Machine Learning Predictions of Spectroscopic Properties and Carbonyl Reactivity From A Database of Charge Density Descriptors

By

Kiran Kumar Donthula

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
in
Computational Science

Middle Tennessee State University

December 2019

Dissertation Committee:

Dr. Preston J. MacDougall, Committee Chair

Dr. Cen Li

Dr. John Wallin

Dr. Anatoliy Volkov

I dedicate this research to my loving,

Father and Mother,

who made me believe in myself and whose love, encouragement, and guidance make me able to achieve such success and honor.

I love you, Amma and Nanna, now and forever.

ACKNOWLEDGEMENTS

First, I would like to thank and praise the God for his blessings to pursue a Ph.D. in Computational Science which was one of the biggest dreams in my life.

I would like to express my sincere gratitude to my research advisor, **Prof. Preston J. MacDougall**. He is a role model of the greatest scientists. I feel extremely fortunate to have him as my mentor. I thank him for his enormous support, kindness, encouragement, great respect, and remarkable ability to find time in spite of being very busy whenever I have approached him to discuss anything. I again thank him for providing a unique opportunity to work on this interesting project.

I am very thankful to my wonderful doctoral committee members: **Dr. John Wallin**, **Dr. Anatoliy Volkov**, and **Dr. Cen Li** for being part of my Ph.D. committee. I appreciate all their time, support, guidance, and helpful suggestions regarding Ph.D. work at MTSU. I gratefully acknowledge **Dr. Anatoliy Volkov** for his guidance, help, and knowledge. I would like to thank my Master's advisors **Dr. Ngee Sing Chong** and **Dr. Ooi** for their constant support. I would also like to thank **Dr. Judith Iriarte-Gross** for her encouragement and providing an opportunity to participate in WISTEM and EYH programs.

At last but not least I want to express my gratitude and appreciation to my family for encouraging and supporting me my entire life and for their sacrifices made on my behalf. I would especially like to thank my father **Damodar** and mother **Padma** for teaching me what it means to be a good human being. A very special thanks to my beautiful wife **Shruthi** who stood beside me everyday and encouraged when it seemed impossible. A special thanks to my lovely son **Kushansh Vidhu** for all his unconditional love. I would like to thank my brother **Raj kumar**, his wife **Shravani**, and their baby **Rishank Vidhu** for their encouragement and support.

I would also like to thank my in-laws **Venu Gopal** and **Sujatha**, my brother-in-laws **Rahul** and **Praneeth**, my sisters **Swathi** and **Sowmya**, and my nephew **Dhruv Athiradh** for their love, best wishes, and support.

I wish to avail myself of this opportunity to express a sense of gratitude and love to my friends for providing me the right environment and moral support I needed for my work. A very special thanks to my friends for helping and encouraging me to complete this work especially **Sachintha Pitigala** and **Mathew Wang**. I would also like to thank my best friend **Srikanth Manda** for his longstanding and unconditional belief in my abilities and immeasurable support. I would also like to thank all my friends here and back home in India who supported me during my studies.

Finally, I gratefully acknowledge Middle Tennessee State University, Department of Computational Science, and Department of Chemistry for providing me this opportunity and all the financial support.

ABSTRACT

Carbonyl compounds are important to study because of their biological and industrial significance. A database of critical point descriptors for valence-shell charge concentrations and depletions of carbon atoms in a range of aldehydes, ketones, imides, and amides has been created. For each critical point, the database contains data related to the probability distribution of electrons (value of the total electron density, ρ, at bond critical points which have been correlated with bond strength). This includes, data related to the curvature of ρ at maxima and minima in carbon's valence shell of charge concentration (VSCC) ($\nabla^2\rho(\mathbf{r})$ and Hessian eigen values, which have been correlated with chemical reactivity). For both types of critical points, radii from the enveloped carbon nucleus are included in the database.

Artificial neural networks (ANNs) are strong tools for predicting nonlinear functions, and they are used in this study to both leverage charge density-based descriptors and learn about their relative chemical significance. An ANN prediction scheme was developed for the spectroscopic properties and interaction energies of carbonyl compounds, based on the topological properties of electron density obtained from QTAIM (The input data necessary for training and testing the proposed ANN scheme was data obtained from Quantum Theory of Atoms In Molecules.). In 2009, Balabin and Lomakina [1] used three-layer feed-forward artificial neural networks, with back propagation, to predict density functional theory (DFT) energies that are comparable to those obtained with large basis set using lower-level energy values as training data. These studies, and others, indicate that data-mining techniques, used in conjunction with artificial neural networks, can be productively applied in the prediction of properties that would otherwise be computationally expensive and time-consuming to calculate.

For our study, we have selected 225 small systems of carbonyl group-containing molecules as a training set, with each molecule containing 18 bond critical point descriptors and 30 Laplacian critical point descriptors. These properties were used to train ANN for predicting C=O stretching frequencies and ^{13}C chemical shifts. Additional properties, such as intermolecular interaction energies with nucleophiles are also estimated. Predictions are made using the Laplacian critical point data, as well as the bond critical point data, both separately and combined. The study was carried out using the leave-one-out cross validation method. Expected Mean Absolute Percent Errors (MAPE) and Mean Absolute Errors (MAE) are compared between these three data sets. The calculated MAPE for neural network predictions of ^{13}C shifts and C=O stretching frequencies are 1.38, 0.53. MAEs for neural network predictions of covalent and van der Waals interaction energies are 3.44 kcal/mol and 4.78 kcal/mol. Here, all molecular wave functions have been generated using Gaussian 09 [2], and electron density analysis is done using programs AIMAll [3] and DenProp [4].

For the stretch-test we chose the *E. coli.* enzyme D-fructose-6-phosphate aldolase (FSA) [5], which catalyzes a nucleophilic addition reaction of a carbon nucleophile (ketone) to a carbon electrophile (aldehyde). The covalent interaction energy between a nucleophile and an electrophile within the binding pocket of an enzyme (FSA) is predicted by our ANN with an absolute error of 3.2 kcal/mol.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER 1 **Introduction** 1

1.1 Background of Machine-Learning in Molecular Modeling 2

1.2 Quantum Theory of Atoms in Molecules (QTAIM) 15

1.3 Introduction to Aldehydes and Ketones 24

1.4 Spectroscopic Properties of Aldehydes and Ketones 28

1.4.1 IR Spectroscopy 28

1.4.2 ^{13}C NMR Spectroscopy 30

1.5 Interaction Energies 35

1.6 Artificial Neural Networks (ANN) 37

1.7 Computational Chemistry in Drug Design 43

1.7.1 Structure-Based Computer Aided Drug Design (SBDD) . . . 44

1.7.2 Ligand-Based Computer Aided Drug Design (LBDD) 49

CHAPTER 2 **Methodology** 57

2.1 Electronic Structure Methods 57

2.2 Electron Density Analysis 58

2.3 Development of the ANN Model 60

2.4 Datasets 63

CHAPTER 3 **Results and Discussion** 68

3.1 Using Artificial Neural Networks in Conjunction with Topological Analysis of Charge Distributions 68

3.2 Predicting Spectroscopic Properties Using Artificial Neural Networks . 73

3.3 Predicting Covalent and Van der Waals Interaction Energies Using ANNs 86

3.4 Leveraging Machine-Learning with Small Molecules to Predict Ligand Interactions with Large Molecules 97

CHAPTER 4 **Conclusions** 112

BIBLIOGRAPHY 115

Appendices 156

CHAPTER A The Molecules Used in This Study 157

CHAPTER B NMR and IR Predicted Values 164

LIST OF TABLES

Table 1 – Sample BCP input data for 2-methylbutanal 64

Table 2 – Sample LCP Input Data for 2-methylbutanal 64

Table 3 – Mean absolute percent errors (MAPEs) of predicted values for ^{13}C chemical shifts and C=O stretching frequencies when ANN trained using **default** parameters of the network 77

Table 4 – The combination of compounds used in the dataset to create the ANN model 79

Table 5 – Comparison of MAPE of predicted values of ^{13}C chemical shifts and C=O stretching frequencies when using **optimized** parameters of ANN for three experiments 81

Table 6 – Comparison of optimized parameters used in all the three experiments in the prediction of spectroscopic properties 82

Table 7 – MAEs of predicted covalent $\triangle E_{int}$ and van der Waals $\triangle E_{int}$ using the **default** parameters for training ANN 92

Table 8 – MAEs of predicted covalent $\triangle E_{int}$ and van der Waals $\triangle E_{int}$ using the **optimized** parameters for training ANN 93

Table 9 – Comparison of the optimized parameters used in all the three experiments in the prediction of $\triangle E_{int}$ values 94

Table 10 – Comparison of covalent $\triangle E_{int}$ in the binding pocket of the FSA enzyme105

Table 11 – Comparison of van der Waals $\triangle E_{int}$ values between FSA binding pocket and the nucleophile 107

Table 12 – Predicted NMR shifts 165

LIST OF FIGURES

Figure 1 – The VSCC of carbonyl carbon of acetone 20

Figure 2 – Structure of an aldehyde and ketone 25

Figure 3 – (A) Resonance effect, (B) Inductive effect, explicitly showing the bond polarity with partial charges, (C) Nucleophilic addition 25

Figure 4 – The angle of nucleophilic attack on carbonyl group 26

Figure 5 – The infrared spectrum of acetone 28

Figure 6 – Representation of a natural neuron 38

Figure 7 – Representation of an artificial neuron 38

Figure 8 – Representation of an artificial neural network 39

Figure 9 – Labelling scheme for the series of carbonyl compounds 58

Figure 10 – Representation of an artificial neural network 62

Figure 11 – Molecular graph of 2-methylbutanal 65

Figure 12 – The Laplacian of electron density of 2-methylbutanal 66

Figure 13 – The Laplacian of electron density of formaldehyde 71

Figure 14 – Geometry used in the current study 88

Figure 15 – Interaction energy curve of acetone and F^- interaction 88

Figure 16 – The Laplacian ($\nabla^2\rho$=0 surface) of the total charge density for the covalent interaction between acetone and fluoride ion 89

Figure 17 – The Laplacian ($\nabla^2\rho$=0 surface) of the total charge density for the van der Waals interaction between acetone and fluoride ion 90

Figure 18 – The binding pocket of FSA with substrate 3-hydroxypropanal 100

Figure 19 – The Laplacian ($\nabla^2\rho$=0 surface) of the total charge density for the covalent interaction between carbonyl carbon and fluoride ion 103

Figure 20 – The Laplacian ($\nabla^2\rho$=0 surface) of the total charge density for the covalent interaction between carbonyl carbon and fluoride ion 104

Figure 21 – Labeling scheme used for binding pocket wall 107

Figure 22 – The Laplacian ($\nabla^2\rho$=0 surface) of the total charge density for the 11-mer polypeptide cyclosporin A 111

Figure 23 – Molecular structures of compounds used in our study 158

CHAPTER 1

Introduction

Chemists have been trying to develop more and more realistic and valuable models of atoms and molecules for centuries. In the early 1800s, Dalton described atoms as tiny, inseparable, indestructible particles which have certain mass, size, and chemical behavior. Later, J. J. Thompson determined that atoms are composed of spheres of evenly distributed positive and negative charges. This became popular as the "plum pudding" model of the atom.

Rutherford, Bohr, Schrödinger, and others extended and enhanced the earlier models to give us our current working model of the atom. The noticeable point is that models change and progress over time. Today we have a wide collection of molecular models available to mimic and study the processes and interactions that take place at atomic level. Due to advances in computer and software technologies over the last few decades, computational chemists routinely perform formerly unattainable simulations. Computer technology also enables advanced data analysis approaches to be used, and facilitates visualization methods that greatly enhance our understanding of the chemical world.

Chemistry permeates every aspect of our life, from the advancement of new drugs, to the food that we eat, and materials we use on a day-to-day basis. Chemistry depend on empirical information based on creative and strenuous experimentation that leads to discoveries of compounds and materials with the desired properties as well as evermore efficient methods to synthesize them. Many innovations in chemistry are guided by searching huge databases of computational or experimental molecular structures and properties by using ideas based on chemical similarity [6]. The structure and properties of molecules can be obtained from quantum mechanics. So, chemical discovery should be based on fundamental quantum principles. Indeed, quantum-mechanical calculations and machine-learning (ML) have currently been merged and are aiming towards the goal of sped-up discovery of chemicals with desired properties [7, 8, 9].

The motivation behind our study is to explore the use of ML methods to maximally leverage high-resolution topological information of molecular fragments to efficiently and accurately model arbitrarily complex chemical systems. In order to prove this principle, small chemical systems containing carbonyl compounds have been selected as a training set. An additional enzyme active site has been modeled with a cluster and used as a biomolecular application test. Carbonyl compounds were chosen for our study because of their biological importance, as the carbonyl group is present in all amino acids, nucleic acids, natural esters, reducing sugars *etc.* [10, 11, 12]. The probability of finding a carbonyl group in a binding site in protein-ligand interaction studies is also high [13].

In this study, we attempt to predict chemical properties of a test molecule by utilizing the knowledge of chemical information (training data) for similar molecules. This can be achieved using ML algorithms [14]. Specifically, artificial neural networks (ANN) possess this capability [15]. The chemical properties and biological activity of compounds can be studied using their molecular and electronic structure information derived from quantum-mechanical calculations. We postulate that useful predictions of chemical properties can be made efficiently based only on the electronic structure represented by a small number of variables containing condensed chemical information obtained from the theory of Quantum Theory of Atoms in Molecules (QTAIM) developed by Prof. Richard F. W. Bader *et al.* [16].

1.1 Background of Machine-Learning in Molecular Modeling

A much-quoted description of ML by Arthur Samuel in 1959 is “it is a field of study that gives computers the ability to learn without being explicitly programmed ” [17]. ML systems are those which learn from data to build a model, and that model is then applied for further studies. Instead of being particularly programmed to solve a unique problem, these algorithms depend on given data to produce statements about new data. The simplest example for a ML algorithm would be regression: based on finite number of examples, a function is inferred which enables predictions for new samples.

ML is commonly categorized into four types: supervised, semi-supervised, unsupervised, and reinforcement learning [14]. Although all ML types have been applied in chemical research, supervised learning has so far been used most frequently [14], and we have also used supervised algorithms in our study. This popularity may be due to its heuristic and intuitive approach to learning, which is similar to scientist's way of gaining insights into structure-property relationships. In supervised prediction model, examples are pairs of input x and label y, for example molecules and their energy and the task is to predict the label of new examples, that is, to learn the function f:$x \rightarrow y$. If x and y are continuous variables, the mapping is a regression; if they are discrete, then it is a classification [14]. To train and optimize model f, we can utilize a large group of supervised ML algorithms to approximate the output value for a given input. There are many ML algorithms exist, however the ones utilized most frequently in the cheminformatics belong to two large families, kernel-based ML [18] and ANN [19].

ML has been successfully employed in a wide variety of fields, including recommender systems [20], brain computer interfaces [21], robotics [22], web searching [23], spam filters [24], credit scoring [25], stock trading [26], drug design [27], cheminformatics [28], speech recognition [29], image recognition [30], and many other applications. In a similar manner, the application of data mining and machine-learning tools in different fields of science, especially in many branches of theoretical and computational chemistry, has been emerging in the last few years (for reviews, see Refs. [7, 14, 31, 32]).

The concept of "chemical space" proposes a representation of the molecules and their properties in form of a geographical map [33]. And this theoretically possible chemical space is astronomically vast. To get such a map, one initially creates a property space by assigning dimensions to series of molecular descriptors. Each molecule is placed in this multidimensional property space utilizing the descriptor values as positional coordinates, as first presented by Pearlman and Smith [34]. A huge number of various molecular descriptors exist, and any combination of these descriptors might be chosen to create a property space formally having tens to hundreds of different dimensions, from which a chemical space can be derived. Given some molecule, represented by its number of electrons and set of nuclei at their equilibrium geometries, one can commonly predict its observable properties using *ab initio* quantum chemical methods such as CCSD(T) in a sufficiently large basis. This is possible for small molecules, and density functional theory (DFT) can be used even though it is less reliable for larger ones. But even DFT is not fast enough to search the entire chemical compound space, the size of which increases in combination with the number of atoms and distinct elements. Hence, a significant problem is to search chemical compound space to find new drugs with desired functionalities.

The fundamental property of a molecule is its ground-state energy. In addition, there are also many interesting properties at the ground-state configuration, for example, ionization potentials, dipole moments, and vibrational frequencies. Some of these properties can be extracted from the same electronic structure calculation from which the molecule's energy was obtained, while others need some extra computation. Given the impossibility of computing desired properties of all possible molecules, it is interesting to inquire whether a ML algorithm trained on known examples can be used to predict the properties of all possible molecules at significantly lower computational cost [35]. Provided that this is true, chemical compound space can be searched orders of magnitude faster. Various groups are therefore formulating procedures to do this.

For example, Rupp and co-workers [7] successfully developed a ML model based on nonlinear statistical regression for the prediction of atomization energies of organic molecules. The atomization energy is an important molecular property which gives information on the stability of a molecule with respect to the atoms that constitute it. These energy values can be obtained experimentally [36], or by using computational methods [37]. The molecules considered for their study were obtained from a generated database (GDB), a library which contains 10^9 stable and synthetically accessible organic molecules. To build their model, a subset containing 7165 small organic molecules was obtained from the GDB, each contained up to seven "heavy" atoms that include C, N, O, or S, and were all saturated with hydrogen atoms. Each molecule was represented in the form of a matrix, which contains information of atomic coordinates and nuclear charges, called the Coulomb matrix. The mean absolute errors (MAE) were found to be reduced from 17 kcal/mol to 10 kcal/mol when they increased the size of the training set from 500 to 7000 molecules. They found that training on information from 15% of the molecules allows predictions for the remaining 85% with an accuracy of approximately 15 kcal/mol. The best result in their work was getting errors close to $\approx$ 10 kcal/mol without doing quantum- mechanical calculations. Their method by-passed solving the Schrödinger equation and reduced calculation time by several orders of magnitude to only several seconds per molecule.

A subsequent publication from Hansen *et al.* [31], studied atomization energies of molecules in their ground-state equilibrium geometry. In this study, they tried to improve the ML prediction accuracies of atomization energies using more precise and suitable ML approaches compared to the one explored by Rupp *et al.* As in Rupp *et al.*, this group used the same dataset containing 7165 molecules represented in the form of the Coulomb matrix. In order to apply ML, a molecule's structural information needs to be represented in an appropriate form, a vector of numbers. Hansen *et al.* encountered two difficulties while doing the vectorial representation of structural information. First, the dimensionalities of the Coulomb matrices are different from molecule to molecule due to changes in the number of atoms present in each molecule. The second issue was that there is no particular ordering of atoms in the Coulomb matrix. Therefore, one can obtain many different

Coulomb matrices for the same molecule, while the energies of these configurations remain the same. To solve the first problem, each matrix was padded with zeros, which made the size of Coulomb matrices of all the molecules equal. To overcome the second problem, they investigated three representations derived from the Coulomb matrix, and these representations included eigen-spectrum representation, sorted Coulomb matrices, and randomly sorted Coulomb matrices.

In the eigen-spectrum representation, each molecule was encoded as a vector of eigenvalues which is invariant in terms of permutations of rows and columns of the matrix. It also reduced the dimensionality of the matrix. In sorted Coulomb matrices, a unique ordering of the atoms was done by rearranging the matrix in such a way that the rows of the matrix were ordered based on their norm. They have used Eucledian L2- norm, in which the square root of the sum of the squared vectors is calculated. To generate randomly sorted training data, the Coulomb matrix was constructed based on a random ordering of the atoms with a vector containing the norm of each row. ML algorithms were trained on the three kinds of datasets mentioned above to predict atomization energies. Their results indicated that the prediction performance was influenced by the representation of a molecule. Among these three representations of molecular data, randomly sorted Coulomb matrices reduced the prediction error from 10 kcal/mol to 3 kcal/mol.

In 2013, Montavon and Rupp developed a ML model that simultaneously predicts a variety of molecular electronic properties from a single query [38]. The predicted properties include atomization energy, polarizability, ionization potential, electron affinity, and excitation energies. In their model, an artificial neural network was trained on a database of *ab initio* calculation results for 7000 stable small organic molecules. Once the neural network had been trained, the prediction of properties for a molecule was made in approximately 100 milliseconds rather than spending hours or days for making reliable quantum-chemical computations. They also observed a systematic reduction of errors correlating with an increase in training set size. There are several examples of machine-learning applications in the approximation of Gibbs energies of formation [39], ionization energies [40], electron affinities [41], and many other properties such as biological properties [42], and chromatographic properties [43].

Applying ML algorithms in the field of computer-aided drug design has a long history [44, 28], notably in quantitative structure-activity relationship (QSAR) applications [45]. In QSAR, the biological activity of a compound is predicted as output using various physical, chemical, and topological properties as input data. In the early 2000s, the literature shows that a wide variety of descriptors were used in QSAR analysis [46, 47, 48]. QSAR has several different applications, namely toxicity prediction [49, 50, 51], physical property prediction, biological activity prediction, antiviral activity prediction, and inhibitor predictions for cancer treatment [52].

Patra *et al.* [53], used ANN to develop a QSAR model which was then used to predict new potent drug candidates for diabetes mellitus. In the case of diabetes mellitus, chronic formation of cataracts is possible and it is important to inhibit the aldose reductase (AR) enzyme in the presence of anti-oxidants. In our study, we have also chosen the *E. Coli.* enzyme D-fructose-6-phosphate aldolase (FSA) as a proof-of-principle test which will be explained in the Discussion section. They chose flavonoids for their study because of their promising AR inhibitory effect and because they are also strong anti-oxidants. They tried to predict the aldose reductase inhibition and strong anti-oxidant activities of flavonoids. In their study, different flavonoids were used to train the network. The training data contained two kinds of molecular descriptors, namely electronegativity (which is an empirical parameter) and molar volume of functional groups, to find the biological activity of the compounds. In contrast, our approach uses measurable or computed well-defined physical properties to train ANNs. In general, an ANN contains one input layer, one or more hidden layers, and an output layer [54]. Neurons in the input layer respond to training data that is provided into the network and passes the weighted output to the hidden layer. The output neurons receive the weighted output from the hidden layer and produces the final output of the network. The predictions of the network can be improved by adjusting different parameters [55]. More details regarding the architecture of ANN and different parameters of the network will be discussed in Section 1.6. Patra *et al.* have done several experiments by changing one of the parameters, specifically the number of neurons present in the hidden layer, to yield better predictions. They found that better predictions were observed when they used 8 hidden neurons. After studying 6561 compounds, they found 10 compounds which exhibit both the AR inhibition and anti-oxidant effects. But, it was not confirmed experimentally.

In 2007, Galicia *et al.* [56], tried to predict the anti-nociceptive activity of several morphinan molecules (configurationally related to morphine) using QSAR analysis. Different quantum chemical and structural descriptors, such as sum of atomic distances between atoms of molecules, were used to model its biological activity. Initially, they used multiple linear regression modeling to select the most relevant QSAR model and then applied ANN to optimize the training results to improve the predicted results. A total of 37 compounds containing a total of 1488 molecular descriptors, which included quantum chemistry descriptors, were used to train the model. Once the multiple linear regression model was built, those were optimized using a feed-forward neural network.

From these studies, one overarching lesson is that it is very important to choose suitable descriptors to represent a molecule in the best possible way to yield better predictions from ML algorithms. A successful QSAR model predicts **not only** the biological activity of a compound, but **also** gives information on what kind of properties (descriptors) were important to build a better model. Our research will likewise shed light on whether bond topological data or valence shell charge concentration (VSCC) topological data are more predictive of the spectroscopic properties and interaction energies that we consider for carbonyl compounds. Both types of data are routinely calculated or measured, by numerous groups around the world, without a clear understanding of which type of data is best suited to different chemical situations.

Nantasenamat *et al.* [57], made a comparative investigation of three types of descriptors in the predictions of spectral properties. They were the first group that tried to computationally predict the spectral properties of green fluorescent protein (GFP), which is an autofluorescent protein containing 238 amino acid residues obtained from the outer dermal layer of the Pacific Northwest Jellyfish. Within the context of molecular orbital theory, the absorption peaks result from an electronic transition between highest occupied molecular orbital (HOMO) to the lowest unoccupied molecular orbital (LUMO). This electronic transition is accompanied by a charge transfer that leads to a change in the electron density distribution. The authors used non-observable HOMO and LUMO energies as training data in the predictions of the desired properties. Three software programs namely RECON [58], E-DRAGON [59], and Spartan'04 [60] were used to produce different descriptors accounting for electron densities, orbital energies, and quantum chemical properties. The predictive ability of these three kinds of descriptors obtained from these programs were evaluated by training the ANN using default network parameters. After comparing their predictive performance, they concluded that the quantum chemical descriptors produced by Spartan'04 gave better predictions.

Use of QTAIM descriptors in the construction of QSAR

In rational drug design, the property or biological activity of a compound is to be predicted from the molecular structure. This has led the field of quantitative structure activity/property relationship (QSAR/QSPR) to evolve new ways to encode chemical structure for use with advanced chemometric and ML methods [61]. These methods should create a model that is easy to understand. It is best accomplished by representing the 3D molecular structure in an efficient and unique way. Although a variety of descriptors have been suggested to define molecules in a QSPR manner, it is important to select appropriate kinds of descriptors that relate to the property being studied. In our study, the focus is on investigation of the potential applicability of topological descriptors that are well-defined within the Quantum Theory of Atoms In Molecules (QTAIM) [62] as training data to predict chemical properties using ML algorithms. Previous research has shown the successful applications of QTAIM descriptors in ML [63].

Alsberg *et al.* [64] was the earliest group to attempt a new approach called struQT to represent the molecular structure using QTAIM descriptors for use in QSAR/QSPR modeling. In this approach, each molecule was again represented in the form of a matrix, but where each row contained bond critical point (BCP) information in terms of spatial and electronic properties. Spatial properties include XYZ coordinates of the BCP, whereas the electronic properties include the values of electron density, ρ, the Laplacian ($\nabla^2\rho$), and the ellipticity parameters. These parameters are those most commonly reported in the literature to summarize the nature of the corresponding bond. For example, electron density at BCP determines the bond order, the Laplacian of the electron density distinguishes two broad classes of bonds (shared vs closed-shell interactions), and the ellipticity detects the σ or π character of a bond. The applicability of this kind of new structure representation was tested by Alsberg *et al.* to predict the wavelength (λ_{max}) of the UV absorption maximum of the first excitation for a set of anthocyanidin compounds. A total of 18 compounds were considered in their study, among which three compounds were used as a validation set. After performing partial least squares regression analysis, the model was applied on the three unseen validation samples. The three mean absolute errors for validation samples 6'-Hydroxyflavylium, 4'-Dihydroxyflavylim, 5,7,3',4',5'-Tetrahydroxyflavylium are 43.9, 0.01, and 5.2 nm respectively. They pointed out that the first compound was an exceptional case based on previous studies, due to which these errors were highly deviated.

The molecular property pK_a $(-logK_a)$ has several applications in different areas, such as medicinal chemistry, biochemistry, pharmaceutical chemistry, and drug development [65]. It is important to calculate these values to determine whether the drug molecule will diffuse through the membranes and various physical barriers such as blood-brain barriers, or not. Popelier and Chaudhry [66] developed a QSAR model using QTAIM descriptors to estimate the pK_a values of a set of aliphatic carboxylic acids, anilines and phenols. The acid-ionization constants, K_a, of compounds are important in terms of determining their pharmacokinetic properties such as protonation states of weak acids and bases at physiological pH levels. Two QSAR models were constructed using the partial least squares methods. The first model was based on topological descriptors and the second one was based on bond lengths. Topological properties such as electron density, the Laplacian, and ellipticity values evaluated at BCP were included in the descriptor matrix for each of the acid, aniline, and phenol molecules. The second, empirical descriptor matrix was constructed using equilibrium bond lengths. A set of 40 carboxylic acids, 36 anilines, and 19 phenols were used in this study. The best model was obtained when they used BCP descriptors, as demonstrated by comparison of correlation coefficient (r^2) and the cross-validated correlation quotient (q^2) values.

Predicting NMR Chemical Shifts

Buttingsrud *et al.* [67], investigated the ability of descriptors based on BCPs in the electron density for predicting theoretically computed proton chemical shifts in a series of substituted benzene compounds. Based on these compounds, four different datasets were created with varying complexity. Datasets 1 and 2 were constructed with only either fluoro- or chloro-substituted benzene compounds. Dataset 3 was prepared with both fluoro- and chloro-substituted benzene rings. The final dataset was obtained by using benzene compounds substituted with various electron-withdrawing and donating groups such as cyano, formyl, amino, hydroxy, methoxy, and methyl groups. The BCP values of electron density, the Laplacian, and ellipticity were used as descriptors with a partial least square regression to develop a QSPR model.

For each dataset, five different models were created based on the number of BCPs used to generate the model. The simplest model was created using only the descriptors of a BCP between the hydrogen atom of interest and the benzene ring. The next model included BCPs connected to the carbon atom bonded to the studied hydrogen atom. Similarly, three more models were built by including more bonds further away from the hydrogen atom. These models were compared to determine whether the most local BCP information is sufficient enough for accurate predictions, or if including more distant bonds is necessary. The results were quantified by root mean squared error of prediction (RMSEP) and cross-validated squared correlation coefficient (q^2) values.

The simplest model for fluoro-substituted benzene compounds (dataset 1) gave good predictions (q^2 = 0.98) of chemical shifts. A very minor improvement in these results was observed by including the additional BCPs to build the model. In chloro-substituted compounds, the simplest model gave poor results (q^2 = 0.1) and these results were improved (q^2 = 0.95) by including more BCPs. For dataset 3, better results (q^2 = 0.7) were observed when the simplest model was used and these were improved after adding more critical points. Finally, the regression with the simplest model was (q^2 = 0.91) improved when they added more BCPs for complex dataset. These studies show the potential of using BCP properties to accurately predict molecular properties, but also that the local properties chosen were not sufficient to obtain the best predictions.

Use of QTAIM bond properties in the construction of QSPR

Buttingsrud *et al.* [68] tested the validity of using BCP properties-based descriptors for building a reliable QSPR model to predict atomic polar tensors (a matrix of gradients of the molecular dipole moment). Several other studies also showed that the descriptors obtained from QTAIM can be successfully used in the field of QSAR/QSPR. Popelier and coworkers [69] developed a method called quantum topological molecular similarity (QTMS) that has been tested for its capability of producing predictive QSAR models. This method starts with the construction of a vector using properties determined at the BCPs. The properties include the electron density at the BCP (ρ_{BCP}), three principal curvatures evaluated at the BCP (λ_1, λ_2, and λ_3), the Laplacian of ρ_{BCP}, bond ellipticity (ε), and the kinetic energy densities ($K(r), G(r)$) at the BCP. The versatility of BCP descriptors has been demonstrated by its ability to predict several physicochemical and pharmacological properties. For example, QTMS descriptors were used for the prediction of hepatocyte toxicity of phenols [70], prediction of basicities of 125 pyridine derivatives [71], prediction of toxicity of aromatic aldehydes to the ecologically important species *Tetrahymena pyriformis* [72], the ecotoxicological hazards of nitroaromatics to the species *Saccharomyces cerevisiae* [73], and many more. To the best of our knowledge, no groups have yet investigated QSAR/QSPR or ML approaches to predicting chemical reactivity using descriptors or training data comprised of VSCC topological data (critical points in $\nabla^2\rho$).

Density Functional Theory (DFT)

The idea of using the electron density as the fundamental descriptor for electronic structure calculations has its roots in 1927 with the theories of Thomas [74] and Fermi [75]. Yet it was not until 1964 that Hohenberg and Kohn provided the mathematical foundations for using the electron density to replace the wavefunction as the main descriptor for a system of electrons [76]. Their important results are outlined as two theorems that are the foundation of DFT, and for which Walter Kohn received the Nobel prize in 1998 [77]. The results from the Hohenberg-Kohn theorems are comparable to wavefunction theory results, but with the electron density playing the prominent role instead of the wavefunction. From the computational point of view, the DFT emerges as a transformative method due to its excellent balance of speed and accuracy. In addition, these methods offer great advantages for large systems like polymers, proteins, *etc*... The number of publications on “density functional theory” has been growing at an astonishing rate. These methods account for about 90 percent of computational chemistry publications; from applications in biochemistry, to materials science [78].

In recent years DFT has been widely used for investigating the chemical properties of different kinds of molecules [79, 80, 81, 82]. In comparison to many other traditional quantum-mechanical techniques, DFT-based calculations give very satisfactory results and use less computational time. Kohn and Hohenberg [76] postulated the existence of a unique functional which determines the ground state energy from the density, exactly. This theorem was the foundation of DFT and launched a global search for the most accurate and general purpose functional.

In a simplified form DFT methods divide the total electronic energy into separate terms according to

$$E = E^T + E^V + E^J + E^{XC} \tag{1}$$

where E^T is the kinetic energy for each electron, E^V is the potential energy associated with nucleus-electron attraction and nucleus-nucleus repulsion, E^J is the electron-electron repulsion, and E^{XC} is the exchange-correlation which corrects overestimation of electron-electron repulsion. All four energy terms in Eq. (1), except for the nucleus-nucleus repulsion, are functions of the electron density. The key success of DFT lies heavily on the formulation of E^{XC}, which is approximated by integrals including primarily the spin densities. The E^{XC} can be expressed as the sum of the two separate parts:

$$E^{XC} = E^X + E^C \tag{2}$$

In Eq. (2), E^X is the exchange energy corresponding to same-spin electron-electron interactions and E^C is the correlation energy corresponding to mixed-spin electron-electron interactions. There are many forms of several functionals which have been developed and implemented to analyze chemical properties of molecules [83]. Some functionals were developed from fundamental quantum mechanics, and some were developed by fitting them to experimental results. These distinct kinds of approaches are attributed to *ab initio* and semiempirical DFT methods, respectively.

The selection of an appropriate exchange-correlation functional relies greatly upon the system of interest and the availability of computational resources. These functionals are divided into four general categories: local density approximation (LDA) [76, 84], generalized gradient approximation (GGA) [85, 86, 87], meta-GGA [88], and hybrid functionals [89]. Some of the commonly used functionals are B3LYP [86, 89], B3P86 [90, 86], and PW91 [91]. The B3LYP method utilizes a three-parameter formulation (3) developed by Axel D. Becke (B) [85] combined with the correlation formula defined by Lee, Yang, and Parr (LYP) [86, 89]. The B3LYP is called a hybrid [92] functional since it combines Hartree-Fock exchange with one or more exchange and correlation functionals in a weighted fashion. In this work B3LYP was chosen because hybrid functionals represent the most frequently used approach, accounting for nearly 75% of all DFT functionals cited in recent literature over the past 20 years.

1.2 Quantum Theory of Atoms in Molecules (QTAIM)

The Quantum Theory of Atoms In Molecules was pioneered by Richard Bader and co-workers beginning in the early 1980s [93]. It continues to be developed and extended. For over a century, the transferability of both atomic and functional group properties has been experimentally confirmed [94]. In addition, the evidence that atoms and functional groups act similarly from one molecule to another molecule has been serving in the advancement of chemistry [95]. Indeed, this is the underlying basis for the utility of the Periodic Table and the organization of Organic Chemistry text books into chapters on the different functional groups. However, it is necessary to define and understand an atom **within a molecule** if we are to employ the Schrödinger equation or the quantum theories described thus far. The QTAIM provides a solution, partitioning a molecular system into constituent atoms, based on the electron density. A molecule can be divided into a set of atoms using the topology of the electron density. Atomic properties, such as energy, dipole moment and charge, can then be obtained by integrating their respective operators over the atomic volumes. The resulting atomic properties can be added up to get the value of that property for the whole system. The power of QTAIM is that it is possible to uniquely divide any electronic property into individual atomic contributions [93].

One of the fundamental postulates of quantum mechanics is that the wavefunction (ψ) is the central quantity, which contains all the dynamic information about a molecular system. Analysis of the wavefunction is not always simple due to the high dimensionality when treating a many-particle system. One property that can be obtained from the wavefunction is the electron density, which forms the fundamental basis of the theory of atoms in molecules in addition to its foundational role in DFT, the latter being limited to molecules and isolated atoms. QTAIM provides proper definitions for valuable chemical entities such as atoms, bonds, and functional groups. This theory is established around a quantum observable, the electronic charge density. It evolved as a theory of molecular and condensed phase electronic structure, in which inter-atomic surfaces are defined to be where the gradient of electron density radiates outward from BCPs. These are called zero-flux surfaces. Volumes defined by such surfaces are identified by properties that are well-defined and additive. For example, the energy of a molecule is given by the sum of energies of regions bounded by zero-flux surfaces. These distinct volumes partition the charge density of a molecule into space-filling regions, each of which typically surrounds a single nucleus, which is often called the “topological atom” , or simply “atom” .

QTAIM is a powerful tool in the prediction of molecular properties and reactivity based on their structures [62]. The central idea of the theory lies in the partitioning of electron density, which can be determined experimentally using X-ray diffraction or calculated using quantum mechanics [96]. The mathematical analysis of the electron density’s topology is an important part of QTAIM and it can be conveniently described using its gradient vector field and critical points [97]. The electron density, $\rho(\mathbf{r})$, and its critical points provide complete information of molecular structure and bonding [93]. The electron density can be analysed in terms of the stationary points in its gradient field, (i.e. minima, maxima and saddle points), which are called critical points. The gradient of the electron density is zero at such points.

The characterisation of a critical point in the electron density distribution is done by studying the Hessian matrix. The Hessian matrix is a 3x3 array of nine second derivatives of $\rho(\mathbf{r})$, which are evaluated at the critical point $\mathbf{r_c}$. The matrix is denoted by H($\mathbf{r_c}$) and written as:

$$\mathbf{H}(r_c) = \begin{pmatrix} \frac{\partial^2\rho}{\partial x^2} & \frac{\partial^2\rho}{\partial x\partial y} & \frac{\partial^2\rho}{\partial x\partial z} \\ \frac{\partial^2\rho}{\partial y\partial x} & \frac{\partial^2\rho}{\partial y^2} & \frac{\partial^2\rho}{\partial y\partial z} \\ \frac{\partial^2\rho}{\partial z\partial x} & \frac{\partial^2\rho}{\partial z\partial y} & \frac{\partial^2\rho}{\partial z^2} \end{pmatrix} \tag{3}$$

As the Hessian matrix is real and symmetric, it can be diagonalized to obtain the associated eigenvalues and eigenvectors. The procedure of finding the eigenvectors defines the axes of a new coordinate system where the off-diagonal elements of H($\mathbf{r_c}$) are all zero. We are only interested in critical points with non-zero eigenvalues. The associated matrix is denoted by:

$$\mathbf{H}(\mathbf{r_c}) = \begin{pmatrix} \frac{\partial^2 \rho}{\partial x^2} & 0 & 0 \\ 0 & \frac{\partial^2 \rho}{\partial y^2} & 0 \\ 0 & 0 & \frac{\partial^2 \rho}{\partial z^2} \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \tag{4}$$

where λ_1, λ_2, and λ_3 are the eigenvalues of the Hessian matrix.

The critical points are classified based on their rank (number of non-zero eigenvalues) and signature (sum of the signs of eigenvalues) of non-zero eigenvalues of the Hessian matrix evaluated at the critical point. The four types of stable critical points in ρ having three non-zero eigenvalues are:

- (3,-3) = Nuclear attractor (NA), 3 negative curvatures: ρ is a local maximum,
- (3,-1) = Bond critical point (BCP), a saddle point: ρ is a maximum in two dimensions and a minimum in one,
- (3, +3) = Cage critical point (CCP), ρ is a local minimum,
- (3, +1) = Ring critical point (RCP), a saddle point: ρ is a minimum in two dimensions and a maximum in one.

BCPs are topologically well-defined points that play an important role in describing the molecular structure. They establish the presence of bonding between the atoms via either sharing of valence electrons, ionic, or van der Waals (non-covalent) interactions [98]. Previous studies have shown that the characteristics of a bond, such as bond order, bond energy, and bond π character, can be obtained by studying several properties of the electron density at the BCP [99, 100, 101]. The two broad classes of bonds can also be classified based on the values of the Laplacian of the electron density at the BCP. For example, BCPs with relatively high $\rho(\mathbf{r_c})$ values and a negative Laplacian value indicate a shared (covalent) interaction, whereas BCPs with a relatively low $\rho(\mathbf{r_c})$ value, and a positive value indicate a closed-shell (ionic, or van der Waals) interaction [100].

In 1923 G. N. Lewis proposed a generalized theory of chemical bonding (an acid-base theory) based on the behavior of electron pairs [102]. According to Lewis and related models such as the Valence Shell Electron Pair Repulsion Model (VSEPR) [103, 104], there are bonding and nonbonding pairs of electrons in the valence shell of an atom. The electrons in the outermost shell of an atom are called valence electrons, and they are important to study when characterizing the chemical properties of an atom. But these electron pairs cannot be seen in the topology of the electron density. The desired properties can be obtained by studying the information-rich and derived function of the electron density, the Laplacian $\nabla^2\rho(\mathbf{r})$. In 1984, Bader, MacDougall, and Lau were the first to use topological properties of the Laplacian in relation to models of electronic structure and chemical reactivity. This led to a physical basis for the VSEPR model [105] as well as orbital models of chemical reactivity. The Laplacian of the electron density can be written as:

$$\nabla^2\rho = \frac{\partial^2\rho}{\partial x^2} + \frac{\partial^2\rho}{\partial y^2} + \frac{\partial^2\rho}{\partial z^2} \tag{5}$$

In general, the Laplacian of a function measures whether the function is locally concentrated or depleted, which can be easily proved via finite difference formulae. Thus, the Laplacian of a scalar field, such as the charge density, enables one to determine the regions where the field is locally concentrated ($\nabla^2\rho(\mathbf{r}) < 0$) and depleted ($\nabla^2\rho(\mathbf{r}) > 0$). Connections to the Lewis model of electron pairing was introduced into the QTAIM theory through this property. The Laplacian of the charge density also recovers the electronic shell structure of an isolated atom in terms of an alternating pairs of shells of charge concentration followed by charge depletion [106, 107, 98].

The outermost shell of an atom is called valence shell and it is divided into two regions. The inner region of the valence shell is the valence shell of charge concentration (VSCC), whereas the outer region is called the valence shell of charge depletion (VSCD). For a free atom, the VSCC contains a sphere over whose surface electronic charge is maximally and uniformly concentrated. This VSCC is distorted once an atom enters into chemical combination with other atoms. The distorted shell of charge concentration has maxima, minima, and saddle points. The maxima have been shown to correspond to the localized pairs of electrons, which are assumed in the Lewis model [98, 108, 109].

Lewis extended the Brønsted-Lowry acid-base theory [110] and defined an acid as an electron-pair acceptor (electrophile), and a base as an electron-pair donor (nucleophile). The Laplacian, defined in terms of charge concentrations and depletions, can also predict the sites of nucleophilic and electrophilic attack in a variety of systems [108]. These sites of attack in a molecule correlate respectively with the sites of maximum charge concentration and charge depletion in the VSCCs of reacting atoms. A local charge concentration corresponds to a nucleophilic reactive site (Lewis base), whereas a charge depletion corresponds to an electrophilic reactive site (Lewis acid). The Lewis model encompasses many types of chemical reactivity through the concept of acid-base reactions. The trajectory of Lewis acid-base reactions can also be predicted by aligning the local maxima (lumps) in the VSCC of a base with the local minima (holes) in the VSCC of an acid [111]. Figure 1 shows a sample VSCC of carbonyl carbon of acetone.

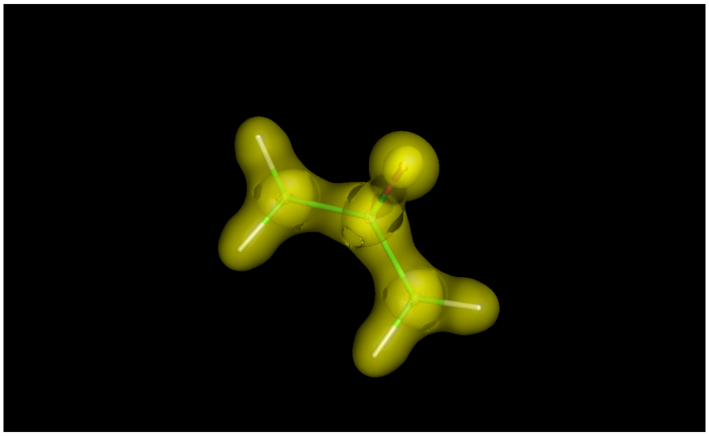


Figure 1: The VSCC of carbonyl carbon of acetone

During the late 1970s and 1980s, Burgi and Dunitz explained the principle of structure correlation [111]. They collected as many structures as possible containing the structural fragments of interest. Each structure provided a snapshot of the fragment in a specific environment, and these were then ordered into a sequence corresponding to a gradual deformation of the fragment corresponding roughly to evolution along a "reaction path". This method has been applied to map several reaction paths[111]. For example, from crystal structures containing an amino nitrogen (nucleophile) in the proximity of a carbonyl group (electrophile), they found that N approaches C=O at an average angle of 110° during a simulated nucleophilic addition reaction. These observations were interpreted as implying a preferred angle of approximately 110° for the approach of the nucleophilic nitrogen lone pair to the electrophilic carbonyl carbon.

The full topology of the Laplacian is usually studied in terms of critical points to reveal hidden features of electron density such as electron shells. As with BCPs, at each critical point $\nabla(\nabla^2\rho(\mathbf{r}))$=0, and they are defined by their rank and signature. The Laplacian also has four kinds of non-degenerate critical points of the rank 3: maxima (3, -3), minima (3, +3), and two types of saddle points (3, +1) and (3, -1). But now, since **negative** Laplacian indicates concentration, (3,+3) and (3,-1) CPs correspond to maximum concentration (lumps) and minimum concentration (holes) in the VSCC, respectively. The four types of stable critical points in the Laplacian having three non-zero eigenvalues are:

- (3, +3) = local minimum, bonded and non-bonded charge concentrations,
- (3, +1) = a saddle point, linking charge concentrations by unique pair of gradient paths,
- (3, -3) = local maximum, charge depletion, often seen in VSCDs, but absent in VSCCs.
- (3, -1) = a ring critical point, found at point of least charge concentration within a VSCC.

According to Bader *et al.* [105] (3, +1) critical points in $-\nabla^2\rho$ correspond to regions of charge deficit on carbon atom and these critical points corresponds to the centers of nucleophilic attack. They found that the carbonyl carbons in formaldehyde and acrolein contain (3, +1) critical points, which form angles of 111 and 109° with the C=O bond axis. Based on these topological features, they predicted that the carbonyl carbon is approached by a nucleophile at angle of approximately 110° with the C=O bond axis, in excellent agreement with the crystallographic studies of Burgi and Dunitz [111]. They also found the same pattern of critical points in the valence-shell charge concentration of acetaldehyde.

Advantages of Using QTAIM Descriptors

In the development of new drugs, it is argued that better modeling can be accomplished on the basis of well-defined physical descriptors [112]. The suggested descriptors should have simple physicochemical interpretations and provide valuable physicochemical insight. The descriptors derived from the topological analysis of the electron density which are experimentally accessible scalar fields, fundamentally related to all molecular ground-state properties. The electron density is an important property of atoms, molecules, and condensed phases of matter. On the basis of DFT, Hohenberg-Kohn proposed a theorem which is basically the reversal of Schrödinger's statement: the ground-state charge density maps to a unique ground-state wavefunction, from which one can recover the number of electrons in the system (N), as well as external potential ($V_{ext}(\mathbf{r})$), the potential of the interaction between the electrons and nuclei. Since the ground-state wavefunction is uniquely determined by the ground-state charge density, so are all observable properties of the system, including the ground-state molecular energy. This study demonstrates how few electron density descriptors can be used to predict spectroscopic properties and interaction energies.

The idea of molecular similarity permeates much of medicinal chemistry, particularly in the conceptualization of new drug molecules[113]. The observations laid out earlier raise the question of whether molecular similarity can be evaluated objectively. Similarity is a relational concept that can only be formulated for a given quality or property, such as molecular geometry, molecular weight and empirical formula. There is no shortage of molecular similarity studies applied to various properties, whether amino acid sequences, chemical 2D graphs, geometric superimpositions, or molecular electron density superimpositions. But what approach captures the maximum extent of molecular similarity, and why? Is the response to this inquiry is a function of the issue or is it universal? It is argued in this work that the appropriate response is unique: the molecular electronic density captures and determines all the properties of the molecule. In practice, however, the functional relationship between the density and several properties is frequently unknown or known only approximately. The properties derived from the topology of the electron density hold much guarantee, as their choice eliminates a significant source of modeling uncertainty. The properties derived from the electron density can produce strong correlations that can be used to predict the properties of unknown compounds. At the same time, valuable physicochemical insight is shown by these density-derived properties. The charge density can also be obtained experimentally [114], and as evidenced by the success of DFT, it contains necessary information embedded in it.

In general, it is possible to obtain valuable information not only from the properties of a molecule as a whole, but also from the properties of an atom within that molecule [115]. This information helps researchers better understand the specific role an atom plays in that molecule’s chemistry. The concept of studying the properties of a molecule based on the properties of individual atoms within that molecule is one of the cornerstones of the chemistry [116]. There are many different methods used to obtain the properties of a part of a molecule, which often produce conflicting results. The reason is that they are based on methods that apply arbitrary conditions to the partitioning of the molecular wavefunction.

Researchers have no way to test the validity of the results obtained when using such arbitrary methods. Fortunately, QTAIM provides new insights into the chemistry of atoms by looking from a different perspective. This approach allows one to partition a molecular property into its atomic contributions in a non-arbitrary manner, based on the physics of an open system [93]. These principles can then be utilized by researchers to interpret observed chemical behavior as a function of the individual atomic contributions.

With the ever-increasing quality of experimental data, more and finer details of the electron density are observed using the most widely used Stewart–Hansen–Coppens multipole model [117, 118]. The model also provides the means for observing a high degree of atom transferability in similar chemical environments and for building pseudoatom databases. Due to the transferability of atomic fragments, excellent agreement with the measured density can be achieved by simple addition of fragment densities from data bases such as Koritsanszky *et al.* [119].

1.3 Introduction to Aldehydes and Ketones

The carbonyl group (C=O) is among the most important functional groups in organic chemistry due to its ubiquitous presence in natural products and its versatile reactivity [120]. If the carbonyl group is united with only hydrogens or carbon atoms, then the compounds are known as aldehydes or ketones (Figure 2). In aldehydes, at least one hydrogen atom is bonded to the carbonyl group and these are shown by the general formula RCHO. In ketones, two carbon atoms bond to the carbonyl group and these are shown by the formula RCOR’ (either or both R or R’ may be aliphatic or aromatic) [121].

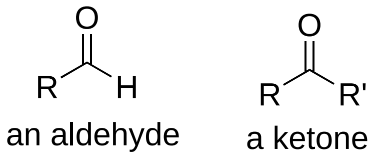


Figure 2: Structure of an aldehyde and ketone

Aldehydes and ketones have great importance in both biological chemistry (they are found in proteins, carbohydrates, starch, and DNA) [122] and in synthetic organic chemistry [123]. Carbonyl compounds contain a polarized reactive carbon-oxygen double bond which creates a partial positive charge on the carbonyl carbon atom and a partial negative charge on the oxygen atom. Due to this electron rich doubly bonded C=O group, the molecule is able to participate in a variety of chemical reactions such as oxidation, reduction, condensation, and addition reactions.

The carbonyl group acquires electrophilicity from both the resonance and inductive effects [120]. One of the resonance structures in Figure 3 show a positive charge on the carbon atom, indicating that the carbon atom is electron deficient. The inductive effect is an electronic effect due to the polarization of sigma bonds within a molecule, which in turn is due to the difference in electronegativity between the atoms. This effect also shows that the carbon atom is deficient in electron density. As a result, this carbon atom is electrophilic in nature and is susceptible to attack by a nucleophile (Nu). Figure 3 depicts all three effects.

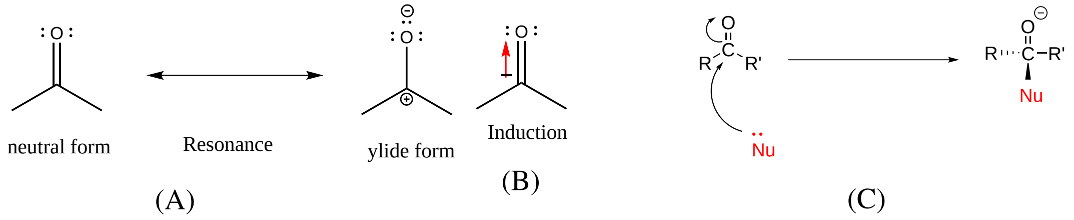


Figure 3: (A) Resonance effect, (B) Inductive effect, explicitly showing the bond polarity with partial charges, (C) Nucleophilic addition

The most important bioorganic reactions of carbohydrates, peptides, carboxylic acid esters and anhydrides, acetals and hemiacetals, ketals and hemiketals, proteins, and lipids involve nucleophilic addition to the C=O group [124]. In these reactions, the nucleophile attacks positively polarized carbon atoms of the C=O group at an angle of approximately 110° to the plane of the carbonyl group and transforms sp^2 hybridized C into sp^3 C by forming a tetrahedral carbon complex as product, transition-state, or intermediate [125]. As a stretch test of our proposed ML model for predicting interaction energies, a nucleophilic addition reaction of a carbon nucleophile (ketone) to a carbon electrophile (aldehyde), catalyzed by an enzyme, is described later in the dissertation.

After analysing high-resolution crystal structures of small molecules, Bürgi, Dunitz and Shefter [111] noted that the interaction occurs between nucleophiles and carbonyl groups with a preference for attacking angles ranging from 102° to 114°. Their work focused on studying the addition reaction pathways of a nucleophile (O or N) to a carbonyl group. Figure 4 demonstrates this angle of nucleophilic attack. The probability of addition will be influenced by the angle of nucleophilic approach during an intermolecular interaction. In general, aldehydes are more reactive compared to ketones toward nucleophilic attack, which can be explained in terms of both steric and electronic effects.

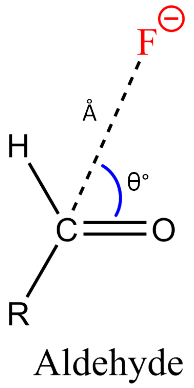


Figure 4: The angle of nucleophilic attack on carbonyl group

Steric effects. Ketones have two alkyl groups which contribute to steric hindrance in the transition state. Whereas in aldehydes, the transition state is less crowded because of the presence of only one alkyl group, and therefore has a lower energy barrier to addition.

Electronic effects. Alkyl groups are electron-donating groups. Ketones contain two alkyl groups and thus can stabilize the ylide resonant structure's positive charge on the carbon atom of the carbonyl group. Whereas in aldehydes, the positive charge is less stabilized compared to ketones since there is only one electron donating group.

Chemical Reactivity

The first theorem of Density Functional Theory (DFT) [76] states that "all properties of all states are formally determined by the ground state charge density of a system" . Hence it should be possible to locate where the electrophilic or nucleophilic attack will occur in a molecule merely based on the ground state charge density distribution. Chemical reactivity can also be studied based on the properties of the electronic charge distributions. We have already briefly discussed how the Lewis and VSPER models were connected to the properties of the local charge concentrations of the Laplacian distribution [108]. However, the Lewis model encompasses chemical reactivity as well, through the concept of acid-base reactions [126]. The positions of local charge concentrations and depletions that correspond to sites of nucleophilic and electrophilic reactive sites, respectively, are identified as critical points in the VSCC of the base and the acid. The alignment of a charge concentration on the base with a charge depletion on the acid predicts the geometry of approach of the reactants. This also enables one to predict the angle of nucleophilic attack. The location of holes in the VSCC of a carbonyl carbon determines the position of nucleophilic attack at this atom [108]. These geometric predictions are consistent with our studies of interaction energies discussed later.

1.4 Spectroscopic Properties of Aldehydes and Ketones

Aldehydes and ketones have unique IR (infrared) and NMR (nuclear magnetic resonance) spectral properties. Sir William Herschel was one of the first scientists who observed infrared radiation in the early 19^{th} century [127]. However, it was only in the early 20^{th} century when chemists started taking advantage of this technique [128]. A common example of a molecule with a carbonyl group is acetone, whose IR spectrum is shown in Figure 5. The carbonyl stretching frequency is at 1716 cm^{-1}, and is labeled as A in the figure.

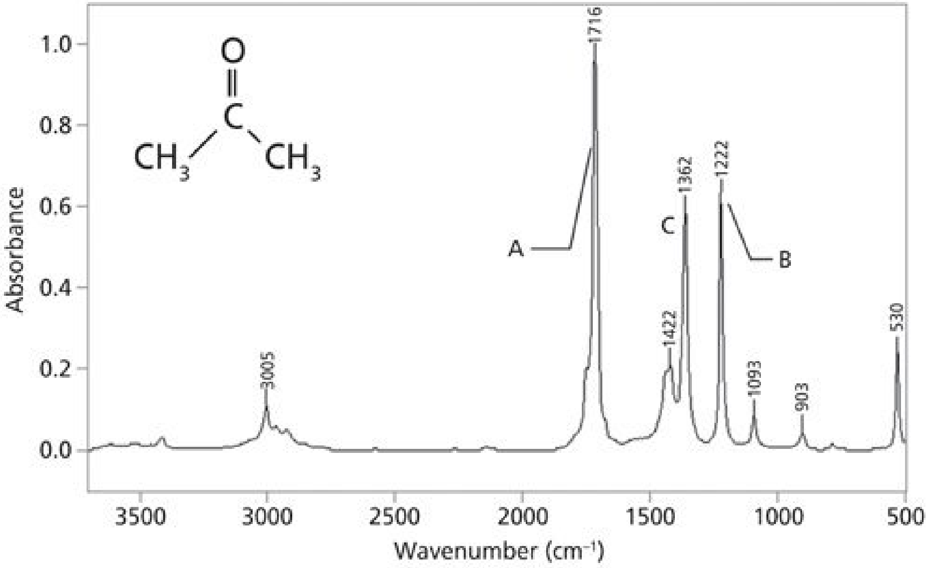


Figure 5: The infrared spectrum of acetone

1.4.1 IR Spectroscopy

IR spectroscopy is one of the most widely used spectroscopic techniques for the characterization of materials in different areas of research [129]. It became one of the crucial interrogative tools in chemistry, physics, biology, and material sciences, whenever structural characterization is of prime importance [130]. An interesting quality of IR spectroscopy is ability to study samples in any state, such as powders, crystals, semicrystals, liquids, pastes, films, fibers, and gases [131].

IR spectroscopy is a widely used method in analytical chemistry for structural characterization of molecules [132]. In many applications of IR spectroscopy, not only qualitative but quantitative analysis is also needed, for example to determine the concentration of molecular species. It has a wide range of applications from analyzing small molecules to complex samples in various fields including pharmaceutical [133], food [134], textile industries [135], biomedical research [136], forensic sciences [137], and disease detection [138, 139]. It is based on the vibrations of the bonds in a molecule. IR spectra are considered as molecular fingerprints since each molecule produces a unique infrared spectrum. A characteristic absorbance frequency is observed for specific functional groups. Carbonyl groups usually appear at about 1700 cm^{-1}.

The exact position of the peak depends on the chemical environment of the C=O functional group. The fluctuation of peak position, called a frequency shift [129], occurs due to the gradual change in the vibrational frequencies of a chemical bond. For example, conjugation of carbonyl group with more electronegative substituents tends to increase the vibrational frequency to a higher energy value, a blue shift. Whereas substituents which donate π electron density into the carbonyl group tend to lower the vibrational frequency to a lower energy value, a red shift. Peak assignments of an IR spectrum to a specific carbonyl functional group is not an easy task. Thus, IR spectroscopy must be combined with other spectroscopic methods such as NMR spectroscopy (which we will also discuss in Section 1.4.2) to identify the differences between carbonyl functional groups [140].

The vibrational modes of a molecule are detected experimentally using infrared spectroscopy. This helps in determining the molecular structure and environment. To acquire such valuable information, it is essential to find out what vibrational motion belongs to each peak in the spectrum. This process of assigning peaks is quite difficult because of the vast number of closely spaced peaks observed even in simple molecules. To help this process, theoretical calculations of vibrational frequencies are required, and these are mostly computed from quantum-mechanical calculations using the harmonic oscillator approximation [141]. The harmonic oscillator approximation is extensively used for calculating molecular vibrational frequencies because more accurate methods need very large amounts of computational time.

Fundamentally, the harmonic vibrational frequencies ignore the effects of vibrational anharmonicity, due to which the calculated frequencies tend to be 10% larger than the experimental frequencies [142]. One possible solution to this problem can be using a scaling factor [143], which brings theory into better agreement with experiment. Numerous solutions have been recommended to adjust the calculated vibrational frequencies for better agreement with experiment [144, 145]. These consist of rescaling all the computed frequencies with a single scale factor [145], rescaling the high and low frequencies individually with different scale factors [146], and rescaling the appropriate force constants in the Hessian matrix [147]. Although scaling factors improve agreement in the high-frequency region, they lead to higher differences between calculated and experiment frequencies in the lower frequency region. Therefore, in our study we have chosen to use the uncorrected harmonic frequencies from the *ab initio* calculations.

1.4.2 ^{13}C NMR Spectroscopy

The first NMR experiment was done by Rabi in early 1937 [148], and the first application of NMR in bulk materials were performed, respectively, by the Purcell group on paraffin [149]. Since then, ^{13}C NMR spectroscopy has become an indispensable tool for all areas of chemistry. The important role of this technique arises from the valuable information that can be extracted, which spans both structure and dynamics [150]. In this technique, the molecules are identified and characterized based on the chemical shifts in the frequency of radiation emitted during transitions between spin states of the nuclei present in the compound [151]. The basic information that one can obtain from NMR spectroscopy are chemical shifts, coupling constants, and relaxation rates [152]. The chemical environment of a given nucleus can be determined by finding the values of chemical shifts and coupling constants to other peaks. The comparison of computed chemical shifts with experimentally-determined chemical shifts provides valuable information in deducing the correct structure of an unknown compound [153]. In recent years, the theoretical calculation of NMR properties significantly enhanced the experimental work [154]. Similarly, comparison of a synthesized compound's ^{13}C NMR spectrum with a theoretically calculated spectrum can reveal the success of the synthesis in synthetic organic chemistry [155].

In NMR spectroscopy, we use a quantity called the chemical shift, which is the change of nuclear shielding of a target nucleus with respect to a reference nucleus [156]. For example, tetramethylsilane (TMS) is used as a reference molecule in both ^{1}H and ^{13}C NMR studies. In this case, the frequencies of nuclei in a spectrometer are obtained by comparing and normalizing the frequencies against the frequency of TMS in the spectrometer [157]. Chemical shifts are determined and expressed relative to the reference molecule in the dimensionless unit of parts-per-million (ppm). Normally, ^{13}C NMR spectra are recorded across the range of 0-210 ppm. The resonance of the single type of carbon-13 nucleus in TMS appears at 0.0 ppm, which is used as the reference standard. The environment of the resonating nucleus influences the frequency of the resonance. For example, electron withdrawing groups move chemical shifts to higher frequency values (downfield shift), whereas electron donation groups to lower frequency values (upfield shift) [157]. In ^{13}C NMR spectroscopy, carbonyl carbons of both aldehydes and ketones have characteristic frequency shifts in the range of 190-200 ppm and 205-220 ppm, respectively.

In general, carbon chemical shifts can be estimated from large databases of compounds with known chemical shifts [158]. For obtaining chemical shifts of a small molecule, there are a few options that can be considered. These options include: finding a similar compound with known chemical shifts from the database; deriving rules from chemical shifts of known compounds in the database with varying substituents; and training machine-learning methods using known chemical shifts.

Predicted chemical spectra can be used in assisting the structural elucidation of an unknown compound [159]. In a similar way, these spectra can be used to reduce the number of structure verification experiments by filtering possible stereoisomers [160]. Another application is the calculation of chemical properties such as LogP, which is a component of Lipinski's Rule of 5 used to predict drug-likeness of a compound based on predicted chemical shifts [161]. In 1984, Kalchhauser and Robien introduced a computer program called CSEARCH for the analysis of ^{13}C NMR spectra [162]. This program predicts and automatically assigns carbon chemical shifts. A database containing 8000 spectra was created from the literature. Here chemical shifts are predicted based on the HOSE (Hierarchical Organisation of Spherical Environments) code approach [163]. In this approach, the program starts at the carbon atom whose chemical shift is to be predicted, looks one bond away from this carbon and tries to search for this environment in the database. If this is successful, then it moves to further atoms until it reaches the boundary of the molecule.

Satoh *et al.* [164] introduced a ^{13}C NMR chemical shift prediction system CAST/CNMR. This system is based on a database containing 733 compounds and their three-dimensional structural information, along with NMR chemical shift data. Chemical shifts are predicted by comparing the structural information of a query molecule within the database. In the prediction procedure, CAST/CNMR searches the database and finds molecules having similar partial structures around the carbon of interest in the query structure. The predicted shift is the average of all the ^{13}C NMR shift values of the hit molecules.

Previously Used Methods in Predicting Chemical Shifts

Some widely used methods for predicting chemical shifts include empirical methods, *ab initio* calculations, and machine-learning methods such as artificial neural networks (ANNs).

Ab initio methods

Electronic structure calculations of magnetic behavior in molecules are present in the literature dating back to 1937 [165]. In the early 1950s, Ramsey presented a series of pioneering papers delineating equations used to compute NMR parameters [166]. Over the following years, several methods to calculate NMR parameters evolved within the quantum chemistry community which were of great interest to experimentalists [167, 168]. In general, two major approaches are used for the calculations, which include wavefunction-based methods and Density Functional Theory-based methods. *Ab initio* calculations determine ^{13}C NMR chemical shifts by computing magnetic properties of a given substance. The gauge-including atomic orbital (GIAO) [154] method is one of the commonly used approaches for calculating nuclear magnetic shielding tensors. It has been proven in many instances that the results obtained by using GIAO were more accurate compared to the ones calculated with other approaches [169]. DFT methods usually produce good results at relatively low computational cost [170]. Due to this, DFT methods have been used in the study of metal complexes [171], large organic compounds [172], and organometallic compounds [173], where *ab initio* methods are cost-prohibitive.

In 2007, Bagno and Saielli [152] summarized computational work using the Amsterdam Density Functional (ADF) suite [174] and Gaussian 03 [175] to find chemical shifts of different chemical elements in various compounds. In 2002, Giampaolo *et al.* [176] performed Hartree-Fock calculations of ^{13}C NMR chemical shift of low-polarity compounds. These calculations were used as a tool to support the structural interpretation of NMR data of low-polarity natural products. In their method, GIAO-calculated chemical shift values of optimized structures were used to draw linear correlation plots of calculated versus experimental data. They obtained linear correlation coefficient (r) of around 0.995. Another example is given by Cimino *et al.* [177], who investigated the application of different quantum chemistry approaches and basis sets in calculating the ^{13}C NMR chemical shifts of 15 low-polarity natural products. The corrected mean absolute error (CMAE) for about 50 different chemical shift calculations was reported to range from 1.49 ppm to 3.35 ppm.

Database methods

Empirical methods relating atomic structural descriptors to ^{13}C NMR chemical shifts have been used to accurately predict the ^{13}C NMR spectra for compounds whose chemical shifts are not known. These approaches rely on large data sets of known compounds with assigned chemical shifts. The advancement in modern computer systems makes it possible to store and search among large numbers of chemical structures along with their chemical shifts. Examples include Spectral Database for Organic Compounds [178] ($\approx 130,000$ ^{13}C chemical shifts), CSEARCH database [179] ($\approx 4,000,000$ ^{13}C chemical shifts), ACD/CNMR ($\approx 2,160,000$ ^{13}C chemical shifts), NMRShiftDB [180] ($\approx 200,000$ ^{13}C chemical shifts), and BIORAD KnowItAll database ($\approx 3,500,000$ ^{13}C chemical shifts). To predict the chemical shift for each carbon atom, the database is searched for structures containing carbon atoms with a similar chemical environment. The most challenging part in this method is to find the best way of encoding the chemical environment of an atom that can be easily searched. Examples include the Hierarchically Ordered Spherical description of Environment code (HOSE) [163] and SMILES [181] code. However, there are some drawbacks in using each of these methods, which include the size of the database and its access time. The storage space needs to be increased with increasing number of molecules represented, which in turn increases the access time.

Machine-Learning Methods

Previous studies have shown that machine-learning approaches, such as artificial neural networks (ANNs), have the capability to predict ^{13}C NMR chemical shifts with impressive accuracy [182, 183]. ANNs have been used to predict chemical shifts of several different classes of compounds such as alkanes [184], acrylonitrile co-polymers [185], trisaccharides [186], and many more small organic molecules [187].

1.5 Interaction Energies

It is found that there are four kinds of interactions in nature, which include strong interactions, weak interactions, electromagnetic interactions, and gravitational forces. The strong and weak interactions occur because of short-range forces, which can be seen between protons, neutrons and other fundamental particles. Gravitational forces are present with all mass systems. According to general theory of relativity, this interaction emerges from the distortion of space. The electromagnetic interactions occur between atomic and sub-atomic systems, which results in the formation of atoms and molecules.

Among these four interactions, only electromagnetic interactions are essentially important to molecular systems, in which the interaction range of strong and weak forces is very short ($<10^{-5}$ nm) and gravitational forces are very weak. The formation of covalent and noncovalent bonds in chemistry are due to both classical and quantum-mechanical electromagnetic interactions. According to molecular quantum mechanics, covalent chemical bonding between a pair of interacting atoms occur due to the overlap of partially filled occupied orbitals. These interactions were first illustrated by Heitler and London in 1927 [188].

There is another type of interaction between molecules, which results in the formation of molecular complexes. Because there are no breaking or making of covalent bonds in formation of these complexes, these are called as noncovalent interactions, or van der Waals interactions [189]. These interactions are generally found in molecular clusters and biomolecular systems which play an important role in processes such as phase changes, protein folding, molecular recognition, and enzyme-substrate binding. These interactions are usually weak in comparison to covalent interactions. For large molecules, they can be intramolecular as well as intermolecular. However, in comparison to covalent interactions, and because of their weakness, the noncovalent intramolecular interactions are difficult to model accurately. Therefore, it is important to study new methods of modeling noncovalent intermolecular interactions, which have importance in many fields of chemistry and physics. Noncovalent intermolecular interactions are classified into four kinds: electrostatic, induction, dispersion, and exchange. The first one is originated from interaction between two permanent multipoles (electrostatic), the second one from interaction between a permanent multipole and an induced multipole (induction), the third one from interaction between instantaneous multipoles (dispersion), and the last one from the overlap of occupied orbitals (exchange). The sum of these four different intermolecular energies yields the total intermolecular interaction energy.

The calculation of interaction energies between proteins and ligands in drug design is very important and is used to find the strength of drug binding between, for example, a protein and a ligand. There are several different methods available in calculating these energies. However, some of these methods are extremely expensive and time-consuming. There are 3 primary methods for calculating interaction energies: empirical force-field methods, semi-empirical methods, and *ab initio* methods. Among these, the most accurate method and also the most time-consuming is *ab initio*, based on solving the Schrödinger equation. These interactions can be studied in two different approaches in *ab initio* schemes: the supermolecular method and the perturbation method. In the perturbation method, the interaction energy is evaluated using perturbation theory which treats the interaction between subsystem wavefunctions as perturbations on each other. In the supermolecular method, the interaction energy is obtained from the difference between the energies of the complex and the total energy of the isolated molecules. Among these two methods, supermolecular method is the most widely used for calculating interaction energies. As such, we have used the supermolecular method to calculate interaction energies in our study.

1.6 Artificial Neural Networks (ANN)

ANNs are inspired by biological nervous systems [190] found in animals as shown in Figure 6. The algorithms try to mimic the brain and are programmed to function like biological neural systems [191]. Biological neurons consist of a cell body which contains a nucleus that governs the cell activity. To the left of cell 1, many fine threads called dendrites or receivers are shown in Figure 6. These provide input signals to the cell. To the right in Figure 6, one longer thread called an axon or transmitter is shown. This carries the output signals to cell 2. Impulses can be transmitted unchanged or modified by synapses. A synapse, which is the junction between the neurons, can change the strength of the connection between neurons and cause excitation or inhibition of another neuron. The result is an intelligent brain that possesses capabilities of learning, prediction, and recognition.

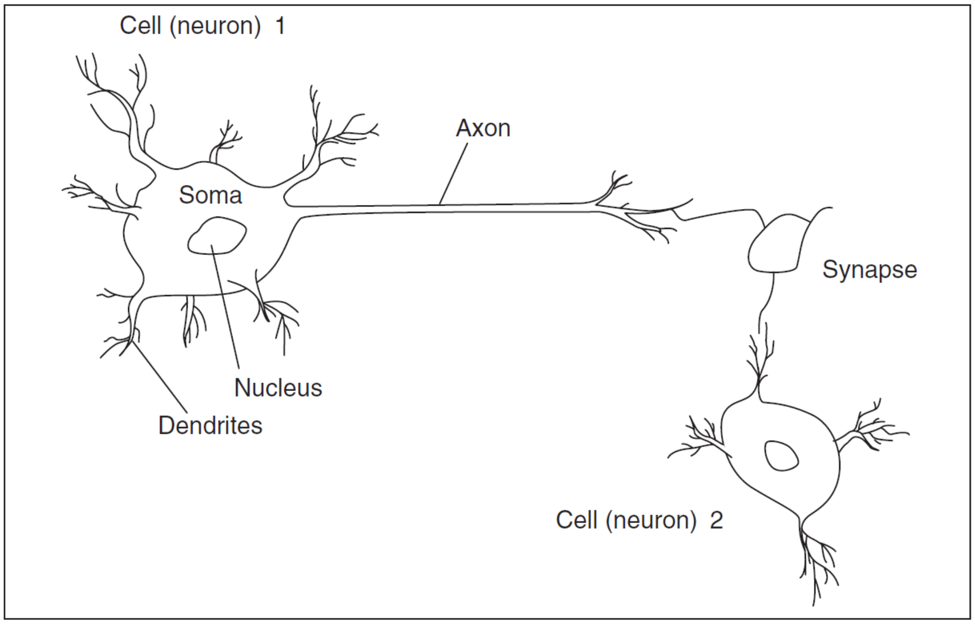


Figure 6: Representation of a natural neuron

It is important to remember that neurons in the ANN are an abstract representation of biological neurons. ANN consists of inputs for the neuron, associate weights for the inputs, transfer functions for the neuron, and the output of the neuron (Figure 7).

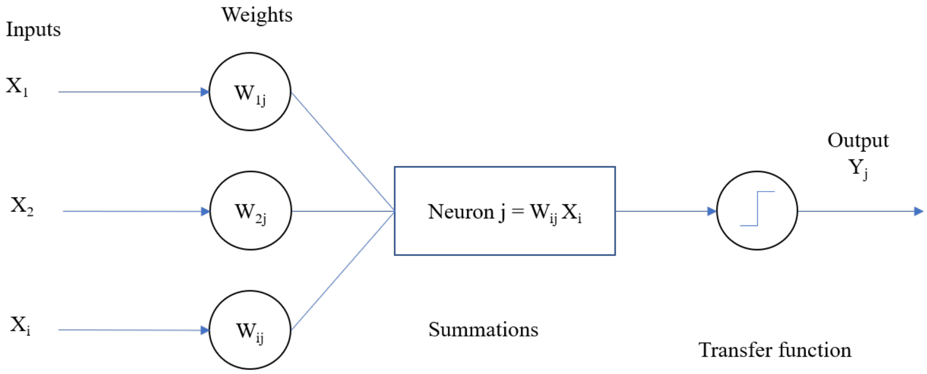


Figure 7: Representation of an artificial neuron

A collection of these neurons forms an artificial neural network as shown in Figure 8. Over the past decade, ANNs have had huge success in machine-learning and data-mining applications [192]. Some examples of applications include medical diagnoses, risk evaluation of insurance and loan applicants, image classification, and predicting the structure of chemical compounds [193].

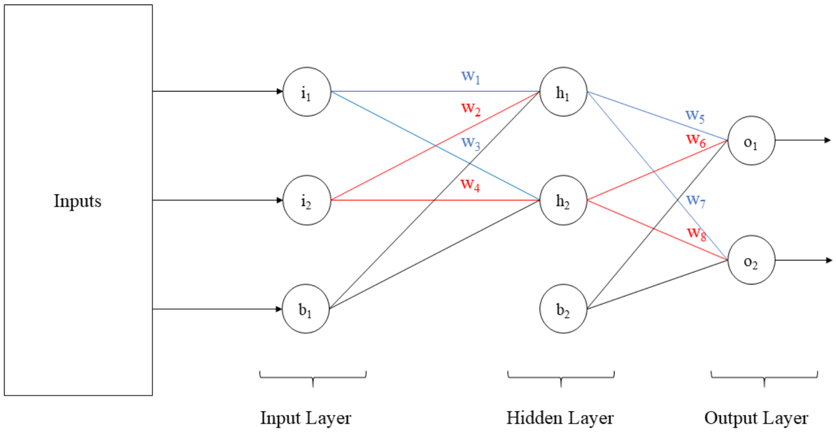


Figure 8: Representation of an artificial neural network

The most important property of ANNs is the ability of a network to “learn” from its environment and improve its performance. As a result, the method is often referred to as Machine-Learning (ML) [194]. The ANN consists of an input layer, one or more hidden layers, and an output layer. The input signal propagates layer-by-layer in the forward direction and these networks are commonly called multilayer perceptrons (MLP) [194]. MLP with the back-propagation learning method is one of the most successfully used methods in chemistry and drug design because of its well-defined and explicit set of equations for weight corrections. This is a supervised learning algorithm in which the network is trained with training data, and where the expected outputs are provided to train the algorithm [194]. The learning consists of both a forward pass and a backward pass. In the forward pass, the input vector is applied to the input layer and these input values are modified by a fixed weight, and its effect passes through the network layer-by-layer. Finally, output produced by the network is compared with the desired output to calculate the error signal. This error signal is then back-propagated in the backward pass to adjust the weights in such a way that the actual output value moves closer to the desired output value according to an error correction rule. One complete forward and backward pass through the network is called an epoch.

The following section describes the back-propagation algorithm. Figure 8 demonstrates signal propagating through the network, where symbols i_1 and i_2 represent input neurons, h_1 and h_2 represent hidden neurons, o_1 and o_2 represent output neurons, and $w_1 - w_8$ represent weights of connections between layers. Primarily, these weights are initialized with random values. In the forward pass, the total net input to each hidden layer neuron (net_{h_1} and net_{h_2}) and activation function (out_{h_1} and out_{h_2}) are calculated by formulae in Equations 6 through 9. Here we show calculations for only one neuron [195].

$$net_{h_1} = w_1 \times i_1 + w_2 \times i_2 + b_1 \tag{6}$$

$$out_{h_1} = \frac{1}{1+e^{-net_{h_1}}} \tag{7}$$

This process will be repeated for output layer neurons, using outputs obtained from hidden layer neurons as inputs. The output for o_1 neuron is calculated by [195],

$$net_{o_1} = w_5 \times out_{h_1} + w_6 \times out_{h_2} + b_2 \tag{8}$$

$$out_{o_1} = \frac{1}{1+e^{-net_{o_1}}} \tag{9}$$

We can now calculate the error for each output neuron using Equation 10 [195]. Once the output is calculated from all the output neurons, then the total error will be calculated using Equation 11 [195],

$$E_{o_1} = \frac{1}{2}(target_{o_1} - out_{o_1})^2 \tag{10}$$

$$E_{total} = E_{o_1} + E_{o_2} \tag{11}$$

In the backward pass, all the weights are updated to minimize the error for each output neuron. By applying the chain rule, we can calculate how the change in w_5 will affect the total error by doing partial differentiation of total error with respect to w_5 [195].

$$\frac{\partial E_{total}}{\partial w_5} = \frac{\partial E_{total}}{\partial out_{o_1}} \times \frac{\partial out_{o_1}}{\partial net_{o_1}} \times \frac{\partial net_{o_1}}{\partial w_5} \tag{12}$$

Alternatively, we have $\frac{\partial E_{total}}{\partial out_{o_1}}$ and $\frac{\partial out_{o_1}}{\partial net_{o_1}}$ which can be written as $\frac{\partial E_{total}}{\partial net_{o_1}}$, also known as δo_1. Substituting these in Equation 12 yields

$$\frac{\partial E_{total}}{\partial w_5} = \delta o_1 \times \frac{\partial net_{o_1}}{\partial w_5} \tag{13}$$

To reduce the error, the value of $\frac{\partial E_{total}}{\partial w_5}$ is subtracted from the current weight (w_5). The updated weight for w_5 is w^{+5}, which is then calculated by [196],

$$w_5^+ = w_5 - \eta \times \frac{\partial E_{total}}{\partial w_5} \tag{14}$$

In equation 14, η is the learning-rate parameter [195] and its value ranges between 0 and 1. This parameter controls the magnitude of the changes applied to the weights. The above-mentioned procedure will be followed to calculate the total net input to each hidden layer neuron, activation functions, and errors for each output neuron, and the weights will be adjusted by repeating equations 6 through 14 until all the neurons get updated weights.

The ANN generates a model which depends on the architecture, that is, the number of layers, the number of neurons in each layer, and the way neurons are connected. The topology or architecture of a network is related to the number of neurons in the hidden layer. It has a significant effect on the prediction accuracy and hence should be optimized. The performance of the ANNs depends on several parameters, which include learning rate (η), momentum (m), number of epochs (N), number of hidden layers (HL), and number of hidden neurons (H) [195]. These parameters are defined and discussed below.

Learning Rate

The learning rate controls the magnitude of the changes made to the weights in each iteration of training and its value ranges between 0 and 1. This parameter needs to be carefully tuned. If the learning rate is too large, there is a chance of overshooting a good answer, and then on the next iteration it will be undershooting and get into an oscillating pattern where training never converges. If the learning rate is small, then the training is slow, which causes a smaller change in the weights and the optimization takes longer. The momentum parameter finds the amount of influence from the previous iteration on the present one and its purpose also to speed up the training process with the reduced risk of oscillating.

Epoch

During the process of training, many examples of relevant input/output combinations are presented to the network. Each example consisting of a pair with the input values and the corresponding target output values. The patterns are sequentially presented to the network in an iterative manner; the weights being updated during the process to adapt the network to the required behavior. This process of iteration continues until the connection weight values allow the network to perform the desired mapping. Each presentation of the whole pattern set is called as an epoch.

1.7 Computational Chemistry in Drug Design

The maintenance and further advancement of a strong and powerful drug discovery pipeline is important for fighting against new diseases. The process of discovering and developing new therapeutic solutions is extremely complex, expensive, and time consuming. The typical discovery and development of a drug, from lead identification to clinical trials, can take approximately 14-15 years of time with a total cost of 2.6 billion US dollars [197].

The human body is complex chemical machinery since it contains thousands of chemicals, from many classes of compounds, such as carbohydrates, proteins, fats, etc.., all of which can undergo many possible chemical changes. These reactions also interact with each other in numerous systems. Every process in the body is a kind of chemical transformation that leads to pain, movements, thought processes, feelings, and many more complex and simple changes. The human body has also been equipped with all the essential chemical components, numerous enzymes and neurotransmitters for the balanced and appropriate working of all the life-maintaining processes. Still, sometimes bioprocesses fail to act due to various exogenous or endogenous factors. Thus, administering external aids, which we call drugs or medicines, becomes necessary to restore the normal functioning. Due to recent developments, such as combinatorial chemistry and high-throughput screening technology, drug developers have created an environment to expedite the drug discovery process by allowing vast libraries of compounds to be screened and synthesized in a shorter amount of time [198, 199]. However, because of low efficiency and high failure rates in drug discovery, several different approaches are usually pursued simultaneously. Computer-aided drug design (CADD) is among the augmented approaches for reaching medicinal goals.

Drug design research has gained significant insights from computational approaches

[200]. The drug discovery process involves several different steps in order for a drug to reach the market, and these are discussed in the later sections. These steps include target identification and validation, lead identification and optimization, and finally pre-clinical and clinical development. In a drug discovery pipeline, CADD is used to filter-out small sets of active compounds from large compound libraries. It is used to guide the optimization of lead compounds and also to design novel compounds. CADD can be classified into two categories; namely structure-based and ligand-based methods.

1.7.1 Structure-Based Computer Aided Drug Design (SBDD)

The growing availability of structural information and inexpensive high-performance computing platforms have expanded the applicability of SBDD methods and provided a more rational initial point for the long process of drug discovery and development. This process is iterative and advances through multiple cycles before an optimized ligand enters Phase I clinical trials. In general, these methods are defined by the application of computational algorithms in combination with experimental data, either to evaluate the binding in terms of affinity, or to design novel molecules that are anticipated to bind the target molecule with specificity and affinity. Successes have been reported for many compounds that went through clinical trials and received FDA approval to enter the market [201]. In the early 1990s one of the first HIV-1 protease targeted drugs, Saquinavir, was developed using SBDD methods [202]. Amprenavir is another HIV-1 protease targeted drug, which was also developed using SBDD [203].

Target Identification and Validation

The first step in the drug discovery process is the identification of drug targets which can be enzymes, ion channels, various types of receptors, transporters or various other targets. A drug shows its therapeutic action when it binds to its biological targets, called receptors. Receptors are often proteins which contain an active site, or binding site, where the binding of a drug molecule (ligand) occurs. In order to design a good ligand, it is important and challenging to know the atomic-level structure of the receptor and identify active sites of the target protein. Target identification can be achieved using genomic and proteomic approaches, which are considered as laborious and time-consuming [204]. To overcome this problem, computational methods have been developed as a useful alternative. A detailed computerized 3D model of a protein's structure enables one to extensively study the structure and dynamics of its potential drug receptors.

Structure-based drug design depends on the ability to determine and analyze the 3D structures of target proteins, which can be obtained experimentally through X-ray crystallography or NMR techniques. After the structure is determined, it is often saved (entered) to a public database. The Protein Data Bank (PDB) [205] and The Cambridge Structural Database (CSD) [206] are among the most commonly used databases for storing and obtaining protein structures that can be used for docking studies. PDB contains approximately 120,000 protein structures determined mainly using X-ray crystallography, but also some using NMR spectroscopy. X-ray crystallographic study is possible only when the target protein can be crystallized. However, some proteins, such as membrane proteins, are not easily crystallized [207]. Thus, experimental methods are not always successful in finding their structures. In these circumstances, computational methods play a prominent role [208, 209]. These are based on comparative modeling of target proteins in which target structure is predicted based on a template with a similar sequence; i.e., proteins with nearly identical sequences have very similar structures. There are various programs and web servers available that automate the comparative modeling process; e.g., PSIPRED [210] and MODELER [211].

In the absence of experimentally determined structures, the potential drug target's binding site and molecular function could be obtained via structural comparison with a well-characterized protein for which the structure and biochemical function are known. Sequence comparison is one of the existing methods used to find the alignments between a nucleotide or protein sequence against a database containing experimental 3D structures of known protein sequences. Once the homologous (similar sequence) protein is identified, then the 3D model of the target protein will be constructed. There are several computer programs available that perform the process of comparative modeling automatically, providing a foundation for drug design by structure. Homology modeling, *ab initio* folding, and threading approaches are a few of the methods used for protein structure predictions [207].

Once a 3D structure of the protein is obtained, then the next important step in SBDD is to find the binding pocket on that protein. The binding site information is often obtained from co-crystal structures of the target protein or a closely related protein with a bound ligand. If the co-crystal structure is not available, then computational methods like POCKET, SURFNET, Q-SITEFINDER, etc... can also be used for binding pocket identification [212, 213]. These methods can be divided into three classes: 1) geometric methods that use geometric algorithms to identify the cavities on a protein's surface, 2) energy-based approaches that calculate van der Waals, electrostatic, hydrogen binding, hydrophobic, and hydrophilic interaction energies of probes with the pocket, and 3) molecular dynamics-based methods that use multiple conformations of the target protein to predict likely binding sites. After identifying the binding pocket, one important characteristic called binding pocket volume needs to be calculated. This information will give insight on excluding ligands which are too bulky to fit in the pocket during the lead identification process.

Molecular Docking

In addition to finding a target protein, it is also necessary to find the optimal interaction mode between the potential receptor and the small molecule probes (ligand molecules). The term docking is used for computational schemes that attempt to predict the structure of an intermolecular complex formed between a receptor and a ligand. Details of how these intermolecular complexes form are necessary for successful drug design. Most of the time, the receptor is a protein, and the ligand can either be another protein or a small molecule such as a potential drug.

Molecular docking is one of the widely used techniques in lead optimization and “hit” identification [214]. The molecular docking approach can be used to model the interaction between a target protein and a ligand molecule at the atomic level. This allows us to characterize the behavior of ligand molecules in the binding site of target proteins. The first docking program, DOCK, was developed by the Kuntz group in 1982 [215]. It worked based on optimizing the degree of structural complementarity between the target protein and ligand molecule.

Docking methods can be classified as either rigid body or flexible, based on the representation of protein (receptor) and ligand molecules during the docking process. In the early studies, ligand-receptor binding mechanisms were based on the lock-and-key theory proposed by Fischer, in which both the ligand and receptor were considered as rigid bodies. Then Koshland took the lock-and-key theory a step further with induced-fit theory [216, 217]. This theory stated that the binding site of the receptor was continually reshaped during protein-ligand interactions. It also suggested that the ligand should be considered as flexible during docking. As a result, the binding events could be characterized more accurately compared to rigid-body docking. Usually, rigid-body docking simulations are applied to screen a large database during an initial virtual screening process. However, flexible docking methods are still necessary for optimizing ligand poses obtained from an initial rigid docking procedure. With the availability of faster computers and also inexpensive clusters of computers, flexible docking procedures are becoming more commonplace. Some example programs that include ligand flexibility are Glide [218], FRED [219], AutoDock [220], GOLD [221], and FlexX [222], but these are just a few examples of many available docking programs. Some of the successful docking applications in drug discovery campaigns that used docking software include: FK506 immunophilin in 2006 using DOCK, aurora kinases inhibitors in 2006, cytochrome P450 inhibitors in 2011, and falcipain inhibitors in 2011 using Glide.

Docking algorithms produce a vast number of potential poses of the ligand bound to the receptor. A scoring function distinguishes the correct poses from incorrect poses in a reasonable amount of computation time. These functions estimate the binding affinity between the protein and ligand rather than calculating it based on some assumptions and simplifications. These functions can be classified as: force-field-based, empirical, or knowledge-based scoring functions. After identifying the target protein, it is necessary to confirm whether the correct target has been identified or not. Validation processes such as reliable and suitable animal models, or gene targeting and expression tools, help researchers find any unwanted or adverse reactions due to binding of the drug to a secondary target.

Lead Identification and Optimization

A lead compound is a chemical compound having the basic structural requirements for showing the necessary pharmacological or biological activity. The identification of lead compounds and optimization of pharmacological properties are the focal points of early-state drug discovery. Currently, most pharmaceutical industries use high-throughput screening (HTS) as a means to identify new lead compounds [223]. Using HTS, active compounds such as antibodies or genes which regulate a specific biomolecular pathway may be identified [223]. Although HTS is a commonly used method in the pharmaceutical industry, there are some disadvantages of this method [223]; namely the high cost and time-demanding nature of the process have led to the increasing employment of SBDD with the use of complementary computational methods. A lead compound should also have many possible structural variations for further improvement of the binding, or for further enhanced action. Increasing the affinity of a drug molecule towards its target protein will enhance its potency. Binding affinities can be calculated from running ensembles of molecular dynamics simulations. High throughput docking, informatics, and docking simulations are a few computer-based techniques that help to identify a lead compound.

Once a small molecule has been identified as a lead compound, it must be evaluated before going to the next stages. It is essential to realize that the ranking given by the scoring function is not always indicative of a true binding constant, since the model of protein-ligand interaction is intrinsically an approximation. Furthermore, both the protein and ligand flexibility as well as solvent effects are not accurately described. In general, several molecules that score well during the docking run are also examined in further tests since even the top-scored molecules could fail the *in vitro* assays. Lead compounds are first assessed using computer visualization tools and can often be optimized for increased affinity. Leads are also examined for oral availability using Lipinski's "Rule of 5", which states that possible leads should have less than five hydrogen bond donors and less than ten hydrogen bond acceptors, a molecular weight less than 500 Da, and a calculated log of the partition coefficient less than 5 [224]. There are also other factors such as chemical and metabolic stability and the ease of synthesis can also play an important role in making the decision to proceed with a particular lead candidate. Finally, lead compounds reach into the wet lab for biochemical evaluation.

1.7.2 Ligand-Based Computer Aided Drug Design (LBDD)

The ligand-based drug design method relies on knowledge of ligand molecules that interact with a target of interest. The structures of the ligand molecules that are known to interact with the target protein are collected and used as reference structures in these methods. The major goal of this approach is to identify and extract the important physicochemical properties responsible for these interactions, and also discard the information which is not relevant to the interactions. This method is also considered as an indirect approach for drug discovery since it does not require the structural information of the target of interest. Some popular approaches of LBDD are pharmacophore modeling, molecular similarity search, and quantitative structure-activity relationship (QSAR) modeling [225].

LBDD techniques apply different computational algorithms for describing properties of ligand molecules based on the biological function to be predicted. Molecular properties or descriptors can be structural as well as physicochemical, depending on the complexity of the problem. The molecular descriptors can be defined in terms of geometry, volume, surface area, molecular weight, ring content, bond distances, bond angles, interatomic distances, electronegativities, topological charge indices, polarizabilities, functional group composition, aromaticity indices, solubility, octanol/water partition coefficient, partial charges, number of hydrogen bond donors etc... [226, 227, 228, 229, 230]. These descriptors can themselves be obtained in several different ways, such as quantum-mechanical calculations, molecular mechanics, or graph theoretical methods. Based on "dimensionality" of the chemical representation of these descriptors, they are classified as: one-dimensional (1D), which includes scalar physicochemical properties such as molecular weight; two-dimensional (2D) molecular constitution-derived descriptors; and three-dimensional (3D) molecular conformation-derived descriptors.

Pharmacophore Modeling

The concept of a pharmacophore was first introduced by Ehrlich in 1909 [231]. A pharmacophore is a partial molecular framework that carries the essential features responsible for the biological activity of a drug compound. The pharmacophore model can be established in either a ligand-based or structure-based manner. Pharmacophore models may be built using only knowledge of the structural features of active ligand molecules when limited or no structural information of target protein is available. In cases where 3D structural information of the target protein is known, then the active site information can also be used in producing the models. The pharmacophore can be defined using structural features such as acidic groups, basic groups, hydrogen bond acceptors, hydrogen bond donors, partial charges, aliphatic hydrophobic moieties, and aromatic hydrophobic moieties. However, the models built using hydrogen bond donors and acceptors, plus acidic or basic residues are found to be most effective [232]. There are several programs such as DISCO, GASP etc… available to generate pharmacophore models [233]. The overexpression of murine double minute 2 oncoprotein (MDM2), which inhibits p53 tumor supressor, is responsible for approximately 50% of all human cancers. Reactivation of MDM2-p53 integration has been appeared to be a novel approach for improving caner cell death. Bowman *et al.* [234] generated a pharmacophore model based on hydrogen-bond donor sites and hydrophobic sites of the active site of MDM2. They used snapshots from a molecular dynamic simulation of MDM2 bound to p53 tumor suppressor. The resulting structures were used in generating the pharmacophore model. A virtual screening of a library containing 35,000 compounds determined 27 hits. After testing in a binding assay, four compounds were identified as true hits.

Molecular Similarity Searches

Molecular fingerprints are a way of representing a molecule's structure in digital form. Binary digits represent either presence or absence of particular features in the molecule. The fingerprint method allows rapid structural comparison between molecular structures. In molecular similarity searches, fingerprint methods are used to find novel compounds based on the knowledge of physical and chemical similarity to known drugs for the target protein. These similarity search methods are simple yet effective, since molecules with similar structure behave similar in terms of binding properties [6]. For example, a G-protein-coupled receptor GPR30 specific agonist which activates GPR30 was developed using similarity searches.

Quantitative Structure-Activity Relationship (QSAR)

QSAR is a computational method that finds a mathematical relation between structural features of the ligand molecules that bind to a target, and their corresponding biological activity [235, 236]. Molecules with similar structures are presumed to have similar biological activity within this method [237]. QSAR models have been used successfully on several drug targets, such as renin [238], thrombin [239], and carbonic anhydrase [240]. Several different kinds of 2D and 3D QSAR models were developed over the last decades. These methods differ in terms of the chemical descriptors and mathematical approaches used in creating the models. QSAR relationships can also be used to predict the activity of new drug molecules.

In order to assess the activity of drug molecules, many different physical and chemical properties can be used. Among these, half-maximal inhibitory concentrations (IC50) and inhibition constants (K_i) are the most commonly used measures. IC50 values are used to measure a drug's efficacy; it is the amount of drug necessary to inhibit a biological process by half. Whereas K_i values are equilibrium constants used to quantify the inhibitory potency of the inhibitor. QSAR models can be used to study the positive or negative influence of a specific descriptor of a drug molecule on its activity. In classical 2D QSAR models, physical and chemical properties such as geometric, steric, electronic, and hydrophobic features of compounds are correlated with their biological activity. Whereas in 3D QSAR models, in addition to features used in 2D models, quantum chemical features are also used. In recent years, various machine-learning algorithms are also being used in the development of QSAR models [241].

QSAR models are built by collecting a group of active ligands which bind to the desired target protein, and then their activities are identified via literature searches, database screening, and high-throughput screening experiments. The next step is to find the structural or physicochemical properties most affecting biological activity. Later, a mathematical model is generated to find the relationship between those properties and their biological activity. Finally, the model is applied to predict the activity of test compounds in the database. The success of a QSAR model depends on selecting the descriptors that result in a mathematical relationship that is successful in predicting biological activity. It is also important to use a chemically diverse sampling space as the training set in order to develop a suitable model, so that potential hits will not be missed while screening the library. Statistical methods such as multivariable linear regression are used in linear QSAR models when the activity/descriptor relation is linear. This helps one to pick molecular descriptors that are important in predicting the biological activity of interest. However, the relation is not always linear. In that scenario, machine-learning tools such as artificial neural networks (ANNs) are used to generate QSAR models [45]. After finding the right descriptors to build the QSAR model, these models can be validated using cross-validation methods. Some successful applications of QSAR in drug discovery include Zolmitriptan, Norfloxacin, and Losartan [242].

Virtual Screening

Ligand-based virtual screening approaches are used when there is little or no structural information available for the therapeutic target [243]. These tools require knowledge of active ligand molecules with some biological activity against a target. Ligand-based approaches include compound classification methods and machine-learning algorithms. Virtual screening (VS) can be done using chemical similarity searches and virtual docking, or by identifying compounds by predicted biological activity through QSAR studies or pharmacophore modeling. In VS, libraries of commercially available drug-like compounds are computationally screened against a database containing targets of known structure. The compounds which are predicted to bind well to the target are experimentally tested [244].

The screening process produces a small set of molecules called hits, and these are subjected to ranking methods. The ranking procedure compares the similarity between hits and a query compound. The similarity may be in terms of biological activity or the optimal docking pose for each ligand bound to the target protein. Usually, these initial hits are subjected to higher level computational techniques for further screening procedures. This procedure may not produce a drug compound which is ready for clinical studies, but at least it provides insight into leads that have not previously been associated with a target. The money-saving advantage of utilizing computational schemes in the lead optimization phase of drug development is significant since it reduces the number of compounds that must by synthesized and tested *in vitro* [245].

Virtual screening utilizes high-performance computing to accelerate the screening process of large chemical databases when finding the ligand molecules to be synthesized [246]. In order to perform VS, a library must be designed that includes a wide variety of sizes and features. There are three kinds available; general libraries designed to screen against any target, targeted libraries which can be used for specific target, and focused libraries designed for a family of related targets. For example, Fink *et al.* [247] generated a database, referred to as GDB, containing 26.4 million possible organic structures using C, N, O, and F atoms up to 11 atoms in total. Later, a database called GDB-13 was created by Blum and coworkers in 2009. This database includes C, N, O, S and Cl atoms and contains around 930 million compounds. These databases are frequently used in performing VS.

Similarly, apart from developing a ligand database for VS, it is also important to represent molecular structure in a way that can be efficiently read and stored by computer systems. In 1985, Weinninger [181] designed a chemical notation method called SMILES (Simplified Molecular Input Line Entry System) based on the principles of molecular graph theory. This method permits rigorous structure representation by use of a very small and natural grammar. This is one of the most commonly used methods for storage and retrieval of compounds across multiple computer platforms.

Drug Metabolism and ADMET Properties

In addition to finding target proteins and lead compounds, optimizing drug metabolism and pharmacokinetic properties such as ADMET (absorption, distribution, metabolism, excretion, and toxicity) are also important for the success of any drug candidate. After lead discovery and optimization, there is significant consideration given to improving the compound's ADMET properties without losing its therapeutic activity. The prediction of a ligand molecule's ADMET properties can help in decision-making and provide valuable information in the development of a computational model. There are several different *in silico* methods available for evaluating ADMET properties of ligand molecules based on simple empirical rules. These methods include structure-based ones to study the interaction of a lead compound with the target proteins involved, and also some ligand-based ones to study key properties using quantitative structure property relation (QSPR) models.

The absorption of a drug molecule depends both on its permeability through the intestine walls and also on its solubility in water [248]. Thus, predictions of permeability and solubility are important in lead optimization [249]. An orally administered drug with poor solubility and high dissolution rate will be excreted without entering the blood stream. This causes the drug to be incapable of producing the desired action and can even cause biological side-effects. To find the solubility experimentally, it requires the drug to be synthesized which is a time-consuming process. However, predicting solubility computationally is fast, thus reducing costs and also time. Various statistical and mathematical models have been developed in the last few decades and are available for calculating ADMET properties to predict the behavior of lead compounds.

1.8 Aims and Thesis Content

The primary aim of this project was to develop an ANN model trained on electron density properties to predict a broad range of molecular properties of carbonyl compounds. As a proof of principle, our initial plan was to use the bond critical point data and key charge density descriptors based on topological features in the Laplacian of the charge density to train ANNs for the prediction of desired properties. The goal is to predict both the spectroscopic properties and interaction energies between carbonyl compounds and a model nucleophile. Optimization of this approach constitutes the major part of this thesis.

In Chapter 2, the methodology for obtaining descriptors from quantum- mechanical calculations is described along with the procedure followed to develop the ANN models. Chapter 3 briefly introduces the important features of QTAIM employed in this work, as well as documenting the results in terms of MAPEs obtained in predicting the required properties. The second overarching goal of evaluating the relative and relevant information content of the topological properties of the total charge density *vs.* its Laplacian distribution is also discussed. Finally, included are some application of these models for larger systems like proteins.

CHAPTER 2

Methodology

All the calculations reported in this dissertation were executed on the MTSU Department of Chemistry's Linux-based VOLTRON 19-node cluster. Among 19 nodes, there are 10 nodes each with 2X quad-core Intel Xeon E5450 3.0 GHz cpus and 48 GB of RAM. The remaining 9 nodes, each contains 2 X 12-core AMD Opteron 6348 2.8 GHz cpus and 128 GB of RAM.

2.1 Electronic Structure Methods

Spartan'10 [250] was used to build structures of carbonyl compounds including aldehydes, ketones, imides, and amides in order to obtain initial cartesian coordinates of all nuclei prior to geometry optimization using *ab initio* methods. All *ab initio* electronic structure calculations were carried out using the *Gaussian09* program [2]; the level of theory utilized was Density Functional Theory (DFT) treatment of electron correlation, using the B3LYP (B=Becke, 3=three-parameter, and LYP=Lee-Yang-Parr) hybrid functional [86, 89]. This level of theory was used to obtaining fully optimized geometries and relative energies of carbonyl compounds in their ground state [86, 89]. The widely used B3LYP was used in our study because it is considered to be one of the most well-balanced and accurate functionals for a wide variety of applications [79, 80, 81, 82]. The Pople split-valence double-ζ basis set (6-31+G*) [251, 252, 253] was employed to describe the atoms of the carbonyl compounds. After obtaining the optimized geometry coordinates, single-point wavefunctions were calculated using M05-2X [254] hybrid meta functional with the 6-311++G** basis set [251].

The geometry optimizations of the carbonyl+fluoride complexes, as well as the isolated molecules, were carried out using standard convergence criteria of 10^{-8} au at the SCF (self-consistent field) level with the B3LYP/6-31+G* set. The convergence criteria of Max

Force=4.5D-4, RMS Force=3.0D-4, Max Disp=1.8D-3, and RMS Disp=1.2D-3 was used for geometry optimization. The optimized geometries and relative covalent interaction energies were calculated using C1-F1 bond distance of 1.7 Å, F1-C1-O1 angle of 110°, and F1-C1-O1-C2 dihedral angle of 90°, as the **starting** geometry parameters. (See Figure 9, where the atoms of a carbonyl compounds are shown with labelling scheme.) To obtain van der Waals interaction energies, geometry optimizations were **started** using C1- F1 bond distance of 3.4 Å, F1-C1-O1 angle of 145°, and the F1-C1-O1-C2 dihedral angle of 90°. The rest of the carbonyl compound's geometry used starting values obtained for the isolated molecule.

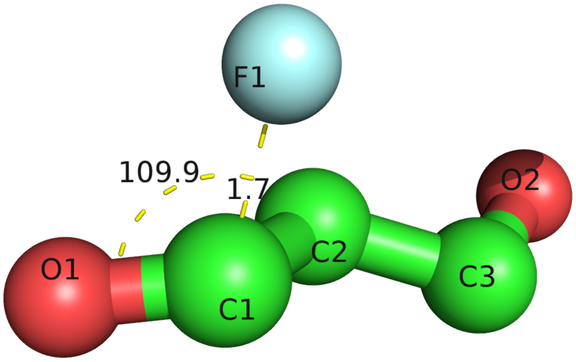


Figure 9: Labelling scheme for the series of carbonyl compounds (hydrogens not shown)

2.2 Electron Density Analysis

The typical approach to quantum chemistry utilizes the wave function ψ as the central quantity. The reason is that once we obtain ψ we can get all dynamical information about this specific state of our target system. A standard example of this approach is the Hartree-Fock approximation. However, the utilization of these quantum-mechanical calculations on macromolecules continues to pose a great challenge for computational chemists. The major limitation of *ab initio* methods is the scaling problem, since the computational cost of these methods increases considerably as the size of the system increases. For instance, HF calculation scales as N^4 since it depends on 4N variables, three spatial and one spin variable for each of the N electrons. The systems we are studying in chemistry contain many atoms and many more electrons. Hence, any wave function based approach quickly reaches an unmanageable size. To circumvent this problem, one can obtain the energy and other properties of interest from a less complicated quantity, the electron density, as the central variable. This is the idea at the heart of DFT.

The electron density is a physical observable through which many chemical and physical properties of the system can be related. It can be determined by experimental methods such as X-ray diffraction [114], or it may come from *ab initio* calculations [255]. The past decade has witnessed enormous methodological developments in X-ray crystallography which has become the preferred technique for the determination of structures of biological macromolecules at atomic scale by taking benefit from the major advances in scientific fields as diverse as biochemistry, molecular biology, computer science, synchrotron physics, and lately robotics. Today, X-ray crystallography can address the determination of complex three dimensional structures of macromolecules, very rapidly. Presently, more than 25 crystal structures are deposited daily in the Protein Data Bank (http://www.rcsb.org) [256].

With the development of quantum chemistry, computationally obtaining the electron density of molecular systems is becoming a routine task. In our study, we obtained the electron density from quantum chemistry calculations. After obtaining the wave function files from the above mentioned calculations, they were imported into *AIMQB*, the driver for QTAIM calculations, and the results were visualized using *AIMStudio* [3]. Topological analysis of the electron density in the carbonyl compounds, within the formalism of Bader's Quantum Theory of Atoms In Molecules (QTAIM) method, was carried out with the *AIMAll* [257] package to determine the position of the BCPs around the carbonyl carbon. The determination of electron density properties such as electron density at the BCP (ρ_{BCP}), the three principal curvatures evaluated at the BCP (λ_1, λ_2, and λ_3), the Laplacian of the charge density at the BCP ($\nabla^2 \rho_{BCP}$), and distance of the BCP from carbonyl carbon nucleus were derived from the *AIMAll* output files. The topological properties of the Laplacian of the charge density, LCP data, were determined using the program *Denprop* [4]. The above mentioned electron density properties were extracted using an in-house Python script.

In this study, we examined a set of 225 carbonyl compounds composed of 108 aldehydes, 86 ketones, 25 amides, and 6 imides. Their molecular structures are shown in Appendix A. The corresponding experimental ^{13}C Nuclear Magnetic Resonance (NMR) chemical shifts (ppm) and C=O vibrational stretching frequency values (cm^{-1}) were collected from a spectral database of organic compounds library (SDBS) provided by the Japanese National Institute of Advanced Industrial Science and Technology (AIST) [178]. This website provides several different kinds of spectra for a large number of organic compounds recorded by different techniques. The interaction energies ($\triangle E_{int}$) between carbonyl compounds and fluoride ion (F$^-$) were calculated using the supermolecular approach,

$$\triangle E_{interaction} = E_{interacting\ reactants} - E_{CC} - E_F \tag{15}$$

where $E_{interacting\ reactants}$, E_{CC}, and E_F are total energies of the interacting reactants, carbonyl compound, and nucleophile, respectively.

2.3 Development of the ANN Model

The machine-learning package *WEKA* (*version 3.6.13*) [258] was used in the Artificial Neural Network (ANN) predictions of ^{13}C chemical shifts and C=O vibrational frequencies of carbonyl compounds, as well as for predictions of interaction energies of carbonyl+fluoride complexes. WEKA stands for Waikato Environment for Knowledge Analysis. This program was developed at the University of Waikato in New Zealand, and the software is freely available [258]. In this study, the multilayer perceptron method, based on the back-propagation algorithm, has been used (discussed in Section 1.6). Our network (I-H-O) had an input layer (I), a hidden layer (H) and an output layer (O). The neural network configuration in this study is as shown in Figure 10. Each layer contains a certain number of artificial neurons, which are equal to the number of input and output values. While training our network with the BCP data, for example, the configuration of our network can be described by the short notation 18-9-1, where each number specifies the number of neurons in one layer starting from the input layer. For training the network with LCP data, and combined BCP and LCP data, the configurations were 30-15-1 and 48-24-1, respectively.

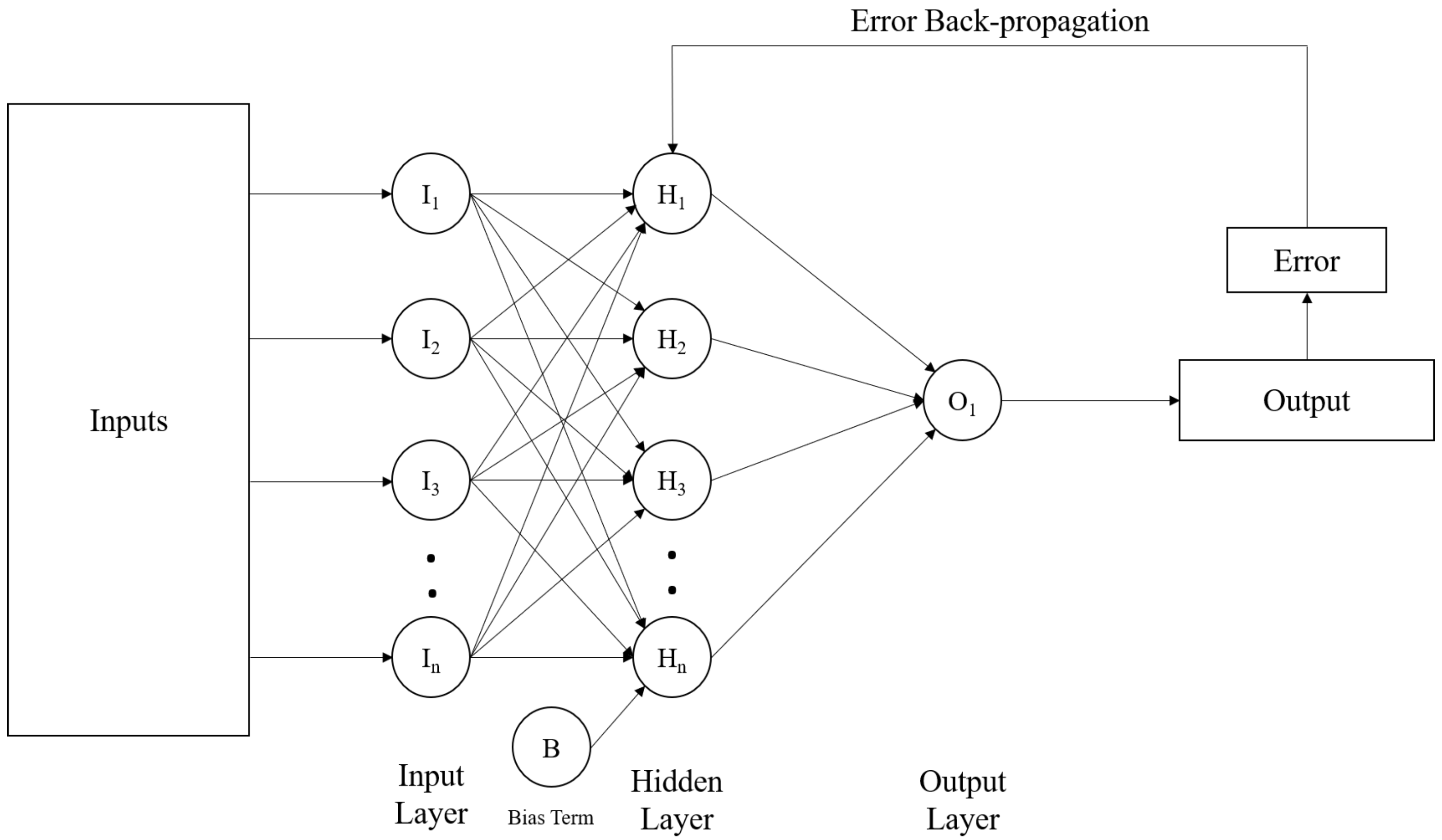


Figure 10: Representation of an artificial neural network used in this study

There are several different parameters that can be adjusted to yield better predictions. These parameters include learning rate (η), momentum (m), number of epochs (N), number of hidden layers (HL) and number of hidden neurons (H). We have tested several network configurations by changing these parameters. In the initial neural network experiments, to train the model, we used default parameters of WEKA: specifically, η=0.3, m=0.2, N=500, and H=a, where (a=[number of input neurons+number of output neurons]/2). Later, we performed several experiments by changing parameters one at a time. Initially, η was changed from 0.001 to 0.300 with an increment of 0.001 while keeping other parameters constant. Once we obtained the best η value, it was then held constant and other parameters such as momentum, number of epochs, and number of hidden layer neurons were changed to find their optimum parameters.

The performance of the trained model on the test dataset is a good indication of its capacity to predict out-of-sample events of the study domain. Thus, an important step in the ANN model development process is the splitting of available data into training, test and validation datasets. We used the leave-one-out cross-validation technique, in which the whole dataset was divided into 225 samples, 224 samples were used for *training* and one sample for *testing*. This process is repeated 225 times, with each sub-sample used exactly once as the testing sample.

2.4 Datasets

The critical point data for all the studied molecules were collected in a matrix, D, of dimension (225 x n_c), where n_c is the number of critical point data per molecule. Thus, each row of D contains critical point data (d_i) of one molecule. A total of nine separate datasets, with each dataset containing critical point descriptors for all 225 molecules, were used to train ANNs to predict ^{13}C chemical shifts, C=O stretching frequencies, and the interaction energy values. The nine datasets include three of each of the following: bond critical points

(BCPs), Laplacian critical points (LCPs), and combined (BCP and LCP) datasets. Each of the previously mentioned datasets also contains one of the following: a class label for experimental ^{13}C chemical shifts, C=O stretching frequencies, or theoretical interaction energy values. A sample BCP and LCP input data of 2-methylbutanal molecule with ^{13}C chemical shift value as a class label can be seen in Tables 1 and 2, respectively.

Figures 11 and 12 show the two kinds of critical points considered in this study. Figure 11 contains the molecular graph of 2-methylbutanal in which the bond critical points in the gradient field of ρ are denoted as green spheres. The critical point properties around the C1 atom of the C1-O2 carbonyl group are collected in Tables 1 and 2.

Table 1: Sample BCP input data for 2-methylbutanal

Distance (au) from carbonyl C	λ_1(au)	λ_2(au)	λ_3(au)	ρ(au)	$\nabla^2\rho$(au)	^{13}C shift (ppm)
0.78	-1.050	-1.020	2.050	0.405	-0.019	205.2
1.48	-0.497	-0.473	0.361	0.253	-0.609	
1.36	-0.728	-0.722	0.511	0.273	-0.939	

Table 2: Sample LCP input data for 2-methylbutanal

Distance (au) from carbonyl C	λ_1(au)	λ_2(au)	λ_3(au)	ρ(au)	$\nabla^2\rho$(au)	^{13}C shift (ppm)
1.035	-1.084	-0.663	9.877	0.046	0.135	205.2
1.035	-1.120	-0.669	9.839	0.050	0.135	
0.954	4.769	5.138	25.911	-1.085	0.298	
0.980	6.479	6.729	18.622	-1.213	0.299	
0.984	5.229	9.111	17.177	-1.069	0.433	

Figure 12 shows (3, -1) critical points in the gradient field of $\nabla^2\rho$ located above and below the C1 atom in pink spheres and (3, +3) critical points in blue sphere around the C1 atom. Only these types of critical points were used in this study because of the following considerations.

The topology of the Laplacian of the charge density allows one to recover the chemical model of localized bonded and non-bonded electron pairs and to characterize local concentrations and depletions of the electronic distribution. A local charge depletion in the valence-shell of an atom is defined by a minimum in $\nabla^2\rho$, a (3, -1) critical point. Whereas a local charge concentration within the VSCC is defined by a (3, +3) critical point [105]. It has already been shown that the regions of local charge concentration and depletion as defined by the Laplacian of ρ, correctly predict the sites of electrophilic and nucleophilic attack, respectively, in a variety of systems [259, 260]. Thus, we made an attempt to use these critical point properties of carbonyl compounds to train our ANN for predicting the spectroscopic properties and interaction energies.

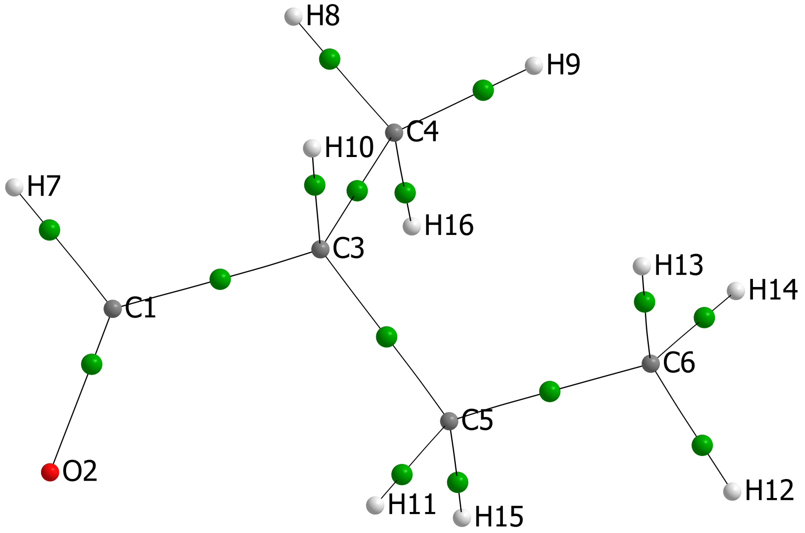


Figure 11: Molecular graph of 2-methylbutanal

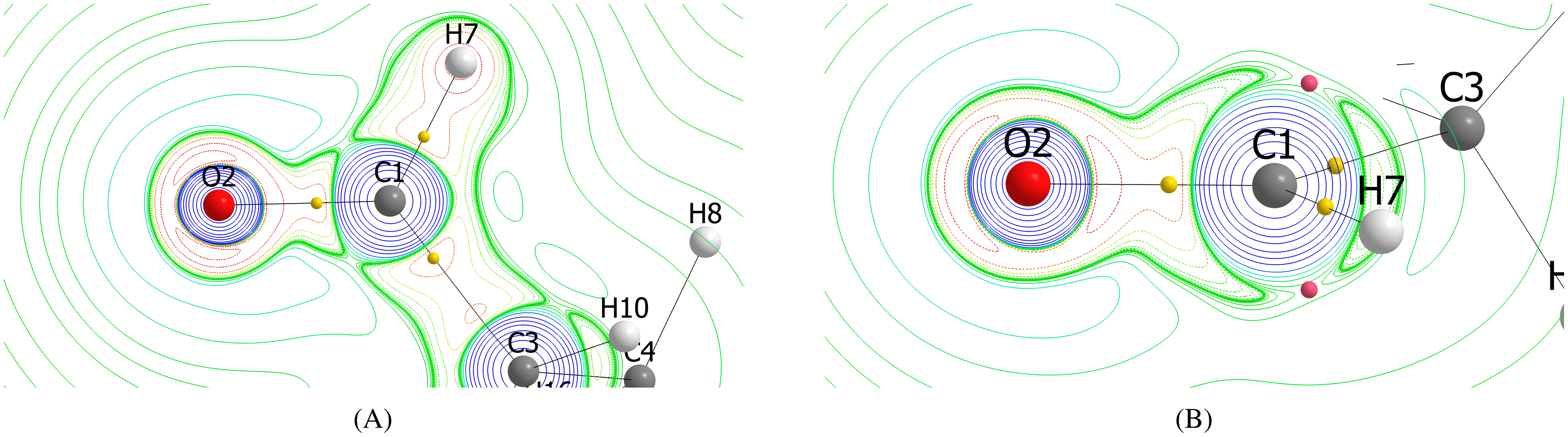


Figure 12: A contour diagram of the Laplacian distribution in 2-methylbutanal. The dashed (solid) lines denote regions of charge concentration (depletion). Starting at a zero contour, contour values change in steps of ±2x10^{n}, ±4x10^{n}, and ±8x10^{n} with n beginning at -3 and increasing in steps of unity. (A) The yellow spheres represent (3, +3) critical points. (B) The pink spheres represent (3, -1) critical points.

Prediction Performance

In our study, the performances of ANNs were measured in terms of the mean absolute error (MAE) and mean absolute percent error (MAPE) as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |P_i - A_i| \tag{16}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \frac{|P_i - A_i|}{A_i} \times 100 \tag{17}$$

In the above-mentioned equations, n represents the number of samples in the dataset, A_i denotes the actual value and P_i represents the predicted value obtained from the ANN output.

CHAPTER 3

Results and Discussion

3.1 Using Artificial Neural Networks in Conjunction with Topological Analysis of Charge Distributions

The primary over-arching goal of this dissertation is to test the hypothesis that topological properties of the electron density of molecules can be used to train artificial neural networks (ANNs) to efficiently predict a broad range of molecular properties. As a proof of principle, we have investigated the abilities of bond critical point data, as well as key charge density descriptors based on topological features in the Laplacian of the charge density [105], to train ANNs for the prediction of spectral properties and interaction energies of carbonyl compounds. In addition to predicting spectral and interaction properties, the secondary over-arching goal of the dissertation is an attempt to gauge the different information content in these two scalar distributions; ρ and $\nabla^2\rho$. This goal will be discussed in Section 3.3.

Within the context of the Quantum Theory of Atoms In Molecules, QTAIM, the properties of electron density (ρ) contains the necessary and sufficient information to define molecular structure and characterize bonding properties [261, 262]. The topological theory of molecular structure has demonstrated connection between topological properties of a molecule's charge distribution and the fundamental concepts underlying the idea of molecular structure, and it has also added the mathematical power of René Thom's Catastrophe Theory to describe the possible mechanisms of *structural change* [261]. As mentioned earlier, QTAIM defines molecular structure and its change in terms of the morphology of the molecular charge distribution. From this topological definition of molecular structure, one can also obtain the *bond paths* which connect the atoms [93]. The characterization of the bond paths (which are also called *atomic interaction lines* in cases where non-covalent interactions exist) using well-defined properties such as ellipticity and bond-bending, adds further correspondence between topological properties of molecular charge distributions and familiar chemical models such as π-bonding delocalization and bond strain [100, 263].

The properties of the electron density evaluated at the BCP, including the density, Laplacian, and distance to the nuclei have all been used previously to extract chemical information on the bond such as its strength, order, polarity etc.. According to Bader *et al.*, the properties at this point can summarize the interaction between two atoms. For example, the strength of the bond or bond order can be correlated with the magnitude of the electron density at the BCP [100]. In general, the electron density at the BCP for covalent bonds is more than 0.20 au, but less than 0.10 au for closed shell interactions including ionic, van der Waals and hydrogen bonding. A study by Grabowski [264] evaluated the properties at BCPs in systems with a large variety of hydrogen bonding interactions ranging from extremely strong to extremely weak interactions in various chemical environments. He found that these parameters correlate well with the strength of the bond.

One of the most valuable electronic properties at the BCP is the Laplacian of electron density, $\nabla^2\rho(\mathbf{r})$. The sign of the Laplacian of $\rho(\mathbf{r})$ determines regions of electronic charge concentration ($\nabla^2\rho(\mathbf{r}) < 0$) and depletion ($\nabla^2\rho(\mathbf{r}) > 0$). Covalent bonds are commonly associated with the overlap of the valence shell charge concentrations of bonded atoms, producing an accumulation of charge at the BCP, thus represented by a negative $\nabla^2\rho(\mathbf{r})$. On the other hand, in closed-shell interactions (ionic bonds or hydrogen bonds or van der Waals), the interaction occurs between two electronic systems with the outermost electronic shells filled, and these are characterized by a positive $\nabla^2\rho(\mathbf{r})$.

The Laplacian of the charge density also plays an important role throughout the QTAIM, being secondary only to the central role of the charge density itself. As shown in Figure 13, it recovers the shell model of electronic structure in terms of a corresponding number of pairs of alternating shells of charge concentration ($\nabla^2\rho(\mathbf{r}) < 0$) and depletion ($\nabla^2\rho(\mathbf{r}) > 0$)

associated with each quantum shell. The topology of the Laplacian of the electron density does provide a faithful mapping of the bonded and non-bonded electron pairs as anticipated on the basis of the Lewis model. For any given Lewis structure, the successful and widely used VSEPR model predicts the arrangement of shared and unshared electron pairs around the atom. Bader *et al.* [105], Bader and MacDougall [108], and MacDougall [109] in the studies of molecular geometry and reactivity observed that the number and relative sizes of the maxima in the VSCC of bonded atoms, as determined by finding the extrema in the Laplacian, correlate directly with the localized bonded and nonbonded pairs of electrons evoked in the Lewis and Gillespies's VSEPR models of the electronic structure of the atom. Thus, one can conclude that the Laplacian of the charge density does provide the physical basis for the Lewis and VSEPR models [265, 93]. The properties of the Laplacian of charge density reproduces not just the geometrical aspects of the Lewis model, but additionally recovers a physical basis for his definition of acid-base reactions. A nonbonded charge concentration is a Lewis base or nucleophile, while a charge depletion is a Lewis acid or electrophile. In the event that two reactants approach each other in a Lewis acid-base-type reaction, their relative orientation can be anticipated by corresponding topological features in the Laplacian functions of their electron density. Charge concentrations of one molecule can be considered to be complementary to depletions of the other. As a result, local features of the Laplacian can be utilized as probable descriptors and predictors of molecular recognition and complementarity.

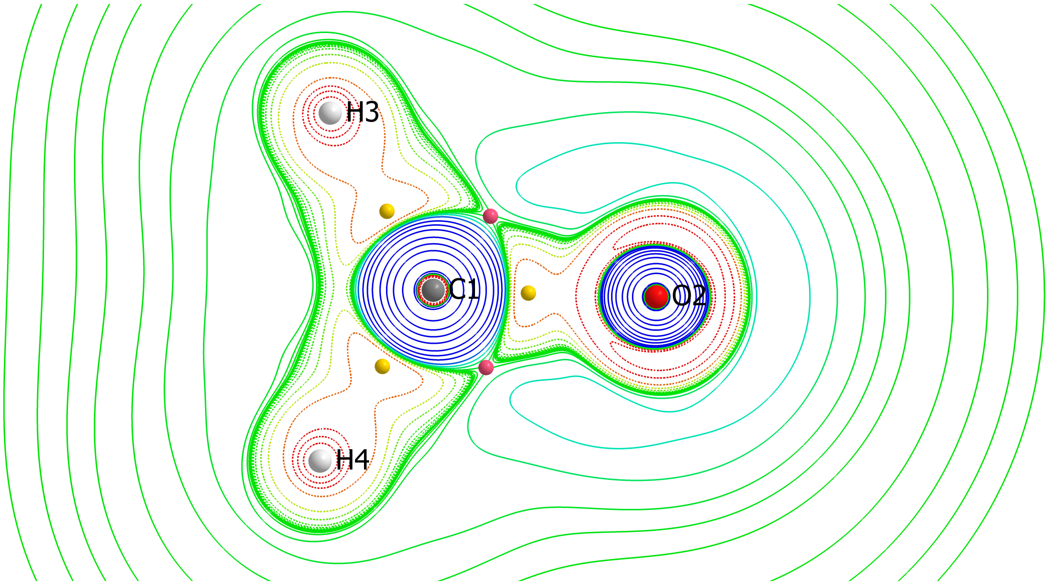


Figure 13: A contour diagram of the Laplacian distribution in formaldehyde. The dashed (solid) lines denote regions of charge concentration (depletion). Starting at a zero contour, contour values change in steps of $\pm 2\text{x}10^n$, $\pm 4\text{x}10^n$, and $\pm 8\text{x}10^n$ with n beginning at -3 and increasing in steps of unity. The pink spheres represent (3,-1) critical points and yellow spheres represent (3,+3) critical points.

Molecular orbitals (in particular, the Highest Occupied Molecular Orbital, or HOMO, and the Lowest Unoccupied Molecular Orbital, or LUMO) and their properties are very useful and important parameters for quantum chemistry and have been extremely successful in rationalizing trends in molecular structure and chemical reactivity [266]. The frontier (HOMO/LUMO) molecular orbital model demonstrates the procedure in which the molecule interacts with other species [267]. The classification of molecules as electron donors and acceptors were first suggested by Gilbert Lewis in 1923 [102]. Accordingly, such electron acceptors and donors are generally referred to as “Lewis acids and bases”. Translating the idea of Lewis acidity and basicity into terms of molecular orbitals, Lewis acids are electrophilic molecules that have a low lying LUMO, while Lewis bases are nucleophilic and have a high energy HOMO. Koopmans’ theorem [268] suggests that the HOMO energy is associated with the corresponding ionization potential in a molecule. Similarly, it suggests that the LUMO energy is associated with electron affinity. This assumption is valid only in the context of restricted Hartree-Fock theory in which it is assumed that the orbitals of the ion are similar to those of neutral molecule (the frozen orbital approximation).

Local properties are greatly desirable in establishing a reactivity-oriented description of molecular systems. Electron density distribution is essential for understanding chemical reactivity, and nucleophilic or electrophilic attacks can be rationalized based on electrostatic interactions. Moreover, the change in electron density under the influence of an approaching reagent is also of major importance. In 1954 Fukui *et al.* [269] have observed for the first time the importance of frontier orbitals (HOMO and LUMO) as principal factors governing both electrophilic and nucleophilic reactions. They developed the frontier electron theory of reactivity in conjugated molecules. This theory was successfully applied to electrophilic, nucleophilic, and radical reactions in various aromatic and other conjugated molecules, and simple justification for this theory was also given. This theory begins with the reasonable idea that the less tightly bound electrons in a molecule should have greater reactivity influence than the more tightly bound electrons.

Finally, one considers the HOMO-LUMO gap, i.e., the energy difference between HOMO and LUMO, as a principal quantity in molecular orbital theory. The gap between HOMO and LUMO has been used as a conventional measure of kinetic stability [270, 271, 272]. A large HOMO-LUMO gap indicates high kinetic stability and low chemical reactivity. Since it is energetically not favorable to extract electrons from a low-lying HOMO or add electrons to a high-lying LUMO. From the previously given examples in several studies [108], it was clear that the regions of charge concentration and depletion in the Laplacian distribution correspond with the regions where HOMO and LUMO are concentrated, respectively. It was demonstrated that the charge concentrations determine the sites of electrophilic attack which correlate with the regions where HOMO is most concentrated whereas the charge depletions (holes) determine the sites of nucleophilic attack which correlate with the regions of space where the LUMO is most concentrated. Molecular orbitals are mathematically arbitrary and cannot be experimentally observed. While the Laplacian of the charge density is observable experimentally.

Information Content in the Topologies of ($\rho(\mathbf{r})$) and ($\nabla^2\rho(\mathbf{r})$)

As we have described in section 1.2, the topological properties of $\rho(\mathbf{r})$ recover the conventional network of bonds in molecules, and even some of their characteristics, such as partial π -character; in essence the molecular structure. Whereas the topological properties of $\nabla^2\rho$ seem more closely connected to models that relate the electronic configuration and distribution of electronic charge to the shape and reactive properties of molecules; in essence; key aspects of the electronic structure. Molecular properties such as vibrational frequencies and NMR chemical shifts are dependent of both molecular and electronic structure, as does chemical reactivity. Nevertheless, it can be instructive to learn which is dominant for a given molecular property. A second over-arching goal of this dissertation is to employ machine-learning to evaluate the relative and relevant information content of the topological properties of the total charge density (ρ) *vs.* its Laplacian distribution ($\nabla^2\rho$).

3.2 Predicting Spectroscopic Properties Using Artificial Neural Networks

The carbonyl stretching frequencies are influenced by both inductive and resonance effects. The frequency increases when electron-withdrawing substituents are added and decreases with electron-donating substituents. Since the stretching frequencies are directly proportional to the bond force constants, increase in wave numbers indicate higher formal bond orders of the carbonyl bond. A study conducted by Nummert *et al.* have shown that the infrared carbonyl stretching frequency of compounds is sensitive to changes in the substituents [273]. They have inspected 22 phenyl esters of substituted benzoic acids. For instance, it was observed that the frequency increased from 1742.6 cm^{-1} to 1749.0

and 1746.9 cm^{-1} when substituent is an electron withdrawing groups such as CN and F, respectively. Whereas the frequency decreased from 1742.6 cm^{-1} to 1741.9 and 1740.7 cm^{-1} when substituent is electron donating group like CH_3 and NH_2, respectively. The properties of the electron density at a BCP provide information on bond order, bond energy, and bond character [99], so we expect there to be a relation between the BCP properties and stretching frequencies.

The ^{13}C NMR chemical shift is the position of the signal in an NMR spectrum relative to a standard, usually tetramethylsilane. The position of this signal is essential for structure elucidation of organic molecules by NMR. Different chemical shifts will be observed for different carbon-13 nuclei in the functional groups within a molecule, depending on their bonds and the atoms attached. Electron density near the nucleus is one of the factors which cause chemical shifts to be increased or decreased. Lower chemical shift values can be observed when neighboring atoms have high electron density and there are electron-donors bonded to the atom, shielding the probed nucleus from an external magnetic field. Higher chemical shift values can be observed when the electron density near the probed nucleus is lowered by electron-withdrawing substituents. Nummert *et al.* [274] performed a study on the influence of substituent effects on the carbonyl carbon ^{13}C NMR chemical shifts in substituted phenyl benzoates. They found that the chemical shift value decreased when they used electron donating substituents and the value increased with the electron donating substituents. Since chemical shifts are dependent on the electron density, we expect these chemical shifts will be accurately predicted when we train the ANNs with properties of electron density; but *which* properties?

To test whether the critical point descriptors of the charge density and/or the Laplacian of the charge density are relevant and crucial towards the prediction of spectral properties of carbonyl compounds, we again used critical point descriptors as a training set in our model. A total of 225 carbonyl compounds (see Appendix A) belonging to aldehydes, ketones, amides, and imide functional groups were selected, and their experimental values of ^{13}C chemical shifts and C=O stretching frequencies were collected from the website of SDBS [178]. To maintain the structural diversity, we tried to include carbonyl compounds containing several different side chains in their structure. The chemical structures of 225 molecules were obtained by varying the substituents of carbonyl compounds RCOR'. Those substituents include electron withdrawing groups (EWGs), electron donating groups (EDGs), and neutral atoms. Different combinations of substituents were used to build the molecules required for our study. Some examples of EWGs include CF_3, CCl_3, NO_2, CN, F, Cl. EDGs include CH_3, OCH_3, OH, OR, and CH_2. The neutral group example is H. In our study, we have included molecules for which experimental NMR and IR values are available in SDBS. The number of atoms in the molecules ranges from 7 to 31 atoms.

The artificial neural network (ANN) requires the numerical description of the chemical environment of the carbon atom of interest for predicting ^{13}C NMR chemical shifts. An appropriate description should meet certain conditions to be considered as input data for neural network. Thus, the input vector for each molecule should be constant in length for describing their properties. The ANN model was developed to predict chemical properties of carbonyl compounds based on the critical point data of charge density and its Laplacian distribution. Considering that 225 compounds were studied in this work, the dimensions of the data matrix D for LCP, BCP, and combined datasets are (225x30), (225x18), and (225x48), respectively. This matrix was used as an input for the prediction of different properties of carbonyl compounds.

As mentioned in section 1.6, there are several different parameters in ANNs that can be adjusted to yield better predictions. In the initial artificial neural network experiments, we used default parameters of WEKA, such as η=0.3, m=0.2, N=500, H=a. To evaluate the ability of the model to predict spectroscopic properties, the mean absolute percent errors (MAPEs) were used as a measure of prediction error. These errors were calculated using equation 17 and are presented in Table 3.

Table 3: Mean absolute percent errors (MAPEs) of predicted values for ^{13}C chemical shifts and C=O stretching frequencies when ANN trained using **default** parameters of the network

Type	property	Experiment 1			Experiment 2			Experiment 3		
		LCP	BCP	Combined	LCP	BCP	Combined	LCP	BCP	Combined
aldehydes	NMR (%)	7.5	7.19	6.52	1.90	1.51	1.87	2.36	1.89	2.04
	IR (%)	1.56	2.21	2.10	0.70	0.54	0.70	0.87	0.61	0.94
ketones	NMR (%)	2.80	9.19	2.51	2.27	2.18	2.19	2.15	3.11	2.33
	IR (%)	1.58	2.36	1.10	0.61	0.60	0.61	0.56	0.69	0.63
imides	NMR (%)	7.79	1.19	2.94	2.64	2.28	3.04	2.50	2.34	2.34
	IR (%)	1.77	1.23	0.95	0.64	0.48	0.83	0.24	0.15	0.29
amides	NMR (%)	12.90	13.9	10.80	2.79	1.47	3.44	2.52	1.88	2.36
	IR (%)	1.37	1.56	2.44	1.44	0.74	1.42	0.98	0.59	1.13

Optimization of Machine-Learning Parameters

The fundamental issue in utilizing ANNs is the parameter tuning. However, there is no precise method to select optimal parameters for the ANNs [55]. For that reason, we designed a set of experiments to study the influence of different parameters on the performance of the ANNs trained with the back-propagation algorithm: the learning rate, momentum, number of epochs, number of neurons in the hidden layer. The learning rate was varied between 0.001 and 0.3 with an increment of 0.001, and we have used the following combinations of momentum, number of epochs, and number of hidden layer neurons, respectively: (0.1-1.0:1.0), (500-3000:100), and (10-20:1). In the above-mentioned combinations, the first two numbers show the range and last number shows the step size.

Effect of Training Data on Prediction Accuracy

The dataset contains a total of 225 carbonyl compounds among which 108 aldehydes, 86 ketones, 6 imides, and 25 amides. We performed three experiments to create different ANN models. From these experiments, we assume that the performance of the model will be influenced by the size of the training set and type of compounds included in the training set.

Experiment 1

Our dataset contains four types of carbonyl compounds. From this, four types of datasets were prepared to develop the ANN models using the combination of compounds shown in Table 4. We followed the same procedure as mentioned above to develop the model with optimized parameters. Once the parameters were optimized, the model then applied to the test data to predict the desired properties of molecules present in the test data. For example, dataset 1 contains 117 compounds which include ketones, imides, and amides. We used leave-one-out cross-validation technique, in which the whole dataset was divided into 117 samples, 116 samples were used for training and one sample for testing. This process is repeated 117 times, with each sub-sample used as the testing sample exactly once. During this process, the parameters have been optimized as in the above-mentioned procedure. Once we obtained the network with optimal parameters, it is then applied to the test data containing aldehydes to predict the required properties. The similar procedure was followed for the remaining datasets in this experiment and the calculated MAPE is presented in Table 5.

Table 4: The combination of compounds used in the dataset to create the ANN model

Datasets	Compounds used to develop the ANN model	Test data
1	Ketones, imides, amides	Aldehydes
2	Aldehydes, imides, amides	Ketones
3	Aldehydes, ketones, amides	Imides
4	Aldehydes, ketones, imides	Amides

Experiment 2

In this experiment we used a larger dataset to train the model compared to the other two experiments. Here the network was trained on all four kinds of compounds. The dataset containing 225 molecules was divided into 225 samples. Among which 224 samples were used as a training set and one sample as a testing sample. The MAPE data are shown in Table 5.

Experiment 3

During the process of developing the model, the ANN was trained on only one kind of compound and predicted the properties of similar kinds of molecules only. For example, dataset 1 contains 108 compounds which are aldehydes. We used leave-one-out cross-validation technique, in which the whole dataset was divided into 108 samples, 107 samples were used for training and one sample for testing. This process is repeated 108 times, with each sub-sample used exactly once as the testing sample. During this process, the parameters are optimized as in the above-mentioned procedure. A similar procedure was followed for remaining datasets in this experiment and the calculated MAPE data are included in Table 5. The optimum parameters for predicting ^{13}C NMR shifts and C=O stretching frequencies for all three experiments are shown in Table 6.

Table 5: Comparison of MAPE of predicted values of ^{13}C chemical shifts and C=O stretching frequencies when using **optimized** parameters of ANN for three experiments

Type	property	Experiment 1			Experiment 2			Experiment 3		
		LCP	BCP	Combined	LCP	BCP	Combined	LCP	BCP	Combined
aldehydes	NMR (%)	12.4	7.43	7.03	1.18	1.24	1.07	1.25	1.32	1.17
	IR (%)	1.53	1.19	1.70	0.61	0.54	0.55	0.54	0.51	0.55
ketones	NMR (%)	1.73	7.91	2.21	1.49	1.92	1.61	1.45	2.18	1.54
	IR (%)	0.81	7.33	2.68	0.41	0.51	0.42	0.40	0.59	0.41
imides	NMR (%)	4.27	2.26	1.10	1.36	2.36	0.93	2.15	2.02	1.44
	IR (%)	2.59	1.04	1.87	0.23	0.43	0.43	0.19	0.06	0.17
amides	NMR (%)	3.51	10.8	4.25	2.60	1.39	2.14	2.07	1.57	1.43
	IR (%)	3.17	1.93	2.89	0.95	0.59	0.93	0.60	0.50	0.60

Table 6: Comparison of optimized parameters used in all the three experiments in the prediction of spectroscopic properties

	Type	Property	LCP			BCP			Combined		
			η [a]	N [b]	H [c]	η [a]	N [b]	H [c]	η [a]	N [b]	H [c]
Exp 1	aldehydes	NMR	0.016	2500	6	0.124	900	5	0.012	1000	5
		IR	0.062	500	7	0.036	1900	11	0.006	500	5
	ketones	NMR	0.016	500	5	0.060	1500	9	0.004	600	5
		IR	0.077	800	7	0.163	3100	12	0.009	500	7
	imides	NMR	0.024	600	6	0.052	3200	9	0.014	800	17
		IR	0.005	1400	9	0.121	500	12	0.006	1100	15
	amides	NMR	0.007	1200	6	0.067	1400	9	0.004	1100	10
		IR	0.004	500	9	0.031	500	5	0.004	600	11
Exp 2		NMR	0.019	900	5	0.041	2500	17	0.006	2300	10
		IR	0.040	500	6	0.052	2500	16	0.014	500	11
Exp 3	aldehydes	NMR	0.012	500	8	0.009	500	10	0.006	500	11
		IR	0.004	1200	7	0.045	2500	9	0.009	500	7
	ketones	NMR	0.012	500	6	0.011	2100	4	0.004	800	8
		IR	0.008	600	5	0.013	1000	5	0.004	600	10
	imides	NMR	0.023	500	4	0.021	500	6	0.037	500	1
		IR	0.001	500	4	0.027	500	5	0.015	500	1
	amides	NMR	0.148	700	4	0.057	1300	6	0.042	500	6
		IR	0.001	500	1	0.008	500	5	0.001	500	2

[a] Learning rate
[b] Number of epochs
[c] Number of hidden neurons

Discussion of The Results

As the first step, we trained our ANN model using default parameters (η=0.3, m=0.2, N=500, and H=a) to predict ^{13}C NMR chemical shifts of carbonyl compounds. Producing an optimal ANN model is an important part of modeling with artificial neural networks, which is needed to harness the maximum benefit from the computational intelligence of the network. The accuracy of the results improved for all the three experiments after tuning the parameters (Table 6) when compared to using default parameters of the network. We observed that the accuracy of the models improved with optimal parameters compared to default parameters of the network. For this purpose, we performed several experiments by changing the parameters one at a time as discussed in section 2.3. After tuning the parameters, the models produced a maximum of 37% reduction in the MAPEs when compared to ANN trained using default parameters.

To investigate whether the accuracy of the model is influenced by the type of molecules included in the training set, the ANN was trained using three different ways (see section 3.2). After comparing the MAPEs of three experiments, we found that the models performed better when the ANN was trained on all four kinds of carbonyl compounds before applying on the test data. We observed higher MAPEs in experiment 1, where the model was applied on test data containing new kind of carbonyl compounds which were not seen by the network as compared to the other two experiments. After comparing the results between the three experiments, experiment 2 gave smaller MAPEs of 1.54, 1.45, and 1.38 for predicting ^{13}C chemical shifts when the ANN trained on BCP, LCP, and combined datasets respectively. A slightly higher MAPEs were obtained when ANN was trained using BCP datasets in the predictions of NMR shifts. We also observed that in most cases combined dataset improved the prediction accuracy of the network. As we expected, the LCP dataset provided more accurate results for predicting ^{13}C chemical shifts, since the values of chemical shift depends on the chemical environment of a given nucleus and this information is well captured in the Laplacian of the charge density. These findings are comparable with other ML models developed specifically for predicting chemical shifts. For example, the Meiler *et al.* group [187] used artificial neural networks for predicting ^{13}C chemical shifts trained on 1.3 million molecules and achieved a root mean square deviation (RMSD) of 1.3 ppm. They have developed descriptors based on the atom types and the chemical environments of all atoms numerically. Atom types were determined using their element number, hybridization state, and number of bonded hydrogen atoms. The chemical environments of carbon atoms were described by sorting the atoms in spheres and counting the occurrence of every atom type in each sphere. In our approach we achieved RMSDs of 4.0 ppm, 4.4 ppm, 3.9 ppm, when we used LCP, BCP, and combined datasets respectively to train the ANN. These results are obtained using only 225 molecules.

For the second test we trained our ANN on the same datasets (BCP, LCP, and combined) for predicting C=O stretching frequencies. We obtained similar results in terms of prediction accuracies as observed in NMR chemical shift predictions. The MAPEs were reduced by using optimized parameters when compared to default parameters. For example, the network with the tuned parameters produced a maximum of 28% reduction in the MAPEs when compared to ANN trained using default parameters. In experiment 2 for predicting C=O stretching frequencies, we got MAPEs of 0.53, 0.56, and 0.54 when the model trained using BCP, LCP, and combined datasets, respectively.

Summary

In this study, the employed ANN model was able to predict spectroscopic properties such as C=O stretching frequencies and ^{13}C NMR chemical shifts of carbonyl compounds with acceptable accuracy. The NMR chemical shifts were predicted more accurately when we used the Laplacian critical point descriptors as compared to charge density only descriptors. For C=O stretching frequencies, bond critical point descriptors provided more accurate results in comparison with the Laplacian critical point descriptors. This intuitively supports the initial hypothesis that different sets of topological data may be better suited for predicting different molecular properties. Stretching frequencies are primarily associated with molecular structure, as BCPs are defined by topological properties in $\rho(\mathbf{r})$. Whereas chemical shifts are primarily associated with electronic structure, which are coarsely related to topological properties of $\nabla^2\rho(\mathbf{r})$. Since neither of these types of molecular properties are *exclusively* related to molecular or electronic structure, it makes sense that the combined dataset gave more accurate results compared to when we used separate datasets. These applications of ANNs have shown that it is possible to predict spectroscopic properties with a minimum amount of input data (30 numerical descriptors for LCP and 18 for BCP). In a study conducted by the Meiler group [187] for predicting NMR shifts using ANNs, the network was trained on 180 descriptors. For future studies, the accuracy of our model can be improved by providing more topological data which can easily be added to our current descriptors. Apart from adding chemical insight into the information contained in different types of experimentally accessible topological data, a significant reduction in prediction times is also potentially a major benefit of the current method. After creating the ANN model by training only once, which takes about few hours on a computer with i7 processor and 8 GB RAM, they run a prediction about several times faster (within few seconds) and independent of direct access to the datasets.

Another important conclusion which can be drawn from ML models is recognizing that the important descriptors strongly correlated with different spectroscopic and chemical properties. In this study, both BCP properties and the LCP properties are used and compared to one another. After reviewing the results in terms of MAPE, BCP properties have highest contribution to predicting C=O stretching frequencies, whereas the Laplacian critical points have highest contribution to predicting ^{13}C chemical shifts. In conclusion, this study shows that ANN trained on QTAIM properties is potentially applicable for the fast and reliable prediction of spectroscopic properties of carbonyl compounds.

3.3 Predicting Covalent and Van der Waals Interaction Energies Using ANNs

The predictions of spectral properties were a proof-of-principle experiments, whereas the intermolecular interactions for large molecules is the major goal of our study as it plays a key role in computational chemistry applications in the drug design process. The prediction of protein-ligand binding energies is of central interest in CADD, but it is still difficult to accomplish a high degree of accuracy. Here, we report the prediction of both covalent and van der Waals intereaction energies using ANNs trained on ρ and/or $\nabla^2\rho$ critical point descriptors. A nucleophilic addition reaction between a fluoride ion and a carbonyl group was taken as an initial model for a chemical interaction in our investigation, and the interaction energies ($\triangle E_{int}$) were calculated for both strong (covalent bond formation) and weak (van der Waals) interactions for our set of 225 carbonyl-containing molecules. Here we used a supermolecular approach [275] in calculating $\triangle E_{int}$ values between carbonyl-containing molecules and fluoride ion using the following equation:

$$\triangle E_{interaction} = E_{CC+F^-} - E_{CC} - E_{F^-} \tag{18}$$

where E_{CC+F^-} , E_{CC}, and E_{F^-} are total energies of the interacting reactants (carbonyl compound+F$^-$ complex) and non-interacting (free), carbonyl compound, and nucleophile (fluoride ion), respectively.

Initially, acetone interacting with fluoride ion was taken as an example for finding the initial coordinates to be used in calculating interaction energies for 225 molecules in our dataset. For this, the equilibrium geometry and relative energy of each isolated monomer, and the total energy of the complex, were obtained via geometry optimizations at B3LYP, HF, and MP2 levels of theory [276] using the 6-31+G* basis set [251, 252, 253]. We took the bond distance of C1-F1 (R_{C1-F1}) as the reaction coordinate, while the remaining parameters of the complex geometry were optimized (keeping the coordinate R_{C1-F1} fixed at different values in the range of $1.3 \leq R_{C1-F1} \leq 3.5$ Å with the step size of 0.1 Å). At each R_{C1-F1} coordinate, both the F1-C1-O1 angle of 90° and F1-C1-O1-C2 dihedral angle of 90° were used as starting parameters for geometry optimizations (Figure 14). These partial geometry optimizations of the complex (CC+ F^{-}) were also performed at B3LYP, HF, and MP2 levels using the same 6-31+G* basis set. Figure 15 shows the interaction energy plotted as a function of the distance between carbonyl C and fluoride ion.

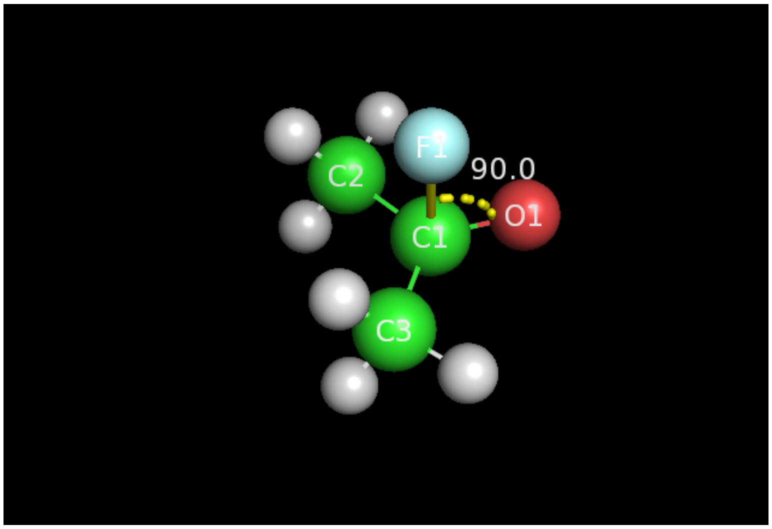


Figure 14: Geometry used in the current study

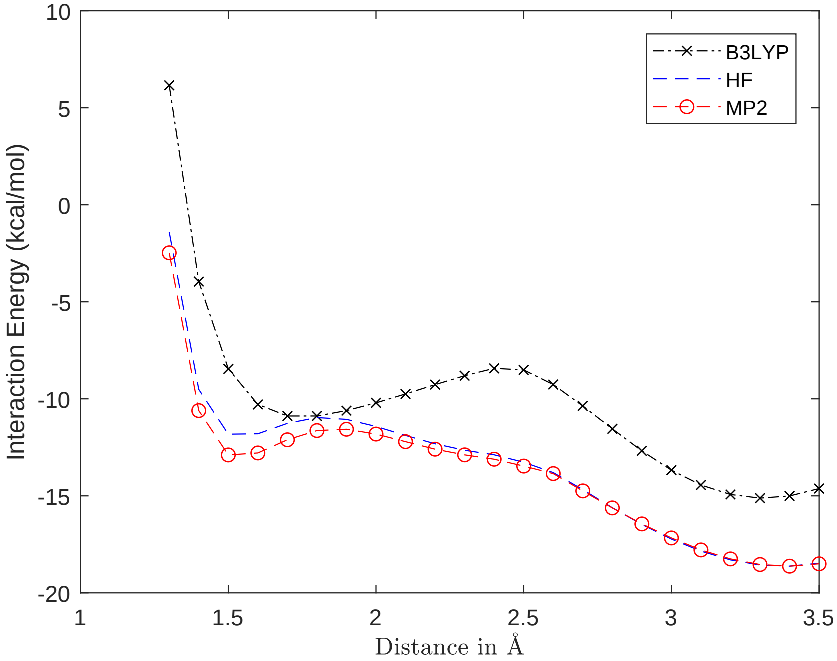


Figure 15: Interaction energy curve of acetone and F^- interaction calculated with the 6-31+G* basis set using different levels of theory.

From these calculations we observed two stationary points at the C1-F1 distances in the neighborhood of 1.7 Å and 3.4 Å. The optimized geometries around 1.7 Å distance all show that the angle of the nucleophilic approach ($\angle FCO$) is around 111 ± 1 °, as shown in studies by Burgi and Dunitz [111]. They have analysed six crystal structures with intramolecular interactions between N and C=O. The angle of 107± 6° was observed in their studies. The long range interactions all occurred at a nucleophilic approach with an average angle of 146 ± 1 °. The product formed between the carbonyl C and fluoride ion at a bond distance of 1.7 Å is conventionally called a "covalent interaction" (see Figure 16), even though the properties at the BCP are characteristic of a closed-shell interaction. This has been observed for most "covalent" bonds to fluorine, as in F_2 and CH_3F [108, 98]. The other equilibrium geometry, formed at 3.4 Å, is a van der Waals interaction (see Figure 17).

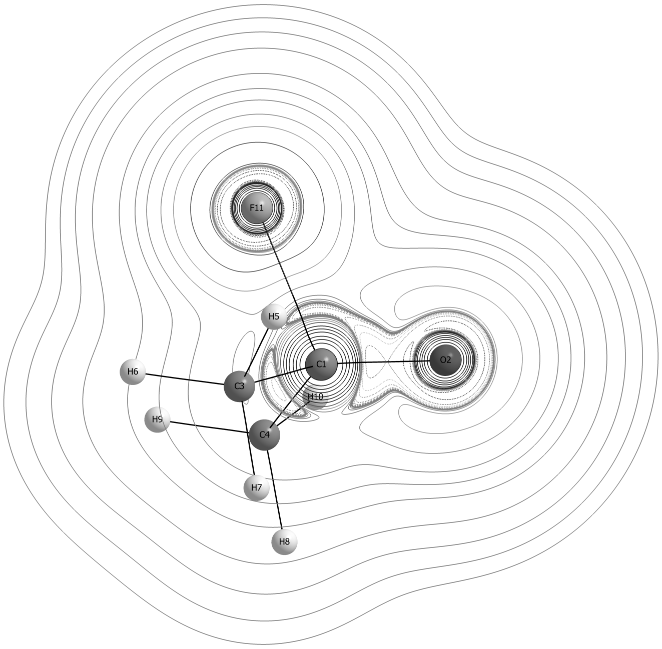


Figure 16: A contour diagram of the Laplacian distribution of covalent interaction between acetone and fluoride ion. The fluoride ion approaches C=O bond axis at angle of 110°. The dashed (solid) lines denote regions of charge concentration (depletion). Starting at a zero contour, contour values change in steps of ±2x10^{n}, ±4x10^{n}, and ±8x10^{n} with n beginning at -3 and increasing in steps of one.

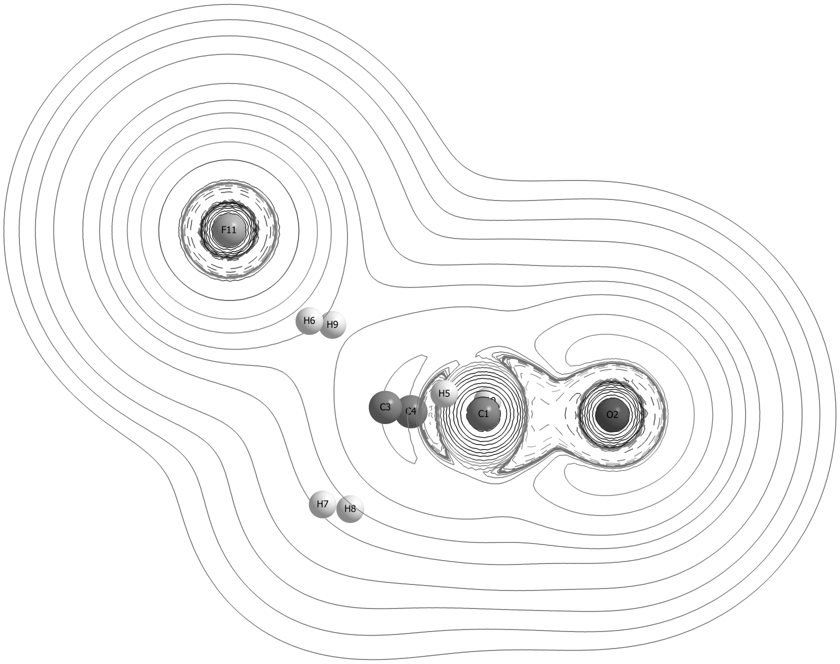


Figure 17: A contour diagram of the Laplacian distribution of the van der Waals interaction between acetone and fluoride ion. The fluoride ion approaches C=O bond axis at angle of 145°. The dashed (solid) lines denote regions of charge concentration (depletion). Starting at a zero contour, contour values change in steps of $\pm 2\text{x}10^n$, $\pm 4\text{x}10^n$, and $\pm 8\text{x}10^n$ with n beginning at -3 and increasing in steps of one.

For calculating covalent $\triangle E_{int}$ values of 225 molecules in the dataset, based on our initial studies for acetone reported above, we started our calculations using the F1-C1 bond distance of 1.7 Å, the F1-C1-O1 angle of 110°, and the F1-C1-O1-C2 dihedral angle of 90° as a starting geometry. Then, full geometry optimizations were performed. For calculating van der Waals $\triangle E_{int}$ values, and again based on our initial studies for acetone reported above,we used the F1-C1 bond distance of 3.4 Å, the F1-C1-O1 angle of 145°, and the F1-C1-O1-C2 dihedral angle of 90° as a starting geometry. These energies were extracted using an in-house Python script and used as a class label in the dataset to train the ANN.

The BCP, LCP, and combined datasets used in this work are the same as those that were used to develop models for predicting spectroscopic properties (Section 3.2), but here the theoretical interaction energy is used as training property (class label). We performed the same three experiments mentioned in the section 3.2. For all three experiments, the mean absolute errors (MAEs) were calculated using Equation 16. The MAEs of predicted $\triangle E_{int}$ values are summarized in Table 7. These results were obtained with the default parameters, η=0.3, m=0.2, N=500, H=a to train the ANN. Among these three experiments, the second and third experiments performed well in comparison to the first experiment. To improve the accuracy of predictions, we again performed several experiments as described in the Section 3.2 by changing the parameters of the ANN one at a time. The obtained results are presented in Table 8. Also, the optimum parameters obtained after tuning the network are shown in Table 9.

We also compared these results in terms of MAPEs. The best MAPEs for predicting covalent and van der Waals $\triangle E_{int}$ values obtained in experiment 2 are 6.4 % and 9.2 % when the network was trained on LCP and BCP datasets, respectively. In general, the interaction energies are much harder to predict than spectroscopic properties. For instance, Hartree-Fock overestimates the binding energy of H_2 by more than 100%!

In the study by Jenness *et al.* [277], the authors evaluated various theoretical approaches for calculating $\triangle E_{int}$ values between a water molecule and a series of linear acenes, particularly benzene, anthracene, pentacene, heptacene, and nonacene. They explored long-range interactions between these molecules. The theoretical methods included in their study are DFT-SAPT [278], Grimme *et al.* [279, 280] schemes of DFT-D2, DFT-D3, and the van der Waals density functionals (vdW-DF) of Lundqvist *et al.* [281]. These results were compared to those from wave function-based methods such as second-order Møller-Plesset perturbation theory (MP2) [282], coupled-cluster with single, double, and perturbative triple excitations (CCSD(T)) [283, 284, 285], and the spin-component-scaled MP2 (SCS-MP2)[286]. All the calculations were carried out with the MOLPRO *ab initio* package (version 2009.1)[287]. The calculated interaction energies between water and, the listed acenes, using the DFT-SAPT method, are -3.20, -3.34, -3.21, -3.21, and -3.21 kcal/mol, respectively. The mean absolute errors relative to the DFT-SAPT method were

Table 7: MAEs of predicted covalent $\triangle E_{int}$ and van der Waals $\triangle E_{int}$ using the **default** parameters for training ANN

Type	Property	Experiment 1			Experiment 2			Experiment 3		
		LCP	**BCP**	**Combined**	**LCP**	**BCP**	**Combined**	**LCP**	**BCP**	**Combined**
aldehydes	Covalent $\triangle E_{int}$ (kcal/mol)	4.17	4.74	8.14	3.63	4.06	3.77	3.51	3.55	4.04
	van der Waals $\triangle E_{int}$ (kcal/mol)	8.62	9.79	10.5	5.79	5.82	6.06	5.70	5.63	6.68
ketones	Covalent $\triangle E_{int}$ (kcal/mol)	12.5	7.02	6.31	4.33	3.93	3.52	4.86	6.11	6.34
	van der Waals $\triangle E_{int}$ (kcal/mol)	14.1	5.87	10.3	6.57	6.70	6.39	6.83	4.91	6.85
imides	Covalent $\triangle E_{int}$ (kcal/mol)	20.9	11.0	11.1	4.31	4.83	3.58	2.92	4.10	3.19
	van der Waals $\triangle E_{int}$ (kcal/mol)	18.0	7.19	33.4	5.03	8.26	5.95	6.23	7.61	8.19
amides	Covalent $\triangle E_{int}$ (kcal/mol)	21.8	21.4	15.1	11.6	6.32	11.8	11.8	9.22	9.02
	van der Waals $\triangle E_{int}$ (kcal/mol)	27.4	16.4	25.7	16.7	6.51	9.92	11.9	5.24	7.29

Table 8: MAEs of predicted covalent $\triangle E_{int}$ and van der Waals $\triangle E_{int}$ using the **optimized** parameters for training ANN

Type	Property	Experiment 1			Experiment 2			Experiment 3		
		LCP	**BCP**	**Combined**	**LCP**	**BCP**	**Combined**	**LCP**	**BCP**	**Combined**
aldehydes	Covalent $\triangle E_{int}$ (kcal/mol)	9.93	4.32	8.51	2.66	3.22	2.81	2.50	2.78	2.55
	van der Waals $\triangle E_{int}$ (kcal/mol)	13.4	6.33	12.3	5.12	4.53	4.77	5.19	4.33	4.31
ketones	Covalent $\triangle E_{int}$ (kcal/mol)	9.28	8.43	8.59	3.27	3.84	3.28	3.14	4.42	3.40
	van der Waals $\triangle E_{int}$ (kcal/mol)	6.01	8.66	6.79	4.34	4.53	4.54	4.36	4.51	4.47
imides	Covalent $\triangle E_{int}$ (kcal/mol)	14.9	10.5	13.2	5.25	2.98	2.81	1.89	2.07	1.98
	van der Waals $\triangle E_{int}$ (kcal/mol)	14.5	6.72	12.6	4.74	8.36	5.13	5.12	4.97	5.23
amides	Covalent $\triangle E_{int}$ (kcal/mol)	16.6	11.3	11.4	7.16	5.81	7.62	5.85	5.42	5.62
	van der Waals $\triangle E_{int}$ (kcal/mol)	10.4	11.1	20.2	7.18	5.84	5.41	7.85	5.27	5.19

Table 9: Comparison of the optimized parameters used in all the three experiments in the prediction of $\triangle E_{int}$ values

	Type	Property	LCP			BCP			Combined		
			LR	N	HN	LR	N	HN	LR	N	HN
Exp 1	aldehydes	Covalent $\triangle E_{int}$ (kcal/mol)	0.012	500	10	0.017	500	10	0.091	500	6
		van der Waals $\triangle E_{int}$ (kcal/mol)	0.012	500	10	0.048	1000	9	0.034	500	10
	ketones	Covalent $\triangle E_{int}$ (kcal/mol)	0.006	500	7	0.020	2500	5	0.006	1600	4
		van der Waals $\triangle E_{int}$ (kcal/mol)	0.022	500	7	0.053	900	7	0.006	500	7
	imides	Covalent $\triangle E_{int}$ (kcal/mol)	0.010	500	10	0.014	500	9	0.006	900	5
		van der Waals $\triangle E_{int}$ (kcal/mol)	0.027	500	15	0.062	700	9	0.031	500	15
	amides	Covalent $\triangle E_{int}$ (kcal/mol)	0.007	500	15	0.023	2500	16	0.020	500	4
		van der Waals $\triangle E_{int}$ (kcal/mol)	0.017	500	8	0.050	500	11	0.020	500	10
Exp 2		Covalent $\triangle E_{int}$ (kcal/mol)	0.011	500	5	0.024	2500	12	0.013	500	20
		van der Waals $\triangle E_{int}$ (kcal/mol)	0.012	500	8	0.026	1900	13	0.010	1500	5
Exp 3	aldehydes	Covalent $\triangle E_{int}$ (kcal/mol)	0.012	500	8	0.009	500	10	0.010	1700	10
		van der Waals $\triangle E_{int}$ (kcal/mol)	0.004	1200	7	0.045	2500	9	0.005	500	5
	ketones	Covalent $\triangle E_{int}$ (kcal/mol)	0.012	500	6	0.011	2100	4	0.006	500	10
		van der Waals $\triangle E_{int}$ (kcal/mol)	0.008	600	5	0.013	1000	5	0.004	500	6
	imides	Covalent $\triangle E_{int}$ (kcal/mol)	0.023	500	4	0.021	500	6	0.002	500	2
		van der Waals $\triangle E_{int}$ (kcal/mol)	0.001	500	4	0.027	500	5	0.001	800	4
	amides	Covalent $\triangle E_{int}$ (kcal/mol)	0.148	700	4	0.057	1300	6	0.001	500	2
		van der Waals $\triangle E_{int}$ (kcal/mol)	0.001	500	1	0.008	500	5	0.002	500	4

calculated for all the other methods. For example, MAEs of 1.76, 0.38, 3.41, 0.16, 3.44, 0.03, and 0.14 kcal/mol were obtained for the same systems using PBE, PBE+D2, revPBE, revPBE+D2, BLYP, BLYP+D2, and vdW-DF1 methods, respectively. In our study, on all 225 carbonyl compounds, we obtained MAEs for predicted van der Waals $\triangle E_{int}$ values of 4.80, 5.06, and 4.78 kcal/mol when we trained ANN with BCP, LCP, and combined datasets, respectively. In an another study, Bose and co-workers [288] employed ML to predict $\triangle E_{int}$ values in water clusters. In the case of water dimer and trimer $\triangle E_{int}$ values, the reported root mean square errors (RMSE) are 0.12 and 0.34 kcal/mol, respectively. Our optimal predictions for van der Waals $\triangle E_{int}$ values have RMSEs of 6.46, 6.53, and 6.17 kcal/mol when we trained ANN on BCP, LCP, and combined datasets, respectively.

Interpretation of Results

In this part of our investigation, we again found support for our initial hypothesis that different types of topological data are better-suited to predicting certain molecular properties. In this case, $\triangle E_{int}$ values, which are intuitively related to models of *electronic structure* (such as HOMO-LUMO gap, partial atomic charges, resonance etc...) are found to be more accurately predicted by LCP properties in comparison with BCP descriptors. As discussed in earlier studies [265], the Laplacian of electron density recovers both the Lewis-VSPER model of electron pairs and chemical reactivity through the concept of acid-base reaction. These discoveries are based on physical observables. Science is based on observable things. Chemists used molecular orbital models (HOMO and/or LUMO) to predict where the nucleophilic attack occurs. These are purely imaginary and are not measurable. But, in QTAIM, we can use charge concentrations and depletions in the VSCCs of the respective base and acid atoms to predict the positions of nucleophilic attack. In our study, we are using observable properties to train the model.

It is unclear why the combined data set was not best in predicting interaction energies.

We also found that the smaller values of learning rate coefficients provided better results. This tells us that the network learns better when we take smaller steps to adjust the weights. The performance of the network can be improved in several ways. **In general, the ANN will make better predictions if the network is trained well**. This can be achieved by increasing the size of the training set, or providing more information for each molecule in terms of its descriptors. **We only used topological data for a single atom**. We could augment our training data by including topological information from the nearby atoms, thus taking into account substituent effects directly, rather than indirectly. As we have seen, the *type of descriptor* is also important. There may be other types of topological data, which we have not considered, that result in more accurate predictions of other molecular properties. In addition, as reported by the majority of researchers in this area, non-topological descriptors are also useful. Many of these are not experimentally observable, or uniquely defined, such as HOMO-LUMO gaps, and they would not contribute to the second over-arching goal of this study: assessing the informational content of different types of topological data. The ANN performance can also be improved by changing different parameters of the network, such as learning-rate, number of hidden layers, number of hidden layer neurons, and number of epochs. In our study, we have improved the predicted accuracy of the model by changing these parameters of the network.

After comparing the performance of our model in terms of MAEs with other methods, such as first-principle electronic structure calculations and ML methods, we have noticed that our model is in reasonable agreement with other studies. In addition to predicting $\triangle E_{int}$ values, we are also interested in testing our model on larger systems, such as proteins which is an important goal of computational drug design. The application of *ab initio* electronic structure calculations utilizing wave function-based approaches to describe intermolecular interactions in biomolecular systems is challenging because the systems are large. Quantum-mechanical methods for calculating $\triangle E_{int}$ values are computationally expensive and scale poorly with respect to the system size, e.g., coupled cluster singles doubles (CCSD) scale as N^6, and MP2 scales as N^5, where N represents a measure of the system size [289, 290]. For that reason, there is a need to find an alternative approach that can provide excellent accuracy at lower computational cost and time. With this goal in mind, in the following section we extend our study to calculation of the $\triangle E_{int}$ values within the drug binding pocket of an enzyme.

3.4 Leveraging Machine-Learning with Small Molecules to Predict Ligand Interactions with Large Molecules

One of the essential properties of all living organisms is the process of metabolism, by which organic compounds are synthesized and broken down. The metabolism of a whole cell is a remarkably complex system, yet it can be divided into subsystems and pathways, which are comprised of multiple biochemical reactions that change one compound into another. Metabolic reactions are like any other chemical reaction, which involve consumption of substrates and production of products. Many essential reactions in metabolism are too slow on their own, and to serve their biological roles these reactions need to be sped up by biological catalysts called enzymes. For example, the enzyme orotidine 5'-phosphate decarboxylase enhances the rate of the decarboxylation reaction of orotic acid by 10^{17} times [291], otherwise it would take 78 million years to complete in neutral aqueous solution at room temperature!

Enzymes speed up the reaction rate by reducing the activation energy, selectively binding to the substrate, and modifying it into products [292]. Initially, the substrate binds to the enzyme at the active site and starts the catalytic process. The active site is the specific region of an enzyme which communicates directly with the substrate. In regular enzymatic reactions, substrates transform to products after passing through the transition state. During this process, the electron distribution in various chemical bonds of the substrate molecule are altered in a way that eventually leads to the formation of products. There are two commonly-used models for enzyme-substrate interaction: the lock-and-key model and the induced-fit model. In the lock-and-key model, the substrate exactly fits in the active site of an enzyme [293]. Whereas in induced-fit model, the confirmations of both the enzyme and substrate molecules alter upon binding of the substrate [294].

In the study of biochemical processes, understanding mechanism of enzyme-catalysed reactions is extremely important. This knowledge helps to develop new drugs and design novel protein catalysts. The proof-of-principle test of our ANN model for predicting interaction energies of reacting molecules involved small molecules, which did not require large amounts of computational time for even the most accurate *ab initio* method that we used (MP2). However, calculations on enzymes containing 16,000 or more atoms at the MP2 level of theory is computational prohibitive. In addition, anything approaching an "exact" calculation, such as Complete active space self-consistent field (CAS-SCF), is essentially impossible. For a "stretch-test" of the usefulness of our ANN model for predicting interaction energies of large molecules, based on small molecules, we have chosen to model the active site of the *E. coli* enzyme D-fructose-6-phosphate aldolase (FSA) [5], which catalyzes a nucleophilic addition reaction of a carbon nucleophile (ketone) to a carbon electrophile (aldehyde). Aldolases are lyases that typically catalyse one of the most fundamental reactions in organic chemistry: the addition of a keto donor to an aldehyde acceptor molecule, resulting in the formation of a new carbon-carbon bond. In biological systems, aldol condensation and cleavage reactions play crucial roles in sugar metabolic pathways such as glycolysis and gluconeogenesis. Glycolysis is a cytoplasmic pathway which extracts energy from glucose by splitting into three-carbon compounds called pyruvates, and generating energy.

The calculation of interaction energies for nucleophilic addition reactions within the binding pocket of an enzyme is computationally demanding. CPU time for a single interaction energy calculation within the binding pocket of an enzyme on our VOLTRON cluster (168 processors) can take up to 7 hours, or even days, depending on the size of the enzyme, the level of electronic structure theory used, and the degree of statistical sampling that is required for thermodynamic purposes. Since, we obtained promising results of MAEs of around 3.44 kcal/mol in predicting the interaction energies of nucleophilic addition reactions of small molecules, we attempted to apply the same methodology in predicting interaction energies within the binding pocket of an enzyme. The ANN model we have developed uses only very local properties that are experimentally accessible. In future, we can easily add transferable fragment densities available at pseudo-atom databases developed by Koritsanzsky *et al.* [119] to increase the speed.

In this investigation, the training data for our ANN model is limited to the topological properties of a *single atom*; the carbonyl carbon. Thus, in principle, the size of the enzyme is irrelevant to our model. The properties of the entire system are reflected in the charge distributions of the atoms directly involved in the reaction. This is related to the fundamental principle of DFT, in which any property of the system of interacting particles can be obtained as a functional of the ground state density. The interaction energies are calculated using the following method:

$$\triangle E_{interaction} = E_{pocket+reactants(CC+F^-)} - E_{pocket+CC} - E_{pocket+F^-} + E_{pocket} \tag{19}$$

where $E_{interaction}$, $E_{pocket+reactants}$, $E_{pocket+CC}$, $E_{pocket+F^-}$, and E_{pocket} are the interaction energy of the reactants in the binding pocket, total energies of the interacting reactants in the pocket, carbonyl compound in the pocket, nucleophile in the pocket, and just the pocket, respectively.

The molecule 3-hydroxypropanal that we are interested in studying as the substrate in the binding pocket of the FSA enzyme is shown Figure 18. This molecule was used as a substrate since this class of enzymes are highly selective catalysts which speed up only specific reactions [295]. The topological properties of electron density of 3-hydroxypropanal in the binding pocket of an enzyme were computed in terms of bond critical points and Laplacian critical points in the VSCC of the carbonyl carbon. As discussed in Chapter 2 and Section 3.1, this critical point data was used as test data for predicting the interaction energies of fluoride ion and the carbonyl carbon of 3-hydroxypropanal within the binding pocket of the FSA enzyme.

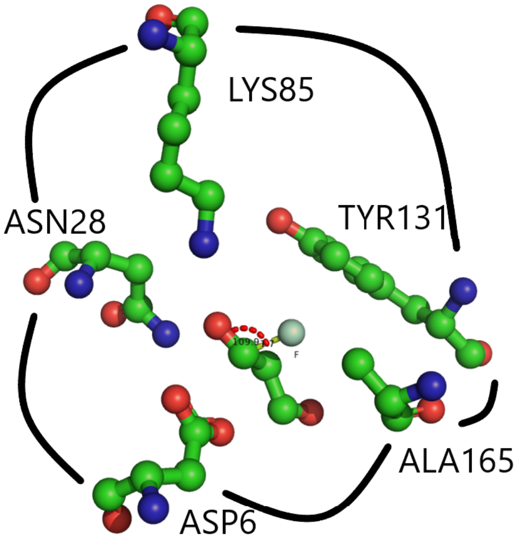


Figure 18: The binding pocket of FSA with substrate 3-hydroxypropanal (H-atoms omitted)

Computational Procedure Used for the Stretch-Test

The initial coordinates for the binding pocket of the chosen macromolecule, which in our stretch-test is the FSA enzyme, were taken from Protein Data Bank website (entry code: 1L6W) [5, 256]. In nature, this protein often has glycerol as a bound ligand, and which is the case in the single-crystal X-ray diffraction structure used in our stretch-test [296]. In our stretch-test model, the 3-hydroxypropanal ligand was placed in the FSA pocket by preserving the key ligand-receptor contacts where glycerol is found in the crystal structure. For example, one hydroxyl group in glycerol hydrogen-bonds to residues Asn 28 and Asp 6.

We maintained the same bonding between the hydroxyl group of 3-hydroxypropanal and these active site residues as a starting geometry similar to glycerol in the pocket.

In equation 19, the energy E_{pocket} was obtained by performing geometry optimization of the binding pocket while keeping the pocket cluster fixed in the experimental arrangement. The energy $E_{pocket+CC}$ was obtained by performing a complete geometry optimization of the 3-hydroxypropanal within the fixed binding pocket. During this calculation, the 3-hydroxypropanal was allowed to move, while the rest of the pocket molecules were fixed. Because, a tradeoff between accuracy and computational cost is unavoidable, we performed calculations for obtaining interaction energies by neglecting the conformational flexibility of the binding pocket of protein. Typically, interaction energies calculated with such an approach have average deviations from experiment of 2 kcal/mol or more [297, 298]. $E_{pocket+F^-}$ was obtained following the same procedure, i.e., by placing nucleophile in the pocket and allowing it to move, while the active-site pocket cluster was fixed. The energy $E_{pocket+reactants}$ was obtained by performing geometry optimization of the reactants within the pocket. During this calculation, the reactants were allowed to move, while the pocket geometry was fixed. All these calculations were performed at the same B3LYP/6-31+G* level of theory as previously mentioned in Section 3.3.

Stretch-Test Results

To explore whether the ANN model that was developed for predicting small-molecule interaction energies can be “stretched” and used for estimation/prediction of interaction energies of arbitrarily large molecules, instead of computing the interaction energy of the substrate reacting with the nucleophile within the binding pocket of an enzyme using a very expensive and time-consuming *ab initio* quantum mechanics, we employ our previously described ANN. In this stretch-test, the ANN model is tested on critical point descriptors generated when the carbonyl-containing 3-hydroxypropanal is located inside the enzyme pocket. Although the descriptors of the active site cluster are calculated via *ab initio* methods, the ANN greatly speeds up the interaction energy predictions. Quantum-mechanical calculations of such energies are extremely computationally intensive jobs for most biological systems, with 10-100 thousand atoms present in a typical enzyme. Nucleic acids, which can also be drug targets, can have billions of atoms in one chromosome.

Just as the energy changes in chemistry are only a small fractions of the total energy of the system, so the changes induced by the interactions between atoms in the charge density are only small ‘ripples’ in the total density. However, these small changes are amplified and made evident in the Laplacian of the charge density which can be extracted in terms of the Laplacian charge density critical points. The Laplacian distribution of 3-hydroxypropanal inside the binding pocket of the enzyme is shown in Figure 19. We can observe that most of the bonding charge concentration is on oxygen, in addition to its lone pair concentrations. It is also obvious that there is a substantial depletion in the regions above and below the carbon. A nucleophile can find facile reaction, attacking the carbon from above the molecular plane. The position of the (3, +1) critical point with respect to the C=O bond axis provides a possible angle of attack for the nucleophile of 110.2°, which is the same angle of approach of a nucleophile to a carbonyl found in experiment [111] and it can be observed in Figure 20.

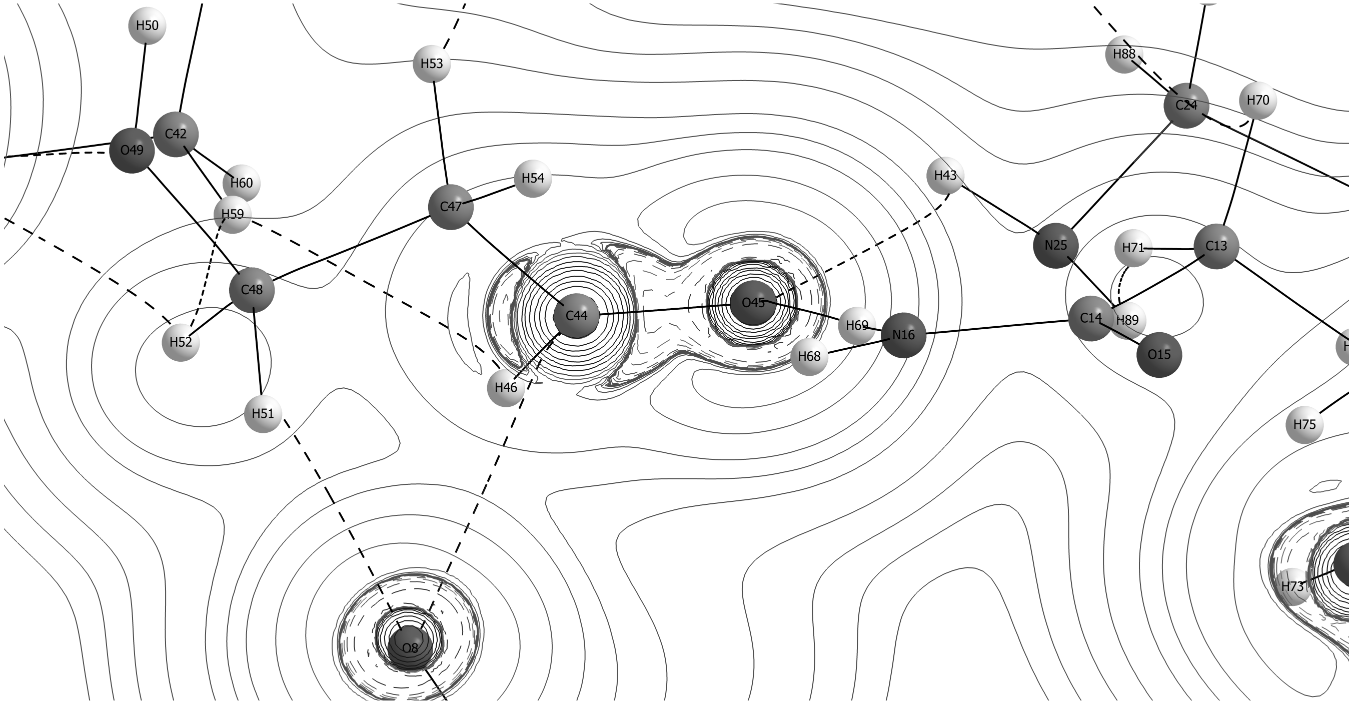


Figure 19: A contour diagram of the Laplacian distribution of 3-hydroxypropanal within the binding pocket of FSA enzyme. The dashed (solid) contour lines denote regions of charge concentration (depletion). Starting at a zero contour, contour values change in steps of ±2x10^n, ±4x10^n, and ±8x10^n with n beginning at -3 and increasing in steps of one. The more-or-less linear lines are atomic interaction lines; with solid paths indicating covalent bonds, and dashed paths indicating van der Waals interactions.

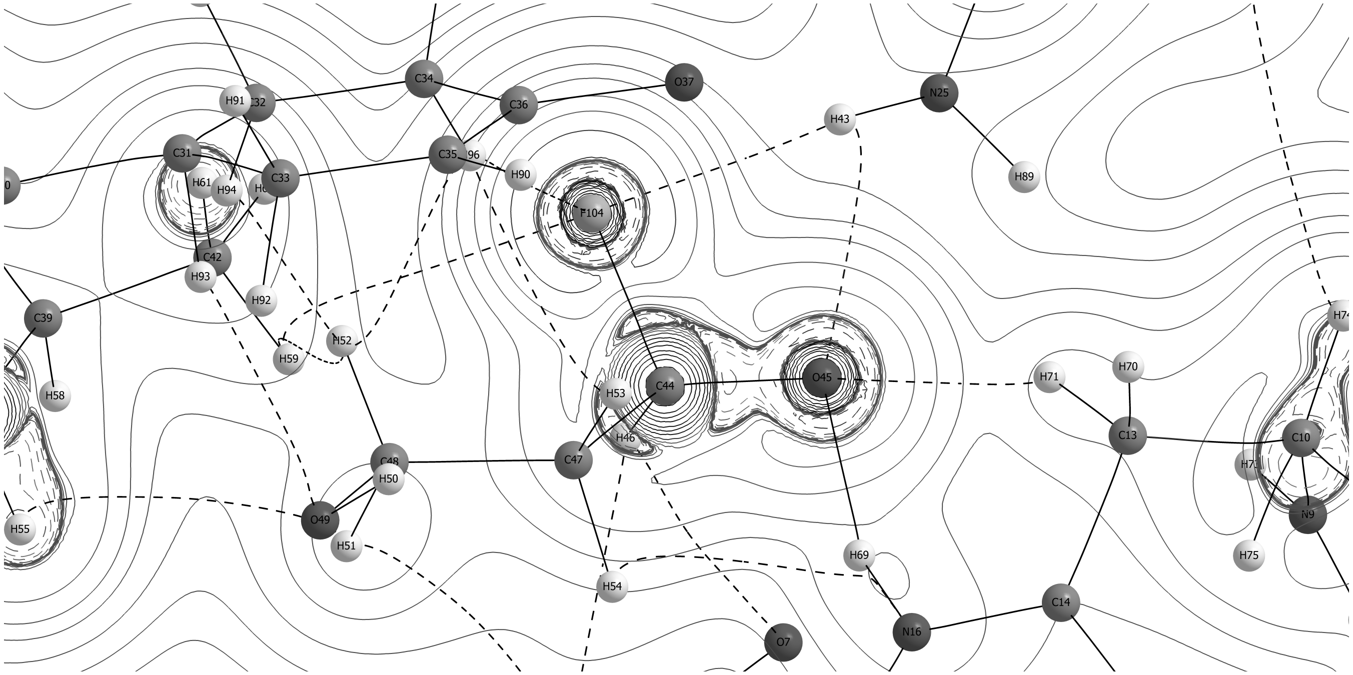


Figure 20: A contour diagram of the Laplacian distribution of covalent interaction between 3-hydroxypropanal and fluoride ion within the binding pocket of FSA enzyme. The fluoride ion approaches C=O bond axis at angle of 110.2°. The dashed (solid) contour lines denote regions of charge concentration (depletion). Starting at a zero contour, contour values change in steps of $\pm 2\text{x}10^n$, $\pm 4\text{x}10^n$, and $\pm 8\text{x}10^n$ with n beginning at -3 and increasing in steps of one. The more-or-less linear lines are atomic interaction lines; with solid paths indicating covalent bonds, and dashed paths indicating van der Waals interactions.

For calculating interaction energy ($\triangle E_{int}$) values between 3-hydroxypropanal and fluoride ion inside the binding pocket of the FSA enzyme, we used a similar starting geometry parameters as mentioned in Section 3.3. After performing the geometry optimizations, for the covalent interaction, the nucleophile approached carbonyl C at an angle of around 110.2° and at a C-F distance of 1.6 Å as shown Figure 20. In Table 10 we compare *ab initio* calculated covalent interaction energy and the ANN predicted energy when we used the three kinds of datasets (BCP, LCP, and combined) as mentioned in Section 2.4. The predicted values were obtained when we used optimized parameters of the network. These results tells us that the ANN trained on LCP data produced smaller errors when compared to the other two datasets.

Table 10: Comparison of covalent $\triangle E_{int}$ in the binding pocket of the FSA enzyme

DFT/B3LYP/6-31+G*	ANN predicted (kcal/mol)		
ab initio calculated (kcal/mol)	BCP	LCP	Combined
-22.7	-11.9	-19.7	-14.9

For calculating the van der Waals $\triangle E_{int}$ values, again we used similar starting geometry parameters used in Section 3.3. During this geometry optimization, instead of interacting with the ligand carbonyl, the fluoride ion nucleophile approached the carbonyl carbon of second residue (labeled as residue-2, Figure 21) of the pocket cluster. This kind of interactions has significance in the drug discovery process. Many enzymes modify the mechanism by which a reaction continues by making covalent bonds to the ligand molecules. The chemically reactive groups on the surface of the binding pocket of an enzyme's active site are often directly involved in converting ligands to product molecules. There are various examples of enzyme-catalyzed reactions that undergo mechanisms involving the formation of a covalent intermediate between the enzyme and the ligand molecule. One familiar example is the formation of a Schiff base by the condensation of an amine with a carbonyl. As such, we decided to design new calculations that would test the applicability of our model given these new constraints.

Cluster Used As A Model of the Binding Pocket

The binding pocket model contains five residues and these residues are labeled as residue-1 through residue-5 (Figure 21). Here, we considered the whole cluster as a carbonyl compound and tried to predict the interaction energy between the cluster and the fluoride ion. All the residues contain carbonyl groups, however we chose the C=O group of residue-2 to model the nucleophilic addition reaction, since this group is involved in interactions with the ligand molecule. Before performing geometry optimizations, hydrogen atoms were added to cap the amino acid fragments in the active site cluster. van der Waals $\triangle E_{int}$ values between fluoride ion and carbonyl C of the pocket were calculated using the equation 20:

$$\triangle E_{interaction} = E_{cluster+F^-} - E_{cluster} - E_{F^-} \tag{20}$$

where $E_{cluster+F^-}$, $E_{cluster}$, and E_{F^-} are total energies of the interacting (cluster+F$^-$ complex) and non-interacting reactants, cluster, and nucleophile (fluoride ion), respectively.

The energy of the cluster, $E_{cluster}$, was obtained from the geometry optimization of the cluster. These calculations were carried out by keeping the whole cluster frozen except for the C=O bond of residue-2. In order to obtain the total energy of the complex, $E_{cluster+F^-}$, geometry optimization was performed with the whole cluster fixed except for carbon, oxygen, and fluoride ion. The LCPs and BCPs of the carbonyl carbon were computed and used as a test dataset for predicting the van der Waals $\triangle E_{int}$ values. The ANN predicted energies as compared with the calculated ones are shown in Table 11. These results show that the BCP data is better for predicting van der Waals $\triangle E_{int}$ values and it is hard to make any conclusions on this. Although, B3LYP is not accurate for calculating these energies, we are not trying to predict these energies using DFT. In our study, we are trying to predict these energies using ML trained on charge density descriptors. Future research could investigate whether charge density descriptors derived from functionals that are better in predicting van der Waals energies provide a better training data for ML.

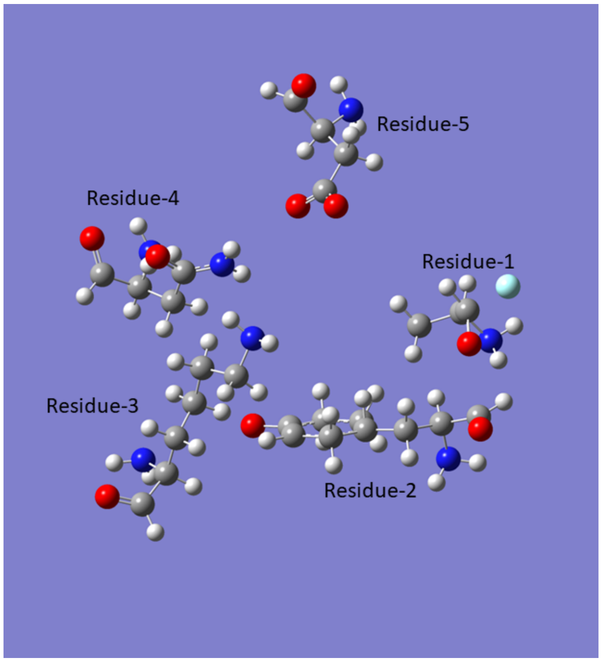


Figure 21: Labeling scheme used for binding pocket wall

Table 11: Comparison of van der Waals $\triangle E_{int}$ values between FSA binding pocket and the nucleophile

DFT/B3LYP/6-31+G*	ANN predicted (kcal/mol)		
ab initio calculated (kcal/mol)	BCP	LCP	Combined
-27.6	-38.6	-39.3	-42.7

Constructing Molecular Charge Densities From Databases

The QTAIM descriptors used here are often determined by experimentally via X-ray diffraction techniques, since there are high-resolution electron charge density studies on numerous crystal structures, and Bader's QTAIM theory is often applied to extract important chemical information. This has grown into a new area of research termed "Quantum Crystallography" [299, 300]. The term "quantum crystallography" was first introduced by Massa, Huang, and Karle in 1995 for methods that take advantage of "crystallographic information to enhance quantum-mechanical calculations and the information derived from them" [301]. There are conferences dedicated almost exlusively to this field, Sagamore and European Charge Density meetings. Koritsanszky and Coppens [302], in a review of applications of the topological analysis to high-resolution X-ray densities, have shown numerous examples where the BCP and LCP descriptors were obtained reliably from X-ray diffraction experiments. The experimental determination of charge density distribution is a challenging and complex task. For instance, it is not always possible to find a crystal good enough for such investigations. Furthermore, incomplete collection of X-ray scattering intensities, especially at high diffraction angle, can hinder the data refinement process. For these and other reasons, such as high-temperature conditions, diffraction data quality is frequently not sufficient to get reliable charge density results.

Among the possible models suggested in the literature, the multipolar expansion is by far the most selected one. According to Stewart [303], the total electron density could be projected onto atom-centered electron density approximations (pseudoatoms). Later, the development of the popular Hansen-Coppens (HC) formalism [117] offered highly effective tools to extract accurate charge density distributions from high-resolution experimental X-ray and electron diffraction data. In the Hansen-Coppens formalism, the atomic electron density is divided into three components

$$\rho_{\kappa}(r) = P_c\rho_c(r) + P_v\kappa^3\rho_v(\kappa r) + \sum_{l=0}^{l_{max}} \kappa^{'3} R_l(\kappa^{'} r) \sum_{m=0}^{l} P_{lm\pm} d_{lm\pm}(\theta,\phi) \tag{21}$$

where ρ_c and ρ_v are the spherically-averaged core and valence electron densities, respectively. The parameter κ represents the contraction or expansion of the spherical valence shell. The last part indicates the aspherical contributions to the valence density (deformation density). The parameter $\kappa^{'}$ represents the contraction or expansion of the deformation functions.

The valence deformation density is expanded in density functions made up of a radial part $R_l(\kappa' r)$ and the density-normalized spherical harmonic functions $d_{lm}(\theta, \phi)$.

Due to systematic experimental errors, constraints in multipole pseudoatom model, or lack of precise “phase angle” for scattered radiation, as well as large unpredictability of hydrogen atom positions from thermal motion, the confidence in experimental charge density might be compromised [304, 305, 306]. This is particularly problematic in the case of macromolecules, which, as discussed above, has great importance from biological point of view. Due to the much larger number of electrons, they are generally also harder to model using computational methods. Consequently, an effort was undertaken to determine whether it was possible to reproduce the charge density of larger systems based on the rapidly expanding number of high-quality data sets for an ever increasing database of small molecules, that could serve as a giving the starting point for assessing electrostatic properties and topological analysis of new molecules with limited data. One solution came in 1991, when Brock *et al.* [307] explored the transferability of atomic multipolar parameters (pseudoatom-based models) *between* various molecules. Indeed, the impetus for this transferability of models of atoms, and fragments of molecules, is rooted in the very nature of organic chemistry, where functional groups are presumed to have largely transferable properties between molecules in a given class of compounds, such as carbonyl compounds, gave the idea that atoms under identical chemical conditions frequently do not differ much in terms of charge density description, when present in different molecules. These perceptions started the production of databases of aspherical atom models. At present, we have three established databases available. Volkov *et al.* [308, 309] developed a considerably more modern pseudoatom database, which can be used to generate accurate charge distributions, as well its Laplacian, from just the nuclear coordinates of an X-ray experiment. Figure 22 presents an example, where the total charge density of the 11-mer polypeptide cyclosporin A is accurately predicted by rapidly construction, starting only from nuclear coordinates and the atomic database. No *ab initio* calculation was required to obtain the Laplacian distribution for cyclosporin A [310]. The “holes in the VSCC ” of carbonyl carbons as defined by MacDougall and Bader [311] can also be seen, and these appear similar to the very accurate *ab initio* calculations we performed (Figure 1).

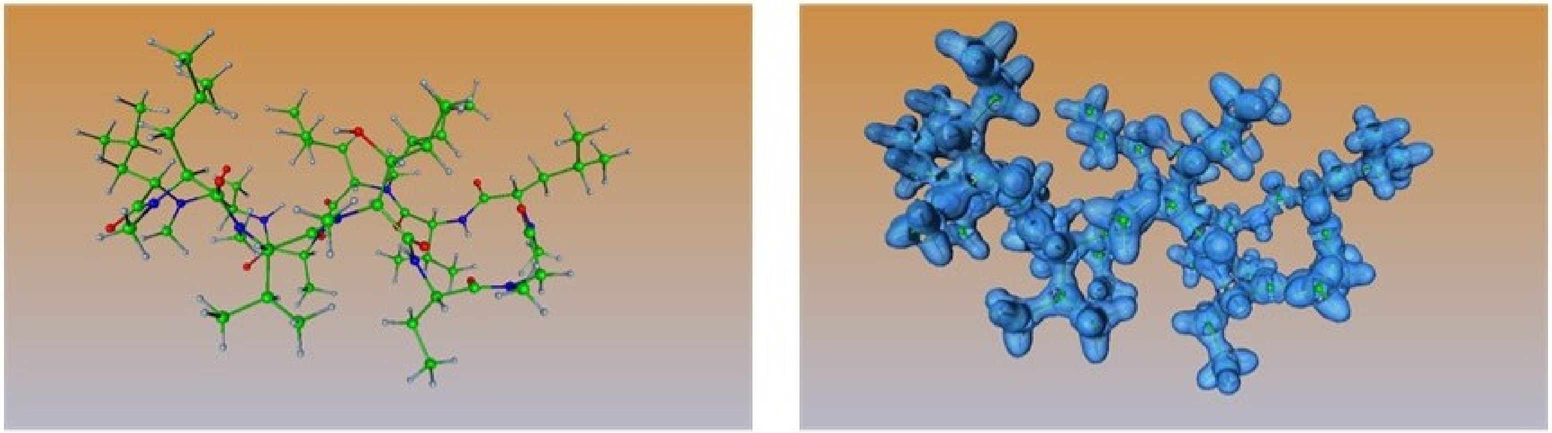


Figure 22: The total charge density of the 11-mer polypeptide cyclosporin A [310]

CHAPTER 4

Conclusions

The use of artificial neural networks (ANNs) is now well-established across science, particularly in chemistry. ANNs possess numerous advantages in terms of speed, convenience, and suitability for circumstances in which no sufficient analytical model is available. Use of Machine-Learning (ML) to enhance or replace conventional quantum-mechanical calculations has been rising in the last few years. The main advantage is that the computational cost of training a neural network is a fraction of the cost of DFT calculations, which are in turn post-Hartree-Fock methods such as MP2, MP3, and CCSD. With this in mind, we have proposed an ANN-based model to map a relationship between QTAIM descriptors and spectroscopic properties and interaction energies of carbonyl compounds in aldehydes, ketones, imides, and amides.

We have introduced a ML model, trained on a database of a small number of key electron density properties for few hundreds of carbonyl group containing molecules. Our model predicts spectroscopic properties and interaction energies with surprising accuracy, given the small set of training data per molecule, compared to the thousands of integrals that must be calculated in any *ab initio* method. In this study, ANNs trained on two kinds of descriptors to select the optimal ones for predicting each property. Electron density bond critical point descriptors and topological features of the Laplacian of the charge density distribution were used to train the model. We find that our ML approach is able to predict C=O stretching frequencies, NMR shifts, and interaction energies of various carbonyl compounds in a highly reliable manner. The excellent accuracy is combined with high computational efficiency, reducing the overall computation time by several orders of magnitude. With extensive experiments, for predicting C=O stretching frequencies and NMR shifts, our best models produced mean absolute percent errors of 0.52 and 1.45 when the models were trained on LCP and BCP datasets, respectively. We also found that smaller learning rates produced better results.

Our model produced mean absolute errors of 3.5 kcal/mol and 4.8 kcal/mol for predicting covalent interaction energies and van der Waals interaction energies when we trained our ANNs on combined datasets. Thus, the presented model could be efficiently employed for predicting spectroscopic properties and interaction energy values ($\triangle E_{int}$) that we are looking for. In this proof-of-principle investigation, we have studied spectroscopic properties and $\triangle E_{int}$ values of smaller chemical systems containing less than 30 atoms. However, much larger systems can be handled by our ML approach at little additional cost. Furthermore, we demonstrate the ability of ANNs to selected properties of macromolecules based only on the information contained in single atom. For a “stretch test” of performance of our ANN model, we obtained an absolute error of 3.1 kcal/mol while predicting the covalent interaction energy for a nucleophilic addition reaction between 3-hydroxypropanal and fluoride ion within the binding pocket of the D-Fructose-6-phosphate aldolase enzyme. This error can be reduced in future studies by increasing the size of the dataset or by including more descriptors of the neighboring atoms. The above findings are not restricted to the predictions of the studied properties, but can also be applied in a broader sense.

Finally, the present research confirms the fact that ANNs can be trained from a modest amount of data to accurately predict the required properties. We believe that we can improve this model further by increasing the size of the dataset or by including more descriptors. Current work is underway to expand and test our ANN approach for the prediction of NMR shifts, C=O stretching frequencies, and $\triangle E_{int}$ values with an expanded set of chemical descriptors, if necessary, to obtain results with greater accuracy. We believe that the proposed method can also be extended to other functional groups. In this study, as a proof-of-principle we have selected fluoride ion as a nucleophile. However in future studies, we can apply the same methodology for other nucleophiles by incorporating their LCP descriptors in the dataset. A natural extension to this work can be creation of a database containing both the BCP and LCP descriptors of atoms from various functional groups. Using the database, the required ANN models can be developed for the prediction of both spectroscopic properties and $\triangle E_{int}$ values.

BIBLIOGRAPHY

[1] Balabin, R. M.; Lomakina, E. I. Neural Network Approach to Quantum-Chemistry Data: Accurate Prediction of Density Functional Theory Energies. *J. Chem. Phys.* **2009**, *131*, 074104-074108. https://doi.org/10.1063/1.3206326.

[2] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; et al. *Gaussian 09*; Gaussian, Inc.: Wallingford, CT., 2009.

[3] AIMALL (Version 13.11.04), Keith, T. A., TK Gristmill Software, Overland Park KS, USA, (aim.tkgristmill.com). 2013.

[4] Volkov, A.; Koritsanszky, T.; Chodkiewicz, M.; King, H. F. On the Basis-Set Dependence of Local and Integrated Electron Density Properties: Application of a New Computer Program for Quantum-Chemical Density Analysis. *J. Comput. Chem.* **2009**, *30*, 1379-1391. https://doi.org/10.1002/jcc.21160.

[5] Thorell, S.; Schurmann, M.; Sprenger, G. A.; Schneider, G. Crystal Structure of Decameric Fructose-6-Phosphate Aldolase from *Escherichia coli* Reveals Inter-Subunit Helix Swapping as a Structural Basis for Assembly Differences in the Transaldolase Family. *J. Mol. Biol.* **2002**, *319*, 161-171. https://doi.org/10.1016/S0022-2836(02)00258-9.

[6] Johnson, A.H.; Maggiora, G. M. Concepts and Applications of Molecular Similarity. *J. Mol. Struct.* **1992**, *269*, 376-377. https://doi.org/10.1016/0022-2860(92)85011-5.

[7] Rupp, M.; Tkatchenko, A.; Müller, K. R.; Von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301. https://doi.org/10.1103/PhysRevLett.108.058301.

[8] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087-2096. `https://doi.org/10.1021/acs.jctc.5b00099`.

[9] Snyder, J. C.; Rupp, M.; Hansen, K.; Müller, K. R.; Burke, K. Finding Density Functionals with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 253002. `https://doi.org/10.1103/PhysRevLett.108.253002`.

[10] Pauling, L.; Corey, R. B. Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets. *Proc. Natl. Acad. Sci.* **1951**, *37*, 729-740. `https://doi.org/10.1073/pnas.37.11.729`.

[11] Pauling, L.; Corey, R. B.; Branson, H. R. The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. *Proc. Natl. Acad. Sci.* **1951**, *37*, 205-211. `https://doi.org/10.1073/pnas.37.4.205`.

[12] Watson, J. D.; Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature.* **1953**, *171*, 737-738. `https://www.nature.com/articles/248765a0`.

[13] Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53*, 5061-5084. `https://doi.org/10.1021/jm100112j`.

[14] Rupp, M.; Von Lilienfeld, O. A.; Burke, K. Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry. *J. Chem. Phy.* **2018**, *148*, 241401. `https://doi.org/10.1063/1.5043213`.

[15] Behler, J. Neural Network Potential-Energy Surfaces in Chemistry: A Tool for Large-Scale Simulations. *Phys. Chem. Chem. Phy*. **2011**, *13*, 17930-17955. https://doi.org/10.1039/c1cp21668f.

[16] Popelier P. L. A. *Atoms in Molecules. An Introduction*. Pearson Education: Harlow, Great Britain, 2000.

[17] Wikipedia. Machine Learning. URL (https://en.wikipedia.org/wiki/Machine_learning).

[18] Hofmann, T.; Schölkopf, B.; Smola, A. J. Kernel Methods in Machine Learning. *Ann. Statist*. **2008**, *36*, 1171-1220. https://doi.org/10.1214/009053607000000677.

[19] Mitchell B.O., J. B. O. Machine Learning Methods in Chemoinformatics. *WIREs Comput. Mol. Sci*. **2014**, *4*, 468-481. https://doi.org/10.1002/wcms.1183.

[20] Recommender Systems Handbook; **2011**. https://doi.org/10.1007/978-0-387-85820-3.

[21] Lemm, S.; Blankertz, B.; Dickhaus, T.; Müller, K. R. Introduction to Machine Learning for Brain Imaging. *Neuroimage* **2011**, *56*, 387-399. https://doi.org/10.1016/j.neuroimage.2010.11.004.

[22] Bredeche, N.; Shi, Z.; Zucker, J. D. Perceptual Learning and Abstraction in Machine Learning: An Application to Autonomous Robotics. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev*. **2006**, *36*, 172-181. https://doi.org/10.1109/TSMCC.2006.871139.

[23] Chau, M.; Chen, H. A Machine Learning Approach to Web Page Filtering Using Content and Structure Analysis. *Decis. Support Syst*. **2008**, *44*, 482-494. https://doi.org/10.1016/j.dss.2007.06.002.

[24] Guzella, T. S.; Caminhas, W. M. A Review of Machine Learning Approaches to Spam Filtering. *Expert Syst. Appl.* **2009**, *36*, 10206-10222. https://doi.org/10.1016/j.eswa.2009.02.037.

[25] Huang, C. L.; Chen, M. C.; Wang, C. J. Credit Scoring with a Data Mining Approach Based on Support Vector Machines. *Expert Syst. Appl.* **2007**, *33*, 847-856. https://doi.org/10.1016/j.eswa.2006.07.007.

[26] Chen, Y.; Mabu, S.; Shimada, K.; Hirasawa, K. A Genetic Network Programming with Learning Approach for Enhanced Stock Trading Model. *Expert Syst. Appl.* **2009**, *36*, 12537-12546. https://doi.org/10.1016/j.eswa.2009.05.054.

[27] Lavecchia, A. Machine-Learning Approaches in Drug Discovery: Methods and Applications. *Drug Discovery Today*. **2015**, *20*, 318-331. https://doi.org/10.1016/j.drudis.2014.10.012.

[28] Varnek, A.; Baskin, I. Machine Learning Methods for Property Prediction in Chemoinformatics: Quo Vadis? *J. Chem. Inf. and Model.* **2012**, *52*, 1413-1437. https://doi.org/10.1021/ci200409x.

[29] Deng, L.; Li, X. Machine Learning Paradigms for Speech Recognition: An Overview. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 1060-1089. https://doi.org/10.1109/TASL.2013.2244083.

[30] Wernick, M. N.; Yang, Y.; Brankov, J. G.; Yourganov, G.; Strother, S. C. Machine Learning in Medical Imaging. *IEEE Signal Process Mag.* **2014**, *27*, 25–38. https://ieeexplore.ieee.org/document/5484160.

[31] Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O. A.; Tkatchenko, A.; Muller, K. R. Assessment and Validation of Machine

Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput*. **2013**, *9*, 3404-3419.https://pubs.acs.org/doi/10.1021/ct400195d.

[32] Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep Learning for Computational Chemistry. *J. Comput. Chem*. **2017**, *38*, 1291-1307. https://doi.org/10.1002/jcc.24764.

[33] Haggarty, S. J.; Clemson, P. A.; Wong, J. C.; Schreiber, S. L. Mapping Chemical Space Using Molecular Descriptors and Chemical Genetics: Deacetylase Inhibitors. *Comb. Chem. High Throughput Screen*. **2004**, *7*, 669-676. https://doi.org/10.2174/1386207043328319.

[34] Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discov. Des*. **1998**, *2*, 339-153. https://doi.org/10.1007/0-306-46857-3_18.

[35] Von Lilienfeld, O. A. Quantum Machine Learning in Chemical Compound Space. *Angew. Chemie. Int. Ed.* **2018**, *57*, 4164-4169. https://doi.org/10.1002/anie.201709686.

[36] Kordis, J.; Gingerich, K. A. Atomization Energy and Standard Heat of Formation of Gaseous Diatomic Arsenic. *J. Chem. Eng. Data* **1973**, *18*, 135-136. https://doi.org/10.1021/je60057a023.

[37] Gu, J.; Leszczynski, J. Atomization Energies, Formation Enthalpies, Bond Dissociation Energies, and Adiabatic Electron Affinities of the PFn/PFn- Series, n = 1-6. *J. Phys. Chem. A* **1999**, *103*, 7856-7860. https://doi.org/10.1021/jp990792n.

[38] Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K. R.; Von Lilienfeld, O. A. Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New J. Phys*. **2013**, *15*, 095003. https://doi.org/10.1088/1367-2630/15/9/095003.

[39] Hu, L. H.; Wang, X. J.; Wong, L. H.; Chen, G. H. Combined First-Principles Calculation and Neural-Network Correction Approach for Heat of Formation. *J. Chem. Phys.* **2003**, *119*, 11501-11507. `https://doi.org/10.1063/1.1630951`.

[40] Rupp, M.; Ramakrishnan, R.; Von Lilienfeld, O. A. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *J. Phys. Chem. Lett.* **2015**, *6*, 3309-3313. `https://doi.org/10.1021/acs.jpclett.5b01456`.

[41] Wang, X. J.; Wong, L. H.; Hu, L. H.; Chan, C. Y.; Su, Z.; Chen, G. H. Improving the Accuracy of Density-Functional Theory Calculation: The Statistical Correction Approach. *J. Phys. Chem. A* **2004**, *108*, 8514-8525. `https://doi.org/10.1021/jp047263q`.

[42] Hemmateenejad, B.; Safarpour, M. A.; Miri, R.; Taghavi, F. Application of Ab Initio Theory to QSAR Study of 1,4-Dihydropyridine-Based Calcium Channel Blockers Using GA-MLR and PC-GA-ANN Procedures. *J. Comput. Chem.* **2004**, *25*, 1495-1503. `https://doi.org/10.1002/jcc.20066`.

[43] Bucinski, A.; Wnuk, M.; Gorynski, K.; Giza, A.; Kochanczyk, J.; Nowaczyk, A.; Baczek, T.; Nasal, A. Artificial Neural Networks Analysis Used to Evaluate the Molecular Interactions between Selected Drugs and Human α1-Acid Glycoprotein. *J. Pharm. Biomed. Anal.* **2009**, *50*, 591-596. `https://doi.org/10.1016/j.jpba.2008.11.005`.

[44] Zupan, J.; Gasteiger, J.; *Neural Networks in Chemistry and Drug Design*, 2nd ed., Wiley-VCH: New York, 1999.

[45] Devinyak, O.; Lesyk, R. 5-Year Trends in QSAR and Its Machine Learning Methods. *Curr. Comput. Aided-Drug Des.* **2016**, *12*, 265-271. `https://doi.org/10.2174/1573409912666160509121831`.

[46] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, Wiley-VCH: Weinheim, Germany, 2000. https://doi.org/10.1002/9783527613106.

[47] Rucker, G.; Rucker, C. Counts of All Walks as Atomic and Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 683-695. https://doi.org/10.1021/ci00015a005.

[48] Galvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. Charge Indexes. New Topological Descriptors. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 520-525. https://doi.org/10.1021/ci00019a008.

[49] Toropov, A. A.; Benfenati, E. QSAR Models for Daphnia Toxicity of Pesticides Based on Combinations of Topological Parameters of Molecular Structures. *Bioorganic Med. Chem.* **2006**, *14*, 2779-2788. https://doi.org/10.1016/j.bmc.2005.11.060.

[50] Khadikar, P. V.; Phadnis, A.; Shrivastava, A. QSAR Study on Toxicity to Aqueous Organisms Using the PI Index. *Bioorganic Med. Chem.* **2002**, *10*, 1181-1188. https://doi.org/10.1016/S0968-0896(01)00375-3.

[51] Agrawal, V. K.; Khadikar, P. V. QSAR Prediction of Toxicity of Nitrobenzenes. *Bioorganic Med. Chem.* **2001**, *9*, 3035-3040. https://doi.org/10.1016/S0968-0896(01)00211-5.

[52] Riahi, S.; Pourbasheer, E.; Dinarvand, R.; Ganjali, M. R.; Norouzi, P. QSAR Study of 2-(1-Propylpiperidin-4-yl)-1H-Benzimidazole-4-Carboxamide as PARP Inhibitors for Treatment of Cancer. *Chem. Biol. Drug Des.* **2008**, *72*, 575-584. https://doi.org/10.1111/j.1747-0285.2008.00739.x.

[53] Patra, J. C.; Chua, B. H. Artificial Neural Network-Based Drug Design for Diabetes Mellitus Using Flavonoids. *J. Comput. Chem.* **2011**, *32*, 555-567. https://doi.org/10.1002/jcc.21641.

[54] Hassoun, M. *Fundamentals of Artificial Neural Networks*; MIT Press: Cambridge, MA, 1995.

[55] Bashiri, M.; Farshbaf Geranmayeh, A. Tuning the Parameters of an Artificial Neural Network Using Central Composite Design and Genetic Algorithm. *Sci. Iran.* **2011**, *18*, 1600-1608. https://doi.org/10.1016/j.scient.2011.08.031.

[56] Ramirez-Galicia, G.; Garduno-Juarez, R.; Hemmateenejad, B.; Deeb, O.; Deciga-Campos, M.; Moctezuma-Eugenio, J. C. QSAR Study on the Antinociceptive Activity of Some Morphinans. *Chem. Biol. Drug Des.* **2007**, *70*, 53-64. https://doi.org/10.1111/j.1747-0285.2007.00530.x.

[57] Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Tansila, N.; Naenna, T.; Prachayasittikul, V. Prediction of GFP Spectral Properties Using Artificial Neural Network. *J. Comput. Chem.* **2007**, *28*, 1275-1289. https://doi.org/10.1002/jcc.20656.

[58] RECON. Version 5.5. Rensselaer Polytechnic Institute: Troy, New York, U.S.A. URL (http://www.drugmining.com/files/RECON/recondoc/WinRecon.html)

[59] Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; et al. Virtual Computational Chemistry Laboratory - Design and Description. *J. Comput. Aided. Mol. Des.* **2005**, *19*, 453-463. https://doi.org/10.1007/s10822-005-8694-y.

[60] Spartan'04, Wavefunction, Inc.: Irvine, CA, 2004.

[61] Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discovery Today*. **2018**, *23*, 1538-1546. https://doi.org/10.1016/j.drudis.2018.05.010.

[62] Bader, R. F. W. A Quantum Theory of Molecular Structure and Its Applications. *Chem. Rev.* **1991**, *91*, 893-928. `https://doi.org/10.1021/cr00005a013`.

[63] Hawe, G. I.; Alkorta, I.; Popelier, P. L. A. Prediction of the Basicities of Pyridines in the Gas Phase and in Aqueous Solution. *J. Chem. Inf. Model.* **2010**, *50*, 87-96. `https://doi.org/10.1021/ci900396k`.

[64] Alsberg, B. K.; Marchand-Geneste, N.; King, R. D. A New 3D Molecular Structure Representation Using Quantum Topology with Application to Structure-Property Relationships. *Chemom. Intell. Lab. Syst.* **2000**, *54*, 75-91. `https://doi.org/10.1016/S0169-7439(00)00101-5`.

[65] Manallack, D. The pKa Distribution of Drugs: Application to Drug Discovery. *Perspect. Med. Chem.* **2011**, *1*, 25-38. `https://doi.org/10.1177/1177391X0700100003`.

[66] Chaudry, U. A.; Popelier, P. L. A. Estimation of pKa Using Quantum Topological Molecular Similarity Descriptors: Application to Carboxylic Acids, Anilines and Phenols. *J. Org. Chem.* **2004**, *69*, 233-241. `https://doi.org/10.1021/jo0347415`.

[67] Buttingsrud, B.; Alsberg, B. K.; Astrand, P. O. An Investigation of Descriptors Based on the Critical Points in the Electron Density by Building Quantitative Structure-Property Relationships for Proton Chemical Shifts. *J. Mol. Struct. THEOCHEM* **2007**, *810*, 15-24. `https://doi.org/10.1016/j.theochem.2007.01.031`.

[68] Buttingsrud, B.; Ryeng, E.; King, R. D.; Alsberg, B. K. Representation of Molecular Structure Using Quantum Topology with Inductive Logic Programming in Structure-Activity Relationships. *J. Comput. Aided. Mol. Des.* **2006**, *20*, 361-373. `https://doi.org/10.1007/s10822-006-9058-y`.

[69] O'Brien, S. E.; Popelier, P. L. A. Quantum Molecular Similarity. 3. QTMS Descriptors. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 764-775. https://doi.org/10.1021/ci0004661.

[70] Roy, K.; Popelier, P. L. A. Exploring Predictive QSAR Models for Hepatocyte Toxicity of Phenols Using QTMS Descriptors. *Bioorganic Med. Chem. Lett*. **2008**, *18*, 2604-2609. https://doi.org/10.1016/j.bmcl.2008.03.035.

[71] Hawe, G. I.; Alkorta, I.; Popelier, P. L. A. Prediction of the Basicities of Pyridines in the Gas Phase and in Aqueous Solution. *J. Chem. Inf. Model*. **2010**, *50*, 87-96. https://doi.org/10.1021/ci900396k.

[72] Kar, S.; Harding, A. P.; Roy, K.; Popelier, P. L. A. QSAR with Quantum Topological Molecular Similarity Indices: Toxicity of Aromatic Aldehydes to Tetrahymena Pyriformis. *SAR QSAR Environ. Res*. **2010**, *21*, 149-168. https://doi.org/10.1080/10629360903568697.

[73] Roy, K.; Popelier, P. L. A. Exploring Predictive QSAR Models Using Quantum Topological Molecular Similarity (QTMS) Descriptors for Toxicity of Nitroaromatics to *Saccharomyces cerevisiae*. *QSAR Comb. Sci*. **2008**, *27*, 1006-1012. https://doi.org/10.1002/qsar.200810028.

[74] Thomas, L. H. The Calculation of Atomic Fields. *Math. Proc. Cambridge Philos. Soc*. **1927**, *23*, 542-548. https://doi.org/10.1017/S0305004100011683.

[75] Fermi, E. Eine Statistische Methode Zur Bestimmung Einiger Eigenschaften Des Atoms Und Ihre Anwendung Auf Die Theorie Des Periodischen Systems Der Elemente. *Zeitschrift fur Phys*. **1928**, *48*, 73-79. https://doi.org/10.1007/BF01351576.

[76] Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev*. **1964**, *136*, 864-871. https://doi.org/10.1103/PhysRev.136.B864.

[77] Kohn, W. Nobel Lecture: Electronic Structure of Matter - Wave Functions and Density Functional. *Rev. Mod. Phys.* **1999**, *71*, 1253-1266. https://doi.org/10.1103/revmodphys.71.1253.

[78] Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Molecular Physics*. **2017**, *115*, 2315-2372. https://doi.org/10.1080/00268976.2017.1333644.

[79] Zhao, Y.; Truhlar, D. G. Density Functionals with Broad Applicability in Chemistry. *Acc. Chem. Res.* **2008**, *41*, 157-167. https://doi.org/10.1021/ar700111a.

[80] Tirado-Rives, J.; Jorgensen, W. L. Performance of B3LYP Density Functional Methods for a Large Set of Organic Molecules. *J. Chem. Theory Comput.* **2008**, *4*, 297-306. https://doi.org/10.1021/ct700248k.

[81] Ren, Y.; Wolk, J. L.; Hoz, S. The Performance of Density Function Theory in Describing Gas-Phase SN2 Reactions at Saturated Nitrogen. *Int. J. Mass Spectrom.* **2002**, *221*, 59-65. https://doi.org/10.1016/S1387-3806(02)00894-1.

[82] Claes, L.; François, J. P.; Deleuze, M. S. Theoretical Study of the Internal Elimination Reactions of Xanthate Precursors. *J. Comput. Chem.* **2003**, *24*, 2023-2031. https://doi.org/10.1002/jcc.10358.

[83] Plumley, J. A.; Dannenberg, J. J. A Comparison of the Behavior of Functional/Basis Set Combinations for Hydrogen-Bonding in the Water Dimer with Emphasis on Basis Set Superposition Error. *J. Comput. Chem.* **2011**, *32*, 1519-1527. https://doi.org/10.1002/jcc.21729.

[84] Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, 1133-1138. `https://doi.org/10.1103/PhysRev.140.A1133`.

[85] Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38*, 3098-3100. `https://doi.org/10.1103/PhysRevA.38.3098`.

[86] Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37*, 785-789. `https://doi.org/10.1103/PhysRevB.37.785`.

[87] Perdew, J. P. Density-Functional Approximation for the Correlation Energy of the Inhomogeneous Electron Gas. *Phys. Rev. B* **1986**, *33*, 8822-8824. `https://doi.org/10.1103/PhysRevB.33.8822`.

[88] Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta–Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91*, 146401. `https://doi.org/10.1103/PhysRevLett.91.146401`.

[89] Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648-5652. `https://doi.org/10.1063/1.464913`.

[90] Becke, A. D. A New Mixing of Hartree-Fock and Local Density-Functional Theories. *J. Chem. Phys.* **1993**, *98*, 1372. `https://doi.org/10.1063/1.464304`.

[91] Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, Molecules, Solids, and Surfaces: Applications of the Generalized Gradient Approximation for Exchange and Correlation. *Phys. Rev. B* **1992**, *46*, 6671-6687. `https://doi.org/10.1103/PhysRevB.46.6671`.

[92] Becke, A. D. Perspective: Fifty Years of Density-Functional Theory in Chemical Physics. *J. Chem. Phys.* **2014**, *140*, 18A301. https://doi.org/10.1063/1.4869598.

[93] Bader, R. F. W. *Atoms in Molecules- A Quantum Theory*; Clandendon press: Oxford, UK, 1994.

[94] Bader, R. F. W.; Bayles, D. Properties of Atoms in Molecules: Group Additivity. *J. Phys. Chem. A* **2000**, *104*, 5579-5589. https://doi.org/10.1021/jp9943631.

[95] Bader, R. F. W.; Popelier, P. L. A.; Keith, T. A. Theoretical Definition of a Functional Group and the Molecular Orbital Paradigm. *Angew. Chem. Int. Ed. Engl.* **1994**, *33*, 620-631. https://doi.org/10.1002/anie.199406201.

[96] Coppens, P.; Hermansson, K. X-Ray Charge Densities and Chemical Bonding. *Phys. Today* **1998**, *51*, 66. https://doi.org/10.1063/1.882350.

[97] Gillespie, R. J.; Popelier, P. L. A. *Chemical Bonding and Molecular Geometry From Lewis to Electron Densities*; Oxford University Press: New York, 2001.

[98] Bader, R. F. W.; Essen, H. The Characterization of Atomic Interactions. *J. Chem. Phys.* **1984**, *80*, 1943. https://doi.org/10.1063/1.446956.

[99] Wiberg, K. B.; Bader, R. F.; Lau, C. D. Theoretical Analysis of Hydrocarbon Properties. 2. Additivity of Group Properties and the Origin of Strain Energy. *J. Am. Chem. Soc.* **1987**, *109*, 1001-1012. https://doi.org/10.1021/ja00238a005.

[100] Bader, R. F. W.; Slee, T. S.; Cremer, D.; Kraka, E. Description of Conjugation and Hyperconjugation in Terms of Electron Distributions. *J. Am. Chem. Soc.* **1983**, *105*, 5061-5068. https://doi.org/10.1021/ja00353a035.

[101] Popelier, P. L. A. Quantum Molecular Similarity. 1. BCP Space. *J. Phys. Chem. A* **1999**, *103*, 2883-2890. `https://doi.org/10.1021/jp984735q`.

[102] Lewis, G. N. The Atom and the Molecule. *J. Am. Chem. Soc*. **1916**, *38*, 762-785. `https://doi.org/10.1021/ja02261a002`.

[103] Gillespie, R. J.; Nyholm, R. S. Inorganic Stereochemistry. *Q. Rev. Chem. Soc*. **1957**, *11*, 339-380. `https://doi.org/10.1039/qr9571100339`.

[104] Gillespie, R. J. *Molecular Geometry*; Van Nostrand Reinhold: London, 1972.

[105] Bader, R. F. W.; MacDougall, P. J.; Lau, C. D. H. Bonded and Nonbonded Charge Concentrations and Their Relation to Molecular Geometry and Reactivity. *J. Am. Chem. Soc*. **1984**, *106*, 1594-1605. `https://doi.org/10.1021/ja00318a009`.

[106] Sagar, R. P.; Ku, A. C. T.; Smith, V. H.; Simas, A. M. The Laplacian of the Charge Density and Its Relationship to the Shell Structure of Atoms and Ions. *J. Chem. Phys*. **1988**, *88*, 4367. `https://doi.org/10.1063/1.453796`.

[107] Shi, Z.; Boyd, R. J. The Shell Structure of Atoms and the Laplacian of the Charge Density. *J. Chem. Phys*. **1988**, *88*, 4375. `https://doi.org/10.1063/1.454711`.

[108] Bader, R. F. W.; MacDougall, P. J. Toward a Theory of Chemical Reactivity Based on the Charge Density. *J. Am. Chem. Soc*. **1985**, *107*, 6788-6795. `https://doi.org/10.1021/ja00310a007`.

[109] MacDougall, P. J. The Laplacian of the Electronic Charge Distribution. Ph.D. Dissertation, McMaster University, Canada, 1989.

[110] Atkins, P.; De Paula, J.; Keeler, J. *Atkins' Physical Chemistry*; Oxford university press: Oxford, 2018.

[111] Burgi, H. B.; Dunitz, J. D.; Shefter, E. Chemical Reaction Paths. IV. Aspects of O ··· C = O Interactions in Crystals. *Acta Crystallogr. Sect. B* **1974**, *B30*, 1517-1527. https://doi.org/10.1107/s0567740874005188.

[112] Stalke, D. Meaningful Structural Descriptors from Charge Density. *Chem. Eur. J.* **2011**, *17*, 9264-9278. https://doi.org/10.1002/chem.201100615.

[113] Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 3186-3204. https://doi.org/10.1021/jm401411z.

[114] Coppens, P. *X-Ray Charge Densities and Chemical Bonding*; Oxford University Press: New York, 1997.

[115] Bader, R. F. W.; Beddall, P. M. Virial Field Relationship for Molecular Charge Distributions and the Spatial Partitioning of Molecular Properties. *J. Chem. Phys.* **1972**, *56*, 3320. https://doi.org/10.1063/1.1677699.

[116] Bader, R. F. W.; Henneker, W. H.; Cade, P. E. Molecular Charge Distributions and Chemical Binding. *J. Chem. Phys.* **1967**, *46*, 3341. https://doi.org/10.1063/1.1841222.

[117] Hansen, N. K.; Coppens, P. Testing Aspherical Atom Refinements on Small-molecule Data Sets. *Acta Crystallogr. Sect. A* **1978**, *A34*, 909-921. https://doi.org/10.1107/S0567739478001886.

[118] Stewart, R. F. Electron Population Analysis with Rigid Pseudoatoms. *Acta Crystallogr. Sect. A* **1976**, *A32*, 565-574. https://doi.org/10.1107/S056773947600123X.

[119] Volkov, A.; Li, X.; Koritsanszky, T.; Coppens, P. Ab Initio Quality Electrostatic Atomic and Molecular Properties Including Intermolecular Energies from a

Transferable Theoretical Pseudoatom Databank. *J. Phys. Chem. A* **2004**, *108*, 4283-4300. https://doi.org/10.1021/jp0379796.

[120] Carroll, F. A. *Perspectives on Structure and Mechanism in Organic Chemistry*; John Wiley & Sons: Hoboken, NJ, 2011.

[121] Vale, N.; Rulka, A.; Hudson, R.; Jones, M. *Biomedical Chemistry: Current Trends and Developments*; De Gruyter Open: Berlin, 2015. https://doi.org/10.1515/9783110468755.

[122] Lehninger, A. L.; Nelson, D. L.; Cox, M. M. *Lehninger Principles of Biochemistry*, 6th ed.; W. H. Freeman: New York, 2013.

[123] Wong, C. H.; Danielle, W. *Enzymes in Synthetic Organic Chemistry*; Elsevier: Oxford, U.K., 1994.

[124] Carey, F. A.; Sundberg, R. J. *Advanced Organic Chemistry: Part A*; Plenum Press: New York, 1990.

[125] Otera, J. *Modern Carbonyl Chemistry*; Wiley: Weinheim, 2000.

[126] Lewis, G.N. Acids and Bases. *Journal of the Franklin Institute*. **1938**, *226*, 293-313.

[127] Barr, E. S. Historical Survey of the Early Development of the Infrared Spectral Region. *Am. J. Phys*. **1960**, *28*, 42-54. https://doi.org/10.1119/1.1934975.

[128] Smith, B. C. *Fundamentals of Fourier Transform Infrared Spectroscopy*; CRC press: Boca Raton, Fl, 1996.

[129] Staurt, B. *Infrared Spectroscopy: Fundamentals and Applications*; John Wiley and Sons, Ltd.: West Sussex, England, 2004.

[130] Lindon, J. C.; Tranter, G. E.; Koppenaal, D. *Encyclopedia of spectroscopy and spectrometry*; Academic Press: Oxford, 2016.

[131] McDonald, R. S. Review: Infrared Spectrometry. *Anal. Chem.* **1986**, *58*, 1906-1925. https://doi.org/10.1021/ac00122a003.

[132] McKelvy, M. L.; Britt, T. R.; Davis, B. L.; Gillie, J. K.; Graves, F. B.; Lentz, L. A. Infrared spectroscopy. *Anal. Chem.* **1998**, *70*, 119-178.

[133] Peinado, A.; Hammond, J.; Scott, A. Development, Validation and Transfer of a Near Infrared Method to Determine in-Line the End Point of a Fluidised Drying Process for Commercial Production Batches of an Approved Oral Solid Dose Pharmaceutical Product. *J. Pharm. Biomed. Anal.* **2011**, *54*, 13-20. https://doi.org/10.1016/j.jpba.2010.07.036.

[134] Zandi-Atashbar, N.; Hemmateenejad, B.; Akhond, M. Determination of Amylose in Iranian Rice by Multivariate Calibration of the Surface Plasmon Resonance Spectra of Silver Nanoparticles. *Analyst* **2011**, *136*, 1760-1766. https://doi.org/10.1039/c0an00863j.

[135] Blanco, M.; Coello, J.; Eustaquio, A.; Itturriaga, H.; Maspoch, S. Development and Validation of Methods for the Determination of Miokamycin in Various Pharmaceutical Preparations by Use of near Infrared Reflectance Spectroscopy. *Analyst* **1999**, *124*, 1089-1092. https://doi.org/10.1039/a901774g.

[136] Barer, R.; Cole, A. R. H.; Thompson, H. W. Infra-Red Spectroscopy with the Reflecting Microscope in Physics, Chemistry and Biology. *Nature*. **1949**, *163*, 198-201. https://doi.org/10.1038/163198a0.

[137] Cascant, M. M.; Rubio, S.; Gallello, G.; Pastor, A.; Garrigues, S.; Guardia, M. de la. Burned Bones Forensic Investigations Employing near Infrared Spectroscopy. *Vib. Spectrosc.* **2017**, *90*, 21-30. https://doi.org/10.1016/j.vibspec.2017.02.005.

[138] Kondepati, V. R.; Keese, M.; Mueller, R.; Manegold, B. C.; Backhaus, J. Application of Near-Infrared Spectroscopy for the Diagnosis of Colorectal Cancer in Resected Human Tissue Specimens. *Vib. Spectrosc*. **2007**, *44*, 236-242. https://doi.org/10.1016/j.vibspec.2006.12.001.

[139] Backhaus, J.; Mueller, R.; Formanski, N.; Szlama, N.; Meerpohl, H. G.; Eidt, M.; Bugert, P. Diagnosis of Breast Cancer with Infrared Spectroscopy from Serum Samples. *Vib. Spectrosc*. **2010**, *52*, 173-177. https://doi.org/10.1016/j.vibspec.2010.01.013.

[140] Mobaraki, N.; Hemmateenejad, B. Structural Characterization of Carbonyl Compounds by IR Spectroscopy and Chemometrics Data Analysis. *Chemom. Intell. Lab. Syst*. **2011**, *109*, 171-177. https://doi.org/10.1016/j.chemolab.2011.08.011.

[141] Wilson, E.B.; Decius, J.C; Cross, P.C. *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*; Dover: New York, 1980.

[142] Merrick, J. P.; Moran, D.; Radom, L. An Evaluation of Harmonic Vibrational Frequency Scale Factors. *J. Phys. Chem. A* **2007**, *111*, 11683-11700. https://doi.org/10.1021/jp073974n.

[143] Pople, J. A.; Scott, A. P.; Wong, M. W.; Radom, L. Scaling Factors for Obtaining Fundamental Vibrational Frequencies and Zero-Point Energies from HF/6–31G* and MP2/6–31G* Harmonic Frequencies. *Isr. J. Chem*. **1993**, *33*, 345. https://doi.org/10.1002/ijch.199300041.

[144] Halls, M. D.; Velkovski, J.; Schlegel, H. B. Harmonic Frequency Scaling Factors for Hartree-Fock, S-VWN, B-LYP, B3-LYP, B3-PW91 and MP2 with the Sadlej

PVTZ Electric Property Basis Set. *Theor. Chem. Acc.* **2001**, *105*, 413-421. https://doi.org/10.1007/s002140000204.

[145] Alecu, I. M.; Zheng, J.; Zhao, Y.; Truhlar, D. G. Computational Thermochemistry: Scale Factor Databases and Scale Factors for Vibrational Frequencies Obtained from Electronic Model Chemistries. *J. Chem. Theory Comput.* **2010**, *6*, 2872-2887. https://doi.org/10.1021/ct100326h.

[146] Otwinowski, Z.; Borek, D.; Majewski, W.; Minor, W. Multiparametric Scaling of Diffraction Intensities. *Acta Cryst.* **2003**, *A59*, 228-234. https://doi.org/10.1107/S0108767303005488.

[147] Allen, A. E. A.; Payne, M. C.; Cole, D. J. Harmonic Force Constants for Molecular Mechanics Force Fields via Hessian Matrix Projection. *J. Chem. Theory Comput.* **2018**, *14*, 274-281. https://doi.org/10.1021/acs.jctc.7b00785.

[148] Rabi, I. I.; Zacharias, J. R.; Millman, S.; Kusch, P. A New Method of Measuring Nuclear Magnetic Moment. *Phys. Rev.* **1938**, *53*, 318. https://doi.org/10.1103/PhysRev.53.318.

[149] Purcell, E. M.; Torrey, H. C.; Pound, R. V. Resonance Absorption by Nuclear Magnetic Moments in a Solid. *Phys. Rev.* **1946**, *69*, 37. https://doi.org/10.1103/PhysRev.69.37.

[150] Markwick, P. R. L.; Malliavin, T.; Nilges, M. Structural Biology by NMR: Structure, Dynamics, and Interactions. *PLoS Comput. Biol.* **2008**, *4*, e1000168. https://doi.org/10.1371/journal.pcbi.1000168.

[151] Gunther H. *NMR Spectroscopy: Basic Principles, Concepts and Applications in Chemistry*; Wiley: Weinheim, 2013.

[152] Bagno, A.; Saielli, G. Computational NMR Spectroscopy: Reversing the Information Flow. *Theor. Chem. Acc.* **2007**, *117*, 603-619. https://doi.org/10.1007/s00214-006-0196-z.

[153] Stærk, D.; Chapagain, B. P.; Lindin, T.; Wiesman, Z.; Jaroszewski, J. W. Structural Analysis of Complex Saponins of Balanites Aegyptiaca by 800 MHz 1H NMR Spectroscopy. *Magn. Reson. Chem.* **2006**, *44*, 923-928. https://doi.org/10.1002/mrc.1879.

[154] Kaupp, M.; Buhl, M.; Malkin, V. G.; editors. *Calculation of NMR and EPR Parameters: Theory and Applications*; Wiley-VCH: Verlag, 2004.

[155] Jackman, L. M.; Sternhell, S. *Application of Nuclear Magnetic Resonance Spectroscopy in Organic Chemistry*; Pergamon Press: New York, 1969.

[156] Silverstein, R. M.; Bassler, G. C. *Spectrometric Identification of Organic Compounds*, 4th ed.; John Wiley & Sons, Inc.: New York, 1981.

[157] Mehring M. *Principles of High Resolution NMR in Solids*; Springer Verlang: Berlin, 1983.

[158] Jeannerat, D. Human- and Computer-Accessible 2D Correlation Data for a More Reliable Structure Determination of Organic Compounds. Future Roles of Researchers, Software Developers, Spectrometer Managers, Journal Editors, Reviewers, Publisher and Database Managers toward Artificial-Intelligence Analysis of NMR Spectra. *Magn. Reson. Chem.* **2017**, *55*, 7-14. https://doi.org/10.1002/mrc.4527.

[159] Meiler, J.; Sanli, E.; Junker, J.; Meusinger, R.; Lindel, T.; Will, M.; Maier, W.; Köck, M. Validation of Structural Proposals by Substructure Analysis and 13C NMR Chemical Shift Prediction. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 241-248. https://doi.org/10.1021/ci010294n.

[160] Elyashberg, M. E.; Blinov, K. A.; Williams, A. J. The Application of Empirical Methods Of13C NMR Chemical Shift Prediction as a Filter for Determining Possible Relative Stereochemistry. *Magn. Reson. Chem.* **2009**, *47*, 333-341. https://doi.org/10.1002/mrc.2396.

[161] Schnackenberg, L. K.; Beger, R. D. Whole-Molecule Calculation of Log P Based on Molar Volume, Hydrogen Bonds, and Simulated 13C NMR Spectra. *J. Chem. Inf. Model.* **2005**, *45*, 360-365. https://doi.org/10.1021/ci049643e.

[162] Kalchhauser, H.; Robien, W. CSEARCH: A Computer Program for Identification of Organic Compounds and Fully Automated Assignment of Carbon-13 Nuclear Magnetic Resonance Spectra. *J. Chem. Inf. Comput. Sci.* **1985**, *58*, 103-108. https://doi.org/10.1021/ci00046a010

[163] Bremser, W. Hose - a Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355-365. https://doi.org/10.1016/S0003-2670(01)83100-7.

[164] Satoh, H.; Koshino, H.; Uzawa, J.; Nakata, T. CAST/CNMR: Highly Accurate 13C NMR Chemical Shift Prediction System Considering Stereochemistry. *Tetrahedron.* **2003**, *59*, 4539-4547. https://doi.org/10.1016/S0040-4020(03)00662-8.

[165] Van Vleck, J. H. On the Anisotropy of Cubic Ferromagnetic Crystals. *Phys. Rev.* **1937**, *52*, 1178-. https://doi.org/10.1103/PhysRev.52.1178-1198.

[166] Ramsey, N. F. Theory of Molecular Hydrogen and Deuterium in Magnetic Fields. *Phys. Rev.* **1952**, *85*, 60-65. https://doi.org/10.1103/PhysRev.85.60.

[167] Bifulco, G.; Dambruoso, P.; Gomez-Paloma, L.; Riccio, R. Determination of Relative Configuration in Organic Compounds by NMR Spectroscopy and Computational Methods. *Chem. Rev.* **2007**, *107*, 3744-3779. https://doi.org/10.1021/cr030733c.

[168] Harper, J. K.; Doebbler, J. A.; Jacques, E.; Grant, D. M.; Von Dreele, R. B. A Combined Solid-State NMR and Synchrotron X-Ray Diffraction Powder Study on the Structure of the Antioxidant (+)-Catechin 4.5-Hydrate. *J. Am. Chem. Soc.* **2010**, *132*, 2928-2937. https://doi.org/10.1021/ja907671p.

[169] Maitra, N. T.; K. Burke, H.; Appel, E. K. U.; Leeuwen, R. V. Reviews in Modern Quantum Chemistry: A Celebration of the Contributions of R. G. Parr, edited by K. D. Sen, *World Scientific, Amsterdam*. **2002**, 1186-1225.

[170] Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115*, 2315-2372. https://doi.org/10.1080/00268976.2017.1333644.

[171] Tekarli, S. M.; Drummond, M. L.; Williams, T. G.; Cundari, T. R.; Wilson, A. K. Performance of Density Functional Theory for 3d Transition Metal-Containing Complexes: Utilization of the Correlation Consistent Basis Sets. *J. Phys. Chem. A* **2009**, *113*, 8607-8614. https://doi.org/10.1021/jp811503v.

[172] Tirado-Rives, J.; Jorgensen, W. L. Performance of B3LYP Density Functional Methods for a Large Set of Organic Molecules. *J. Chem. Theory Comput.* **2008**, *4*, 297-306. https://doi.org/10.1021/ct700248k.

[173] Schultz, N. E.; Zhao, Y.; Truhlar, D. G. Density Functional for Inorganometallic and Organometallic Chemistry. *J. Phys. Chem. A* **2005**, *109*, 11127-11143. https://doi.org/10.1021/jp0539223.

[174] Te Velde, G. T.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. Chemistry with ADF. *J. Comput. Chem.* **2001**, *22*, 931-967. https://doi.org/10.1002/jcc.1056.

[175] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A., Cheeseman, J. R.; Montgomery Jr. J.A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M. *Gaussian 03*, Revision C. 02; Gaussian. Inc.: Wallingford, CT, 2004.

[176] Barone, G.; Gomez-Paloma, L.; Duca, D.; Silvestri, A.; Riccio, R.; Bifulco, G. Structure Validation of Natural Products by Quantum-Mechanical GIAO Calculations of 13C NMR Chemical Shifts. *Chem. Eur. J.* **2002**, *8*, 3233-3239. https://doi.org/10.1002/1521-3765(20020715)8:14<3233::AID-CHEM3233>3.0.CO;2-0.

[177] Cimino, P.; Gomez-Paloma, L.; Duca, D.; Riccio, R.; Bifulco, G. Comparison of Different Theory Models and Basis Sets in the Calculation of 13C NMR Chemical Shifts of Natural Products. *Magn. Reson. Chem.* **2004**, *42*, S26-S33. https://doi.org/10.1002/mrc.1410.

[178] Saito, T.; Yamaji, T.; Hayamizu, K.; Yanagisawa, M.; Yamamoto, O.; Matsuyama, S.; Wasada, N.; Someno, K.; Kinugasa, S.; Tamura, T.; et al. SDBSWeb.

[179] Kalchhauser, H.; Robien, W. CSEARCH: A Computer Program for Identification of Organic Compounds and Fully Automated Assignment of Carbon-13 Nuclear Magnetic Resonance Spectra. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 103-108. https://doi.org/10.1021/ci00046a010.

[180] Steinbeck, C.; Kuhn, S. NMRShiftDB - Compound Identification and Structure Elucidation Support through a Free Community-Built Web Database. *Phytochemistry* **2004**, *65*, 2711-2717. https://doi.org/10.1016/j.phytochem.2004.08.027.

[181] Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31-36. https://doi.org/10.1021/ci00057a005.

[182] Meiler, J.; Meusinger, R.; Will, M. Fast Determination of 13C NMR Chemical Shifts Using Artificial Neural Networks. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1169-1176. https://doi.org/10.1021/ci000021c.

[183] Ivanciuc, O.; Rabine, J. P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J. P. 13C NMR Chemical Shift Prediction of the Sp3 Carbon Atoms in the α Position Relative to the Double Bond in Acyclic Alkenes. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 587-598. https://doi.org/10.1021/ci9601574.

[184] Svozil, D.; Kvasnicka, V.; Pospichal, J. Introduction to Multi-Layer Feed-Forward Neural Networks. *Chemom. Intell. Lab. Syst*. **1997**, *39*, 43-62. https://doi.org/10.1016/S0169-7439(97)00061-0.

[185] Kaur, J.; Brar, A. S. An Approach to Predict the 13C NMR Chemical Shifts of Acrylonitrile Copolymers Using Artificial Neural Network. *Eur. Polym. J*. **2007**, *43*, 156-163. https://doi.org/10.1016/j.eurpolymj.2006.09.014.

[186] Clouser, D. L,; Jurs, P. C. Simulation of the 13C Nuclear Magnetic Resonance Spectra of Trisaccharides Using Multiple Linear Regression Analysis and Neural Networks. *Carbohydrate research*. **1995** *271*, 65-77. https://doi.org/10.1016/0008-6215(95)00051-T

[187] Meiler, J.; Maier, W.; Will, M.; Meusinger, R. Using Neural Networks for 13C NMR Chemical Shift Prediction-Comparison with Traditional Methods. *J. Magn. Reson*. **2002**, *157*, 242-252. https://doi.org/10.1006/jmre.2002.2599.

[188] Heitler, W.; London, F. Wechselwirkung Neutraler Atome Und Homoopolare Bindung Nach Der Quantenmechanik. *Zeitschrift fur Phys*. **1927**, *44*, 455-472. https://doi.org/10.1007/BF01397394.

[189] Desiraju, G. R.; Vittal, J. J.; Ramanan, A. *Crystal Engineering: A Textbook*; World Scientific: Singapore, 2011.

[190] Guenther, F. H. *Neural Networks: Biological Models and Applications*; Oxford: International Encyclopedia of the Social & Behavioral Sciences. 2001, 10534-7.

[191] McCulloch, W. S.; Pitts, W. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bull. Math Biophys*. **1943**, *5*, 115-133.

[192] Kamruzzaman, S. M.; Jehad Sarkar, A. M. A New Data Mining Scheme Using Artificial Neural Networks. *Sensors* **2011**, *11*, 4622-4647. `https://doi.org/10.3390/s110504622`.

[193] Kononenko. I.; Kukar, M. *Machine Learning and Data Mining*; Horwood Publishing: Chichester, UK, 2007.

[194] Haykin, S. *Neural Networks and Learning Machines*; Pearson Prentice Hall, NJ, 2009.

[195] Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature*. **1986**, *323*, 533-536. `https://doi.org/10.1038/323533a0`

[196] Leondes, C. T. *Optimization techniques*; Academic Press: California, 1998.

[197] Moses, H.; Dorsey, E. R.; Matheson, D. H. M.; Thier, S. O. Financial Anatomy of Biomedical Research. *J. Am. Med. Assoc*. **2005**, *294*, 1333-1342. `https://doi.org/10.1001/jama.294.11.1333`.

[198] Lahana, R. How Many Leads from HTS? *Drug Discov. Today* **1999**, *4*, 447. `https://doi.org/10.1016/s1359-6446(99)01393-8`.

[199] Lobanov, V. Using Artificial Neural Networks to Drive Virtual Screening of Combinatorial Libraries. *Drug Discovery Today: BIOSILICO*. **2004**, *2*, 149-156. `https://doi.org/10.1016/S1741-8364(04)02402-3`.

[200] Joseph-Mccarthy, D. Computational Approaches to Structure-Based Ligand Design. *Pharmacol. Ther.* **1999**, *84*, 179-191. `https://doi.org/10.1016/S0163-7258(99)00031-5`.

[201] Martinez, J. D.; Parker, M. T.; Fultz, K. E.; Ignatenko, N. A.; Gerner, E. W.; Abraham. J.; Wiley, I. J. *Burger's Medicinal Chemistry and Drug Discovery*. 2003.

[202] Craig, J. C.; Duncan, I. B.; Hockley, D.; Grief, C.; Roberts, N. A.; Mills, J. S. Antiviral Properties of Ro 31-8959, an Inhibitor of Human Immunodeficiency Virus (HIV) Proteinase. *Antiviral Res.* **1991**, *16*, 295-305. `https://doi.org/10.1016/0166-3542(91)90045-S`.

[203] Kim, E. E.; Baker, C. T.; Dwyer, M. D.; Murcko, M. A.; Rac, B. G.; Tung, R. D.; Navia, M. A. Crystal Structure of HIV-1 Protease in Complex with VX-478, a Potent and Orally Bioavailable Inhibitor of the Enzyme. *J. Am. Chem. Soc.* **1995**, *117*, 1181-1182. `https://doi.org/10.1021/ja00108a056`.

[204] Hajduk, P. J.; Huth, J. R.; Tse, C. Predicting Protein Druggability. *Drug Discov. Today* **2005**, *10*, 1675-1682. `https://doi.org/10.1016/S1359-6446(05)03624-X`

[205] Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, Jr. E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-Based Archival File For Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535-42. `https://doi.org/10.1016/S0022-2836(77)80200-3`

[206] Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. Sect. B* **2016**, *72*, 171-179. https://doi.org/10.1107/S2052520616003954.

[207] Vyas, V. K.; Ukawala, R. D.; Ghate, M.; Chintha, C. Homology Modeling a Fast Tool for Drug Discovery: Current Perspectives. *Indian J. Pharm. Sci.* **2012**, *74*, 1. https://doi.org/10.4103/0250-474X.102537.

[208] Becker, O. M.; Dhanoa, D. S.; Marantz, Y.; Chen, D.; Shacham, S.; Cheruku, S.; Heifetz, A.; Mohanty, P.; Fichman, M.; Sharadendu, A.; et al. An Integrated in Silico 3D Model-Driven Discovery of a Novel, Potent, and Selective Amidosulfonamide 5-HT1A Agonist (PRX-00023) for the Treatment of Anxiety and Depression. *J. Med. Chem.* **2006**, *49*, 3116-3135. https://doi.org/10.1021/jm0508641.

[209] Budzik, B.; Garzya, V.; Shi, D.; Walker, G.; Woolley-Roberts, M.; Pardoe, J.; Lucas, A.; Tehan, B.; Rivero, R. A.; Langmead, C. J.; et al. Novel N-Substituted Benzimidazolones as Potent, Selective, CNS-Penetrant, and Orally Active M1 MAChR Agonists. *ACS Med. Chem. Lett.* **2010**, *1*, 244-248. https://doi.org/10.1021/ml100105x.

[210] Buchan, D. W. A.; Ward, S. M.; Lobley, A. E.; Nugent, T. C. O.; Bryson, K.; Jones, D. T. Protein Annotation and Modelling Servers at University College London. *Nucleic Acids Res.* **2010**, *38*, W563-W568. https://doi.org/10.1093/nar/gkq427.

[211] Marti-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. Comparative Protein Structure Modeling of Genes and Genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291-325. https://doi.org/10.1146/annurev.biophys.29.1.291.

[212] Laurie, R.; Alasdair, T.; Jackson, R. M. Methods for the Prediction of Protein-Ligand Binding Sites for Structure-Based Drug Design and Virtual Ligand Screening. *Curr. Protein and Pept. Sci.* **2006** *7*, 395-406. https://doi.org/10.2174/138920306778559386

[213] Henrich, S.; Salo-Ahen, O. M. H.; Huang, B.; Rippmann, F.; Cruciani, G.; Wade, R. C. Computational Approaches to Identifying and Characterizing Protein Binding Sites for Ligand Design. *J. Mol. Recognit.* **2010**, *23*, 209-219. https://doi.org/10.1002/jmr.984.

[214] Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. Lead Discovery Using Molecular Docking. *Curr. Opin. Chem. Biol.* **2002**, *6*, 439-446. https://doi.org/10.1016/S1367-5931(02)00339-3.

[215] Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269-288. https://doi.org/10.1016/0022-2836(82)90153-X.

[216] Koshland, D. E. Correlation of Structure and Function in Enzyme Action. *Science*. **1963**, *142*, 1533-1541. https://doi.org/10.1126/science.142.3599.1533.

[217] Hammes, G. G. Multiple Conformational Changes in Enzyme Catalysis. *Biochemistry*. **2002**, *41*, 8221-8228. https://doi.org/10.1021/bi0260839.

[218] Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects. *J. Med. Chem.* **2006**, *49*, 534-553. https://doi.org/10.1021/jm050540c.

[219] McGann, M. FRED Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2011**, *51*, 578-596. https://doi.org/10.1021/ci100436p.

[220] Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639-1662. https://doi.org/10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B.

[221] Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins Struct. Funct. Genet.* **2003**, *52*, 609-623. https://doi.org/10.1002/prot.10465.

[222] Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX Incremental Construction Algorithm for Protein- Ligand Docking. *Proteins Struct. Funct. Genet.* **1999**, *37*, 228-241. https://doi.org/10.1002/(SICI)1097-0134(19991101)37:2<228::AID-PROT8>3.0.CO;2-8.

[223] Broach, J. R.; Thorner, J. High-Throughput Screening for Drug Discovery. *Nature*. **1996**, *384*, 14-16.

[224] Lipinski, C. A.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3-25. https://doi.org/10.1016/S0169-409X(96)00423-1

[225] Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476-488. https://doi.org/10.1002/minf.201000061.

[226] Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**. https://doi.org/10.1021/ja00226a005.

[227] Randic, M. Representation of Molecular Graphs by Basic Graphs. *J. Chem. Inf. Comput. Sci*. **1992**, *110*, 5959-5967. https://doi.org/10.1021/ci00005a010.

[228] Schuur, J. H.; Selzer, P.; Gasteiger, J. The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 334-344. https://doi.org/10.1021/ci950164c.

[229] Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. MS-WHIM, New 3D Theoretical Descriptors Derived from Molecular Surface Properties: A Comparative 3D QSAR Study in a Series of Steroids. *J. Comput. Aided. Mol. Des*. **1997**, *11*, 79-92. https://doi.org/10.1023/A:1008079512289.

[230] Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. Deriving the 3D Structure of Organic Molecules from Their Infrared Spectra. *Vib. Spectrosc*. **1999**, *19*, 151-164. https://doi.org/10.1016/s0924-2031(99)00014-4.

[231] Ehrlich, P. Über Den Jetzigen Stand Der Salvarsantherapie Mit Besonderer Berücksichtigung Der Nebenwirkungen Und Deren Vermeidung. *The Collected Papers of Paul Ehrlich*. **1960**, 393-404. https://doi.org/10.1016/b978-0-08-009056-6.50042-7.

[232] Lin, S.K. Pharmacophore Perception, Development and Use in Drug Design. Edited by Osman F. Güner. *Molecules* **2000**, *5*, 987-989. https://doi.org/10.3390/50700987.

[233] Patel, Y.; Gillet, V. J.; Bravi, G.; Leach, A. R. A Comparison of the Pharmacophore Identification Programs: Catalyst, DISCO and GASP. *J. Comput. Aided. Mol. Des*. **2002**, *16*, 653-681. https://doi.org/10.1023/A:1021954728347.

[234] Bowman, A. L.; Nikolovska-Coleska, Z.; Zhong, H.; Wang, S.; Carlson, H. A. Small Molecule Inhibitors of the MDM2-P53 Interaction Discovered by Ensemble-Based Receptor Models. *J. Am. Chem. Soc.* **2007**, *129*, 12809-12814. https://doi.org/10.1021/ja073687x.

[235] Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *7*, 395-399. https://doi.org/10.1021/jm00334a001.

[236] Hansch, C.; Fujita, T. $\rho - \sigma - \pi$ Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616-1626. https://doi.org/10.1021/ja01062a035.

[237] Sirci, F.; Napolitano, F.; Pisonero-Vaquero, S.; Carrella, D.; Medina, D. L.; di Bernardo, D. Comparing Structural and Transcriptional Drug Networks Reveals Signatures of Drug Activity and Toxicity in Transcriptional Responses. *npj Syst. Biol. Appl.* **2017**, *3*, 1-12. https://doi.org/10.1038/s41540-017-0022-3.

[238] Gupta, S. Quantitative Structure-Activity Relationships of Renin Inhibitors. *Mini-Reviews Med. Chem.* **2005**, *3*, 315-321. https://doi.org/10.2174/1389557033488051.

[239] Kontogiorgis, C.; Hadjipavlou-Litina, D. Quantitative Structure - Activity Relationships (QSARs) of Thrombin Inhibitors: Review, Evaluation and Comparative Analysis. *Curr. Med. Chem.* **2005**, *10*, 525-577. https://doi.org/10.2174/0929867033457935.

[240] Hansch, C.; McClarin, J.; Klein, T.; Langridge, R. A Quantitative Structure-Activity Relationship and Molecular Graphics Study of Carbonic Anhydrase Inhibitors. *Mol. Pharmacol.* **1985**, *27*, 493-498.

[241] Zhang, L.; Tan, J.; Han, D.; Zhu, H. From Machine Learning to Deep Learning: Progress in Machine Intelligence for Rational Drug Discovery. *Drug Discov. Today* **2017**, *22*, 1680-1685. https://doi.org/10.1016/j.drudis.2017.08.010.

[242] Clark, D. E. What Has Computer-Aided Molecular Design Ever Done for Drug Discovery? *Exp. Opin. Drug Discov*. **2006**, *1*, 103-110. https://doi.org/10.1517/17460441.1.2.103.

[243] Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M. Novel Technologies for Virtual Screening. *Drug Discov. Today* **2004**, *9*, 27-34. https://doi.org/10.1016/S1359-6446(04)02939-3.

[244] Lavecchia, A.; Giovanni, C. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Curr. Med. Chem.* **2013**, *20*, 2839-2860. https://doi.org/10.2174/09298673113209990001.

[245] Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759-4767. https://doi.org/10.1021/jm001044l.

[246] Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *J. Chem. Inf. Model.* **2006**, *46*, 462-470. https://doi.org/10.1021/ci050348j.

[247] Fink, T.; Bruggesser, H.; Reymond, J. L. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons. *Angew. Chem. Int. Ed*. **2005**, *44*, 1504-1508. https://doi.org/10.1002/anie.200462457.

[248] Bergström, C. A. S. In Silico Predictions of Drug Solubility and Permeability: Two Rate-Limiting Barriers to Oral Drug Absorption. *Basic and Clin. Pharmacol.*

and Toxicol. **2005**, *96*, 156-161. https://doi.org/10.1111/j.1742-7843.2005.pto960303.x.

[249] Fagerberg, J. H.; Karlsson, E.; Ulander, J.; Hanisch, G.; Bergström, C. A. S. Computational Prediction of Drug Solubility in Fasted Simulated and Aspirated Human Intestinal Fluid. *Pharm. Res*. **2015**, *32*, 578-589. https://doi.org/10.1007/s11095-014-1487-z.

[250] Spartan 2010 Wavefunction, Inc., Irvine, C. A., 2010.

[251] Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XX. A Basis Set for Correlated Wave Functions. *J. Chem. Phys*. **1980**, *72*, 650. https://doi.org/10.1063/1.438955.

[252] Frisch, M. J.; Pople, J. A.; Binkley, J. S. Self-Consistent Molecular Orbital Methods 25. Supplementary Functions for Gaussian Basis Sets. *J. Chem. Phys*. **1984**, *80*, 3265. https://doi.org/10.1063/1.447079.

[253] Chandrasekhar, J.; Andrade, J. G.; von Rague Schleyer, P. Efficient and Accurate Calculation of Anion Proton Affinities. *J. Am. Chem. Soc*. **1981**, *103*, 5609-5612. https://doi.org/10.1021/ja00408a074.

[254] Zhao, Y.; Schultz, N. E.; Truhlar, D. G. Design of Density Functionals by Combining the Method of Constraint Satisfaction with Parametrization for Thermochemistry, Thermochemical Kinetics, and Noncovalent Interactions. *J. Chem. Theory Comput*. **2006**, *2*, 364-382. https://doi.org/10.1021/ct0502763.

[255] Clemente, D. A. Electron Distributions and the Chemical Bond. *Inorganica Chim. Acta*. **1983**, *73*, 145. https://doi.org/10.1016/s0020-1693(00)90840-5.

[256] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N. The Protein Data Bank, *Nucleic Acids Res.* **2000**, *28*, 235-242. https://doi.org/10.1093/nar/28.1.235.

[257] Keith, T. A. AIMAll (Version 17.01.25).

[258] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software. ACM SIGKDD Explor. Newsl. 2009. https://doi.org/10.1145/1656274.1656278.

[259] Hupf, E.; Olaru, M.; Raţ, C. I.; Fugel, M.; Hubschle, C. B.; Lork, E.; Grabowsky, S.; Mebs, S.; Beckmann, J. Mapping the Trajectory of Nucleophilic Substitution at Silicon Using a Peri-Substituted Acenaphthyl Scaffold. *Chem. Eur. J.* **2017**, *23*, 10568-10579. https://doi.org/10.1002/chem.201700992.

[260] Bikbaeva, Z. M.; Ivanov, D. M.; Novikov, A. S.; Ananyev, I. V.; Bokach, N. A.; Kukushkin, V. Y. Electrophilic-Nucleophilic Dualism of Nickel(II) toward Ni···I Noncovalent Interactions: Semicoordination of Iodine Centers via Electron Belt and Halogen Bonding via σ-Hole. *Inorg. Chem.* **2017**, *56*, 13562-13578. https://doi.org/10.1021/acs.inorgchem.7b02224.

[261] Bader, R. F. W.; Nguyen-Dang, T. T.; Tal, Y. A Topological Theory of Molecular Structure. *Rep. Prog. Phys.* **1981**, *44*, 893. https://doi.org/10.1088/0034-4885/44/8/002.

[262] Biegler-konig, F. W.; Bader, R. F. W.; Tang, T. H. Calculation of the Average Properties of Atoms in Molecules. II. *J. Comput. Chem.* **1982**, *3*, 317. https://doi.org/10.1002/jcc.540030306.

[263] Wiberg, K. B.; Bader, R. F.; Lau, C. D. Theoretical Analysis of Hydrocarbon Properties. 1. Bonds, Structures, Charge Concentrations, and Charge Relaxations. *J. Am. Chem. Soc.* **1987**, *109*, 985-1001. https://doi.org/10.1021/ja00238a004.

[264] Grabowski, S. J. Ab Initio Calculations on Conventional and Unconventional Hydrogen Bonds-Study of the Hydrogen Bond Strength. *J. Phys. Chem. A* **2001**, *105*, 10739-10746. https://doi.org/10.1021/jp011819h.

[265] Bader, R. F.; Gillespie, R. J.; MacDougall, P. J. A Physical Basis for the VSEPR Model of Molecular Geometry. *J. Am. Chem. Soc.* **1988**, *110*, 7329-7336. https://doi.org/10.1021/ja00230a009.

[266] Subedi, S.; Nakarmi, J. J. Vibrational Frequency Analysis of CH3Cl Molecule; Ab Initio Study. *Himal. Phys.* **2015**, *5*, 142-145. https://doi.org/10.3126/hj.v5i0.12902.

[267] Fleming, I. *Frontier Orbitals and Organic Chemical Reactions*; Wiley: London, 1976.

[268] Koopmans, T. Uber Die Zuordnung von Wellenfunktionen Und Eigenwerten Zu Den Einzelnen Elektronen Eines Atoms. *Physica* **1934**, *1*, 104-113. https://doi.org/10.1016/S0031-8914(34)90011-2.

[269] Fukui, K.; Yonezawa, T.; Shingu, H. A Molecular Orbital Theory of Reactivity in Aromatic Hydrocarbons. *J. Chem. Phys.* **1952**, *20*, 722. https://doi.org/10.1063/1.1700523.

[270] Aihara, J. I. Reduced HOMO-LUMO Gap as an Index of Kinetic Stability for Polycyclic Aromatic Hydrocarbons. *J. Phys. Chem. A* **1999**, *103*, 7487-7495. https://doi.org/10.1021/jp990092i.

[271] Aihara, J. I. Weighted HOMO-LUMO Energy Separation as an Index of Kinetic Stability for Fullerenes. *Theor. Chem. Acc.* **1999**, *102*, 134-138. https://doi.org/10.1007/s002140050483.

[272] Yoshida, M.; Aihara, J. ichi. Validity of the Weighted HOMO-LUMO Energy Separation as an Index of Kinetic Stability for Fullerenes with up to 120 Carbon Atoms. *Phys. Chem. Chem. Phys.* **1999**, *1*, 227-230. https://doi.org/10.1039/a807917j.

[273] Nummert, V.; Travnikova, O.; Vahur, S.; Leito, I.; Piirsalu, M.; Maemets, V.; Koppel, I.; Koppel, I. A. Influence of Substituents on the Infrared Stretching Frequencies of Carbonyl Group in Esters of Benzoic Acid. *J. Phys. Org. Chem.* **2006**, *19*, 654-663. https://doi.org/10.1002/poc.1112.

[274] Nummert, V.; Piirsalu, M.; Maemets, V.; Vahur, S.; Koppel, I. A. Effect of Ortho Substituents on Carbonyl Carbon 13C NMR Chemical Shifts in Substituted Phenyl Benzoates. *J. Phys. Org. Chem.* **2009**, *22*, 1155-1165. https://doi.org/10.1002/poc.1569.

[275] Hobza, P.; Zahradnik, R. *Intermolecular Complexes: The Role of van der Waals Systems in Physical Chemistry and in the Biodisciplines*; Elsevier Science Ltd, 1988.

[276] Moller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618-622. https://doi.org/10.1103/PhysRev.46.618.

[277] Jenness, G. R.; Karalti, O.; Al-Saidi, W. A.; Jordan, K. D. Evaluation of Theoretical Approaches for Describing the Interaction of Water with Linear Acenes. *J. Phys. Chem. A* **2011**, *115*, 5955-5964. https://doi.org/10.1021/jp110374b.

[278] Hesselmann, A.; Jansen, G.; Schutz, M. Density-Functional Theory-Symmetry-Adapted Intermolecular Perturbation Theory with Density Fitting: A New Efficient

Method to Study Intermolecular Interaction Energies. *J. Chem. Phys.* **2005**, *122*, 014103. https://doi.org/10.1063/1.1824898.

[279] Grimme, S. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**. https://doi.org/10.1002/jcc.20495.

[280] Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104. https://doi.org/10.1063/1.3382344.

[281] Lee, K.; Murray, E. D.; Kong, L.; Lundqvist, B. I.; Langreth, D. C. Higher-Accuracy van der Waals Density Functional. *Phys. Rev. B* **2010**, *82*, 081101. https://doi.org/10.1103/PhysRevB.82.081101.

[282] Szabo, A.; Ostlund, N. S.*Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*; Dover: Mineola, New York, 1996.

[283] Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. Reprint of: A Fifth-Order Perturbation Comparison of Electron Correlation Theories. *Chem. Phys. Lett.* **2013**, *589*, 37-40. https://doi.org/10.1016/j.cplett.2013.08.064.

[284] Hampel, C.; Peterson, K. A.; Werner, H. J. A Comparison of the Efficiency and Accuracy of the Quadratic Configuration Interaction (QCISD), Coupled Cluster (CCSD), and Brueckner Coupled Cluster (BCCD) Methods. *Chem. Phys. Lett.* **1992**, *190*, 1-12. https://doi.org/10.1016/0009-2614(92)86093-W.

[285] Deegan, M. J. O.; Knowles, P. J. Perturbative Corrections to Account for Triple Excitations in Closed and Open Shell Coupled Cluster Theories. *Chem. Phys. Lett.* **1994**, *227*, 321-326. https://doi.org/10.1016/0009-2614(94)00815-9.

[286] Grimme, S. Improved Second-Order Moller-Plesset Perturbation Theory by Separate Scaling of Parallel- and Antiparallel-Spin Pair Correlation Energies. *J. Chem. Phys*. **2003**, *118*, 9095-9102. `https://doi.org/10.1063/1.1569242`.

[287] Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schutz, M.; Celani, P.; Gyorffy, W.; Kats, D.; Korona, T.; Lindh, R.; Mitrushenkov, A.; Rauhut, G.; Shamasundar, K. R.; Adler, T. B.; Amos, R. D.; Bennie, S. J.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Heeselmann, A.; Hetzer, G.; Hrenar, T.; Jansen, G.; Koppl, C.; Lee, S. J. R.; Liu, Y.; Lloyd, A. W.; Ma, Q.; Matta, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklab, A.; O'Neill, D. P.; Palmieri, P.; Peng, D.; Pfluger, K.; Pitzer, R.; Reiher, M.; Shiozaki, T.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Welborn, M. *MOLPRO*, Version 2015.1, a Package of Ab Initio Programs. URL(http://www.molpro.net/), 2015.

[288] Bose, S.; Dhawan, D.; Nandi, S.; Sarkar, R. R.; Ghosh, D. Machine Learning Prediction of Interaction Energies in Rigid Water Clusters. *Phys. Chem. Chem. Phys*. **2018**, *20*, 22987-22996. `https://doi.org/10.1039/C8CP03138J`.

[289] Purvis, G. D.; Bartlett, R. J. A Full Coupled-Cluster Singles and Doubles Model: The Inclusion of Disconnected Triples. *J. Chem. Phys*. **1982**, *76*, 1910. `https://doi.org/10.1063/1.443164`.

[290] Head-Gordon, M.; Pople, J. A.; Frisch, M. J. MP2 Energy Evaluation by Direct Methods. *Chem. Phys. Lett*. **1988**, *153*, 503-506. `https://doi.org/10.1016/0009-2614(88)85250-3`.

[291] Radzicka, A.; Wolfenden, R. A Proficient Enzyme. *Science*. **1995**, *267*, 90-93. `https://doi.org/10.1126/science.7809611`.

[292] Pauling, L. Molecular Architecture and Biological Reactions. *Chem. Eng. News* **1946**, *24*, 1375-1377. https://doi.org/10.1021/cen-v024n010.p1375.

[293] Fischer, E. Einfluss Der Configuration Auf Die Wirkung Der Enzyme. II. *Ber. Dt. Chem. Ges.* **1894**, *27*, 2985-2993. https://doi.org/10.1002/cber.189402703169.

[294] Koshland, D. E. Correlation of Structure and Function in Enzyme Action. *Science*. **1963**, *142*, 1533. https://doi.org/10.1126/science.142.3599.1533.

[295] Guclu, D.; Szekrenyi, A.; Garrabou, X.; Kickstein, M.; Junker, S.; Clapes, P.; Fessner, W. D. Minimalist Protein Engineering of an Aldolase Provokes Unprecedented Substrate Promiscuity. *ACS Catal.* **2016**, *6*, 1848-1852. https://doi.org/10.1021/acscatal.5b02805.

[296] Schurmann, M.; Sprenger, G. A. Fructose-6-Phosphate Aldolase is a Novel Class I Aldolase from *Escherichia coli* and is Related to a Novel Group of Bacterial Transaldolases. *J. Biol. Chem.* **2001**, *276*, 11055-11061. https://doi.org/10.1074/jbc.M008061200.

[297] Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46*, 2287-2303. https://doi.org/10.1021/jm0203783.

[298] Wang, R.; Lai, L.; Wang, S. Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comput. Aided. Mol. Des.* **2002**, *16*, 11-26. https://doi.org/10.1023/A:1016357811882.

[299] Clinton, W. L.; Massa, L. J. Determination of the Electron Density Matrix from X-Ray Diffraction Data. *Phys. Rev. Lett.* **1972**, *29*, 1363. https://doi.org/10.1103/PhysRevLett.29.1363.

[300] Jayatilaka, D.; Grimwood, D. J. Wavefunctions Derived from Experiment. I. Motivation and Theory. *Acta Crystallogr. Sect. A* **2001**, *A57*, 76-86. https://doi.org/10.1107/S0108767300013155.

[301] Massa, L.; Huang, L.; Karle, J. Quantum Crystallography and the Use of Kernel Projector Matrices. *Int. J. Quantum Chem.* **1995**, *29*, 371. https://doi.org/10.1002/qua.560560841.

[302] Koritsanszky, T. S.; Coppens, P. Chemical Applications of X-Ray Charge-Density Analysis. *Chem. Rev.* **2001**, *101*, 1583-1628. https://doi.org/10.1021/cr990112c.

[303] Stewart, R. F. VII. Vibrational Averaging of X-Ray Scattering Intensities. *Isr. J. Chem.* **1977**, *16*, 137-143. https://doi.org/10.1002/ijch.197700023.

[304] Volkov, A.; Coppens, P. Critical Examination of the Radial Functions in the Hansen-Coppens Multipole Model through Topological Analysis of Primary and Refined Theoretical Densities. *Acta Crystallogr.* **2001**, *A57*, 395-405. https://doi.org/10.1107/S0108767301002434.

[305] Volkov, A.; Abramov, Y. A.; Coppens, P. Density-Optimized Radial Exponents for X-Ray Charge-Density Refinement from Ab Initio Crystal Calculations. *Acta Crystallogr.* **2001**, *A57*, 272-282. https://doi.org/10.1107/S0108767300018547.

[306] Coppens, P.; Volkov, A. The Interplay between Experiment and Theory in Charge-Density Analysis. *Acta Crystallogr. Sect. A Found. Crystallogr.* **2004**, *A60*, 357-364. https://doi.org/10.1107/S0108767304014953.

[307] Brock, C. P.; Dunitz, J. D.; Hirshfeld, F. L. Transferability of Deformation Densities among Related Molecules: Atomic Multipole Parameters from Perylene for Improved

Estimation of Molecular Vibrations in Naphthalene and Anthracene. *Acta Crystallogr.* **1991**, *B47*, 789-797. https://doi.org/10.1107/S0108768191003932.

[308] Volkov, A.; Koritsanszky, T.; Coppens, P. Combination of the Exact Potential and Multipole Methods (EP/MM) for Evaluation of Intermolecular Electrostatic Interaction Energies with Pseudoatom Representation of Molecular Electron Densities. *Chem. Phys. Lett.* **2004**, *391*, 170-175. https://doi.org/10.1016/j.cplett.2004.04.097.

[309] Volkov, A.; Messerschmidt, M.; Coppens, P. Improving the Scattering-Factor Formalism in Protein Refinement: Application of the University at Buffalo Aspherical-Atom Databank to Polypeptide Structures. *Acta Crystallogr. Sect. D* **2007**, *D63*, 160-170. https://doi.org/10.1107/S0907444906044453.

[310] Volkov, A. Personal Communication. Middle Tennessee State University, TN, 2009.

[311] MacDougall, P. J.; Bader, R. F. W. Atomic Properties and the Reactivity of Carbenes. *Can. J. Chem.* **1985**, *64*, 1496-1508. https://doi.org/10.1139/v86-246.

Appendices

APPENDIX A

The Molecules Used in This Study

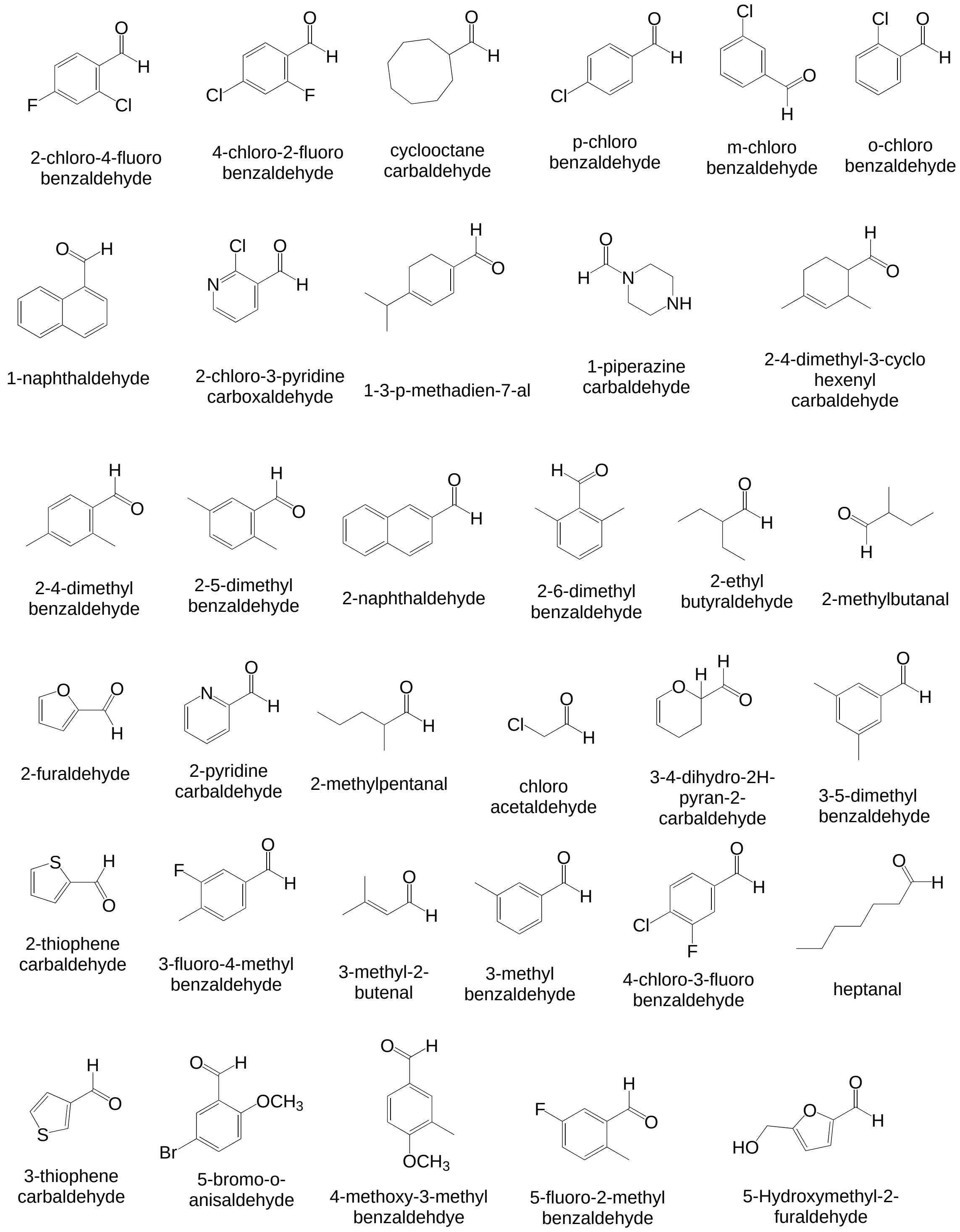

2-chloro-4-fluoro benzaldehyde
4-chloro-2-fluoro benzaldehyde
cyclooctane carbaldehyde
p-chloro benzaldehyde
m-chloro benzaldehyde
o-chloro benzaldehyde
1-naphthaldehyde
2-chloro-3-pyridine carboxaldehyde
1-3-p-methadien-7-al
1-piperazine carbaldehyde
2-4-dimethyl-3-cyclo hexenyl carbaldehyde
2-4-dimethyl benzaldehyde
2-5-dimethyl benzaldehyde
2-naphthaldehyde
2-6-dimethyl benzaldehyde
2-ethyl butyraldehyde
2-methylbutanal
2-furaldehyde
2-pyridine carbaldehyde
2-methylpentanal
chloro acetaldehyde
3-4-dihydro-2H-pyran-2-carbaldehyde
3-5-dimethyl benzaldehyde
2-thiophene carbaldehyde
3-fluoro-4-methyl benzaldehyde
3-methyl-2-butenal
3-methyl benzaldehyde
4-chloro-3-fluoro benzaldehyde
heptanal
3-thiophene carbaldehyde
5-bromo-o-anisaldehyde
4-methoxy-3-methyl benzaldehdye
5-fluoro-2-methyl benzaldehyde
5-Hydroxymethyl-2-furaldehyde

Figure 23: Molecular structures of compounds used in our study

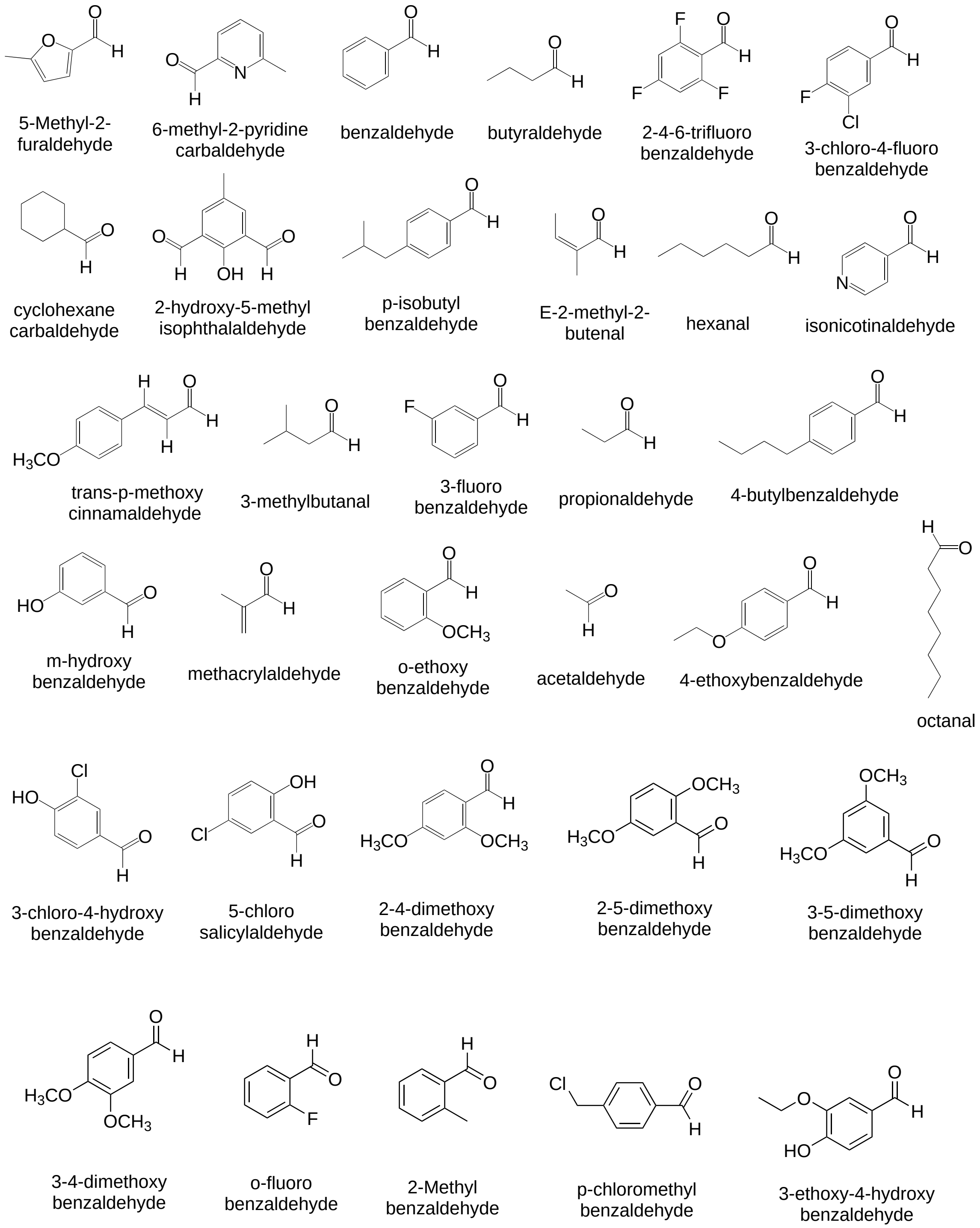


Figure 23 (Continued.)

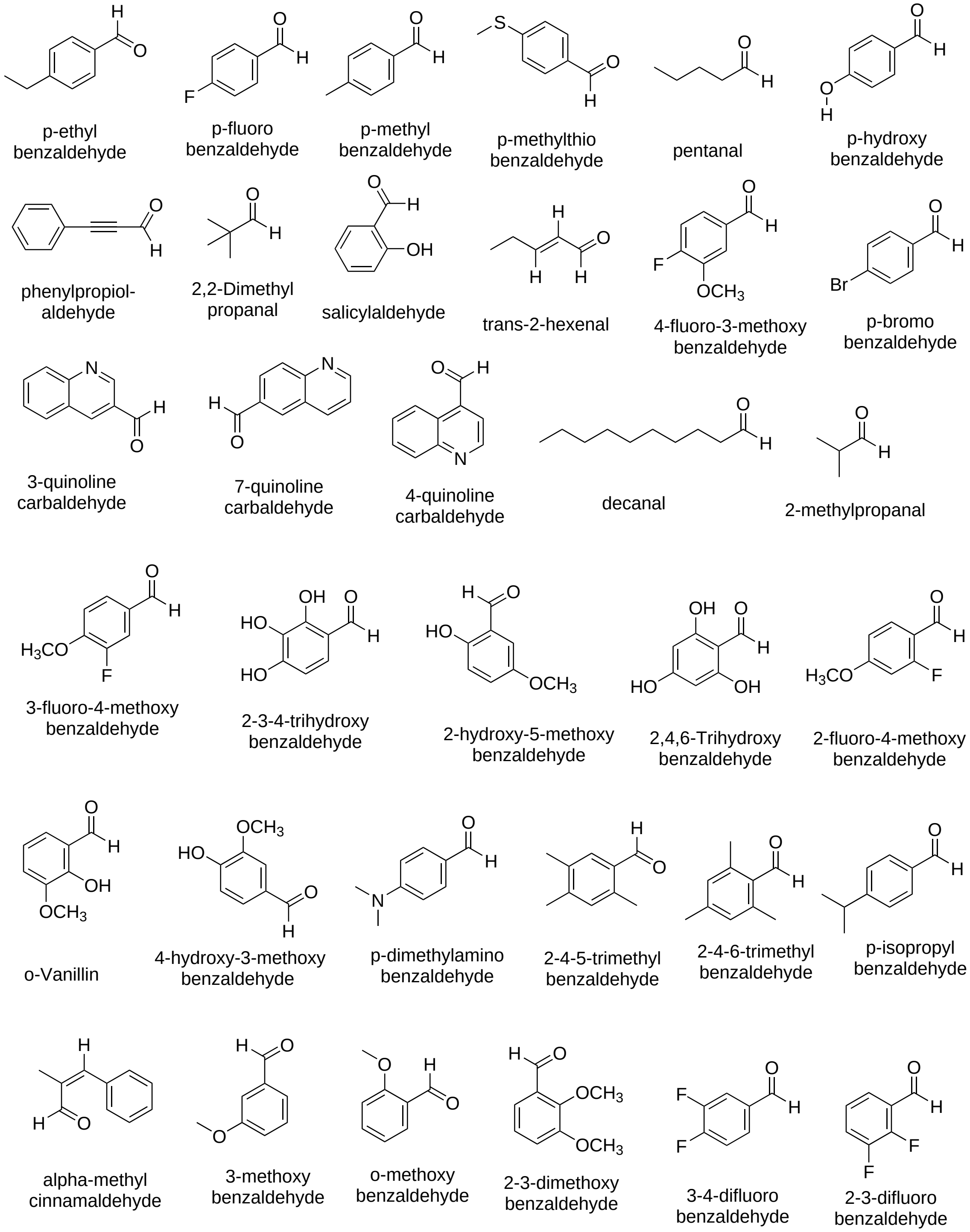

p-ethyl benzaldehyde
p-fluoro benzaldehyde
p-methyl benzaldehyde
p-methylthio benzaldehyde
pentanal
p-hydroxy benzaldehyde
phenylpropiol-aldehyde
2,2-Dimethyl propanal
salicylaldehyde
trans-2-hexenal
4-fluoro-3-methoxy benzaldehyde
p-bromo benzaldehyde
3-quinoline carbaldehyde
7-quinoline carbaldehyde
4-quinoline carbaldehyde
decanal
2-methylpropanal
3-fluoro-4-methoxy benzaldehyde
2-3-4-trihydroxy benzaldehyde
2-hydroxy-5-methoxy benzaldehyde
2,4,6-Trihydroxy benzaldehyde
2-fluoro-4-methoxy benzaldehyde
o-Vanillin
4-hydroxy-3-methoxy benzaldehyde
p-dimethylamino benzaldehyde
2-4-5-trimethyl benzaldehyde
2-4-6-trimethyl benzaldehyde
p-isopropyl benzaldehyde
alpha-methyl cinnamaldehyde
3-methoxy benzaldehyde
o-methoxy benzaldehyde
2-3-dimethoxy benzaldehyde
3-4-difluoro benzaldehyde
2-3-difluoro benzaldehyde

Figure 23 (Continued.)

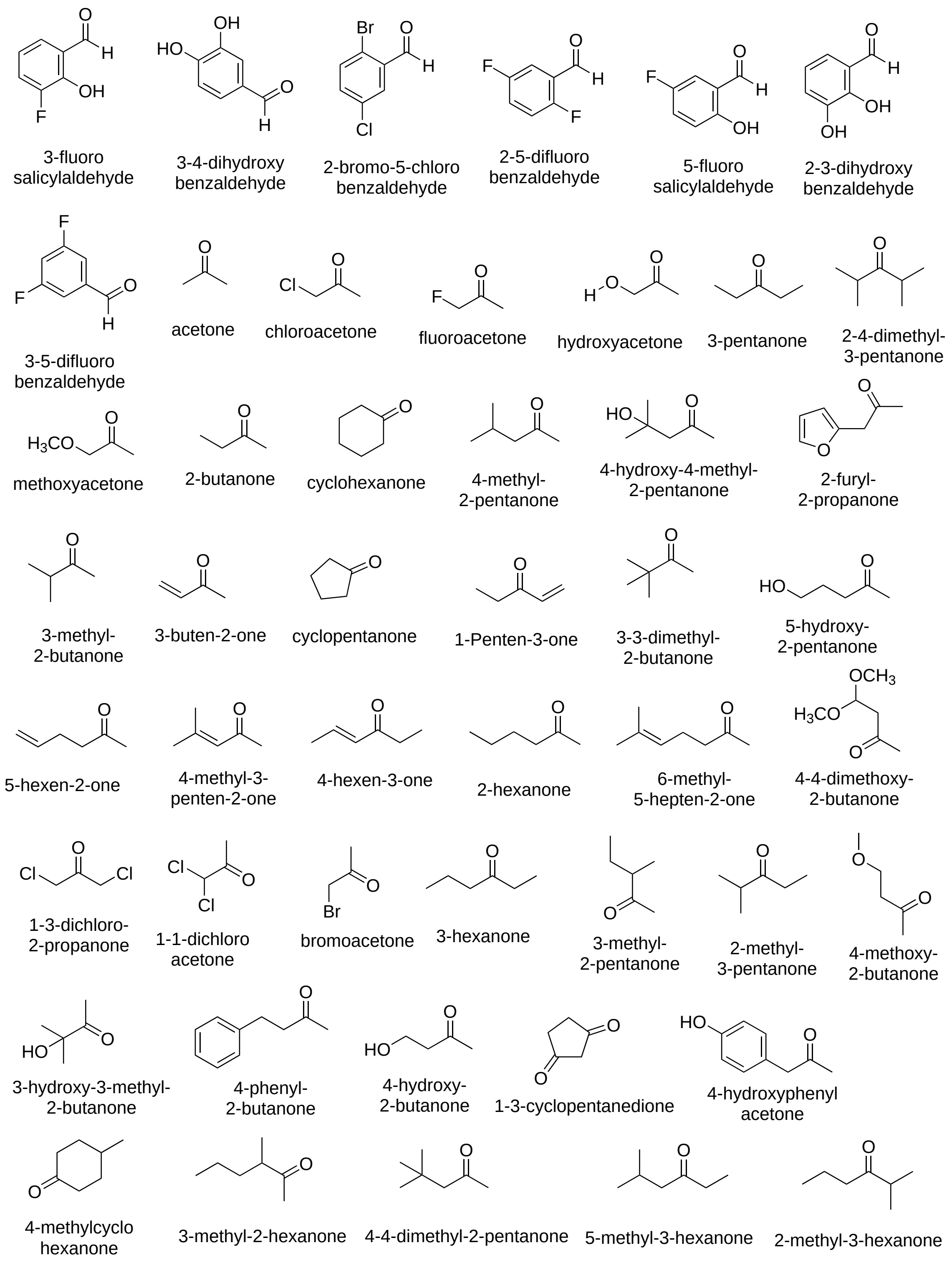


Figure 23 (Continued.)

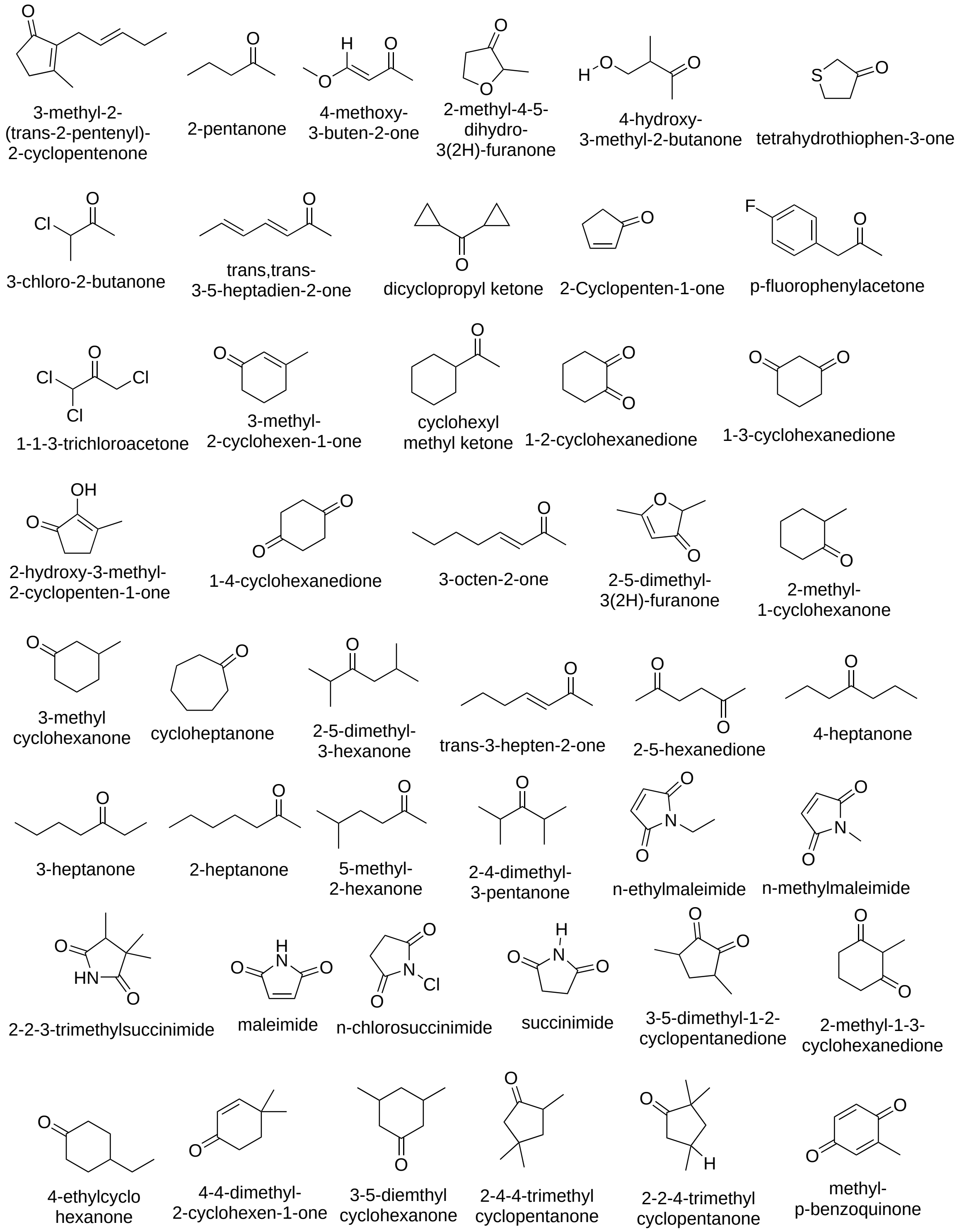
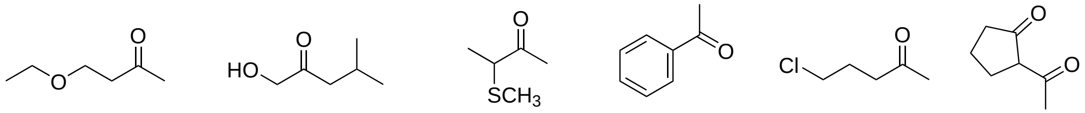


Figure 23 (Continued.)



4-ethoxy-2-butanone 1-hydroxy-4-methyl-2-pentanone 3-methylthio-2-butanone acetophenone 5-chloro-2-pentanone 2-hydroxy-1-cyclopentenyl methyl ketone

Figure 23(Continued.)

APPENDIX B

NMR and IR Predicted Values

Table 12: Predicted NMR shifts

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol01	2-chloro-4fluorobenzaldehyde	188.11	190.09	1.98	186.00	-2.11	188.11	0.00
mol02	4-chloro-2-fluorobenzaldehyde	185.95	191.33	5.38	189.98	4.03	190.45	4.50
mol03	cyclooctanecarbaldehyde	204.60	207.69	3.09	206.57	1.97	209.47	4.87
mol04	p-chlorobenzaldehyde	190.77	191.64	0.87	190.60	-0.17	191.28	0.51
mol05	m-chlorobenzaldehyde	190.76	192.35	1.59	191.46	0.70	192.38	1.62
mol06	o-chlorobenzaldehyde	189.46	192.50	3.04	191.86	2.40	191.84	2.38
mol07	2-chloro-3-pyridinecarboxaldehyde	189.18	191.03	1.85	192.15	2.97	190.47	1.29
mol08	1-3-p-methadien-7-al	192.40	190.16	-2.24	191.96	-0.44	192.26	-0.15
mol09	1-piperazinecarbaldehyde	160.81	167.51	6.70	170.49	9.68	165.40	4.59
mol10	2-4-dimethyl-3-cyclohexenylcarbaldehyde	205.86	191.53	-14.33	189.69	-16.17	190.23	-15.63
mol11	2-4-dimethylbenzaldehyde	192.23	195.04	2.81	192.68	0.45	193.71	1.48

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol12	2-5-dimethylbenzaldehyde	192.78	195.71	2.93	193.18	0.40	193.76	0.98
mol13	2-6-dimethylbenzaldehyde	193.57	194.91	1.34	193.79	0.22	193.26	-0.31
mol14	2-naphthaldehyde	192.01	191.91	-0.10	191.72	-0.29	191.90	-0.12
mol15	1-naphthaldehyde	193.25	194.34	1.09	191.57	-1.68	192.37	-0.89
mol16	2-ethylbutyraldehyde	205.64	204.39	-1.25	204.54	-1.10	205.87	0.23
mol17	2-furaldehyde	177.93	182.45	4.52	183.16	5.23	182.55	4.62
mol18	2-methylbutanal	205.20	204.40	-0.80	204.36	-0.84	205.53	0.33
mol19	Pyridine-2-aldehyde	193.34	194.95	1.61	195.83	2.49	194.13	0.79
mol20	2-methylpentanal	205.19	204.47	-0.72	205.29	0.10	205.65	0.46
mol21	2-thiophenecarbaldehyde	182.97	185.17	2.20	182.75	-0.22	182.30	-0.67
mol22	3-3-dimethyl-2-butanone	213.81	213.40	-0.41	212.61	-1.20	211.09	-2.72

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol23	3-4-dihydro-2H-pyran-2-carbaldehyde	202.03	200.19	-1.84	203.35	1.32	201.51	-0.53
mol24	3-5-dimethylbenzaldehyde	192.75	192.24	-0.51	192.18	-0.57	192.63	-0.12
mol25	4-chloro-3-fluorobenzaldehyde	189.77	191.39	1.62	189.30	-0.47	190.38	0.61
mol26	3-fluoro-4-methylbenzaldehyde	190.80	191.42	0.62	190.91	0.11	191.26	0.46
mol27	3-methyl-2-butenal	191.03	191.78	0.75	189.41	-1.62	190.32	-0.71
mol28	3-methylbenzaldehyde	192.47	193.87	1.40	191.80	-0.67	192.81	0.34
mol29	3-thiophenecarbaldehyde	184.91	189.61	4.70	188.37	3.46	188.48	3.57
mol30	4-4-dimethoxy-2-butanone	205.10	208.45	3.35	209.24	4.14	208.53	3.43
mol31	4-methoxy-3-methylbenzaldehyde	190.95	189.96	-0.99	194.08	3.13	193.98	3.03
mol32	5-fluoro-2-methylbenzaldehyde	191.15	192.29	1.14	190.54	-0.61	191.55	0.40
mol33	5-hydroxymethyl-furfural	178.00	181.91	3.91	182.83	4.83	183.25	5.25
mol34	5-methyl-2-furaldehyde	176.81	182.30	5.49	182.39	5.58	181.98	5.17

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol35	6-methyl-2-pyridinecarbaldehyde	193.59	194.86	1.27	194.23	0.64	194.19	0.60
mol36	acetone	206.55	206.14	-0.41	206.27	-0.28	205.89	-0.66
mol37	benzaldehyde	192.28	186.86	-5.42	191.91	-0.37	192.97	0.69
mol38	butyraldehyde	202.80	200.79	-2.01	201.10	-1.70	202.01	-0.79
mol39	chloroacetaldehyde	193.99	195.11	1.12	197.24	3.25	196.55	2.56
mol40	chloroacetone	200.29	202.09	1.80	205.87	5.58	204.54	4.25
mol41	cyclohexanecarbaldehyde	204.73	204.15	-0.58	204.29	-0.44	205.60	0.87
mol42	3-chloro-4-fluorobenzaldehyde	189.35	191.05	1.70	189.83	0.48	190.90	1.55
mol43	2-4-6-trifluorobenzaldehyde	183.13	182.82	-0.31	182.48	-0.65	181.85	-1.28
mol44	E-2-methyl-2-butenal	195.09	194.39	-0.71	192.28	-2.81	195.21	0.12
mol45	fluoroacetone	204.77	199.73	-5.04	199.90	-4.87	198.97	-5.80

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol46	trans-4-methoxycinnamaldehyde	193.41	188.53	-4.88	186.25	-7.16	187.21	-6.20
mol47	heptanal	202.83	200.98	-1.85	200.89	-1.94	201.18	-1.65
mol48	hexanal	202.87	200.98	-1.90	201.25	-1.62	201.95	-0.92
mol49	hydroxyacetone	208.23	207.22	-1.01	213.32	5.09	211.80	3.57
mol50	4-pyridinecarboxaldehyde	191.65	192.61	0.96	190.10	-1.55	190.66	-0.99
mol51	3-methylbutanal	202.71	201.34	-1.37	201.11	-1.60	201.26	-1.45
mol52	3-fluorobenzaldehyde	191.04	192.08	1.04	191.01	-0.03	191.47	0.43
mol53	m-hydroxybenzaldehyde	192.82	193.10	0.28	191.90	-0.92	192.45	-0.37
mol54	4-butylbenzaldehyde	191.85	190.97	-0.88	192.24	0.39	192.35	0.50
mol55	methacrylaldehyde	194.60	194.89	0.29	192.57	-2.03	194.31	-0.29
mol56	methoxyacetone	206.50	206.51	0.01	210.98	4.48	210.42	3.92
mol57	p-isobutylbenzaldehyde	192.05	191.02	-1.03	192.08	0.03	192.19	0.14

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol58	2-hydroxy-5-methylisophthalaldehyde	191.99	187.83	-4.16	185.12	-6.87	187.49	-4.50
mol59	3-ethoxy-4-hydroxybenzaldehyde	191.22	192.60	1.38	192.09	0.87	192.12	0.90
mol60	2-5-dimethoxybenzaldehyde	189.37	192.35	2.98	192.89	3.52	193.31	3.94
mol61	3-5-dimethoxybenzaldehyde	191.74	191.77	0.03	192.77	1.03	193.02	1.28
mol62	3-4-dimethoxybenzaldehyde	190.70	190.21	-0.49	191.95	1.25	191.82	1.12
mol63	o-fluorobenzaldehyde	187.34	190.15	2.81	190.03	2.69	190.24	2.90
mol64	2-Methylbenzaldehyde	192.33	191.11	-1.22	190.20	-2.13	191.74	-0.59
mol65	octanal	193.64	205.39	11.75	201.90	8.26	203.10	9.46
mol66	p-chloromethylbenzaldehyde	191.39	191.22	-0.17	191.24	-0.15	191.55	0.16
mol67	p-ethylbenzaldehyde	191.89	191.03	-0.86	192.23	0.34	192.24	0.35
mol68	p-fluorobenzaldehyde	190.54	190.53	-0.01	191.36	0.82	191.51	0.97

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol69	p-hydroxybenzaldehyde	191.26	189.89	-1.37	191.84	0.58	191.69	0.43
mol70	p-methylbenzaldehyde	191.74	190.83	-0.91	192.20	0.46	192.19	0.45
mol71	p-methylthiobenzaldehyde	191.11	190.05	-1.06	191.44	0.33	191.37	0.26
mol72	pentanal, valeraldehyde	202.83	200.48	-2.35	201.31	-1.53	201.87	-0.96
mol73	phenylpropiolaldehyde	176.71	180.70	3.99	170.71	-6.00	172.95	-3.76
mol74	2-2-dimethylpropanal	205.83	208.34	2.51	204.55	-1.28	206.70	0.87
mol75	propionaldehyde	203.21	203.18	-0.04	200.64	-2.58	200.97	-2.24
mol76	salicylaldehyde	196.54	191.92	-4.62	192.68	-3.86	191.62	-4.92
mol77	trans-2-hexenal	194.01	188.85	-5.16	189.90	-4.11	189.81	-4.20
mol78	4-fluoro-3-methoxybenzaldehyde	190.68	192.34	1.66	191.88	1.20	191.83	1.15
mol79	3-quinolinecarbaldehyde	190.71	190.11	-0.60	190.56	-0.15	190.66	-0.06
mol80	7-quinolinecarbaldehyde	192.00	191.09	-0.91	192.39	0.39	193.01	1.01

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol81	4-quinolinecarbaldehyde	192.61	193.53	0.92	189.82	-2.79	190.86	-1.75
mol82	cyclohexanone	211.56	193.71	-17.85	208.94	-2.62	207.69	-3.87
mol83	4-ethoxybenzaldehyde	190.66	189.82	-0.84	191.82	1.16	191.74	1.08
mol84	o-ethoxybenzaldehyde	189.59	190.33	0.74	192.73	3.14	192.30	2.71
mol85	acetaldehyde	199.93	198.87	-1.06	199.40	-0.53	199.29	-0.64
mol86	3-chloro-4-hydroxybenzaldehyde	190.27	189.69	-0.58	191.20	0.93	191.80	1.53
mol87	5-chlorosalicylaldehyde	195.43	189.10	-6.33	191.52	-3.91	191.08	-4.35
mol88	2-4-dimethoxybenzaldehyde	188.14	189.18	1.04	193.48	5.34	192.83	4.69
mol89	decanal	202.55	200.15	-2.40	200.79	-1.76	201.02	-1.53
mol90	2-methylpropanal	204.86	202.02	-2.85	204.68	-0.18	204.96	0.10
mol91	2-2-3-trimethylsuccinimide	183.90	177.95	-5.95	188.80	4.90	190.53	6.63

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol92	maleimide	171.36	173.50	2.14	169.99	-1.37	171.56	0.20
mol93	4-methyl-2-pentanone	208.57	208.23	-0.35	210.55	1.98	210.66	2.09
mol94	2-4-dimethyl-3-pentanone	218.32	213.86	-4.46	216.51	-1.81	217.03	-1.29
mol95	n-chlorosuccinimide	179.33	172.33	-7.01	178.95	-0.38	177.50	-1.83
mol96	n-ethylmaleimide	170.66	174.71	4.05	174.26	3.60	171.08	0.42
mol97	n-methylmaleimide	170.84	174.57	3.73	172.79	1.95	170.11	-0.73
mol98	2-butanone	209.28	207.27	-2.01	209.56	0.28	209.46	0.18
mol99	3-methyl-2-butanone	212.49	212.28	-0.21	211.37	-1.12	211.16	-1.33
mol100	3-pentanone	212.07	210.48	-1.59	212.77	0.70	213.26	1.19
mol101	succinimide	177.81	175.65	-2.16	180.00	2.19	178.09	0.28
mol102	2-fluoro-p-anisaldehyde	185.95	188.86	2.91	190.34	4.39	188.49	2.54
mol103	3-fluoro-4-methoxybenzaldehyde	189.74	188.31	-1.43	190.76	1.02	189.49	-0.25

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol104	2-3-4-trihydroxybenzaldehyde	193.28	191.21	-2.07	191.81	-1.47	191.36	-1.92
mol105	2-4-6-trihydroxybenzaldehyde	190.87	195.37	4.50	199.03	8.16	200.31	9.44
mol106	2-hydroxy-5-methoxybenzaldehyde	196.14	189.20	-6.94	194.02	-2.12	192.76	-3.38
mol107	3-methoxysalicylaldehyde	196.59	192.21	-4.38	192.00	-4.59	192.09	-4.50
mol108	4-hydroxy-3-methoxybenzaldehyde	191.21	192.52	1.31	192.03	0.82	192.09	0.88
mol109	p-dimethylaminobenzaldehyde	190.11	187.95	-2.16	191.52	1.41	190.35	0.24
mol110	2-4-5-trimethylbenzaldehyde	192.39	188.99	-3.40	190.66	-1.73	190.61	-1.78
mol111	2-4-6-trimethylbenzaldehyde	192.82	194.02	1.20	191.14	-1.68	192.62	-0.20
mol112	p-isopropylbenzaldehyde	191.85	188.88	-2.97	191.54	-0.31	189.59	-2.26
mol113	3-buten-2-one	198.82	199.74	0.92	197.72	-1.10	198.53	-0.29
mol114	cyclopentanone	220.16	209.77	-10.39	214.23	-5.93	213.95	-6.21

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol115	1-penten-3-one	201.14	209.59	8.45	201.80	0.66	201.30	0.16
mol116	4-hydroxy-4-methyl-2-pentanone	210.70	206.77	-3.93	209.86	-0.84	209.50	-1.20
mol117	2-furyl-2-propanone	204.11	209.49	5.38	209.39	5.28	210.12	6.01
mol118	6-methyl-5-hepten-2-one	208.48	208.21	-0.27	210.01	1.53	210.75	2.27
mol119	5-hydroxy-2-pentanone	209.91	207.17	-2.74	206.15	-3.77	208.13	-1.78
mol120	1-3-dichloro-2-propanone	195.01	203.34	8.33	198.19	3.18	198.92	3.91
mol121	1-1-dichloroacetone	194.60	197.13	2.53	200.71	6.11	198.95	4.35
mol122	bromoacetone	199.59	201.19	1.60	195.49	-4.10	198.58	-1.01
mol123	alpha-methylcinnamaldehyde	195.37	193.34	-2.03	193.91	-1.46	196.15	0.78
mol124	2-3-dimethoxybenzaldehyde	189.97	193.38	3.41	193.29	3.32	193.79	3.82
mol125	3-4-difluorobenzaldehyde	189.70	187.63	-2.07	189.79	0.09	189.76	0.06
mol126	2-3-difluorobenzaldehyde	186.19	190.65	4.46	189.25	3.06	189.50	3.31

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol127	3-5-difluorobenzaldehyde	189.42	187.85	-1.57	191.61	2.19	192.40	2.98
mol128	2-5-difluorobenzaldehyde	186.07	190.43	4.36	189.03	2.96	189.23	3.16
mol129	5-fluorosalicylaldehyde	195.47	190.76	-4.71	190.44	-5.03	190.38	-5.09
mol130	3-fluorosalicyclaldehyde	196.33	190.66	-5.67	190.37	-5.96	190.26	-6.07
mol131	3-4-dihydroxybenzaldehyde	191.02	188.42	-2.60	191.05	0.03	190.51	-0.52
mol132	2-3-dihydroxybenzaldehyde	196.87	191.64	-5.23	191.24	-5.63	191.30	-5.57
mol133	3-methoxybenzaldehyde	192.09	192.75	0.66	191.47	-0.62	192.50	0.41
mol134	o-methoxybenzaldehyde	189.55	192.91	3.36	193.59	4.04	192.94	3.39
mol135	5-hexen-2-one	207.87	207.72	-0.15	207.12	-0.75	208.98	1.11
mol136	4-methyl-3-penten-2-one	198.35	201.41	3.06	199.45	1.10	199.40	1.05
mol137	4-hexen-3-one	200.54	203.53	2.99	201.93	1.39	204.13	3.59

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol138	2-hexanone	209.00	208.10	-0.90	208.84	-0.17	208.84	-0.17
mol139	3-methyl-2-pentanone	212.77	211.78	-0.99	209.54	-3.23	208.31	-4.46
mol140	3-hexanone	211.65	210.04	-1.61	213.69	2.04	213.71	2.06
mol141	2-methyl-3-pentanone	215.26	213.83	-1.44	213.43	-1.84	216.10	0.84
mol142	4-methoxy-2-butanone	206.79	207.04	0.25	209.14	2.35	208.63	1.84
mol143	3-hydroxy-3-methyl-2-butanone	212.71	220.04	7.33	220.61	7.90	220.83	8.12
mol144	4-phenyl-2-butanone	207.72	208.91	1.19	208.67	0.95	211.06	3.34
mol145	4-hydroxyphenylacetone	208.40	208.46	0.06	207.40	-1.00	208.31	-0.09
mol145	4-hydroxyphenylacetone	206.18	208.03	1.85	207.12	0.94	208.10	1.92
mol146	p-fluorophenylacetone	189.00	199.46	10.46	199.02	10.02	199.02	10.02
mol148	2-bromo-5-chlorobenzaldehyde	190.47	189.25	-1.22	186.45	-4.02	188.41	-2.06
mol149	5-bromo-o-anisaldehyde	188.16	191.09	2.93	191.55	3.39	191.81	3.65

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol150	p-bromobenzaldehyde	190.96	189.98	-0.98	190.96	0.00	191.23	0.27
mol151	2-cyclopenten-1-one	210.45	192.97	-17.48	202.86	-7.59	201.51	-8.94
mol152	3-methyl-2-(trans-2-pentenyl)-2-cyclopentenone	208.96	199.28	-9.68	206.54	-2.42	208.20	-0.77
mol153	2-pentanone	208.93	207.55	-1.39	208.66	-0.27	208.42	-0.52
mol154	4-hydroxy-2-butanone	209.34	206.06	-3.28	206.40	-2.94	204.68	-4.66
mol155	1-3-cyclopentanedione	197.59	207.45	9.86	209.13	11.54	208.70	11.11
mol156	4-methoxy-3-buten-2-one	197.21	198.64	1.43	199.30	2.09	200.34	3.13
mol157	2-methyl-4-5-dihydro-3(2H)-furanone	216.16	207.20	-8.96	213.79	-2.37	212.40	-3.76
mol158	4-hydroxy-3-methyl-2-butanone	213.02	208.99	-4.03	209.07	-3.95	209.79	-3.23
mol159	tetrahydrothiophen-3-one	212.79	208.86	-3.93	211.51	-1.28	210.59	-2.20
mol160	3-chloro-2-butanone	202.95	203.58	0.63	203.65	0.70	204.40	1.45

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol161	trans,trans-3-5-heptadien-2-one	198.76	199.73	0.97	197.12	-1.64	198.27	-0.50
mol162	dicyclopropyl ketone	210.68	197.20	-13.48	206.64	-4.04	203.91	-6.77
mol163	3-methyl-2-cyclohexen-1-one	199.28	199.94	0.66	200.00	0.72	202.46	3.18
mol164	cyclohexyl methyl ketone	211.99	209.98	-2.01	210.84	-1.15	211.48	-0.51
mol165	1-2-cyclohexanedione	194.72	194.70	-0.02	193.46	-1.26	195.60	0.88
mol166	1-3-cyclohexanedione	193.02	198.74	5.72	199.12	6.10	199.96	6.94
mol167	2-hydroxy-3-methyl-2-cyclopenten-1-one	203.76	193.82	-9.94	195.07	-8.70	197.17	-6.59
mol168	1-4-cyclohexanedione	208.29	207.19	-1.10	206.29	-2.00	206.75	-1.54
mol169	3-octen-2-one	198.64	201.57	2.93	199.57	0.93	200.88	2.24
mol170	2-5-dimethyl-3(2H)-furanone	205.58	190.69	-14.89	197.03	-8.55	196.43	-9.15
mol171	2-methyl-1-cyclohexanone	213.28	211.17	-2.11	211.43	-1.85	212.65	-0.63
mol172	3-methylcyclohexanone	211.45	208.76	-2.69	208.07	-3.38	208.68	-2.77

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol173	cycloheptanone	214.96	210.57	-4.39	210.13	-4.83	211.32	-3.64
mol174	4-methylcyclohexanone	211.93	208.69	-3.24	208.56	-3.37	209.09	-2.84
mol175	trans-3-hepten-2-one	198.62	198.97	0.35	198.95	0.33	199.57	0.95
mol176	2-5-hexanedione	206.87	206.81	-0.06	211.44	4.57	208.19	1.32
mol177	4-heptanone	211.06	208.92	-2.14	212.50	1.44	212.70	1.64
mol178	3-heptanone	211.75	210.58	-1.17	211.97	0.22	208.95	-2.80
mol179	2-heptanone	209.04	205.84	-3.20	207.29	-1.75	203.62	-5.42
mol180	5-methyl-2-hexanone	209.22	207.63	-1.59	208.40	-0.82	208.06	-1.16
mol181	2-4-dimethyl-3-pentanone	218.32	214.73	-3.59	216.89	-1.43	217.56	-0.76
mol182	5-methyl-3-hexanone	211.37	210.87	-0.51	211.62	0.25	214.32	2.95
mol183	2-methyl-3-hexanone	214.70	210.90	-3.80	212.88	-1.83	212.04	-2.66

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol184	3-methyl-2-hexanone	212.68	210.93	-1.76	210.14	-2.54	212.62	-0.07
mol185	4-4-dimethyl-2-pentanone	208.67	209.42	0.75	206.52	-2.15	209.35	0.68
mol186	2-5-dimethyl-3-hexanone	214.30	214.73	0.43	220.30	6.00	229.64	15.34
mol187	4-ethoxy-2-butanone	207.28	207.29	0.01	208.69	1.41	208.35	1.07
mol188	1-hydroxy-4-methyl-2-pentanone	209.50	211.42	1.92	209.49	-0.01	208.92	-0.58
mol189	3-methylthio-2-butanone	204.54	208.11	3.57	205.16	0.62	206.86	2.32
mol190	acetophenone	197.85	200.30	2.45	202.45	4.60	202.52	4.67
mol191	5-chloro-2-pentanone	207.42	207.21	-0.21	207.01	-0.41	207.04	-0.38
mol192	methyl-p-benzoquinone	187.66	196.49	8.83	183.57	-4.09	189.46	1.80
mol193	2-hydroxy-1-cyclopentenyl methyl ketone	204.83	196.34	-8.49	204.29	-0.54	199.92	-4.91
mol194	2-methyl-1-3-cyclohexanedione	187.05	200.38	13.33	204.14	17.09	205.84	18.79
mol195	3-5-dimethyl-1-2-cyclopentanedione	206.41	193.06	-13.35	196.60	-9.81	198.17	-8.24

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol196	4-4-dimethyl-2-cyclohexen-1-one	199.61	200.93	1.32	199.87	0.26	201.88	2.27
mol197	3-5-diemthylcyclohexanone	211.21	209.31	-1.90	211.89	0.68	211.38	0.17
mol198	2-2-4-trimethylcyclopentanone	223.19	213.53	-9.66	217.41	-5.78	215.60	-7.59
mol199	2-4-4-trimethylcyclopentanone	221.47	212.83	-8.64	216.88	-4.59	217.38	-4.10
mol200	4-ethylcyclohexanone	212.18	208.55	-3.63	210.80	-1.38	210.75	-1.43
mol201	formamide	162.83	163.29	0.46	167.75	4.92	168.93	6.10
mol202	N-methylformamide	166.49	164.13	-2.36	168.13	1.64	165.27	-1.23
mol203	acetamide	178.06	171.23	-6.83	170.00	-8.06	170.02	-8.04
mol204	Urea	159.49	155.65	-3.84	140.16	-19.33	151.31	-8.18
mol205	acrylamide	171.49	168.12	-3.37	162.59	-8.91	163.70	-7.79
mol206	N,N-dimethylformamide	162.60	161.82	-0.78	170.17	7.57	167.74	5.14

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol207	propionamide	177.68	175.14	-2.54	172.75	-4.93	174.86	-2.82
mol208	N-methylacetamide	171.77	170.47	-1.30	168.29	-3.48	171.66	-0.11
mol209	N-ethylformamide	164.92	163.80	-1.12	169.09	4.17	161.34	-3.58
mol210	methacrylamide	171.17	172.22	1.05	172.03	0.86	174.34	3.17
mol211	crotonamide	172.13	171.53	-0.60	167.47	-4.66	165.56	-6.57
mol212	N,N-dimethylacetamide	170.49	169.17	-1.32	172.72	2.23	172.88	2.39
mol213	n-propylformamide	165.62	164.03	-1.59	167.63	2.01	165.43	-0.19
mol214	n-isopropylformamide	163.73	164.13	0.40	167.75	4.02	164.53	0.80
mol215	butyramide	174.27	176.77	2.50	173.45	-0.82	176.20	1.93
mol216	isobutyramide	180.19	176.19	-4.00	176.21	-3.98	176.46	-3.73
mol217	N-ethylacetamide	170.66	176.99	6.33	171.86	1.20	175.79	5.13
mol218	N-methylpropionamide	175.35	176.37	1.02	172.31	-3.04	180.42	5.07

Continued on next page

Table 12 – continued from previous page

			BCP-NMR		LCP-NMR		COMBINED-NMR	
		Exp.	Pre.	Error	Pred.	Error	Pred.	Error
mol219	oxamide	162.37	172.71	10.34	172.29	9.92	172.94	10.57
mol220	N,N-dimethylacrylamide	166.43	163.93	-2.50	165.15	-1.28	165.44	-0.99
mol221	N-methyl-N-vinylacetamide	169.17	170.31	1.14	170.91	1.74	168.87	-0.31
mol222	N-(hydroxymethyl)acrylamide	166.94	165.95	-1.00	163.58	-3.37	166.88	-0.06
mol223	diacetamide	172.47	172.80	0.33	173.57	1.10	174.70	2.23
mol224	N-tert-butylformamide	163.44	161.65	-1.79	161.51	-1.93	163.08	-0.36
mol225	N,N-diethylformamide	162.16	162.48	0.32	165.77	3.61	158.41	-3.75