Development of Molecular Mechanics Methods to Cover Conjugated Drug-Like Molecules for Structure Based Drug Design

Candide Champion

Department of Chemistry

McGill University

Montréal, Québec, Canada

August 2019

A thesis submitted to McGill University in partial fulfillment of the requirements of the M.Sc. degree

Abstract

Considerable resources and time are required to bring a new drug to the market, in a multidisciplinary process involving structural biologists, synthetic chemists, pharmacologists, among many other experts. It has long been recognized that computation could alleviate costs and human-involvement; and computational tools are now applied to virtually all stages of the drug discovery process. From rigorous statistical analyses of large sets of data or employment of newly emerging artificial intelligence techniques to predict absorption, distribution, metabolism, excretion and toxicity (ADMET) properties among others, to more physically grounded methods simulating the structural and dynamic features of molecular systems. The focus of this thesis will be directed towards this latter class, as we investigate molecular mechanics (MM) models, in which a computationally affordable classical (as opposed to quantum) description of molecules, allows for high-throughput applications and the simulation of large biomolecular systems. Notable applications include virtual screening of large libraries of potential inhibitors, and molecular dynamics simulations of entire proteins or nucleic acids, which can also provide insights onto macromolecule-ligand binding. First, we will describe the inner-workings of MM, and review the current state-of-the-art, as well as most recent contributions to improve these models. Our research group has recently been involved in the development of a new method called H-TEQ, in which we have challenged the traditional representation of molecules in MM using atom types and base our predictions on quantified implementations of well-established chemical principles. Conjugated chains and aromatic rings are privileged scaffolds in drug design, and 84% of FDA approved pharmaceuticals contain at least one nitrogen atom. In that context, any successful application of MM based methods in drug design is bound to pay particular attention to these moieties. In the second part of this thesis, we report our most recent contributions to H-TEQ, in which we extend its domain of application to small organic molecules containing unsaturations. In the third part of this thesis, we describe current shortcomings of MM to describe molecules in which nitrogen atoms are bound to π-systems and share potential solutions which could resolve those issues.

Résumé

Afin d'amener un nouveau composé pharmaceutique sur le marché, des investissement considérable (en ressources et en temps) sont nécessaires, dans un processus impliquant biologistes structurels, chimistes organiciens, pharmacologistes, parmi de nombreux autres experts. Il est reconnu depuis un certain temps que l'utilisation d'outils informatiques pourrait réduire le recours à la main d'œuvre, ainsi que les coûts. Dès à présent, pratiquement chaque étape de la conception de nouveaux médicaments fait appel à ces nouveaux outils. De l'analyse statistique rigoureuse de large jeux de données, ou l'emploi émergent de l'intelligence artificielle permettant de prédire certaines propriétés moléculaires (e.g. absorption, distribution, métabolisme, excrétion et toxicité), jusqu'à l'utilisation de méthodes reposant sur des principes physiques simulant les caractéristiques structurelles et dynamiques de systèmes moléculaires. Dans cette thèse, nous nous intéresserons particulièrement à cette seconde catégorie, en explorant les modèles classiques (et non quantiques) de mécanique moléculaire (MM) auxquels des coûts de calculs avantageux permettent des applications haut-débit et de simuler de larges systèmes biomoléculaires. Les principales applications des modèles MM incluent le criblage virtuel de bibliothèques chimiques afin identifier de potentiels inhibiteurs, ainsi que les simulations par dynamique moléculaire de protéines ou d'acides nucléiques permettant d'éclaircir les mécanismes d'adhésion entre ligand et macromolécule. Dans un premier temps, nous dresserons un bilan sur l'état-de-l'art actuel des méthodes MM, tout en détaillant leur fonctionnement, puis nous décrirons les plus récentes contributions apportées par la communauté scientifique en vue d'améliorer leur performance. Notre groupe de recherche a récemment développé une nouvelle technique appelée H-TEQ, dans laquelle nous omettons la représentation traditionnelle de molécules à travers les « types d'atomes » (liée à plusieurs défaillances majeures), et dont les prédictions sont fondées sur une quantification de principes chimiques qualitatifs établis depuis plusieurs décennies. La conception de nouveaux médicaments fait appel à de nombreux fragments moléculaires privilégiés (e.g. chaines conjuguées, cycles aromatiques) et 84% des composés thérapeutiques approuvés par la FDA contiennent au moins un atome d'azote. Dans ce contexte, l'application des méthodes MM lors de la conception de nouveaux médicaments se doit de porter une attention particulière lors de la paramétrisation des fragments mentionnés ci-dessus, afin d'obtenir des résultats fiables. Dans

la deuxième partie de cette thèse, nous rapportons nos plus récentes contributions apportées à H-TEQ, dont nous étendons le domaine d'application aux molécules organiques désaturés. Dans la troisième partie, nous décrivons les défaillances des méthodes MM pour décrire les molécules contenant un azote lie à un système conjugué, puis proposons de potentielles solutions permettant de résoudre ces insuffisances.

Acknowledgements

First and foremost, I would like to thank my friends, family and my partner Katherine for the continuous support they have given me throughout this degree. At times, research can prove to be particularly frustrating, and it is thanks to your encouragements and help that I have been able to complete this degree and enjoy myself doing so.

I would like to thank Nicolas Moitessier for allowing to enter the fascinating field of computational chemistry without any prior knowledge of programming. Thank you for your guidance over the entire course of my research, for providing your scientific insights when I most needed them, and for proofreading this thesis.

I would like to particularly thank Stephen J. Barigye, who helped me tremendously as I first learned to program and initiated my research. Your constructive comments and criticisms have kept me motivated and surely made me a better researcher. Our long and captivating scientific discussions will remain some of my best memories from this degree. I would also like to thank Wanlei Wei with whom it has been a pleasure to collaborate over these past few months, as well as everyone else who has been part of our group during these two years. Thank you, Anne, Jessica, Jiaying, Juan, Julia, Leo, Mihai, Naëla and Sharon.

Finally, I would also like to thank Alex Wahba and every member of McGill's chemistry outreach group. With you all, it has been a great pleasure to participate in demonstrations and share our passion of science with others.

Contribution of Authors

This thesis includes work that was conducted by the author as well as a manuscript containing work involving other authors.

Chapter 1: Work has been completed entirely by the author.

Chapter 2: This project was based on previous work by Stephen J. Barigye, following a methodology developed by Zhaomin Liu, Paul Labute and Nicolas Moitessier. Wanlei Wei contributed to the analysis of the data. I performed the vast majority of the calculations, wrote a program to analyse the data, carried-out most of the analysis, and wrote the manuscript presented in this chapter.

Chapter 3: Work has been completed entirely by the author.

Chapter 4: Work has been completed entirely by the author.

Table of contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

ADMET	Absorption, Distribution, Metabolism, Excretion and Toxicity
AI	Artificial Intelligence
AMBER	Assisted Modeling Building with Energy Refinement
BLW	Block-Localized Wavefunction
CADD	Computer-Aided Drug Design
CGenFF	CHARMM General FF
CHARMM	Chemistry at HARvard Molecular Mechanics
EDA	Energy Decomposition Analysis
EDG	Electron Donating Group
ESP	Electrostatic Potential
EWG	Electron Withdrawing Group
FEP	Free Energy Perturbation
FF	Force Field
GAFF	General AMBER Force Field
GPU	Graphical Processing Unit
GROMOS	GROningen MOlecular Simulation
H-TEQ	Hyperconjugation for Torsional Energy Quantification
LBDD	Ligand Based Drug Design
LJ	Lennard-Jones
LP	Lone Pair of electrons
MD	Molecular Dynamics
MM	Molecular Mechanics
MP2	Møller-Plesset (perturbation theory) – level 2
NMA	*N*-Methylacetamide
NBO	Natural Bond Orbital
OPLS	Optimized Potential for Liquid Simulations
PES	Potential Energy Surface
QM	Quantum Mechanics
RMSE	Root Mean Square Error
SBDD	Structure Based Drug Design
TS	Transition State
vdW	van der Waals

Chapter 1

1 Current Status of Molecular Mechanics Based Methods for Drug Discovery

1.1 Molecular Mechanics in Drug Discovery

1.1.a Computational Methods in Drug Discovery

In 2004, Jorgensen stated that while computers couldn't design drugs single-handedly, they permeated practically all stages of drug discovery (DD).[1] Computational tools have steadily spread since this observation, as a result of the continued increase in computational power, and the ever-growing availability of large amounts of data. DD is a non-linear process containing multiple feedback-loops, in which all facets of a potential therapeutic compound are examined concomitantly.[2] This idea has been reinforced by the introduction of computational tools which can predict properties on wide libraries of compounds when access to physical samples are limited (i.e. low throughput or expensive experiments). For example, it is now common to estimate a molecule's "drug-likeness" (based on Lipinski's rules or more advanced schemes),[3] and/or evaluate absorption, distribution, metabolism, excretion and toxicity (ADMET) properties[4] earlier in the DD process, which could reduce the fraction of pharmacokinetics or toxicity related failures in clinical phases.[5] More importantly, a compound's ability to interact with a biological target can be assessed in computer aided drug design (CADD) following two main approaches known as ligand based drug design (LBDD) and structure based drug design (SBDD).

In SBDD, knowledge of the 3-D structure of a designated target is used to search for potential inhibitors possessing significant structural and chemical complementarity.[6] Central to the SBDD paradigm is the estimation of rigorous macromolecule-drug binding affinities; the accuracy of those predictions rely on multiple factors such as the level of detail of the structural model (subatomic, atomic, coarse-grained),[7] a proper sampling of accessible molecular conformations,[8] as well as the accuracy of the mathematical functions (potentials) used to compute a molecule's

(or molecular complex's) potential energy surface (PES).[9, 10] A wide array of methods allowing to predict binding affinities are now accessible to researchers; from rapid but approximate scoring functions employed during preliminary searches through large libraries of compounds (virtual high throughput screening),[11] to the more accurate and computationally costly free energy perturbation (FEP)[12, 13] calculations generally performed on the most promising candidates only, among other methods.[14]

Molecular potentials can be computed following two fundamental approaches. On one hand, quantum mechanical (QM) techniques could provide a very accurate depiction of the PES of molecules as they consider electronic interactions at the subatomic level explicitly. However, such methods cannot be carried to high-throughput tasks or evaluate the energetics of large molecules (e.g. entire proteins in solvent), due to their restrictive computational costs. In this context, molecular mechanics (MM) methods have emerged in which the PES is calculated using simplified potentials aimed to approximate the QM (and experimental) energy surface, while reducing computational costs by several orders of magnitude.

The major focus of this review will be on the current status of atomistic MM methods with an emphasis on the modeling of small drug-like molecules. We will first give a general description of how molecular systems are treated in MM, followed by a discussion of the current liabilities of MM, and how the research community has addressed them in recent years. It is out of the scope of this review to discuss the large array of computational methods applicable to DD and we recommend the following reviews to any interested reader.[1, 15]

1.1.b General Description of Molecular Mechanics

In MM, molecular systems are somewhat described as a set of **beads (or points)** interconnected by **springs**, where beads represent atoms in the molecule and springs represent the different interactions beads are subject to (Figure 1.1). Each bead is assigned an "atom type" based on the element, hybridization state and direct chemical environment of the atom that it represents. Every atom type is associated to a set of parameters which determines its behavior. For example, a carbon atom in ethane could be assigned the hypothetical atom type "C_{sp3}", while a carbon atom in an alkene would be assigned "C_{sp2}" and a carbon atom in benzene would be assigned "$C_{aromatic}$".

It becomes evident that describing larger portions of chemical space, while considering subtle differences in chemical environments and properties, requires the addition of more atom types. For example, in the most recent version of AMBER, the protonated histidine residue is assigned different atom types than its unprotonated counterpart.[16] However, it is a known drawback that the addition of too many atom types both increases the cost of the parametrization process and is linked to the redundancy of some parameters.[17] These limitations will be discussed in more detail throughout this review.

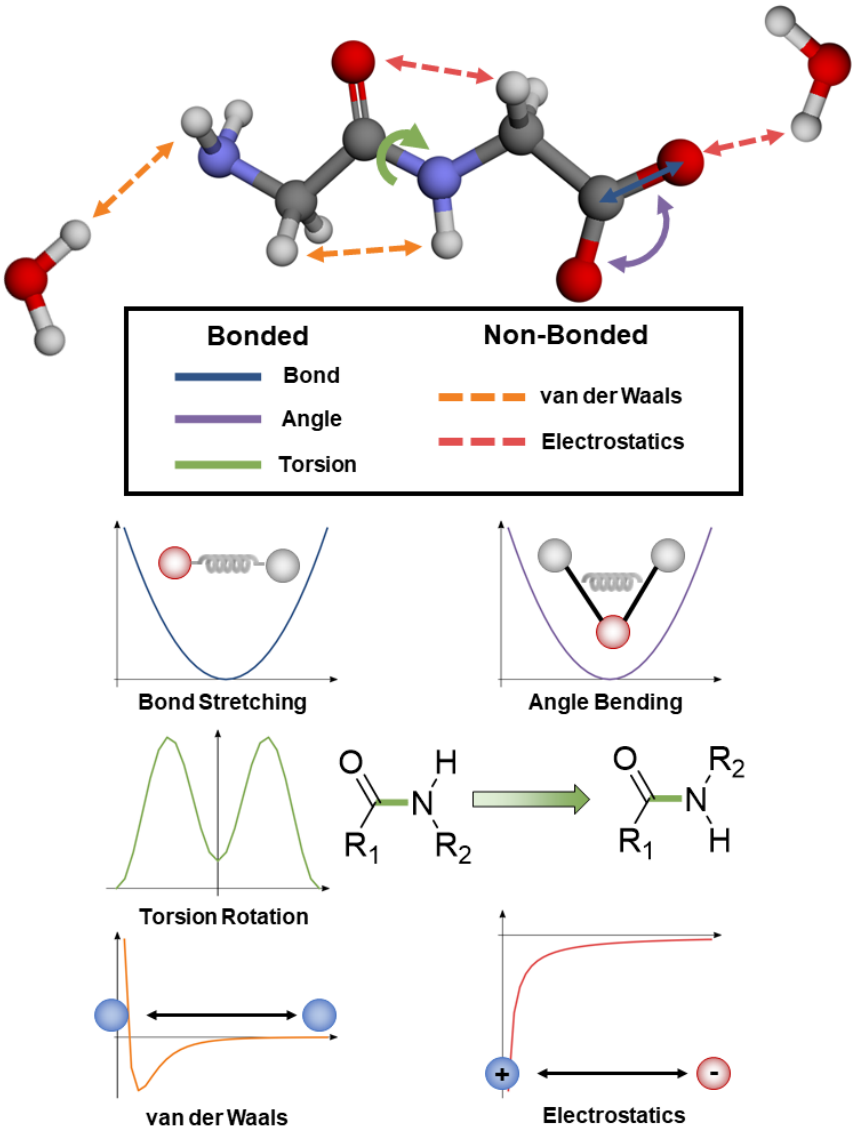


Figure 1.1. Interactions included in common molecular mechanics models to calculate the potential energy of a molecule. As an example, glycylglycine is shown in the presence of two water molecules. Not all interactions are shown to reduce visual clutter.

The overall energy of the system is calculated as a linear combination of multiple components, each associated to an underlying interaction as exemplified by Eqs. 1.1 to 1.7. These contributions can be split into two categories, bonded interactions (bonds, angles, torsions, out-of-plane) which are calculated for atoms within the same molecule, and non-bonded interactions (van der Waals and electrostatics) which are calculated for pairs of atoms separated by 3 or more bonds (intramolecular) or pairs of atoms in different molecules (intermolecular). Each term in the equation uses two or more parameters (shown in **bold**) which vary depending on the atom types of the beads involved in that interaction. Parameters are obtained from precomputed tables. In case parameters for a specific molecule are missing, programs can either assign generic parameters (developed from a similar molecule), with no guarantee that these will transfer well, or exit without being able to provide any information. The equations (which may differ from those shown here) used to calculate the multiple interactions and the corresponding set of parameters are referred to as force fields (FFs).

$$\mathrm{E_{total}} = \underbrace{\mathrm{E_{bonds}} + \mathrm{E_{angles}} + \mathrm{E_{torsions}} + \mathrm{E_{out-of-plane}}}_{\boldsymbol{bonded}} + \underbrace{\mathrm{E_{vdW}} + \mathrm{E_{electrostatic}}}_{\boldsymbol{non-bonded}} \tag{1.1}$$

$$E_{bonds} = \boldsymbol{K_r}\left(r - \boldsymbol{r_{eq}}\right)^2 \tag{1.2}$$

$$E_{angles} = \boldsymbol{K_\theta}\left(\theta - \boldsymbol{\theta_{eq}}\right)^2 \tag{1.3}$$

$$E_{torsion} = \sum_{n=1}^{N} \boldsymbol{V_n}\left(1 + cos(n\varphi + \boldsymbol{\delta})\right) \tag{1.4}$$

$$E_{out-of-plane} = \boldsymbol{K_\omega}\left(\omega - \boldsymbol{\omega_{eq}}\right)^2 \tag{1.5}$$

$$E_{vdW} = \sum_{pairs\ i,j} \boldsymbol{\varepsilon_{ij}} \left[\left(\frac{\boldsymbol{R_{min,i,j}}}{r_{i,j}} \right)^{12} - \left(\frac{\boldsymbol{R_{min,i,j}}}{r_{i,j}} \right)^{6} \right] \tag{1.6}$$

$$E_{electrostatics} = \sum_{pairs\ i,j} \frac{q_i q_j}{4\pi\varepsilon_0 r_{i,j}} \quad (1.7)$$

The connectivity of a molecule is pre-set, and bond breaking/forming events cannot occur within the molecular mechanics framework. A notable exception to that rule is ReaxFF,[18] in which connectivity is recalculated for each conformation, and set depending on interatomic distances. While this FF has seen applications in various contexts,[19-21] it cannot yet be applied to study large biologically relevant macromolecules as simulations are subject to complications as the number of possible reactions increases (e.g. large proteins).[22, 23] In addition to the fixed molecular connectivity imposed by MM methods, lone pairs of electrons are usually not explicitly considered (there is no lone pairs bead). Hence, the many interactions lone pairs can undergo (conjugation, lp-lp repulsion, etc.) are thus generally considered by the heteroatom holding the lone pair. This simplification can become problematic when modeling nitrogen containing compounds which we will further discuss in this thesis.

1.1.c Main Domains of Application of Force Fields

MM methods trace their roots back to the 1970's with Allinger's MM1 force field for hydrocarbons.[24] While computational capabilities have and continue to increase exponentially (allowing expensive QM calculations to be used in rare cases[25]), MM methods remain at the forefront due to their applicability to high-throughput tasks such as docking and computationally intensive tasks such as molecular dynamics (MD), which constitutes the two major applications of MM in the realm of SBDD.

In docking, a potential ligand is placed within a biological target's active site (protein, or nucleic acid), in what is known as a pose (binding mode), which is subsequently analyzed by a scoring function. Multiple poses for the same ligand are evaluated, thereby providing insights onto receptor-ligand binding modes.[26] It is important to note that not all scoring functions rely on FFs in their calculations.[8] The strength of docking originates from its extremely cheap computational cost, allowing the scan of large libraries of compounds (10^9 potential ligands), it is therefore routinely used in virtual screening methods.[27] However, docking only gives an account of static

receptor-ligand pair binding affinities, whereas binding events are known to be subject to dynamic changes (induced fit).[28]

Dynamic (fast) events can be tracked in MD, as molecular systems are simulated following Newton's laws of motion. The position of every atom along the overall trajectory is updated after small finite timesteps (femtoseconds), by calculating the forces acting on each atom using a potential energy function.[29] Although MD simulations have been carried out using accurate QM potentials,[25] cheaper MM potentials can carry simulations of larger systems (e.g. entire proteins or even entire virus[30]) in explicit solvent and for increased time scales (microsecond).[31] Increase in computational power combined with MM potentials has thus allowed to monitor biological events which could not previously be captured by MD such as protein folding,[32] motion of ions through channels[33] and dynamics of membrane transporters.[34] More accurate receptor-ligand binding affinities can also be acquired from MD simulations using techniques such as FEP calculations.[13] Hybrid QM/MM potentials have also been employed within MD simulations, which treat regions of interest (e.g. protein active site and ligand) at the more accurate QM level, while accounting for the less important regions (i.e. rest of the protein) with faster MM potentials.[35]

1.1.d Parametrization of Force Fields

The performance of a FF depends on both the specific potential used to compute the energy and the validity of the parameters. While all FFs do not share the same potential, it can be argued that the quality of the parameters prevail on the potential function used.[17] Indeed, the simple potential shown in Eq. 1.1, used by most common FFs in SBDD (see section 1.1.e), has been applied with great success to model enzyme-inhibitor interactions,[36] formation of lipid bilayers,[37] and protein folding (in all 3, non-bonded interactions prevail).[38] Hence, the particular methodology employed during the parametrization (from which its parameters arise) is one of the major differences between FFs and their resulting performance. We will only give a short account of the parametrization process, as a recent review by *Vanommeslaeghe et al.*,[17] describes it in great detail. Parametrization is complex,[39] and the limited attention we will grant this topic should not be misunderstood as a token of its simplicity.

First, the list of all atom types that will be used to describe molecular systems needs to be chosen. Ideally, each element will have multiple possible atom types depending on its hybridization state and direct chemical environment. The introduction of more atom types in principle allows to discern more subtle differences between molecular moieties, at the cost of a more extensive parametrization. FF development hence relies heavily on the transferability of parameters (associated to atom types) developed on molecules with specific chemical properties and environments, to different molecules with similar environments and properties (Figure 1.2).

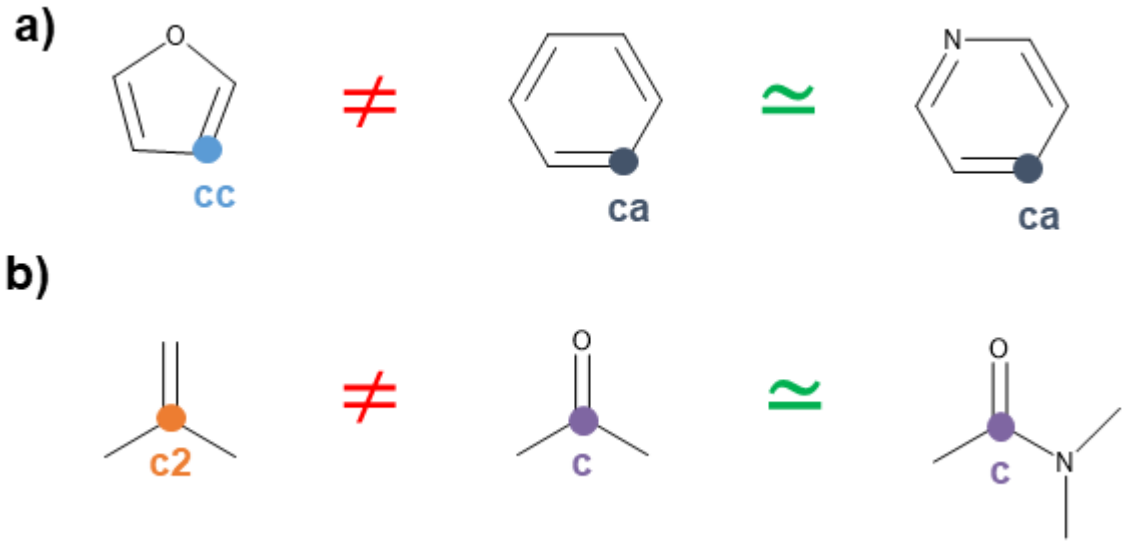


Figure 1.2. Transferability of atom types from a molecule to another. In both example, GAFF atom types are chosen to demonstrate how carbon atoms are assigned different atom types depending on their chemical environment. "≠" denotes situations in which carbon atoms are too different, and a different atom type is hence assigned. "≃" symbol denotes situations where the chemical environment is similar enough to retain the same atom type.

While this reliance on transferability is particularly fit for the modeling of peptides and nucleic acids as they are built from a few repeating units, the vastness of drug-like molecule chemical space[40] limits the possibility for few (e.g. hundreds of) atom types to describe all possible drug-like molecules. An attempt to incorporate a description of the entire periodic table was carried by *Rappe et al.* with the universal force field (UFF).[41] However, applications of UFF in condensed phase simulations of small molecules have shown poor performance,[42] and no scientific articles were found using UFF to model proteins. This exemplifies how accuracy can be lost at the cost of covering a larger portion of chemical space. While the general philosophy behind UFF is not at

fault, the parametrization required to describe accurately the entire periodic table (including biopolymers and organic molecules) is far too costly and cannot be easily carried out. Parametrization for SBDD applications are thus focused only on the relevant elements: C, N, O, H, P, S, halogens and metal ions (e.g. Cu^{2+}, Zn^{2+}, Mg^{2+}) that usually bind to proteins and nucleic acids. It is important to note that new atom type descriptions can always be added to the FF later on in the development process when particular functional groups are found to be improperly modeled. Commonly used FFs are hence built upon over the years as more parameters are generated and existing parameters are perfected; rather than remodeled from scratch, which would require tremendous efforts.

Once the atom types in the FF description have been determined, a training set of molecules is generated for which parameters will be optimized. A validation set (containing different molecules) usually accompanies that training set, to confirm the transferability of parameters to similar molecules.[43] Target data then has to be collected (or calculated) for which the parameters are optimized to reproduce. This optimization is an iterative process, during which parameters are refined from an initial guess until they reproduce the target data with sufficient accuracy. Considering the additive nature of each term in the FF, each part is optimized separately and in a specific order. Hard degrees of freedom (bonds, angles) are optimized first, as they greatly influence the other degrees of freedom (vdW, electrostatics, torsions). Each part of the FF is optimized using different forms of target data such as experimental spectroscopic data (IR, H-NMR),[44] thermodynamic properties (ρ, ΔH_{vap}, ΔG_{solv}),[12] or QM calculations.[45] Interestingly, it has been found that using QM calculated IR vibrations could replace experimental values.[46, 47] In order to reproduce experimental thermodynamic properties (used to parametrize non-bonded parameters), simulated values must be acquired by carrying short MD simulations, which are ultimately compared to experimental values. Torsional parameters are obtained by fitting to QM torsional energy profiles, which require many conformations to be optimized at the QM level. Both aforementioned methods are computationally expensive, hence the availability of target data is the limiting factor determining the size of the training set and ultimately, the transferability of the FF.

Overall, to generate physically meaningful and transferable parameters, it is necessary to use as large and diverse a training set as possible, as well as correct types of target data and

methodologies. For example, the TIP4P-Ew water model was parametrized by using thermodynamic properties (ρ, ΔH_{vap}) over a range of temperatures (235.5-400K) and was found to reproduce other properties which were not used during the training (e.g. heat capacities, self-diffusion coefficient).[48] On the other hand, non-bonded terms in MMFF94 were essentially parametrized by using gas phase QM calculations, which were later found to be less reliable than thermodynamic properties which provide a better account of condensed phase properties and dynamic effects.[49]

1.1.e Common Force Fields in Structure Based Drug Discovery

The most common set of force fields known as class I FFs include the AMBER,[50, 51] CHARMM,[52, 53] GROMOS[54, 55] and OPLS[56, 57] series which calculate potentials using an equation similar to Eq. 1.1. Class II/III FFs use more complex equations, including for example higher order terms for the bond and angle energies (respecting behavior out of equilibrium) and cross-terms which describe the interplay between two motions (e.g. a bond stretching as an angle bends). Common FFs in these classes include CFF,[58, 59] the MM series,[60-63] and MMFF94.[64, 65] While these FFs describe molecular systems more accurately, they require a more extensive parametrization, ultimately limiting the applicability to cover the extremely large chemical space of small drug-like molecules. Furthermore, in SBDD related applications, correctly modeling low-energy conformations has been deemed more important than less probable high-energy conformers. Ligands usually bind to receptors in one of their lowest-in-energy conformations (although this has been challenged)[66, 67] and over the course of a room temperature MD simulation molecules do not see their angles and bonds vary far from equilibrium. Prediction of accurate energy barriers for rotation around a bond are essential however, as overestimated energy barriers would restrict molecules from performing crucial motions.[68] All in all, class II FFs are not as relevant to SBDD applications as they have been largely parametrized from gas phase QM calculations, lack parameters for biological macromolecules and have smaller applicability domains, they have however been applied with great success to predict vibrational and Raman spectra.[59, 63]

In order to be applied to SBDD projects, it is mandatory to have an accurate description of macromolecules (proteins, RNA, etc.), potential binders (e.g. small organic molecules), and most importantly of the intermolecular interactions between them.[69] An essential fact that we have

omitted thus far is that the description of different classes of molecules (proteins, nucleic acids, lipids, small organic molecules) are not performed by the same part of the FF. Using AMBER as an example, each class mentioned above is described by its specifically tailored FF.[16, 70, 71] The importance of accurate intermolecular interactions between molecules from two different classes (e.g. protein and drug) requires complementarity from both parts of the FF. Therefore, it is not possible to use just any protein FF with just any organic FF to understand the interactions between a protein and a drug. This has led to the development of FFs geared to describe small organic molecules working in synergy with existing highly optimized force fields for biomolecules. To that end, AMBER developers have issued the generalized amber force field (GAFF),[43] CHARMM includes the CHARMM general force field (CGenFF)[72], and GROMOS' latest parameter set (54A8[73]) which covers biopolymers is supplemented by an automated topology builder (ATB[74]) for drug-like compounds. Note that computations are required to generate ATB parameters, as opposed to GAFF or CGenFF in which parameters are directly available; ATB is hence similar to the other automatic parameter generating toolkits we will discuss later in this review. On the other hand, OPLS3e[75] includes parameters for both small organics and amino acids within the same FF. A newer version of GAFF, called GAFF2 has been available since the distribution of the AMBER16 package, although the manuscript detailing changes that were implemented is still in preparation.[76]

Table 1.1. Summary of the common class I force fields for structure based drug design.

Family	Molecules covered	Name of latest version
AMBER	proteins, nucleic acids	ff14SB[16]
	small organics	GAFF[43]
	lipids	Lipid14[71]
	carbohydrates	GLYCAM06[77]
CHARMM	proteins, nucleic acids, lipids	CHARMM36[78]
	small organics	CGenFF[72]
GROMOS	proteins, nucleic acids, lipids, small organics	54A8[73]
	small organics	ATB2.0[74]
OPLS	proteins, small molecules	OPLS3e[75]
	nucleic acids	OPLS-AA/M[79]

Note: A complete list detailing all versions of these four FF families can be found in this review by Riniker.[80]

The fact that FFs are tuned to accurately model one specific class of molecules highlights one the biggest hurdles in FF development, that of *transferability*. The current state-of-the-art of FFs in drug discovery relies heavily on the transferability of parameters developed on one set of molecules, to model other similar molecules. A very good example of lack of transferability can be provided by the recent interest in intrinsically disordered proteins (IDPs) which lack a clearly defined three-dimensional folded structure. It has recently been shown that traditional parameters for proteins where not transferable to IDPs, and new parameters had to be developed.[81, 82] This effort to allow the modeling of IDPs is particularly interesting as no new atom type descriptions were introduced in the FF, and the new parameters performed well on both traditional proteins and IDPs. It is important to keep in mind however that prior to that addition, IDPs were not properly modeled, hence an accurate representation of particular molecules depends strongly on the composition of the training set. This issue of parameter transferability is most exacerbated for small drug-like molecules considering the vastness of the chemical space. In 2015, authors of the OPLS force field estimated that 33% of drug-like molecules were missing at least one torsion parameters despite the vast training set used to develop it.[83] In their latest publication to date, attempting to cover an even wider range of drug-like chemical space, the authors do not estimate the current % coverage of drug-like chemical space of their method.[75]

1.1.f Water Models in Force Fields

It is well established that water plays important roles in biological events and does not act simply as an inert solvent.[84] In that respect, a proper treatment of water molecules needs to be included in computational methods to reflect those interactions, with a proper balance between accuracy and computational cost. FEP calculations rely heavily on accurate calculations of solvent-solute interactions to predict binding energies,[85] and are hence run in explicit solvent. On the other hand, docking large libraries of compounds is usually performed using implicit solvation models.[86] Many explicit water models have been developed over the years and are discussed here,[87, 88] the number of interactions sites included being the major difference between models. SPC[89] and TIP3P[87] models include 3 interaction sites (on each atom), while TIP4P[87] includes an additional "dummy-atom" next to the oxygen, and TIP5P[90] includes two additional sites which can be attributed to the lone pairs (Figure 1.3). It is out of the scope of this review to discuss extensively

in what ways water models differ from one and more information on implicit and hybrid models can be found here.[91] A comparative study by *Nguyen et al.* highlights the importance of choosing the correct biomolecular FF/solvation model pair.[92] This is not surprising as parameters for the biological FF are generated using a specific water model during the parametrization of non-bonded parameters. Therefore, these FFs need to be applied with the same water model, AMBER, CHARMM and OPLS typically use TIP3P or TIP4P while GROMOS uses SPC waters.

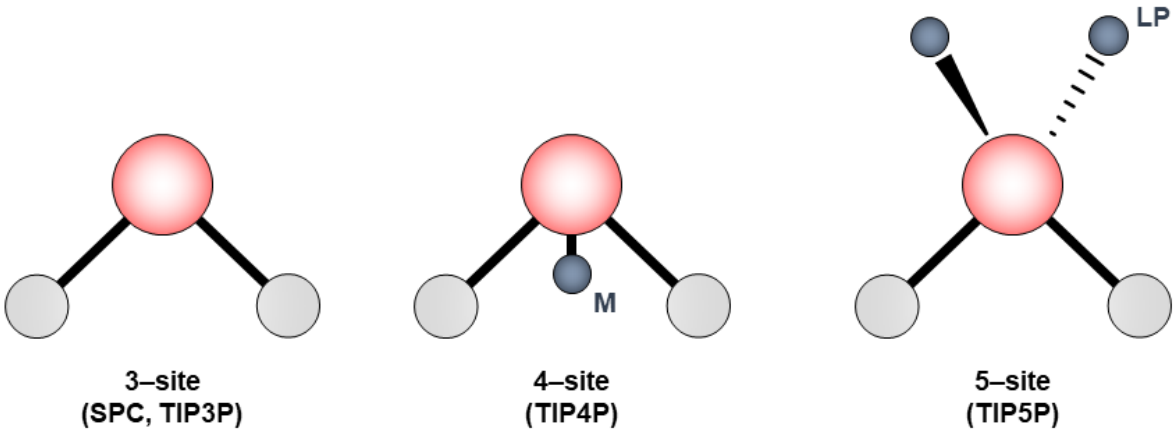


Figure 1.3. Different water models in Molecular Mechanics. The additional charge site in the TIP4P model is referred to as M because the particle is "massless". The additional charge sites in the TIP5P model are referred to as LP (lone pairs).

We will now turn our attention to recent efforts directed towards the improvement of FFs, keeping the modeling of small drug-like molecules at the center of our discussion. First, we will discuss how the potential energy equation used in class I FFs has been challenged, more specifically how treatment of the non-bonded interactions (van der Waals and electrostatics) could be improved by using more physically meaningful models. Then, we will discuss areas of FF development which, might have not received enough attention and may require notable improvements.

1.2 van der Waals Interactions

The van der Waals (vdW) contribution to the energy contains an attractive and repulsive component. The former can be attributed to London dispersion forces; as any two atoms approach each other, the dynamic nature of their electrons leads to the formation of temporary dipoles eventually allowing weak attractive dipole-dipole interactions to occur. This weak attractive force is in competition with a strong repulsion (Pauli exclusion, electrons in close proximity) at very close distances.

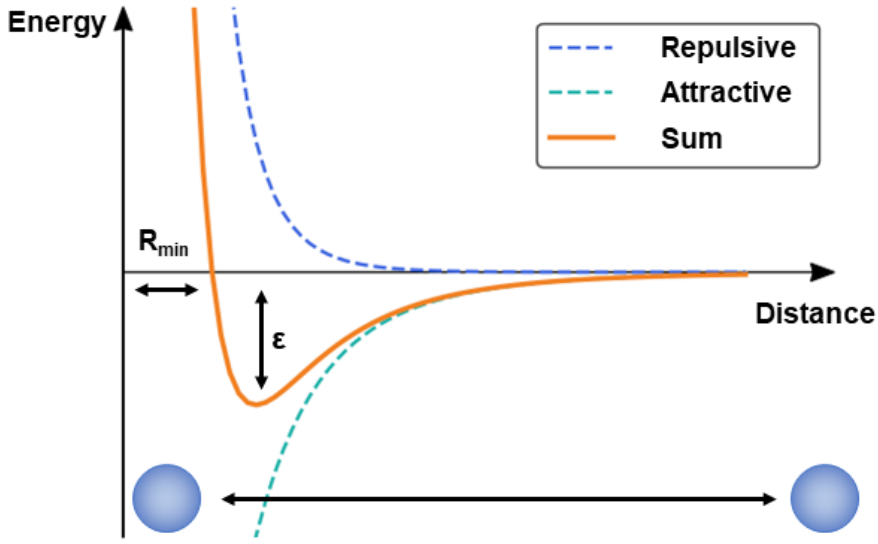


Figure 1.4. Lennard Jones 12-6 potential used to model the van der Waals interactions in class I force fields. Blue spheres represent atoms.

$$E_{Buf-14-7} = \sum_{pairs\ i,j} \boldsymbol{\varepsilon_{ij}} \left(\frac{\mathbf{1} + \boldsymbol{\delta}}{\frac{r_{i,j}}{R_{min,i,j}} + \boldsymbol{\delta}} \right)^{7} \left(\frac{1 + \boldsymbol{y}}{\left(\frac{r_{i,j}}{R_{min,i,j}}\right)^{7} + \boldsymbol{y}} - 2 \right) \tag{1.8}$$

Since vdW interactions need to be calculated for every pair of atoms (alongside the electrostatic contribution), it makes up for a large portion of the computational cost associated with the MM energy. Hence the vdW interaction has typically been modeled using the Lennard Jones (LJ) 12-6 potential (Figure 1.4), although limitations arise such as an overestimated repulsion at

very short distances. To correct for this limitation, some FFs scale down the vdW interaction of atoms separated by 3 bonds, or use a separate set of parameters ($\boldsymbol{\varepsilon}$, $\mathbf{R_{min}}$) for 1,4 interactions. It has been noted however that the derived $1/R^6$ dependence of the dispersion is obtained by resorting to approximations during the derivation and should not be taken as an end all be all rule to follow.[93, 94] Other functional forms have been developed including the exponential-6 form which has not found widespread application due to the increase in computational cost and the need for additional parameters. On the other hand, the buffered 14-7 potential (Buf-14-7, Eq. 1.8) used by MMFF94[95] and AMOEBA (which we will discuss further in the next section),[96] has received more attention. While parameters in MMFF94 were incorrect as a result of fitting to gas phase data only, reparameterization using more appropriate strategies,[97-99] has shown improved performance of the Buf-14-7 potential over the standard LJ 12-6 for organic molecules and nucleic acid base pair stacking.[100] The importance of fitting to condensed phase data can be explained by the fact that vdW interactions in a molecule are not simply pair-wise (as the MM functional calculates them), but also contain many-body effects linked to the chemical environment.[101] These many body interactions are considered implicitly when parameters are fit to experimental liquid properties since these macroscale liquid properties are the result of all interactions at the atomic scale, including many-body interactions.[102]

For the calculation of the vdW energy, each atom type holds two parameters, one related to the well depth ($\boldsymbol{\varepsilon}$), and the other to the distance at which attraction and repulsion cancel each other out ($\mathbf{R_{min}}$). Since vdW interactions involve two atoms, parameters for the interaction are obtained by combining the $\boldsymbol{\varepsilon}$ and $\mathbf{R_{min}}$ (associated to both atom types) using mixing rules such as taking the geometric mean for both (OPLS, GROMOS), and Lorentz-Berthelot mixing rules (AMBER, CHARMM). The Lorentz-Berthelot rules were shown to overestimate well depth (and ultimately liquid densities),[103] although this study was carried out on rare gases only. In the previously mentioned article,[100] *Riu et al.* have compared in detail the impact of multiple mixing rules on different potentials (LJ 12-6, Buf-14-7 and Buckingham) and found that the Buf-14-7 potential coupled with refined mixing rules[104] performed better than current implementations in AMBER and CHARMM. The authors do mention however that the study was performed only using sophisticated QM calculations (SAPT) and serves only as a starting point which should be validated/refined with MD simulations to reproduce liquid properties. The Buf-14-7 potential

contains two additional buffering parameters **δ** and **γ**, which have been set to constant values in the current version of AMOEBA to limit the number of parameters which need to be developed. Overall, the different potentials and mixing rules used explains the non-transferability of vdW parameters from a FF to another, and by extension the non-transferability of torsional parameters.

To conclude this section, we would like to argue that using simplified potentials such as the LJ 12-6 should not remain the state-of-the-art solely for their computational cost advantage in an environment were computational capabilities continue to grow exponentially. Considering the importance of the vdW interaction in biological phenomena,[105, 106] more research should be carried out to provide the best account of vdW forces, ultimately improving the accuracy of binding energy calculations in SBDD related applications. In 2004, MacKerell stated in a review[107] that the LJ 12-6 potential remained adequate for room temperature MD simulations, although this claim was not backed up by any particular study, but rather on the number of successful applications of FFs found in the literature. While the general consensus that the LJ 12-6 potential is sufficient for SDBB applications, it has recently been challenged by the re-apparition of the Buf-14-7 potential which could ultimately provide a more accurate treatment of vdW interactions,[100] allowing to probe more sensitive biological phenomena. Considering the interdependence of FF terms (e.g. vdW, torsions, electrostatics), a considerable modification to one of the terms would require a complete reparameterization of the FF, which explains why AMBER, CHARMM, GROMOS and OPLS have not modified the potential handling the vdW interactions. Finally, errors stemming from one part of the FF (vdW, electrostatics) are usually counterbalanced by other parts of the FF, hence a comparison of the performance of an isolated term doesn't necessarily translate into a better FF overall. This is referred to as the self-consistency of FFs.

1.3 Electrostatic Interactions and Polarizable Models

Class I FFs consider electrostatic interactions using a simple Coulombic additive potential, between fixed point charges located at the center of mass of every atom. The potential is "additive" as the contribution to the energy is calculated as a sum of all pairwise monopole-monopole interactions, without considering effects charges can have on one another (e.g. polarization, correlation, charge transfer), multi-body effects, and a with very crude representation of the charge distribution. As opposed to other interactions, class I FFs do not rely on parameters to calculate electrostatic effects, however they rely on the assignment of partial charges to every atom. We will first present how traditional FFs allocate these partial point charges, followed by a discussion of the wide variety of alternative polarizable models which have emerged to describe electronic interactions more rigorously.

1.3.a Partial Charge Fitting Schemes

Partial charges need to be assigned to all atoms in a molecule before a simulation and are kept constant over the course of the simulation, regardless of conformational changes which are known to impact charge distribution.[108, 109] One of the most widely used schemes to generate partial charges is to fit charges to electrostatic potentials (ESP) computed with ab initio QM methods.[110] This approach has two known drawbacks; not assigning identical partial charges to chemically equivalent atoms, and poorly predicting the partial charges of buried atoms. These limitations were addressed by including restrictions during the fitting process, giving birth to the restricted electrostatic potential (RESP) charging scheme, which predicted DNA base pairing energies and solvation free energies of small molecules more accurately.[111, 112] The RESP scheme remains computationally expensive, and while tabulated parameters exist for proteins and nucleic acids (within AMBER),[51, 113] the method is generally not widely used for small drug-like molecules.

Rule based methods such as bond charge increment (BCI) and electronegativity equalization have been developed with high-throughput applications in mind (orders of magnitude faster than methods relying on QM calculations such as RESP). In BCI (used by CGenFF),[114] formal charges are first assigned to each atom, and then redistributed in a stepwise process based on the atom types, until a convergence criterion is met. Electronegativity equalization methods

follow a similar approach were charges are redistributed until all atoms reach the same electronegativity.[115, 116]

Another option to generate partial charges are semi-empirical methods such as AM1,[117] PM3[118] and the CM1-CM5 series[119-121] which provide comparable accuracies to RESP models, with reduced computational costs. These semi-empirical models have also been combined with rule based methods, the most notable example being AM1-BCC[122, 123] available within AMBER. The performance of AM1-BCC to reproduce free energies of solvation (organic molecules), hydrogen bonding energies (base pair dimers and organic molecules) has been extensively monitored and found to perform on par with RESP while considerably reducing computational costs, making it an excellent candidate to generate partial charges for small drug-like molecules. AM1-BCC charges were also found to be transferable to the OPLS_2005 FF (reproduces solvation free energies well), except for polar molecules (e.g. amides, amines, ethers, etc.).[12] To correct for these classes of molecules, another semi-empirical method with rule-based corrections was thus developed to work in conjunction with the OPLS_2005 FF called CM1A-BCC.[12] Authors do note that a self-consistent FF (i.e. all other parameters generated with CM1A-BCC charges during training), would perform better, and the CM1A-BCC scheme was carried to later versions of OPLS. Indeed, these improvements have been observed in the newest version to date (OPLS3e[75]), in which authors focused on the ability of the FF to predict protein-ligand binding affinities (low RMSE of ~ 1 kcal·mol^{-1} on a very large set of 393 small molecules with different binding partners).

None of the methods presented above are perfect however, and research in this field remains particularly active.[124, 125] On the other hand, an inherent limitation of the Coulombic potential used by class I FFs is that it neglects multi-body effects and polarization. While some of the charging schemes presented (particularly the newer) try to incorporate these effects implicitly,[112, 125] a radically different approach consists in having a potential energy function which can treat polarization explicitly.

1.3.b Polarizable Electrostatic Models

The notion that the simplistic point charge model used to describe electrostatic interactions could be improved by partially representing electron delocalization using multiple charge sites has been demonstrated before any application in the context of FFs. For example, in 1990, *Hunter et al.* showed that distributing charges in aromatic systems using tripoles (positive charge on atoms, negative charge above and below every atom in the π-system) gave a better account of π-π stacking than describing the system with a single point charge on every atom.[126] Point charges are equivalent to a symmetric spherical distribution of the charge density around the point charge, thereby neglecting any potential anisotropy of the charge distribution, affecting hydrogen bonding,[127] halogen bonding (σ-hole),[128] π-π stacking[129] and cation-π interactions,[130] which are ubiquitous in biological systems. In addition to the limitation associated with a poor description of charge localization, non-polarizable models are notoriously deficient in two major aspects: **1**) by retaining the same partial charges on atoms throughout simulations, in spite of the known dependence of charges on conformations[131] and **2**) by neglecting the influence of inter/intramolecular effects on the charge distribution, which are central in protein-ligand interactions,[132] protein-protein interactions during folding,[133] and energetics of ion conduction through channels,[134, 135] among others.

Although polarizable methods could provide more accurate depictions of biological processes, they require a more extensive parametrization and have been estimated to be 3 to 10 times more computationally expensive (depending on the specific implementation of polarization).[136] The computational drawback has been a long-lasting argument against polarizable methods, however recent hardware and software improvements such as the implementation of MD codes on GPU architectures,[137] and the introduction of the particle mesh Ewald (PME) algorithm to evaluate electrostatic energies,[138] have somewhat alleviated these computational cost restrictions. We will now turn our attention towards the various attempts to include electronic polarization explicitly in FFs. First, by describing briefly how polarization is implemented in those models and finally discuss applications and outlooks of these new methods.

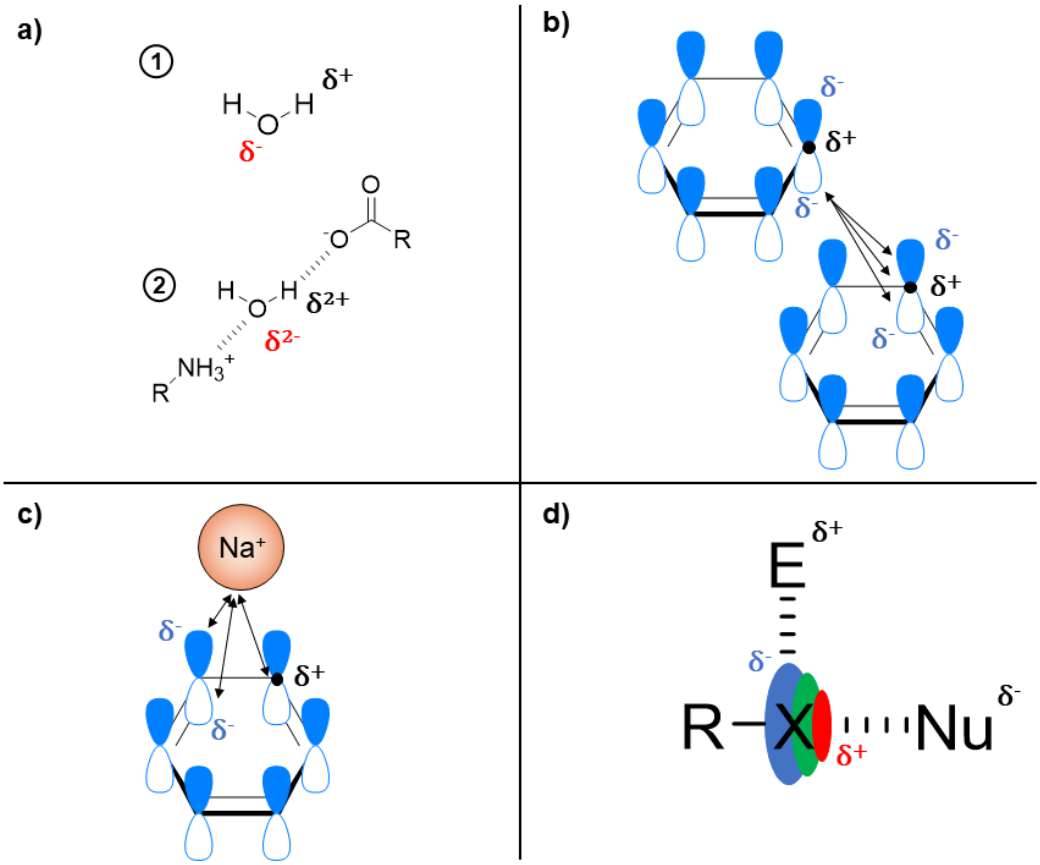


Figure 1.5. Different interactions in which polarization plays an important role: a) hydrogen bonding, b) π-π stacking, c) cation-π interactions and d) halogen bonding. In d) X represents a halogen (F, Cl, Br, I), E is an electrophile and Nu is a nucleophile.

1.3.c Fluctuating Charge and Drude Oscillator Models

Polarization is the process by which the charge distribution in a molecule changes in response to the environment.[139] The fluctuating charge (FQ), and Drude oscillator models incorporate polarization while retaining the simple Coulombic potential of non-polarizable FFs. They both allow charges to vary over the course of a simulation (thereby targeting problem **1** mentioned above), but only the Drude model considers the anisotropic nature of the electron distribution thus additionally addressing problem **2**.

In the FQ model, the charges are simply reassigned after each simulation step, to reflect the effect of the environment on the charge distribution. The movement of charges is calculated using charge equilibration (Qeq) schemes based on the notion of electronegativity equalization, which are analogous to those employed to derive partial charges.[140] An implementation of a FQ model in CHARMM is available and covers a wide range of molecules (proteins, lipids, carbohydrates), although the method has not yet been parameterized for small drug-like molecules yet.[141, 142]

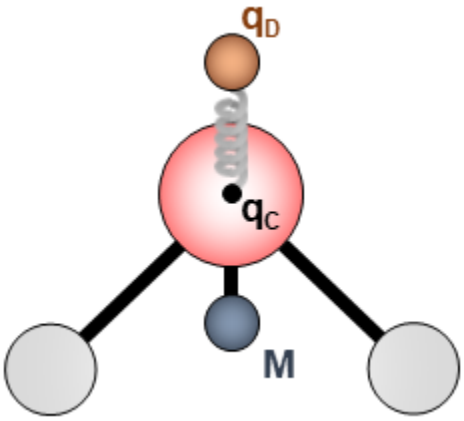


Figure 1.6. Drude oscillator model for a water molecule. Note that the water model is built from TIP4P (hence the additional M charge site). q_C is the charge at the oxygen nucleus, q_D is the charge of the Drude particle attached to the oxygen. For graphical purposes, q_D is shown far from the atomic nucleus, however in practice Drude particles are close to the nucleus.

In the Drude oscillator model, polarization is included through the addition of charges sites. More specifically, every non-hydrogen atom holds two-point charges (i.e. dipole description), which are connected by a spring (Figure 1.6). The first charge is fixed at the atomic nucleus, while the other (Drude particle) is free to move in response to the external electric field. An advantage of this fluctuating dipole model is that it can be interpreted chemically (i.e. charges represent the nucleus and electron density respectively). Nevertheless, the description of the charge density by a single point remains a clear simplification. In order to limit the computational cost, Drude particles are typically assigned to non-hydrogen atoms only. An implementation of that method can be found in CHARMM (Drude-2013), which currently covers proteins, DNA, lipids, and carbohydrates.[143, 144] A few small organic molecules were also included during the parametrization, although the FF cannot be (and has not been) applied to any diverse set of drug-like molecules yet.

1.3.d Polarization through Multipole Expansion

Another approach aimed at describing anisotropic features of the electron density is through the inclusion of polarization using atomic multipole moments.[145] These techniques tackle both problems **1** and **2** discussed above but require the addition of a polarization term to the FF equation to cover charge-dipole, dipole-dipole interactions (and more if the multipole expansion is carried further).

The first comprehensive implementation of a polarizable FF including multipole expansion (up to dipoles), called ff02 was issued by *Cieplak et al.* within the AMBER package in 2002.[146] Subsequent efforts to provide a more robust parameterization were later shown to improve the accuracy of amino acid intermolecular interaction energies.[147] More recently, a few applications of ff02 to probe protein-protein[148] and protein-ligand[149] interactions were found, however the ligands were short peptide chains, highlighting the absence of parameters for small drug-like molecules.

Arguably, the most popular FF including multipolar electrostatics (up to quadrupoles) is AMOEBA in which electrostatics are calculated as a sum of permanent and induced multipoles. AMOEBA is the polarizable FF which currently covers the largest portion of chemical space. Initially developed for water,[150] the FF was supplemented over the years allowing the treatment of ions,[151] proteins,[152] nucleic acids,[153] and small organic molecules.[96] The performance of AMOEBA towards small organic compounds was assessed: In the gas phase, dimer equilibrium structures and dimer binding energies conformed to QM results. Condensed phase properties (ρ, ΔH_{vap}) were found to be well reproduced, most notably the hydration free energies of 27 molecules are in close agreement to experimental values (RMSE = 0.69 kcal·mol^{-1}).[96] However, the FF hasn't been fully automated, and assigning partial charges and parameters still requires manual involvement, which for now prohibits any high-throughput application.

1.3.e Beyond Point Charge Representations

More recently, approaches to describe polarization while capturing the three-dimensional nature of the electron distribution have emerged. *Donchev et al.* have issued QMPFF which uses a decaying exponential function to describe part of the electrostatic energy,[154] and *Naserifar et al.* have used Gaussian functions in RexPoN, a reactive and polarizable FF.[155-157] The exponential and Gaussian type functions are analogous to the Slater and Gaussian type orbitals used as basis functions in QM calculations (Figure 1.7). Both of these FFs are particularly interesting because they challenge the traditional FF parametrization strategy where condensed phase properties are used as target data. Indeed, both QMPFF and RexPoN use only gas phase QM derived properties as training data, as opposed to condensed phase properties as discussed previously. Authors of QMPFF argue that including polarization in the FF potential energy function, making it more physically grounded leads to a greater transferability from gas phase to condensed phase. This claim has been confirmed by the excellent agreement of calculated condensed properties of water by RexPoN with experimental properties. Additionally, promising quantitative agreement with experiment of protein-ligand binding affinities were obtained by QMPFF,[158] although its performance was only compared to MMFF94 on a very small subset of molecules.

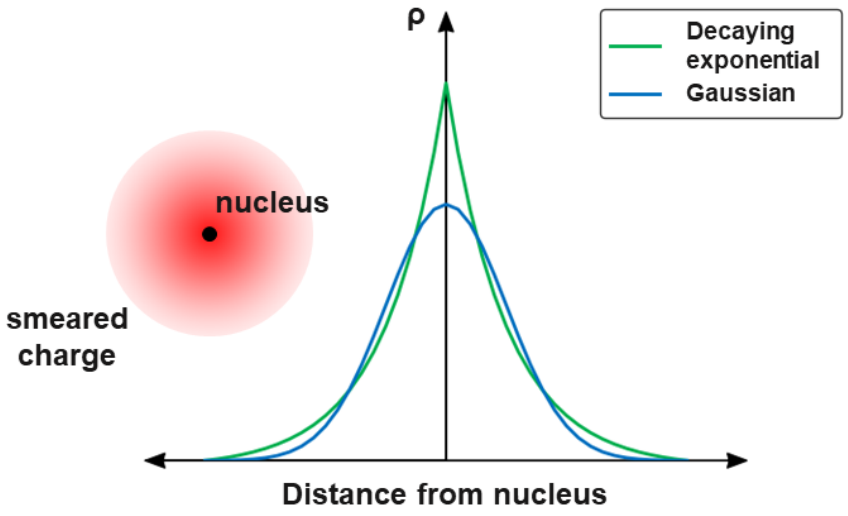


Figure 1.7. Three-dimensional nature of the electron distribution using decaying exponential and gaussian functions.

Overall, QMPFF has not received a lot of attention since its initial publication in 2005, probably because the parametrization process is particularly costly, and has hence not been extended to many molecules. Finally, efforts by *Naserifar et al.* in deriving the RexPoN force field are very recent (2018), and applications of the method by other groups were not found. It should be noted that inclusion of exponentials in the potential energy function considerably increases the associated computational costs (more so than computing more charge/charge pairs with the inclusion of multipoles), hence the applicability of these methods to carry out high throughput tasks remains to be validated.

1.3.f Successful Applications of Polarizable Force Fields and Perspectives

Altogether, class I FFs cannot describe electrostatic interactions accurately due to the inherent limitation of the simple point charge model. Polarizable FFs can provide a gain in accuracy at the cost of a more complex parametrization and longer simulation times. Polarizable methods have repeatedly been shown to provide accurate results in multiple contexts such as prediction of ligand-protein binding energies,[159-161] and stable MD simulations of proteins.[152, 162] A more physically grounded description of electrostatics also leads to a better transferability of the associated parameters, as demonstrated by the amino acid parameters in AMOEBA.[152] Furthermore, the Drude model within CHARMM was used to perform long scale (μs) simulations of two small proteins (ubiquitin and CspA), showcasing that long scale simulations are possible.[163]

On one hand, researchers building non-polarizable methods argue that the computational cost increase and more extensive parametrization associated with polarizable models and comparable accuracies obtained thus far do not justify the use of such methods. For example, the Drude-CHARMM and non-polarizable CHARMM36 models displayed similar trajectories when simulating 10 proteins for 100 ns.[162] The ability of AMOEBA to predict the solvation free energy of various small molecules is multiple solvents (acetonitrile, chloroform, DMSO and toluene), was also found to be in par with GAFF, a non polarizable FF.[164] On the other hand, polarizable FFs developers argue that polarizable FFs have not received the same extensive parametrization and are expected to outperform their non-polarizable counterparts once proper parametrization is achieved. A study by *Lindorff-Larsen et al.* supports this idea as they show that more recent FFs perform better than older FFs, simply because more appropriate parameters are available.[165] Hence

a comparable accuracy with minimal parameter refinement should be seen as a success, and polarizable models are expected to describe relevant biological events more accurately once they have been specifically parametrized to do so (e.g. recent addition of parameters for halogenated ligands).[166]

There are currently no extensive comparative studies (polarizable vs. non-polarizable FFs) in the literature performed by authors that are not directly involved in the development of polarizable methods. Similarly, most applications of these new FFs have been performed within the laboratories they have been developed in. Comparative studies performed by 3rd parties are highly encouraged to diagnose areas in which polarizable FFs could be improved without bias or in which they outperform non-polarizable FFs. Overall, the accuracy of polarizable FFs are expected to grow as more parameters are introduced. To this day, polarizable FFs have not been extensively parametrized to cover drug-like molecule space and cannot be applied to high-throughput SBDD tasks. Finally, it would not be surprising to see both types of FFs applied in different contexts, polarizable FFs could be used to provide very accurate binding free energies (e.g. FEP calculations) as molecular recognition is known to be sensitive to polarization. Whereas more demanding simulations of very large proteins, or flexible proteins such as IDPs[167] requiring a more extensive sampling of conformational space, could be performed using faster non-polarizable FFs.

1.4 Torsional Parameters and Transferability

It should not have eluded to our reader that during our extensive discussion of electrostatics in FFs, we reported that many of these methods were "not yet applicable to drug-like molecules". By that, we meant that parameters from one class (or more) were missing to describe the PES of these molecules accurately. In fact, torsional parameters are the main culprit, as bonds and angles are generally assumed to be fully covered,[80] vdW parameters are obtained using the mixing rules detailed previously and electrostatics do not rely on parameters *per se*. As of today, non-polarizable methods also fail to cover chemical space completely (although to a lesser extent). As previously stated, authors of OPLS3 estimated their FF to cover only 2/3rd of drug-like molecules, although the training set used consisted of 6,500 molecules (only surpassed by their latest

publication in which 20,000 torsional profiles were generated and parameterized).[75, 83] Generally, not all classes of molecules are subject to missing parameters. Biological molecules such as proteins and nucleic acids consist of repeating units (i.e. amino acids and base pairs) which eases their parametrization. Once each building block has been assigned parameters, all polymers built from these repeating units can be simulated. In fact, the parametrization for these groups is more often performed with greater care than drug-like molecules (e.g. using more robust QM functionals, sampling larger portions of conformational space).

On the other hand, drug-like molecule space is far more diverse[40] and its sheer size posses problems to the FF parametrization based on atom types. The general trend has been to increase the number of atom types to cover more diverse functional groups,[80] however this cannot be carried indefinitely. In fact, millions of different torsion parameters would be required to fully cover a FF containing 50 atom type definitions (far less than the 139 currently in CHARMM/CGenFF or 124 in OPLS3). To date, the largest scale parametrization (20,000 torsional profiles) is still orders of magnitude smaller than the millions of parameters required, which remains (and will continue to be) computationally prohibitive. More realistically, not all of these different torsions need to be specifically parametrized, and researchers have relied on the assumption that parameters obtained for particular functional groups would be *transferable* to other comparable functional groups (i.e. within similar chemical environments). Although, in practice the validity of this assumption can be questioned.

As a temporary solution to that problem, multiple automated toolkits have emerged, allowing to generate missing parameters (e.g. GAAMP,[168] ffTK,[169] Paramfit[170] and Parmscan[171]). These tools are particularly fit for purposes were few parameters need to be obtained, for example to parametrize a ligand prior to an FEP calculation to predict a protein-ligand binding affinity. On the other hand, these methods cannot be carried to high-throughput tasks (e.g. parametrizing large libraries of potential ligands), as they rely on time consuming QM calculations. While these tools permit previously inaccessible applications of FFs, they do not attempt to solve the problem of poor parameter transferability and poor coverage of chemical space at a fundamental level. To address this foundational problem, general FFs were built to specifically cover small molecule space (e.g. GAFF, CGenFF), however their accuracies and coverage remain questionable. Indeed,

in one of our recent studies, we have shown that the agreement between QM and GAFF torsional profiles on a diverse set of small organic molecules were very unsatisfactory (average RMSE of 1.74 kcal·mol^{-1} for 1,000 profiles).[172]

A number of technical problems can arise from the traditional conception of FFs based on atom types. For example, if new hydration free energy data suggests that a new atom type needs to be introduced to model the vdW interactions of a particular moiety, then new valence parameters to describe all of the bonds, angles and torsions these new atom types can be involved in are immediately required. Often these new parameters were copied from "parent" parameters without a proper basis for these choices, and led to unnecessary redundancies.[173] *Mobley et al.* have discussed how the choice of atom type definitions has traditionally been determined from chemical intuition without a rigorous basis supporting these choices.[173] They have suggested replacing human annotated features (here atom types) by fully automated ones in their newly developed method SMIRNOFF. Although the label designating each atom is still referred to as an atom type in their scheme, the way in which they are assigned (through direct chemical perception) is radically different and does not involve subjective rules. More concretely, SMIRNOFF uses the SMIRKS language to describe all atoms by strings of characters. Each SMIRKS string contains information about the atom (i.e. element, hybridization, connectivity) and can be further "decorated" to describe the chemical environment up to two bonds away. Allowing for example to discern between a sp^3 carbon atom bonded to electron withdrawing elements (e.g. -CF_3), and an sp^3 carbon bonded to carbons/hydrogens (e.g. -CH_3), which is something traditional atom typing methodologies did not account for (Figure 1.8). Further, when generating bond-types, angle-types, etc. by combining the atom types which make them up, SMIRNOFF also considers the bond order (single, double, triple) as part of the representation, as opposed to traditional FFs (bond orders are implied from the atom types, leading to known problems[173]). As a result, the parameter file in SMIRNOFF is roughly 300 lines long as opposed to the roughly 6,500 lines in GAFF. The few parameters in their method were obtained from previous FFs (namely GAFF and Parm@Frosst[174]) and SMIRNOFF was found to perform on par with GAFF to predict the hydration free energies of 642 small molecules. Novel parametrization methodologies to work in accordance with the new SMIRKS representation of molecules are currently under development.[175]

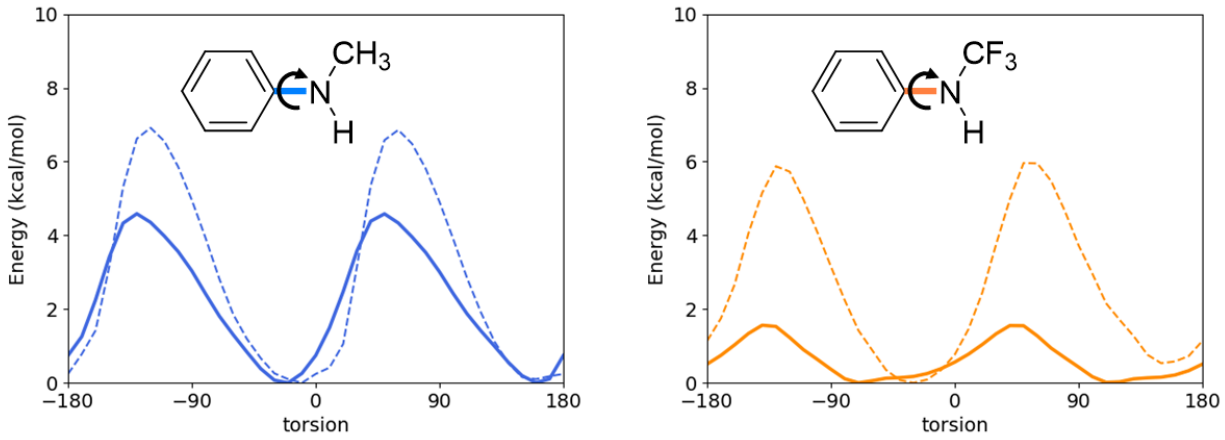


Figure 1.8. Comparison of the torsional profiles of two anilines. Bold line corresponds to QM energy (MP2/6-311+G**) and dashed line corresponds to GAFF energy. The same GAFF atom types are assigned to atoms involved in the torsion (namely: ca, nh, c3 and hn) which is highlighted by the fact both GAFF (dashed) profiles are very similar. This is not in accordance with QM calculations however, in which inductive effects lead to a very large decrease in the torsional barrier's height.

The poor transferability of torsional parameters can in part be attributed to the fact they are usually expected to empirically correct for errors emerging from other parts of the FFs, most notably 1,4 non-bonded interactions. Atom types are also responsible for the non-transferability of torsional parameters as portions of an atom's chemical environment are neglected when assigning atom types. Using the same example as above, in GAFF the carbon in a -CH_3 or a -CF_3 moiety would be assigned the same **c3** atom type (representing all sp^3 carbons), which does not reflect the strong inductive effects fluorine atoms have on the carbon, ultimately affecting torsional profiles (Figure 1.8). It is generally believed that functions reflecting physically grounded interactions would be more transferable.[158] In this context, our lab is currently developing a novel FF which does not rely on atom types called H-TEQ.[172, 176] First, we have shown that hyperconjugation was the major contributor to the torsional profiles of small saturated molecules. Then, we have replaced the classical torsional term (in GAFF) by a hyperconjugation term and found that H-TEQ reproduced the torsional energy profiles of about 1,000 small organic molecules more accurately than GAFF. As expected, a more physically grounded torsional term (based on hyperconjugation) led to a better transferability of our method. We have observed one main drawback from H-TEQ, its reliance on other terms of the FF (we kept other terms as calculated by GAFF during our analyses). Specifically, our method did not empirically correct for inaccurate 1,4 non-bonded

interactions. Nevertheless, we expect H-TEQ to be more robust when applied in conjunction with more accurate vdW and electrostatic terms (e.g. polarizable models). H-TEQ is currently applicable to virtually all saturated small molecules. In chapter 2 of this thesis, we will describe our efforts to extend the coverage of H-TEQ to molecules containing unsaturations.

Overall, atom type based FFs have been employed with great success to simulate proteins and nucleic acids. However, the vastness of drug-like molecule space and poor transferability of torsional parameters remains a major challenge for high-throughput SBDD applications relying on FFs. Articles in the literature discussing the link between atom typing and parameter transferability are scarce. Likewise, to the best of our knowledge, only two techniques (SMIRNOFF and H-TEQ) moving away from an atom type based parameter assignment have emerged as potential substitutes. Both FFs are relatively new and have not been extensively tested yet. Nonetheless they appear as promising candidates to cover greater portions of chemical space with greater accuracies, provided that continuous efforts are conducted to supplement their current abilities.

1.5 Additional Liabilities of Force Fields

The majority of efforts in FF development have been directed towards torsions and electrostatic interactions, as both were recognized to require considerable refinement. Additionally, as the number of applications of FFs in various context grows, additional limitations come to light. For example, the modeling of nitrogen containing compounds is particularly challenging. *Vitaku et al.* have observed that 84% of all FDA approved drugs contain at least one nitrogen atom, 59% contain at least a nitrogen heterocycle, and on average a drug contains 2.3 nitrogen atoms.[177] Considering the prevalence of nitrogen containing drugs, any successful SBDD application is bound to pay particular attention to these moieties during the parametrization. A good example is the inclusion of an additional charge site to nitrogen atoms (corresponding to a lone pair) within aromatic rings in OPLS3,[83] to better represent their electrostatic interactions. In chapter 3 of this thesis, we will discuss in detail how current FFs fail to model molecules containing conjugated nitrogen substituents and offer potential solutions to correct for these drawbacks.

Furthermore, increases in computational power have allowed for longer MD simulations to be carried, and deficiencies of FFs with respect to the kinetics of biomolecules have become apparent.[80, 178-180] Precise experimental thermodynamic data being more readily available explains why FF have generally used it over kinetic data for parametrization. Hence, using more experimental kinetic data during the parametrization of FFs should naturally improve their ability to simulate biochemical events.

1.6 Conclusions

Molecular mechanics based methods are widely applied to SBDD projects. The popularity of FFs and their simple functional form over QM methods is largely due to their suitable accuracies and greatly reduced computational costs. It is now anticipated that the more accurate functionals of polarizable FFs would provide deeper insights into drug-target binding modes, once proper parametrization is undergone, ultimately improving the reliability of FFs methods in the design of new pharmaceutical compounds. By the same token, it is likely that the current state-of-the-art additive class I FFs continue to see widespread use, particularly to study larger molecular systems, for longer periods of time. The development of novel methodologies which do not rely on atom types are also expected to provide an increased transferability of parameters, and a greater coverage of chemical space would allow to scan more varied libraries of potential therapeutic compounds. Finally, additional insights onto the weaknesses of FFs will be obtained as they are applied in more projects, ultimately guiding the next rounds of refinement.

References

1. Jorgensen, W. L., The Many Roles of Computation in Drug Discovery. *Science* **2004**, 303, 1813-1818.

2. MacCoss, M.; Baillie, T. A., Organic Chemistry in Drug Discovery. *Science* **2004**, 303, 1810-1813.

3. Tian, S.; Wang, J.; Li, Y.; Li, D.; Xu, L.; Hou, T., The Application of in Silico Drug-Likeness Predictions in Pharmaceutical Research. *Adv. Drug Deliv. Rev.* **2015**, 86, 2-10.

4. Daina, A.; Michielin, O.; Zoete, V., Swissadme: A Free Web Tool to Evaluate Pharmacokinetics, Drug-Likeness and Medicinal Chemistry Friendliness of Small Molecules. *Sci. Rep.* **2017**, 7, 42717.

5. Hay, M.; Thomas, D. W.; Craighead, J. L.; Economides, C.; Rosenthal, J., Clinical Development Success Rates for Investigational Drugs. *Nat. Biotechnol.* **2014**, 32, 40.

6. Kuntz, I. D., Structure-Based Strategies for Drug Design and Discovery. *Science* **1992**, 257, 1078-1082.

7. Dans, P. D.; Walther, J.; Gomez, H.; Orozco, M., Multiscale Simulation of DNA. *Curr. Opin. Struct. Biol.* **2016**, 37, 29-45.

8. Wang, Z.; Sun, H. Y.; Yao, X. J.; Li, D.; Xu, L.; Li, Y. Y.; Tian, S.; Hou, T. J., Comprehensive Evaluation of Ten Docking Programs on a Diverse Set of Protein-Ligand Complexes: The Prediction Accuracy of Sampling Power and Scoring Power. *Phys. Chem. Chem. Phys.* **2016**, 18, 12964-12975.

9. Halgren, T. A., Potential-Energy Functions. *Curr. Opin. Struct. Biol.* **1995**, 5, 205-210.

10. Lazaridis, T.; Karplus, M., Effective Energy Functions for Protein Structure Prediction. *Curr. Opin. Struct. Biol.* **2000**, 10, 139-145.

11. Zoete, V.; Grosdidier, A.; Michielin, O., Docking, Virtual High Throughput Screening and in Silico Fragment-Based Drug Design. *J. Cell. Mol. Med.* **2009**, 13, 238-248.

12. Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W., Prediction of Absolute Solvation Free Energies Using Molecular Dynamics Free Energy Perturbation and the Opls Force Field. *J. Chem. Theory Comput.* **2010**, 6, 1509-1519.

13. Jiang, W.; Roux, B., Free Energy Perturbation Hamiltonian Replica-Exchange Molecular Dynamics (Fep/H-Remd) for Absolute Ligand Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2010**, 6, 2559-2565.

14. Ballester, P. J.; Mitchell, J. B. O., A Machine Learning Approach to Predicting Protein–Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* **2010**, 26, 1169-1175.

15. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, 66, 334-395.

16. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., Ff14sb: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99sb. *J. Chem. Theory Comput.* **2015**, 11, 3696-3713.

17. Kenno, V.; Olgun, G.; Alexander, D. M., Jr., Molecular Mechanics. *Curr. Pharm. Des.* **2014**, 20, 3281-3292.

18. van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A., Reaxff: A Reactive Force Field for Hydrocarbons. *J. Phys. Chem. A* **2001**, 105, 9396-9409.

19. Nielson, K. D.; van Duin, A. C. T.; Oxgaard, J.; Deng, W. Q.; Goddard, W. A., Development of the Reaxff Reactive Force Field for Describing Transition Metal Catalyzed Reactions, with Application to the Initial Stages of the Catalytic Formation of Carbon Nanotubes. *J. Phys. Chem. A* **2005**, 109, 493-499.

20. Chenoweth, K.; Cheung, S.; van Duin, A. C. T.; Goddard, W. A.; Kober, E. M., Simulations on the Thermal Decomposition of a Poly(Dimethylsiloxane) Polymer Using the Reaxff Reactive Force Field. *J. Am. Chem. Soc.* **2005**, 127, 7192-7202.

21. Rahaman, O.; van Duin, A. C. T.; Goddard, W. A.; Doren, D. J., Development of a Reaxff Reactive Force Field for Glycine and Application to Solvent Effect and Tautomerization. *J. Phys. Chem. B* **2011**, 115, 249-261.

22. Monti, S.; Corozzi, A.; Fristrup, P.; Joshi, K. L.; Shin, Y. K.; Oelschlaeger, P.; van Duin, A. C. T.; Barone, V., Exploring the Conformational and Reactive Dynamics of Biomolecules in Solution Using an Extended Version of the Glycine Reactive Force Field. *Phys. Chem. Chem. Phys.* **2013**, 15, 15062-15077.

23. Liu, J.; Li, X. X.; Guo, L.; Zheng, M.; Han, J. Y.; Yuan, X. L.; Nie, F. G.; Liu, X. L., Reaction Analysis and Visualization of Reaxff Molecular Dynamics Simulations. *J. Mol. Graph. Model.* **2014**, 53, 13-22.

24. Allinger, N. L. Calculation of Molecular Structure and Energy by Force-Field Methods. In *Advances in Physical Organic Chemistry*, Gold, V.; Bethell, D., Eds.; Academic Press: 1976; Vol. 13, pp 1-82.

25. Tuckerman, M.; Laasonen, K.; Sprik, M.; Parrinello, M., Ab-Initio Molecular-Dynamics Simulation of the Solvation and Transport of Hydronium and Hydroxyl Ions in Water. *J. Chem. Phys.* **1995**, 103, 150-161.

26. Taylor, R. D.; Jewsbury, P. J.; Essex, J. W., A Review of Protein-Small Molecule Docking Methods. *J. Comput. Aid. Mol. Des.* **2002**, 16, 151-166.

27. Walters, W. P.; Stahl, M. T.; Murcko, M. A., Virtual Screening—an Overview. *Drug Discov. Today* **1998**, 3, 160-178.

28. Csermely, P.; Palotai, R.; Nussinov, R., Induced Fit, Conformational Selection and Independent Dynamic Segments: An Extended View of Binding Events. *Trends Biochem. Sci.* **2010**, 35, 539-546.

29. Durrant, J. D.; McCammon, J. A., Molecular Dynamics Simulations and Drug Discovery. *Bmc Biol.* **2011**, 9.

30. Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K., Molecular Dynamics Simulations of the Complete Satellite Tobacco Mosaic Virus. *Structure* **2006**, 14, 437-449.

31. Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K., Ten-Microsecond Molecular Dynamics Simulation of a Fast-Folding Ww Domain. *Biophys. J.* **2008**, 94, L75-L77.

32. Best, R. B.; Hummer, G.; Eaton, W. A., Native Contacts Determine Protein Folding Mechanisms in Atomistic Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, 110, 17874-17879.

33. Jogini, V.; Roux, B., Dynamics of the Kv1.2 Voltage-Gated K+ Channel in a Membrane Environment. *Biophys. J.* **2007**, 93, 3070-3082.

34. Khalili-Araghi, F.; Gumbart, J.; Wen, P.-C.; Sotomayor, M.; Tajkhorshid, E.; Schulten, K., Molecular Dynamics Simulations of Membrane Channels and Transporters. *Curr. Opin. Struc. Biol.* **2009**, 19, 128-137.

35. Senn, H. M.; Thiel, W., Qm/Mm Methods for Biomolecular Systems. *Angew. Chem. Int. Ed.* **2009**, 48, 1198-1229.

36. Buch, I.; Giorgino, T.; De Fabritiis, G., Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, 108, 10184-10189.

37. Tobias, D. J.; Tu, K.; Klein, M. L., Atomic-Scale Molecular Dynamics Simulations of Lipid Membranes. *Curr. Opin. Colloid Interface Sci.* **1997**, 2, 15-26.

38. Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W., Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, 330, 341-346.

39. Vanommeslaeghe, K.; Yang, M. J.; MacKerell, A. D., Robustness in the Fitting of Molecular Mechanics Parameters. *J. Comput. Chem.* **2015**, 36, 1083-1101.

40. Dobson, C. M., Chemical Space and Biology. *Nature* **2004**, 432, 824-828.

41. Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M., Uff, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, 114, 10024-10035.

42. Martin, M. G., Comparison of the Amber, Charmm, Compass, Gromos, Opls, Trappe and Uff Force Fields for Prediction of Vapor-Liquid Coexistence Curves and Liquid Densities. *Fluid Ph. Equilibria* **2006**, 248, 50-55.

43. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, 25, 1157-1174.

44. Showalter, S. A.; Brüschweiler, R., Validation of Molecular Dynamics Simulations of Biomolecules Using Nmr Spin Relaxation as Benchmarks: Application to the Amber99sb Force Field. *J. Chem. Theory Comput.* **2007**, 3, 961-975.

45. Zgarbova, M.; Otyepka, M.; Sponer, J.; Mladek, A.; Banas, P.; Cheatham, T. E.; Jurecka, P., Refinement of the Cornell Et Al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.* **2011**, 7, 2886-2902.

46. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D., Charmm General Force Field: A Force Field for Drug-Like Molecules Compatible with the Charmm All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, 31, 671-690.

47. Pulay, P.; Fogarasi, G.; Pongor, G.; Boggs, J. E.; Vargha, A., Combination of Theoretical Ab Initio and Experimental Information to Obtain Reliable Harmonic Force Constants. Scaled Quantum Mechanical (Qm) Force Fields for Glyoxal, Acrolein, Butadiene, Formaldehyde, and Ethylene. *J. Am. Chem. Soc.* **1983**, 105, 7037-7047.

48. Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T., Development of an Improved Four-Site Water Model for Biomolecular Simulations: Tip4p-Ew. *J. Chem. Phys.* **2004**, 120, 9665-9678.

49. Kaminski, G.; Jorgensen, W. L., Performance of the Amber94, Mmff94, and Opls-Aa Force Fields for Modeling Organic Liquids. *J. Phys. Chem.* **1996**, 100, 18010-18013.

50. Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P., A New Force-Field for Molecular Mechanical Simulation of Nucleic-Acids and Proteins. *J. Am. Chem. Soc.* **1984**, 106, 765-784.

51. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1996**, 118, 2309-2309.

52. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M., Charmm: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, 4, 187-217.

53. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, 102, 3586-3616.

54. Scott, W. R. P.; Hunenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Kruger, P.; van Gunsteren, W. F., The Gromos Biomolecular Simulation Program Package. *J. Phys. Chem. A* **1999**, 103, 3596-3607.

55. Daura, X.; Mark, A. E.; van Gunsteren, W. F., Parametrization of Aliphatic Chn United Atoms of Gromos96 Force Field. *J. Comput. Chem.* **1998**, 19, 535-547.

56. Damm, W.; Frontera, A.; TiradoRives, J.; Jorgensen, W. L., Opls All-Atom Force Field for Carbohydrates. *J. Comput. Chem.* **1997**, 18, 1955-1970.

57. Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J., Development and Testing of the Opls All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, 118, 11225-11236.

58. Lifson, S.; Warshel, A., Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and N-Alkane Molecules. *J. Chem. Phys.* **1968**, 49, 5116-5129.

59. Hwang, M. J.; Stockfisch, T. P.; Hagler, A. T., Derivation of Class Ii Force Fields. 2. Derivation and Characterization of a Class Ii Force Field, Cff93, for the Alkyl Functional Group and Alkane Molecules. *J. Am. Chem. Soc.* **1994**, 116, 2515-2525.

60. Allinger, N. L., Conformational Analysis. 130. Mm2. A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms. *J. Am. Chem. Soc.* **1977**, 99, 8127-8134.

61. Allinger, N. L.; Yuh, Y. H.; Lii, J. H., Molecular Mechanics. The Mm3 Force Field for Hydrocarbons. 1. *J. Am. Chem. Soc.* **1989**, 111, 8551-8566.

62. Allinger, N. L.; Chen, K.; Lii, J.-H., An Improved Force Field (Mm4) for Saturated Hydrocarbons. *J. Comput. Chem.* **1996**, 17, 642-668.

63. Nevins, N.; Lii, J.-H.; Allinger, N. L., Molecular Mechanics (Mm4) Calculations on Conjugated Hydrocarbons. *J. Comput. Chem.* **1996**, 17, 695-729.

64. Halgren, T. A., Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of Mmff94. *J. Comput. Chem.* **1996**, 17, 490-519.

65. Halgren, T. A., Merck Molecular Force Field. Iii. Molecular Geometries and Vibrational Frequencies for Mmff94. *J. Comput. Chem.* **1996**, 17, 553-586.

66. Wang, Q.; Pang, Y.-P., Preference of Small Molecules for Local Minimum Conformations When Binding to Proteins. *PLoS One* **2007**, 2, e820-e820.

67. Perola, E.; Charifson, P. S., Conformational Analysis of Drug-Like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization Upon Binding. *J. Med. Chem.* **2004**, 47, 2499-2510.

68. Goursot, A.; Mineva, T.; Vásquez-Pérez, J. M.; Calaminici, P.; Köster, A. M.; Salahub, D. R., Contribution of High-Energy Conformations to Nmr Chemical Shifts, a Dft-Bomd Study. *Phys. Chem. Chem. Phys.* **2013**, 15, 860-867.

69. Nerenberg, P. S.; Head-Gordon, T., New Developments in Force Fields for Biomolecular Simulations. *Curr. Opin. Struc. Biol.* **2018**, 49, 129-138.

70. Pérez, A.; Marchán, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M., Refinement of the Amber Force Field for Nucleic Acids: Improving the Description of A/Γ Conformers. *Biophys. J.* **2007**, 92, 3817-3829.

71. Dickson, C. J.; Madej, B. D.; Skjevik, A. A.; Betz, R. M.; Teigen, K.; Gould, I. R.; Walker, R. C., Lipid14: The Amber Lipid Force Field. *J. Chem. Theory Comput.* **2014**, 10, 865-879.

72. Vanommeslaeghe, K.; MacKerell, A. D., Automation of the Charmm General Force Field (Cgenff) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, 52, 3144-3154.

73. Reif, M. M.; Hünenberger, P. H.; Oostenbrink, C., New Interaction Parameters for Charged Amino Acid Side Chains in the Gromos Force Field. *J. Chem. Theory Comput.* **2012**, 8, 3705-3723.

74. Koziara, K. B.; Stroet, M.; Malde, A. K.; Mark, A. E. J. J. o. C.-A. M. D., Testing and Validation of the Automated Topology Builder (Atb) Version 2.0: Prediction of Hydration Free Enthalpies. *J. Comput. Aid. Mol. Des.* **2014**, 28, 221-233.

75. Roos, K.; Wu, C. J.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L. L.; Abel, R.; Friesner, R. A.; Harder, E. D., Opls3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J. Chem. Theory Comput.* **2019**, 15, 1863-1874.

76. Wang, J., Development of the Second Generation of the General Amber Force Field. **2017**

77. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; Gonzalez-Outeirino, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J., Glycam06: A Generalizable Biomolecular Force Field. Carbohydrates. *J. Comput. Chem.* **2008**, 29, 622-55.

78. Huang, J.; MacKerell Jr, A. D., Charmm36 All-Atom Additive Protein Force Field: Validation Based on Comparison to Nmr Data. *J. Comput. Chem.* **2013**, 34, 2135-2145.

79. Robertson, M. J.; Qian, Y.; Robinson, M. C.; Tirado-Rives, J.; Jorgensen, W. L., Development and Testing of the Opls-Aa/M Force Field for Rna. *J. Chem. Theory Comput.* **2019**, 15, 2734-2742.

80. Riniker, S., Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. *J. Chem. Inf. Model.* **2018**, 58, 565-578.

81. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell Jr, A. D., Charmm36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2016**, 14, 71.

82. Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; de Groot, B. L.; Grubmüller, H., Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **2015**, 11, 5513-5524.

83. Harder, E.; Damm, W.; Maple, J.; Wu, C. J.; Reboul, M.; Xiang, J. Y.; Wang, L. L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A., Opls3: A Force Field Providing Broad Coverage of Drug-Like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, 12, 281-296.

84. Tait, M. J.; Franks, F., Water in Biological Systems. *Nature* **1971**, 230, 91-94.

85. Kollman, P., Free Energy Calculations: Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, 93, 2395-2417.

86. Zou, X.; Yaxiong; Kuntz, I. D., Inclusion of Solvation in Ligand Binding Free Energy Calculations Using the Generalized-Born Model. *J. Am. Chem. Soc.* **1999**, 121, 8033-8043.

87. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, 79, 926-935.

88. Jorgensen, W. L.; Tirado-Rives, J., Potential Energy Functions for Atomic-Level Simulations of Water and Organic and Biomolecular Systems. *Proc. Natl. Acad. Sci.* **2005**, 102, 6665-6670.

89. Glättli, A.; Daura, X.; Gunsteren, W. F. v., Derivation of an Improved Simple Point Charge Model for Liquid Water: Spc/a and Spc/L. *J. Chem. Phys.* **2002**, 116, 9811-9828.

90. Mahoney, M. W.; Jorgensen, W. L., A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid, Nonpolarizable Potential Functions. *J. Chem. Phys.* **2000**, 112, 8910-8922.

91. Skyner, R. E.; McDonagh, J. L.; Groom, C. R.; van Mourik, T.; Mitchell, J. B. O., A Review of Methods for the Calculation of Solution Free Energies and the Modelling of Systems in Solution. *Phys. Chem. Chem. Phys.* **2015**, 17, 6174-6191.

92. Nguyen, T. T.; Viet, M. H.; Li, M. S., Effects of Water Models on Binding Affinity: Evidence from All-Atom Simulation of Binding of Tamiflu to a/H5n1 Neuraminidase. *Sci. World J.* **2014**, 2014, 536084-536084.

93. Hart, J. R.; Rappé, A. K., Van Der Waals Functional Forms for Molecular Simulations. *J. Chem. Phys.* **1992**, 97, 1109-1115.

94. Halgren, T. A., The Representation of Van Der Waals (Vdw) Interactions in Molecular Mechanics Force Fields: Potential Form, Combination Rules, and Vdw Parameters. *J. Am. Chem. Soc.* **1992**, 114, 7827-7843.

95. Halgren, T. A., Merck Molecular Force Field. Ii. Mmff94 Van Der Waals and Electrostatic Parameters for Intermolecular Interactions. *J. Comput. Chem.* **1996**, 17, 520-552.

96. Ren, P. Y.; Wu, C. J.; Ponder, J. W., Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *J. Chem. Theory Comput.* **2011**, 7, 3143-3161.

97. Wang, J.; Cieplak, P.; Li, J.; Hou, T.; Luo, R.; Duan, Y., Development of Polarizable Models for Molecular Mechanical Calculations I: Parameterization of Atomic Polarizability. *J. Phys. Chem. B* **2011**, 115, 3091-3099.

98. Wang, J.; Tingjun, H., Application of Molecular Dynamics Simulations in Molecular Property Prediction I: Density and Heat of Vaporization. *J. Chem. Theory Comput.* **2011**, 7, 2151-2165.

99. Jorgensen, W. L.; Madura, J. D.; Swenson, C. J., Optimized Intermolecular Potential Functions for Liquid Hydrocarbons. *J. Am. Chem. Soc.* **1984**, 106, 6638-6646.

100. Qi, R.; Wang, Q.; Ren, P., General Van Der Waals Potential for Common Organic Molecules. *Bioorg. Med. Chem.* **2016**, 24, 4911-4919.

101. Reilly, A. M.; Tkatchenko, A., Van Der Waals Dispersion Interactions in Molecular Materials: Beyond Pairwise Additivity. *Chem. Sci.* **2015**, 6, 3289-3301.

102. Jorgensen, W. L.; Tirado-Rives, J., The Opls [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, 110, 1657-1666.

103. Delhommelle, J.; MilliÉ, P., Inadequacy of the Lorentz-Berthelot Combining Rules for Accurate Predictions of Equilibrium Properties by Molecular Simulation. *Mol. Phys* **2001**, 99, 619-625.

104. Waldman, M.; Hagler, A. T., New Combining Rules for Rare Gas Van Der Waals Parameters. *J. Comput. Chem.* **1993**, 14, 1077-1084.

105. Gilson, M. K.; Zhou, H.-X., Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys.* **2007**, 36, 21-42.

106. DiStasio, R. A., Jr.; von Lilienfeld, O. A.; Tkatchenko, A., Collective Many-Body Van Der Waals Interactions in Molecular Systems. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, 109, 14791-14795.

107. Mackerell Jr., A. D., Empirical Force Fields for Biological Macromolecules: Overview and Issues. *J. Comput. Chem.* **2004**, 25, 1584-1604.

108. Stouch, T. R.; Williams, D. E., Conformational Dependence of Electrostatic Potential Derived Charges of a Lipid Headgroup: Glycerylphosphorylcholine. *J. Comput. Chem.* **1992**, 13, 622-632.

109. Reynolds, C. A.; Essex, J. W.; Graham Richards, W., Errors in Free-Energy Perturbation Calculations Due to Neglecting the Conformational Variation of Atomic Charges. *Chem. Phys. Lett* **1992**, 199, 257-260.

110. Singh, U. C.; Kollman, P. A., An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem.* **1984**, 5, 129-145.

111. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The Resp Model. *J. Phys. Chem.* **1993**, 97, 10269-10280.

112. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A., Application of Resp Charges to Calculate Conformational Energies, Hydrogen Bond Energies, and Free Energies of Solvation. *J. Am. Chem. Soc.* **1993**, 115, 9620-9631.

113. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P., A Point-Charge Force Field for Molecular

Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* **2003**, 24, 1999-2012.

114. Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D., Automation of the Charmm General Force Field (Cgenff) Ii: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **2012**, 52, 3155-3168.

115. Rappe, A. K.; Goddard, W. A., Charge Equilibration for Molecular-Dynamics Simulations. *J. Phys. Chem.* **1991**, 95, 3358-3363.

116. No, K. T.; Grant, J. A.; Jhon, M. S.; Scheraga, H. A., Determination of Net Atomic Charges Using a Modified Partial Equalization of Orbital Electronegativity Method. 2. Application to Ionic and Aromatic Molecules as Models for Polypeptides. *J. Phys. Chem.* **1990**, 94, 4740-4746.

117. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P., Development and Use of Quantum Mechanical Molecular Models. 76. Am1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, 107, 3902-3909.

118. Stewart, J. J. P., Optimization of Parameters for Semiempirical Methods Ii. Applications. *J. Comput. Chem.* **1989**, 10, 221-264.

119. Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. J. J. o. C.-A. M. D., Class Iv Charge Models: A New Semiempirical Approach in Quantum Chemistry. *J. Comput. Aid. Mol. Des.* **1995**, 9, 87-110.

120. Thompson, J. D.; Cramer, C. J.; Truhlar, D. G., Parameterization of Charge Model 3 for Am1, Pm3, Blyp, and B3lyp. *J. Comput. Chem.* **2003**, 24, 1291-1304.

121. Marenich, A. V.; Jerome, S. V.; Cramer, C. J.; Truhlar, D. G., Charge Model 5: An Extension of Hirshfeld Population Analysis for the Accurate Description of Molecular Interactions in Gaseous and Condensed Phases. *J. Chem. Theory Comput.* **2012**, 8, 527-541.

122. Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I., Fast, Efficient Generation of High-Quality Atomic Charges. Am1-Bcc Model: I. Method. *J. Comput. Chem.* **2000**, 21, 132-146.

123. Jakalian, A.; Jack, D. B.; Bayly, C. I., Fast, Efficient Generation of High-Quality Atomic Charges. Am1-Bcc Model: Ii. Parameterization and Validation. *J. Comput. Chem.* **2002**, 23, 1623-1641.

124. Mukherjee, G.; Patra, N.; Barua, P.; Jayaram, B., A Fast Empirical Gaff Compatible Partial Atomic Charge Assignment Scheme for Modeling Interactions of Small Molecules with Biomolecular Targets. *J. Comput. Chem.* **2011**, 32, 893-907.

125. Cerutti, D. S.; Rice, J. E.; Swope, W. C.; Case, D. A., Derivation of Fixed Partial Charges for Amino Acids Accommodating a Specific Water Model and Implicit Polarization. *J. Phys. Chem. B* **2013**, 117, 2328-2338.

126. Hunter, C. A.; Sanders, J. K. M., The Nature Of .Pi.-.Pi. Interactions. *J. Am. Chem. Soc.* **1990**, 112, 5525-5534.

127. Lommerse, J. P. M.; Price, S. L.; Taylor, R., Hydrogen Bonding of Carbonyl, Ether, and Ester Oxygen Atoms with Alkanol Hydroxyl Groups. *J. Comput. Chem.* **1997**, 18, 757-774.

128. Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P., Halogen Bonding: The Σ-Hole. *J. Mol. Model* **2007**, 13, 291-296.

129. Martinez, C. R.; Iverson, B. L., Rethinking the Term "Pi-Stacking". *Chem. Sci.* **2012**, 3, 2191-2201.

130. Caldwell, J. W.; Kollman, P. A., Cation-.Pi. Interactions: Nonadditive Effects Are Critical in Their Accurate Representation. *J. Am. Chem. Soc.* **1995**, 117, 4177-4178.

131. Williams, D. E., Alanyl Dipeptide Potential-Derived Net Atomic Charges and Bond Dipoles, and Their Variation with Molecular Conformation. *Biopolymers* **1990**, 29, 1367-1386.

132. Hensen, C.; Hermann, J. C.; Nam, K.; Ma, S.; Gao, J.; Höltje, H.-D., A Combined Qm/Mm Approach to Protein−Ligand Interactions: Polarization Effects of the Hiv-1 Protease on Selected High Affinity Inhibitors. *J. Med. Chem.* **2004**, 47, 6673-6680.

133. van der Vaart, A.; Bursulaya, B. D.; Brooks, C. L.; Merz, K. M., Are Many-Body Effects Important in Protein Folding? *J. Phys. Chem. B* **2000**, 104, 9554-9563.

134. Allen, T. W.; Andersen, O. S.; Roux, B., Energetics of Ion Conduction through the Gramicidin Channel. *Proc. Natl. Acad. Sci.* **2004**, 101, 117-122.

135. Patel, S.; Davis, J. E.; Bauer, B. A., Exploring Ion Permeation Energetics in Gramicidin a Using Polarizable Charge Equilibration Force Fields. *J. Am. Chem. Soc.* **2009**, 131, 13890-13891.

136. Baker, C. M.; Best, R. B., Matching of Additive and Polarizable Force Fields for Multiscale Condensed Phase Simulations. *J. Chem. Theory Comput.* **2013**, 9, 2826-2837.

137. Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S., Accelerating Molecular Dynamic Simulation on Graphics Processing Units. *J. Comput. Chem.* **2009**, 30, 864-872.

138. Darden, T.; York, D.; Pedersen, L., Particle Mesh Ewald: An N·Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, 98, 10089-10092.

139. Baker, C. M., Polarizable Force Fields for Molecular Dynamics Simulations of Biomolecules. *WIREs Comput. Mol. Sci.* **2015**, 5, 241-254.

140. Patel, S.; Brooks III, C. L., Charmm Fluctuating Charge Force Field for Proteins: I Parameterization and Application to Bulk Organic Liquid Simulations. *J. Comput. Chem.* **2004**, 25, 1-16.

141. Patel, S.; Mackerell Jr., A. D.; Brooks III, C. L., Charmm Fluctuating Charge Force Field for Proteins: Ii Protein/Solvent Properties from Molecular Dynamics Simulations Using a Nonadditive Electrostatic Model. *J. Comput. Chem.* **2004**, 25, 1504-1514.

142. Bauer, B. A.; Patel, S., Recent Applications and Developments of Charge Equilibration Force Fields for Modeling Dynamical Charges in Classical Molecular Dynamics Simulations. *Theor. Chem. Acc.* **2012**, 131, 1153.

143. Lemkul, J. A.; Huang, J.; Roux, B.; MacKerell, A. D., An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem. Rev.* **2016**, 116, 4983-5013.

144. Baker, C. M.; Anisimov, V. M.; MacKerell, A. D., Jr., Development of Charmm Polarizable Force Field for Nucleic Acid Bases Based on the Classical Drude Oscillator Model. *J. Phys. Chem. B* **2011**, 115, 580-596.

145. Cardamone, S.; Hughes, T. J.; Popelier, P. L. A., Multipolar Electrostatics. *Phys. Chem. Chem. Phys.* **2014**, 16, 10367-10387.

146. Cieplak, P.; Caldwell, J.; Kollman, P., Molecular Mechanical Models for Organic and Biological Systems Going Beyond the Atom Centered Two Body Additive Approximation: Aqueous Solution Free Energies of Methanol and N-Methyl Acetamide, Nucleic Acid Base, and

Amide Hydrogen Bonding and Chloroform/Water Partition Coefficients of the Nucleic Acid Bases. *J. Comput. Chem.* **2001**, 22, 1048-1057.

147. Wang, J.; Cieplak, P.; Li, J.; Wang, J.; Cai, Q.; Hsieh, M.; Lei, H.; Luo, R.; Duan, Y., Development of Polarizable Models for Molecular Mechanical Calculations Ii: Induced Dipole Models Significantly Improve Accuracy of Intermolecular Interaction Energies. *J. Phys. Chem. B* **2011**, 115, 3100-3111.

148. Chen, F.; Liu, H.; Sun, H.; Pan, P.; Li, Y.; Li, D.; Hou, T., Assessing the Performance of the Mm/Pbsa and Mm/Gbsa Methods. 6. Capability to Predict Protein–Protein Binding Free Energies and Re-Rank Binding Poses Generated by Protein–Protein Docking. *Phys. Chem. Chem. Phys.* **2016**, 18, 22129-22139.

149. Chen, J.; Wang, J.; Zhang, Q.; Chen, K.; Zhu, W., Probing Origin of Binding Difference of Inhibitors to Mdm2 and Mdmx by Polarizable Molecular Dynamics Simulation and Qm/Mm-Gbsa Calculation. *Sci. Rep.* **2015**, 5, 17421.

150. Ren, P. Y.; Ponder, J. W., Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *J. Phys. Chem. B* **2003**, 107, 5933-5947.

151. Grossfield, A.; Ren, P.; Ponder, J. W., Ion Solvation Thermodynamics from Simulation with a Polarizable Force Field. *J. Am. Chem. Soc.* **2003**, 125, 15671-15682.

152. Shi, Y.; Xia, Z.; Zhang, J. J.; Best, R.; Wu, C. J.; Ponder, J. W.; Ren, P. Y., Polarizable Atomic Multipole-Based Amoeba Force Field for Proteins. *J. Chem. Theory Comput.* **2013**, 9, 4046-4063.

153. Zhang, C.; Lu, C.; Jing, Z.; Wu, C.; Piquemal, J.-P.; Ponder, J. W.; Ren, P., Amoeba Polarizable Atomic Multipole Force Field for Nucleic Acids. *J. Chem. Theory Comput.* **2018**, 14, 2084-2108.

154. Donchev, A. G.; Ozrin, V. D.; Subbotin, M. V.; Tarasov, O. V.; Tarasov, V. I., A Quantum Mechanical Polarizable Force Field for Biomolecular Interactions. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, 102, 7829-7834.

155. Naserifar, S.; Brooks, D. J.; GoddardIII, W. A.; Cvicek, V., Polarizable Charge Equilibration Model for Predicting Accurate Electrostatic Interactions in Molecules and Solids. *J. Chem. Phys.* **2017**, 146, 124117.

156. Naserifar, S.; GoddardIII, W. A., The Quantum Mechanics-Based Polarizable Force Field for Water Simulations. *J. Chem. Phys.* **2018**, 149, 174502.

157. Oppenheim, J. J.; Naserifar, S.; Goddard, W. A., Extension of the Polarizable Charge Equilibration Model to Higher Oxidation States with Applications to Ge, as, Se, Br, Sn, Sb, Te, I, Pb, Bi, Po, and at Elements. *J. Phys. Chem. A* **2018**, 122, 639-645.

158. Khoruzhii, O.; Donchev, A. G.; Galkin, N.; Illarionov, A.; Olevanov, M.; Ozrin, V.; Queen, C.; Tarasov, V., Application of a Polarizable Force Field to Calculations of Relative Protein–Ligand Binding Affinities. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, 105, 10378-10383.

159. Bell, D. R.; Qi, R.; Jing, Z.; Xiang, J. Y.; Mejias, C.; Schnieders, M. J.; Ponder, J. W.; Ren, P., Calculating Binding Free Energies of Host–Guest Systems Using the Amoeba Polarizable Force Field. *Phys. Chem. Chem. Phys.* **2016**, 18, 30261-30269.

160. Jiao, D.; Golubkov, P. A.; Darden, T. A.; Ren, P., Calculation of Protein-Ligand Binding Free Energy by Using a Polarizable Potential. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, 105, 6290-6295.

161. Jiao, D.; Zhang, J.; Duke, R. E.; Li, G.; Schnieders, M. J.; Ren, P., Trypsin-Ligand Binding Free Energies from Explicit and Implicit Solvent Simulations with Polarizable Potential. *J. Comput. Chem.* **2009**, 30, 1701-11.

162. Lopes, P. E. M.; Huang, J.; Shim, J.; Luo, Y.; Li, H.; Roux, B.; MacKerell, A. D., Polarizable Force Field for Peptides and Proteins Based on the Classical Drude Oscillator. *J. Chem. Theory Comput.* **2013**, 9, 5430-5449.

163. Huang, J.; Lopes, P. E. M.; Roux, B.; MacKerell, A. D., Recent Advances in Polarizable Force Fields for Macromolecules: Microsecond Simulations of Proteins Using the Classical Drude Oscillator Model. *J. Phys. Chem. Lett.* **2014**, 5, 3144-3150.

164. Mohamed, N. A.; Bradshaw, R. T.; Essex, J. W., Evaluation of Solvation Free Energies for Small Molecules with the Amoeba Polarizable Force Field. *J. Comput. Chem.* **2016**, 37, 2749-2758.

165. Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E., Systematic Validation of Protein Force Fields against Experimental Data. *PLoS One* **2012**, 7, e32131.

166. Lin, F.-Y.; MacKerell, A. D., Improved Modeling of Halogenated Ligand–Protein Interactions Using the Drude Polarizable and Charmm Additive Empirical Force Fields. *J. Chem. Inf. Model.* **2019**, 59, 215-228.

167. Baker, C. M.; Best, R. B., Insights into the Binding of Intrinsically Disordered Proteins from Molecular Dynamics Simulation. *WIREs Comput. Mol. Sci.* **2014**, 4, 182-198.

168. Huang, L.; Roux, B., Automated Force Field Parameterization for Nonpolarizable and Polarizable Atomic Models Based on Ab Initio Target Data. *J. Chem. Theory Comput.* **2013**, 9, 3543-3556.

169. Mayne, C. G.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. C., Rapid Parameterization of Small Molecules Using the Force Field Toolkit. *J. Comput. Chem.* **2013**, 34, 2757-2770.

170. Betz, R. M.; Walker, R. C., Paramfit: Automated Optimization of Force Field Parameters for Molecular Dynamics Simulations. *J. Comput. Chem.* **2015**, 36, 79-87.

171. Wang, J.; Kollman, P. A., Automatic Parameterization of Force Field by Systematic Search and Genetic Algorithms. *J. Comput. Chem.* **2001**, 22, 1219-1228.

172. Liu, Z. M.; Barigye, S. J.; Shahamat, M.; Labute, P.; Moitessier, N., Atom Types Independent Molecular Mechanics Method for Predicting the Conformational Energy of Small Molecules. *J. Chem. Inf. Model.* **2018**, 58, 194-205.

173. Mobley, D.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K., Open Force Field Consortium: Escaping Atom Types Using Direct Chemical Perception with Smirnoff V0.1. *bioRxiv* **2018**, 286542.

174. Bayly, C.; McKay, D.; Truchon, J., In; Merck & Co. internal development release: 2011.

175. Zanette, C.; Bannan, C. C.; Bayly, C. I.; Fass, J.; Gilson, M. K.; Shirts, M. R.; Chodera, J. D.; Mobley, D. L., Toward Learned Chemical Perception of Force Field Typing Rules. *J. Chem. Theory Comput.* **2019**, 15, 402-423.

176. Liu, Z. M.; Pottel, J.; Shahamat, M.; Tomberg, A.; Labute, P.; Moitessier, N., Elucidating Hyperconjugation from Electronegativity to Predict Drug Conformational Energy in a High Throughput Manner. *J. Chem. Inf. Model.* **2016**, 56, 788-801.

177. Vitaku, E.; Smith, D. T.; Njardarson, J. T., Analysis of the Structural Diversity, Substitution Patterns, and Frequency of Nitrogen Heterocycles among U.S. Fda Approved Pharmaceuticals. *J. Med. Chem.* **2014**, 57, 10257-10274.

178. Piana, S.; Lindorff-Larsen, K.; Shaw, David E., How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys. J.* **2011**, 100, L47-L49.

179. Vitalini, F.; Mey, A. S. J. S.; Noé, F.; Keller, B. G., Dynamic Properties of Force Fields. *J. Chem. Phys.* **2015**, 142, 084101.

180. Yoo, J.; Aksimentiev, A., Refined Parameterization of Nonbonded Interactions Improves Conformational Sampling and Kinetics of Protein Folding Simulations. *J. Phys. Chem. Lett.* **2016**, 7, 3812-3818.

2 Atom Type Independent Modeling of the Conformational Energy of Benzylic, Allylic, and other Bonds Adjacent to Conjugated Systems

This chapter is reproduced from a manuscript submitted for publication: "Atom Type Independent Modeling of the Conformational Energy of Benzylic, Allylic, and other Bonds Adjacent to Conjugated Systems", **Champion C**., Barigye, S. J., Wei W., Liu Z., Labute P., Moitessier N. *J. Chem. Inf. Model.* (ci-2019-005818)

2.1 Introduction

2.1.a Computational Methods in Drug Discovery and Molecular Mechanics

Computational methods are often quick and cost-effective complements to experiments to identify potential binders to targets of therapeutic interest and/or off-targets. Over the years, computational tools have contributed to many stages of the drug discovery process, from the prediction of drug-likeness[1] following, for example, Lipinski's rule of five or ADME (absorption, distribution, metabolism and excretion) properties[2] using artificial intelligence (AI) or statistical analysis, to physics-based methods providing insights into the structure and dynamics of molecular systems.[3, 4] It is anticipated that Structure-Based Drug Design (SBDD) will have even greater relevance in future Drug Discovery paradigms.[5, 6] The accuracy of predicted drug binding affinities depend on several factors such as the level of detail of the structural model[7] (subatomic, atomic, coarse-grained), the accuracy of the energy potentials computed for molecular conformations,[8, 9] as well as the degree to which all energetically accessible conformations are sampled.[10] In this context, quantum mechanical (QM) methods would provide a very accurate depiction of the energetics of molecular systems, allowing rigorous estimates of ligand-macromolecule binding energies.[11] These methods can, however, not be carried to high-throughput tasks, to scan large

portions of conformational space, or to study large macromolecules, due to their restrictive computational costs. In light of this limitation, molecular mechanics (MM) methods have been developed to evaluate the energetics of molecular systems using simplified potentials with the objective to reproduce experimental data and QM potentials, while reducing computational costs by several orders of magnitude. However, the accuracy of these more empirical MM methods largely depends on the quality of the potentials and parameters of the underlying force fields (FFs).[12, 13]

2.1.b Atom-type based FFs

In MM, the potential energy of a molecular system (e.g. small molecules, proteins, nucleic acids, and complexes) is calculated using a FF corresponding to a set of potential energy functions and its associated precomputed parameters (Eqs. 1.1-1.7). The contributions from each term in a FF can be split into two categories, "bonded" interactions (bonds, angles, torsions, out-of-plane angles) which are calculated for atoms within the same molecule, and "non-bonded" interactions (e.g., van der Waals and electrostatics) which are calculated for pairs of atoms separated by 3 or more bonds (intramolecular) or pairs of atoms in different molecules (intermolecular). It should be noted that the interactions of atoms separated by 3 bonds are therefore of both types: torsions and non-bonded interactions.

2.1.c Transferability

"*Atom types*" are central to most widely applied FFs in SBDD, such as the AMBER,[14, 15] CHARMM,[16, 17] GROMOS,[18, 19] and OPLS[20-23] series. In AMBER protein FF, for example, parameters for aromatic carbons (atom type CA) are different from aliphatic carbons (CT) or carbonyl carbon (C) and other carbon types. However, these definitions are limited to local environments and distant chemical functional groups do not affect the atom type (and set of parameters) assigned to particular moieties, which consequently ignores some electronic effects. For example, electron donating or withdrawing substituents adjacent to aromatic rings are not considered in ring atom types, whereas torsional energy profiles could differ when such substituents are present.

Parametrization of a FF consists in finding the ideal values for all the parameters (shown in bold) associated with each function (Eqs. 1.1-1.7). For example, the bond stretching term (Eq. 1.2) describes the ideal bond distance between two atoms and is characterized by an equilibrium bond length ($\mathbf{r_{eq}}$) and a force constant ($\mathbf{K_r}$). In order to describe all possible bond stretching events, parameters for the equilibrium value and force constants are required for all combinations of two atom types.[12] This parameter fitting process uses experimental (e.g., H-NMR, thermodynamic properties) and/or high-level QM data as reference, which are costly to obtain, ultimately imposing a limit on the size of the training set used to develop parameters. FFs thus rely on the transferability of parameters obtained from molecules in the training set to other similar molecules. The current consensus is that no particular FF could accurately describe the energetics of all possible small drug-like molecules due to the sheer size of the chemical space, and the poor transferability of empirical parameters generated on specific molecules.[24] It is important to keep in mind that not all types of parameter are subject to this lack of transferability; for instance, bonds and angles are generally assumed to be fully covered. However, the authors of OPLS3.0 estimated in 2015 that 33% of drug-like molecules were missing at least one torsion parameter,[21] (a more recent version attempts to solve this limitation[25]) and the treatment of non-bonded interactions has also recently been challenged by the introduction of polarizable Force Fields (e.g. AMOEBA,[26, 27] CHARMM-Drude[28, 29]). Current developments in FF methodologies are hence highly focused towards torsional and non-bonded interactions.

To address the liabilities resulting from poor parameter transferability and/or missing parameters, researchers have followed two main approaches. On one hand, automated toolkits such as GAAMP,[30] ffTK,[31] Paramfit[32] and Parmscan[33] have been developed, allowing to generate accurate parameters for specific molecules of interest from QM data. These user-friendly toolkits are particularly fit for researchers studying the interactions within a ligand/receptor pair using molecular dynamics (MD), since only few parameters need to be generated (usually for the drug-like molecule). However, these tools cannot be carried to high-throughput tasks (e.g. docking libraries of compounds), due to the computational costs associated with the parameter fitting process. While these toolkits allow parameters to be generated for specific studies, they do not solve the problem of parameter transferability. A radically different approach consists in developing MM methods with greater transferabilities without relying on the concept of atom types

to determine parameters. To our knowledge, Mobley et al.'s recent attempt with SMIRNOFF,[34] a FF which uses direct chemical perception instead of traditional atom types to determine parameters, as well as H-TEQ (developed in our lab),[35, 36] are the only methods moving away from the atom type paradigm of FFs. Both SMIRNOFF and H-TEQ, were shown to perform comparably well to GAFF (one of the most widely used FFs for small organic molecules)[37] in reproducing liquid properties and QM torsional profiles respectively. The performance of these methods has not yet been extensively tested in the context of SBDD relevant interactions, due to their very recent releases, and we expect further work to allow these methods to cover most (if not all) possible organic molecules with good accuracy. These efforts are highly encouraging in the future ability of atom type free FFs to rival state-of-the-art FFs towards SBDD applications, without requiring any molecule-specific parametrization.

While our previous versions of H-TEQ focused on saturated compounds, we report here our efforts to incorporate unsaturated compounds.

2.2 Impact of Unsaturations on Torsional Energy

2.2.a Organic Chemistry Principles and Drug Conformational Energy

In organic chemistry, several models have been employed to rationalize the conformational preferences of molecules and stereoselectivity in chemical reactions. For example, the hyperconjugation model is often evoked to rationalize the preference of the staggered conformation in ethane and the anomeric effect in carbohydrate molecules.[38-40] Briefly, hyperconjugation is a stabilizing interaction involving the donation of electrons from a bonding (e.g. σ) to an antibonding (e.g. σ*) orbital, leading to the formation of a new orbital lying lower in energy (Figure 2.1).[41] Two main factors influence the strength of hyperconjugation interactions.[42, 43] First, a greater physical overlap between interacting orbitals leads to a stronger interaction. This overlap is maximized when σ and σ* are in *anti* relationship (Figure 2.1a) rather than *syn* (Figure 2.1b) and is minimal when orbitals are perpendicular to one another. Second, a smaller gap between donating and accepting orbitals energy levels leads to a stronger interaction. For example in fluoroethane, the electronegative fluorine results in a lower lying $\sigma^*_{(C\text{-}F)}$ orbital (Figure 2.1c)

than $\sigma^*_{(C\text{-}H)}$ (Figure 2.1d) and reducing the energy gap ultimately favoring the conformation in which $\sigma_{(C\text{-}H)}$ and $\sigma^*_{(C\text{-}F)}$ (the best acceptor in this molecule) are *anti*.

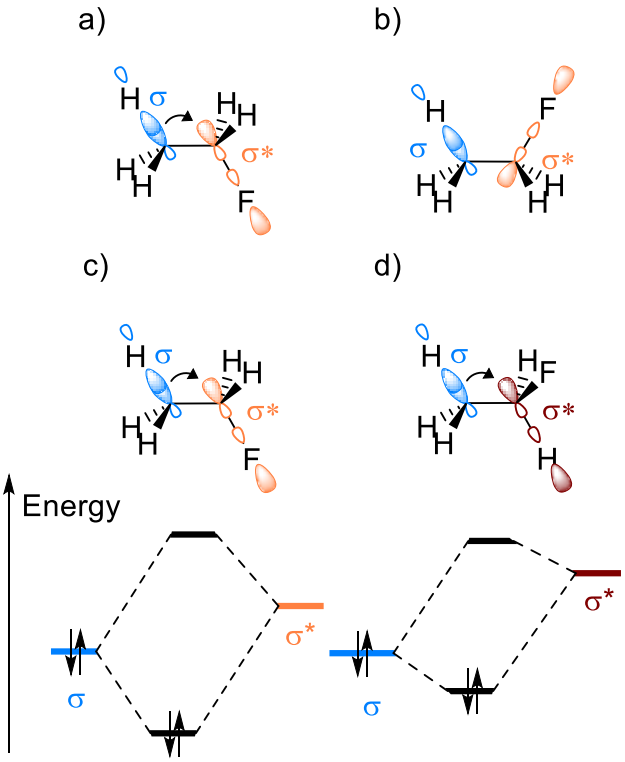


Figure 2.1. Factors influencing the strength of hyperconjugation interactions in the fluoroethane molecule. The orbital overlap is greater when $\sigma_{(C\text{-}H)}$ and $\sigma^*_{(C\text{-}F)}$ are anti (a) rather than syn (b). The energy gap between σ and σ* is smaller between $\sigma_{(C\text{-}H)}$ and $\sigma^*_{(C\text{-}F)}$ (c) than $\sigma_{(C\text{-}H)}$ and $\sigma^*_{(C\text{-}H)}$ (d).

Although qualitative in nature, these chemistry principles are highly transferable since they follow general principles such as the degree of electron donating or electron withdrawing character, presence of lone pairs and the degree of overlap of molecular orbitals. Indeed, we have demonstrated that if these principles are quantified (using simple atomic properties), universal models for computing the torsional energy of molecules could be developed.[35, 36] While our previous studies were focused towards $\sigma \rightarrow \sigma^*$ and $n \rightarrow \sigma^*$ hyperconjugation modes, a large number of drug like molecules contain unsaturations and aromatic ring systems,[44] which exhibit other hyperconjugation modes: $\sigma \rightarrow \pi^*$ and $\pi \rightarrow \sigma^*$, which we will refer to as ***π-hyperconjugation***. These additional hyperconjugation interactions must play an important role in determining the conformational preferences of such moieties. Therefore, the goal of the present manuscript is to

describe our progress in integrating π interaction modes into our H-TEQ method, to guarantee its applicability to torsions in any drug-like molecule, improving the accuracy of FF-based methods for SBDD applications. In this present work, conjugated systems are not considered, and we focus on the prediction of torsional parameters for $C(sp^2)$-$C(sp^3)$ bonds only.

2.2.b Asymmetric Induction and π-Hyperconjugation

As predictive yet qualitative chemical principles, multiple chemical models have been developed to predict diastereoselectivity in nucleophilic addition reactions involving carbonyl groups such as the Cram and Felkin-Anh models.[45, 46] The early Cram model states that the ideal path of attack of a nucleophile towards a carbonyl group, is essentially that minimizing steric hindrance. The more reliable Felkin-Anh model invokes additional electronic effects which control the diastereoselectivity; a strong electron withdrawing group (R^L in Figure 2.2) at the vicinal position oriented antiperiplanar to the incoming nucleophile leads to a favorable $\sigma \rightarrow \sigma^*$ interaction stabilizing the transition state. Furthermore, the angle of attack of the nucleophile is not 90° but ~107° (Burgi-Dunitz angle)[47] which maximizes the alignment of the nucleophile σ orbital with the carbonyl π* orbital, ultimately leading to the bond formation.[4-6] Although these different models disagree as to which of these interactions predominate, and do not predict the same stereochemical outcomes, it remains clear that both steric and electronic effects govern nucleophilic addition reactions on carbonyl centers.

While the Felkin-Anh model is in principle intended to predict the orientation of nucleophilic attacks to the C-α, it can also provide an understanding of the $\pi \rightarrow \sigma^*$ or $\sigma \rightarrow \pi^*$ hyperconjugation propensity. From the Felkin-Anh model, strongly electron-withdrawing (EWD) groups play a role similar to large substituents favoring the alignment of the σ* with π and π* orbital, and thus favoring the $\pi \rightarrow \sigma^*$ hyperconjugation. Indeed, as can be observed in Table .2.1, our calculations with the natural bond orbitals method (NBO) (see Computational Methods) allowed us to quantify the strength of hyperconjugation interactions and showed that strong EWD groups (e.g. fluorine) favor $\pi \rightarrow \sigma^*$ relative to $\sigma \rightarrow \pi^*$. The favorable nucleophilic attack at the carbonyl group may therefore be attributed to the interaction between the $\sigma \rightarrow \pi^*$ resulting in a lower energy LUMO and thus more susceptible to nucleophilic attack. On the other hand, electron donating groups (e.g. CH_3) result in weak σ* receptors and thus $\sigma \rightarrow \pi^*$ hyperconjugation

predominates. In Table 2.1, we present values for the energy gaps and Fock matrix elements obtained from NBO calculations. In short, the Fock matrix elements are related to the overlap between the interacting orbitals (a larger value/overlap corresponds to stronger hyperconjugation), and energy gaps are self-explanatory (smaller energy gap leads to a stronger interaction as discussed). We would recommend the following article[42] to any reader interested in the details of how NBO calculates these values.

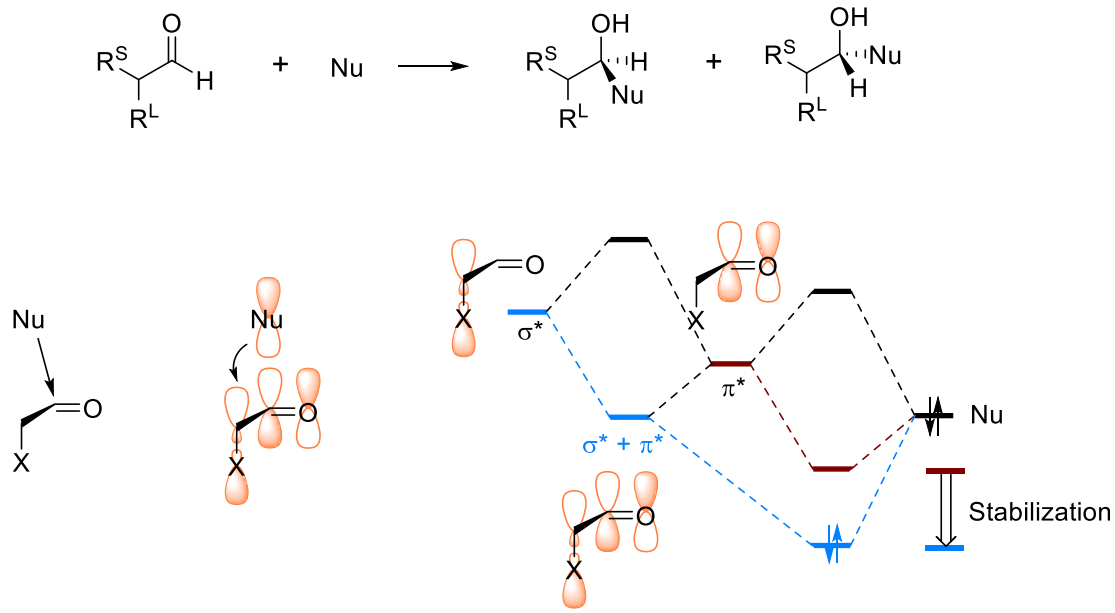
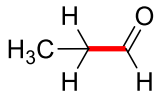
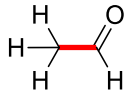
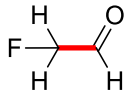
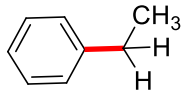
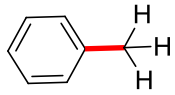
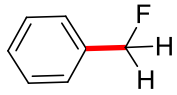


Figure 2.2. Electronic interactions evoked by the Felkin-Anh model to predict conformational preference.

The qualitative models routinely employed by organic chemists are often transferable, as they tend to isolate the predominant factors to simplify the picture. However, in order to translate these qualitative theories into robust quantitative predictions, an inclusion of other weaker interactions might be necessary. The underlying interrogative is: for a given torsion involving σ → π* and π → σ*, should the predominant hyperconjugation mode be exclusively considered or should contributions from both modes be incorporated in our model; this is particularly interesting for cases where comparable magnitudes are observed (e.g. toluene in Table 2.1). Furthermore, should σ → σ* hyperconjugation interactions be neglected as they are smaller in magnitude than π-hyperconjugation interactions? It is important to note that σ → σ* are maximal when the σ and σ* orbitals are *anti* (in plane with the π-system), whereas σ → π* and π → σ* are maximal when

the σ and σ* orbitals are perpendicular to the π-system (Figure 2.3). Hyperconjugation and π-hyperconjugation hence favor different conformations, thus neglecting weaker competing interactions could hinder the predictive ability of our model.

Table 2.1. Energy gap and Fock matrix elements for π→σ* and σ → π* hyperconjugation

		E_{hyp}(kcal/mol)	ΔE [BD – BD*]/(a.u)	F[BD, BD*]/(a.u)
	π → σ*	1.77	1.10	0.039
	σ → π*	5.96	1.03	0.070
	π → σ*	2.28	1.10	0.045
	σ → π*	9.32	0.91	0.082
	π → σ*	4.5	1.01	0.060
	σ → π*	2.3	1.40	0.051
	π → σ*	4.42	0.90	0.056
	σ → π*	4.7	0.98	0.061
	π → σ*	5.22	0.90	0.061
	σ → π*	7	0.86	0.069
	π → σ*	11.68	0.79	0.086
	σ → π*	2.04	1.34	0.047

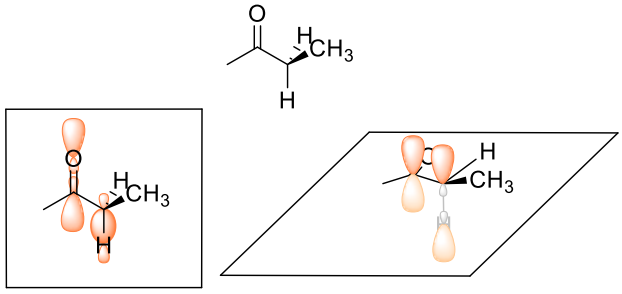


Figure 2.3. Different conformations favor different hyperconjugation modes. σ → σ* is favored when C-H and C=O are coplanar (left), π → σ* is favored when C-H and C=O are perpendicular (right).

2.2.c Hyperconjugation and/or Sterics as Major Torsional Energy Contributors

The conformational flexibility of small drug-like molecules essentially stems from rotation around bonds (i.e. dihedral angles). Hence, an accurate prediction of torsional energy profiles is critical for applications in SBDD. Non-bonded interactions (vdW, electrostatics) cannot be neglected however, as when a bond is rotated, the molecule can reorganize other degrees of freedom in order to minimize steric clash, allow favorable H-bonding or vdW interactions to occur etc.; these additional effects become weaker however as molecules get smaller. Typically, empirical torsional parameters are parametrized last, and are the only term in the FF equation which do not explicitly describe a specific underlying physical interaction. While this empirical nature can make up for errors in non-bonded parameters and improve the accuracy of molecules in the development set, it may be at the root of the poor transferability of torsional parameters.[13] We hence hypothesize that replacing these poorly transferable empirical parameters, by contributions from different hyperconjugation modes ($\sigma \rightarrow \sigma^*$, π-hyperconjugation) will improve the transferability of torsional energies for drug-like molecules. It should be noted that for the purpose of our comparison, the remaining terms of the FF energy will be calculated with the current implementation of GAFF2, hence residual error from the other parts of the FF are expected to be present.

The first step in our approach is to confirm our hypothesis that hyperconjugation interactions will play an important role in the determination of conformational preference. Rotational energy profiles were computed with QM at the MP2/6-311+G** level of theory which is consistent with previous studies.[35, 36] As shown in Figure .4, different π-system reveal varying conformational preferences and radically different rotational profiles (amplitude and periodicity). On one hand, the thiophene and benzene profiles contain two minima (±90°) and two maxima (0°, 180°), whereas the ketone and furan show 3 minima (180°, ±60°) and 3 maxima (0°, ±120°). Inspecting the optimal conformations of each molecule, we notice that for both the benzene and thiophene derivatives, the -CH_3 substituent is positioned such as to maximize the overlap between the $\sigma_{(C\text{-}C)}$ and π/π^* orbitals in the aromatic ring, at the expense of possible interactions between the $\sigma_{(C\text{-}H)}$ and π-orbitals (Figure 2.4). On the other hand, the furan derivative in its preferred conformation shows reduced orbital overlap between the $\sigma_{(C\text{-}C)}$ and the π-orbitals, allowing one of the $\sigma_{(C\text{-}H)}$ to

partially overlap with the π-orbitals. For the ketone, $\sigma_{(C\text{-}C)}$ does not interact at all with the π-orbitals, and both $\sigma_{(C\text{-}H)}$ partially overlap with the π-orbitals. In the benzene example, the preference for ±90° can be attributed to unfavorable steric clash between hydrogens at the ortho position when the methyl group is in plane of the π-system. The thiophene derivative reveals a very similar profile, although it is not expected to be subject to a steric clash of the same magnitude, the hydrogen atom being further away (5-membered rings having inherently different geometries than 6 membered rings).

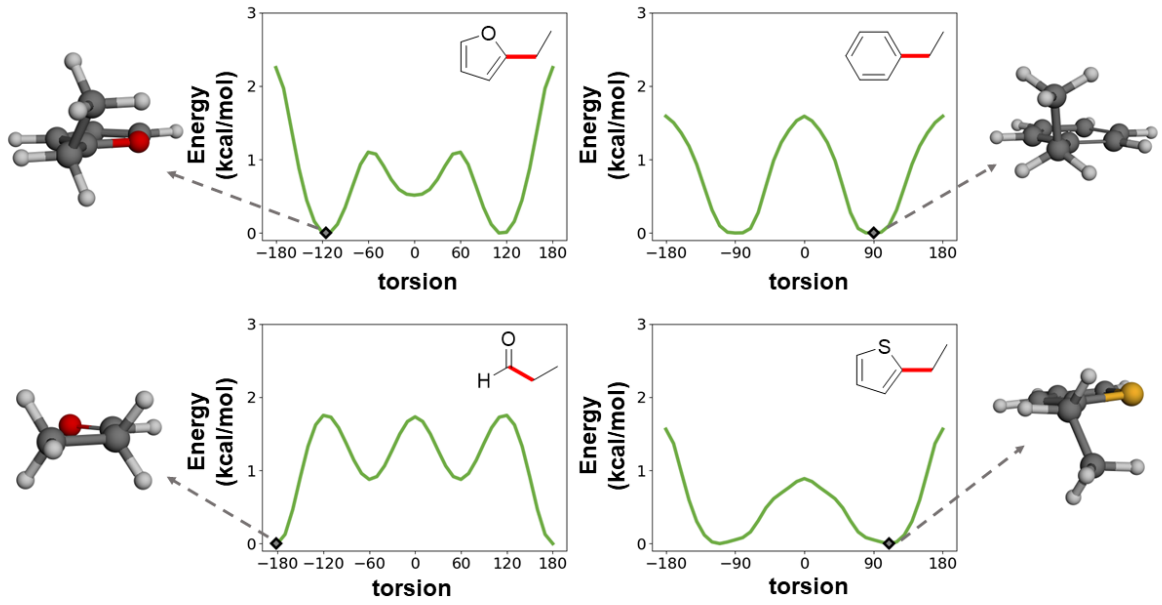


Figure 2.4. Variety of QM torsional profiles is linked to the underlying interactions. Rotated bonds are shown in red.

A notable difference between the benzene and thiophene profiles is the broadness of the low energy region. Clearly, the nature of the π-system is closely linked to which interaction will predominate, and ultimately to which conformation will be preferred. Three hyperconjugation modes are competing in these systems ($\sigma \rightarrow \sigma^*$, $\sigma \rightarrow \pi^*$ and $\pi \rightarrow \sigma^*$, Figure 2.3), and their strength depends on two major factors, the energy level difference of the interacting bonding/antibonding orbital pair and the spatial orbital overlap.[41-43, 48] The nature of the π-system is directly related to the energy levels of π and π* orbitals, as well as their polarization.[49] While it remains unclear at

this stage as to which interaction predominates in each case, we can assess that the same set of interactions (with different electronic effects from various functional groups) may lead to very different profiles (Figure 2.4).

2.2.d Understanding Interactions

While quantum chemistry can provide an accurate depiction of molecules, the information that can be extracted remains limited, and it is sometimes impossible to directly translate results obtained from these theoretical calculations into well understood chemical or physical principles. This discrepancy has thus led to a wide range of QM based methods that decompose the quantum energy into more chemically relevant parts. There are currently three main approaches allowing to dissect delocalization interactions (hyperconjugation in this work): natural bond orbital (NBO) analysis,[50] energy decomposition analysis (EDA)[51] and the block localized wavefunction method (BLW).[52] While these methods are built around similar concepts (a full wavefunction or electron density is compared to a localized construct, and the energy difference between both is assumed to be related to delocalization interactions), they operate quite differently, specifically in the way they generate the localized orbitals. While EDA methods and BLW use non-orthogonal orbitals (which increases the role of steric effects), NBO uses orthogonal orbitals to describe the localized reference.[53] Such decomposition schemes were initially developed to study intermolecular interactions,[54, 55] but have more recently been used to study intramolecular hyperconjugation type interactions.[56] The degree to which the decomposition is performed also varies from method to method and to this day, NBO is the only method which can output an energy value for every bonding/antibonding orbital pair in a molecule. In other methods, hyperconjugation and conjugation energies are agglomerated into a single energy term, hence not giving a chemically relevant picture with the same level of resolution.

Considering the two major interactions present in our systems are $\sigma \rightarrow \pi^*$ and $\pi \rightarrow \sigma^*$, we expect that factors increasing the amplitude of one of them will decrease the amplitude of the other, as a good σ-donor is usually a poor σ-acceptor, and *vice versa* (see Table 2.1).[42] Hence, a full decomposition of the interactions resolving both $\sigma \rightarrow \pi^*$ and $\pi \rightarrow \sigma^*$ seems more valuable, our end goal being to understand and develop rules to explain the factors controlling these interactions. NBO has notably been applied to understand the conformational preference of ethane, by invoking

the predominant role of hyperconjugation,[38] to discern how different elements within pnictogens (N, P, As)[57] and chalcogens (O, S, Se, Te)[58] impacts n → σ* hyperconjugation and ultimately the magnitude of the anomeric effect. NBO has also already been used to study the torsional energy profiles of conjugated systems.[59] Overall, NBO has been employed to understand the conformational preference of a wide range of molecules, explaining these preferences with different hyperconjugation modes, as well as to explore how different elements within a group can impact such interactions; it is therefore particularly fit for our purposes.

2.3 Computational Methods

2.3.a Construction of the Development Set

In order to complement our previously developed H-TEQ method, our objective was to replace the empirical torsional energy found in GAFF2 by a new function which would describe π-hyperconjugation. This function would be based on atomic properties (e.g. treating all carbons atoms in the same way) and from the topology of the molecule. Overall, the function assigns torsional parameters from only atomic properties and topology, without any prior atom typing of the molecules. To that end, we have first constructed a development set (Figure 2.5), covering a large variety of π-systems and saturated groups positioned at a vicinal position, in order to understand the effects of different moieties on torsional profiles, and the underlying interactions giving rise to such profiles. It should be noted that our data set is not built to resemble drug-like molecules, but rather to cover as wide a range of chemical space as possible (-R groups going from very electron withdrawing (e.g. fluorine) to electron donating (e.g. hydroxyl), in order to understand the factors governing the various hyperconjugation modes.

2.3.b Details of the Calculations

First, we have obtained torsional energy profiles for every molecule in our development set using QM at the MP2/6-311+G** level of theory using the software GAMESS-US,[60, 61] which is consistent with our prior studies.[35, 36] In more detail, the torsional profiles were obtained by freezing the desired torsion from -180° to 180° with 10° increments while allowing all other degrees of freedom to optimize. The resulting optimized conformations were used to perform the MM and NBO calculations.

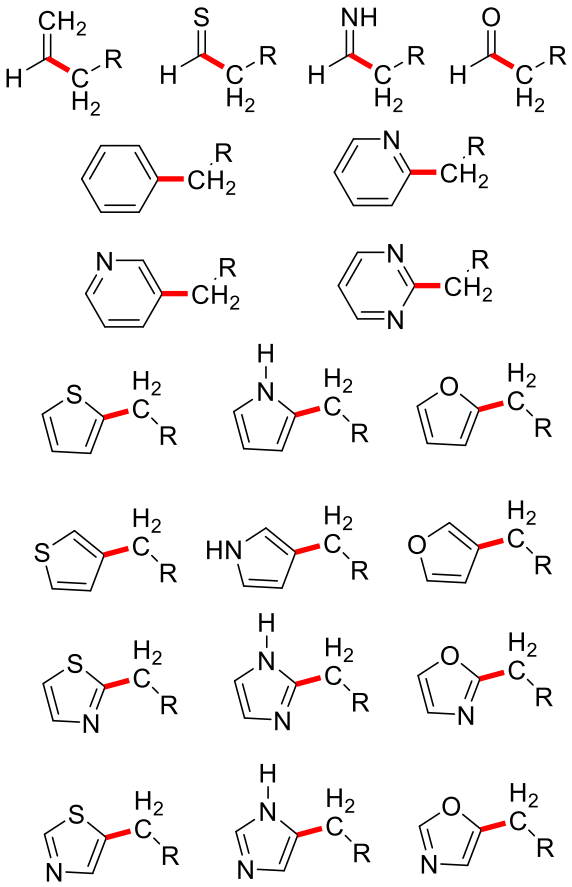


Figure 2.5. Development set of molecules used to study conformational preference of organic molecules containing π-systems. Rotated bonds are shown in red (R = H, F, Cl, CH_3, OH).

MM calculations were performed using the AMBER16 package; GAFF2 atom types were automatically assigned with antechamber, partial charges were assigned using the AM1-BCC method on the global minimum structures and were carried to all other conformations of the same molecule. Finally, the GAFF2 energy was calculated by following the sander routine.

NBO calculations were performed with the NBO 6.0[50] program using the same level of theory and basis set as our QM calculations. To verify whether NBO energies could be used, we first replaced the torsional energy (related to the central rotated bond only) in GAFF2, by hyperconjugation energies (from NBO) related to all of the hyperconjugation modes around the central bond of interest. We then resorted to a scaling of these NBO energies down by factors of 0.25 for $\sigma \rightarrow \sigma^*$ and 0.4 for π-hyperconjugation to minimize the root-mean squared error (RMSE) between QM and scaled NBO profiles (see Appendix 1). Note that the NBO profiles were obtained

by replacing the torsional energy in GAFF2 by scaled NBO hyperconjugation energies. Only the torsional energy related to the central rotated bond were replaced by NBO energies. A Fourier regression of the hyperconjugation profiles of every molecule in the development set was performed, such that each $\sigma \rightarrow \pi^*$ and $\pi \rightarrow \sigma^*$ would be associated to a set of $V_{1\text{-}3}$ parameters.

The RMSE calculations were performed with Eq. 2.1, in which every point in the torsional profile is compared to the QM reference. Prior to the RMSE calculation, profiles were rescaled such that the point with lowest energy was set to 0 kcal/mol. RMSEs being an imperfect measure, cut-offs or Boltzmann weights are sometimes used to discard or scale down the contribution from points of the potential energy surface that are higher in energy.[62] We did not notice any significant difference in RMSEs when using such schemes (Appendix 2), and have kept the simpler RMSE scheme shown in Eq. 2.1.

$$RMSE = \sqrt{\frac{\sum_n \left(E_{MM} - E_{QM}\right)^2}{n}} \tag{2.1}$$

Finally, to compare the performance of our method H-TEQ 3.0, we have replaced the torsional energy in GAFF2 by the equations we have developed (Eqs. 2.2 and 2.3) which are meant to reproduce NBO calculated π-hyperconjugation energies. $\sigma \rightarrow \sigma^*$ hyperconjugation energies were also included by using our previously developed set of equations.[36] It is important to note that only the torsional energies related to the central rotated bond were replaced by H-TEQ 3.0 values. Indeed, our method cannot yet cover all possible chemical groups (notably torsions involving $\pi \rightarrow \pi^*$ and $n \rightarrow \pi^*$ interactions) and some of the molecules in our validation set could not be supported by H-TEQ 3.0. Hence to treat every molecule with the exact same methodology, we kept all of the other (non-central) GAFF2 torsional energies. Generally, rotation around these other torsions is minimal as the central bond is rotated, particularly when the molecule is small. A few exceptions in which the other parts of the molecule reorganized considerably were observed though (Appendix 3).

2.3.c Construction of the Validation Set

To evaluate the performance of H-TEQ 3.0, we have developed a validation set of 50 diverse drug molecules from a larger set previously developed in our lab.[63] In order to reduce the computational cost associated with the obtention of QM torsional profiles, molecules were fragmented, keeping only the most relevant parts. For example, molecules were fragmented at sp^3-sp^3 bonds, replacing parts of the molecule by hydrogens atoms. Furthermore, torsional profiles were obtained by using 15° (rather than 10°) increments, thereby reducing the number of conformations in a profile from 36 to 24. Overall, our validation set does not contain any molecule used to train the model, and a variety of novel π-systems were used (e.g. extended conjugated systems, fused rings). The full set can be found in the supporting information (Appendix 4).

2.4 Results and Discussion

2.4.a Quantifying Hyperconjugation from NBO

The first step to our approach consisted in verifying that NBO hyperconjugation values could be used to replace the torsional energy within GAFF2. When replacing GAFF2 torsional energies by scaled NBO hyperconjugation values, we note a significant decrease in RMSE from 0.84 kcal/mol to 0.55 kcal/mol with respect to the QM reference (full results in Appendix 1). A more detailed account of NBO's performance can be obtained by inspecting the histogram shown in Figure 2.6. As expected from the average values shown in Appendix 1, RMSEs obtained by NBO are lower than those obtained using GAFF2. More interestingly, GAFF2 reveals the presence of two populations, which we can interpret as molecules having been explicitly parameterized (low RMSE), and those for which parameters were transferred from similar molecules but which suffer from poor transferability (larger RMSEs). The development set used herein consists of very small molecules only (< 20 atoms) and we expect the lack of transferability to be further exacerbated in larger, more diverse drug-like molecules, as the probability that more parameters will be sub-optimal is larger, and smaller errors will accumulate. In Figure 2.6, the torsional profiles for 3 molecules are shown using QM, GAFF2 and NBO (replacing the torsion energy of GAFF2). Although for some molecules (e.g. ethylbenzene) the impact of replacing torsional energies by hyperconjugation was low, the QM torsional profile was reproduced much more accurately in the

furan and ketone examples. While the energy barriers remained underestimated in the ketone profile, NBO correctly predicted that the 3 energy minima were not equivalent. Furthermore, the furan profile was modeled with far greater accuracy by NBO, which can be explained by the fact that 5-membered rings are poorly parametrized (some not parameterized at all) in GAFF2, hence calculations often rely on "generic" parameters.

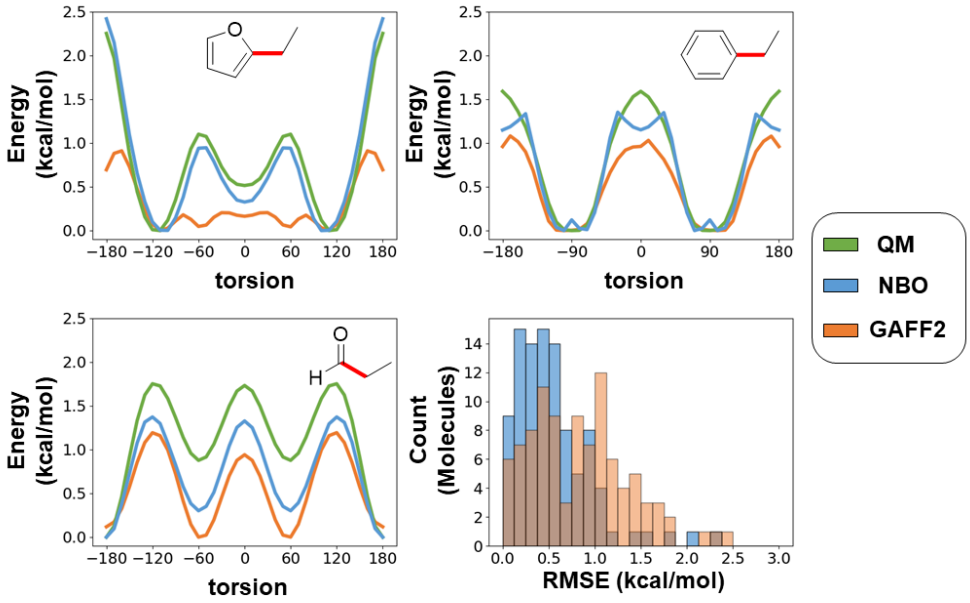


Figure 2.6. Replacing the torsional energy term in GAFF2, by hyperconjugation obtained from NBO (with scaling factors of 0.4 and 0.25). Histogram distribution of the RMSEs of the 98 molecules in our development set between NBO/QM (blue) and GAFF2/QM (orange).

While NBO energies were better at reproducing QM profiles, these calculations cannot be run to generate torsional parameters every time parameters are required. Hence, following an approach used successfully in the development of earlier versions of H-TEQ, our first objective was to understand factors governing the strength of the different interactions based on NBO generated data, and to develop a set of rules based on atomic properties (electronegativity, bond length, aromaticity etc..) which would reproduce NBO interaction energies, and could be calculated on-the-fly. Such a method would allow chemists to generate parameters for any drug-like molecule containing π-systems for use in MD simulations or docking studies.

2.4.b Electronegativity, Aromaticity and π – Hyperconjugation

As previously mentioned, two major factors impact the strength of the σ → π* / π → σ* interactions. First, the energy gap between bonding and antibonding interacting orbitals is directly related to electronegativity. More electronegative elements yield lower-in-energy orbitals; for example, the π and π* orbital energy levels are higher in benzene, than in pyridine. Thus, introduction of heteroatoms into an aromatic ring, or non-aromatic conjugated systems leads to a lowering of the energy levels. The energy levels of π and π* orbitals being quite disparate, we can expect that a lowering of the energy levels would favor interactions unidirectionally (σ → π* or π → σ*), as when the energy gap decreases for one interaction, it is expected to increase for the other.

Secondly, the spatial overlap between interacting orbitals. From heterocyclic chemistry, it is known that the introduction of heteroatoms into aromatic systems results in differences in atomic charges, as well as a greater shielding effect due to heteroatom substituents.[64] It is also expected that more electronegative heteroatoms lead to a polarization of the π orbitals, which strongly impacts the ability of vicinal (σ or σ*) orbitals to overlap and interact. π and π* orbitals have opposite polarization, further reinforcing the fact that as one interaction becomes stronger, the other weakens, as it is impossible for two orbitals with opposite polarizations to have simultaneous strong overlaps with vicinal orbitals. In non-polar π-systems such as alkenes, π and π* orbitals are equally distributed towards both atoms of the double bond. In contrast, introduction of more electronegative heteroatoms (N, O) leads to the polarization of the π orbital towards the heteroatom, which ultimately limits the ability of the π-system to partake in π → σ* donation (lower orbital overlap, increase in energy gap). On the other hand, the π* orbital is polarized towards the carbon atom of the double bond leading to stronger overlap for the σ → π* interaction, resulting in a more prominent acceptor ability. As a rule of thumb, good π acceptors will be poor π donors and *vice versa* (although in some cases both interactions occur with similar magnitudes, Table 2.1). In Figure 2.7, NBO profiles are shown for these specific interactions, indeed we observe that less electronegative elements in the π-system leads to a stronger π → σ*, but weaker σ → π* interaction. The concept of electronegativity is therefore central in understanding the propensity of a system to contribute to π-hyperconjugation.

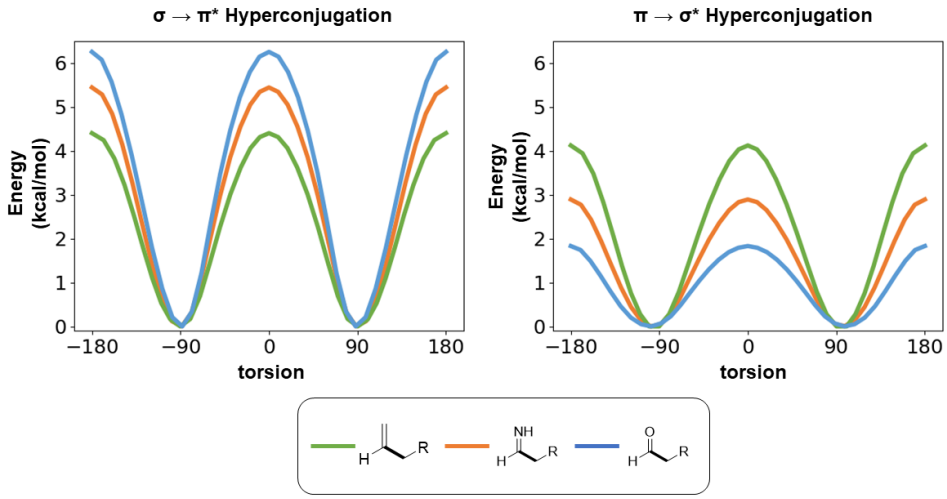


Figure 2.7. Electronegativity of elements in π-systems has an opposing effect for π → σ*(C-R) and σ(C-R) → π* interactions. In this example -R is a methyl group, although this trend extends to all -R groups studied (H, Cl, F, OH, CH_3).

Similarly, the elements involved in the σ group with which the π-system is interacting impact the torsional energy profiles. More electronegative elements act as better donors, and weaker acceptors, and less electronegative elements will be better donors and weaker acceptors. The concept of electronegativity will hence be our major descriptor for π-hyperconjugation.

2.4.c Developing Equations for π-Hyperconjugation

Traditionally, the torsional component of the energy in FFs is calculated using a truncated Fourier series (Eq. 1.4); the number of terms (N) included varies depending on the FF but usually doesn't exceed 6 with 3 being most common. While FFs are empirical in nature, it is essential to understand that each term (V_n) can be interpreted in the context of rotation around a bond and assigned a corresponding chemical meaning. The V_1 terms relates to the *syn* or *anti* preference of two groups, since π-orbitals are somewhat symmetrical (similar density above and below the ring/double bond), the V_1 term should be negligible in our equation. The V_2 term describes the energy cost of rotating around a bond and is related to the strength of the interaction which is maximal at 90° (maximal orbital overlap) and minimal at 0° (no orbital overlap). Finally, the V_3 term can be understood as a correcting factor which can weakly shift the energy barrier, this V_3

term is also related to orbital overlap in the sense that it controls whether it is possible to rotate the bond slightly away from the ideal conformation (±90°) without losing π-hyperconjugation (Appendix 5).

For our purposes, V_2 and V_3 components of the torsion energy were sufficient to describe π-hyperconjugation interactions; π-hyperconjugation torsion profiles obtained from NBO were fitted to derive the V_2 and V_3 parameters for each molecule in the set. Since we are treating both interactions independently, each interaction will be assigned its corresponding V_2 and V_3 value, which will then be summed to describe π-hyperconjugation fully. Furthermore, as we noticed that σ → σ* could not be neglected, we will also add the classical hyperconjugation contribution (weaker), by using our previously developed set of equations.[36] As expected, we found that only the V_2 term was subject to large variations from a molecule to another, hence the V_3 term was assigned a constant value of -0.5 kcal/mol for all molecules. We then concentrated our efforts into the development of an equation to model the V_2 term for both σ → π* and π → σ * interactions, based on our understanding of the effects of electronegativity (χ) on energy levels and polarization of the orbitals involved in these interactions. More specifically how the strengths of these interactions are modulated by the electronegativity of relevant parts of the molecule.

$$V_2(\boldsymbol{\sigma} \rightarrow \boldsymbol{\pi}^*) = \boldsymbol{a}\,\frac{\chi^{\pi 1} + \chi^{\pi 2}}{\chi^{\sigma}} + \boldsymbol{b} \tag{2.2}$$

$$V_2(\boldsymbol{\pi} \rightarrow \boldsymbol{\sigma}^*) = \boldsymbol{c}\,\frac{\chi^{\sigma}}{\chi^{\pi 1} + \chi^{\pi 2}} + \boldsymbol{d} \tag{2.3}$$

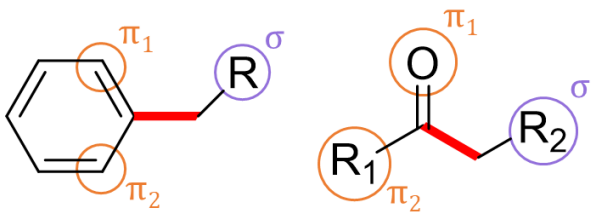


Figure 2.8. Parts of the molecule considered to predict the strength of multiple interactions. The electronegativities of circled atoms are used to calculate the interactions of $\sigma_{(C\text{-}R)} \rightarrow \pi^*$ and $\pi \rightarrow \sigma^*_{(C\text{-}R)}$ (benzene analog) and $\sigma_{(C\text{-}R2)} \rightarrow \pi^*$ and $\pi \rightarrow \sigma^*_{(C\text{-}R2)}$ (ketone).

$$\chi_{group} = \frac{1}{\omega + N}\left(\omega \times \chi_{central} + \sum^{N} \chi_{neighbor}\right) \quad (2.4)$$

It is important to note that the major factors in Eqs. 2.3 and 2.4 are inverses of one another, as effects favoring one interaction, disfavor the other. Here V_2 is shown as proportional to a function based on the electronegativity (χ) of various parts of the molecule (Figure 2.8). The values for electronegativities were obtained from the Pauling scale (Appendix 6)[65] and electronegativity was calculated using the concept of group electronegativity as discussed previously.[35, 36] Indeed for -CF_3 or -CH_3 substituents, we expect the electronegativity of the carbon atom to be much larger in trifluoromethyl than methyl, due to the neighboring chemical environment. Sanderson's electronegativity equilibration principle[66] states that the electronegativity of both atoms in a diatomic system equilibrate to give rise to a new value related to the equilibrium charge distribution in the molecule (this postulate can be extended to all molecules, not simply diatomics). Indeed, while electronegativity measures the ability of an atom to attract electrons towards itself, in polar molecules after the electron density has found its ideal distribution, there is no net flux of electrons away from this optimal distribution; in principle, electronegativity needs to be the same for every atom. A large amount of work has been dedicated to understanding the relationship between electron density and electronegativity, from which researchers have developed many schemes to calculate "group electronegativity" for specific parts of a molecule.[67-70] Although this area of research received a lot of attention in the 80's and 90's, no recent contributions were found in the literature. Some of the schemes for group electronegativity rely on calculating the partial charge of every atom, ultimately requiring QM calculations, and are hence not consistent with our objective to develop a method applicable to high-throughput tasks. We have thus opted to use the simple equation described here by Smith *et al.* (Appendix 7).[70] It is interesting to note that electronegativity equalization methods have also been applied to derive partial charges,[71-73] for example the current implementation of CGenFF (CHARMM force field for drug-like molecules), uses a method which draws from these ideas to generate partial charges.[74]

The strength of the $\sigma \rightarrow \pi^*$ and $\pi \rightarrow \sigma^*$ obtained from NBO correlate well ($r^2 = 0.71$ and 0.81) with the developed rules (Figure 2.9), linear least square regression provided us with the values for a, b, c and d coefficients in Eqs. 2.2 and 2.3. It should be noted that changing the

electronegativity scale (e.g. Pauling units vs. Mulliken units) or weights in the group electronegativity calculation impacted the correlation of our method with NBO derived values. Indeed, while modifications to the equation could in principle improve the accuracy (stronger correlation to NBO) of one of the interactions (e.g. σ → π*), it usually led to a decrease in the correlation with the other (π → σ*). We thus decided to keep the simplest equation (with a weight of 2 for the central atom) and Pauling units as used previously to limit overfitting. Overall, Eqs. 2.2 and 2.3 both use the same electronegativity scales and weights. It should be noted that minor scaling factors were used to differentiate different kinds of π-systems (6-membered, 5-membered, double bonds) such that the same equation could be used for all molecules (Appendix 8).

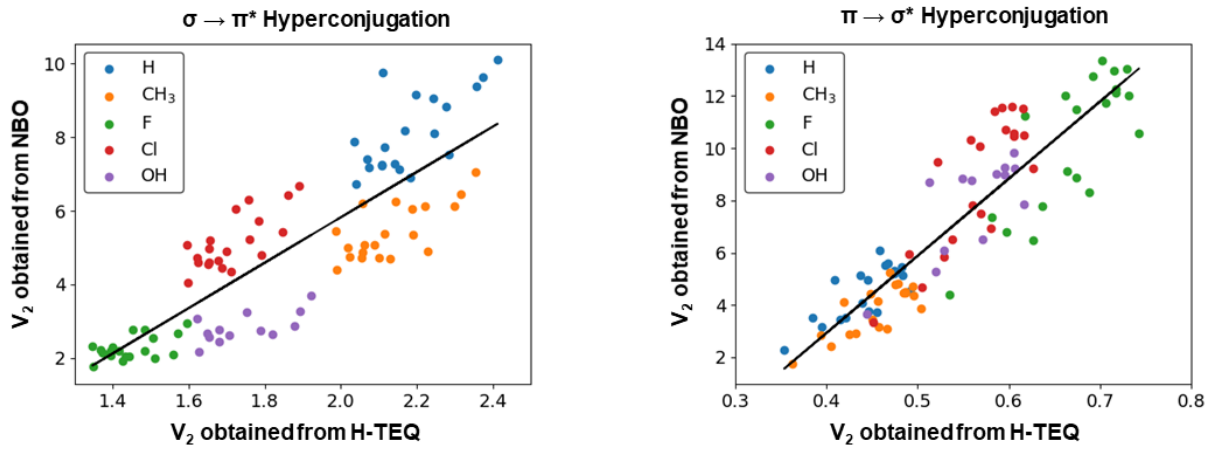


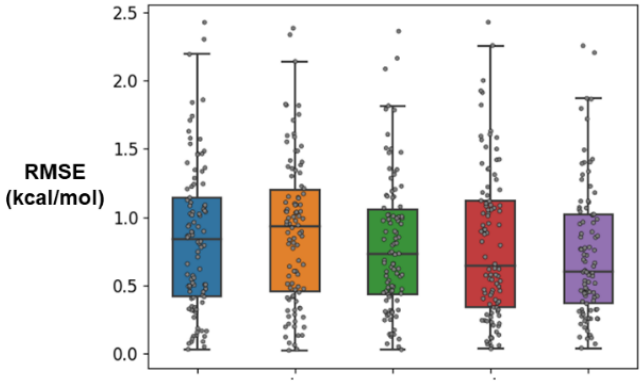
Figure 2.9. Comparison of rules developed (Eqs. 2.2 and 2.3) to describe both π-hyperconjugation modes (σ → π* and π → σ*) with values calculated using NBO analysis. Correlations coefficients obtained are $r^2 = 0.71$ (σ → π*) and $r^2 = 0.81$ (π → σ*).

These equations were implemented into H-TEQ3.0, a program deriving $V_{1\text{-}3}$ parameters for MM calculations. The developed java program also includes the equations from the previous versions of H-TEQ. The parameters (a-d) used in Eqs. 2.2 and 2.3 are shown in Table 2.2. Considering the V_2 values generated for π → σ* and σ → π* are summed to make an overall V_2 term, parameters b and d could be agglomerated into a single parameter. We present them as separate parameters to highlight the fact they were obtained from a linear least square regression around NBO data. To test the transferability of these parameters, we have also performed a bootstrapping analysis (Appendix 9).

Table 2.2. Parameters obtained from the linear regression.

Parameter	Value (kcal/mol)
a	6.16
b	-6.50
c	29.52
d	-8.88

2.4.d Evaluation



Method	GAFF2	H-TEQ 3.0			
Inclusion of V_3 correction	n/a	×	×	✓	✓
Inclusion of σ → σ*	n/a	✓	×	✓	×

Figure 2.10. Performance of GAFF2 and H-TEQ3.0 methods over the development set of 98 molecules. Contributions of σ → σ* hyperconjugation and the V_3 correction factor were also monitored to understand their impact on our method. The black line at the center of each box corresponds to the median value.

The performance of this newly developed method (H-TEQ 3.0) on the development set of molecules was compared alongside GAFF2 against QM energies (Figure 2.10). The contribution of σ → σ* was calculated using the previously published version of H-TEQ 2.0.[36] Overall, our method was found to perform better than GAFF2 when a V_3 correction factor of -0.5 kcal/mol was used. While the inclusion of σ → σ* had a minimal impact on the overall RMSE, the marginally lowest RMSE was found when σ → σ* hyperconjugation was omitted, which contradicted with the results found when replacing raw NBO values (Appendix 1). This discrepancy likely results from the equation modelling σ → σ* hyperconjugation being trained only on sp^3 centers, which

might not be fully transferable to conjugated (sp^2) centers. In conjugated systems, shielding of the σ orbitals by π orbitals is expected, and different geometries of the substituents (109.5° vs. 120°) modifies the ability of orbitals to overlap.

Overall, our method was found to be more accurate than GAFF2 in reproducing QM profiles, doing so without requiring the use of atom type description of the molecules, or any prior parameterization.

2.4.e Performance and Validation

To validate our findings, we have applied the H-TEQ3.0 method to a diverse set of 50 drug molecules (Appendix 4) which does not contain any molecule used to train the model, and a variety of novel π-systems (extended conjugated systems, fused rings, Figure 2.11).

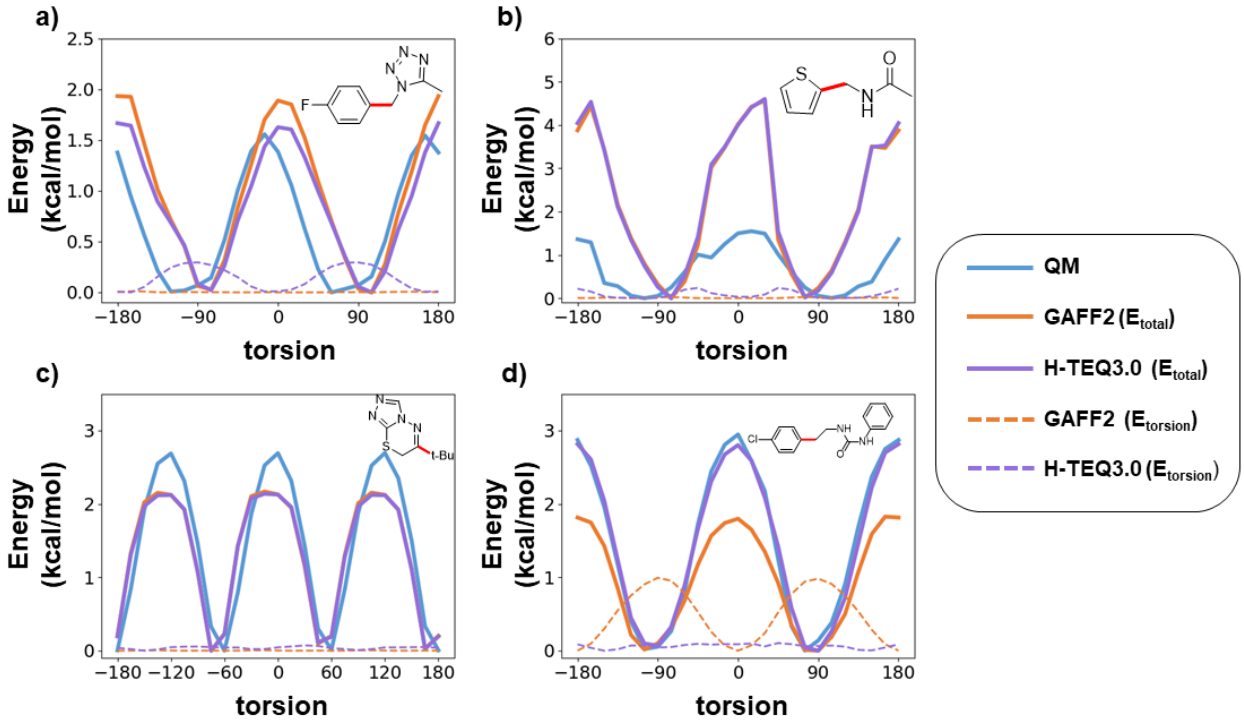
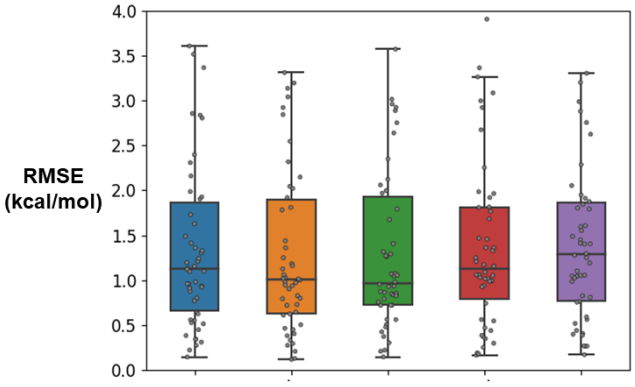


Figure 2.11. Performance of our method on 4 drug-like molecules chosen from the validation set. Full lines correspond to the total energy predicted by each method, dashed lines correspond to the torsional component (of the central bond in red) only. The four molecules shown are (a) 1-(4-fluorobenzyl)-5-methyl-1H-tetrazole, (b) *N*-(thiophen-2-ylmethyl)acetamide, (c) 6-(*tert*-butyl)-7*H*-[1,2,4]triazolo[3,4-b][1,3,4]thiadiazine and (d) 1-(4-chlorophenethyl)-3-phenylurea.

Furthermore, we rotated bonds that were located both in the center and at extremities of the molecules, the former being more important as they lead to the most prominent conformational changes. As for the development set, our method was compared to the widely used GAFF2, and torsional energy was replaced by our equations for π-hyperconjugation and previous equations from H-TEQ 2.0 were employed for σ → σ* hyperconjugation. Again, the effects of the V_3 correcting factor and σ → σ* interactions, were monitored by switching them on/off (Figure 2.12).



Method	GAFF2	H-TEQ 3.0			
Inclusion of V_3 correction	n/a	×	×	✓	✓
Inclusion of σ → σ*	n/a	✓	×	✓	×

Figure 2.12. Performance of GAFF2 and H-TEQ3.0 methods over the validation set of 50 molecules. Contributions of σ → σ* hyperconjugation and the V_3 correction factor were also monitored to understand their impact on our method. One outlier with a large RMSE (~20 kcal/mol) is not shown (Appendix 3). The black line at the center of each box corresponds to the median value.

Regardless of the specific method used (inclusion or not of the V_3, etc.), our method performs on par with the current implementation of GAFF2 (full results in Appendix 10). The V_3 factor was found to slightly negatively impact the accuracy of our method, while the effects from σ → σ* hyperconjugation were found to be minimal. In Table 2.3, results are summarized for the version of H-TEQ including both V_3 and σ → σ*, the slight increase in accuracy over GAFF2 in the development set was lost in the validation set. This should not be confused as a lack of transferability of H-TEQ however, and the larger RMSEs in the validation set result from the larger prevalence of non-bonded interactions as torsions are rotated, in these larger drug-like molecules. Indeed, the non-bonded parameters were calculated using GAFF2, and our method has no impact

on the non-bonded parameters' accuracy. The prevalence of non-bonded interactions is demonstrated in Figure 2.11, indeed GAFF2 and H-TEQ3.0 profiles are very similar, and the torsional component of the energy is weak compared to the overall predicted energy barriers (Fig. 2.11a-d).

Table 2.3. Accuracy of GAFF2 and H-TEQ3.0 to reproduce the torsional profiles over the development and validation sets of molecules.

Method compared to MP2/6-311+G and set of molecules used**	**Average RMSE (kcal/mol)**
GAFF // development set	0.84
GAFF (no torsions) // development set [a]	0.93
H-TEQ3.0 // development set [b]	0.80
GAFF // validation set	1.69
GAFF (no torsions) // validation set [a]	1.78
H-TEQ3.0 // validation set [b]	1.71

[a] The torsional energies related to the central bond were set to 0.
[b] Both $\sigma \rightarrow \sigma^*$ and V_3 were included.

Figure 2.11c shows an example of a correct prediction of GAFF2 and H-TEQ3.0 where the torsional component is equal to 0 as a result of the phase cancelation of the torsion energy, and the weak vdW contribution to the energy alone can correctly predict the energy barriers. In Figure 2.10b, a similar phase cancelation is observed, although the overall profile overestimated the height of the energy barrier by a factor of 2.5. In this case, the flexibility of other parts of the molecule were not modelled well by other energy terms of GAFF2 (vdW, electrostatics). Profiles shown in Figure 2.11a and 2.11d are more interesting. In Figure 2.11a the weak torsional component to the energy predicted by H-TEQ3.0 (out of phase with the overall profile), replacing the null contribution of GAFF2 led to a slightly more accurate profile, while in Figure 2.11d the opposite is seen, an incorrect torsional energy is predicted by GAFF2, which when replaced by a null contribution from H-TEQ led to a much more accurate profile.

An understanding of the magnitude of various π-hyperconjugation modes can explain the origins of the phase cancelation of the torsional terms. Indeed, sp^3 carbons involved in $\sigma \rightarrow \pi^*$ and $\pi \rightarrow \sigma^*$ interactions of interest have 3 substituents which are separated by dihedrals of 120°.

Considering the interactions are essentially modeled by a V_2 term, if the V_2 are of similar magnitude they will cancel out. In the development set, the substituent that was modified could be more electronegative (F, Cl and OH) than the 2 other H atoms on the sp^3 carbon, hence phase cancelation was not observed for these molecules. On the other hand, the majority sp^3 carbon atoms bound to π-systems in drug-like molecules will have two H atoms and another larger group as a substituent, the atom directly attached to the sp^3 carbon being in most cases a carbon atom. The propensity of π-hyperconjugation is similar for both -H and -C($R_1R_2R_3$) unless R_{1-3} are very electronegative, as predicted by NBO calculations (Figure 2.9), which explains why phase cancelation is observed for many drug-like molecules in the validation set. As a null hypothesis, we have also calculated the RMSEs where central torsional energies in GAFF2 were set to 0 and have observed larger RMSEs than when using GAFF2 with torsions or H-TEQ 3.0 (Table 2.3). However, the difference was rather small, which supports the idea that non-bonded interactions play an important role in determining these torsional profiles.

As a result, a major contributor to the energy in bulky drug-like molecules, when a torsion at the center of the molecule is rotated is sterics. Consequently, a correct modeling of non-bonded interactions is of greater importance to correctly predict the conformational energy landscape of such molecules. Polarizable Force Fields are expected to predict these non-bonded interactions with greater accuracy. Methods such as AMOEBA FF, may provide a much more thorough treatment of electrostatic interactions (performed using dipole and quadrupoles moments). However, an automated tool to generate AMOEBA atom types and parameters is yet to be developed, which hindered our ability to perform and include such a comparison in the present study, as atom types and parameters would have to be assigned manually.

2.5 Conclusions

We have shown that replacing the torsional energy calculated by empirical parameters in FFs with more chemically meaningful potentials describing hyperconjugation interactions in conjugated molecules led to accuracies comparable to the widely used GAFF2, without relying on atom types description or parameterization, avoiding common drawbacks known to be associated with these methods. As opposed to previous work performed on saturated molecules,

hyperconjugation is not the predominant factor in determining conformational preference. The self-consistency of FFs (empirical torsion making up for poor non-bonded parameters) thus explains why, for the time being, transferable methods like H-TEQ3.0 do not perform significantly better than empirical methods as long as the other terms (especially non-bonded terms) are not trained concomitantly. The non-transferability of parameters remains a central challenge in FF development, and we expect chemically relevant transferable methods like H-TEQ3.0 to provide more accurate depictions of the energetics of small drug-like molecules, provided that non-bonded interactions are more accurately transcribed. Future research goals include the comparison of our method in MD and docking studies, comparison against a wider range of FFs (including polarizable FFs).

Finally, as the treatment of non-bonded interactions was shown to be problematic in this study, application of the atom type free methodology to describe non-bonded interactions could also be examined, removing entirely the need for atom typing in FFs, ultimately allowing more transferable methods to be applied towards many different SBDD programs.

References

1. Tian, S.; Wang, J.; Li, Y.; Li, D.; Xu, L.; Hou, T., The Application of in Silico Drug-Likeness Predictions in Pharmaceutical Research. *Adv. Drug Deliv. Rev.* **2015**, 86, 2-10.

2. Daina, A.; Michielin, O.; Zoete, V., Swissadme: A Free Web Tool to Evaluate Pharmacokinetics, Drug-Likeness and Medicinal Chemistry Friendliness of Small Molecules. *Sci. Rep.* **2017**, 7, 42717.

3. Durrant, J. D.; McCammon, J. A., Molecular Dynamics Simulations and Drug Discovery. *Bmc Biol.* **2011**, 9.

4. Huang, N.; Jacobson, M. P., Physics-Based Methods for Studying Protein-Ligand Interactions. *Curr. Opin. Drug Disc. Devel.* **2007**, 10, 325-331.

5. Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y.; Bryant, S. H., Structure-Based Virtual Screening for Drug Discovery: A Problem-Centric Review. *AAPS J.* **2012**, 14, 133-41.

6. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, 66, 334-395.

7. Dans, P. D.; Walther, J.; Gomez, H.; Orozco, M., Multiscale Simulation of DNA. *Curr. Opin. Struct. Biol.* **2016**, 37, 29-45.

8. Halgren, T. A., Potential-Energy Functions. *Curr. Opin. Struct. Biol.* **1995**, 5, 205-210.

9. Lazaridis, T.; Karplus, M., Effective Energy Functions for Protein Structure Prediction. *Curr. Opin. Struct. Biol.* **2000**, 10, 139-145.

10. Wang, Z.; Sun, H. Y.; Yao, X. J.; Li, D.; Xu, L.; Li, Y. Y.; Tian, S.; Hou, T. J., Comprehensive Evaluation of Ten Docking Programs on a Diverse Set of Protein-Ligand Complexes: The Prediction Accuracy of Sampling Power and Scoring Power. *Phys. Chem. Chem. Phys.* **2016**, 18, 12964-12975.

11. De Vivo, M., Bridging Quantum Mechanics and Structure-Based Drug Design. *Front. Biosci.* **2011**, 16, 1619-1633.

12. Kenno, V.; Olgun, G.; Alexander, D. M., Jr., Molecular Mechanics. *Curr. Pharm. Des.* **2014**, 20, 3281-3292.

13. Vanommeslaeghe, K.; Yang, M. J.; MacKerell, A. D., Robustness in the Fitting of Molecular Mechanics Parameters. *J. Comput. Chem.* **2015**, 36, 1083-1101.

14. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1996**, 118, 2309-2309.

15. Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P., A New Force-Field for Molecular Mechanical Simulation of Nucleic-Acids and Proteins. *J. Am. Chem. Soc.* **1984**, 106, 765-784.

16. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M., Charmm: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, 4, 187-217.

17. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.;

Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, 102, 3586-3616.

18. Scott, W. R. P.; Hunenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Kruger, P.; van Gunsteren, W. F., The Gromos Biomolecular Simulation Program Package. *J. Phys. Chem. A* **1999**, 103, 3596-3607.

19. Daura, X.; Mark, A. E.; van Gunsteren, W. F., Parametrization of Aliphatic Chn United Atoms of Gromos96 Force Field. *J. Comput. Chem.* **1998**, 19, 535-547.

20. Damm, W.; Frontera, A.; TiradoRives, J.; Jorgensen, W. L., Opls All-Atom Force Field for Carbohydrates. *J. Comput. Chem.* **1997**, 18, 1955-1970.

21. Harder, E.; Damm, W.; Maple, J.; Wu, C. J.; Reboul, M.; Xiang, J. Y.; Wang, L. L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A., Opls3: A Force Field Providing Broad Coverage of Drug-Like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, 12, 281-296.

22. Abel, R.; Harder, E.; Damm, W.; Reboul, M.; Maple, J.; Wu, C. J.; Xiang, J.; Cerutti, D.; Lupyan, D.; Wang, L. L.; Dahlgren, M.; LeBard, D., Opls3 Force Field: An Improved Classical Force Field for the Modeling of Drug-Like Small Molecules, Proteins, Rna, and DNA. *Abstr. Pap. Am. Chem. S.* **2015**, 250.

23. Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J., Development and Testing of the Opls All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, 118, 11225-11236.

24. Kirkpatrick, P.; Ellis, C., Chemical Space. *Nature* **2004**, 432, 823-823.

25. Roos, K.; Wu, C.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L.; Abel, R.; Friesner, R. A.; Harder, E. D., Opls3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J. Chem. Theor. Comput.* **2019**, 15, 1863-1874.

26. Shi, Y.; Xia, Z.; Zhang, J. J.; Best, R.; Wu, C. J.; Ponder, J. W.; Ren, P. Y., Polarizable Atomic Multipole-Based Amoeba Force Field for Proteins. *J. Chem. Theory Comput.* **2013**, 9, 4046-4063.

27. Ren, P. Y.; Wu, C. J.; Ponder, J. W., Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *J. Chem. Theory Comput.* **2011**, 7, 3143-3161.

28. Lopes, P. E. M.; Huang, J.; Shim, J.; Luo, Y.; Li, H.; Roux, B.; MacKerell, A. D., Polarizable Force Field for Peptides and Proteins Based on the Classical Drude Oscillator. *J. Chem. Theory Comput.* **2013**, 9, 5430-5449.

29. Huang, J.; Lopes, P. E. M.; Roux, B.; MacKerell, A. D., Recent Advances in Polarizable Force Fields for Macromolecules: Microsecond Simulations of Proteins Using the Classical Drude Oscillator Model. *J. Phys. Chem. Lett.* **2014**, 5, 3144-3150.

30. Huang, L.; Roux, B., Automated Force Field Parameterization for Nonpolarizable and Polarizable Atomic Models Based on Ab Initio Target Data. *J. Chem. Theory Comput.* **2013**, 9, 3543-3556.

31. Mayne, C. G.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. C., Rapid Parameterization of Small Molecules Using the Force Field Toolkit. *J. Comput. Chem.* **2013**, 34, 2757-2770.

32. Betz, R. M.; Walker, R. C., Paramfit: Automated Optimization of Force Field Parameters for Molecular Dynamics Simulations. *J. Comput. Chem.* **2015**, 36, 79-87.

33. Wang, J.; Kollman, P. A., Automatic Parameterization of Force Field by Systematic Search and Genetic Algorithms. *J. Comput. Chem.* **2001**, 22, 1219-1228.

34. Mobley, D.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K., Open Force Field Consortium: Escaping Atom Types Using Direct Chemical Perception with Smirnoff V0.1. *bioRxiv* **2018**, 286542.

35. Liu, Z. M.; Pottel, J.; Shahamat, M.; Tomberg, A.; Labute, P.; Moitessier, N., Elucidating Hyperconjugation from Electronegativity to Predict Drug Conformational Energy in a High Throughput Manner. *J. Chem. Inf. Model.* **2016**, 56, 788-801.

36. Liu, Z. M.; Barigye, S. J.; Shahamat, M.; Labute, P.; Moitessier, N., Atom Types Independent Molecular Mechanics Method for Predicting the Conformational Energy of Small Molecules. *J. Chem. Inf. Model.* **2018**, 58, 194-205.

37. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, 25, 1157-1174.

38. Pophristic, V.; Goodman, L., Hyperconjugation Not Steric Repulsion Leads to the Staggered Structure of Ethane. *Nature* **2001**, 411, 565-568.

39. Hoffmann, R.; Radom, L.; Pople, J. A.; Schleyer, P. v. R.; Hehre, W. J.; Salem, L., Strong Conformational Consequences of Hyperconjugation. *J. Am. Chem. Soc.* **1972**, 94, 6221-6223.

40. Juaristi, E.; Cuevas, G., Recent Studies of the Anomeric Effect. *Tetrahedron* **1992**, 48, 5019-5087.

41. Alabugin, I. V.; Gilmore, K. M.; Peterson, P. W., Hyperconjugation. *Wiley Interdiscip. Rev. Comput. Mol. Sci* **2011**, 1, 109-141.

42. Alabugin, I. V.; Zeidan, T. A., Stereoelectronic Effects and General Trends in Hyperconjugative Acceptor Ability of Σ Bonds. *J. Am. Chem. Soc.* **2002**, 124, 3175-3185.

43. Alabugin, I. V. Stereoelectronic Effects with Donor and Acceptor Separated by a Single Bond Bridge. In *Stereoelectronic Effects*; 2016.

44. Taylor, R. D.; MacCoss, M.; Lawson, A. D. G., Rings in Drugs. *J. Med. Chem.* **2014**, 57, 5845-5859.

45. Cram, D. J.; Elhafez, F. A. A., Studies in Stereochemistry .10. The Rule of Steric Control of Asymmetric Induction in the Syntheses of Acyclic Systems. *J. Am. Chem. Soc.* **1952**, 74, 5828-5835.

46. Anh, N. T.; Eisenstein, O.; Lefour, J. M.; Tranhuudau, M. E., Orbital Factors and Asymmetric Induction. *J. Am. Chem. Soc.* **1973**, 95, 6146-6147.

47. Burgi, H. B.; Dunitz, J. D., From Crystal Statics to Chemical-Dynamics. *Accounts. Chem. Res.* **1983**, 16, 153-161.

48. Alabugin, I. V.; Bresch, S.; dos Passos Gomes, G., Orbital Hybridization: A Key Electronic Factor in Control of Structure and Reactivity. *J. Phys. Org. Chem.* **2015**, 28, 147-162.

49. Fleming, I., *Molecular Orbitals and Organic Chemical Reactions, Reference Edition*. John Wiley & Sons: NY, 2010.

50. Glendening, E. D.; Landis, C. R.; Weinhold, F., Nbo 6.0: Natural Bond Orbital Analysis Program. *J. Comput. Chem.* **2013**, 34, 1429-1437.

51. von Hopffgarten, M.; Frenking, G., Energy Decomposition Analysis. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, 2, 43-62.

52. Mo, Y. R.; Gao, J. L.; Peyerimhoff, S. D., Energy Decomposition Analysis of Intermolecular Interactions Using a Block-Localized Wave Function Approach. *J. Chem. Phys.* **2000**, 112, 5530-5538.

53. Weinhold, F.; Carpenter, J. E., Some Remarks on Nonorthogonal Orbitals in Quantum Chemistry. *Comput. Theor. Chem.* **1988**, 165, 189-202.

54. Morokuma, K., Molecular Orbital Studies of Hydrogen Bonds .3. C=O H-O Hydrogen Bond in H2co H2o and H2co 2h2o. *J. Chem. Phys.* **1971**, 55, 1236-&.

55. Waller, M. P.; Robertazzi, A.; Platts, J. A.; Hibbs, D. E.; Williams, P. A., Hybrid Density Functional Theory for Π-Stacking Interactions: Application to Benzenes, Pyridines, and DNA Bases. *J. Comput. Chem.* **2006**, 27, 491-504.

56. Fernández, I.; Frenking, G., Direct Estimate of the Strength of Conjugation and Hyperconjugation by the Energy Decomposition Analysis Method. *Chem.: Eur. J.* **2006**, 12, 3617-3629.

57. Carballeira, L.; Pérez-Juste, I., Ab Initio Study and Nbo Interpretation of the Anomeric Effect in Ch2(Xh2)2 (X = N, P, as) Compounds. *J. Phys. Chem. A* **2000**, 104, 9362-9369.

58. Salzner, U.; Schleyer, P. v. R., Generalized Anomeric Effects and Hyperconjugation in Ch2(Oh)2, Ch2(Sh)2, Ch2(Seh)2, and Ch2(Teh)2. *J. Am. Chem. Soc.* **1993**, 115, 10231-10236.

59. Lu, K. T.; Weinhold, F.; Weisshaar, J. C., Understanding Barriers to Internal-Rotation in Substituted Toluenes and Their Cations. *J. Chem. Phys.* **1995**, 102, 6787-6805.

60. Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A., General Atomic and Molecular Electronic-Structure System. *J. Comput. Chem.* **1993**, 14, 1347-1363.

61. Gordon, M. S.; Schmidt, M. W., Advances in Electronic Structure Theory: Gamess a Decade Later. *Theory and Applications of Computational Chemistry: The First Forty Years* **2005**, 1167-1189.

62. Robertson, M. J.; Qian, Y.; Robinson, M. C.; Tirado-Rives, J.; Jorgensen, W. L., Development and Testing of the Opls-Aa/M Force Field for Rna. *J. Chem. Theory Comput.* **2019**, 15, 2734-2742.

63. Campagna-Slater, V.; Pottel, J.; Therrien, E.; Cantin, L. D.; Moitessier, N., Development of a Computational Tool to Rival Experts in the Prediction of Sites of Metabolism of Xenobiotics by P450s. *J. Chem. Inf. Model.* **2012**, 52, 2471-2483.

64. Katritzky, A. R.; Ramsden, C. A.; Joule, J. A.; Zhdankin, V. V. 2.3 - Structure of Five-Membered Rings with One Heteroatom. In *Handbook of Heterocyclic Chemistry (Third Edition)*; Elsevier: Amsterdam, 2010, pp 87-138.

65. Pauling, L., The Nature of the Chemical Bond Iv the Energy of Single Bonds and the Relative Electronegativity of Atoms. *J. Am. Chem. Soc.* **1932**, 54, 3570-3582.

66. Sanderson, R. T., An Interpretation of Bond Lengths and a Classification of Bonds. *Science* **1951**, 114, 670-672.

67. Bratsch, S. G., A Group Electronegativity Method with Pauling Units. *J. Chem. Educ.* **1985**, 62, 101-103.

68. Mullay, J., Calculation of Group Electronegativity. *J. Am. Chem. Soc.* **1985**, 107, 7271-7275.

69. Yang, Z. Z.; Wang, C. S., Atom-Bond Electronegativity Equalization Method .1. Calculation of the Charge Distribution in Large Molecules. *J. Phys. Chem. A* **1997**, 101, 6315-6321.

70. Smith, D. W., Group Electronegativities from Electronegativity Equilibration - Applications to Organic Thermochemistry. *J. Chem. Soc. Faraday Trans.* **1998**, 94, 201-205.

71. Gasteiger, J.; Marsili, M., Iterative Partial Equalization of Orbital Electronegativity - a Rapid Access to Atomic Charges. *Tetrahedron* **1980**, 36, 3219-3228.

72. No, K. T.; Grant, J. A.; Jhon, M. S.; Scheraga, H. A., Determination of Net Atomic Charges Using a Modified Partial Equalization of Orbital Electronegativity Method .2. Application to Ionic and Aromatic-Molecules as Models for Polypeptides. *J. Phys. Chem.* **1990**, 94, 4740-4746.

73. Rappe, A. K.; Goddard, W. A., Charge Equilibration for Molecular-Dynamics Simulations. *J. Phys. Chem.* **1991**, 95, 3358-3363.

74. Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D., Automation of the Charmm General Force Field (Cgenff) Ii: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **2012**, 52, 3155-3168.

3 Predicting the Hybridization of Nitrogen in Force Field Models

3.1 Introduction

3.1.a Prevalence of Nitrogen in Drugs and Biopolymers

In 2014, *Vitaku et al.* have compiled a database containing all U.S. FDA approved pharmaceuticals;[1] out of the 1086 small drug-like molecules present in this set, 84% contain at least one nitrogen atom. Nitrogen is also central in biopolymers; all base pairs which make up nucleic acids contain multiple nitrogen atoms, and amino acids are joined together in proteins by amide bonds.[2] Additionally, compounds which resemble peptides (peptidomimetics), or are directly protein-based such as biologics (e.g. insulin), have been recognized as important classes of therapeutics.[3] Molecular recognition of peptidomimetics (as well as classical drugs) by biopolymers is central to the development of new drugs and requires an accurate understanding of their dynamics and conformation upon binding.[4] Rotation around amide linkages is particularly slow due to the partial double bond character of the C-N bond, and it is well established that the *trans* conformation is preferred in folded proteins to minimize steric clashes.[5] On the other hand, *cis*/*trans* isomerization of the proline residue is one of the rate determining step in protein folding,[6] and *cis*/*trans* isomers exist in a state of dynamic equilibrium in unfolded proteins.[7] Considering the prevalence of nitrogen in drug-like molecules and biopolymers, a rigorous description of the conformations and dynamics of these systems is required in order to provide meaningful descriptions of protein folding and macromolecule-drug interactions in computational structure-based drug design (SBDD) methods.

3.1.b Force Fields in Drug Design

As extensively discussed in this thesis, computational methods are used in virtually all drug discovery projects.[8] For example, virtual screening of libraries of compounds to identify "hits",[9] or more accurate free energy perturbation (FEP), are routinely employed to predict binding

energies of potential therapeutics.[10, 11] While in principle quantum mechanical (QM) provide more trustworthy estimates,[12] they are unsuitable due to their high computational costs, and commonplace applications are usually performed with empirical molecular mechanics (MM) models and their underlying force fields (FFs). Among the many available MM methods, some of the most widely used comprise the AMBER,[13, 14] CHARMM,[15, 16] GROMOS[17, 18] and OPLS[19-22] series, which calculate the potential energy of a molecular system as a sum of bonded and non-bonded interactions (Eq. 3.1) using parameters obtained from precomputed tables.

In MM, every atom in a molecule is assigned an *atom type* based on its element, hybridization state and direct chemical environment. In GAFF (the drug-like molecule FF associated with AMBER),[23] sp^3 nitrogens (amines) are assigned the atom type **n3**, sp^2 nitrogens bound to aromatic rings are assigned the atom type **nh**, and amide nitrogens are assigned the atom type **n**. These atom types determine the behavior of the atom, by dictating which parameters are used to calculate the contribution of each term to the overall energy (Eq. 1.1).

3.1.c Challenges in the Modeling of Nitrogen Containing Compounds

In addition to limitations which have previously been discussed in this thesis (non-transferability of parameters, poor treatment of electrostatic interactions, etc.), the modeling of nitrogen containing molecules is particularly challenging. As will be discussed throughout this report, these problems arise from both limitations in the simple additive functional (Eq. 1.1), and from the implicit treatment of lone pairs (lone pairs are considered by the heteroatom holding the lone pair, not by explicit lone-pairs). Indeed, MM methods do not deal with the electron density explicitly, and the implicit treatment of the many interactions ($n \rightarrow \sigma^*$ hyperconjugation, $n \rightarrow \pi^*$ conjugation, electrostatics) lone pairs partake in is more often insufficient. As early as 1997, *Dixon et al.* have shown that the addition of lone pairs to FFs could improve the modeling of hydrogen bonding, by considering its inherent anisotropic nature.[24] However, direct inclusions of lone pairs in force fields have remained scarce; a few examples include OPLS3 in which off-atom charges are included for aryl nitrogens,[22] the TIP5P water model,[25] and H-TEQ (developed in our lab) which incorporates lone pairs to account for their hyperconjugation interactions.[26] *Allinger et al.* have also highlighted the importance of lone pairs to account for hydrogen-bond directionality with MM4, a class II Force Field,[27] however this FF is not relevant for biomolecular simulations

or SBDD applications, due to its more complex parametrization and greater computational costs.[28] It could also be argued that polarizable methods such as the Drude oscillator models or multipolar expansion models include lone pairs in their refined electrostatic potentials by describing the electron density with additional dynamic charge points.[29, 30] Although, the more sophisticated treatment of electrostatic interactions by polarizable models comes with a greater computational costs (~3-10 times more costly).[31] More recently, *Oroguchi et al.* have shown that a simple inclusion of lone pairs in amino acid residues could improve H-bonding directionality in molecular dynamics (MD) simulations, without significant impact on the computational cost.[32] Overall, while FF explicitly incorporating lone pairs exist, and widely used packages often allow the manual addition of lone pairs in the form of "dummy atoms", most common applications of FFs in SBDD related applications do not include a treatment of lone pairs of electrons explicitly.

We will now discuss the two main challenges associated with the modeling of nitrogen containing compounds. The hybridization state of nitrogen atoms depends on neighbouring chemical environment, in other words, the hybridization state of nitrogen in different molecules (in their most stable conformation) can vary significantly. Nitrogen atoms found in saturated molecules such as trimethyl amine will have a sp^3 hybridization and tetrahedral geometry. In these systems, nitrogen can be a chiral center, and may take both R and S forms (depending on the position of the lone pair), although conversion from one form to the other occurs rapidly if substituents aren't bulky, in a process known as nitrogen inversion. On the other hand, nitrogen atoms bound to π-systems (e.g. peptide bonds) will be found with sp^2 hybridization and planar geometry. The lone pair of electron has a strong p-character, with equal electron density above/below the nitrogen, which maximizes overlap with the π-system. However, the hybridization of nitrogen strongly depends on the nature of the neighbouring π-system and need not be necessarily truly sp^2 (Figure 3.2). For example, *Alabugin et al.* have extensively studied the effects of substituents in anilines on the hybridization of nitrogen, and found that conjugation is in competition with the intrinsic hybridization properties of the amino group (sp^3).[33] Indeed, electron withdrawing substituents on the benzene ring led to an increase in conjugation and more planar (sp^2) nitrogen, while electron donating substituents led to a weaker conjugation and more bent/pyramidal sp^3 nitrogen. It is interesting to note that the amino group intrinsically prefers the sp^3 hybridization state, and it is only in the presence of a strongly accepting π-system (e.g.

carbonyls in peptides, heteroatom containing aromatic rings) that the stabilization due to conjugation (n → π*) is strong enough and leads to a rehybridization of the nitrogen to a sp^2 state.

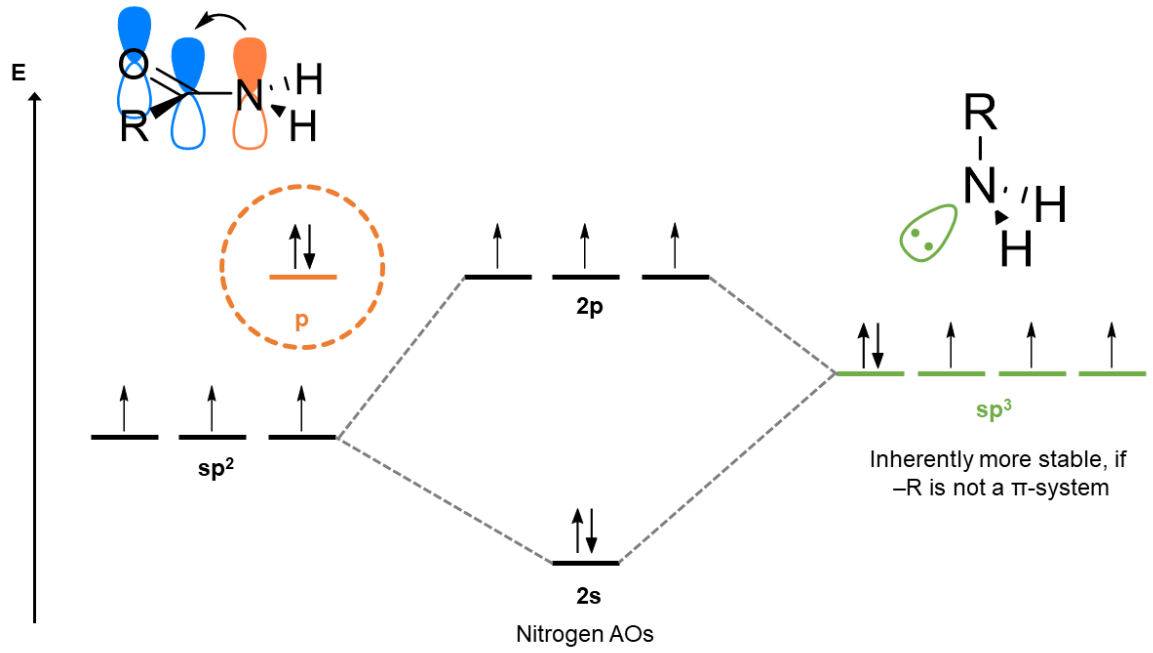


Figure 3.1. Energy Diagram representation of the two extreme hybridization states of nitrogen in organic molecules. The higher in energy p-electrons circled lead to the less stable sp^2 when conjugation in not possible.

In GAFF2 (one of the most commonly used FFs for drug-like molecules), [23] which we will use as a comparison throughout this study, anilines (nitrogens are assigned atom type **nh**) are improperly predicted to be flat, which can be attributed to angle parameters which have equilibrium values of ~120° corresponding to an sp^2 geometry. In fact, the **nh** atom type is assigned to all nitrogen atoms bound to π-systems (with the exception of amides and thioamides), even though molecules have significantly different preferred conformations/hybridizations. Notable distortions from planar sp^2 geometries are also encountered in amides, when important electronic or steric effects are in play, or due to restrictions such as the amide nitrogen being part of a ring (e.g. proline amino acid, β-lactams).[34] These deviations from planarity have profound implications in biological events such as amide bond proteolysis,[35] *cis-trans* isomerization of peptides,[7] and protein splicing.[36] Furthermore, the β-lactam scaffold is present is multiple antibiotic families. (e.g. penicillins, carbapenems, etc.).[37]

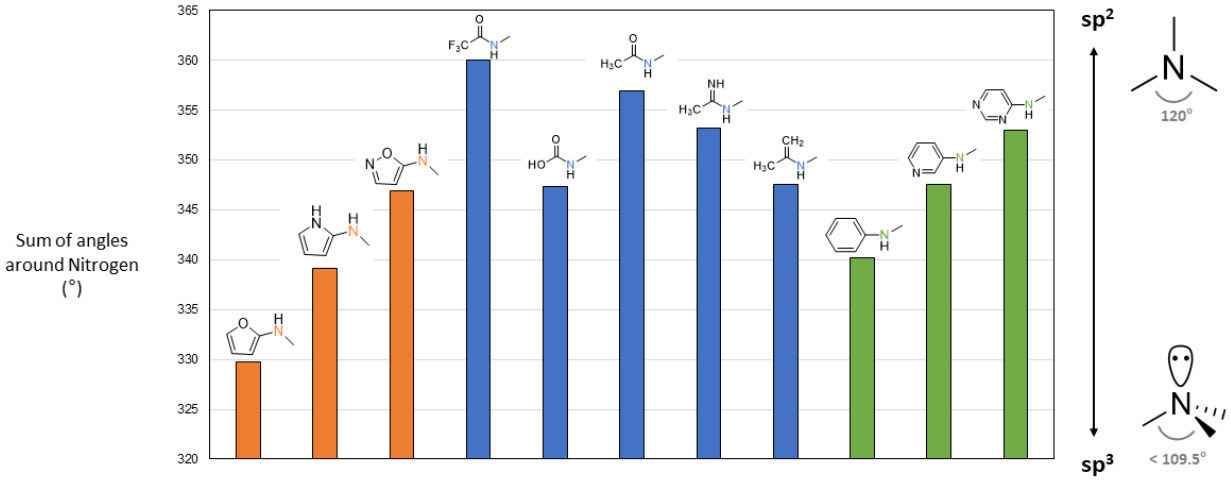


Figure 3.2. Different hybridizations of Nitrogen depending on neighboring chemical environment. Values extracted from QM calculations at the MP2/6-311+G** level of theory, by constraining substituents around nitrogen at various positions (see Methods). Shown are the angle values obtained for the lowest-in-energy structures. Colours highlight the difference in neighbouring π-systems, but are not related to the atom type assigned by GAFF2.

Additionally, the hybridization state of nitrogen can vary in response to rotation around a bond. These changes of hybridization were noticed as early as in the 80's when QM calculations were applied to small amides such as formaldehyde and *N*-methylacetamide (NMA), both in the gas and condensed phases.[38-40] NMA is a small molecule that is often used to model a subset of the properties of polypeptides since it contains the amide bond and is small enough to allow extensive studies at the QM level. It was found that a large energy barrier is associated with rotation from *cis* to *trans* (both sp^2), due to loss of conjugation. The barriers are also slightly larger in water relative to gas phase calculations. More interestingly, the molecule can rotate from *cis* to *trans* following two paths with distinct transition states coined TS_{anti} and TS_{syn} in which nitrogen is sp^3 hybridized (Figure 3.3). The difference in energy between both TSs was initially attributed to electrostatic interactions, although hyperconjugation ($n \rightarrow \sigma^*_{(C\text{-}O)}$) could also be responsible for this disparity. Authors of these previous studies mentioned that subsequent examinations with more accurate methodologies were required as well as more thorough comparison to experimental results. In 2017, *Thakkar et al.* have examined the potential energy surface (PES) of NMA is great detail at the B3LYP/6-311+G** level of theory, confirming the presence of both sp^3 transition

states, while providing a better picture of the actual path taken by NMA during isomerization (slightly different than what is represented in Figure 3.3).[41] However, they omit to discuss which interaction(s) is (are) responsible for the difference in energy between both transition states. The pyramidalization of nitrogen plays a fundamental role in the rate of hydrolysis of amide bonds,[42, 43] and in the cleavage of N-glyosidic bonds (nucleic acids).[44, 45]

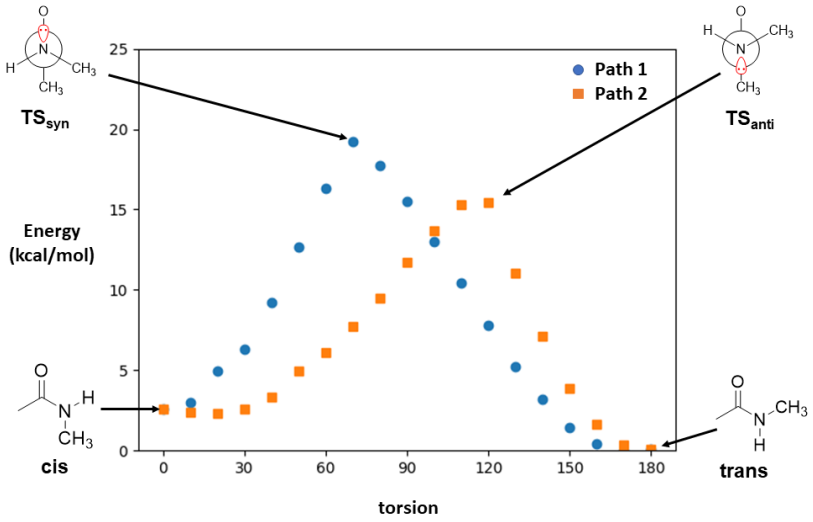


Figure 3.3. *Cis*/*trans* isomerization of the *N*-methylacetamide molecule can take two distinct paths through which different transition states are reached.

To our knowledge, none of the widely used FFs consider the pyramidalization of nitrogen, with the exception of MMFF94 (although over-predicting pyramidalization for some molecules).[46, 47] MMFF94 is outdated however, and infrequently used in SBDD related applications after weaknesses in its parametrization process were recognized, and have not been corrected since.[48] Major hurdles to implement this complex behavior are outlined in the literature such as the large variation of charge distribution with respect to conformation (which non-polarizable FFs cannot include) and the use of redundant coordinates in the FF potential energy (out-of-plane and torsions).[49] *Mobley et al.* also pointed out that traditional atom typing methodologies are not fit to represent the spectrum of possible hybridizations for nitrogen (Figure 3.2), as too many atom types would be required.[50] Furthermore, we believe that the simple additive potential used by FFs might hinder the ability to represent pyramidalization. A more accurate modeling of these nitrogen containing molecules is set to have drastic implications for more

accurate simulations of biological events, as well as for the prediction of binding affinities, considering the omnipresence of nitrogen in drug-like compounds.

3.2 Methods

The first step in our approach consisted in understanding why current FFs performed poorly to predict the conformations (and by extension hybridization) of nitrogen containing molecules. To this end, we constituted a data set of 104 nitrogen containing molecules (Figure 3.4), where the nitrogen atom of interest is placed next to a π-center (sp^2) which can be part of an aromatic ring or a conjugated chain, thereby allowing conjugation to occur. The molecules in our set contain a wide range of π-systems; we have also varied the position on the ring to which the nitrogen substituent is attached, and finally added substituents (R_1 in Figure 3.4) to the nitrogen in order to understand the impact of electron withdrawing (e.g., trifluoromethyl) and electron donating (e.g., hydroxyl) groups on the preferred conformation of nitrogen.

As discussed earlier in this thesis, to ensure an accurate and extensive description of the PES of a molecule, a wide range of conformations needs to be included during the parametrization. The flexibility of small drug-like compounds essentially stems from dihedral angles, and parameters for torsions are typically generated by obtaining torsional energy profiles at the QM level, where a bond is rotated, and the energy is calculated at a few points along that rotation. However, sampling larger portions of conformational space to generate more accurate torsion parameters has been performed in the context of proteins, where Ramachandran (ϕ, ψ) maps are obtained to better reproduce the conformations of peptide backbones.[51] In this example, developers of the AMBER ff14SB generated two-dimensional maps which allowed to discern coupling behavior between adjacent torsions (implicit inclusion in the parameters). In CHARMM, similar maps were obtained and a correction factor cross-term (CMAP) was added to the potential energy function to model the coupling behavior between backbone torsions explicitly.[52, 53] A similar approach (obtaining two-dimensional maps) was carried by *Robertson et al.* in the development of the OPLS-AA/M FF for RNA.[54] On the other hand, we have not found in the literature any application of such extensive searches through conformational space in the context of parametrizing drug-like molecules. This is not surprising as those two-dimensional maps are

computationally demanding (> 100h for the largest molecules in our set), and the vastness of drug-like chemical space has led researchers to favor larger data sets of molecules over a greater sampling of conformational space.[55]

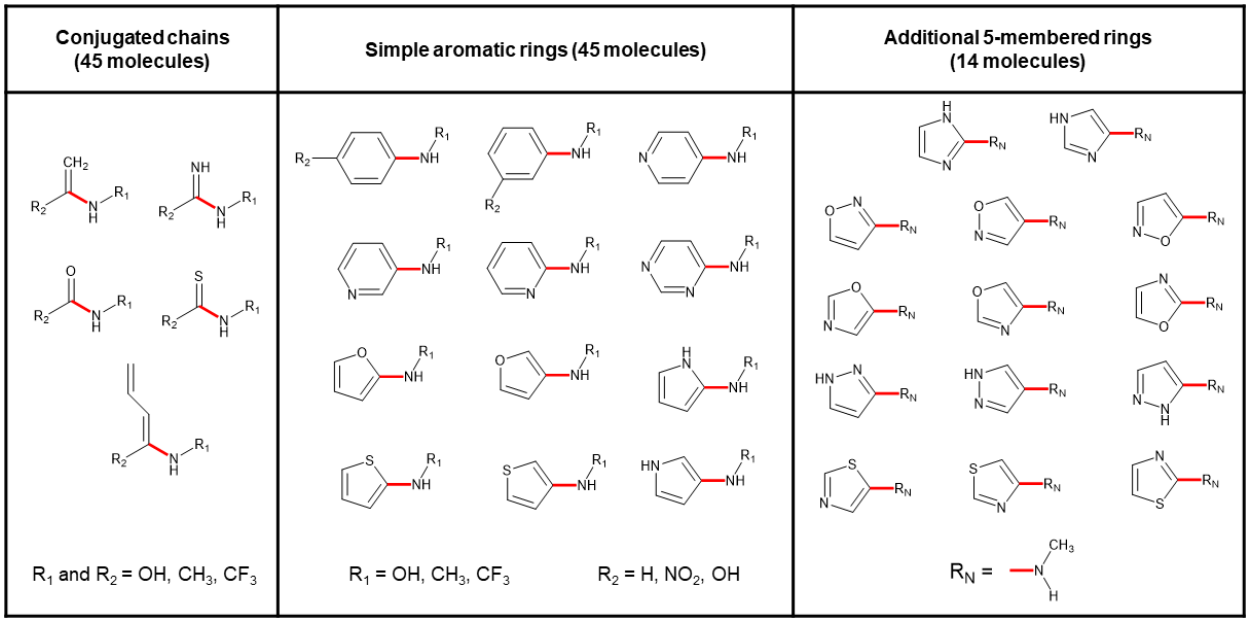


Figure 3.4. Set of molecules used to study the conformational preference and hybridization of nitrogen. Rotated bonds are shown in red.

From this perspective, we decided to obtain two-dimensional maps for each of the molecules in our set, to gain deeper insights into the relationship between hybridization, conformation and chemical environment. The two-dimensional maps were generated by freezing the desired torsion (-180° → 180° with 10° increments), as well as freezing the other substituent on the nitrogen to force the molecule into a specific hybridization state (Figure 3.5). It is important to note that all of the molecules in our set contain a plane of symmetry, hence rotation from -180° → 0° should be equivalent to rotation from 0° → 180°. However, when constraining the $\boldsymbol{\theta_2}$ torsion angle, we incremented it in one direction only, whereas two sp^3 geometries are possible for each $\boldsymbol{\theta_1}$ conformation (Figure 3.5). Therefore, the region 0° → 180° describes one of these possible sp^3 states and the region -180° → 0° describes the other one. Overall, each "2D map" is composed of 288 different conformations, in which two coordinates ($\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$) are frozen and the remaining degrees of freedom in the molecule are optimized at the MP2/6-311+G** level of theory using the software GAMESS-US.[56, 57]

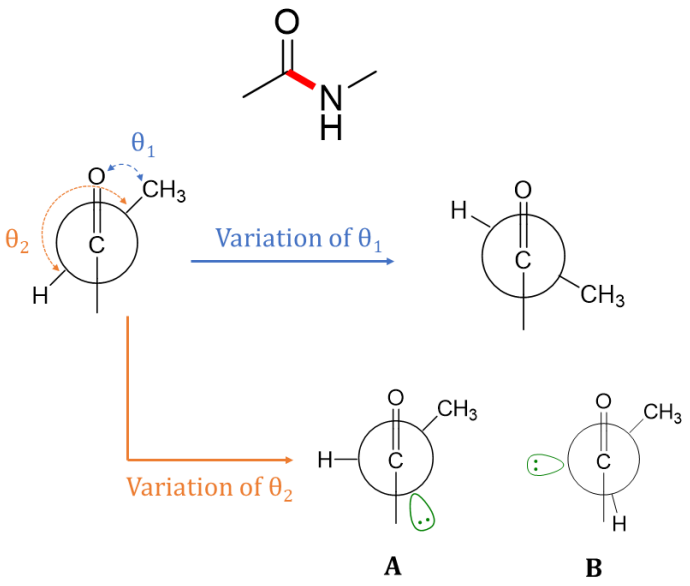


Figure 3.5. Torsion angles varied to generate two-dimensional maps. The θ_1 torsion is scanned from -180° → 180° with 10° increments. For each of these points, the θ_2 torsion if scanned from 110° (slightly more bent than sp^3) to 180° (perfectly sp^2) also using 10° increments. Conformations A and B correspond to the two possible sp^3 states for the same θ_1 value.

It should be noted that due to small size of the molecules under study, the two frozen torsions θ_1 and θ_2 account for virtually all of the conformational change between different points on the map (i.e. there is no major reorganization of other parts of the molecule). Therefore, our maps focus heavily on the nitrogen atom involved in these two torsions, providing details onto their preferred hybridization states. The energy of the 288 conformations were then evaluated using the GAFF2 FF within the AMBER16 package. GAFF2 atom types were assigned using antechamber; charges were assigned using the AM1-BCC method on the lowest energy conformer and were carried to all other conformations. The GAFF2 derived potential energy is computed using the Sander routine. We then proceeded to modify the GAFF2 parameters (and later equation) in order to understand how the MM potential could be modified to reproduce QM derived potentials more accurately.

3.3 Results and Discussion

3.3.a Inadequacy of GAFF2 to Model the Hybridization State of Nitrogen

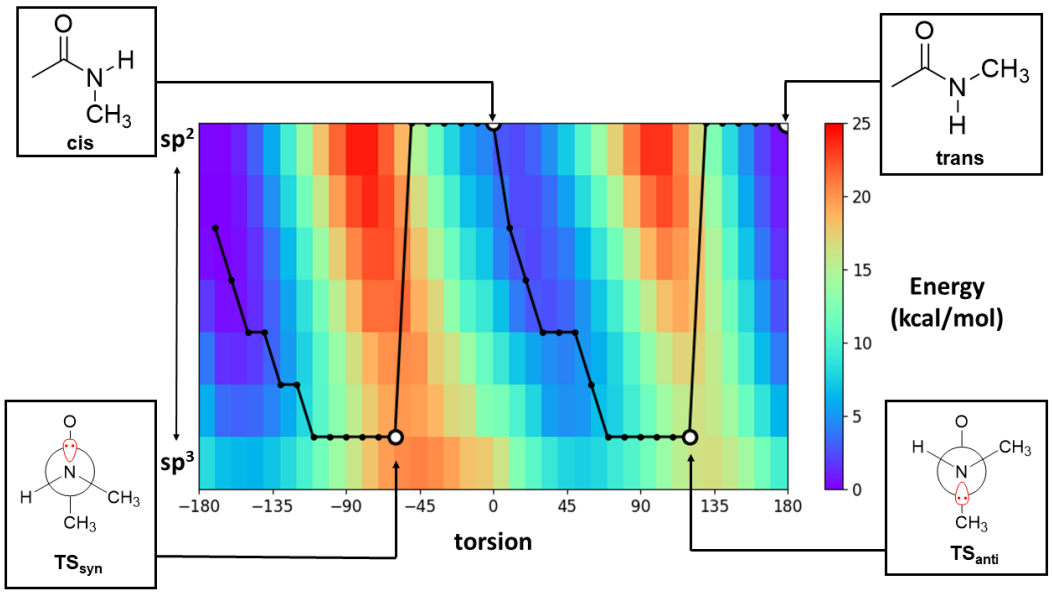


Figure 3.6. 2D map of N-methylacetamide calculated at the MP2/6-311+G** level of theory. The four points circled correspond to both minima (cis and trans) and transition states.

To begin our analysis, we focused our attention on the NMA molecule considering its importance (template for amide bonds) and use in prior studies.[39, 58] Four key conformations corresponding to the two minima and the two TSs described previously, are highlighted in the QM 2D map obtained (Figure 3.6). For each of the 36 increments of the $\boldsymbol{\theta_1}$ angle (from -180° → 180°), we also highlight in black the conformation of lowest energy (w.r.t. the $\boldsymbol{\theta_2}$ angle). Considering we are using discrete increments, the black line connecting these conformations is an approximation of the path taken by the molecule during *cis* → *trans* isomerization, and a more accurate path was reported by *Thakkar et al.*[41] Nevertheless, the black line reflects the variation in hybridization upon rotation; from purely sp^2 (*cis* and *trans*) to sp^3 (both TSs), as observed by other groups.[58] Additionally, it should be noted that the "real" preferred conformation when $\boldsymbol{\theta_1}$ is fixed at a particular value might not be shown on the map since $\boldsymbol{\theta_2}$ is also varied with fixed increments of 10° (e.g. the ideal value of $\boldsymbol{\theta_2}$ when $\boldsymbol{\theta_1}$ is 90° might be 134°, but our maps contains values for 130° and 140° only). Therefore, the geometries of the TSs shown are the closest conformations (to the

TSs described by *Thakkar et al.*) present in our map. The fact that our map only approximates the path from *cis* → *trans* and that TSs are not exact should not be mistaken as liabilities in our methodology, our main objective being to sample as many conformations of the PES as possible to assess current drawbacks in GAFF2. To summarize, the PES of NMA reveals the presence of two large energy barriers (>15 kcal·mol^{-1}) for the rotation of the amide bond. In the regions of low energy (*cis* and *trans*), the sp^2 conformation is preferred, whereas in regions of high energy (TSs) the sp^3 conformation is preferred.

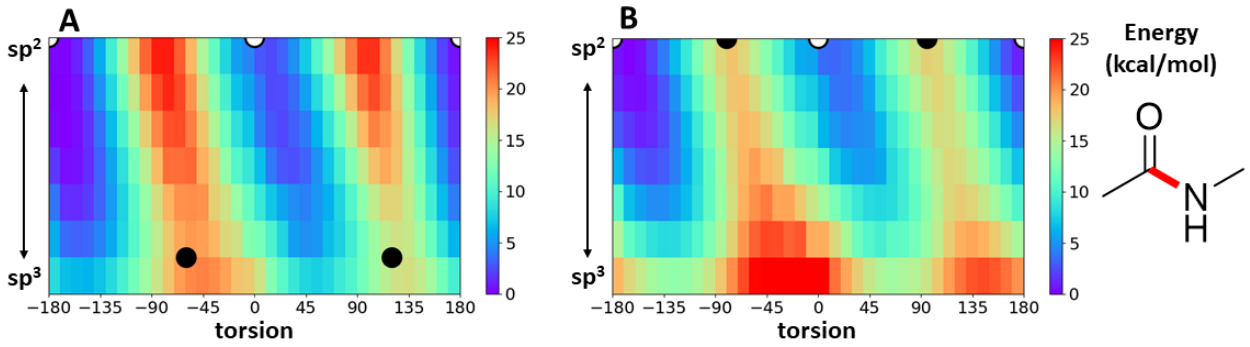


Figure 3.7. 2D maps of NMA calculated with (A) MP2/6-311+G** and (B) GAFF2. White points correspond to energy minima, and black points correspond to transition states (lowest energy points on the energy barrier).

We then compared the 2D maps of NMA (QM vs. GAFF2 in Figure 3.7) and found that GAFF2 failed to correctly reproduce the PES. Indeed, while conformations close to the *cis* and *trans* minima are accurately represented by GAFF2, the sp^3 hybridization state is penalized too greatly (for all values of $\boldsymbol{\theta_1}$). In QM, the saddle points (lowest energy points on the energy barriers corresponding to TSs) are sp^3 hybridized, while GAFF2 incorrectly predicts them to be sp^2 hybridized. Moreover, GAFF2 also penalizes the sp^3 conformers too greatly in the low energy regions (~10 kcal·mol^{-1} in GAFF2 vs. ~ 5 kcal·mol^{-1} in QM). We have accredited these differences to the angle parameters associated to the three angles around the nitrogen atom. Indeed, the angle parameters for the three angles **c-n-c3**, **c-n-hn**, and **c3-n-hn** have equilibrium angle values of 120.7°, 117.5° and 117.7° respectively which too heavily penalized bent sp^3 geometries (angles ~ 109.5°). In fact, simply removing contributions from these three angles led to a decrease in RMSE (3.96 → 3.01 kcal·mol^{-1}), hinting that incorrect angle parameters were at least partially responsible

for the poor prediction, although the RMSE remained large, suggesting that the complex behavior at play might require more than simply reparametrizing angles.

$$Boltzmann\ RMSE\ = \sqrt{\frac{\sum_n \left(E_{MM} - E_{QM}\right)^2 \boldsymbol{e}^{-\left(\frac{\boldsymbol{E_{QM}}}{\boldsymbol{kT}}\right)}}{n}} \tag{3.1}$$

In light of the poor performance of GAFF2 to reproduce the PES of a staple molecule, we sought to investigate its accuracy on our entire set of 104 molecules, by calculating RMSEs between QM and MM maps. Unsurprisingly, deviations between QM and GAFF2 are large (Table 3.1). Authors of the OPLS FFs,[54, 59] have suggested to use Boltzmann scaling factors when fitting parameters to maps containing low and high energy conformers, essentially weighing down the contribution of less significant high energy conformers. The Boltzmann scaling factor (Eq. 3.1) used to complement the traditional RMSE calculation (Eq. 2.1), requires the specification of a temperature. Using too low a temperature results in completely disregarding high energy conformations (Boltzmann factor < 0.1 if E_{confor} > 1.4 kcal·mol^{-1} at 298K), and a temperature of 2000K was found optimal by aforementioned authors. While we report both Boltzmann scaled and unscaled RMSEs in Table 3.1, all RMSEs mentioned in the text (following this comment) will be Boltzmann scaled at a temperature of 2000K.

Table 3.1. Current Performance of GAFF2 to predict the conformational preferences of nitrogen containing molecules.

Method compared to MP2/6-311+G**	**Average RMSE (kcal·mol^{-1})**	**Median RMSE (kcal·mol^{-1})**
GAFF2 2D map [a]	4.44	4.29
GAFF2 2D map [b]	2.40	2.22
GAFF2 1D profile [a]	3.04	2.99
GAFF2 1D profile [b]	2.02	1.86

[a] RMSE calculated without any scaling
[b] RMSE calculated using a Boltzmann scaling factor at 2000K

Overall, GAFF2 cannot reproduce these maps accurately (only 3 molecules with RMSEs < 1 kcal·mol^{-1}), hence we sought to determine which parts of the FF were responsible for these weak predictions. In the case of NMA, we had attributed the disparity between QM and MM to

angle parameters. Thus, to confirm the validity of the torsional parameters assigned to NMA, we extracted the 1D torsional profile (black path in Figure 3.6, Figure 3.8A), which revealed better agreement with QM (RMSE 0.71 kcal·mol^{-1}), supporting our hypothesis that angle parameters were responsible. It is worth reminding our reader that, agreement for these 1D profiles does not mean that the correct hybridization state is assigned overall. On the other hand, torsional parameters were missing to describe molecules in which a nitrogen atom was bound to a 5-membered ring (e.g. Figure 3.8B). In those cases, GAFF2 assigned generic parameters, which resulted in drastically incorrect 1D profiles and by extension 2D maps (Figure 3.9C and 3.9D).

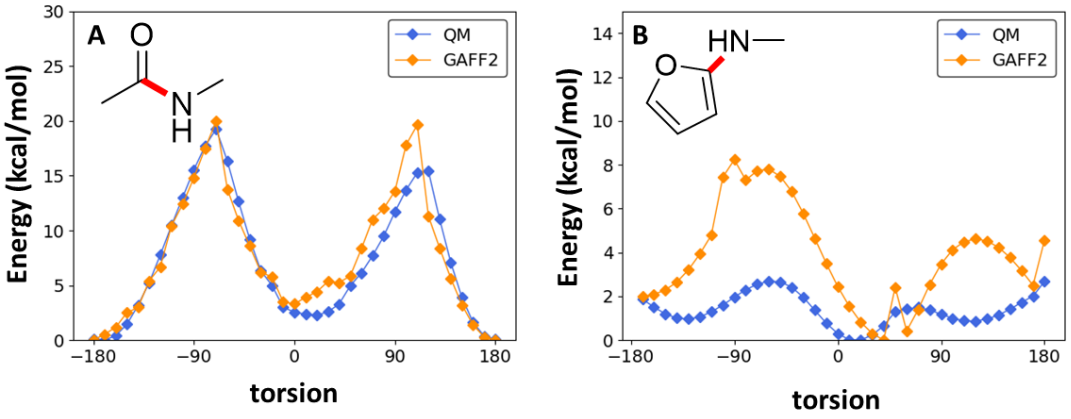


Figure 3.8. 1D torsional profiles of (A) NMA and (B) *N*-methylfuran-2-amine extracted from the 2D maps.

In general, the shapes of the PESs described by our QM 2D maps fell within three distinct categories. The maps of NMA, and other molecules in which the π-system is a conjugated chain (amide, alkene, thioamides, imine, etc.) displayed two regions of low energy (sp^2 favoured) separated by two energy barriers (sp^3 favoured). Molecules within this first category were usually accurately modeled in regions of low energy, but poorly modeled in regions of high energy (TSs predicted to be sp^2 instead of sp^3). Molecules in which a nitrogen atom is bound to a 6-membered π-system constituted our second category; the energy landscapes described by QM were similar to those in category 1. In fact, these maps also contained two regions of low energy separated by two energy barriers (although the magnitude of these barriers were weaker, e.g. ~ 5 kcal·mol^{-1} for *N*-methylaniline, Figure 3.9A). However, in this second category, the preferred hybridization states were now closer to sp^3 for both energy minima and TSs, all of which were incorrectly predicted as

being sp^2 by GAFF2 (Figure 3.9B). Interestingly, we noticed that the shape of the PES of molecules in the second category (two energy barriers, typical for sp^2 centers), and their preferred sp^3 hybridization seemingly contradicted one another. Indeed, torsional profiles involving sp^3 centers (e.g. ethane) typically contain three energy minima and barriers.

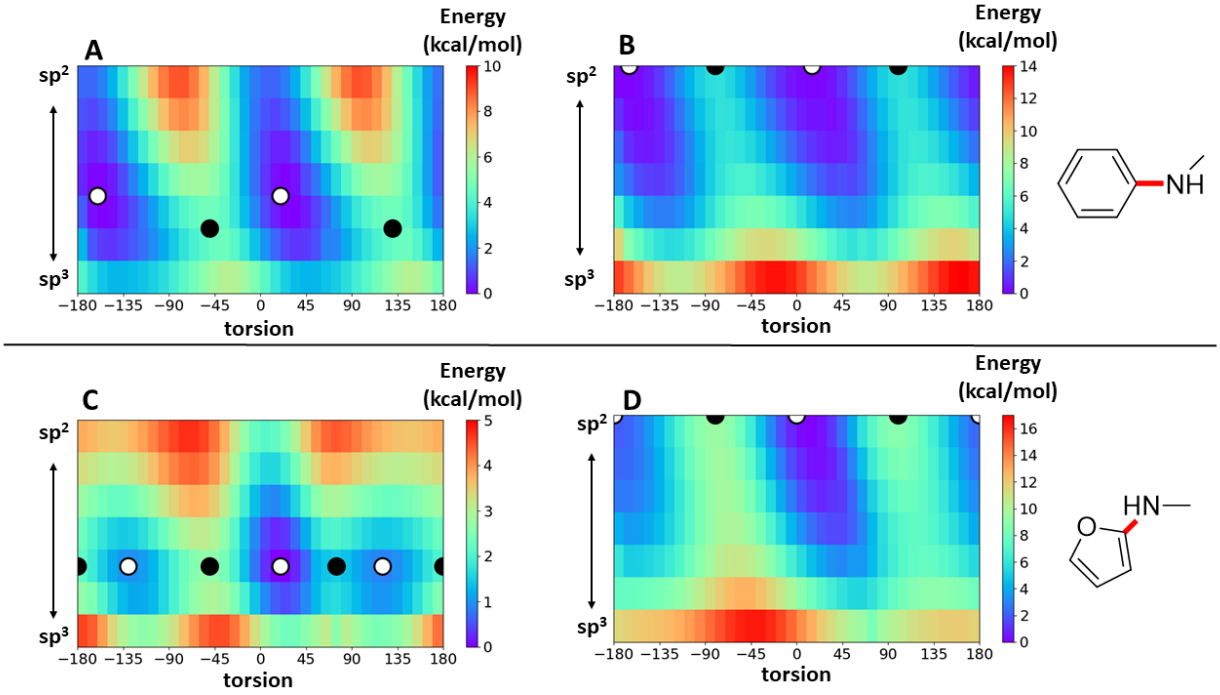


Figure 3.9. 2D maps of N-methylaniline (A = QM, B = GAFF2) and N-methylfuran-2-amine (C = QM, D = GAFF2). Additional points are included on the map to help visualization, white points correspond to local minima and black points correspond to saddle points on energy barriers.

Finally, molecules in which a nitrogen atom is bound to a 5-membered π-system (e.g. *N*-methylfuran-2-amine) comprise the third category, in which the expected threefold profile was observed (Figure 3.8B and 3.9C), further emphasizing the sp^3 nature of these molecules. However, as previously discussed, torsional parameters for molecules within this third category were missing, and generic parameters assigned by GAFF2 led to drastic differences between QM and MM (Figure 3.9C and 3.9D).

3.3.b Improving the Modeling of Nitrogen Containing Molecules in GAFF2.

Our group has followed alternative approaches to tackle computational chemistry problems, our main idea being to directly incorporate well established and transferable chemical principles into computational chemistry programs. For example, we have encoded the Hammond-

Leffler postulate and Curtin-Hammett principle to predict the stereochemical outcome of asymmetric reactions.[60, 61] More recently, we have carried this philosophy to develop a new FF called H-TEQ in which we substituted the torsional energy by equations modeling various hyperconjugation modes, without relying on the concept of atom types.[26, 62] Our latest contribution to H-TEQ has been described in the previous chapter.

From this perspective, we decided to first address the aforementioned limitations found in GAFF2, with the same methodology as previous iterations of H-TEQ (i.e. quantifying hyperconjugation by using data generated from NBO[63]). From our chemical intuition, as well as previous results,[26] we expected the n → π* interaction to predominate, and essentially determine the conformational preferences of the molecules in our set. However, we rapidly realized that we could not rely on NBO to quantify the intensity of n → π* interactions. For instance, NBO predicted a stabilization greater than 90 kcal·mol^{-1} for the n → π* in the NMA molecule, when the QM calculated barrier was found at around 20 kcal·mol^{-1}. Moreover, the n → π* stabilization in *N*-methylaniline was predicted to be around 40 kcal·mol^{-1} which is more than 8 times greater than its QM energy barrier (smaller than 5 kcal·mol^{-1}). Considering all hyperconjugation modes (some of which are out of phase with the n → π*, thereby reducing the overall barrier height), did not lead to a correct match between QM and NBO profiles either. As discussed in the previous chapter, the inner workings of NBO involve constructing a Lewis like representation of the molecule by localizing orbitals. Considering the systems under study are highly delocalized, we questioned the ability of NBO to describe them. Furthermore, too many hyperconjugation modes (i.e. σ → σ*, σ → π*, π → σ*, n → σ* and n → π*) are present in these molecules, and a potential accumulation of small errors could have impaired the predictive capabilities of our method, had we chosen to follow this approach. Overall, we decided to move away from the previously employed methodology, and abandoned NBO as a tool to understand bonding/antibonding interactions in this present context.

In the introduction, we discussed how, in our opinion, the explicit introduction of lone pairs could potentially improve the modeling of nitrogen containing compounds. However, it is difficult to obtain the position of the lone pair from a QM calculation as it is delocalized, hence including a lone-pair bead (described by a point) becomes problematic. Further, new terms for lp-atom-atom

angles and lp-atom "bonds" would have to be developed. To our knowledge, it is not possible to extract this kind of information from QM calculations (i.e. it is not possible to perform a PES scan with the lone pair as one of the coordinates). While we could in principle develop empirical parameters, this would contradict with our initial objective of having a method based on well founded chemistry principles. Another option would be to freeze the lone pair at a set distance from the nitrogen and with set angles (with the nitrogen's other substituents), however this probably wouldn't allow to model the change from sp^2 to sp^3 as the angles mentioned would have to change too. Overall, the inclusion of lone pairs as "dummy-atoms" which participate in valence interactions proves to be particularly challenging, mostly because target data from which meaningful parameters could be fit to cannot be easily obtained.

In light of these difficulties, we then decided to approach the problem from another perspective. By inspecting the PES of NMA in more detail, we noticed an intrinsic link between torsions and angles. Concretely, as the C-N bond is rotated from *cis* (sp^2 geometry, angles around nitrogen are 120°) to *trans*, the molecule passes by a sp^3 favoured region (angles around nitrogen close to 109.5°). In order to reflect the association of two motions, cross-terms can be included in the MM potential, thus we though that incorporating an angle-torsion cross-term could force the molecules in their correct hybridization state, throughout the rotation of the C-N bond. More specifically, we have developed Eqs. 3.2 and 3.3, inspired from the regular angle term (Eq. 1.3), such that conformations in which angles divert from their equilibrium value θ_{eq} are penalized. However, the value of θ_{eq} now varies w.r.t. the torsion as the bond is rotated and we refer to it as θ_{new}. *Mannfors et al.* have described that to calculate proper forces in MD simulations, each energy term needs to be attributed to the correct atoms.[49] In the present context, all 3 angles around nitrogen hold similar values (i.e. they all bend at the same time when becoming sp^3), hence we apply our additional cross-term 3 times (once for each angle) to reflect for that behavior. We also calculate the variation in Eq. 3.3 by summing up all 4 dihedrals (φ), instead of arbitrarily choosing one.

$$E_{cross-term} = \boldsymbol{K_\theta}\,(\theta - \boldsymbol{\theta_{new}})^2 \tag{3.2}$$

$$\boldsymbol{\theta_{new}} = \boldsymbol{\theta_{avg}} + \boldsymbol{\theta_{var}}\left[\sum_{i=1}^{4} cos(2\varphi_i)\right] \tag{3.3}$$

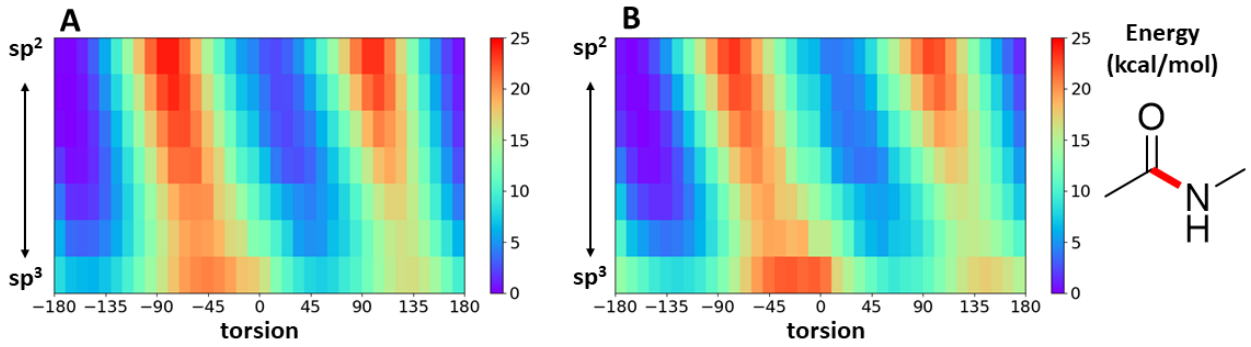


Figure 3.10. 2D maps of NMA at the QM level (A), and with a torsion angle cross-term included in the GAFF2 potential (B).

As expected, the 2D map of NMA was much better reproduced when including our cross-term (RMSE decreases from 1.28 to 0.72 kcal·mol^{-1}, Figure 3.10). Initially we expected our new parameters θ_{avg} and θ_{var} to reflect an underlying physical meaning (i.e. θ_{avg} describes angles in the molecule in average hybridization state and θ_{var} describes the magnitude of the variation in hybridization). Concretely, in the case of NMA where angles vary from ~ 120° to 109.5°, we expected θ_{mid} to be ~115° and θ_{var} to be equal to ~1.25 (5/4 as the cosine varies from -4 to +4). However, the optimal parameters (those which minimized the RMSE), were different from our initial guesses ($\theta_{avg} = 71$ and $\theta_{var} = 1$). As during the optimization all other parameters were kept fixed, we presumed that our method empirically corrected for erroneous parameters in GAFF2 (notably, angle parameters). Implementing the cross-term such that it replaced the original angle terms (instead of supplementing them) led to somewhat more meaningful values ($\theta_{avg} = 105$ and $\theta_{var} = 1$), although θ_{avg} was still far from our initial assumption of 115°, suggesting once again that as we optimized the parameters, the ideal values made-up for pre-existing errors in the FF. Altogether, the addition of a cross-term improves the modeling of NMA, as the TSs now have the correct sp^3 geometry without impacting the low energy region (Figure 3.10).

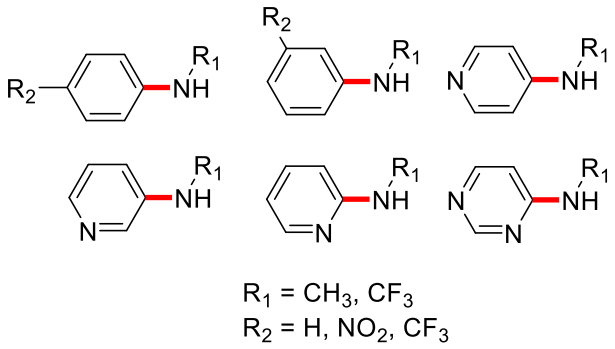


Figure 3.11. Set of anilines, pyridines-amines and pyrimidines-amines used to tune the angle parameters within GAFF2.

After having shown that the PES of NMA could be more rigorously modeled by an angle-torsion cross-term, we sought to apply the cross-term to molecules in the second category (i.e. both minima and TSs are closer to sp^3). When doing so, the optimal value for θ_{var} was 0, indicating that the hybridization did not vary upon rotation (in agreement with QM maps) transforming Eq. 3.2 into an additional angle term, instead of a cross-term. Therefore, we decided to inspect the validity of the angle parameters currently in GAFF2 and examine whether a simple modification to these parameters could reproduce the PES of these molecules more accurately. Concretely, we took the 18 molecules in category two (Figure 3.11), for which the angles around nitrogen are all described by the same three angle types (namely ca-nh-c3, hn-nh-c3 and ca-nh-hn). First, we modified the equilibrium angles by taking values from the saddle points on *N*-methylaniline's QM map, which led to a decrease in RMSE (2.18 kcal·mol^{-1} → 1.81 kcal·mol^{-1}, over the 18 molecules) suggesting that equilibrium angles were incorrect for these molecules. Then, we also inspected whether the force constants could be modified and found that it led to a further decrease in RMSE (2.18 → 1.36 kcal·mol^{-1}), suggesting that new angles parameters could greatly improve the modeling of these molecules. Additionally, we performed a traditional PES scan of these three angles in *N*-methylaniline (Figure 3.12), as generally performed to generate angle parameters, which also led to a decrease in RMSE (2.18 → 1.43 kcal·mol^{-1}). Interestingly, we found that two of the three angles now had θ_{eq} values of 112° and 113°, which corresponds to a hybridization between sp^2 and sp^3. We note that the last angle with a θ_{eq} value of 119° (sp^2) was probably due to a steric clash between the methyl hydrogen, and the ortho-hydrogen.

Table 3.2. Modifications to the angle parameters used to model the 18 molecules shown in Fig. 3.11 (anilines, pyridines, pyrimidines).

Angle (GAFF2 atom types)	**ca-nh-c3**	**hn-nh-c3**	**ca-nh-hn**	**RMSE (kcal·mol^{-1})**
Parameters in GAFF2 (K_θ, θ_{eq})	65.25, 120°	46.42, 116°	48.79, 116°	2.18
Modified parameters A (K_θ, θ_{eq}) [a]	65.25, 112.5°	46.42, 108.5°	48.79, 108.5°	1.81
Modified parameters B (K_θ, θ_{eq}) [b]	25, 112.5°	45, 108.5°	25, 108.5°	1.36
Modified parameters C (K_θ, θ_{eq}) [c]	60, 119°	43.4, 113°	46, 112°	1.43

All values of K_θ are in units of [kcal/(mol·radians^2)]
[a] Values for θ_{eq} taken from the saddle points on the QM 2D map of *N*-methylaniline
[b] Values for θ_{eq} obtained as in [a], values for K_θ were further optimized
[c] Values for K_θ and θ_{eq} obtained from the angle scans shown in Fig. 3.12

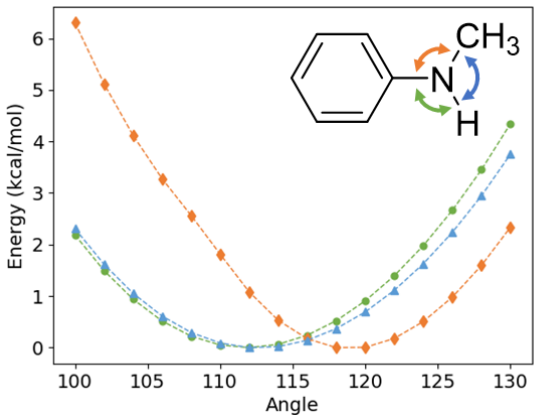


Figure 3.12. PES scan of the three angles around nitrogen of *N*-methylaniline obtained at the MP2/6-311+G** level of theory.

To understand more concretely how the modification of these angle parameters impacted the overall PES, we then plotted the 2D map of *N*-methylaniline using the modified angle parameters (set B in Table 3.2). We observed that the hybridization of both minima and TSs were now predicted to be closer to sp^3 (Figure 3.13), which also extended to the 18 other molecules. On the other hand, the barrier height did not precisely match QM predictions, which suggested that torsional parameters could also be improved. Overall, the new angle parameters we propose here

(notably sets B and C, considering they lead to greater decreases in RMSE) could be directly incorporated into GAFF2. However, we do not argue that these are the best parameters that could describe these molecules. Parameter sets B and C are quite different, and in both cases, the RMSEs remained unsatisfactory to describe the PES of molecules containing conjugated nitrogens.

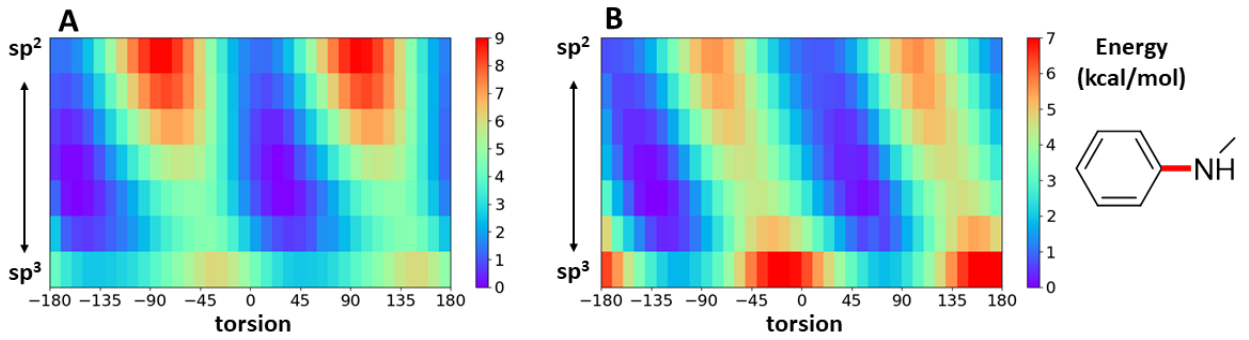


Figure 3.13. 2D maps of *N*-methylaniline at the QM level (A) and using GAFF2 with corrected angle parameters (B).

We expect that a further refinement of the torsional parameters could solve this issue. However, the way these torsional parameters would be developed remains a challenge. Should the torsions be refined after having fixed the angle parameters to new, more accurate values? Should both angles and torsions be optimized simultaneously, from 2D maps in which conformational space is sampled to a greater extent than typical angle and torsion scans? Is it sufficient to look at MP2/6-311+G** quantum data, or should other functionals be employed? Would it be possible to move away from an atom type based approach and include lone pairs explicitly in the FF description of these molecules? All of these technical questions highlight the fact that while the equations to model each interaction in FFs are very simple, complexity arises when multiple degrees of freedom impact one another. In order to efficiently parametrize molecules containing conjugated nitrogens, it is necessary to first answer these questions, and build robust and appropriate parametrization strategies. Then, these methodologies could be used to parametrize all molecules discussed, those within category three being most important considering they are currently missing both angle and torsion parameters. Further, molecules in category one could also benefit from such a re-parametrization, onto which an angle-torsion cross term could ultimately be applied, which we expect would now hold more physically meaningful parameters.

3.4 Conclusions and Future Work

To conclude, we have analyzed in depth the current performance of GAFF2 to predict the conformational preference and hybridization of molecules containing conjugated nitrogen substituents. We have found that GAFF2 performs particularly poorly in that respect and established three different ways in which GAFF2 could be further refined. In our first category, the application of an angle-cross term was found to more accurately model the PES of molecules in which nitrogen is bound to a conjugated chain, where the preferred hybridization states vary along rotation around a torsion. In our second category, a simple refinement of the angle parameters showed promising improvements which could be directly incorporated into GAFF2, although the torsional parameters remained imperfect. Finally, in the third category, current torsional parameters were lacking, and generic parameters led to drastically incorrect predictions of the PES. Overall, the development of new methodologies to parametrize these molecules should allow to generate more physically meaningful parameters. These new methodologies could be applied to all three classes of molecules, and from this new point, the incorporation of an angle-torsion cross-term could be reassessed. Considering the prevalence of nitrogen atoms in drugs, these ameliorations are expected to contribute greatly in the advance of FF based methods for SBDD, by estimating protein-ligand binding affinities and ligand conformations upon binding more truthfully. Further, a more rigorous description of the amide bond should result in more accurate simulations of proteins when using tools such as MD.

In more detail, the development of new methodologies to parametrize these molecules should be the first priority to continue this work. In our opinion, a good hypothesis to try would be to: first obtain more robust angle parameters, then obtain torsional scans in which the molecule is frozen into an sp^2 state from which torsional parameters can be obtained (which should essentially be V_2). Finally, these new parameters can be projected on the 2D maps generated in this work, and if a good match is observed, it can then be concluded that these parameters are adequate. In fact, they would be transferable in conformational space (which is another form of transferability we have not discussed extensively in this thesis). However, if the new parameters are not suitable, the methodology would need to be revisited. Finally, the treatment of 1,4 non-bonded interactions might pose problems to generate meaningful torsional parameters, to avoid or

reduce such limitations an alternative would be to use a polarizable FF (such as AMOEBA) to obtain the rest of the FF terms.

References

1. Vitaku, E.; Smith, D. T.; Njardarson, J. T., Analysis of the Structural Diversity, Substitution Patterns, and Frequency of Nitrogen Heterocycles among U.S. Fda Approved Pharmaceuticals. *J. Med. Chem.* **2014**, 57, 10257-10274.
2. Montalbetti, C. A. G. N.; Falque, V., Amide Bond Formation and Peptide Coupling. *Tetrahedron* **2005**, 61, 10827-10852.
3. Craik, D. J.; Fairlie, D. P.; Liras, S.; Price, D., The Future of Peptide-Based Drugs. *Chem. Biol. Drug. Des.* **2013**, 81, 136-147.
4. Fowler, S. A.; Blackwell, H. E., Structure-Function Relationships in Peptoids: Recent Advances toward Deciphering the Structural Requirements for Biological Function. *Org. Biomol. Chem.* **2009**, 7, 1508-1524.
5. Ramachandran, G. N.; Venkatachalam, C. M., Stereochemical Criteria for Polypeptides and Proteins .4. Standard Dimensions for Cis-Peptide Unit and Conformation of Cis-Polypeptides. *Biopolymers* **1968**, 6, 1255-+.
6. Brandts, J. F.; Halvorson, H. R.; Brennan, M., Consideration of the Possibility That the Slow Step in Protein Denaturation Reactions Is Due to Cis-Trans Isomerism of Proline Residues. *Biochemistry* **1975**, 14, 4953-4963.
7. Nguyen, K.; Iskandar, M.; Rabenstein, D. L., Kinetics and Equilibria of Cis/Trans Isomerization of Secondary Amide Peptide Bonds in Linear and Cyclic Peptides. *J. Phys. Chem. B* **2010**, 114, 3387-3392.
8. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Jr., Computational Methods in Drug Discovery. *Pharmacol Rev* **2014**, 66, 334-395.
9. Kalyaanamoorthy, S.; Chen, Y. P. P., Structure-Based Drug Design to Augment Hit Discovery. *Drug Discov. Today* **2011**, 16, 831-839.
10. Jiang, W.; Roux, B., Free Energy Perturbation Hamiltonian Replica-Exchange Molecular Dynamics (Fep/H-Remd) for Absolute Ligand Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2010**, 6, 2559-2565.
11. Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R., Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, 137, 2695-2703.
12. De Vivo, M., Bridging Quantum Mechanics and Structure-Based Drug Design. *Front. Biosci.* **2011**, 16, 1619-1633.
13. Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P., A New Force-Field for Molecular Mechanical Simulation of Nucleic-Acids and Proteins. *J. Am. Chem. Soc.* **1984**, 106, 765-784.

14. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1996**, 118, 2309-2309.

15. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M., Charmm: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, 4, 187-217.

16. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, 102, 3586-3616.

17. Daura, X.; Mark, A. E.; van Gunsteren, W. F., Parametrization of Aliphatic Chn United Atoms of Gromos96 Force Field. *J. Comput. Chem.* **1998**, 19, 535-547.

18. Scott, W. R. P.; Hunenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Kruger, P.; van Gunsteren, W. F., The Gromos Biomolecular Simulation Program Package. *J. Phys. Chem. A* **1999**, 103, 3596-3607.

19. Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J., Development and Testing of the Opls All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, 118, 11225-11236.

20. Damm, W.; Frontera, A.; TiradoRives, J.; Jorgensen, W. L., Opls All-Atom Force Field for Carbohydrates. *J. Comput. Chem.* **1997**, 18, 1955-1970.

21. Abel, R.; Harder, E.; Damm, W.; Reboul, M.; Maple, J.; Wu, C. J.; Xiang, J.; Cerutti, D.; Lupyan, D.; Wang, L. L.; Dahlgren, M.; LeBard, D., Opls3 Force Field: An Improved Classical Force Field for the Modeling of Drug-Like Small Molecules, Proteins, Rna, and DNA. *Abstr. Pap. Am. Chem. S.* **2015**, 250.

22. Harder, E.; Damm, W.; Maple, J.; Wu, C. J.; Reboul, M.; Xiang, J. Y.; Wang, L. L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A., Opls3: A Force Field Providing Broad Coverage of Drug-Like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, 12, 281-296.

23. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, 25, 1157-1174.

24. Dixon, R. W.; Kollman, P. A., Advancing Beyond the Atom-Centered Model in Additive and Nonadditive Molecular Mechanics. *J. Comput. Chem.* **1997**, 18, 1632-1646.

25. Mahoney, M. W.; Jorgensen, W. L., A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid, Nonpolarizable Potential Functions. *J. Chem. Phys.* **2000**, 112, 8910-8922.

26. Liu, Z. M.; Barigye, S. J.; Shahamat, M.; Labute, P.; Moitessier, N., Atom Types Independent Molecular Mechanics Method for Predicting the Conformational Energy of Small Molecules. *J. Chem. Inf. Model.* **2018**, 58, 194-205.

27. Lii, J.-H.; Allinger, N. L., The Important Role of Lone-Pairs in Force Field (Mm4) Calculations on Hydrogen Bonding in Alcohols. *J. Phys. Chem. A* **2008**, 112, 11903-11913.

28. Kenno, V.; Olgun, G.; Alexander, D. M., Jr., Molecular Mechanics. *Curr. Pharm. Des.* **2014**, 20, 3281-3292.

29. Baker, C. M.; Lopes, P. E. M.; Zhu, X.; Roux, B.; MacKerell, A. D., Accurate Calculation of Hydration Free Energies Using Pair-Specific Lennard-Jones Parameters in the Charmm Drude Polarizable Force Field. *J. Chem. Theory Comput.* **2010**, 6, 1181-1198.

30. Ponder, J. W.; Wu, C. J.; Ren, P. Y.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T., Current Status of the Amoeba Polarizable Force Field. *J. Phys. Chem. B* **2010**, 114, 2549-2564.

31. Baker, C. M.; Best, R. B., Matching of Additive and Polarizable Force Fields for Multiscale Condensed Phase Simulations. *J. Chem. Theory Comput.* **2013**, 9, 2826-2837.

32. Oroguchi, T.; Nakasako, M., Influences of Lone-Pair Electrons on Directionality of Hydrogen Bonds Formed by Hydrophilic Amino Acid Side Chains in Molecular Dynamics Simulation. *Sci. Rep.* **2017**, 7, 15859.

33. Alabugin, I. V.; Manoharan, M.; Buck, M.; Clark, R. J., Substituted Anilines: The Tug-of-War between Pyramidalization and Resonance inside and Outside of Crystal Cavities. *Comput. Theor. Chem.* **2007**, 813, 21-27.

34. Szostak, R.; Aube, J.; Szostak, M., An Efficient Computational Model to Predict Protonation at the Amide Nitrogen and Reactivity Along the C-N Rotational Pathway. *Chem. Commun.* **2015**, 51, 6395-6398.

35. Fischer, G., Chemical Aspects of Peptide Bond Isomerisation. *Chem. Soc. Rev.* **2000**, 29, 119-127.

36. Romanelli, A.; Shekhtman, A.; Cowburn, D.; Muir, T. W., Semisynthesis of a Segmental Isotopically Labeled Protein Splicing Precursor: Nmr Evidence for an Unusual Peptide Bond at the N-Extein-Intein Junction. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, 101, 6397-6402.

37. Birnbaum, J.; Kahan, F. M.; Kropp, H.; Macdonald, J. S., Carbapenems, a New Class of Beta-Lactam Antibiotics: Discovery and Development of Imipenem/Cilastatin. *Am. J. Med* **1985**, 78, 3-21.

38. Wiberg, K. B.; Laidig, K. E., Barriers to Rotation Adjacent to Double Bonds. 3. The C-O Barrier in Formic Acid, Methyl Formate, Acetic Acid, and Methyl Acetate. The Origin of Ester and Amide Resonance. *J. Am. Chem. Soc.* **1987**.

39. Luque, F. J.; Orozco, M., Theoretical-Study of N-Methylacetamide in Vacuum and Aqueous-Solution - Implications for the Peptide-Bond Isomerization. *J. Org. Chem.* **1993**, 58, 6397-6405.

40. Lauvergnat, D.; Hiberty, P. C., Role of Conjugation in the Stabilities and Rotational Barriers of Formamide and Thioformamide. An Ab Initio Valence-Bond Study. *J. Am. Chem. Soc.* **1997**, 119, 9478-9482.

41. Thakkar, B. S.; Svendsen, J. S. M.; Engh, R. A., Cis/Trans Isomerization in Secondary Amides: Reaction Paths, Nitrogen Inversion, and Relevance to Peptidic Systems. *J. Phys. Chem. A* **2017**, 121, 6830-6837.

42. Lopez, X.; Mujika, J. I.; Blackburn, G. M.; Karplus, M., Alkaline Hydrolysis of Amide Bonds: Effect of Bond Twist and Nitrogen Pyramidalization. *J. Phys. Chem. A* **2003**, 107, 2304-2315.

43. Tomic, A.; Kovacevic, B.; Tomic, S., Concerted Nitrogen Inversion and Hydrogen Bonding to Glu451 Are Responsible for Protein-Controlled Suppression of the Reverse Reaction in Human Dpp Iii. *Phys. Chem. Chem. Phys.* **2016**, 18, 27245-27256.

44. Šebera, J.; Trantírek, L.; Tanaka, Y.; Sychrovský, V., Pyramidalization of the Glycosidic Nitrogen Provides the Way for Efficient Cleavage of the N-Glycosidic Bond of 8-Oxog with the Hogg1 DNA Repair Protein. *J. Phys. Chem. B* **2012**, 116, 12535-12544.

45. Sychrovsky, V.; Foldynova-Trantirkova, S.; Spackova, N.; Robeyns, K.; Van Meervelt, L.; Blankenfeldt, W.; Vokacova, Z.; Sponer, J.; Trantirek, L., Revisiting the Planarity of Nucleic Acid Bases: Pyramidilization at Glycosidic Nitrogen in Purine Bases Is Modulated by Orientation of Glycosidic Torsion. *Nucleic Acids Res.* **2009**, 37, 7321-7331.

46. Halgren, T. A., Merck Molecular Force Field .3. Molecular Geometries and Vibrational Frequencies for Mmff94. *J. Comput. Chem.* **1996**, 17, 553-586.

47. Halgren, T. A., Mmff Vii. Characterization of Mmff94, Mmff94s, and Other Widely Available Force Fields for Conformational Energies and for Intermolecular-Interaction Energies and Geometries. *J. Comput. Chem.* **1999**, 20, 730-748.

48. Kaminski, G.; Jorgensen, W. L., Performance of the Amber94, Mmff94, and Opls-Aa Force Fields for Modeling Organic Liquids. *J. Phys. Chem.* **1996**, 100, 18010-18013.

49. Mannfors, B. E.; Mirkin, N. G.; Palmo, K.; Krimm, S., Analysis of the Pyramidalization of the Peptide Group Nitrogen: Implications for Molecular Mechanics Energy Functions. *J. Phys. Chem. A* **2003**, 107, 1825-1832.

50. Mobley, D.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K., Open Force Field Consortium: Escaping Atom Types Using Direct Chemical Perception with Smirnoff V0.1. *bioRxiv* **2018**, 286542.

51. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., Ff14sb: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99sb. *J. Chem. Theory Comput.* **2015**, 11, 3696-3713.

52. Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D., Optimization of the Additive Charmm All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, Ψ and Side-Chain X1 and X2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, 8, 3257-3273.

53. Mackerell Jr, A. D.; Feig, M.; Brooks Iii, C. L., Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations. *J. Comput. Chem.* **2004**, 25, 1400-1415.

54. Robertson, M. J.; Qian, Y.; Robinson, M. C.; Tirado-Rives, J.; Jorgensen, W. L., Development and Testing of the Opls-Aa/M Force Field for Rna. *J. Chem. Theory Comput.* **2019**, 15, 2734-2742.

55. Roos, K.; Wu, C. J.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L. L.; Abel, R.; Friesner, R. A.; Harder, E. D., Opls3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J. Chem. Theory Comput.* **2019**, 15, 1863-1874.

56. Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A., General Atomic and Molecular Electronic-Structure System. *J. Comput. Chem.* **1993**, 14, 1347-1363.

57. Gordon, M. S.; Schmidt, M. W., Advances in Electronic Structure Theory: Gamess a Decade Later. *Theory and Applications of Computational Chemistry: The First Forty Years* **2005**, 1167-1189.

58. Mantz, Y. A.; Branduardi, D.; Bussi, G.; Parrinello, M., Ensemble of Transition State Structures for the Cis-Trans Isomerization of N-Methylacetamide. *J. Phys. Chem. B* **2009**, 113, 12521-12529.

59. Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L., Improved Peptide and Protein Torsional Energetics with the Opls-Aa Force Field. *J. Chem. Theory Comput.* **2015**, 11, 3499-3509.

60. Corbeil, C. R.; Thielges, S.; Schwartzentruber, J. A.; Moitessier, N., Toward a Computational Tool Predicting the Stereochemical Outcome of Asymmetric Reactions: Development and Application of a Rapid and Accurate Program Based on Organic Principles. *Angew. Chem* **2008**, 47, 2635-2638.

61. Weill, N.; Corbeil, C. R.; De Schutter, J. W.; Moitessier, N., Toward a Computational Tool Predicting the Stereochemical Outcome of Asymmetric Reactions: Development of the Molecular Mechanics-Based Program Ace and Application to Asymmetric Epoxidation Reactions. *J. Comput. Chem.* **2011**, 32, 2878-2889.

62. Liu, Z. M.; Pottel, J.; Shahamat, M.; Tomberg, A.; Labute, P.; Moitessier, N., Elucidating Hyperconjugation from Electronegativity to Predict Drug Conformational Energy in a High Throughput Manner. *J. Chem. Inf. Model.* **2016**, 56, 788-801.

63. Glendening, E. D.; Landis, C. R.; Weinhold, F., Nbo 6.0: Natural Bond Orbital Analysis Program. *J. Comput. Chem.* **2013**, 34, 1429-1437.

4 Conclusions and Future Work

4.1 Conclusions

To summarize, virtually all drug discovery (DD) endeavours rely at least at one point on methods utilizing molecular mechanics (MM) potentials to describe molecular systems.[1] The most common applications, molecular dynamics and docking, both rely on accurate force fields (FFs) in order to provide valuable information regarding the dynamics of large biomolecules, and binding affinities between potential ligands and specific biological targets. Much work has been directed towards the improvement of FFs, and the continuous update and refinement of potentials and parameters are expected to further strengthen the reliability of aforementioned predictions, ultimately reducing the costs to bring new pharmaceutical compounds to the market.[2] Most of these efforts have been guided towards the treatment of electrostatic interactions via polarizable models, and the parametrization of torsions to cover drug-like molecule space.[3]

First, we have carried the philosophy of H-TEQ, in which hyperconjugation interactions replace the torsional energy term, to unsaturated molecules (e.g. allylic, benzylic). As opposed to previous studies focusing on saturated molecules,[4, 5] hyperconjugation was no longer the predominant factor determining conformational preference, hence non-bonded interactions (van der Waals, electrostatics) must be accurately transcribed, to reproduce potential energy surfaces (PES) well. We have shown that our method performed on par with GAFF2,[6] one of the most commonly used FFs for small organic molecules, without relying on atom types to assign parameters, which are associated to many known drawbacks.[7] We believe that associating our H-TEQ3.0 equations with polarizable methods such as AMOEBA[8] could lead to better accuracies, and a stronger transferability of parameters which is essential to cover greater portions of drug-like molecule space.

Additionally, we have discussed current drawbacks in GAFF2 limiting its ability to reproduce the correct PES and preferred hybridization states of molecules containing nitrogen atoms bound to π-systems. In this context, we found that we could not rely on NBO[9] to quantify

(hyper)conjugation interactions and discussed difficulties which hindered our ability to carry the H-TEQ methodology towards these molecules. More specifically, our inability to obtain theoretical quantum data to support the explicit inclusion of lone pairs (LPs) with physically sound parameters, suggested that for now, empirical parameters were needed to describe these systems. However, we have determined and classified ways in which GAFF2 led to incorrect predictions, proposed a novel cross-term which could support the change of hybridization of nitrogen centers as bonds are rotated, and shown that corrections to the angle parameters could fix current deficiencies. Altogether, incorrect torsional and angle parameters need to be readjusted, using more suitable methodologies which need to be developed. In that respect, we believe that sampling larger portions of conformational space as we have performed (i.e. 2-D maps) could be more adequate. Once this has been completed, the necessity of the cross-term proposed can be reassessed for which more physically meaningful parameters should be derivable, as the latter would not be empirically correctly for other sources of error.

4.2 Future Work and Perspectives

Overall, more work is required to expand the current coverage of H-TEQ such that it could support all possible drug-like molecules. Work in our lab to include biaryl systems has already been completed, but other torsion types such as those studied in chapter 3, or those involving other kinds of conjugated systems (e.g. aryl-amides) are still missing. The development of more accurate energy decomposition analysis (EDA[10]) tools by theoretical chemists could allow to carry the H-TEQ methodology to molecules discussed in chapter 3. Further, once the H-TEQ method completely covers drug-like chemical space, its equations could be retrained on more rigorously decomposed quantum data, ultimately improving its accuracy. As of now, H-TEQ has been tested by retaining every other (non-torsion) term as found in GAFF currently (MMFF94 and parm@Frosst were also used in our first two studies on saturated compounds).[4, 5] We are thus curious to see how H-TEQ fares when using polarizable methods such as AMOEBA[8] or the Drude-2013[11] FF to calculate other terms (i.e. bonds, angles, vdW, electrostatics). This will be made possible once automated atom typers and parameter assignment protocols are made publicly available. We expect our method to be more accurate in conjunction with polarizable FFs, as current torsional parameters in non-polarizable methods are currently expected to empirically

make up for errors stemming from non-bonded terms, which H-TEQ is fundamentally not designed to take care of.

Additionally, we envision that employing atom type free methodologies such as H-TEQ would be particularly suited to develop more transferable van der Waals parameters, which is also the direction followed by *Mobley et al.* with SMIRNOFF.[7] Technically, an H-TEQ like methodology could be employed to generate bond and angle parameters, although following that route might not be the best allocation of resources as those terms are generally assumed to be well covered. We note that for the sake of compatibility with SMIRNOFF, *Mobley et al.* have planned to reparametrize bonds and angle terms entirely. Finally, H-TEQ should be further validated by carrying concrete applications in the context of SBDD. For example, docking libraries of compounds, and comparing results with other FFs. Or by comparing FEP predicted protein-ligand binding affinities to experimental values. Ultimately, results from various applications will point towards what future rounds of refinement should consist of.

References

1. Jorgensen, W. L., The Many Roles of Computation in Drug Discovery. *Science* **2004**, 303, 1813-1818.

2. Nerenberg, P. S.; Head-Gordon, T., New Developments in Force Fields for Biomolecular Simulations. *Curr. Opin. Struc. Biol.* **2018**, 49, 129-138.

3. Riniker, S., Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. *J. Chem. Inf. Model.* **2018**, 58, 565-578.

4. Liu, Z. M.; Pottel, J.; Shahamat, M.; Tomberg, A.; Labute, P.; Moitessier, N., Elucidating Hyperconjugation from Electronegativity to Predict Drug Conformational Energy in a High Throughput Manner. *J. Chem. Inf. Model.* **2016**, 56, 788-801.

5. Liu, Z. M.; Barigye, S. J.; Shahamat, M.; Labute, P.; Moitessier, N., Atom Types Independent Molecular Mechanics Method for Predicting the Conformational Energy of Small Molecules. *J. Chem. Inf. Model.* **2018**, 58, 194-205.

6. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, 25, 1157-1174.

7. Mobley, D.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K., Open Force Field Consortium: Escaping Atom Types Using Direct Chemical Perception with Smirnoff V0.1. *bioRxiv* **2018**, 286542.

8. Mohamed, N. A.; Bradshaw, R. T.; Essex, J. W., Evaluation of Solvation Free Energies for Small Molecules with the Amoeba Polarizable Force Field. *J. Comput. Chem.* **2016**, 37, 2749-2758.

9. Glendening, E. D.; Landis, C. R.; Weinhold, F., Nbo 6.0: Natural Bond Orbital Analysis Program. *J. Comput. Chem.* **2013**, 34, 1429-1437.

10. von Hopffgarten, M.; Frenking, G., Energy Decomposition Analysis. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, 2, 43-62.

11. Lopes, P. E. M.; Huang, J.; Shim, J.; Luo, Y.; Li, H.; Roux, B.; MacKerell, A. D., Polarizable Force Field for Peptides and Proteins Based on the Classical Drude Oscillator. *J. Chem. Theory Comput.* **2013**, 9, 5430-5449.

Appendix

Appendix 1. Scaling Factors Applied to NBO

When replacing GAFF2 torsional energies by NBO calculated hyperconjugations, we noticed that NBO significantly overestimated energies for all kinds of interactions (Table A1), leading to a high RMSE of 2.01 kcal/mol (computed between NBO predicted profiles and QM profiles) when no scaling factor was applied. Removing the $\sigma \rightarrow \sigma^*$ and comparing only π-hyperconjugation to QM led to a slightly lower RMSE of 1.87 kcal/mol which could erroneously lead us to think $\sigma \rightarrow \sigma^*$ could be neglected to achieve more faithful predictions. When relevant energies were scaled down however, it became apparent that both hyperconjugation and π-hyperconjugation needed to be considered to correctly predict torsional profiles. Overall, with scaling factors of 0.4 for hyperconjugation and 0.25 for π-hyperconjugation, the RMSE for NBO energies substituting the torsional energy was significantly better (0.55 kcal/mol) than using pre-existing torsional parameters within GAFF2 (0.84 kcal/mol). We have used the scaling factors which we found to minimize the RMSE between QM and NBO profiles, having tried various combinations (scaling factors tested from 0 to 1.5, with 0.05 increments).

Table A1. Accuracy of GAFF2 and NBO to reproduce the torsional profiles of the 98 molecules in the development set.

Method compared to MP2/6-311+G**	**Average RMSE (kcal/mol)**
GAFF2	0.84
NBO [a]	2.01
NBO (π-Hyperconjugation only) [a]	1.87
NBO + scaling [a]	0.55
NBO (π-Hyperconjugation only) + scaling [a]	0.75

[a] The torsional term from GAFF2 was replaced by this method, all the other terms were kept.

Appendix 2. Alternative RMSE Calculations

To test the impact of the RMSE equation, and ensure that values obtained could be trusted, we performed RMSE calculations using two additional schemes, which did not reveal any significant difference than the traditional RMSE equation. In the scheme with a cut-off, point in which the QM energy is greater than 10 kcal/mol were not included in the RMSE calculation. In the scheme with Boltzmann weights, the contribution from every point is scaled down by a Boltzmann factor based on the quantum energy (Eq. 3.1).

Table A2. RMSE obtained using different schemes.

Method compared to MP2/6-311+G and set of molecules used**	**Average RMSE (kcal/mol)**	**Average RMSE with cut-off (kcal/mol)**	**Average RMSE with Boltzmann weight (kcal/mol)**
GAFF // development set	0.84	0.84	0.67
H-TEQ 3.0 // development set	0.80	0.80	0.65
GAFF // validation set	1.69	1.67	1.23
H-TEQ 3.0 // validation set	1.71	1.68	1.24

Appendix 3. Outlier in the Validation Set.

One of the molecules in our validation set showed a very large RMSE of > 20 kcal/mol when comparing GAFF2 or H-TEQ to its QM energy profile. A closer look revealed that as we rotated the torsion shown in red (Figure A1), a reorganization of other parts of the molecule occurs. More specifically, the other torsion (shown in blue) rotated leading to weaker conjugation with the neighboring double bond, and the current parameters in GAFF2 strongly penalizes this loss of conjugation whereas QM calculation predicts it to be much weaker.

While our efforts are to describe the rotation around the red torsion (for which we obtained a torsional profile), in this specific case a major rearrangement of other parts of the molecule (blue torsion) were predominant. Since we only replace the parameters for the torsion shown in red with H-TEQ derived values, it is not surprising that the high RMSE is maintained when comparing H-TEQ and QM. The overexaggerated penalty for the loss of conjugation currently predicted by GAFF2 is a major issue we have diagnosed, and research in our group is currently undertaken to incorporate conjugated ($\pi \rightarrow \pi^*$) systems into our H-TEQ method.

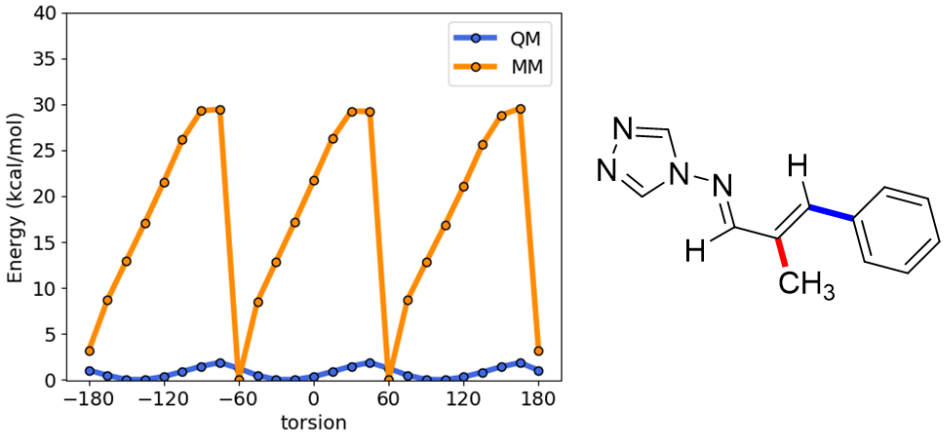


Figure A1. Torsional profile of the outlier found in our validation set. Torsion angle rotated is shown in red. Torsion angle highlighted in blue rearranges leading to the large energy barriers in MM.

Appendix 4. Validation Set Used in our Study.

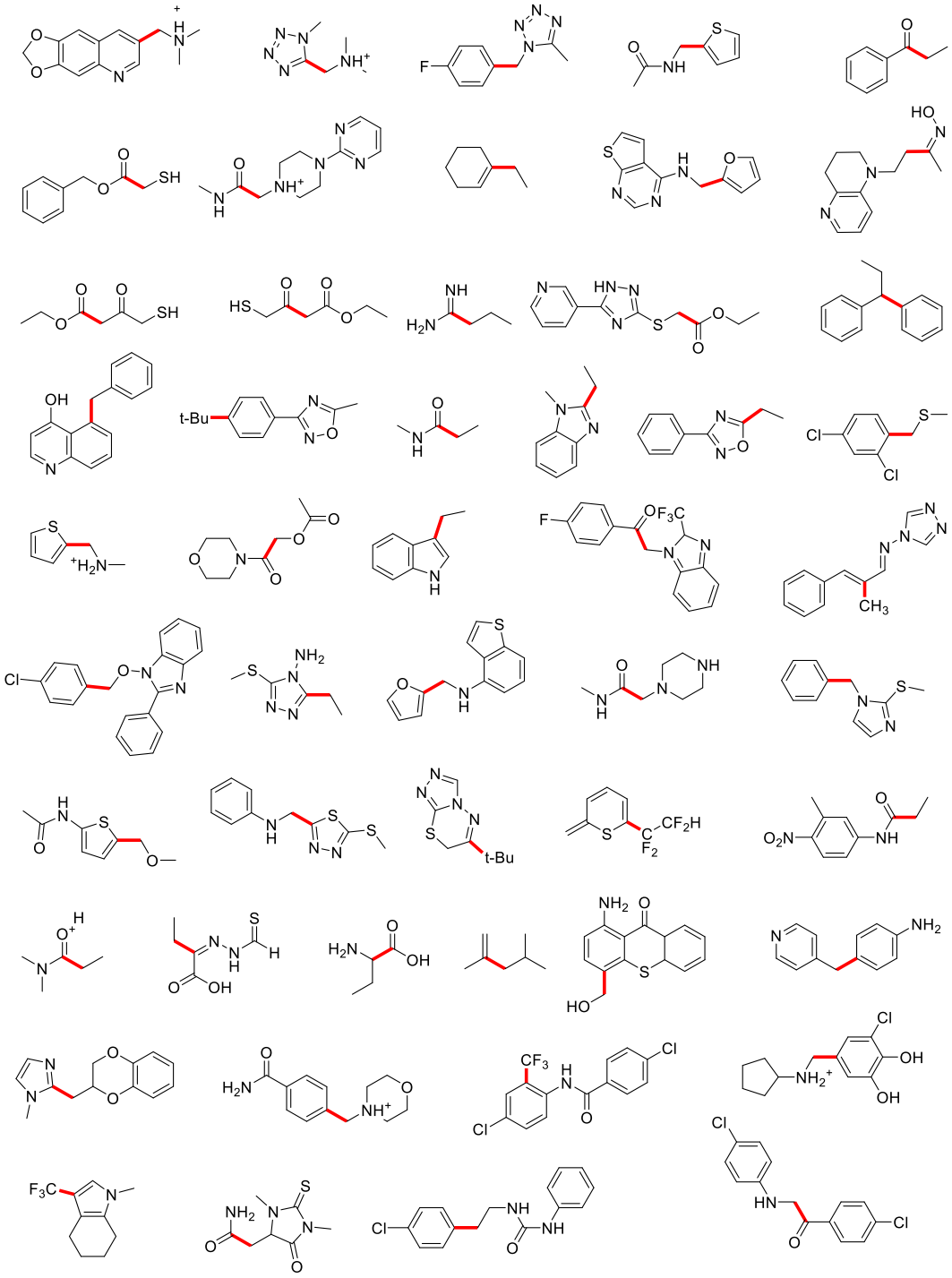


Figure A2. Set of 50 molecules used to validate the ability of H-TEQ 3.0 to describe the torsional energy of drug-like molecules. The rotated bond is shown in red.

Appendix 5. Effect of the V_3 Term on the π-Hyperconjugation Energy Profile.

The effect of changing the strengths of the V_3 correction term on the π-hyperconjugation profile was demonstrated in Figure A3. Each profile shown included a non-zero value for V_2 (here V_2 = 2.5 kcal/mol). The effect of a positive and negative V_3 value on the overall profile was probed by varying V_3 sequentially to be -0.5, 0, and 0.5 kcal/mol. As a result of the introduction of a non-zero V_3 value, the energy profile near the minima (±90°) was shifted (~5°) outwards (±180°) for negative values and inwards for positive values. The impact of V_3 also led to an asymmetry of the two energy barriers. A value for V_1 of the same magnitude could be used to correct for this asymmetry, if needed. In our current study, we chose to keep only V_3, since a counterbalancing V_1 term did not significantly improve the RMSEs with respect to the QM profiles (Table A3).

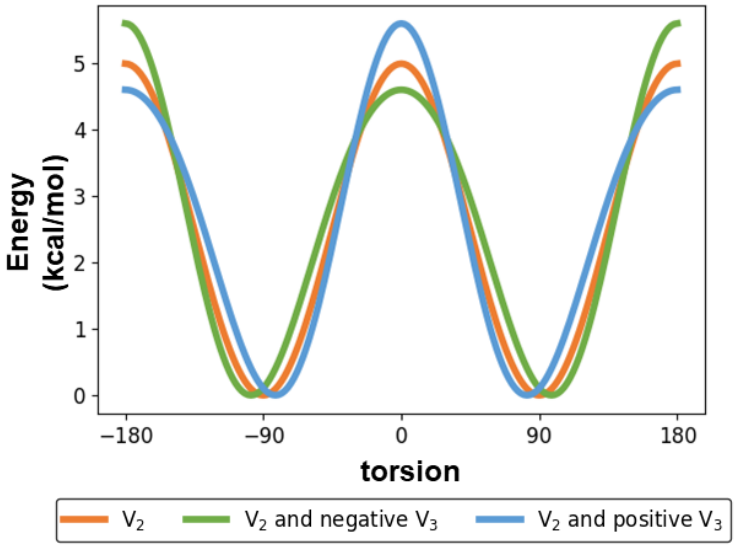


Figure A3. Comparison of π-hyperconjugation energy profiles using no V_3 (orange), a positive V_3 (blue) and a negative V_3 (green).

Table A3. Impact of V_3 and V_1 terms on the RMSEs (development set of 98 molecules), with σ → σ* hyperconjugation also included.

Method compared to MP2/6-311+G and set of molecules used**	**Average RMSE (kcal/mol)**
GAFF // development set	0.84
H-TEQ3.0 // development set	
Inclusion of V_3	0.80
Inclusion of V_3 and V_1	0.80

Appendix 6. Pauling Electronegativity Values used to Determine H-TEQ 3.0 Parameters.

Table A4. Pauling Electronegativity values of common elements.

Element	**Electronegativity (a.u.)**
H	2.2
B	2.04
C	2.55
Si	1.90
N	3.04
P	2.19
O	3.44
S	2.58
F	3.98
Cl	3.16
Br	2.96

Appendix 7. Group Electronegativity

Our group electronegativity assignment scheme not only considered the *central* atom, but also covalently bound neighboring atoms. The central atoms were accounted for with a greater weight ω (Eq. 2.4). The electronegativities of the π-system ($\chi^{\pi 1}$ and $\chi^{\pi 2}$) were calculated by considering both sides of the ring, and the convention was kept for less aromatic molecules (or non-aromatic) such as 5-membered rings or conjugated chains (Figure 2.8). This could be rationalized by inductive effects of the groups attached to the π-system (without forming a double bond); strongly electron-withdrawing groups (EWG) will increase the propensity of σ → π*, and strongly electron-donating groups (EDG) will increase the propensity of π → σ*.

Appendix 8. Scaling factors to Describe π-Hyperconjugation with a Unique Equation.

The left panels in Figures A6 shows the trends found using Eqs. 2.2 and 2.3. The trends for Eq. 2.2 applied to different types of π-systems to model σ → π* correlated well with NBO data (r^2 = 0.71, 0.69 and 0.76 for double bonds, 5-membered rings and 6-membered rings respectively). A minor scaling factor of 0.95 was applied to 5-membered rings leading to a slight increase in correlation for the entire set of molecules (from r^2 = 0.69 to 0.71).

The difference between trends of various types of π-systems was more pronounced for π → σ* hyperconjugation (Eq. 2.3). To correct for these differences, a scaling factor of 1.2 was applied for the double bonds and the same scaling factor of 0.95 was applied to 5-membered rings. Again, had we opted to use separate equations for each type of π-systems, the trends found revealed good correlations (r^2 = 0.70, 0.82, 0.79 for double bonds, 5-membered and 6-membered rings respectively). The right panels in Figure A4 correspond to plots shown in Figure 2.9 of the thesis. The scaling factors abovementioned are used to scale the $\chi^{\pi 1}$ and $\chi^{\pi 2}$ variables in Eqs. 2.2 and 2.3.

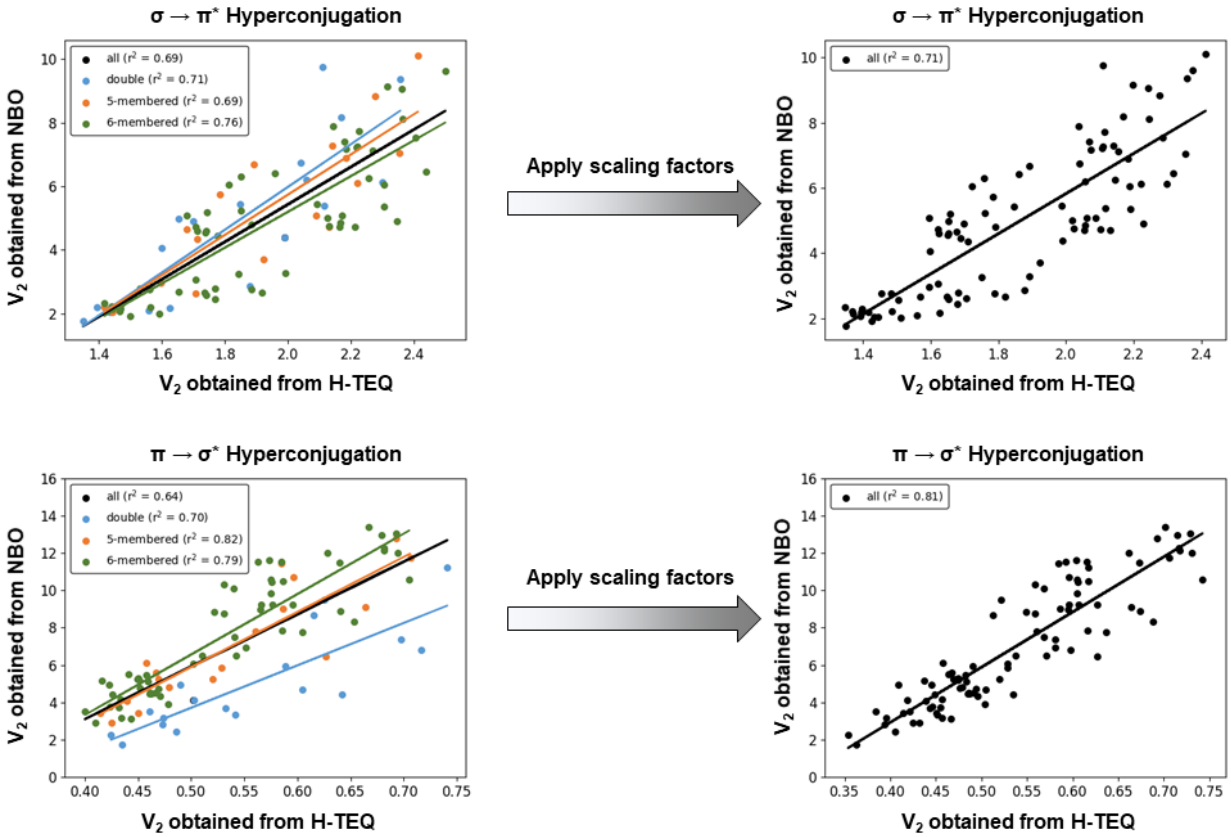


Figure A4. Application of scaling factors to equations used in the modeling of π-hyperconjugation

Appendix 9. Transferability of the a, b, c and d Parameters.

To confirm the transferability of the parameters a, b, c and d obtained from the linear regression around NBO data. We have performed a bootstrapping analysis in which sets of data (of the same size) are randomly generated from the initial data set. In short, points (H-TEQ vs. NBO) are picked randomly to make a new data set which can contain duplicates (or triplicates, etc.) and hence not contain some of the initial points. This kind of analysis allows to control for the presence of outliers in the set. We then obtain new parameters with a linear regression around this new data set. We performed this step 1000 times (and 10000 times), to calculate the average value and standard deviation of these parameters (see Table A5). Overall, we observe that the average value is very close to the value obtained initially, and the standard deviation is small.

Table A5. Results of the bootstrapping analysis.

Parameter	**Initial value from trend (kcal/mol)**	**Bootstrapping 1k iterations average value ± standard deviation (kcal/mol)**	**Bootstrapping 10k iterations average value ± standard deviation (kcal/mol)**
a	6.16	6.14 ± 0.36	6.15 ± 0.37
b	-6.50	-6.47 ± 0.02	-6.48 ± 0.00
c	29.52	29.57 ± 1.39	29.60 ± 1.35
d	-8.88	-8.91 ± 0.04	-8.91 ± 0.01

Appendix 10. Performance of Different Versions of H-TEQ 3.0.

Below, we detail the full results obtained when including (or excluding) $\sigma \rightarrow \sigma^*$ and V_3 correcting terms into the H-TEQ 3.0 method, on both the validation and development sets of molecules.

Table A6. Accuracy of GAFF2 and H-TEQ3.0 over the development and validation sets, by toggling on/off $\sigma \rightarrow \sigma^*$ and the V_3 correction terms.

Method compared to MP2/6-311+G and set of molecules used**	**Average RMSE (kcal/mol)**
GAFF // development set	0.84
H-TEQ3.0 // development set	
Only $\sigma \rightarrow \sigma^*$ is included	0.90
Both $\sigma \rightarrow \sigma^*$ and V_3 are omitted	0.80
Both $\sigma \rightarrow \sigma^*$ and V_3 are included	0.80
Only V_3 is included	0.73
GAFF // validation set	1.69
H-TEQ3.0 // validation set	
Only $\sigma \rightarrow \sigma^*$ is included	1.63
Both $\sigma \rightarrow \sigma^*$ and V_3 are omitted	1.65
Both $\sigma \rightarrow \sigma^*$ and V_3 are included	1.71
Only V_3 is included	1.76