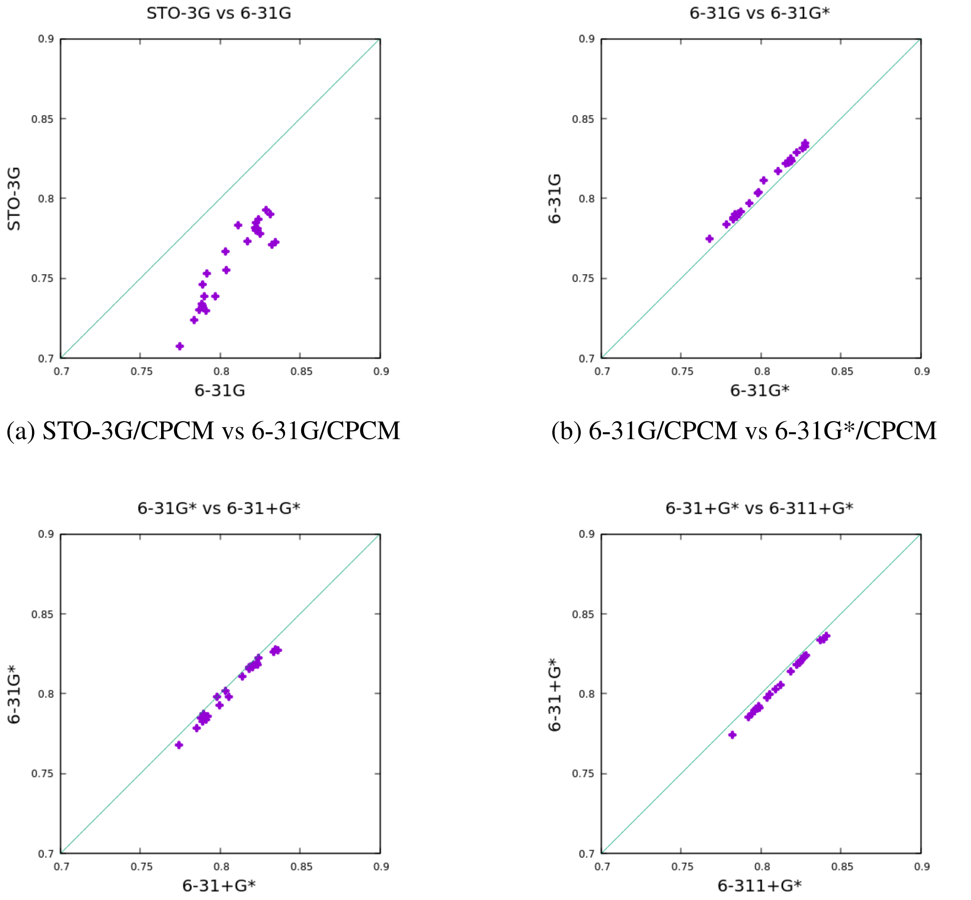
(e) B3LYP vs M062X with

(f) ωb97xd vs M062X

Figure 3.8: Comparisons of different methods for ϕ_S of alcohols

In this study, a few basis sets (STO-3G, 6-31G, 6-31G*, 6-31+G*, 6-311+G*) were used to investigate basis set effects. The results are presented in Figure 3.9.

From Figure 3.9 it can be seen that the minimal basis, STO-3G, differs significantly from 6-31G. In general, STO-3G gives a lower overlap between D/A densities than split valence basis sets, and the relation is not linear. Figure 3.9b, Figure 3.9c, and Figure 3.9d demonstrate that Pople basis sets are very consistent with each other, and double zeta or triple zeta basis sets do not have a significant difference in results. Furthermore, adding polarization and diffuse functions (i.e., * and +, respectively) in basis sets does not affect the results, so there is no need to use more extensive basis sets. It seems 6-31G might be used safely.



(a) STO-3G/CPCM vs 6-31G/CPCM

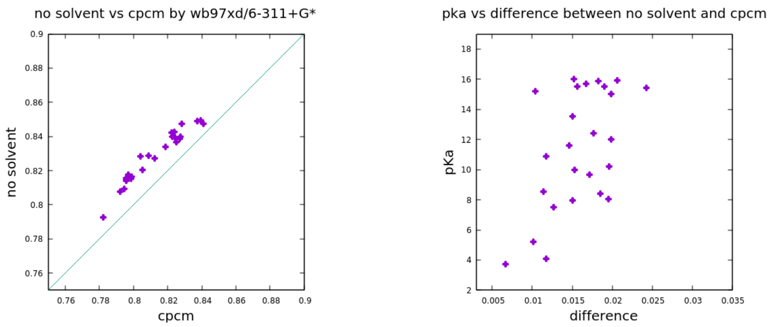
(b) 6-31G/CPCM vs 6-31G*/CPCM

(c) 6-31G*/CPCM vs 6-31+G*/CPCM

(d) 6-31+G*/CPCM vs 6-311+G*/CPCM

Figure 3.9: Comparison of different basis sets for ϕ_S of alcohols with ωb97xd method.

Figure 3.10a demonstrates the comparison of ϕ_S, which comes from the gas phase and implicit solvation model (CPCM) by using the same method and basis set, which is ωb97xd/6-311+G* for alcohols. ϕ_S is higher in the gas phase than CPCM, and they have a nearly linear relationship. Also, Figure 3.10b shows pKa versus the difference between vacuum and solvent ϕ_S results. There is no clear correlation between pKa versus the difference in ϕ_S with and without solvent, but it is seen that the difference tends to increase with increasing pKa.



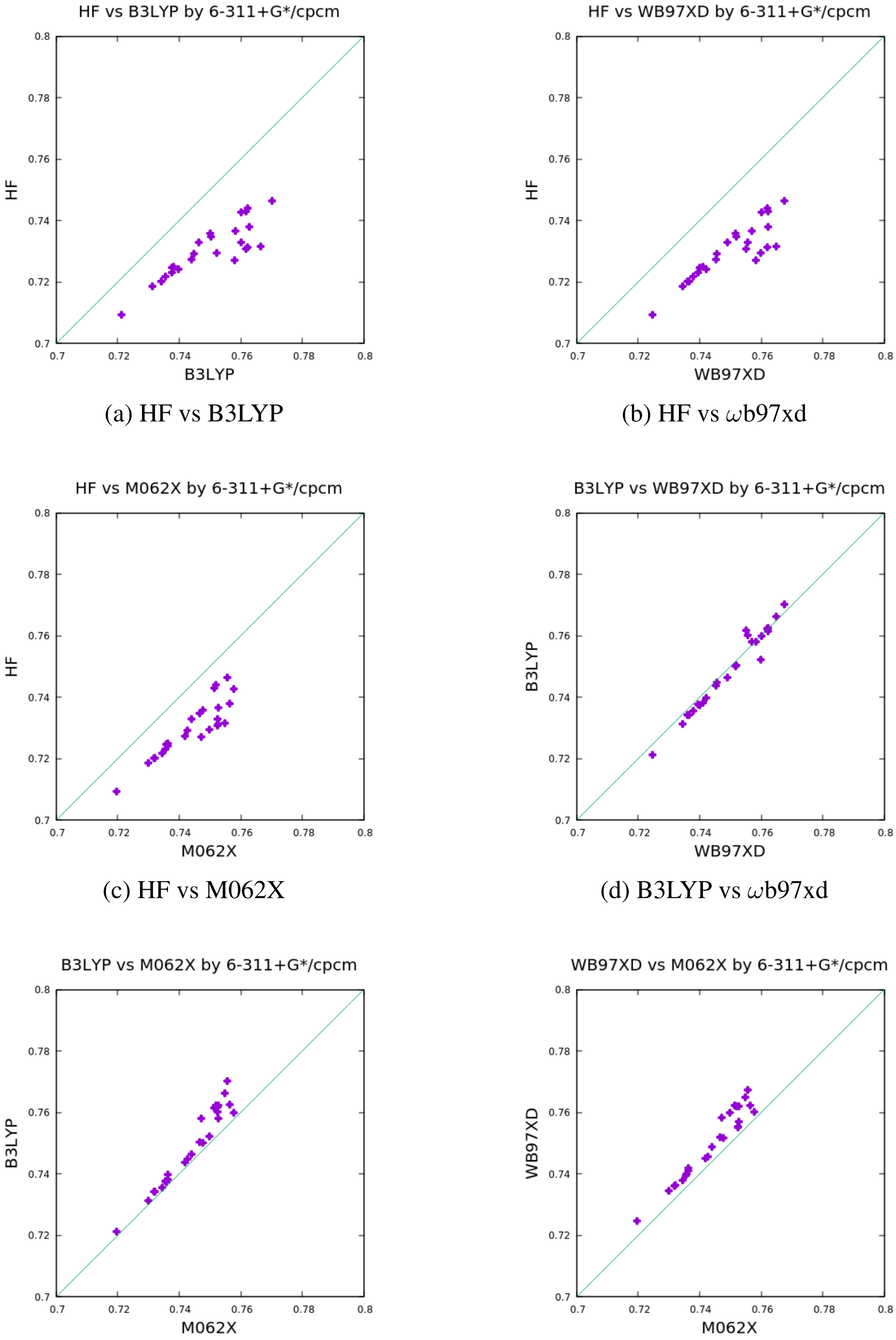
(a) Comparison of gas phase and solvent (CPCM) results

(b) pKa versus the difference between gas phase and solvent (CPCM) results

Figure 3.10: Comparison of gas phase and solvation model (CPCM) ϕ_S results and correlation between their difference and pKa by ωb97xd/6-311+G* for alcohols.

The results in Figure 3.11 for thiols demonstrate almost the same trend as alcohols. The same methods and basis sets are also used to observe correlation for thiols. As in alcohols, ϕ_S is functional independent, but HF and DFT results are different.

The investigations of the basis set effects for thiols revealed that STO-3G and 6-31G basis sets provide similar results, but there are few exceptions. From Figure 3.12b, one can see that adding polarization functions does not affect the results. However, Figure 3.12c shows that adding diffuse functions changes the results significantly. Finally, from 3.12d, one can observe that double zeta or triple zeta basis sets do not differ, and no need to use triple zeta basis functions. The findings suggest that selecting a basis set with diffuse functions for thiols is crucial.



(a) HF vs B3LYP

(b) HF vs ωb97xd

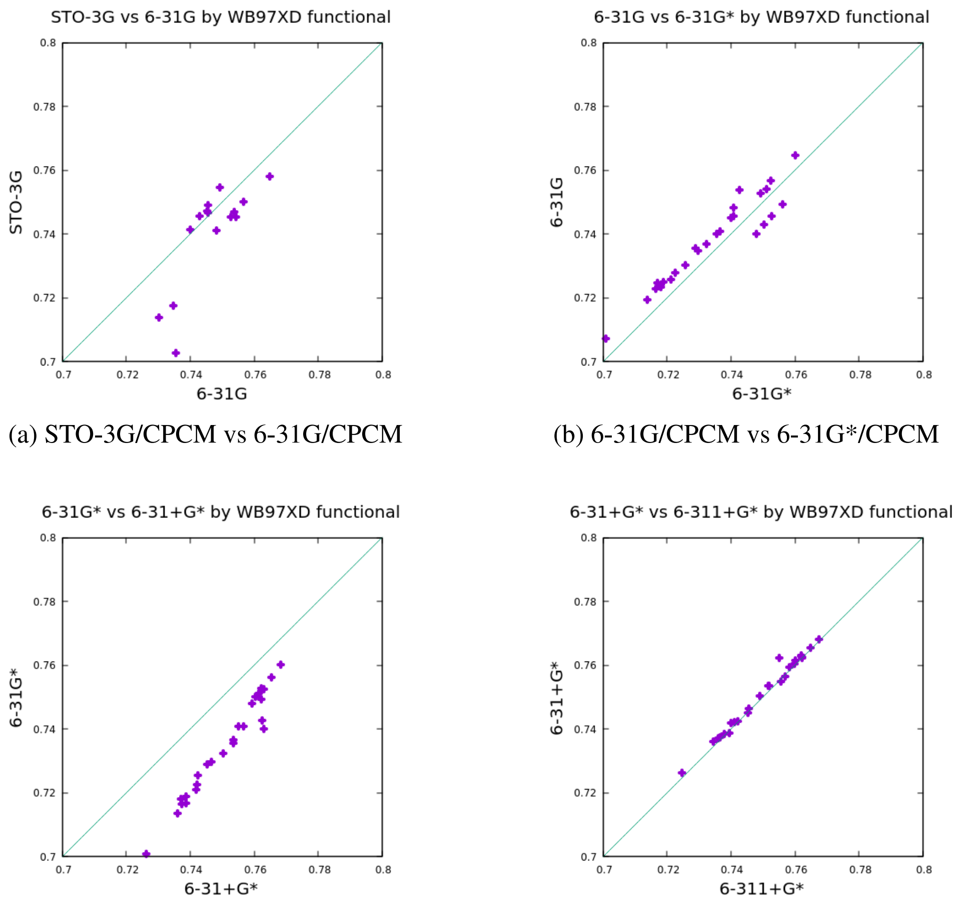
(c) HF vs M062X

(d) B3LYP vs ωb97xd

(e) B3LYP vs M062X

(f) ωb97xd vs M062X

Figure 3.11: Comparisons of different methods for ϕ_S of thiols



(a) STO-3G/CPCM vs 6-31G/CPCM

(b) 6-31G/CPCM vs 6-31G*/CPCM

(c) 6-31G*/CPCM vs 6-31+G*/CPCM

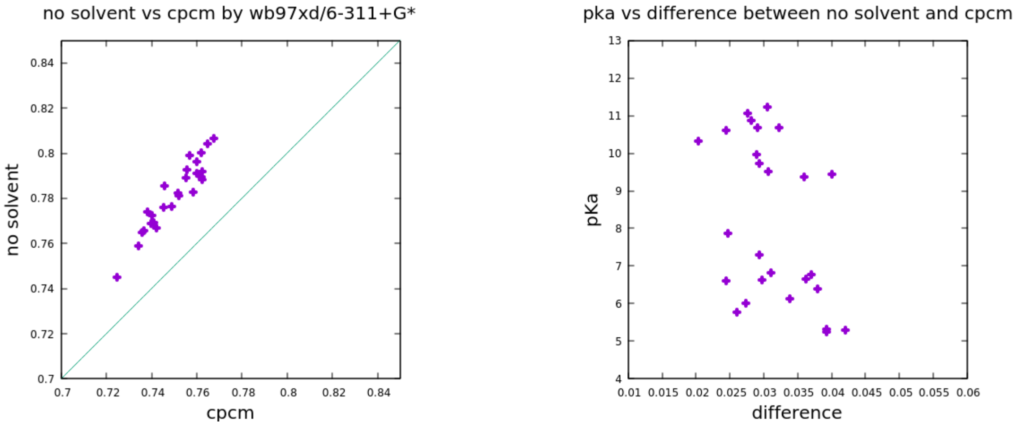
(d) 6-31+G*/CPCM vs 6-311+G*/CPCM

Figure 3.12: Comparison of different basis sets for ϕ_S of thiols with ωb97xd method.

The comparison of the gas phase and CPCM ϕ_S using the same method and basis set, ωb97xd/6-311+G* for thiols, is shown in Figure 3.13.

As in the alcohols, ϕ_S is higher in the gas phase than with CPCM, and there is a linear relationship between them. Also, Figure 3.13 demonstrates pKa versus the difference between vacuum and solvent ϕ_S graph.

There is no clear correlation between pKa and the difference. However, for thiols, the difference does not increase with increasing pKa as in alcohols. Here, there is an almost constant difference range for all thiols in the set, whether high or low pKa.



(a) Comparison of gas phase and solvent (CPCM) results

(b) pKa versus the difference between gas phase and solvent (CPCM) results

Figure 3.13: Comparison gas phase and solvation model (CPCM) results and correlation between their difference and pKa by ωb97xd/6-311+G* for thiols.

3.5 Investigations of other descriptors

From the findings of ϕ_S, one can conclude that ϕ_S indeed captures the variation of pKa within one family of molecules (i.e., either alcohols or thiols), but is not enough to gather all chemical functions within the same equation. We thus investigated other descriptors extracted from the D/A analysis. Mesra provides other terms together with D/A density overlap. Two of them are ϑ, and qCT. Both ϑ, and qCT are related with charge involved in the transition from the initial to final state, so they were investigated as possible descriptors. In the present study, the other investigated descriptors (different from ϕ_S) together with ϑ and qCT are (ϑ - qCT), (ϑ - qCT)*ϕ_S and ϕ_S*(ϑ - qCT)/ϑ. The formulas for ϕ_S, ϑ, and qCT were given in the Mathematical details section.

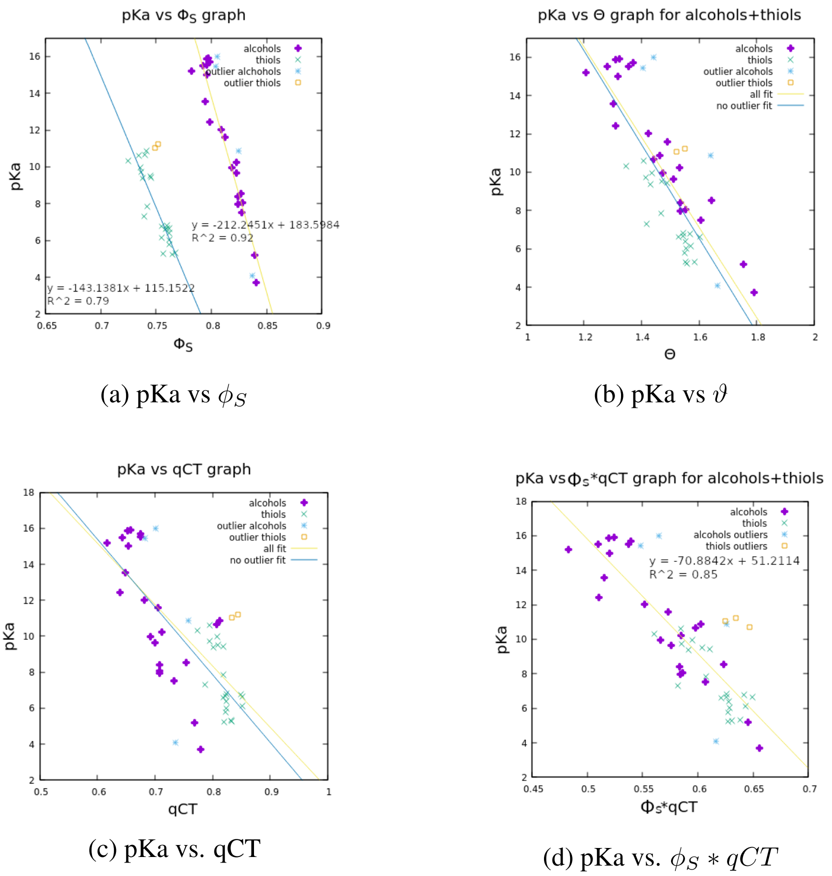
The findings are shown in Table 3.3 which contains both alcohols and thiols in the increasing pKa order. In Table 3.3, the results that belong to thiols are represented as colorful rows. Table 3.3 contains experimental pKa, ϑ, qCT, ϕ_S, ϕ_S*qCT, predicted pKa (from pka vs. ϕ_S*qCT results) and Δ(pKa). In the tables, the molecules with a star (*) were considered as outliers and have not been included in correlation constant or error calculations.

Table 3.3: Alcohols+Thiols Training Set: IUPAC Nomenclature, Experimental pKa (pKa_{exp}), ϑ, qCT, ϕ_S, ϕ_S*qCT, Predicted pKa (pKa_{pred}), predicted pKa - experimental pKa ($\Delta(pKa))$. Molecules with * were considered as outliers for this analysis and MAE and SE are calculated without outlier molecules.

$Molecules$	pKa_{exp}**[38][52]**	ϑ	qCT	ϕ_S	$\phi_S * qCT$	pKa_{pred}	$\Delta(pKa))$
2,6-dinitrophenol	3.71	1.7908	0.7795	0.8409	0.6554	4.75	+1.04
2,4-dinitrophenol*	4.09	1.6659	0.7362	0.8373	0.6164	7.52	+3.43
2,5-dinitrophenol	5.21	1.7553	0.7693	0.8392	0.6456	5.45	+0.24
3-nitrobenzenethiol	5.24	1.5577	0.8205	0.7649	0.6276	6.73	+1.49
5-mercaptouracil	5.30	1.5528	0.8334	0.7569	0.6308	6.50	+1.20
1-4-mercaptophenyl ethane-1-one	5.33	1.5859	0.8321	0.7674	0.6386	5.95	+0.62
3-chlorobenzenthiol	5.78	1.5517	0.8233	0.7623	0.6276	6.72	+0.94
4-bromobenzene-thiol	6.02	1.5550	0.8249	0.7620	0.6286	6.66	+0.64
4-chlorobenzenthiol	6.14	1.5726	0.8519	0.7551	0.6432	5.62	-0.52
3-methoxy-benzene-thiol	6.39	1.5543	0.8245	0.7621	0.6283	6.67	+0.28
benzenethiol	6.61	1.5299	0.8192	0.7583	0.6212	7.19	+0.57
2-methylbenzenethiol	6.64	1.6032	0.8523	0.7623	0.6497	5.16	-1.48
3-methylbenzenethiol	6.66	1.5461	0.8235	0.7602	0.6260	6.84	+0.18
4-methoxybenzenethiol	6.78	1.5707	0.8502	0.7555	0.6423	5.68	-1.10
4-methylbenzenethiol	6.82	1.5444	0.8240	0.7599	0.6262	6.83	+0.00
2-2-2-trifluoro ethane 1-thiol	7.30	1.4187	0.7874	0.7396	0.5824	9.93	+2.63
2,5-dichlorophenol	7.51	1.6066	0.7334	0.8273	0.6067	8.21	+0.70
prop-1-ene-2-thiol	7.86	1.4692	0.8185	0.7420	0.6073	8.16	+0.30
4-cyanophenol	7.96	1.5344	0.7084	0.8243	0.5839	9.82	+1.86
3-hydroxyquinoline	8.06	1.5529	0.7082	0.8281	0.5864	9.64	+1.58
3-methylsulfonylphenol	8.40	1.5352	0.7077	0.8240	0.5831	9.88	+1.48
5-hydroxyquinolin	8.54	1.6424	0.7537	0.8271	0.6234	7.02	-1.52
2-ethoxyethanethiol	9.38	1.4315	0.8017	0.7380	0.5917	9.27	-0.11
phenylmethanethiol	9.43	1.4889	0.8192	0.7456	0.6108	7.92	-1.51
3-mercaptopropane-1-2-diol	9.51	1.4720	0.8105	0.7452	0.6040	8.40	-1.11
3-metoxyphenol	9.65	1.5123	0.7001	0.8226	0.5759	10.40	+0.74
2-mercaptoethanol	9.72	1.4162	0.7954	0.7365	0.5858	9.69	-0.03
prop-2-ene-1-thiol	9.96	1.4370	0.8089	0.7360	0.5954	9.01	-0.95
phenol	9.97	1.4754	0.6915	0.8187	0.5661	11.08	+1.11
4-tertbutylphenol	10.23	1.5331	0.7114	0.8225	0.5851	9.74	-0.50
methanethiol	10.33	1.3474	0.7736	0.7248	0.5607	11.47	+1.14
ethanethiol	10.61	1.4074	0.7959	0.7344	0.5845	9.78	-0.83
butane-1-thiol	10.67	1.4436	0.8077	0.7401	0.5977	8.84	-1.83
2-propanethiol	10.86	1.4635	0.8129	0.7411	0.6024	8.51	-2.35
2-4-6-trimethylphenol*	10.87	1.6413	0.7586	0.8250	0.6258	6.85	-4.02
2-methylpropane-2-thiol*	11.05	1.5230	0.8339	0.7489	0.6245	6.95	-4.10
2-methyl-2-butanethiol*	11.22	1.5500	0.8437	0.7517	0.6342	6.26	-4.96
2-Trifluoromethyl-2-propanol	11.60	1.4907	0.7054	0.8124	0.5730	10.59	-1.01
2,2,2-trichloroethanol	12.02	1.4229	0.6821	0.8089	0.5517	12.10	+0.08
2,2,2-trifluoroethanol	12.43	1.3107	0.6389	0.7989	0.5104	15.03	+2.60
2-propyn-1-ol	13.55	1.3027	0.6488	0.7944	0.5154	14.68	+1.13
2-methoxyethanol	15.00	1.3186	0.6540	0.7957	0.5204	14.32	-0.68
methanol	15.20	1.2077	0.6174	0.7821	0.4828	16.88	+1.79
phenylmethanol*	15.44	1.4046	0.6824	0.8039	0.5486	12.33	-3.11
ethanol	15.50	1.2821	0.6434	0.7922	0.5097	15.08	-0.42
2-buten-1-ol	15.52	1.3565	0.6749	0.7959	0.5372	13.14	-2.38
2-propanol	15.70	1.3702	0.6751	0.7986	0.5391	13.00	-2.70
1-propanol	15.87	1.3108	0.6529	0.7957	0.5195	14.39	-1.48
1-buthanol	15.92	1.3239	0.6574	0.7969	0.5239	14.08	-1.84
tert-butanol*	16.00	1.4433	0.7010	0.8054	0.5646	11.19	-4.81
						Mean Absolute Error (MAE) =	1.11
						Standard Error (SE) =	0.48

In Figure 3.14, the results of investigations of other descriptors are shown. Figure 3.14a reminds the results obtained in the previous section for pKa vs. ϕ_S graph. ϕ_S is the overlap between the D/A densities. It is a normalized descriptor that indicates the locality of the charge redistribution, but not the amount of charge involved. Our first descriptor, ϕ_S, gives a good correlation with pKa for alcohols and thiols separately, but there is no general correlation between them. Figure 3.14b shows pKa vs. ϑ graph. ϑ is the total charge involved in the transition (de-protonation for our purposes). The values of ϑ for all molecules in the current study can be seen in Table 3.3. As shown, pKa increases with decreasing ϑ. ϑ for alcohols is found to be slightly greater than it is for thiols. pKa gives a rather poor correlation with ϑ by $R^2 = 0.75$ without outlier molecules (in the presence of outlier molecules R^2 equals to 0.63). Figure 3.14c shows pKa vs. qCT graph. The descriptor qCT is the charge that is effectively moved from the initial to the final state during de-protonation. The moved charge for thiols is greater than for alcohols. The total charge, ϑ, gives a correlation, but there is no nice correlation for the amount of moved charge as can be seen from correlation constant. ϑ, and qCT give information about the amount of charge involved.

Figure 3.14d shows the pKa vs. ϕ_S*qCT graph. ϕ_S gives information about the locality of the charge reorganization. Thus, ϕ_S*qCT is (locality of the reorganization)*(amount of charged). As shown, pKa increases with decreasing ϕ_S*qCT. It is noteworthy that neither ϕ_S nor qCT do not give a good correlation with pKa of alcohols and thiols combined, but the descriptor ϕ_S*qCT correlates well with pKa of both chemical functions with $R^2 = 0.85$.



(a) pKa vs ϕ_S

(b) pKa vs ϑ

(c) pKa vs. qCT

(d) pKa vs. $\phi_S * qCT$

Figure 3.14: Investigations of correlation between pKa and other descriptors

This study calculated all R^2 values, predicted pKas, mean absolute errors, and standard errors by removing these outlier molecules. Figure 3.14 also shows the results from Table 3.3 with and without outlier molecules. The outlier molecules are the followings: 2,4-dinitrophenol, 2-4-6-trimethyl phenol, phenyl methanol, tert-butanol, 2-methylpropane-2-thiol, and 2-methyl-2-butanethiol. They give the highest difference between measured pKa and predicted pKa values (which come from pKa vs. ϕ_S*qCT graph).

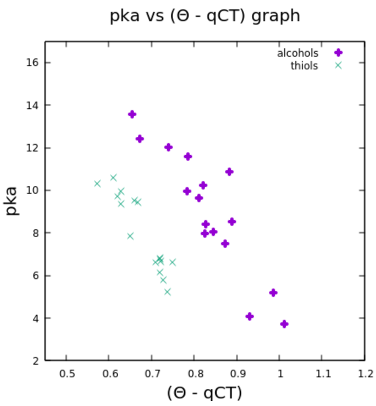
In the de-protonated form, after removing H^+, the molecule gets negatively charged. In the presence of this extra -1 charge, substituent groups or resonance effects help the molecule to stabilize itself. For de-protonation reactions, the presence of electron-withdrawing groups can help stabilization, but the electron-donating groups may cause a more unstable form.

For example, R groups in organic chemistry have electron-donating effects. In the presence of several R substitutions in a molecule, the de-protonated form might not be very stable and gives a higher pKa value. One of the outlier molecules is 2,4,6-Trimethylphenol. Due to having three methyl groups, the de-protonated form of the molecule may not be very stable. It is also valid for phenyl methanol and tert-butanol because phenyl is an electron donating group too. Instead of removing the effect of an extra negative charge (coming from de-protonation), the substituted electron-donating group provides more electrons and increases the negative charge on the molecule. If there is no other way to stabilize the molecule itself, such as having electron-withdrawing groups or resonance effect, these molecules are outlier in our training sets.

However, the Nitro group, (NO_2^-), has the electron-withdrawing character. In phenol, the presence of two nitro groups provides stabilization, but the position of groups might also be important. Our training set includes 2,6-dinitrophenol, 2,5-dinitro phenol, and 2,4-dinitrophenol. 2,4-dinitro phenol is outlier due to the position of nitro groups (on the ortho and para positions). Among them, the most distant nitro group to the OH group is 2,4-dinitro phenol. Since the electron-withdrawing group are farther from the negative charge in 2,4-dinitro phenol, it might be an outlier due to this reason. The long distance effects might not be captured well by the functional and/or the basis set. Also, there is a H bond possibility between H of OH and O or N of nitro groups. It should be thought about more carefully.

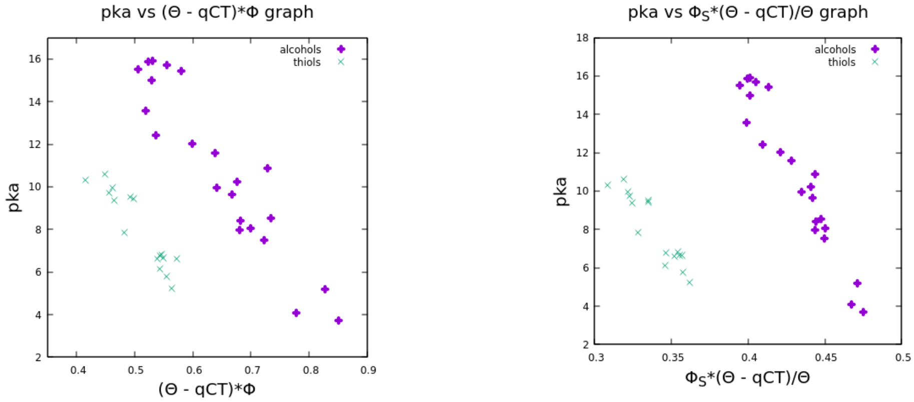
For two outliers thiols, the reason looks similar. In 2-methylpropane-2-thiol and 2-methyl-2-butanethiol, there are not enough stabilization ways for the negative charge that comes from de-protonation. In the presence of electron-donating R groups, the molecules might not be stable. We could conduct a deeper study to see how these molecules that are outliers in this model would behave with other functionals/methods/basis sets. We did not assign methanol or methanethiol as an outlier, but they are small and do not have any other functional group to help stabilize the negative charge, they are not very stable in de-protonated form. From Figure 3.14a, one can see that the presence or absence of outlier molecules does not change the pKa vs. ϕ_S trend. Figure 3.14a also demonstrates that for alcohols D/A density overlap are significantly greater than thiols. Figure 3.14b shows pKa vs. ϑ graphs both in the presence and absence of outlier molecules. The Figure has two fitting results, one is in the presence of outlier molecules as yellow line shows and the another is in the absence of outlier molecules as blue line shows. The correlation constant with outlier is $R^2 = 0.63$, without outlier $R^2 = 0.75$. As the graphs shows, there is a correlation. However, the correlation constant ($R^2 = 0.75$) is still not enough for precise pKa predictions. Figure 3.14c also has two correlation results, one is for with outliers (yellow line) another is without outlier molecules (blue line). qCTs of thiols are higher than alcohols and R^2 value improves from 0.48 to 0.62 by removing outlier molecules. However, the correlation between pKa and qCT is still low to make reliable pKa predictions.

Figure 3.15a is pKa vs. (ϑ - qCT) graph. The descriptor (ϑ - qCT) is the amount of non-moved charge (stayed at its position) during the de-protonation process. As shown, pKa increases with decreasing (ϑ - qCT) too, and, it makes two distinct classes. There is no correlation, but in this case, alcohols' (ϑ - qCT) are higher than thiols. Figure 3.15b is pKa vs. (ϑ - qCT)*ϕ_S graph. As shown, pKa increases with decreasing (ϑ - qCT)*ϕ_S. The descriptor (ϑ - qCT)*ϕ_S is (non-moving charge)*(overlap between D/A densities). Alcohols and thiols give two distinct classes for this descriptor as well. So, there is no correlation, and again the values for alcohols are higher than thiols. Finally, Figure 3.15c is pKa vs. ϕ_S(ϑ - qCT)/ϑ graph. The descriptor ϕ_S(ϑ - qCT)/ϑ is ϕ_S*(normalized non-moved charge). In the figure, pKa increases with decreasing ϕ_S(ϑ - qCT)/ϑ. The descriptor ϕ_S(ϑ - qCT)/ϑ, which gives lower values for thiols, does not give any correlation either.



(a) pKa vs (ϑ - qCT)

Figure 3.15: Investigations of correlation between pKa and other descriptors



(b) pKa vs (ϑ - qCT)*ϕ_S

(c) pKa vs ϕ_S*(ϑ - qCT)/ϑ

Figure 3.15: Investigations of correlation between pKa and other descriptors

CHAPTER 4

DISCUSSIONS

Our study showed that the analysis of detachment and attachment densities in the context of a de-protonation reaction contain information about the reactivity of the acidic function. Our investigation focused on correlating descriptors extracted from the D/A analysis with pKa showed that ϕ_S gives a correlation with correlation coefficients of 0.93 and 0.87 for alcohols and thiols, respectively.

Moreover, two different theories, namely HF and three different DFT functional, were compared for both alcohols and thiols. The findings for alcohols show that DFT and HF give distinct results, but B3LYP, ωb97xd, and M062X give almost the same values. Also, all these three DFT functional always give higher ϕ_S than HF. It might be because HF contains less interaction and provides lower ϕ_S.

The comparison of HF and DFT in thiols is very similar too. ϕ_S coming from HF is always lower than DFTs, and three DFT functionals are consistent with each other as in the alcohols. However, the ϕ_S trend among HF and DFTs for alcohols was almost parallel; there was a regular trend between ϕ_S of HF and DFTs, though ϕ_S that come from HF are lower. We do not see that much regular trend between these two theories for thiols. The results are more dispersed. From these, one can say that the descriptor does not look functional dependent for both alcohols and thiols, though HF and DFT differ.

Investigations of basis set dependency of the ϕ_S gave different results for alcohols and thiols. STO-3G, 6-31G, 6-31G*, 6-31+G*, and 6-311+G* basis sets were used to understand whether basis sets affect the D/A overlap. In alcohols, all Pople basis sets are consistent with each other by giving almost the same values, but STO-3G calculates lower D/A overlap than Pople's. For alcohols, one can conclude that 6-31G is enough to use; there is no need for polarization and diffuse functions or split valance triple or more zeta basis sets.

The trend is not so clear for thiols because there are some exceptions. However, one can see that 6-31+G* (double zeta) gives almost the same value as 6-311+G* (triple zeta), so there is no need to use triple zeta. There are, however, some exceptions by looking at 6-31G and 6-31G*, we can say the addition of polarization functions might not affect the results much. However, from 6-31G* to 6-31+G*, diffuse functions change the results significantly because 6-31G* calculates lower D/A overlap than 6-31+G*.

In the current study, Pople's basis sets give greater ϕ_S than minimal basis, STO-3G for both alcohols and thiols (though there are some exceptions for thiols). Generally, STO-3G gives a rough representation of a system because it includes only necessary orbitals to contain the electrons of the atoms, and each orbital is represented by a single basis function. For example, for carbon atom electron configuration is $1s^2 2s^2 2p^2$, and by minimal basis set, we only need five basis functions; two are for 1s and 2s orbitals, and other three are for 2p orbitals. Due to this, in general, minimal basis sets are not flexible enough to represent a system. Larger basis sets are more flexible and give better accuracy though their usage is computationally more expensive.

Larger basis sets contain more basis functions per orbital than minimal ones. In chemistry, most of the time, valance orbitals play a more critical role in reactions than core orbitals, and providing more flexibility to valance orbitals might help to define a system much better. Having only one basis function per core orbitals but for valance orbitals, two or more basis functions is the idea behind the "split valance" concept. For instance, if there are two basis functions per valance orbitals, the basis set is called "double zeta." Similarly, when there are three basis functions per valance orbitals, the basis set is called "triple zeta." 6-31G* is an example of double zeta basis set, and 6-311+G* is an example of a triple zeta basis set. Usage of split valance basis sets improves the representation of a system. It gives more flexibility to the valance region of the system and provides more accuracy.

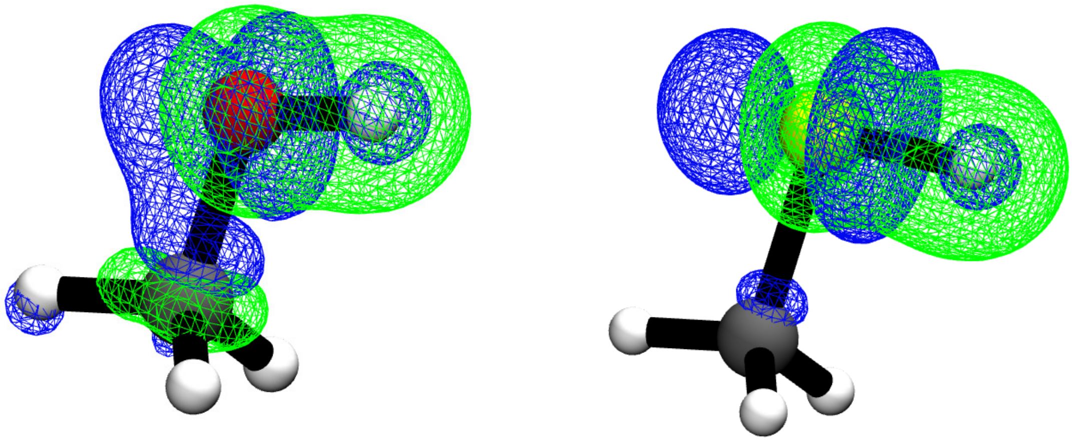
Also, basis sets can have diffuse functions, which provide slower decay over the space, and they might be very beneficial to represent the systems with negative charges such as anions and Rydberg states. Basis sets that contain diffuse functions take up more space, and with functions of neighbor atoms, they give more overlap. It might be a reason why STO-3G gives lower ϕ_S.

The present study used the CPCM solvation model (water as a solvent) to capture more realistic results. However, the comparison of the ϕ_S of CPCM and gas phase was also made. The results for alcohols show that D/A overlap is greater in the gas phase than in the solvent. Also, the difference between vacuum ϕ_S and CPCM ϕ_S increases with increasing pKa. Again, the gas phase gives a greater ϕ_S than CPCM for thiols, but there is no clear trend between the difference and pKa. It seems the difference has a constant range which does not change with pKa.

By using pKa vs. ϕ_S graphs, pKa predictions were also made for alcohols and thiols and compared with the experimental results. Table 3.1 and Table 3.2 shows the experimental pKa's, ϕ_S, predicted pKa values, and the difference between predicted pKa's and experimental ones. Mean Absolute Error (MAE) and Standard Error (SE) were calculated to compare the predicted and measured numbers. The results, unfortunately, are not very good. The predicted numbers are generally higher than the experimental ones for thiols, but the trend is not straightforward for alcohols. Also, the MAE and SE for alcohols are much greater than the thiols. It might be due to the training set for thiols containing fewer molecules which also do not have as broad a pKa range as alcohols, and the experimental pKa's might not be perfect. With all these, it seems our descriptor is not capable of predicting pKa's of different chemical functions in one single equation.

Figure 4.1 shows detachment/attachment density for methanol and methanethiol with the same methods, basis set, and isovalue. We showed in Figure 3.7 that ϕ_S correlates very well with pKa for alcohols and thiols, but a correlation of both of them in the same equation is not possible. Figure 4.1a and Figure 4.1b show how different D/A densities are on methanol and methanethiol. These two are notably different, and one of their differences can be used as a second descriptor.

Our descriptor ϕ_S is a normalized quantity that changes between 0 and 1. However, by having only ϕ_S, we lose the information about the total charge involved in the



(a) (b)

Figure 4.1: Comparisons of D/A densities of methanol and methanethiol

transition (de-protonation in our case) (namely ϑ), and we can have more information about the process by introducing the total charge (ϑ) or the amount of moved charge (qCT).

Thus, ϑ and qCT are two quantities related to charge, and in this study they were also searched. The formulas for calculations of ϑ and qCT can be seen at the end of the Mathematical details section. The investigated descriptors are ϑ, qCT and their combinations; ϕ_S*qCT, ϑ - qCT, (ϑ - qCT)*ϕ_S, and ϕ_S*(ϑ - qCT)/ϑ.
In an attempt to design a descriptor that contains information about both how much charge is involved in the reorganization and how this charge is reorganized, we explored a few combinations of ϕ_S, ϑ, and qCT. Overall, the descriptor ϕ_S*qCT gave the most promising results. Unlike all other attempts, ϕ_S*qCT lets us predict the pKa of alcohols and thiols in a single equation. The correlation coefficient obtained without outliers is 0.85, which is fair, and the mean and standard errors are 1.11 and 0.48 respectively. The the best of our knowledge, this is the first time a single equation based on electronic descriptors can be used to predict the pKa of different chemical functions.

CHAPTER 5

CONCLUSION

In the present study, we repurposed the D/A density calculation, which initially was for the excitation process, for de-protonation reaction. Descriptors extracted from this analysis (i.e., ϕ_S, ϑ, and qCT) were investigated as potential predictors of pKa for a series of alcohols and thiols. This study investigates aliphatic molecules, namely alcohols and thiols, with a broad pKa range. The calculations were performed with hf, b3lyp, ωb97xd, and m062x theories and STO-3G, 6-31G, 6-31G*, 6-31+G*, and 6-311+G* basis sets with CPCM solvation method in water and by using ωb97xd functional with 6-311+G* basis set gas phase and CPCM results were compared. The findings suggest that ϕ_S is not functional dependent for both alcohols and thiols, but HF and DFT give different results. Basis set dependency for alcohols shows that STO-3G and Pople basis sets give different results, but Pople basis sets are consistent, and 6-31G is enough to use; there is no need to use polarization or diffuse functions or triple zeta. However, the results of basis set investigations of thiols are somehow distinct. The type of Pople Basis sets differs in thiols. Adding diffuse functions changes the results; however, adding polarization functions (though there are some exceptions for thiols) or triple zeta does not affect the results much notably. Thus, there is no need to use triple zeta basis sets or those containing polarization functions. One should be careful to add diffuse functions.

From pKa vs. ϕ_S graphs, we showed that pKa and our descriptor ϕ_S are correlated very well with $R^2 = 0.93$ and 0.87 for alcohols and thiols, respectively, but alcohols and thiols can not correlate together. Several studies have already proven that it is not possible to use only one descriptor to achieve a global correlation. Similarly, we also need a second or more descriptor(s) to achieve a general descriptor of alcohols and thiols correlating in the same equation.

Some new descriptors, different from our first descriptor ϕ_S, were also investigated in the Investigations of other descriptors section. Among all our trials ϕ_S*qCT give a general correlation with $R^2 = 0.85$ for a general set containing both alcohols and thiols set. By disregarding a few outliers, our predictions give an accuracy of about +/- 2 pKa units. This result is already fair by itself, but the fact that our descriptor can predict the pKa of two different chemical functions in a single equation is, to the best of our knowledge, a first in the field. Moreover, we now have a better understanding of the fundamental electronic properties hidden behind the pKa values of molecules. We can indeed dissect the ϕ_S*qCT descriptor into contributions of charge redistribution locality and amount of charge involved in the redistribution upon de-protonation. We trust that these results will eventually help the rational design of novel molecules with tuned acidity.

In the present study, the reference pKa values come from different sources. Thus, the error in experimental pKa might also affect the correlation here and might be the source of the outliers observed in our fits. Future studies involving experiments would be valuable in order to design a trustful and consistent set of experimental pKa's. Although this study explores alcohols and thiols only, it is expected that the descriptor can be applied to other classes of compounds. Also, all the alcohols in the present study have only one OH (hydroxyl) group, but symmetric or non-symmetric polyhydroxy alcohols (and thiol equivalents), such as diols and triols, can be investigated too. The next step will naturally be to include carboxylic acids to the set. Ultimately, we aim at probing the acidity of any proton, even aromatic or aliphatic ones.

Finally, all calculations performed in this work were done in a static manner, considering only a single conformation of the molecules. The molecules were chosen such that their conformational space is easily guessable, yet each molecule in the set is expected to have a different dynamic behaviour. Future steps in the project should involve the dynamic of the molecules. One way to take it into account could be performed molecular dynamics simulations and to extract an ensemble of structures on which our protocol could be applied. The averaged results might give better agreements with experimental values.

REFERENCES

[1] M. Namazian and S. Halvani, “Calculations of pka values of carboxylic acids in aqueous solution using density functional theory,” *The Journal of Chemical Thermodynamics*, vol. 38, no. 12, p. 1495–1502, 2006.

[2] L. Z. Benet and J. E. Goyan, “Potentiometric determination of dissociation constants,” *Journal of Pharmaceutical Sciences*, vol. 56, no. 6, p. 665–680, 1967.

[3] A. Apelblat, “Dissociation constants and limiting conductances of organic acids in water,” *Journal of Molecular Liquids*, vol. 95, no. 2, p. 99–145, 2002.

[4] D. Waldron-Edward, “The micro determination of acid and base dissociation constants by paper electrophoresis,” *Journal of Chromatography A*, vol. 20, p. 556–562, 1965.

[5] A. Loewenstein and S. Meiboom, “Rates and mechanisms of protolysis of di- and trimethylammonium ions studied by proton magnetic resonance,” *The Journal of Chemical Physics*, vol. 27, no. 5, p. 1067–1071, 1957.

[6] J. Quijada, G. López, R. Versace, L. Ramírez, and M. L. Tasayco, “On the nmr analysis of pka values in the unfolded state of proteins by extrapolation to zero denaturant,” *Biophysical Chemistry*, vol. 129, no. 2-3, p. 242–250, 2007.

[7] G. R. Grimsley, J. M. Scholtz, and C. N. Pace, “A summary of the measured pk values of the ionizable groups in folded proteins,” *Protein Science*, 2008.

[8] K. J. Nelson, D. Parsonage, A. Hall, P. A. Karplus, and L. B. Poole, “Cysteine pka values for the bacterial peroxiredoxin ahpc,” *Biochemistry*, vol. 47, no. 48, p. 12860–12868, 2008.

[9] C. Horvath, W. Melander, and I. Molnar, “Liquid chromatography of ionogenic substances with nonpolar stationary phases,” *Analytical Chemistry*, vol. 49, no. 1, p. 142–154, 1977.

[10] S. G. Schulman and A. C. Capomacchia, “Variations of fluorescence quantum yields with ph or hammett acidity. near equilibrium vs nonequilibrium excited state proton exchange,” *The Journal of Physical Chemistry*, vol. 79, no. 14, p. 1337–1343, 1975.

[11] J. Reijenga, A. van Hoof, A. van Loon, and B. Teunissen, “Development of methods for the determination of pkavalues,” *Analytical Chemistry Insights*, vol. 8, 2013.

[12] R. Casasnovas, J. Ortega-Castro, J. Frau, J. Donoso, and F. Muñoz, “Theoretical pkacalculations with continuum model solvents, alternative protocols to thermodynamic cycles,” *International Journal of Quantum Chemistry*, vol. 114, no. 20, p. 1350–1363, 2014.

[13] M. Abul Kashem Liton, M. Idrish Ali, and M. Tanvir Hossain, “Accurate pka calculations for trimethylaminium ion with a variety of basis sets and methods combined with cpcm continuum solvation methods,” *Computational and Theoretical Chemistry*, vol. 999, p. 1–6, 2012.

[14] M. Namazian, M. Zakery, M. R. Noorbala, and M. L. Coote, “Accurate calculation of the pka of trifluoroacetic acid using high-level ab initio calculations,” *Chemical Physics Letters*, vol. 451, no. 1-3, p. 163–168, 2008.

[15] M. D. Liptak, K. C. Gross, P. G. Seybold, S. Feldgus, and G. C. Shields, “Absolute pka determinations for substituted phenols,” *Journal of the American Chemical Society*, vol. 124, no. 22, p. 6421–6427, 2002.

[16] J. F. Satchell and B. J. Smith, “Calculation of aqueous dissociation constants of 1,2,4-triazole and tetrazole: A comparison of solvation models,” *Phys. Chem. Chem. Phys.*, vol. 4, no. 18, p. 4314–4318, 2002.

[17] K. C. Gross, P. G. Seybold, Z. Peralta-Inga, J. S. Murray, and P. Politzer, “Comparison of quantum chemical parameters and hammett constants in correlating pka values of substituted anilines,” *The Journal of Organic Chemistry*, vol. 66, no. 21, p. 6919–6925, 2001.

[18] K. C. Gross, P. G. Seybold, and C. M. Hadad, “Comparison of different atomic

charge schemes for predicting pka variations in substituted anilines and phenols," *International Journal of Quantum Chemistry*, vol. 90, no. 1, p. 445–458, 2002.

[19] A. M. Baptista, P. J. Martel, and S. B. Petersen, "Simulation of protein conformational freedom as a function of ph: Constant-ph molecular dynamics using implicit titration," *Proteins: Structure, Function, and Genetics*, vol. 27, no. 4, p. 523–544, 1997.

[20] J. A. Wallace and J. K. Shen, "Predicting pka values with continuous constant ph molecular dynamics," *Methods in Enzymology*, p. 455–475, 2009.

[21] S. G. Itoh, A. Damjanović, and B. R. Brooks, "Ph replica-exchange method based on discrete protonation states," *Proteins: Structure, Function, and Bioinformatics*, vol. 79, no. 12, p. 3420–3436, 2011.

[22] Y. Meng, D. Sabri Dashti, and A. E. Roitberg, "Computing alchemical free energy differences with hamiltonian replica exchange molecular dynamics (h-remd) simulations," *Journal of Chemical Theory and Computation*, vol. 7, no. 9, p. 2721–2727, 2011.

[23] J. M. Swails and A. E. Roitberg, "Enhancing conformation and protonation state sampling of hen egg white lysozyme using ph replica exchange molecular dynamics," *Journal of Chemical Theory and Computation*, vol. 8, no. 11, p. 4393–4404, 2012.

[24] C. A. Fitch, D. A. Karp, K. K. Lee, W. E. Stites, E. E. Lattman, and E. B. García-Moreno, "Experimental pka values of buried residues: Analysis with continuum methods and role of water penetration," *Biophysical Journal*, vol. 82, no. 6, p. 3289–3304, 2002.

[25] D. Bashford and M. Karplus, "Pka's of ionizable groups in proteins: Atomic detail from a continuum electrostatic model," *Biochemistry*, vol. 29, no. 44, p. 10219–10225, 1990.

[26] J. Antosiewicz, J. McCammon, and M. K. Gilson, "Prediction of ph-dependent properties of proteins," *Journal of Molecular Biology*, vol. 238, no. 3,

p. 415–436, 1994.

[27] J. Antosiewicz, J. A. McCammon, and M. K. Gilson, "The determinants of pkas in proteins," *Biochemistry*, vol. 35, no. 24, p. 7819–7833, 1996.

[28] V. Dillet, R. L. Van Etten, and D. Bashford, "Stabilization of charges and protonation states in the active site of the protein tyrosine phosphatases: a computational study," *The Journal of Physical Chemistry B*, vol. 104, no. 47, p. 11321–11333, 2000.

[29] R. Anandakrishnan, B. Aguilar, and A. V. Onufriev, "H++ 3.0: Automating pk prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations," *Nucleic Acids Research*, vol. 40, no. W1, 2012.

[30] H. Li, A. D. Robertson, and J. H. Jensen, "Very fast empirical prediction and rationalization of protein pka values," *Proteins: Structure, Function, and Bioinformatics*, vol. 61, no. 4, p. 704–721, 2005.

[31] D. C. Bas, D. M. Rogers, and J. H. Jensen, "Very fast prediction and rationalization of pka values for protein-ligand complexes," *Proteins: Structure, Function, and Bioinformatics*, vol. 73, no. 3, p. 765–783, 2008.

[32] M. H. Olsson, C. R. Søndergaard, M. Rostkowski, and J. H. Jensen, "Propka3: Consistent treatment of internal and surface residues in empirical pka predictions," *Journal of Chemical Theory and Computation*, vol. 7, no. 2, p. 525–537, 2011.

[33] K. S. Alongi and G. C. Shields, "Theoretical calculations of acid dissociation constants: A review article," *Annual Reports in Computational Chemistry*, p. 113–138, 2010.

[34] S. Zhang, J. Baker, and P. Pulay, "A reliable and efficient first principles-based method for predicting pka values. 2. organic acids," *The Journal of Physical Chemistry A*, vol. 114, no. 1, p. 432–442, 2009.

[35] S. Zhang, J. Baker, and P. Pulay, "A reliable and efficient first principles-based method for predicting pka values. 1. methodology," *The Journal of Physical*

Chemistry A, vol. 114, no. 1, p. 425–431, 2009.

[36] A. P. Harding and P. L. Popelier, “Pka prediction from an ab initio bond length: Part 2—phenols,” *Physical Chemistry Chemical Physics*, vol. 13, no. 23, p. 11264, 2011.

[37] G. Caballero-García, G. Mondragón-Solórzano, R. Torres-Cadena, M. Díaz-García, J. Sandoval-Lira, and J. Barroso-Flores, “Calculation of vs,max and its use as a descriptor for the theoretical calculation of pka values for carboxylic acids,” *Molecules*, vol. 24, no. 1, p. 79, 2018.

[38] I. Ugur, A. Marion, S. Parant, J. H. Jensen, and G. Monard, “Rationalization of the pka values of alcohols and thiols using atomic charge descriptors and its application to the prediction of amino acid pka’s,” *Journal of Chemical Information and Modeling*, vol. 54, no. 8, p. 2200–2213, 2014.

[39] Z. P. Haslak, S. Zareb, I. Dogan, V. Aviyente, and G. Monard, “Using atomic charges to describe the pka of carboxylic acids,” *Journal of Chemical Information and Modeling*, vol. 61, no. 6, p. 2733–2743, 2021.

[40] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry*. Dover Publications, 1996.

[41] C. C. Roothaan, “New developments in molecular orbital theory,” *Reviews of Modern Physics*, vol. 23, no. 2, p. 69–89, 1951.

[42] A. Marion, *Molecular dynamics using a semiempirical quantum force field: development and applications to systems of biological interest*. PhD thesis, Université de Lorraine, 2014.

[43] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, “General performance of density functionals,” *The Journal of Physical Chemistry A*, vol. 111, no. 42, p. 10439–10452, 2007.

[44] N. Mardirossian and M. Head-Gordon, “Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals,” *Molecular Physics*, vol. 115, no. 19, p. 2315–2372, 2017.

[45] M. Caricato, G. W. Trucks, M. J. Frisch, and K. B. Wiberg, "Electronic transition energies: A study of the performance of a large range of single reference density functional and wave function methods on valence and rydberg states compared to experiment," *Journal of Chemical Theory and Computation*, vol. 6, no. 2, p. 370–383, 2010.

[46] T. Etienne, X. Assfeld, and A. Monari, "Toward a quantitative assessment of electronic transitions' charge-transfer character," *Journal of Chemical Theory and Computation*, vol. 10, no. 9, p. 3896–3905, 2014.

[47] T. Etienne, X. Assfeld, and A. Monari, "New insight into the topology of excited states through detachment/attachment density matrices-based centroids of charge," *Journal of Chemical Theory and Computation*, vol. 10, no. 9, p. 3906–3914, 2014.

[48] T. Etienne, "Mesra software."

[49] G. Bruhn, E. R. Davidson, I. Mayer, and A. E. Clark, "Löwdin population analysis with and without rotational invariance," *International Journal of Quantum Chemistry*, vol. 106, no. 9, p. 2065–2072, 2006.

[50] M. Head-Gordon, A. M. Grana, D. Maurice, and C. A. White, "Analysis of electronic transitions as the difference of electron attachment and detachment densities," *The Journal of Physical Chemistry*, vol. 99, no. 39, p. 14261–14270, 1995.

[51] J. Wong, D. A. Walthall, and J. I. Brauman, "Determination of the acidity of long chain alcohols using infrared multiple photon dissociation," *The Journal of Physical Chemistry A*, vol. 106, no. 51, p. 12299–12304, 2002.

[52] B. Tehan, E. Lloyd, M. Wong, W. Pitt, J. Montana, D. Manallack, and E. Gancia, "Estimation of pka using semiempirical molecular orbital methods. part 1:application to phenols and carboxylic acids.," *Quantitative Structure-Activity Relationships*, vol. 21, no. 5, p. 457–472, 2002.