Accelerating Drug Discovery through Method Developments in Nucleic Acid Docking and Force Fields for Small Molecules

Wanlei Wei

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Chemistry,

McGill University

Montréal, Québec, Canada August 2020

I dedicate this work to my family.

Abstract

The field of computational chemistry has grown rapidly and is becoming increasingly more sophisticated and accurate over the past few years and decades. Owing to its multidisciplinary nature, it has been applied to a variety of different fields, including drug discovery and medicinal chemistry. Traditionally, drug discovery has been a costly and laborious task, requiring upwards of billions of dollars and decades for each approved pharmaceutical. As a result, increases in efficiency would greatly benefit humanity and the pharmaceutical industry for treating various diseases. To this end, molecular docking methods attempt to reduce the time and labour costs by evaluating the binding of small molecules to proteins or nucleic acid receptors *in silico*. This is done during the initial drug discovery stages, where billions of pharmaceuticals in a ligand library could be screened virtually, quickly, and inexpensively. Despite the imperfect accuracy of *in silico* docking, its usefulness arises from its ability to enrich compound libraries, allowing medicinal chemists to select drug candidates, which are most promising for synthesis and testing. This has been applied to a variety of medicinal chemistry projects involving protein targets.

Recently, it has been realized that DNAs and RNAs are excellent targets of pharmaceuticals, as evidenced by the discovery of two molecules, branaplam and ribocil. This has ushered in a new era for pharmaceutical companies, known as the "RNA gold rush". Unfortunately, biochemical differences between nucleic acids and proteins have made it challenging to accurately dock molecules to the former. One of these biochemical differences is attributed to the role of water molecules found and involved in the binding of ligand molecules to nucleic acids, which are rarely found in protein binding sites. Consequently, to improve the accuracy of nucleic acid docking, a method to predict the positions of these water molecules was developed. This method, SPLASH'EM, places water molecules based on a previous statistical survey of water hotspots in existing nucleic acid structures from the Protein Data Bank and also uses a specialized force field. To our knowledge, this method is the first fully automated procedure for water placement and is the only well-validated method to-date, for water placement for nucleic acids. SPLASH'EM was found to have achieved the highest accuracy to-date for predicting tightly-bound water molecules in nucleic acid-ligand complexes.

In a parallel effort, further efforts to improve the enrichment capabilities of docking programs was undertaken. This was done by improving the description of the conformational energy landscapes of ligands binding to proteins and nucleic acids. Traditionally, molecular mechanics force fields have been used in place of quantum mechanics during docking, due to their low computational costs, despite losses in accuracy. In particular, the torsional parameters in these molecular mechanics force fields were identified as being poor, as they relied upon a limited number of parameters, called atom-types, which were found to poorly transfer between different molecules. Consequently, a conceptually novel approach was taken to predict torsional parameters, based on chemical principles, without reliance on atom-types. This method, called H-TEQ, was found to significantly outperform the accuracy of existing methods, including GAFF2, MMFF94, and MAB on small druglike molecules. The adoption of H-TEQ to enhance the description of torsional parameters, along with better descriptions of van der Waals and electrostatic interactions, would allow enhancements to the ability of docking programs to correctly identify active compounds, during drug discovery projects.

Résumé

Le domaine de la chimie computationnelle s'est rapidement développé et est devenu de plus en plus sophistiqué et précis au fil des dernières années et décennies. Grâce à sa nature pluridisciplinaire, elle a été utilisée dans des domaines variés, tels que la découverte de médicaments et la chimie médicinale. Traditionnellement, la découverte de médicaments est une tache coûteuse et laborieuse, nécessitant des milliards de dollars et des décennies pour chaque médicament approuvé. En conséquence, une amélioration de l'efficacité bénéficierait à l'humanité et à l'industrie pharmaceutique dans le traitement de diverses maladies. Dans cette optique, les méthodes d'amarrage moléculaire tentent de réduire le temps et le travail nécessaires en évaluant la liaison de petites molécules à des protéines ou des acides nucléiques in silico. Ceci est réalisé lors des étapes initiales de découverte de médicaments, où des milliards de produits pharmaceutiques, dans une bibliothèque de ligands, sont passés au crible virtuellement, rapidement et de façon peu coûteuse. Malgré la précision imparfaite de l'amarrage in silico, son utilité vient de sa capacité à enrichir les bibliothèques de molécules, permettant aux chimistes médicinaux de choisir des molécules candidates plus prometteuses en termes de synthèse et de d'essai.

Il a récemment été découvert que l'ADN et l'ARN sont d'excellentes cibles pour les médicaments, comme le prouve la découverte de deux molécules : le branaplam et le ribocil. Ceci a marqué l'entrée des compagnies pharmaceutiques dans une nouvelle ère appelée la ' ruée vers l'or ARN'. Malheureusement, les différences biochimiques entre acides nucléiques et protéines ont rendu l'amarrage de molécules à ces dernières peu précis. Une des différences biochimiques a été attribuée au rôle des molécules d'eau impliquées dans la liaison de ligands aux acides nucléiques, qui sont rarement présentes aux sites de liaison des protéines. Par conséquent, afin d'améliorer la précision de l'amarrage aux acides nucléiques, une méthode a été développée afin de prédire la position de ces molécules d'eau. Cette méthode, SPLASH'EM, est une méthode basée sur un champ de force, et possède jusqu'à présent la meilleure précision.

Dans un effort parallèle, des efforts supplémentaires ont été réalisés afin d'améliorer l'enrichissement des capacités de programmes d'amarrage. Ceci a été réalisé en améliorant la description du paysage énergétique conformationnel des ligands liés aux protéines et aux acides nucléiques. Traditionnellement, les champs de force de mécanique moléculaire ont été utilisés à la

place de la mécanique quantique lors de l'amarrage, grâce à leur faible coût computationnel, malgré des pertes de précision. En particulier, les paramètres de torsion dans ces champs de force de mécanique moléculaire ont été identifiés comme étant médiocres, étant donné qu'ils s'appuient sur un nombre limité de paramètres, appelés type d'atomes, qui sont mal transférés d'une molécule à l'autre. Par conséquent, une nouvelle approche conceptuelle a été utilisée afin de prédire les paramètres de torsion, basé sur des principes chimiques, sans s'appuyer sur les types d'atomes. Cette méthode, appelée H-TEQ, a permis de surclasser la précision de méthodes existantes, telles que GAFF2, MMFF94 et MAB sur des petites molécules semblables à des médicaments. L'utilisation d'H-TEQ afin d'améliorer la description des paramètres de torsion, ainsi que des meilleures descriptions des interactions de Vander Waals et électrostatiques, permettrait d'améliorer l'enrichissement des capacités des programmes d'amarrage, lors de projets de découverte de médicaments.

Traduit de l'anglais par Quentin Gaydon

Acknowledgments

I would like to thank those closest to me, especially my family and loved ones for their continuous support throughout my PhD. Their constant care during these four year's time has prevented me from stubbornly overthinking scientific problems at hand and distracted me long enough to stop me from going mad.

I would especially like to thank Prof. Nicolas Moitessier, for recruiting me from my M.Sc. degree, and giving me one-of-a-kind experience during these past four years. With your encouragement as a mentor, I truly believe that I have widened, as well as deepened, my knowledge in drug discovery and chemical sciences. I would also like to thank my M.Sc. mentors Prof. James Gauld and Prof. Gerald Monard for their instilling in me a passion for computational chemistry. Their mentorship has served as way-finders in times of difficulty.

I would also like to thank Dr. Stephen J. Barigye and Candide Champion, who I collaborated with extensively over the years. I fondly remember our meeting at the local Presse Café on l'Avenue du Parc for fruitful and lively discussions. I also want to thank everyone else who was a part of our group during these years: Anne, Jessica, Jiaying, Juan, Julia, Leo, Mihai, Naëla, Sharon, Chris Sitko, Chris Wang, Caitlin, Anna, Paolo, Jerry, Sylvain, Kevin, and Michelle.

Table of Contents

1 Introduction to Molecular Mechanics Force Fields for Simulating Nucleic Acids and Applications to Drug Discovery and Dynamics 1
1.1 Preface 1
1.2 Introduction 2
1.3 Nucleic Acids Functions 2
1.4 Nucleic Acid Structure 3
1.5 Nucleic Acids as Pharmaceutical Targets 8
1.6 Use of Molecular Mechanics Force Fields in Computational Chemistry 11
1.7 Molecular Mechanics Force Fields for Modelling Nucleic Acids 16
1.8 Evolution of Modern Nucleic Acid Force Fields 20
1.9 State-of-the-Art DNA Force Fields 22
1.10 State-of-the-Art RNA Force Fields 23
1.11 Treatment of Hydrogen Bonding 24
1.12 Treatment of π-π Stacking 25
1.13 Treatment of Noncanonical Nucleic Acids and Ligands 27
1.14 Future Directions in FF Development and Need for Comprehensive Testing 27
1.15 Conclusion 28
1.16 References 29
2 Predicting Positions of Bridging Water Molecules in Nucleic Acid–Ligand Complexes 45
2.1 Preface 45
2.2 Introduction 46
2.3 Theory and Implementations 49
2.4 Results and Discussion 62
2.5 Conclusion 68
2.5 References 68
3 Torsional Energy Barriers of Biaryls could be Predicted by Electron-richness/deficiency of Aromatic Rings; Advancement of Molecular Mechanics toward Atom-Type Independence 75
3.1 Preface 75
3.2 Introduction 76
3.3 Understanding Chemical Origins 80
3.4 Computational Methods 85

3.5 Results and Discussion 88
3.6 Conclusion 102
3.7 References 104
4 Use of Extended-Hückel Descriptors for Rapid and Accurate Predictions of Conjugated Torsional Energy Barriers 111
4.1 Preface 111
4.2 Introduction 112
4.3 Computational Methods 119
4.4 Results and Discussion 123
4.5 Conclusions 133
4.6 References 134
5 Influence of Molecular Mechanics Torsional Parameters on Docking Accuracies 141
5.1 Preface 141
5.2 Introduction 142
5.3 Computational Methods 146
5.4 Results and Discussion 149
5.5 Conclusion 158
5.6 References 159
6 Conclusion and Closing Remarks 163
6.1 Conclusion 163
6.2 Closing Remarks 165
6.3 References 166
Appendix A: Supplementary Information for "Predicting Positions of Bridging Water Molecules in Nucleic Acid-Ligand Complexes" 168
A1 From Statistics to Binding Free Energy 168
A2 Hydrogen Bonding Potentials of Various Polar Atoms with Water 169
A3 Obtaining an Angle Term 169
A4 Removal of Solvent Exposed Waters 170
A5 Validation Dataset for Water Placement 171
A6 Initial Placement of Water Molecules 172
Appendix B: Supplementary Information for "Torsional Energy Barriers of Biaryls could be Predicted by Electron-richness/deficiency of Aromatic Rings; Advancement of Molecular Mechanics toward Atom-Type Independence" 174

B1 Overall π-electronegativity Modulates Strength of Conjugation 174
B2 Strength of Conjugation 174
B3 Substituent Effects on Torsional Energy Barriers of Biaryls 175
B4 Accuracy of H-TEQ 4.0 and GAFF2 Compared to QM Torsional Energy Profile: Training Set- Number of Occurrences vs. RMSE. 180
B5 Accuracy of H-TEQ 4.0 and GAFF2 Compared to QM Torsional Energy Profile: Validation Set- Number of Occurrences vs. RMSE. 182
Appendix C: Supplementary Information for "Use of Extended-Hückel Descriptors for Rapid and Accurate Predictions of Conjugated Torsional Energy Barriers" 185
C1 Results of Subunit π-orbital Analysis 185
C2 Qualitative Trends in V_2 202
C3 V_2 of QM vs. V_2 from Sum of All $\sigma \rightarrow \sigma^*$ hyperconjugation obtained from NBO 203
C4 Sample torsional profiles of GAFF2 vs. QM 204
C5 Classification of Errors for H-TEQ 4.5 205
C6 Accuracy of H-TEQ 4.5, GAFF2, MAB, and MMFF94 Compared to QM Torsional Energy 206
Appendix D: Supplementary Information for "Influence of Molecular Mechanics Torsional Parameters on Docking Accuracies" 221
D1 PDB IDs of Structures Used for Nucleic Acid Self-Docking 221
D2 PDB IDs of Structures Used for Protein Self-Docking 221

List of Figures

Figure 1.1. Canonical Watson-Crick (left) compared to Hoogsteen (right) base pairing. The guanine nucleotide is in the anti- and syn-conformation in the Watson-Crick and Hoogsteen base pairing, respectively. 5

Figure 1.2. Definition of various torsional angles found in nucleic acids. 6

Figure 1.3. All possible pseudorotational angles and selected conformations of the nucleic acid aldopentose sugar are shown. Red labels indicate conformations frequently found in naturally occurring nucleic acids. Superscripted and subscripted numbers preceding E or T indicate endo- and exo-positions, respectively. 7

Figure 1.4. A-, B-, and Z-forms of DNA are shown. Each type of helix has different geometrical properties, such as width, rise per base, and directionality of turn. RNA could also adopt these geometries. 8

Figure 1.5. Crystal structure of ribocil (ball and stick model) bound to the aptamer domain of FMN riboswitch (cartoon and wireframe model, PDB ID: 5C45). 9

Figure 1.6. Small molecular RNA inhibitors and pharmaceuticals are shown. A) Ribocil is an inhibitor of FMN riboswitch, which was found to disrupt bacterial riboflavin homeostasis. B) Branaplam is a promising drug candidate for spinal muscular atrophy. C) Tetracycline is an antibiotic known to bind bacterial rRNA. 10

Figure 1.7. Small molecule DNA inhibitors and pharmaceuticals are shown. A) Acyclovir is an essential drug, which causes early termination during viral replication in many diseases. B) [16]phenN_4 are promising G4-binders which could have antitumor abilities. C) Proflavine is a bacteriostatic compound, which was used for sterilization. D) 1,4-butanediol dimethanesulfonate is an akylating agent used during chemotherapy. 11

Figure 1.8. Various uses of MM in computational chemistry are shown. A) A hallmark investigation by Auffinger et al into RNA dynamics and solvation by classical MD simulations found that water molecules localized with high frequency in certain spatial regions (light blue).[93] B) An MD study into the equilibrium of A- and B-DNA in different solvation environments hinted at the possible role of cations and phosphates as the key driving force behind these transitions.[66] Higher concentration of water and ethanol induces A- and B-DNA, respectively. C) FEP study of purine riboswitches reproduced experimental binding free energy differences between a variety of ligands.[77] D) Molecular docking methods could be used to quickly and inexpensively screen for pharmaceutically promising drug candidates (right) against a receptor of interest (left) during the initial drug discovery process.[82] E) QM/MM simulations have been applied to investigate chemical reactions and/or dynamics. For example, the conformational dynamics of nucleosides could be investigated. The molecular region of interest (QM) and surrounding solvents (MM) are shown in ball-and-stick and wire representations, respectively. 15

Figure 1.9. Potential energy functions found in classical MM FFs. More advanced FFs also use these terms, in addition to more complex terms. 18

Figure 1.10. Newman projection of the torsional angles, α (left) and γ (right): A) as sampled by parm99 and B) found in nature and sampled by parmbsc0. 21

Figure 1.11. Timeline and evolution of modern type I nucleic acid FFs. Selected references are shown in yellow outlines. 22

Figure 1.12. Directionality of hydrogen bonding with respect to two water molecules are shown. Although water molecule at position B is expected to be more stable than A, most classical FFs assume same strengths of interaction due to neglect of directionality..... 25

Figure 1.13. π–π stacking interaction present in nucleic acids, which important for base stacking and ligand binding. Offset, rather than T- or Y-shaped π–π stacking, are most abundant in nucleic acids. In FFs, π–π stacking is not explicitly treated, but treated by a combination of electrostatic and van der Waals interactions. 26

Figure 2.1. Water molecule (red) interacting with U19, G27, U28 and Lividomycin, in the binding site of 16S-rRNA (PDB: 2ESJ[19]). 48

Figure 2. 2. Method flowchart designed for SPLASH'EM. 50

Figure 2.3. Statistics information collected on the occurrences of water molecules near: Adenine N1 (A), N3 (B), N6 (C), and N7 (D) at various angles (-90° to 90°) and distances (0 to 10 Å along the x-axis) within the plane of the base. Statistics were similarly collected outside the plane at different "pitch" angles. In each voxel, frequencies are indicated as a total percentage of waters collected. Statistics for other hydrogen bond acceptors/ donors at various pitch angles could be found in the supplementary information..... 51

Figure 2.4. Hydrogen bonding potential in QM (red) as compared to the developed PWFF (blue) for a linear distance scan for Ado-N1 (A) and Ado-H62 (B). The corresponding chemical structures are shown on the right, with distances labeled in dashed lines (magenta). 53

Figure 2.5. Model used for evaluating the energy at A) MP2/6-31+G (d,p) and B) MM, with united atom representation of water molecule. Arrows depict the path taken during the potential energy scan. C) Energy as a function of distance, depicting the enthalpic energy calculated in QM (red), van der Waals energy (blue), calculated using Amber ff95, and the difference of the former and latter (green), which is the hydrogen bonding potential. 54

Figure 2.6. Simplified subunits used to probe the relationship between enthalpy and angles. I) pyridine, II) aniline, III) tetrahydrofuran, and IV) 1-methylpyridin-4(1H)-one are shown. Polar atoms of interest are in red, while angle bisectors are depicted by dashed lines. 55

Figure 2.7. (A) Angular contribution function implemented into PWFF for various atom types. (B-E) Energy profile as a function of deviation from the angle bisector, obtained by E_{QM}-E_{VDW} for both in-plane and out-of-plane angles, with respect to the ring, as compared to the developed PW potential, (B) pyridine, (C) aniline, (D) tetrahydrofuran and (E) 1-methylpyridin-4(1H)-one are shown. For acceptors with more than one lone pair (ie. tetrahydrofuran and 1-methylpyridin-4(1H)-one), both in-plane and out-of-plane PW potential are shown. 57

Figure 2.8. A) H-Bond potentials between $CH_3NH_4^+$ and H_2O (in QM) when the latter is: not part of a bridge (green), in a bridged to $(CH_3)_2PO_4^-$(blue), and in a bridge to another $CH_3NH_4^+$ (red). B) Corresponding schematics are shown in (i-iii), respectively. 58

Figure 2.9. Crystallographic water position (green circle) compared to possible alternative water positions (blue circles) overlaid with the 2F0-Fc electron density map. EDIA score is given adjacent to each water. (A). Placed waters are supported by electron density but are situated away (distances of 1.6 and 1.7 Å) from the crystallographic water. (B). Placed waters are not supported by electron density but are in close proximity to the crystallographic water (distance of 1.0 Å). 63

Figure 2.10. 2'-Deoxyguanosine bound to each binding site of the dimeric 2'-Deoxyguanosine riboswitch (PDB: 3SKI). Shown in A) and B) are the two co-crystallized ligands in each binding site, and their associated 2F0-Fc electron densities. 64

Figure 2.11. Comparison of (A) (i) crystallographic water position in red (ii) to water molecules placed by PWFF in green and PWFFa in blue. Comparison of (B) (i) crystallographic position in red to (ii) to water molecules placed by PWFFa in blue and pPWFFa in orange. H-bonds are denoted by black lines while weak H-bonds are denoted by orange lines..... 66

Figure 3.1. Commercially available drug molecules that possess biaryl fragments, with the torsion bonds of interest labelled in red. 76

Figure 3.2. Two distinct conformations of fluoroethane: (a) $\theta_{H\text{-}C\text{-}C\text{-}F} = 180^\circ$ and (b) $\theta_{H\text{-}C\text{-}C\text{-}F} = 0^\circ$, are shown along with bonding (blue) and antibonding (red) orbitals participating in $\sigma_{C\text{-}H} \rightarrow \sigma^*_{C\text{-}F}$ hyperconjugation depicted. For simplicity, not all bonds nor orbitals are shown. 81

Figure 3.3. $\sigma \rightarrow \sigma^*$ hyperconjugation in a) ethane and b) fluoroethane with selected orbitals (bonding orbital is shown in blue, while antibonding orbital is shown in red) and energies shown. In both cases, the σ-bond orbital energy level is the same, while the σ*-antibonding orbital energy for fluoroethane is lower than ethane. This causes a greater amount of hyperconjugative stabilization for fluoroethane compared to ethane. 82

Figure 3.4. (a) Chemical structure of 2-(pyridin-2-yl)oxazole (with the torsional bond of interest highlighted in red) is shown with hyperconjugation of orbitals involving: (b and c) $\sigma \rightarrow \sigma^*$, (d) $\sigma \rightarrow \pi^*$, and (e) $\pi \rightarrow \sigma^*$, and (f and g) $\pi \rightarrow \pi^*$ conjugation. For simplicity, a maximum of one σ-, σ*-, π-, and π*-orbitals are shown in each subfigure. 83

Figure 3.5. A qualitative molecular orbital diagram of: a) biphenyl molecule, b) 2-phenylthiophene, and c) 2,2'-bithiophene are shown, along with a major interaction, involving $\psi 3 \rightarrow \psi 5^*$ conjugation. The right and left rings act as the donor and acceptor, respectively in this figure. For simplicity, not all possible transitions are shown. 84

Figure 3.6. 131 biaryl molecules used for obtaining the torsion energy within the development set. The number on the left of each scaffold indicates the number of molecules contained in that category. 86

Figure 3.7. 100 molecules of the compiled validation set. The bonds of interest are shown in red. 87

Figure 3.8. V_2 of 131 biaryl molecules were plotted against the total π-electron density ($\chi_{\pi tot}$), and categorized into six distinct groups, which differed by the atomic identity of the central bond. Smaller V_2 signifies stronger conjugation strength. The R^2 for each linear regression is also shown in the legends. 90

Figure 3.9. V_2 of 131 biaryl molecules were plotted against the squared difference of π-electron density between the two connected aromatic moieties $(\Delta\chi_\pi)^2$. Biaryls were separated into: A) neutral central bonds: C-C, C-N, and N-N, and B) positively charged central bonds: C-N+, N-N+, and N+-N+. C-N+ and N-N+ was further grouped by the identity of the central atom and number of atoms on each side of the central bond. Smaller V_2 signifies stronger conjugation strength. The R^2 for each linear regression is also shown in the legends. 91

Figure 3.10. Biaryls used to probe the effects of substituents on its torsional energy barrier. The torsional bonds of interest are depicted in red. 93

Figure 3.11. The torsional energy to rotation of biphenyl (red), and biphenyl substituted with strong electron-withdrawing and electron donating groups as computed by MP2/6-311++G(d,p) level of theory. 94

Figure 3.12. The energy terms extracted from a torsion scan of 2,2'-bipyridine. The red represents the sum of the GAFF2 van der Waals and electrostatics terms. The blue and green are V_1 and V_2, respectively, which were obtained from the torsion term as computed by $E_{QM} - (E_{vdw} + E_{ele})$. .. 96

Figure 3.13. A smoothed histogram indicates the number of MM energy profile predictions made by H-TEQ 4.0 (pink) and GAFF2 (blue) within the training set for 131 biaryls. The RMSEs of these predictions are shown on the x-axis. The QM energy profile was used as the reference, with values closer to 0 kcal·mol^{-1} representing more accurate predictions. 98

Figure 3.14. A smoothed histogram indicates the number of MM energy profile predictions made by H-TEQ 4.0 (pink) and GAFF2 (blue) within the validation set for 100 biaryls. The RMSEs of these predictions are shown on the x-axis. The QM energy profile was used as the reference, with values closer to 0 kcal·mol^{-1} representing more accurate predictions. 101

Figure 3.15. The torsion profiles of nine representative biaryl molecules (a-f). For each molecule, the torsion profiles as calculated by QM profile (red), HTEQ 4.0 (blue), and GAFF2 (green) are shown. The reference dihedral angle at 0˚ is marked by a series of four asterisks (*). 102

Figure 4.1. Commercially available drug molecules, which possess extensively conjugated moieties. The torsional bond of interest is highlighted in red..... 112

Figure 4.2. The torsional bond of interest (in red) of hexatriene (left) compared to that of 1,1-difluorohexa-1,3,5-triene (right) are shown. 115

Figure 4.3. A qualitative depiction of hyperconjugation of involving a) $\sigma_{C\text{-}H} \rightarrow \sigma^*_{C\text{-}H}$ and b) $\sigma_{C\text{-}H} \rightarrow \sigma^*_{C\text{-}F}$. The more electronegative fluorine has an overall lower lying σ*-antibonding orbital than hydrogen. Consequently, b) is stabilized to a greater degree than a). 117

Figure 4.4. Hyperconjugation involving n → σ*C-OH interaction in the two ring conformations of D-glycopyranose. In a), the hydroxyl antibonding orbital is not properly positioned to accept the electron density donation from the lone pair. In b), the hydroxyl group is axial, leading to significant orbital overlap between the lone pair electrons on the ring oxygen and the hydroxyl oxygen antibonding orbital. This gives noticeable stabilization in energy. For simplicity, only the atoms of interest are shown..... 118

Figure 4.5. Hyperconjugation involving $\sigma_{C\text{-}H} \rightarrow \sigma^*_{C\text{-}F}$ in 1,2-diflouroethane is shown at two different torsion angles. In a), there is very little overlap between the two orbitals, leading to weaker overlap, while in b), there is significant orbital overlap, leading to noticeable stabilization in energy. For simplicity, only the atoms of interest are shown. 118

Figure 4.6. (a) Chemical structure of oxazole-2-carbaldehyde (with the torsional bond of interest highlighted in red) is shown with hyperconjugation of orbitals involving: (b and c) σ → σ*, (d) σ →π*, and (e) π → σ *, and (f and g) π → π* conjugation. For simplicity, a maximum of one σ-, σ*-, π-, and π*-molecular orbitals are shown in each subfigure. h) A qualitative molecular orbital diagram of oxazole-2-carbaldehyde is shown, along with the major conjugation interaction. In this case, the conjugation interaction involving the smallest energy overlap is depicted. 119

Figure 4.7. Various Molecules included in the development set, categorized into three distinct groups: acyclic, aryl, and biaryl conjugated molecules. The torsional bond of interest is highlighted in red. 121

Figure 4.8. One hundred new molecules of the compiled validation set. Only molecules containing conjugated aryl and linear moieties are shown. The torsional bonds of interest are shown in red. 122

Figure 4.9. Schematic showing the subunit π-orbital analysis process. The conjugated molecule is separated at the torsional bond of interest and capped with hydrogen atoms. Subsequently, the π-MOs of both subunits are compared to determine the donor and acceptor side of the original molecule. 123

Figure 4.10. a) V_2 of 219 unsubstituted, conjugated aryls plotted against the number of electron lone pairs on the β-atom with respect to the ring. b) V_2 of 13 unsubstituted conjugated acyclic molecules, plotted against the absolute difference of electron lone pairs on the terminal position (ie. α- and α'-atom to the central atoms). A representative chemical structure is also depicted, with the labels for each atomic position of interest. 124

Figure 4.11. (a-c) Observed trends between various Extended Hückel Theory properties and V_2 strengths of the difference between QM and nonbonded interactions obtained from GAFF (similar trends and accuracies were obtained with MAB and MMFF94). In each subplot, V_2 is plotted against a) π-bond order, b) sum of atomic π-charges and c) product of atomic π-charges of the central atoms. In total, 684 various conjugated molecules were taken into consideration. d) Correlation between the predicted and measured V_2, based on these descriptors (a-c). 125

Figure 4.12. a) The donation of electron density of various oxyanions into the central atoms decreased their π-charges. b) The donation of the nitrogen electron lone pair away from the central nitrogen into the ring increased its net π-charge. The torsion of interest is in red..... 127

Figure 4.13. a) A smoothed histogram and b) a boxplot indicating the number of MM energy profile predictions made by H-TEQ 4.5 (pink), MMFF94 (blue), MAB (green), and GAFF2 (grey) for the 200 molecules in the validation set. The QM energy profile was used as the reference, with RMSE values closer to 0.0 kcal·mol^{-1} representing more accurate predictions. 130

Figure 4.14. Nine representative structures and their corresponding torsional energy scans as calculated by MP2/6-311+G** (red), H-TEQ 4.5 (blue), and MMFF94 (green) for: a) 1-chloro-4-(3,3-dimethylcycloprop-1-en-1-yl)benzene, b) methyl benzoate, c) (R)-3-(4-(4-methyl-6-oxo-1,4,5,6-tetrahydropyridazin-3-yl)phenyl)-1H-imidazol-3-ium, d) [2,2'-bipyridine]-1,1'-diium, e) 3-phenyl-1,2,5-oxadiazole 2-oxide, f) N-((1S,2S)-2-chlorocyclopropyl)benzamide, g) 4-nitro-1,2,5-oxadiazol-3-amine, h) amino(5-amino-1H-1,2,4-triazol-3-yl)methaniminium, and i) 2-amino-2-thioxoethane(dithioperoxo)imidic acid. The reference torsional bond of interest is marked in red. The reference angle at 0° is marked by a series of four asterisks (*). 132

Figure 5.1. Hyperconjugation interactions are responsible for the observed conformational preferences in small molecules. A) The $\sigma_H \rightarrow \sigma_F^*$ hyperconjugation causes the gauche effect in 1,2-difluoroethane. B) Similarly, $\eta \rightarrow \sigma_O^*$ causes the anomeric effect in (R)-tetrahydro-2H-pyran-2-ol. In both cases, donation of electron density from occupied (red) to unoccupied (blue) orbitals occur..... 145

Figure 5.2. 51 drug molecules were subjected to conformational energy analysis. The 50 drugs above and 1 drug below the dotted line represents those with and without steric clashes in MM, respectively. 149

Figure 5.3. The accuracy of protein self-docking on 289 protein-ligand complexes were performed with GAFF (red), $GAFF_{Tor:\ H\text{-}TEQ}$ (blue), $GAFF_{Tor:0}$ (green), and $GAFF_{Tor:0\ ELE:0}$ (orange) shown. 151

Figure 5.4. Examples of protein binders which were more accurately docked using: GAFF and $GAFF_{Tor:H\text{-}TEQ}$ (a-c) and $GAFF_{Tor:0}$ (d-f). 152

Figure 5.5. The accuracy of nucleic acid self-docking on 178 nucleic acid-ligand complexes were performed with GAFF (red), $GAFF_{Tor:\ H\text{-}TEQ}$ (blue), $GAFF_{Tor:0}$ (green), and $GAFF_{Tor:0\ Ele:0}$ (orange) shown. 153

Figure 5.6. A conformation of fluconazole, which registered a steric clash using GAFF2 LJ 12-6 potential, but not QM and MMFF94 phased LJ 14-7 potential. 154

Figure 5.7. Correlation of MAE between GAFF2 VDW and various other methods. The squared correlation coefficient (R^2) are also shown..... 155

Figure A1. Hydrogen bonding potential in QM (blue) as compared to the developed MM FF (red) for a linear distance scan along the angle bisector of Ado-N1(A) and Ado-H62 (B). The MM potential is sum of the 6-5-3 Lennard-Jones hydrogen bonding potential and van der Waals 12-6

Lennard-Jones potential. Functional form of the developed 6-5-3 Lennard-Jones potential for sp^2 aromatic nitrogen (C) and amine hydrogen (D) used to describe hydrogen bonding potential between polar atoms and water. 169

Figure A2. Mapping of bulk waters by rolling a spherical probe along the surface of the nucleic acid. Placed waters in the light green region were determined to be bulk solvents, and consequently removed..... 171

Figure A3. Testing set for water placement. 171

Figure A4. The top *N* most frequently populated position(s) around each individual polar atom of nucleic acid and ligand was used to place waters. This initial population of particle waters contained those which correctly predicted the crystallographic positions and also those which incorrectly predicted their positions. The coverage refers to the number of correctly predicted crystallographic waters by each population, irrespective of the number of incorrectly predicted water molecules. 172

Figure A5. The accuracy of SPLASH'EM water placement method as decomposed into five different bins of equal sizes (46 PDB waters). 173

Figure C1. Orbital coefficient and molecular orbital energies of 1,2,3-triazine as optimized by CCSD(T)/cc-pVTZ. 185

Figure C2. Orbital coefficient and molecular orbital energies of 4H-1,2,4-triazole as optimized by CCSD(T)/cc-pVTZ. 186

Figure C3. Orbital coefficient and molecular orbital energies of benzene as optimized by CCSD(T)/cc-pVTZ. 186

Figure C4. Orbital coefficient and molecular orbital energies of furan as optimized by CCSD(T)/cc-pVTZ. 187

Figure C5. Orbital coefficient and molecular orbital energies of imidazole as optimized by CCSD(T)/cc-pVTZ. 187

Figure C6. Orbital coefficient and molecular orbital energies of isothiazole as optimized by CCSD(T)/cc-pVTZ. 188

Figure C7. Orbital coefficient and molecular orbital energies of isoxazole as optimized by CCSD(T)/cc-pVTZ. 188

Figure C8. Orbital coefficient and molecular orbital energies of oxazole as optimized by CCSD(T)/cc-pVTZ. 189

Figure C9. Orbital coefficient and molecular orbital energies of pyrazine as optimized by CCSD(T)/cc-pVTZ. 190

Figure C10. Orbital coefficient and molecular orbital energies of 1H-pyrazole as optimized by CCSD(T)/cc-pVTZ. 191

Figure C11. Orbital coefficient and molecular orbital energies of pyridazine as optimized by CCSD(T)/cc-pVTZ. 191

Figure C12. Orbital coefficient and molecular orbital energies of pyridine as optimized by CCSD(T)/cc-pVTZ. 192

Figure C13. Orbital coefficient and molecular orbital energies of pyridmine as optimized by CCSD(T)/cc-pVTZ. 193

Figure C14. Orbital coefficient and molecular orbital energies of pyrrole as optimized by CCSD(T)/cc-pVTZ. 194

Figure C15. Orbital coefficient and molecular orbital energies of 1,3,5-triazine as optimized by CCSD(T)/cc-pVTZ. 194

Figure C16. Orbital coefficient and molecular orbital energies of 1,2,4,5-tetrazine as optimized by CCSD(T)/cc-pVTZ. 195

Figure C17. Orbital coefficient and molecular orbital energies of 1,3,4-thiadiazole as optimized by CCSD(T)/cc-pVTZ. 196

Figure C18. Orbital coefficient and molecular orbital energies of thiazole as optimized by CCSD(T)/cc-pVTZ. 196

Figure C19. Orbital coefficient and molecular orbital energies of thiophene as optimized by CCSD(T)/cc-pVTZ. 197

Figure C20. Orbital coefficient and molecular orbital energies of 2H-1,2,3-triazole as optimized by CCSD(T)/cc-pVTZ. 198

Figure C21. Orbital coefficient and molecular orbital energies of 1H-1,2,4-triazole as optimized by CCSD(T)/cc-pVTZ. 199

Figure C22. Orbital coefficient and molecular orbital energies of 1H-1,2,3-triazole as optimized by CCSD(T)/cc-pVTZ. 200

Figure C23. Orbital coefficient and molecular orbital energies of methylenephosphane as optimized by CCSD(T)/cc-pVTZ. 200

Figure C24. Orbital coefficient and molecular orbital energies of methylenesilane as optimized by CCSD(T)/cc-pVTZ. 201

Figure C25. Orbital coefficient and molecular orbital energies of methanethial as optimized by CCSD(T)/cc-pVTZ. 201

Figure C26. Orbital coefficient and molecular orbital energies of ethylene as optimized by CCSD(T)/cc-pVTZ. 201

Figure C27. Orbital coefficient and molecular orbital energies of formaldehyde as optimized by CCSD(T)/cc-pVTZ. 202

Figure C28. Orbital coefficient and molecular orbital energies of methanimine as optimized by CCSD(T)/cc-pVTZ. 202

Figure C29. Observed trends in V_2. In general, V_2 increased from left to right, signifying a decrease in torsional energy barrier. 202

Figure C30. V_2 of QM vs. $\sigma \rightarrow \sigma^*$ hyperconjugation obtained from NBO. 203

Figure C31. V_2 of QM-$MM_{nonbonded}$ vs. $\sigma \rightarrow \sigma^*$ hyperconjugation obtained from NBO. 203

Figure C32. V_1 of QM-$MM_{nonbonded}$ of GAFF2 vs. $\sigma \rightarrow \sigma^*$ hyperconjugation obtained from NBO. 204

Figure C33. Comparison between the torsional energy profiles of a) (S)-6-(4-(1H-pyrrol-3-yl)phenyl)-5-methyl-4,5-dihydropyridazin-3(2H)-one and b) 2-amino-2-thioxoethane(dithioperoxo)imidic as predicted by QM (red) and GAFF2 (blue)..... 204

Figure C34. RMSE of H-TEQ 4.5 vs. magnitude of V_1 obtained from QM-$MM_{nonbonded}$ of GAFF2 are plotted for all molecules with RMSEs greater than 1.5 kcal·mol^{-1}. The green dotted line is used to classify molecules based on the performance of H-TEQ 4.5. Molecules to the right of this line were erroneous due to a missing V_1 term, while molecules to the left were inaccurate due to imperfections in the V_2 term. 205

Figure C35. Torsional energy profile of [2,2'-bipyridine]-1,1'-diium, whereby the 1-4 electrostatics and 1-4 Van der Waals energies were multiplied by a factor of 6.5. 205

List of Tables

Table 2.1. PW hydrogen bonding parameters. 60

Table 2.2. Polarization factors assigned to H-bond potentials. 61

Table 2.3. PW placement accuracy of current developments 64

Table 3.1. Previous Developments of H-TEQ. 80

Table 3.2. V_2 for Various Substituted 1,1'-Biphenyl Molecules 94

Table 3.3. Associated Parameters to Reproduce V_2 for Various Categories of Biaryls 97

Table 5.1. MAE (in kcal·mol^{-1}) of the energies of 50 common pharmaceuticals calculated using various MM methods compared to the reference QM. Only drugs without extensive steric clashes were included. Bold indicates most accurate MM method. 156

Table C1. Profile. All RMSEs are reported in kcal/mol. 206

Table C2. Ideal V_1 and V_2 values for 200 druglike molecules using GAFF2, MAB, and MMFF94. All values are reported in kcal/mol. 214

List of Equations

Equation 1.1. Equation for calculating the pseudorotational angle of sugars. 6
Equation 1.2. Potential energy functions of standard MM FFs. 16
Equation 2.1. Hydrogen Potential Function of PWFF. 52
Equation 2.2. Equation used for fitting and parametrizing PWFF. 54
Equation 2.3. The hydrogen bonding potential for PWFFa. 56
Equation 2.4. Hydrogen bonding potential for pPWFFa. 59
Equation 3.1. Potential energy function of standard MM FFs. 78
Equation 3.2. The torsional energy function of MM FFs. 78
Equation 3.3. Equation for calculating the π-electronegativity of a single aryl group. 88
Equation 3.4. Equation for calculating the total π-electronegativity of a biaryl molecule. 88
Equation 3.5. Equation for calculating the π-electronegativity difference within a biaryl molecule 88
Equation 3.6. Equation for calculating V2 based on electron-richness/deficiency. 96
Equation 4.1. Potential energy function of standard MM FFs. 115
Equation 4. 2. Potential energy function of standard torsions in MM FFs. 115
Equation 4.3. Function to predict V_2 based on various EHT descriptors. 127
Equation 5.1. Functional form found in type I MM FFs. 143

List of Abbreviations

3D-RISM	3D Reference Interaction Site Model
A	Adenine
ALMO-EDA	Absolutely Localized Molecular Orbitals-Energy Decomposition Analysis
AM1-BCC	Austin Model 1 - Bond Charge Corrections
AMOEBA	Atomic Multipoles Optimized Energetics for Biomolecular Applications
BLW	Block-localized Wavefunction
BO	Bond Order
C	Cytosine
CCSD	Coupled Cluster Single-Double
CCSDT	Coupled Cluster Single-Double-Triple
CG	Coarse Grain
CHARMM	Chemistry at Harvard Macromolecular Mechanics
CPU	Central Processing Unit
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DNA	Deoxyribose Nucleic Acid
EDIA	Electron Density Score for Individual Atoms
EHT	Extended-Huckel Theory
EWG	Electron Withdrawing Group
FEP	Free Energy Perturbation
FF	Force Field
FITTED	Flexibility Induced Through Targeted Evolutionary Description
FMN	Flavin Mononucleotide
FMO	Frontier Molecular Orbital Theory
G	Guanine
GAAMP	General Automated Atomic Model Parametrerization
GAFF	General Amber Force Field
GAMESS	The General Atomic and Molecular Electronic Structure System
GPU	Graphic Processing Unit
HBA	Hydrogen Bond Acceptor

HBD	Hydrogen Bond Donor
HDAC	Histone Deacetylase
HF	Hartree-Fock
HIV	Human Immunodeficiency Virus
HSAB	Hard and Soft Acids and Bases
H-TEQ	Hyperconjugation for Torsional Energy Quantification
HTS	High Throughput Screening
ITC	Isothermal Titration Calorimetry
JAWS	Just Add Water Molecules
LED	Local Energy Decomposition
LJ	Lennard-Jones
MAE	Mean Absolute Error
MD	Molecular Dynamics Simulations
MM	Molecular Mechanics
MMFF94	Merck Molecular Force Field 94
MO	Molecular Orbital
MOE	Molecular Operating Environment
MP2	Møller–Plesset Perturbation Theory
NBO	Natural Bonding Orbitals
NEDA	Natural Energy Decomposition
NMR	Nuclear Magnetic Resonance
OL	Olomouc
OPLS	Optimized Potentials for Liquid Simulations
OPLS-AA	Optimized Potentials for Liquid Simulations - All Atom
P450	Cytochrome P450
PDB	Protein Data Bank
pPWFFa	polarizable Particle Water Force Field with angle corrections
PW	Particle Water
PWFF	Particle Water Force Field
PWFFa	Particle Water Force Field with angle corrections
QM	Quantum Mechanics

QMDFF	Quantum Mechanically Derived Force Field
QUBEKit	QUantum mechanical BEspoke Kit
RESP	Restrained Electrostatic Potential
RMSD	Root Mean Square Deviation
RMSE	Roomt Mean Square Error
RNA	Ribose Nucleic Acid
SAM	S-Adenosyl Methionine
SAMPL	Statistical Assessment of the Modeling of Proteins and Ligands
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SBDD	Structure-based Drug Design
SDF	Structure-data File
SPLASH'EM	Solvation Potential Laid Around Statistical. Hydration on Entire Macromolecules
SVL	Scientific Vector Language
SZMAP	Solvent-Zap-Map
TIP3P	Three-site Transferrable Intermolecular Potential
TIP3P	Thymine
U	Uracil
UTR	Untranslated Region
VDW	van der Waals
WHO	World Health Organization
XNA	Xeno Nucleic Acid

1 Introduction to Molecular Mechanics Force Fields for Simulating Nucleic Acids and Applications to Drug Discovery and Dynamics

1.1 Preface

This chapter introduces the discipline of molecular mechanics force fields as pertinent to the simulation of nucleic acids and its applications to the field of drug discovery. Despite the usefulness of innovations in computational chemistry over the years, there are still many challenges in the field. Some of these challenges are presented in this chapter, with subsequent **Chapters 2-5**, proposing solutions to these problems and addressing these key issues. The advancements made in these subsequent chapters should be useful for future drug discovery projects.

This chapter is based on work from: Wei, W.; Moitessier, N., Status of Molecular Mechanics Force Fields for Nucleic Acid Modelling and Simulations. *Manuscript in Preparation.*

Author contributions: I performed the literature review, analysis, and write-up of the manuscript. Prof. Nicolas Moitessier and I designed the layout of this review article.

1.2 Introduction

Nucleic acids are biomolecules that are found across all domains of life and viruses, which play crucial roles in storing genetic information, signaling, and catalyzing essential, life-sustaining reactions.[1] The importance of nucleic acids in living organisms has inspired the development of many *in silico* methods in computational chemistry to better predict various biochemical and physiological properties. In recent years, scientific endeavors in nucleic acid research has dramatically improved our understanding, and given rise to new potential therapies to combat diseases.[2, 3]

The aim of this review is to introduce and discuss the utility, historic developments, and the status of molecular mechanics (MM) force fields (FFs) for modelling and simulating nucleic acids. These topics are introduced in several stages. Firstly, the basic structures and functions of nucleic acids will be discussed. This would be helpful for familiarizing the reader to subsequent parts of this review. Secondly, promising nucleic acid therapies and pharmaceuticals are examined. Thirdly, a wide range of computational chemistry methods for studying nucleic acids are summarized. These *in silico* methods would be focused on those which make use of MM FFs. Fourthly, a more in-depth inspection of MM FFs, including historical developments and the state-of-the-art, is performed. Lastly, this review points at shortcomings in current nucleic acid FFs and possible future directions and developments.

1.3 Nucleic Acids Functions

Nucleic acids are found across all domains of life and viruses. All living organisms store their genetic information in the form of deoxyribonucleic acids (DNAs), wrapped into folded chromosomes, which act as the blueprint to direct cellular activities and functions. In contrast, ribonucleic acids (RNAs) are used to store genetic information for a few RNA viruses including influenza,[4] SARS-CoV-2,[5] and HIV.[6] In living organisms, RNAs are primarily involved in the protein synthesis processes.[7] For a protein molecule to be synthesized, the genetic information found in DNA must first be read and transcribed into an messenger RNA (mRNA) molecule by various transcription proteins.[8] Subsequently, the mRNA is sent to and interpreted by the ribosomal RNA (rRNA), which is a major component of the protein translation process.[9]

Aminoacyl-transfer RNAs (aa-tRNAs) are couriers which deliver the correct amino acids to the rRNA during protein translation.[10]

Besides their involvement in protein synthesis, RNAs also regulate gene expression using various mechanisms. MicroRNAs (miRNA)[11] and X-inactive specific transcript (Xist)[12] are naturally occurring nucleic acid molecules involved in the gene silencing pathways for specific chromosomal elements. Following their discoveries, RNA interference (RNAi) has been an active area of therapeutic research aimed at employing artificially introduced small interfering RNAs (siRNAs) for highly specific and targeted gene-silencing.[13] The involvement of RNAs in many key areas of biology has led various researchers to propose the "RNA world" hypothesis. It postulated that in the past, all cellular activities were performed by RNAs prior to the emergence of proteins.[14] In fact, the discovery of ribozymes,[15] RNAs capable of catalyzing biochemical reactions in 1982, and that of riboswitches,[16-18] cis-regulatory elements found on the mRNA of certain species in 2002, further reinforced the idea that life may have been possible prior to the evolution of proteins.

1.4 Nucleic Acid Structure

Structurally, DNAs and RNAs are made up of repeating nucleotide subunits with three distinct components: a nitrogenous base, an aldopentose sugar, and a phosphate group.[19] Five common nitrogenous bases exist. While adenine (**A**), guanine (**G**), and cytosine (**C**) are found in both DNA and RNA, thymine (**T**) is exclusively found in DNA while uracil (**U**) is found in RNA. A sixth nucleobase, hypoxanthine (**I**) is also frequently found in tRNAs.[20] Nucleobase modifications, such as methylation, are also found in both DNA and RNA; these fields of study are termed epigenetics and epitranscriptomics, respectively.[21, 22] Nitrogenous bases could be classified into two groups, purines (eg. A, G, and I) and pyrimidines (C, T, and U), based on their chemical structures. While the former is bicyclic and aromatic, the latter possesses only a single aromatic ring.

In most structures of DNA and RNA, Watson-Crick base pairing occurs.[23] Complementary nitrogenous bases hydrogen bond with each other to create a double helix of a consistent width. While **C** pairs with **G** via three hydrogen bonds, **A** pairs with **T** (in DNA) and **U** (in RNA) via two hydrogen bonds (**Figure 1.1**). Besides Watson-Crick base pairing, other non-canonical base pairing could also be found, including Hoogstein base pairing often found in DNA and RNA G-quadruplexes (DNA-G4s and RNA-G4s). In these structures, the glycosidic bond connecting the

nitrogenous base to the aldopentose sugar is rotated 180° from its normal anti geometry, leading to the cis conformation. Unlike Watson-Crick base pairing, Hoogstein base pairing always has two hydrogen bonding interactions. Wobble base pairing, which are base pairing not fulfilling Watson-Crick rules, is frequently found in tRNA molecules for codon-anticodon recognition during protein translation.[20, 24, 25] In fact, the third base pair during codon recognition is able to bind via Wobble base pairing. The recognition of these pairs, in addition to Watson-Crick, allows each codon to bind to and recruit more than one tRNA. Wobble base pairing occurs most frequently between **G** and **U**; **I** and **U**; **I** and **A**; **A** and **C**; and **I** and **C**. In addition to these, other base pairing modes also exists, including reverse-Watson-Crick base pairing, in which the $\mathbf{U}_{\mathbf{O4}}$ rather than $\mathbf{U}_{\mathbf{O2}}$ hydrogen bonds with $\mathbf{A}_{\mathbf{N6}}$.[25]

Base stacking interactions between adjacent nucleobases also confers additional stability upon the overall nucleic acid structure. This π-π stacking interaction could be caused by the unique π-cloud surrounding an aromatic system and plays an important role during intercalator binding between DNA or RNA bases.

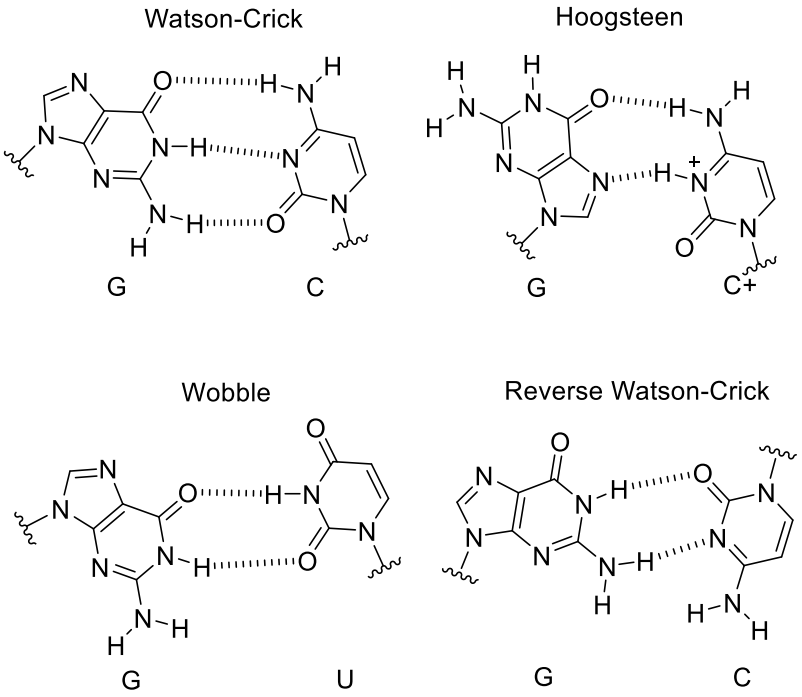


Figure 1.1. Canonical Watson-Crick (left) compared to Hoogsteen (right) base pairing. The guanine nucleotide is in the anti- and syn-conformation in the Watson-Crick and Hoogsteen base pairing, respectively.

Aldopentose sugars of nucleotides are connected to the nucleobases via the β-glycosidic bond, called the χ torsion angle (Figure 1.2). While D-2-deoxyribose sugars are present in DNAs, RNA nucleotides possess D-ribose sugars. Despite differing by just a 2'-OH group, it has major implications in stability and function. Structurally, the sugars of nucleic acids adopt distinct conformations, described using the pseudorotation phase angle (*P*), based on the torsion angles of the furanose ring, which are labelled $\delta_{0\text{-}4}$.[26, 27] Ψ could be calculated according to equation 1.1 and could range between 0° and 360°. Although all values of Ψ are possible, only two regions are heavily populated in naturally occurring nucleic acids (Figure 1.3). Ψ values between 0° and 36° correspond to C3'-endo conformation, which almost all nucleotides of A-form DNA and RNA adopt. This is also called the Northern sugar pucker conformation due to its position on the pseudorotational compass. In contrast, the commonly occurring B-forms of DNA and of the rarer RNA is called the Southern sugar pucker or C2'-endo conformation and possesses Ψ values between 144° and 180°. The nucleotides of Z-form DNA and RNA could adopt both the Northern and Southern sugar pucker conformations. It should be noted that while both A- and B-forms of

nucleic acids are right-handed helices, Z-forms are left-handed. Chemically, it is known that Ψ is determined by a combination of steric, electrostatic, hydrogen bonding, hyperconjugation, and solvation effects.[28-30] Synthetically modified nucleic acids, called xeno nucleic acids (XNA), could modulate the conformational preferences of sugar puckering.[31, 32] Consequently, this approach is currently being pursued by research groups around the world in order to obtain the desired conformations and functions in various therapeutic areas (e.g. RNAi).[33] In nature, sugar puckering has important implications for recognition by various protein machinery, including those involved in replication, transcription, and translation.[34, 35]

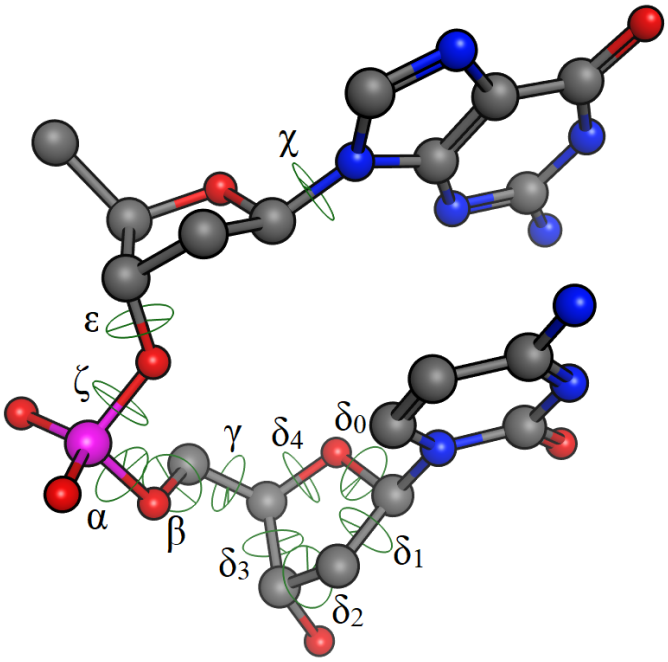


Figure 1.2. Definition of various torsional angles found in nucleic acids.

Equation 1.1. Equation for calculating the pseudorotational angle of sugars.

$$\Psi = \tan^{-1}\frac{-\delta_0+\delta_1-\delta_3+\delta_4}{2\delta_2\,(\sin 36°+\sin 72°)}$$

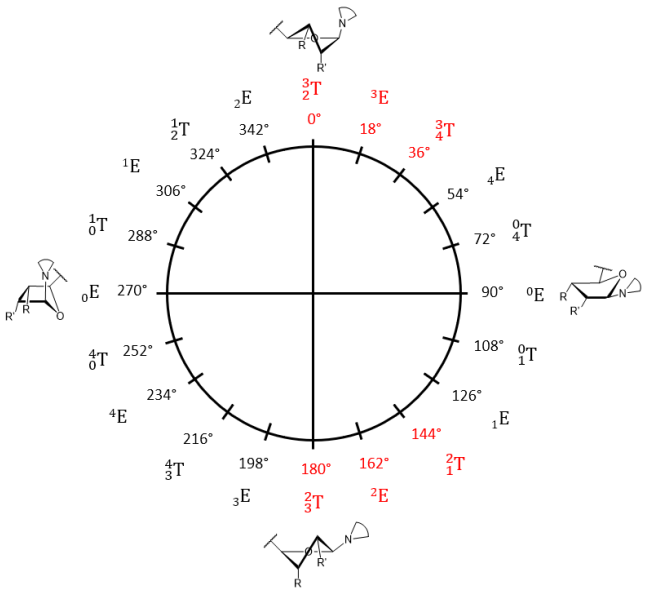


Figure 1.3. All possible pseudorotational angles and selected conformations of the nucleic acid aldopentose sugar are shown. Red labels indicate conformations frequently found in naturally occurring nucleic acids. Superscripted and subscripted numbers preceding E or T indicate endo- and exo-positions, respectively.

The sugar subunits in nucleic acids are attached via their 5'- and 3'-carbon of the sugar atoms by phosphodiester bonds of phosphate groups. Torsional bond of O3'-P-O5'-C5' is known as the α torsion angle, while the P-O5'-C5'-C4' is known as the β torsion angle (Figure 1.2). Similarly, the torsional bond of O5'-P-O3'-C3' is known as ζ torsion angle while the P-O3'-C3'-C4' is known as the ε torsion angle. In nucleic acids, phosphate groups are important for ligand recognition and binding. Like aldopentose sugars, phosphate dihedral angles could vary depending on the nucleic acid state (i.e. type-A, type-B, or type-Z conformations, Figure 1.4). In type-A nucleic acids, the distance between the 5'- and 3'-O are smaller than in that of the B-form, reducing the distance between adjacent nucleotides. Even within the same states, variations in phosphate dihedral angles could occur, known as A_I, A_{II}, B_I, B_{II}, Z_I, and Z_{II} substates of DNA and RNA.[36-38] Although A_I and A_{II} have similar Ψ and sugar puckering, they differ in their α, β, and γ torsional angles.[38] In the A_I substate, average torsional angles of 201°, 294°, and 172° were previously observed for α, β, and γ, respectively by Sims et al (2003) based on a statistical survey of the nucleic acid database (NDB).[38, 39] In contrast, A_{II} substate had torsional values of 188°, 145°, and 190° for α, β, and γ, respectively. Of these conformations, the A_I substate is the canonical and more

energetically favorable. The differences between B_I and B_{II} substates correspond to ε and ζ torsional angles, which are coupled. While B_I takes on ε and ζ values between 120°-210° and 235°-295°, respectively; the less energetically stable B_{II} has values between ε = 210°-300° and ζ = 150°-210°, respectively.[37, 38, 40] Often, a single combined metric is used: ε-ζ, which is approximately -90° for B_I and +90° for B_{II} substates. Z_I and Z_{II} populations also differ in the ζ and β of the subsequent nucleotide torsional angles.[41] In nature, frequent transitions between various states and substates have been observed. In fact, the relative populations of these nucleic acid states in simulations have been used to verify their accuracies.

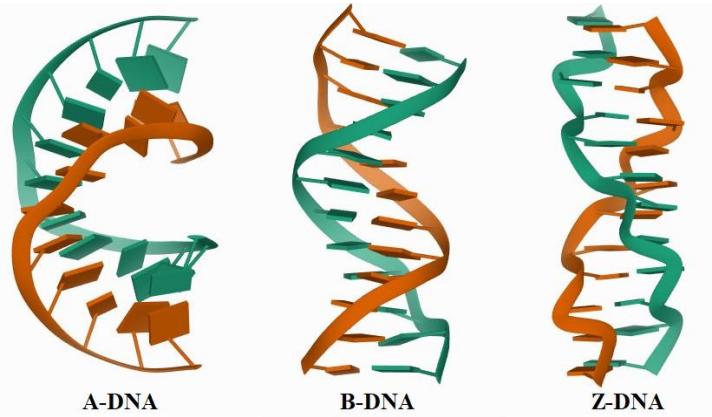


Figure 1.4. A-, B-, and Z-forms of DNA are shown. Each type of helix has different geometrical properties, such as width, rise per base, and directionality of turn. RNA could also adopt these geometries.

1.5 Nucleic Acids as Pharmaceutical Targets

Nucleic acids have been of therapeutic interest for several decades. Recently, computational chemistry methods have been applied to many different areas of therapeutic research to yield useful, insightful, and efficient predictions. Although not exhaustive, this section lists some areas of research that are of high relevance and importance for future pharmaceutical developments against nucleic acid targets. The promise of these therapeutic areas of research have inspired and would continue to inspire future developments in computational chemistry.

RNA riboswitches are promising therapeutic targets for antibiotics due to their high prevalence across members of different bacterial species and their absence in mammalian mRNAs. The binding to these aptamer regions of metabolites[16-18] and druglike[2] molecules was shown to

inhibit either mRNA transcription or protein translation (upregulation upon riboswitch binding is rare). In fact, ribocil was found to be an active inhibitor of the bacterial FMN riboswitch, involved in the riboflavin homeostasis pathway (Figure 1.5 and Figure 1.6A). Another RNA class which show promise as pharmaceutical targets include expanded repeats.[42] In humans, repetitive codons of 20~30 are indicative of normal mRNAs. Unfortunately, however, repetitive trinucleotide segments of several hundreds to thousands indicate neuronal diseases. The former causes the recruitment of excess amount of the cell's splicing machinery, leading to many misspliced protein isoforms. Although a permanent cure is difficult to attain, small molecules were found to bind and lead to the synthesis of correctly spliced proteins.[3, 42] Of notable progress in this area has been the pharmaceutical, branaplam, which was shown to be effective against spinal muscular atrophy, which entered phase-II clinical trials in 2017 (Figure 1.6B). [3, 43] Over the past few decades, antibiotics were shown to inhibit rRNA of bacteria and viruses, thereby interfering with the protein translation machinery.[44] These include tetracyclines and other molecules (Figure 1.6C). Collectively, these promising results have piqued the interest of the scientific community in attempting to target RNAs with pharmaceuticals. Colloquially, this new focus is known as the "RNA gold rush".[45, 46]

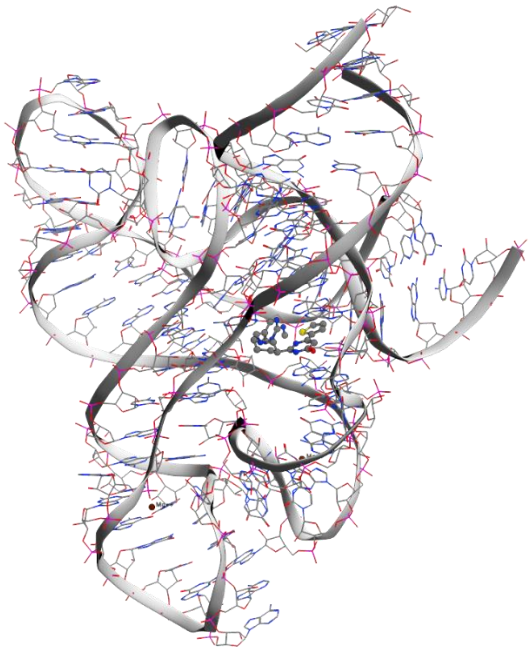


Figure 1.5. Crystal structure of ribocil (ball and stick model) bound to the aptamer domain of FMN riboswitch (cartoon and wireframe model, PDB ID: 5C45).

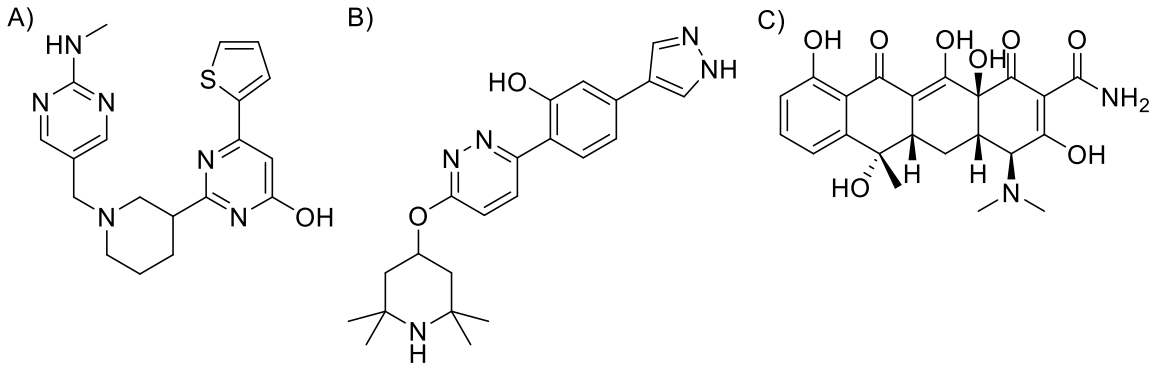


Figure 1.6. Small molecular RNA inhibitors and pharmaceuticals are shown. A) Ribocil is an inhibitor of FMN riboswitch, which was found to disrupt bacterial riboflavin homeostasis. B) Branaplam is a promising drug candidate for spinal muscular atrophy. C) Tetracycline is an antibiotic known to bind bacterial rRNA.

DNAs have also been the subject of interest as pharmaceutical targets for a variety of different diseases, including viral, bacterial, and oncogenic illnesses. During viral infections, the DNA of these pathogens becomes incorporated into that of the host genome. Subsequently, viral DNA polymerase rapidly catalyzes DNA replication to further infect more host cells. Antivirals, such as acyclovir, were found to be effective in blocking DNA polymerases of diseases such as herpes viruses by imitating the structure of nucleosides (Figure 1.7A).[47] These antivirals lack a 3'-OH group, which makes subsequent nucleoside incorporation impossible, terminating DNA replication. In the past, DNA-intercalating bacteriostatic, such as proflavine, were used for wounds sterilization (Figure 1.7C).[43] However, its mutagenic and harmful effects on human epithelial cells discontinued its use.[48] Proflavine disrupts the Watson-Crick base pairing of DNA through intercalation, leading to frameshift mutation or base deletion.[49] Lead optimization is currently being performed to identify compounds with similar properties but greater specificity against bacterial DNAs.[48]

In targeting oncogenes in the human body, several DNA targets are available. One of these are DNA-G4s, found on the ends of chromosomes.[50] As suggested by their names, they are four-stranded helices composed of guanines which are stabilized by Hoogstein base pairing. In normal cells, telomeric regions of chromosomes, where DNA-G4s are found, shorten with each cycle of DNA replication through the natural progression of aging.[51] Unfortunately, hyperactive

telomerases repair and elongate the telomere region in cancer cells, allowing them to achieve immortality.[52] Human therapeutic research efforts are currently underway to prevent telomerase activity by attempting to stabilize DNA-G4 structures. [53-55] This is done by identifying small molecules known to prevent telomerase binding in cancer cells. Potential pharmaceuticals include [16]phenN_4, which is a phenanthroline polyazamacrocycle (Figure 1.7B).[56] Other DNA elements could also be targeted using small molecules, including alkylating agents, which induce chemical modifications to DNAs to kill cancer cells during chemotherapy.[57] These include chemotherapy drug such as 1,4-butanediol dimethanesulfonate (Figure 1.7D). However, this process could have severe side-effects due to the relatively nonspecific localization of these alkylating agents in the human body.

In recent years, DNA gene-editing methods using CRISPR-Cas 9 pathways have gained widespread excitement and attention for their potential to cure genetic diseases.[58-60] However, this technology is still relatively young, and would require many more years of safety testing prior to being widely adopted for use as therapeutics.

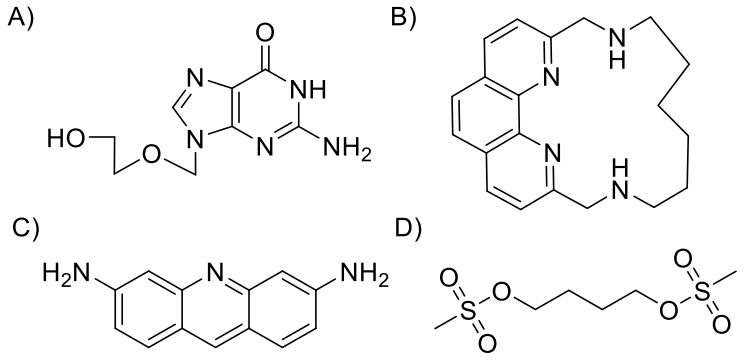


Figure 1.7. Small molecule DNA inhibitors and pharmaceuticals are shown. A) Acyclovir is an essential drug, which causes early termination during viral replication in many diseases. B) [16]phenN_4 are promising G4-binders which could have antitumor abilities. C) Proflavine is a bacteriostatic compound, which was used for sterilization. D) 1,4-butanediol dimethanesulfonate is an akylating agent used during chemotherapy.

1.6 Use of Molecular Mechanics Force Fields in Computational Chemistry

Over the past few decades, the ability to simulate nucleic acids at the atomic resolution has been useful for understanding a variety of different biochemical phenomena. In tandem with

experiments, investigations into nucleic acid dynamics and stabilities using MD simulations have been fruitful and produced many insightful predictions. In addition, the use of computational chemistry methods, employing MM, are rapid and affordable compared to many traditional experimental techniques. With the advent of increasingly more powerful computers, their use and importance are growing. Some historic and hallmark computational studies of nucleic acids employing MM FFs are subsequently described.

Prior to the turn of the century, hydration patterns surrounding DNA and RNA were studied by collecting static X-ray diffraction data.[61] Although useful, this method depended on the resolution of the obtained crystal structures and the residence time of these water molecules (B-factor). Although solution-based NMR structures could provide a native and dynamic depiction of nucleic acids, they do not contain water molecules. To understand the hydration environment of nucleic acids, Auffinger et al (1998) employed molecular dynamics (MD) to simulate the movement of water molecules around RNA nucleotides. From this study, it was discovered that RNA nucleotides were mostly solvated in the plane of the nucleotide (Figure 1.8A),[62] which had important implications for structure-based drug design (SBDD) of novel pharmaceuticals. It is known that desolvation and bridging water molecules are crucial for binding of nucleic acids to proteins and ligands.[46, 63]

In other studies, various groups have used MD simulations to study substate transitions of nucleic acids to better understand the mechanisms affecting their stability.[30, 40, 64, 65] In fact, guided by experimental evidences, Jayaram et al (1998) observed *in silico* that solvation, organization of counterions, and interphosphate repulsion were key factors which facilitated the transition between A- and B-DNA (Figure 1.8B).[66] An understanding of the mechanisms affecting substates of nucleic acids is especially important because of its implications in protein and ligand recognition and binding.[67, 68]

In order to calculate the relative free energies of binding between different ligands, MD simulations could be employed using a technique known as free energy perturbation (FEP) calculations.[69] In this technique, the user defines both the initial and target ligands of interest. At the start of the MD simulation, the initial ligand is incrementally mutated, stepwise, through alchemical transformations into the target ligand by adjusting its associated MM parameters by a coupling parameter, λ. The free energy difference at each step is calculated by the Zwanzig

equation.[69] At the conclusion of the transformation, the free energy difference between the initial and target ligand could be obtained. For a more complete review of FEP, readers are to referred to these other excellent articles.[69-71] Since the landmark use of FEP by Jorgenson and coworkers,[72] it has been applied to investigate a variety of different nucleic acid systems,[73-76] including the binding of different ligands to RNA purine riboswitches (Figure 1.8C).[77] Besides having good correlations with experimentally-obtained results, FEP could give insights into the detailed molecular and energetic mechanisms for ligand binding and discrimination.[77]

Molecular docking is another method, often used in computational chemistry, which could predict the free energy of binding of ligands or proteins to nucleic acids.[46, 54, 78-81] Its main advantage over FEP is the ability to more thoroughly search the conformational space, allowing high energy barriers to be overcome using various docking algorithms. This has the added benefit of being less computationally expensive since docking is not dependent on previous timesteps and quickly achieves convergence. One disadvantage is that the obtained binding free energy may be significantly less accurate than those obtained by FEP since it relies on empirical corrections to conformational entropy. Both docking and FEP suffer in accuracy when receptor flexibility is important which is often the case for highly dynamic RNA molecules. Nevertheless, virtual high-throughput screening (HTS) methods, using molecular docking, has been very useful in the past for their abilities to greatly enrich compound libraries for potential actives (Figure 1.8D).[82, 83] Virtual HTS against RNA riboswitches, to identify promising antibiotics, is currently ongoing in our laboratory.

Hybrid QM/MM calculations, using both quantum mechanics (QM) and MM methods, have been widely used to elucidate chemical reactivities, mechanisms, and dynamics involving nucleic acids (Figure 1.8E).[84, 85] In these multi-scale computations, molecules are usually separated into two or more distinct layers. While the region of greater interest (high layer) is treated using the computationally more expensive QM method, surrounding regions are treated using MM (low layer). The latter confers mechanical and electrostatic effects, which perturbs and is considered by the QM-region. This approach has been used successfully to study the chemical mechanisms of human DNA repair enzymes,[84] peptide bond formation in rRNA,[85] editing process in various aminoacyl-tRNAs,[10, 86-89] and nucleotide demethylation during epigenetic modifications.[90] In the future, with the advent of more computational power, QM/MM-MD simulations would become

more tractable for investigating chemical phenomena, dynamically. Currently, these simulations are still mostly restricted to using semiempirical methods in the high layer due to their high computational cost.[10, 28] In addition, hybrid MM/CG (Molecular Mechanics/Coarse Grain) simulations have also been used to study systems of larger size. In the same manner as QM/MM simulations, the region of interest could be modelled using MM, while the surrounding regions are modelled using a CG model. Although CG models for nucleic acids exist,[91] MM/CG simulations have mainly been applied to membrane proteins, in the context of lipid membranes.[92] However, the option to model nucleic acids using MM/CG simulations remains open in the future. For example, simulations depicting viral nucleic acid release in the context of viral phospholipid bilayer could be performed through MM/CG calculations.

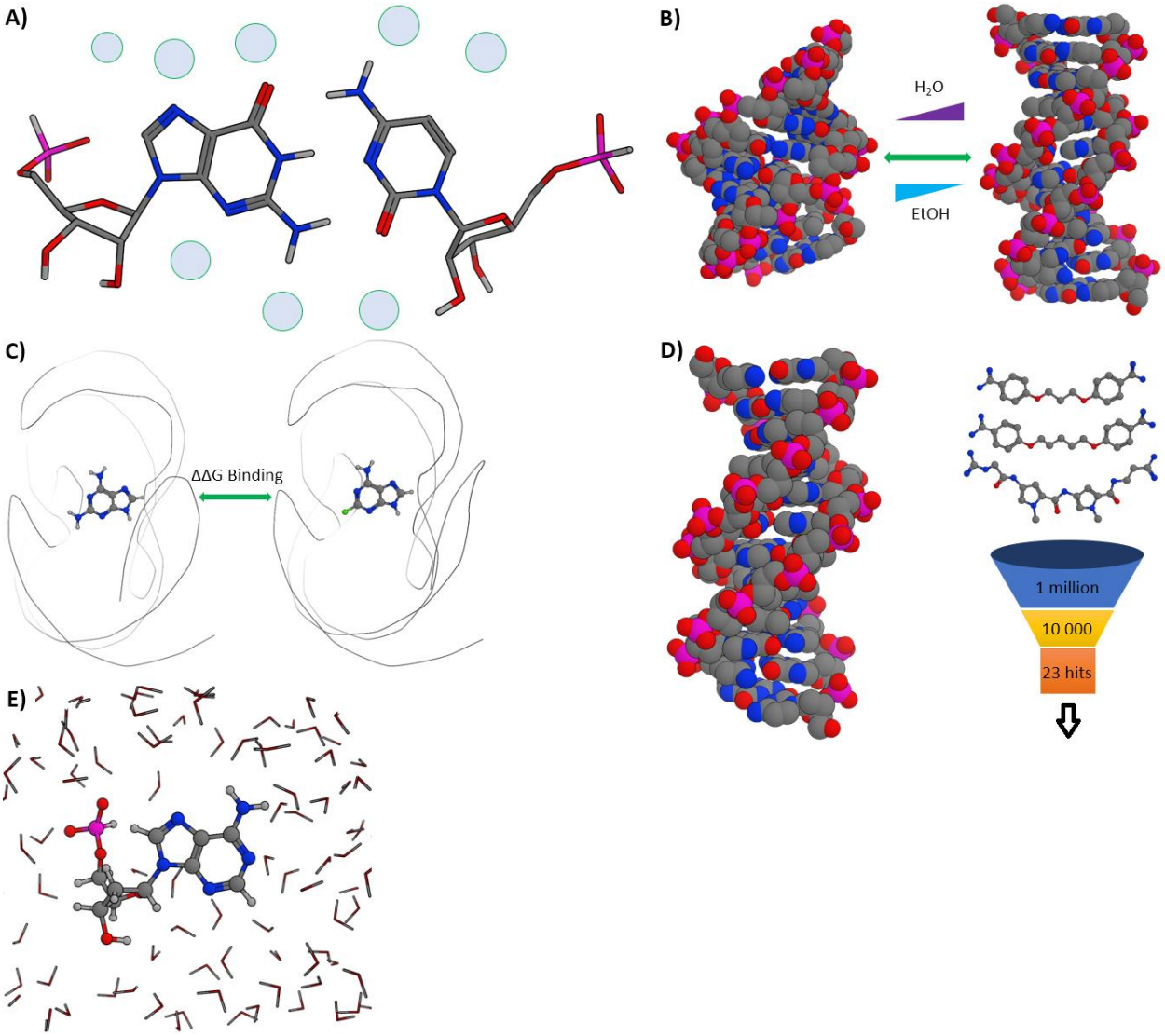


Figure 1.8. Various uses of MM in computational chemistry are shown. A) A hallmark investigation by Auffinger et al into RNA dynamics and solvation by classical MD simulations found that water molecules localized with high frequency in certain spatial regions (light blue).[93] B) An MD study into the equilibrium of A- and B-DNA in different solvation environments hinted at the possible role of cations and phosphates as the key driving force behind these transitions.[66] Higher concentration of water and ethanol induces A- and B-DNA, respectively. C) FEP study of purine riboswitches reproduced experimental binding free energy differences between a variety of ligands.[77] D) Molecular docking methods could be used to quickly and inexpensively screen for pharmaceutically promising drug candidates (right) against a receptor of interest (left) during the initial drug discovery process.[82] E) QM/MM simulations have been applied to investigate chemical

reactions and/or dynamics. For example, the conformational dynamics of nucleosides could be investigated. The molecular region of interest (QM) and surrounding solvents (MM) are shown in ball-and-stick and wire representations, respectively.

1.7 Molecular Mechanics Force Fields for Modelling Nucleic Acids

Although QM calculations are grounded in first principles and have theoretical basis, they are computationally intractable for many purposes. For example, QM optimization, QM-MD, and QM-based scoring functions (during *in silico* molecular docking) are computationally too demanding at the present time for large macromolecules such as DNA double helices and RNA riboswitches. In order to study these larger biological systems, MM FFs were conceptualized and employed. In contrast to QM, MM FFs are a series of empirical, potential energy functions used to reproduce the total energy of a molecular system.[94-98] While many nucleic acid FFs exist, they differ in their complexity and functional forms. They could be categorized as belonging to one of two groups: classical or polarizable FFs.

Classical FFs maintain simpler functional forms, which result in lower computational costs for simulations. For this reason, they have been widely adopted as the method of choice for investigating various biomolecules, including DNA and RNA. Current state-of-the-art classical FFs for modelling nucleic acids are the various updated parametrizations of the original Amber ff94 and ff99.[99] Although CHARMM[34] and OPLS-AA[100, 101] have also been parametrized for nucleic acids, ff99-derived FFs have been more rigorously tested and were found to be more accurate, overall.[102] Although not perfect, various research groups around the world have observed that current implementations of Amber ff99 and CHARMM produced features consistent with experiments (despite more testing being required).[103] Amber, CHARMM, and OPLSS-AA have very similar functional forms. The functional form for Amber FF is shown in Equation 1.1 and illustrated in Figure 1.9.

Equation 1.2. Potential energy functions of standard MM FFs.

$$E_{MM} = \underbrace{\sum_{bond} k_r\left(r - r_{eq}\right)^2 + \sum_{ang} k_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_{tor} \frac{V_n}{2}\left[1 + \cos(\eta\phi - \gamma)\right]}_{bonded} + \underbrace{\sum_{vdw}\left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right) + \sum_{ele}\left(\frac{q_i q_j}{\varepsilon R_{ij}}\right)}_{nonbonded}$$

As could be seen, the first three terms make up the bonded terms, composed of bond stretching, angle bending, and torsional rotation. The latter two potentials are the nonbonded terms, containing van der Waals and electrostatic interactions. In particular, the electrostatic term in classical FFs only considers fixed charges, neglecting the polarizability of molecules, which could vary depending on its chemical environment. For this reason, it is inherently approximate in nature. Central to FFs, is the concept of atom types, which are sets of parameters associated with each atom in the context of its chemical environment and hybridization. For example, these parameters include the strengths and equilibrium distances of the harmonic spring between two directly bonded atoms (i.e. k_r and r_{eq}). It also defines the torsional barrier height and the location of the minimum during a dihedral rotation (ie. v_n and γ). Since DNA and RNA are composed of a limited diversity of monomers (i.e. A, C, T, G, and U), a thorough parametrization of all potentials is possible. The functional form of CHARMM is similar to that of Amber except it has three additional potentials, different sets of parameters, and the decoupling of 1-4 van der Waals parameters from those of other parameters.[99] More specifically, this means that the van der Waals parameters used for 1-4 differ from those of 1-5, 1-6, and others. This is a different approach to that of Amber, which merely applies, by default, a scaling (down) factor of 2.0 and 1.2 for 1-4 van der Waals and electrostatic interactions, respectively. CHARMM uses no scaling factor for the electrostatic interactions. The OPLS-AA FF functional form is nearly identical to that of Amber.

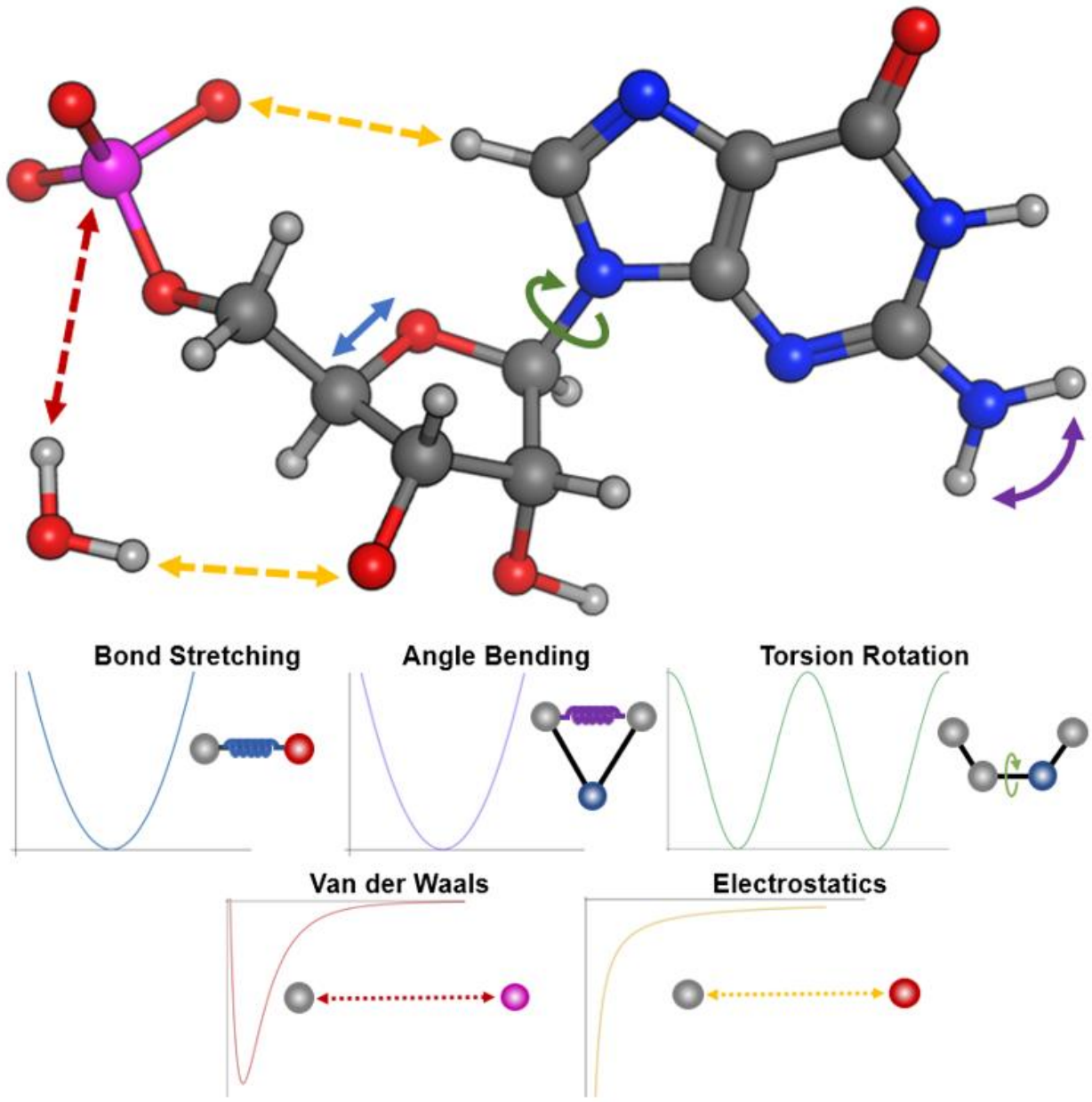


Figure 1.9. Potential energy functions found in classical MM FFs. More advanced FFs also use these terms, in addition to more complex terms.

Polarizable FFs, as their names suggest, are FFs which considers the non-static nature of charge-distribution of atoms in a molecule, by considering their local chemical environment.[104] Polarization have been taken into consideration in various ways, including polarizable point dipoles, fluctuation charge, and Drude oscillator.[105] For example, AMOEBA FF employs the polarizable point dipoles approach, which is able to move in response to external electrostatic

fields, and has been parametrized for nucleic acids.[106] In addition to partial charges, it also places fixed dipole and quadrupole charges on each atom, allowing a more accurate representation of the electrostatic potential of the molecule. In contrast, the Drude-2017 Nucleic Acid FF, originally based on CHARMM, uses two partial charges on each atom.[107, 108] While the first partial charge is stationary and static, the second is dynamic in position and tethered to the center of the atom by a harmonic spring (i.e. charge-on-spring).[109] In both FFs, the total electrostatic energy is calculated by Coulomb's Law, taking into consideration all pairs of partial charges. Besides electrostatics, polarizable FFs also assume that certain motions are correlated and non-additive, leading to the use of cross-terms. For example, AMOEBA employs cross-terms for bond-angle, angle-torsion, and bond-torsion motions. Due to these more complex functional forms, it approximately doubles the computational cost for performing MD simulations. Although in theory, polarizable FFs have a higher ceiling for more accurate simulations due to the ability to have greater control of an extended list of parameters and the ability to tweak them for reproducing chemical phenomena, their accuracies have not yet been proven. In fact, it not clear if current polarizable FFs outperforms classical FFs in terms of the accuracy of nucleic acid simulations.[110] A systematic validation of different FFs amongst a wide variety of different DNA and RNA structures is desperately needed to identify their advantages and deficiencies. Subsequently, future directions in nucleic acid FF development could be determined.

It is also important to describe the conformational energy of small molecules binding to nucleic acids. To this end, many FFs for describing small molecules have also been devised, including the general Amber FF (GAFF),[97] Merck Molecular Force Field (MMFF94),[111] and OPLS,[112]. They are compatible with previously mentioned FFs and have similar functional forms. One problem facing FFs for small molecules is the size of the chemical space, which has been estimated to be on the order of 10^{60}. Consequently, devising individual parameters for each molecule in the chemical space was not a viable option. As a result, well-known general FFs in the past have parametrized only a select few, representative molecules, and subsequently assumed transferability of these parameters for other molecules. However, this led to less accurate parameters since values could vary greatly even for very "similar" molecules. In the recent past, this has led research teams, including ours, to develop atom type free approaches, based on chemical principles and chemical perception.[113, 114] These approaches would helpful for docking and MD simulations.

1.8 Evolution of Modern Nucleic Acid Force Fields

The most widely used and validated FF for simulating nucleic acids are the various re-parameterizations of Amber ff94, which was derived using available experimental and QM calculations.[97] In this original version, torsional parameters of truncated nucleotides (see Figure 1.2) were parametrized against gas-phase computations at the MP2/6-31G* level of theory. Restrained electrostatic potential (RESP), calculated at HF/6-31G*, was used due to its ability to reproduce the electrostatic distribution in solution.

Since its publication, it had been used for the simulations of many nucleic acid structures, but soon revealed several artefacts associated with sugar puckering. The simulation of the DNA dodecamer d(C_5T_5) d(A_5G_5) under physiological conditions resulted in its fluctuation between A- and B-forms of DNA, despite the fact that B-DNA is known to be favored in nature.[115, 116] (For geometric differences between A- and B-DNA, Figure 1.3 and Figure 1.8B could be referenced.) In addition, it was found that ff94 had a tendency for C2'-endo sugar puckering to be underestimated in B-DNA.[117] Conformationally, this meant that during the simulation of DNAs, pseudorotational values ($\delta_{1\text{-}5}$) outside the experimentally observed range of 144° to 180° were found (Figure 1.3). Inaccuracies in the sampled populations of χ torsion angle were also found (Figure 1.2). These artefacts demanded a more thorough parametrization of torsional energy barriers. Consequently, MP2/6-31G* calculations were performed on larger, untruncated nucleotide models, which included the nitrogenous bases. However, DNA and RNA torsional parameters were not separated. This reparameterization became ff99, which became the basis for all present-day Amber nucleic acid FFs (Figure 1.11). It was satisfactorily able to describe the A $\rightarrow$ B subtype transition in DNA duplexes and triplexes under various conditions.[118]

At the turn of the century, state-of-the-art simulations were on the order of 1-10 ns. As computational power became more widely available, longer simulations revealed additional artefacts. In 2004, Várnai et al performed 50 ns of simulation of a B-DNA double helix and discovered α/γ torsions in the gauche$^+$/trans instead of the experimentally observed g^-/g^+ geometry (Figure 1.10).[119] This prompted Orozco and coworkers (2007) to perform extensive reparameterization of the α and γ torsions by performing a 2D QM surface scan at higher levels of theory, leading to the parmbsc0 FF (named after the Barcelona Supercomputing Centre).[118]

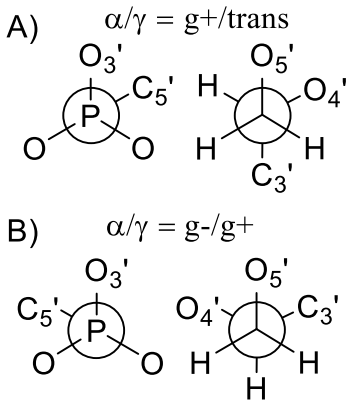


Figure 1.10. Newman projection of the torsional angles, α (left) and γ (right): A) as sampled by parm99 and B) found in nature and sampled by parmbsc0.

The parmbsc0 FF has been used extensively for simulation of both DNA and RNA and was generally robust for low nanosecond simulations. However, advances in computer hardware allowed MD simulations on the order of microseconds, which revealed further deficiencies in the FF.[40, 41, 65, 120, 121] Since its release, research groups around the world have proposed additional modifications. These changes were mostly in the forms of more accurate torsion parameters,[40, 41, 65, 120] although others have also proposed changes to the nonbonded interactions.[121] During this time, FF torsion parameters for DNA and RNA have diverged, due to the biochemical differences between the two. In fact, it was found that further updates to the parameter favorable for RNA simulations lead to issues with DNA.[41, 99, 122]

CHARMM FFs for nucleic acids, on the other hand, have also evolved over the years since the initial release of CHARMM22. As computational power increased over the years, it was realized that CHARMM22 did not reproduce a good balance between A- and B-forms of DNA and RNA.[123] This prompted the optimization and reparameterization of various terms and parameters, which resulted in the newer CHARMM27. This became the basis for all current CHARMM nucleic acid FFs. Like the Amber class of FFs, parameters for DNA and RNA have since diverged due to differences in sugar puckering and solvation due to the 2'-hydroxyl.

OPLS-AA was parametrized for nucleic acid back in the 1990s.[101] While the original functional form was kept, it was also extensively updated recently in a similar manner to both Amber and CHARMM.

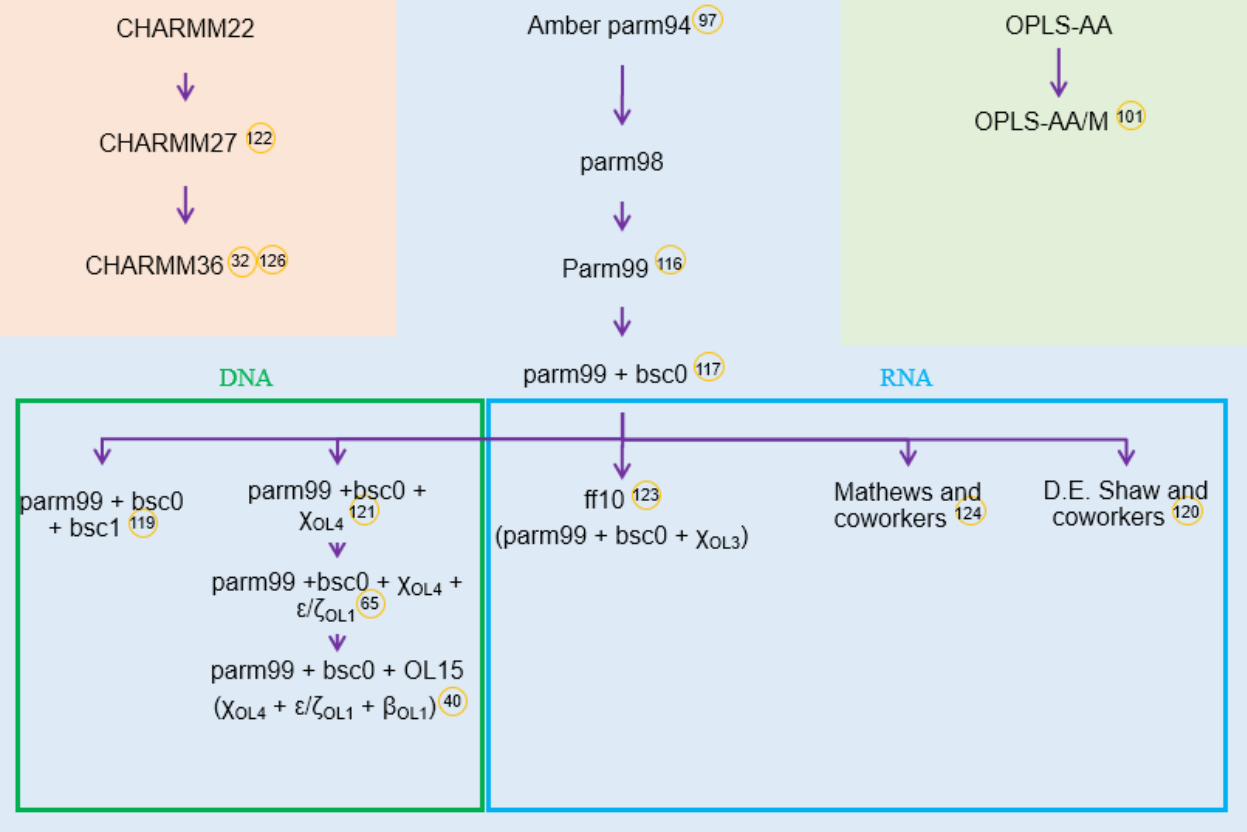


Figure 1.11. Timeline and evolution of modern type I nucleic acid FFs. Selected references are shown in yellow outlines.

1.9 State-of-the-Art DNA Force Fields

Two research groups independently improved the Amber parmbsc0 FF for simulating DNA. Both teams opted to keep the nonbonded interactions of the original ff99. The team led by Jurečka observed that certain noncanonical DNA structures were being modelled poorly using parmbsc0, including DNA G4s and Z-DNAs.[122] To rectify these shortcomings, they performed a reparameterization of the χ torsion. To carry out this task, a higher level of theory was used than previously possible, with the inclusion of implicit solvation. This iteration was named the χ_{OL4} parametrization, which when used together with parmbsc0, improved the description of syn residues, such as parallel and antiparallel DNA G4s. Description of B-DNAs also improved slightly. Although, parmbsc0 with χ_{OL4} correction did not improve Z-DNA description, a follow-up effort in parametrizing the β torsion of DNA addressed this issue.[65] In fact, this

reparameterization, called parmbsc0 β_{OL1}, also bettered the conformational equilibrium of BI/BII and ZI/ZII DNA. In a parallel effort, the ε and ζ torsion parameters were improved by using the same prior approach, which further ameliorated the relative populations of B-I/B-II DNA, through stabilizing the ε/ζ = gauche-/trans geometry.[40] (In fact, this latter strategy was also used for the parametrizing the CHARMM36 parameters for DNA.[34]) A comparison of parmbsc0 with and without $\varepsilon\zeta_{OL1}$ corrections during simulations showed that the use of the latter improved the root-mean-square deviation (RMSD) with respect to the original crystal structure. Taken together, these updated parameters are known as parmbsc0 OL15, and is one of the two FFs currently recommended by the Amber community for simulating DNA.

Concurrently, a team lead by Orozco, also used implicit solvation methods to parametrize various torsional parameters.[120] In particular, new $\delta_{0\text{-}4}$, ε, ζ, and χ torsional parameters were devised, and validated on nearly a hundred distinct DNA molecules, with minimal fraying at the terminal ends, a usual phenomenon in nucleic acid simulations. It also had improved descriptions of noncanonical DNAs, including DNA G4s and hairpin structures. These parameters were combined into parmbsc1, which is the other recommended FF for DNA simulations.

1.10 State-of-the-Art RNA Force Fields

Modelling RNA dynamics is especially challenging, due to its flexible nature. It has been reported that the Amber FFs, ff94, ff99, and parmbsc0, have similar accuracies when employed on RNAs.[41] The application of the latter FF to ribozyme on the order of 50~150 ns was found to unwind the double helix structure into a sense-less ladder structure.[124] This was found to occur mainly due to RNA nucleotides sampling the high-anti region, which was not seen experimentally.[122] Jurečka and coworkers used high-level QM calculations with implicit solvation to reparametrize the χ torsional angle, which was named χ_{OL3}. This modification was found to improve the description of RNA and prevented the ladder-like artifact from occurring. χ_{OL3} was included in Amber ff10. However, its use during DNA simulations had negligible effects.

Using a similar approach, Mathews and coworkers revisited α, β, γ, ε, ζ, and χ torsions of parmbsc0 by simultaneously fitting them using high-level QM calculations with implicit solvation.[125] Umbrella sampling and other methods showed that there was a slight improvement in the population of torsion angles when compared to a set of reference structures from the Protein

Data Bank (PDB).[126] Nevertheless, the simulation of RNA tetramers was found to be improved, whereas previous simulations using FF10 had intercalation artifacts.

Besides the modifications of torsional parameters, a more comprehensive refitting of electrostatic charges and van der Waals interactions in parmbsc0 was also performed by D.E. Shaw and coworkers.[121] This might have been necessary due to the charged nature of nucleic acids, its shielding and polarization effects, which render previous charging schemes less suitable. In fact, these modifications improved the simulation ssRNA, folding and unfolding of RNA duplexes and tetraloops, reversible ligand binding to Guanine Riboswitch, and magnesium-dependent dynamics of SAM-I riboswitches. These FF parameters are available within the standard Amber package.[99]

OPLS-AA was also separately updated for RNA using high-level QM calculations, yielding the specific parameters within the OPLS-AA/M FF. This was a similar strategy to that of Jurečka and coworkers. However, there are minor differences, such as the choice of the reference QM calculations.[101]

Torsional parameters of the 2'-hydroxyl group was also updated in the CHARMM36 FF by Mackerell and coworkers, due to the realization that there was an overemphasis of Watson-Crick base pair opening.[127] QM calculations and subsequent parameter fitting was performed in order to reproduce RNA structural details, including J-couplings, thermodynamic stability, and hydration. However, in general, the smaller scale of validation of RNA FFs compared to DNA, necessitates a more comprehensive testing.

1.11 Treatment of Hydrogen Bonding

In the original implementation of the Amber[128] and CHARMM,[129] hydrogen bonding was a distinct potential energy term, consisting of a Lennard-Jones 12-10 term. However, in later releases, the hydrogen bonding potential energy functions of both FFs were abandoned. Besides reducing the computational costs associated with the extra energy function, Cornell et al. claimed that medium strength hydrogen bonds could be adequately taken into consideration by rolling it into the electrostatics and van der Waals terms.[97] However, these conclusions were based on short, nanoseconds simulations available at the time.

The correct treatment of hydrogen bonding has large structural, dynamic, and ligand-binding implications in nucleic acids. This is not surprising due to the high number of potential

hydrogen bonding sites on nucleotides. Recently, Kührová et al. found that overexaggerated base-phosphate hydrogen bonding interactions resulted in overly stabilized unfolded RNAs, which was an artefact.[130] It also found that ff99bsc0 χ_{OL3} underestimated the base-pairing hydrogen binding potential, which had a propensity for unwinding the RNA, amongst other deficiencies.[110] These findings prompted them to develop a novel hydrogen bonding potential, known as gHBfix.

Evidences also suggest that the treatment of hydrogen bonding in classical FFs, using point charges and Lennard-Jones potential, have reached its limits with regards to their accuracy.[110] In particular, directionality of hydrogen bonding is still lacking in current nucleic acid FFs (Figure 1.12), which was found to be important for other cases, such as scoring the stability of water molecules during DNA/RNA solvation.[46] The fitting procedure during the development of this new FF would be challenging due to our empirical understanding of hydrogen bonding behavior.[131] Nevertheless, improvements in hydrogen bonding in nucleic acid FFs shows promise in improving the accuracy of simulations.

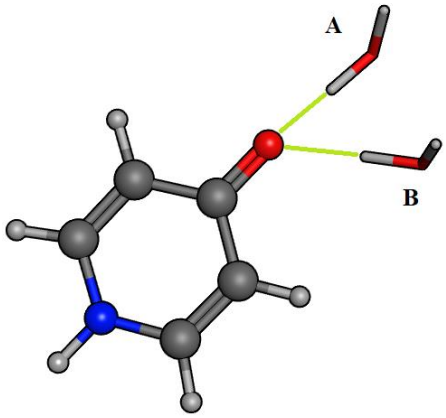


Figure 1.12. Directionality of hydrogen bonding with respect to two water molecules are shown. Although water molecule at position B is expected to be more stable than A, most classical FFs assume same strengths of interaction due to neglect of directionality.

1.12 Treatment of π-π Stacking

Despite the high prevalence of nucleobases capable of π-π stacking, these interactions are not a distinct energy term in current classical FFs (Figure 1.13). Most current FFs treat π-π stacking by using a combination of electrostatics and van der Waals, which approximates this energetic stabilization. However, it is known that significant orbital-orbital interactions are present when

two aromatic rings interact via face-to-face or T-shaped π-π stacking interactions.[132] Due to the nuance interactions of orbitals, current MM methods struggle to accurately predict the preferred orientations or evaluate the energetic stabilization due to base stacking or ligand intercalation.

Attempts at using polarizable FFs, such as AMOEBA, to reproduce the energies of nucleobase stacking at certain geometrical orientations have been investigated.[133] While using multipole expansion and dynamic charge components would be helpful, its accuracy has not been widely validated, either through single point energy calculations or dynamic calculations. Both classical and polarizable FFs neglect to include the charge transfer component, or hyperconjugation effects of π-π stacking interactions, which have previously been found to be significant.[134]

In the future, a more vigorous testing of the ability of FFs to reproduce π-π stacking geometries and energies is required. A careful balance between electrostatic, van der Waals and charge transfer interactions would be desired. Consequently, the addition of a new potential to deal with this charge transfer may be required and improve the overall description of nucleic acids.

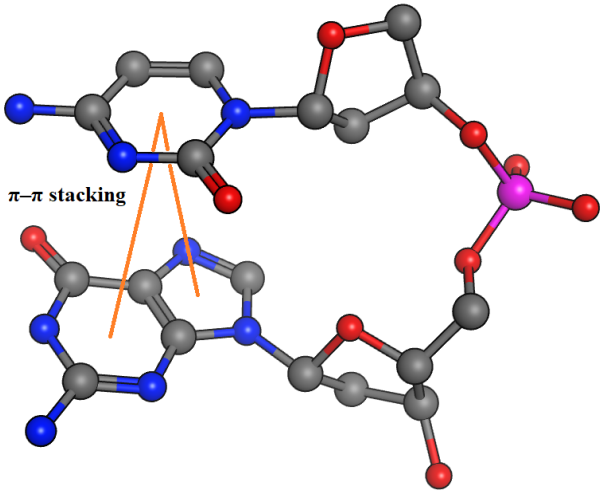


Figure 1.13. π–π stacking interaction present in nucleic acids, which important for base stacking and ligand binding. Offset, rather than T- or Y-shaped π–π stacking, are most abundant in nucleic acids. In FFs, π–π stacking is not explicitly treated, but treated by a combination of electrostatic and van der Waals interactions.

1.13 Treatment of Noncanonical Nucleic Acids and Ligands

Nucleic acid FFs have traditionally focused on the simulation of canonical nucleotides, which are frequently found in nature. However, there exists naturally modified nucleotides and, more recently, artificial nucleotides.[135] In addition, in order to model the dynamics and interaction of these unparametrized nucleotides or the interaction of nucleotides with ligands, FF parameters need to exist for these molecules. Torsions of these nucleotides are expected to differ from those of DNA and RNA, due to biochemical differences, especially hydrogen bonding, electrostatic, steric, and hyperconjugation effects.

From previous sections, it is evident that torsional parameters play a crucial role in DNA and RNA dynamics. Consequently, efforts should also focus on the derivation of specific torsion parameters for modified nucleic acids, as they become increasingly important for treatments and therapeutics. Due to the size of the chemical space, an atom type independent approach for parametrizing glyosidic, sugar, and phosphate torsions should be considered.[82, 113, 136-138] This would be equally applicable to small molecules, which is currently being developed in our laboratory. Previous works on the development of FFs for modified nucleic acids do exist.[139] However, to our knowledge, this effort only focused on the derivation of electrostatic charges, and not torsions. In the future, the ability to simulate all types of nucleic acids would pave the way for designing and making *in silico* predictions of interactions pertinent to various therapies (e.g. interaction of miRNA with its target).

1.14 Future Directions in FF Development and Need for Comprehensive Testing

Due to the empirical nature and simple functional form of classical MM FFs, they are inherently approximate. With increasing computational power and the advent of faster CPUs and GPUs, longer simulations are continuously revealing additional artifacts. Recently, it has been hypothesized that the accuracy limit of classical FFs for simulating nucleic acids is close to being reached, and additional improvements may need to come in the form of more advanced FFs (e.g. polarizable FFs or semiempirical QM methods).

Prior to further development and modification of nucleic acid FFs, an independent and more rigorous test set of nucleic acid structures would be needed. This validation set should be composed of a wide variety of experimental structures (i.e. DNA G4s, RNA Riboswitches, Z-DNA,

and others) under various environmental conditions. This would be especially helpful for validating RNA FFs, since these have not been tested as extensively. Modified amino acids, nucleic acid-ligand, and nucleic acid-protein complexes should also be included. Newer developments in FFs should be tested on this validation set as an objective measure of accuracy.

1.15 Conclusion

Over the past decades, significant progress in the simulation of nucleic acid structures has been made. MM FFs of nucleic acids have come a long way since the 1980s, when only restrained MD simulations were possible. Indeed, the simulation of nucleic acids have been more challenging than proteins, due to the highly charged nature of the former and greater traditional emphasis on the latter. Today, improvements in FFs and in computational power have allowed unrestrained and more accurate simulations of nucleic acids on a timescale of microseconds. Accurate simulations of nucleic acids are important to understand their dynamics, involvement in binding, and interactions. This would be useful for gaining an insight into the atomistic details of nucleic acid-drug interactions, RNA oligonucleotide therapies, and others. The ability for simulated nucleic acids to correctly match known experimental conformations would raise our confidence in them. In the future, this would allow rapid *in silico* predictions toward these previously mentioned applications.

The Amber class of FFs have been the most validated and accurate, to date, for simulating nucleic acids. After decades of development, the FF has now diverged for DNA and RNA as a result of their biochemical differences.[41, 99, 122] These developments were mostly in the form of new torsion parameters, although several groups also tested various nonbonded parameters. For simulations of DNA, two varieties of the Amber FF are recommended: parmbsc1 and parmbsc0 + OL15. For simulations of RNA, validation has not been as thoroughly conducted. However, the FFs from Mathews and coworkers and D.E. Shaw and coworkers have been (self-)reported to accurately reproduce many experimental features. For modelling modified nucleotides, electrostatic parameters do exist for over 100 molecules that are found in nature.[139] However, accurate torsion parameters of these molecules may be lacking due to their nuance chemical differences. Consequently, the use of native torsional parameters, designed for DNA or RNA, would limit the accuracy of MD simulations.

One of the main problems of the field arises from the fact that different research groups use different criteria for judging the accuracy of a FF. A systematic and more objective method should be developed with a range of standard tests for each FF. Future efforts to develop more accurate FFs should first focus on developing a validation set of various DNA and RNA structures for which a wide array of experimental structures exists (e.g. solution-based NMR), similar to the sets used in SAMPL challenges for small molecules. Subsequent FFs could easily test against this validation set to quickly understand their strengths and weaknesses.

1.16 References

1. Blanco, A.; Blanco, G. Chapter 6 - Nucleic Acids. In *Medical Biochemistry*, Blanco, A.; Blanco, G., Eds.; Academic Press: 2017, pp 121-140.

2. Howe, J. A.; Wang, H.; Fischmann, T. O.; Balibar, C. J.; Xiao, L.; Galgoci, A. M.; Malinverni, J. C.; Mayhood, T.; Villafania, A.; Nahvi, A.; Murgolo, N.; Barbieri, C. M.; Mann, P. A.; Carr, D.; Xia, E.; Zuck, P.; Riley, D.; Painter, R. E.; Walker, S. S.; Sherborne, B.; de Jesus, R.; Pan, W.; Plotkin, M. A.; Wu, J.; Rindgen, D.; Cummings, J.; Garlisi, C. G.; Zhang, R.; Sheth, P. R.; Gill, C. J.; Tang, H.; Roemer, T., Selective Small-Molecule Inhibition of an Rna Structural Element. *Nature* **2015**, *526*, 672-677.

3. Palacino, J.; Swalley, S. E.; Song, C.; Cheung, A. K.; Shu, L.; Zhang, X.; Van Hoosear, M.; Shin, Y.; Chin, D. N.; Keller, C. G.; Beibel, M.; Renaud, N. A.; Smith, T. M.; Salcius, M.; Shi, X.; Hild, M.; Servais, R.; Jain, M.; Deng, L.; Bullock, C.; McLellan, M.; Schuierer, S.; Murphy, L.; Blommers, M. J. J.; Blaustein, C.; Berenshteyn, F.; Lacoste, A.; Thomas, J. R.; Roma, G.; Michaud, G. A.; Tseng, B. S.; Porter, J. A.; Myer, V. E.; Tallarico, J. A.; Hamann, L. G.; Curtis, D.; Fishman, M. C.; Dietrich, W. F.; Dales, N. A.; Sivasankaran, R., Smn2 Splice Modulators Enhance U1–Pre-Mrna Association and Rescue Sma Mice. *Nat. Chem. Biol* **2015**, *11*, 511-517.

4. Krammer, F., The Human Antibody Response to Influenza a Virus Infection and Vaccination. *Nature Reviews Immunology* **2019**, *19*, 383-397.

5. Rabi, F. A.; Al Zoubi, M. S.; Kasasbeh, G. A.; Salameh, D. M.; Al-Nasser, A. D., Sars-Cov-2 and Coronavirus Disease 2019: What We Know So Far. *Pathogens* **2020**, *9*, 231.

6. Yang, S. N. Y.; Atkinson, S. C.; Wang, C.; Lee, A.; Bogoyevitch, M. A.; Borg, N. A.; Jans, D. A., The Broad Spectrum Antiviral Ivermectin Targets the Host Nuclear Transport Importin A/B1 Heterodimer. *Antiviral Research* **2020**, *177*, 104760.

7. Gao, C.; Wang, Y., Mrna Metabolism in Cardiac Development and Disease: Life after Transcription. *Physiological Reviews* **2020**, *100*, 673-694.

8. Zhou, Q.; Li, T.; Price, D. H., Rna Polymerase Ii Elongation Control. *Annu. Rev. Biochem* **2012**, *81*, 119-143.

9. Janin, M.; Coll-SanMartin, L.; Esteller, M., Disruption of the Rna Modifications That Target the Ribosome Translation Machinery in Human Cancer. *Molecular Cancer* **2020**, *19*, 70.

10. Wei, W.; Gauld, J. W.; Monard, G., Pretransfer Editing in Threonyl-Trna Synthetase: Roles of Differential Solvent Accessibility and Intermediate Stabilization. *ACS Catal.* **2017**, *7*, 3102-3112.

11. Peng, Y.; Croce, C. M., The Role of Micrornas in Human Cancer. *Signal Transduction and Targeted Therapy* **2016**, *1*, 15004.

12. Colognori, D.; Sunwoo, H.; Kriz, A. J.; Wang, C.-Y.; Lee, J. T., Xist Deletional Analysis Reveals an Interdependency between Xist Rna and Polycomb Complexes for Spreading Along the Inactive X. *Molecular Cell* **2019**, *74*, 101-117.e110.

13. Setten, R. L.; Rossi, J. J.; Han, S.-p., The Current State and Future Directions of Rnai-Based Therapeutics. *Nature Reviews Drug Discovery* **2019**, *18*, 421-446.

14. Ishida, S.; Terasaka, N.; Katoh, T.; Suga, H., An Aminoacylation Ribozyme Evolved from a Natural Trna-Sensing T-Box Riboswitch. *Nat. Chem. Biol* **2020**.

15. Walter, N. G.; Engelke, D. R., Ribozymes: Catalytic Rnas That Cut Things, Make Things, and Do Odd and Useful Jobs. *Biologist (London)* **2002**, *49*, 199-203.

16. Howe, J. A.; Wang, H.; Fischmann, T. O.; Balibar, C. J.; Xiao, L.; Galgoci, A. M.; Malinverni, J. C.; Mayhood, T.; Villafania, A.; Nahvi, A.; Murgolo, N.; Barbieri, C. M.; Mann, P. A.; Carr, D.; Xia, E.; Zuck, P.; Riley, D.; Painter, R. E.; Walker, S. S.; Sherborne, B.; de Jesus, R.; Pan, W.; Plotkin, M. A.; Wu, J.; Rindgen, D.; Cummings, J.; Garlisi, C. G.; Zhang, R.; Sheth, P.

R.; Gill, C. J.; Tang, H.; Roemer, T., Selective Small-Molecule Inhibition of an Rna Structural Element. *Nature* **2015**, *526*, 672-677.

17. Winkler, W.; Nahvi, A.; Breaker, R. R., Thiamine Derivatives Bind Messenger Rnas Directly to Regulate Bacterial Gene Expression. *Nature* **2002**, *419*, 952-956.

18. Mironov, A. S.; Gusarov, I.; Rafikov, R.; Lopez, L. E.; Shatalin, K.; Kreneva, R. A.; Perumov, D. A.; Nudler, E., Sensing Small Molecules by Nascent Rna: A Mechanism to Control Transcription in Bacteria. *Cell* **2002**, *111*, 747-756.

19. Bhagavan, N. V. Chapter 23 - Nucleic Acid Structure and Properties of DNA. In *Medical Biochemistry (Fourth Edition)*, Bhagavan, N. V., Ed.; Academic Press: San Diego, 2002, pp 521-543.

20. Elliott, M. S.; Trewyn, R. W., Inosine Biosynthesis in Transfer Rna by an Enzymatic Insertion of Hypoxanthine. *J. Biol. Chem.* **1984**, *259*, 2407-2410.

21. Shen, L.; Song, C. X.; He, C.; Zhang, Y., Mechanism and Function of Oxidative Reversal of DNA and Rna Methylation. *Annu. Rev. Biochem* **2014**, *83*, 585-614.

22. Liu, N.; Pan, T. Chapter 2 - Rna Epigenetics (Epitranscriptomics). In *Translating Epigenetics to the Clinic*, Laurence, J.; Beusekom, M. V., Eds.; Academic Press: Boston, 2017, pp 19-35.

23. Leontis, N. B.; Westhof, E., Geometric Nomenclature and Classification of Rna Base Pairs. *RNA (New York, N.Y.)* **2001**, *7*, 499-512.

24. Matsuyama, S.; Ueda, T.; Crain, P. F.; McCloskey, J. A.; Watanabe, K., A Novel Wobble Rule Found in Starfish Mitochondria. Presence of 7-Methylguanosine at the Anticodon Wobble Position Expands Decoding Capability of Trna. *J. Biol. Chem.* **1998**, *273*, 3363-3368.

25. Eun, H.-M. 1 - Enzymes and Nucleic Acids: General Principles. In *Enzymology Primer for Recombinant DNA Technology*, Eun, H.-M., Ed.; Academic Press: San Diego, 1996, pp 1-108.

26. Altona, C.; Sundaralingam, M., Conformational Analysis of the Sugar Ring in Nucleosides and Nucleotides. New Description Using the Concept of Pseudorotation. *Journal of the American Chemical Society* **1972**, *94*, 8205-8212.

27. Malek-Adamian, E.; Guenther, D. C.; Matsuda, S.; Martínez-Montero, S.; Zlatev, I.; Harp, J.; Burai Patrascu, M.; Foster, D. J.; Fakhoury, J.; Perkins, L.; Moitessier, N.; Manoharan, R. M.; Taneja, N.; Bisbe, A.; Charisse, K.; Maier, M.; Rajeev, K. G.; Egli, M.; Manoharan, M.; Damha, M. J., 4′-C-Methoxy-2′-Deoxy-2′-Fluoro Modified Ribonucleotides Improve Metabolic Stability and Elicit Efficient Rnai-Mediated Gene Silencing. *Journal of the American Chemical Society* **2017**, *139*, 14542-14555.

28. Burai Patrascu, M.; Malek-Adamian, E.; Damha, M. J.; Moitessier, N., Accurately Modeling the Conformational Preferences of Nucleosides. *Journal of the American Chemical Society* **2017**, *139*, 13620-13623.

29. Huang, M.; Giese, T. J.; Lee, T.-S.; York, D. M., Improvement of DNA and Rna Sugar Pucker Profiles from Semiempirical Quantum Methods. *J. Chem. Theory Comput.* **2014**, *10*, 1538-1545.

30. Pastor, N., The B- to a-DNA Transition and the Reorganization of Solvent at the DNA Surface. *Biophys. J.* **2005**, *88*, 3262-3275.

31. Li, L.; Szostak, J. W., The Free Energy Landscape of Pseudorotation in 3′–5′ and 2′–5′ Linked Nucleic Acids. *Journal of the American Chemical Society* **2014**, *136*, 2858-2865.

32. Suresh, G.; Priyakumar, U. D., Structures, Dynamics, and Stabilities of Fully Modified Locked Nucleic Acid (B-D-Lna and A-L-Lna) Duplexes in Comparison to Pure DNA and Rna Duplexes. *The Journal of Physical Chemistry B* **2013**, *117*, 5556-5564.

33. Yan, S.; Li, X.; Zhang, P.; Wang, Y.; Chen, H.-Y.; Huang, S.; Yu, H., Direct Sequencing of 2′-Deoxy-2′-Fluoroarabinonucleic Acid (Fana) Using Nanopore-Induced Phase-Shift Sequencing (Nipss). *Chemical Science* **2019**, *10*, 3110-3117.

34. Hart, K.; Foloppe, N.; Baker, C. M.; Denning, E. J.; Nilsson, L.; MacKerell, A. D., Optimization of the Charmm Additive Force Field for DNA: Improved Treatment of the Bi/Bii Conformational Equilibrium. *J. Chem. Theory Comput.* **2012**, *8*, 348-362.

35. Zhang, N.; Zhang, S.; Szostak, J. W., Activated Ribonucleotides Undergo a Sugar Pucker Switch Upon Binding to a Single-Stranded Rna Template. *Journal of the American Chemical Society* **2012**, *134*, 3691-3694.

36. Kellas, B. L.; Austoker, J. L.; Beebee, T. J.; Butterworth, P. H., Forms Ai and Aii DNA-Dependent Rna Polymerases as Components of Two Defined Pools of Polymerase Activity in Mammalian Cells. *Eur. J. Biochem.* **1977**, *72*, 583-594.

37. Trieb, M.; Rauch, C.; Wellenzohn, B.; Wibowo, F.; Loerting, T.; Liedl, K. R., Dynamics of DNA: Bi and Bii Phosphate Backbone Transitions. *The Journal of Physical Chemistry B* **2004**, *108*, 2470-2476.

38. Sims, G. E.; Kim, S. H., Global Mapping of Nucleic Acid Conformational Space: Dinucleoside Monophosphate Conformations and Transition Pathways among Conformational Classes. *Nucleic Acids Res.* **2003**, *31*, 5607-5616.

39. Coimbatore Narayanan, B.; Westbrook, J.; Ghosh, S.; Petrov, A. I.; Sweeney, B.; Zirbel, C. L.; Leontis, N. B.; Berman, H. M., The Nucleic Acid Database: New Features and Capabilities. *Nucleic Acids Res.* **2014**, *42*, D114-D122.

40. Zgarbová, M.; Luque, F. J.; Šponer, J.; Cheatham, T. E.; Otyepka, M.; Jurečka, P., Toward Improved Description of DNA Backbone: Revisiting Epsilon and Zeta Torsion Force Field Parameters. *J. Chem. Theory Comput.* **2013**, *9*, 2339-2354.

41. Krepl, M.; Zgarbová, M.; Stadlbauer, P.; Otyepka, M.; Banáš, P.; Koča, J.; Cheatham, T. E.; Jurečka, P.; Šponer, J., Reference Simulations of Noncanonical Nucleic Acids with Different X Variants of the Amber Force Field: Quadruplex DNA, Quadruplex Rna, and Z-DNA. *J. Chem. Theory Comput.* **2012**, *8*, 2506-2520.

42. Donlic, A.; Hargrove, A. E., Targeting Rna in Mammalian Systems with Small Molecules. *Wiley interdisciplinary reviews. RNA* **2018**, *9*, e1477-e1477.

43. Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M., Drugbank 5.0: A Major Update to the Drugbank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074-D1082.

44. Chittapragada, M.; Roberts, S.; Ham, Y. W., Aminoglycosides: Molecular Insights on the Recognition of Rna and Aminoglycoside Mimics. *Perspect Medicin Chem* **2009**, *3*, 21-37.

45. Cross, R., The Rna Drug Hunters Academics, Biotech Start-Ups, and Big Pharma Companies Are Designing Small Molecules That Target Rna. *Chemical & Engineering News* **2017**, *95*, 16-18.

46. Wei, W.; Luo, J.; Waldispühl, J.; Moitessier, N., Predicting Positions of Bridging Water Molecules in Nucleic Acid–Ligand Complexes. *J. Chem. Inf. Model* **2019**, *59*, 2941-2951.

47. Piret, J.; Boivin, G., Resistance of Herpes Simplex Viruses to Nucleoside Analogues: Mechanisms, Prevalence, and Management. *Antimicrob. Agents Chemother.* **2011**, *55*, 459-472.

48. Hridya, V. M.; Hynes, J. T.; Mukherjee, A., Dynamical Recrossing in the Intercalation Process of the Anticancer Agent Proflavine into DNA. *The Journal of Physical Chemistry B* **2019**, *123*, 10904-10914.

49. Gatasheh, M. K.; Kannan, S.; Hemalatha, K.; Imrana, N., Proflavine an Acridine DNA Intercalating Agent and Strong Antimicrobial Possessing Potential Properties of Carcinogen. *Karbala International Journal of Modern Science* **2017**, *3*, 272-278.

50. Han, H.; Hurley, L. H., G-Quadruplex DNA: A Potential Target for Anti-Cancer Drug Design. *Trends in Pharmacological Sciences* **2000**, *21*, 136-142.

51. Shammas, M. A., Telomeres, Lifestyle, Cancer, and Aging. *Curr Opin Clin Nutr Metab Care* **2011**, *14*, 28-34.

52. Jafri, M. A.; Ansari, S. A.; Alqahtani, M. H.; Shay, J. W., Roles of Telomeres and Telomerase in Cancer, and Advances in Telomerase-Targeted Therapies. *Genome Med* **2016**, *8*, 69-69.

53. Zhou, K.; Liu, J.; Xiong, X.; Cheng, M.; Hu, X.; Narva, S.; Zhao, X.; Wu, Y.; Zhang, W., Design, Synthesis of 4,5-Diazafluorene Derivatives and Their Anticancer Activity Via Targeting Telomeric DNA G-Quadruplex. *Eur. J. Med. Chem.* **2019**, *178*, 484-499.

54. Luo, J.; Wei, W.; Waldispühl, J.; Moitessier, N., Challenges and Current Status of Computational Methods for Docking Small Molecules to Nucleic Acids. *European Journal of Medicinal Chemistry* **2019**, *168*, 414-425.

55. Castor, K. J.; Liu, Z.; Fakhoury, J.; Hancock, M. A.; Mittermaier, A.; Moitessier, N.; Sleiman, H. F., A Platinum(Ii) Phenylphenanthroimidazole with an Extended Side-Chain Exhibits Slow Dissociation from a C-Kit G-Quadruplex Motif. *Chemistry – A European Journal* **2013**, *19*, 17836-17845.

56. Carvalho, J.; Quintela, T.; Gueddouda, N. M.; Bourdoncle, A.; Mergny, J.-L.; Salgado, G. F.; Queiroz, J. A.; Cruz, C., Phenanthroline Polyazamacrocycles as G-Quadruplex DNA Binders. *Organic & Biomolecular Chemistry* **2018**, *16*, 2776-2786.

57. Saini, N.; Sterling, J. F.; Sakofsky, C. J.; Giacobone, C. K.; Klimczak, L. J.; Burkholder, A. B.; Malc, E. P.; Mieczkowski, P. A.; Gordenin, D. A., Mutation Signatures Specific to DNA Alkylating Agents in Yeast and Cancers. *Nucleic Acids Res.* **2020**, *48*, 3692-3707.

58. Vakulskas, C. A.; Dever, D. P.; Rettig, G. R.; Turk, R.; Jacobi, A. M.; Collingwood, M. A.; Bode, N. M.; McNeill, M. S.; Yan, S.; Camarena, J.; Lee, C. M.; Park, S. H.; Wiebking, V.; Bak, R. O.; Gomez-Ospina, N.; Pavel-Dinu, M.; Sun, W.; Bao, G.; Porteus, M. H.; Behlke, M. A., A High-Fidelity Cas9 Mutant Delivered as a Ribonucleoprotein Complex Enables Efficient Gene Editing in Human Hematopoietic Stem and Progenitor Cells. *Nat Med* **2018**, *24*, 1216-1224.

59. Heidenreich, M.; Zhang, F., Applications of Crispr-Cas Systems in Neuroscience. *Nat Rev Neurosci* **2016**, *17*, 36-44.

60. Hendel, A.; Bak, R. O.; Clark, J. T.; Kennedy, A. B.; Ryan, D. E.; Roy, S.; Steinfeld, I.; Lunstad, B. D.; Kaiser, R. J.; Wilkens, A. B.; Bacchetta, R.; Tsalenko, A.; Dellinger, D.; Bruhn, L.; Porteus, M. H., Chemically Modified Guide Rnas Enhance Crispr-Cas Genome Editing in Human Primary Cells. *Nat. Biotechnol.* **2015**, *33*, 985-989.

61. Egli, M.; Portmann, S.; Usman, N., Rna Hydration: A Detailed Look. *Biochemistry* **1996**, *35*, 8489-8494.

62. Auffinger, P.; Westhof, E., Hydration of Rna Base Pairs. *J. Biomol. Struct. Dyn.* **1998**, *16*, 693-707.

63. Luo, J.; Wei, W.; Waldispuhl, J.; Moitessier, N., Challenges and Current Status of Computational Methods for Docking Small Molecules to Nucleic Acids. *Eur. J. Med. Chem.* **2019**, *168*, 414-425.

64. Beveridge, D. L.; Dixit, S. B.; Barreiro, G.; Thayer, K. M., Molecular Dynamics Simulations of DNA Curvature and Flexibility: Helix Phasing and Premelting. *Biopolymers* **2004**, *73*, 380-403.

65. Zgarbová, M.; Šponer, J.; Otyepka, M.; Cheatham, T. E.; Galindo-Murillo, R.; Jurečka, P., Refinement of the Sugar–Phosphate Backbone Torsion Beta for Amber Force Fields Improves the Description of Z- and B-DNA. *J. Chem. Theory Comput.* **2015**, *11*, 5723-5736.

66. Jayaram, B.; Sprous, D.; Young, M. A.; Beveridge, D. L., Free Energy Analysis of the Conformational Preferences of a and B Forms of DNA in Solution. *Journal of the American Chemical Society* **1998**, *120*, 10629-10633.

67. Khesbak, H.; Savchuk, O.; Tsushima, S.; Fahmy, K., The Role of Water H-Bond Imbalances in B-DNA Substate Transitions and Peptide Recognition Revealed by Time-Resolved Ftir Spectroscopy. *Journal of the American Chemical Society* **2011**, *133*, 5834-5842.

68. Jayaram, B.; Jain, T., The Role of Water in Protein-DNA Recognition. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 343-361.

69. Jorgensen, W. L.; Thomas, L. L., Perspective on Free-Energy Perturbation Calculations for Chemical Equilibria. *J. Chem. Theory Comput.* **2008**, *4*, 869-876.

70. de Ruiter, A.; Oostenbrink, C., Advances in the Calculation of Binding Free Energies. *Current Opinion in Structural Biology* **2020**, *61*, 207-212.

71. Limongelli, V., Ligand Binding Free Energy and Kinetics Calculation in 2020. *WIREs Computational Molecular Science n/a*, e1455.

72. Jorgensen, W. L.; Ravimohan, C., Monte Carlo Simulation of Differences in Free Energies of Hydration. *The Journal of Chemical Physics* **1985**, *83*, 3050-3054.

73. Jing, Z.; Qi, R.; Thibonnier, M.; Ren, P., Molecular Dynamics Study of the Hybridization between Rna and Modified Oligonucleotides. *J. Chem. Theory Comput.* **2019**, *15*, 6422-6432.

74. Xu, Y.; Villa, A.; Nilsson, L., The Free Energy of Locking a Ring: Changing a Deoxyribonucleoside to a Locked Nucleic Acid. *J. Comput. Chem.* **2017**, *38*, 1147-1157.

75. Bren, U.; Martínek, V.; Florián, J., Decomposition of the Solvation Free Energies of Deoxyribonucleoside Triphosphates Using the Free Energy Perturbation Method. *The Journal of Physical Chemistry B* **2006**, *110*, 12782-12788.

76. Irwin, B. W. J.; Huggins, D. J., Estimating Atomic Contributions to Hydration and Binding Using Free Energy Perturbation. *J. Chem. Theory Comput.* **2018**, *14*, 3218-3227.

77. Sund, J.; Lind, C.; Åqvist, J., Binding Site Preorganization and Ligand Discrimination in the Purine Riboswitch. *The Journal of Physical Chemistry B* **2015**, *119*, 773-782.

78. Delgado Blanco, J.; Radusky, L. G.; Cianferoni, D.; Serrano, L., Protein-Assisted Rna Fragment Docking (Rnax) for Modeling Rna–Protein Interactions Using Modelx. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 24568-24573.

79. He, J.; Tao, H.; Huang, S.-Y., Protein-Ensemble–Rna Docking by Efficient Consideration of Protein Flexibility through Homology Models. *Bioinformatics* **2019**, *35*, 4994-5002.

80. Arnautova, Y. A.; Abagyan, R.; Totrov, M., Protein-Rna Docking Using Icm. *J. Chem. Theory Comput.* **2018**, *14*, 4971-4984.

81. Lang, P. T.; Brozell, S. R.; Mukherjee, S.; Pettersen, E. F.; Meng, E. C.; Thomas, V.; Rizzo, R. C.; Case, D. A.; James, T. L.; Kuntz, I. D., Dock 6: Combining Techniques to Model Rna-Small Molecule Complexes. *RNA (New York, N.Y.)* **2009**, *15*, 1219-1230.

82. Corbeil, C. R.; Englebienne, P.; Yannopoulos, C. G.; Chan, L.; Das, S. K.; Bilimoria, D.; L'Heureux, L.; Moitessier, N., Docking Ligands into Flexible and Solvated Macromolecules. 2. Development and Application of Fitted 1.5 to the Virtual Screening of Potential Hcv Polymerase Inhibitors. *J. Chem. Inf. Model* **2008**, *48*, 902-909.

83. Alam, S.; Khan, F., Virtual Screening, Docking, Admet and System Pharmacology Studies on Garcinia Caged Xanthone Derivatives for Anticancer Activity. *Scientific Reports* **2018**, *8*, 5524.

84. Aboelnga, M. M.; Wetmore, S. D., Unveiling a Single-Metal-Mediated Phosphodiester Bond Cleavage Mechanism for Nucleic Acids: A Multiscale Computational Investigation of a Human DNA Repair Enzyme. *Journal of the American Chemical Society* **2019**, *141*, 8646-8656.

85. Świderek, K.; Marti, S.; Tuñón, I.; Moliner, V.; Bertran, J., Peptide Bond Formation Mechanism Catalyzed by Ribosome. *Journal of the American Chemical Society* **2015**, *137*, 12024-12034.

86. Aboelnga, M. M.; Gauld, J. W., Roles of the Active Site Zn(Ii) and Residues in Substrate Discrimination by Threonyl-Trna Synthetase: An Md and Qm/Mm Investigation. *The Journal of Physical Chemistry B* **2017**, *121*, 6163-6174.

87. Aboelnga, M. M.; Hayward, J. J.; Gauld, J. W., Unraveling the Critical Role Played by Ado762′Oh in the Post-Transfer Editing by Archaeal Threonyl-Trna Synthetase. *The Journal of Physical Chemistry B* **2018**, *122*, 1092-1101.

88. Aboelnga, M. M.; Hayward, J. J.; Gauld, J. W., Enzymatic Post-Transfer Editing Mechanism of E. Coli Threonyl-Trna Synthetase (Thrrs): A Molecular Dynamics (Md) and Quantum Mechanics/Molecular Mechanics (Qm/Mm) Investigation. *ACS Catal.* **2017**, *7*, 5180-5193.

89. Fortowsky, G. B.; Simard, D. J.; Aboelnga, M. M.; Gauld, J. W., Substrate-Assisted and Enzymatic Pretransfer Editing of Nonstandard Amino Acids by Methionyl-Trna Synthetase. *Biochemistry* **2015**, *54*, 5757-5765.

90. Wang, B.; Cao, Z.; Sharon, D. A.; Shaik, S., Computations Reveal a Rich Mechanistic Variation of Demethylation of N-Methylated DNA/Rna Nucleotides by Fto. *ACS Catal.* **2015**, *5*, 7077-7090.

91. Dawson, W. K.; Maciejczyk, M.; Jankowska, E. J.; Bujnicki, J. M., Coarse-Grained Modeling of Rna 3d Structure. *Methods* **2016**, *103*, 138-156.

92. Tarenzi, T.; Calandrini, V.; Potestio, R.; Carloni, P., Open-Boundary Molecular Mechanics/Coarse-Grained Framework for Simulations of Low-Resolution G-Protein-Coupled Receptor–Ligand Complexes. *J. Chem. Theory Comput.* **2019**, *15*, 2101-2109.

93. Auffinger, P.; Westhof, E., Hydration of Rna Base Pairs. *Journal of Biomolecular Structure & Dynamics* **1998**, *16*, 693-707.

94. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26*, 1668-1688.

95. 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.

96. Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M., Charmm: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545-1614.

97. 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 (Vol 117, Pg 5179, 1995). *Journal of the American Chemical Society* **1996**, *118*, 2309-2309.

98. 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. *Journal of the American Chemical Society* **1996**, *118*, 11225-11236.

99. D.A. Case, K. B., I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz,R Harris, S. Izadi, K. Kasava- jhala, A. Kovalenko, R. Krasny, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F. Pan, S. Pantano, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, L. Wilson, R.M. Wolf, X. Wu, D.M. York and P.A. Kollman, Amber 2020. *University of California, San Francisco* **2020**.

100. Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L., Improved Treatment of Nucleosides and Nucleotides in the Opls-Aa Force Field. *Chem. Phys. Lett.* **2017**, *683*, 276-280.

101. 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.

102. Dans, P. D.; Ivani, I.; Hospital, A.; Portella, G.; Gonzalez, C.; Orozco, M., How Accurate Are Accurate Force-Fields for B-DNA? *Nucleic Acids Res.* **2017**, *45*, 4217-4230.

103. Minhas, V.; Sun, T.; Mirzoev, A.; Korolev, N.; Lyubartsev, A. P.; Nordenskiöld, L., Modeling DNA Flexibility: Comparison of Force Fields from Atomistic to Multiscale Levels. *The Journal of Physical Chemistry B* **2020**, *124*, 38-49.

104. Ponder, J. W.; Wu, C.; Ren, P.; 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. *The Journal of Physical Chemistry B* **2010**, *114*, 2549-2564.

105. Inakollu, V. S. S.; Geerke, D. P.; Rowley, C. N.; Yu, H., Polarisable Force Fields: What Do They Add in Biomolecular Simulations? *Current Opinion in Structural Biology* **2020**, *61*, 182-190.

106. 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.

107. Lemkul, J. A.; MacKerell Jr., A. D., Polarizable Force Field for Rna Based on the Classical Drude Oscillator. *J. Comput. Chem.* **2018**, *39*, 2624-2646.

108. Baker, C. M.; Anisimov, V. M.; MacKerell, A. D., Development of Charmm Polarizable Force Field for Nucleic Acid Bases Based on the Classical Drude Oscillator Model. *The Journal of Physical Chemistry B* **2011**, *115*, 580-596.

109. Lemkul, J. A.; MacKerell, A. D., Polarizable Force Field for DNA Based on the Classical Drude Oscillator: I. Refinement Using Quantum Mechanical Base Stacking and Conformational Energetics. *J. Chem. Theory Comput.* **2017**, *13*, 2053-2071.

110. Kührová, P.; Mlýnský, V.; Zgarbová, M.; Krepl, M.; Bussi, G.; Best, R. B.; Otyepka, M.; Šponer, J.; Banáš, P., Improving the Performance of the Amber Rna Force Field by Tuning the Hydrogen-Bonding Interactions. *J. Chem. Theory Comput.* **2019**, *15*, 3288-3305.

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

112. Dahlgren, M. K.; Schyman, P.; Tirado-Rives, J.; Jorgensen, W. L., Characterization of Biaryl Torsional Energetics and Its Treatment in Opls All-Atom Force Fields. *J. Chem. Inf. Model* **2013**, *53*, 1191-1199.

113. Wei, W. L.; Champion, C.; Liu, Z. M.; Barigye, S. J.; Labute, P.; Moitessier, N., Torsional Energy Barriers of Biaryls Could Be Predicted by Electron Richness/Deficiency of Aromatic Rings; Advancement of Molecular Mechanics toward Atom-Type Independence. *J. Chem. Inf. Model* **2019**, *59*, 4764-4777.

114. 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.

115. Feig, M.; Pettitt, B. M., Structural Equilibrium of DNA Represented with Different Force Fields. *Biophys. J.* **1998**, *75*, 134-149.

116. Feig, M.; Pettitt, B. M., Experiment Vs Force Fields: DNA Conformation from Molecular Dynamics Simulations. *The Journal of Physical Chemistry B* **1997**, *101*, 7361-7363.

117. Cheatham, T. E., 3rd; Cieplak, P.; Kollman, P. A., A Modified Version of the Cornell Et Al. Force Field with Improved Sugar Pucker Phases and Helical Repeat. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845-862.

118. 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.

119. Várnai, P.; Zakrzewska, K., DNA and Its Counterions: A Molecular Dynamics Study. *Nucleic Acids Res.* **2004**, *32*, 4269-4280.

120. Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A.; Portella, G.; Battistini, F.; Gelpí, J. L.; González, C.; Vendruscolo, M.; Laughton, C. A.; Harris, S. A.; Case, D. A.; Orozco, M., Parmbsc1: A Refined Force Field for DNA Simulations. *Nature Methods* **2016**, *13*, 55-58.

121. Tan, D.; Piana, S.; Dirks, R. M.; Shaw, D. E., Rna Force Field with Accuracy Comparable to State-of-the-Art Protein Force Fields. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, E1346-E1355.

122. Zgarbová, M.; Otyepka, M.; Sponer, J.; Mládek, A.; Banáš, P.; Cheatham, T. E., 3rd; Jurečka, 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.

123. Foloppe, N.; MacKerell, J., Alexander D., All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data. *J. Comput. Chem.* **2000**, *21*, 86-104.

124. Mlýnský, V.; Banáš, P.; Hollas, D.; Réblová, K.; Walter, N. G.; Šponer, J.; Otyepka, M., Extensive Molecular Dynamics Simulations Showing That Canonical G8 and Protonated A38h+ Forms Are Most Consistent with Crystal Structures of Hairpin Ribozyme. *The Journal of Physical Chemistry B* **2010**, *114*, 6642-6652.

125. Aytenfisu, A. H.; Spasic, A.; Grossfield, A.; Stern, H. A.; Mathews, D. H., Revised Rna Dihedral Parameters for the Amber Force Field Improve Rna Molecular Dynamics. *J. Chem. Theory Comput.* **2017**, *13*, 900-915.

126. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.

127. Denning, E. J.; Priyakumar, U. D.; Nilsson, L.; Mackerell, A. D., Jr., Impact of 2'-Hydroxyl Sampling on the Conformational Properties of Rna: Update of the Charmm All-Atom Additive Force Field for Rna. *J. Comput. Chem.* **2011**, *32*, 1929-1943.

128. 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. *Journal of the American Chemical Society* **1984**, *106*, 765-784.

129. 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.

130. Kührová, P.; Best, R. B.; Bottaro, S.; Bussi, G.; Šponer, J.; Otyepka, M.; Banáš, P., Computer Folding of Rna Tetraloops: Identification of Key Force Field Deficiencies. *J. Chem. Theory Comput.* **2016**, *12*, 4534-4548.

131. Platts, J. A.; Howard, S. T.; Bracke, B. R. F., Directionality of Hydrogen Bonds to Sulfur and Oxygen. *Journal of the American Chemical Society* **1996**, *118*, 2726-2733.

132. Zhao, R.; Zhang, R.-Q., A New Insight into Π–Π Stacking Involving Remarkable Orbital Interactions. *Physical Chemistry Chemical Physics* **2016**, *18*, 25452-25457.

133. Zhang, C.; Bell, D.; Harger, M.; Ren, P., Polarizable Multipole-Based Force Field for Aromatic Molecules and Nucleobases. *J. Chem. Theory Comput.* **2017**, *13*, 666-678.

134. Schneider, W. B.; Bistoni, G.; Sparta, M.; Saitow, M.; Riplinger, C.; Auer, A. A.; Neese, F., Decomposition of Intermolecular Interaction Energies within the Local Pair Natural Orbital Coupled Cluster Framework. *J. Chem. Theory Comput.* **2016**, *12*, 4778-4792.

135. Hoshika, S.; Leal, N. A.; Kim, M.-J.; Kim, M.-S.; Karalkar, N. B.; Kim, H.-J.; Bates, A. M.; Watkins, N. E.; SantaLucia, H. A.; Meyer, A. J.; DasGupta, S.; Piccirilli, J. A.; Ellington, A. D.; SantaLucia, J.; Georgiadis, M. M.; Benner, S. A., Hachimoji DNA and Rna: A Genetic System with Eight Building Blocks. *Science* **2019**, *363*, 884-887.

136. Champion, C.; Barigye, S. J.; Wei, W.; Liu, Z.; Labute, P.; Moitessier, N., Atom Type Independent Modeling of the Conformational Energy of Benzylic, Allylic, and Other Bonds Adjacent to Conjugated Systems. *J. Chem. Inf. Model* **2019**.

137. Liu, Z.; 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.

138. Mobley, D. L.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Slochower, D. R.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K., Escaping

Atom Types in Force Fields Using Direct Chemical Perception. *J. Chem. Theory Comput.* **2018**, *14*, 6076-6092.

139. Aduri, R.; Psciuk, B. T.; Saro, P.; Taniga, H.; Schlegel, H. B.; SantaLucia, J., Amber Force Field Parameters for the Naturally Occurring Modified Nucleosides in Rna. *J. Chem. Theory Comput.* **2007**, *3*, 1464-1475.

2 Predicting Positions of Bridging Water Molecules in Nucleic Acid–Ligand Complexes

2.1 Preface

As briefly touched upon in **Chapter 1**, pharmaceuticals targeting nucleic acids are promising for treating a variety of different diseases. Accurate predictions of nucleic acid binding through *in silico* docking would be useful during the initial drug discovery phase to enrich compound libraries, quickly and inexpensively. Unfortunately, due to the biochemical differences between nucleic acids and proteins, existing docking methods, often designed for the latter, are often inaccurate when applied to the former. This chapter addresses a key shortcoming in contemporary docking programs in modelling ligands binding to nucleic acids: the presence and variability of bridging water molecules and water networks surrounding the ligand. To this end, a new method, for the placement of water molecules in the ligand binding site of nucleic acids was developed. This method places water molecules, based on collected statistics and a newly developed and specialized force field. In the future, this water placement method could be incorporated into a docking program, along with a specialized scoring function capable of considering these placed water molecules. This could potentially improve the enrichment capabilities of docking programs for identifying potential drug candidates.

This chapter is based on work from: Wei, W.; Luo, J.; Waldispühl, J.; Moitessier, N., Predicting Positions of Bridging Water Molecules in Nucleic Acid–Ligand Complexes. J. Chem. Inf. Model 2019, 59, 2941-2951.

Author contributions: I performed the calculations, analysis, and write-up of the manuscript. Prof. Nicolas Moitessier and I were involved in the design of the methods described. Jiaying Luo and Prof. Jérôme Waldispühl gave helpful suggestions, which were incorporated in the experiment.

2.2 Introduction

While proteins have traditionally been the primary target for drug development, interests in nucleic acids as promising targets for drugs and molecular probes has been more recent.[1] The realization that targeting nucleic acids would allow the treatment of previously untreatable diseases has led pharmaceutical companies and academic groups to invest heavily in this area of research, colloquially known as the "RNA gold rush."[2] Two landmark discoveries – Ribocil, an antibiotic agent recently discovered by Merck,[3] and Branaplam, a drug currently undergoing phase II clinical trials by Novartis for spinal muscular atrophy[2, 4] – have shown that drugs targeting nucleic acids could be both effective and selective to their targets.[2]

Ribocil, in particular, was shown to inhibit bacterial cell growth and bind to a regulatory region of a metabolically essential mRNA transcript.[3] These non-coding regions, located in the upstream 5′-untranslated regions (UTRs), called RNA riboswitches, are often involved in the negative feedback loop of bacterial biosynthesis pathways of life-sustaining endogenous metabolites.[5] Riboswitches are attractive targets for antibacterial agents as they are not present in mammalian cells. Consequently, small molecules targeting riboswitches, are a desperately needed repertoire of antibiotics, especially in the current era of increasing antibiotic resistance around the world. Aminoglycosides exhibit favorable nucleic acid binding properties (eg. positively charged ammonium groups), which make them potent antibiotics against bacterial nucleic acids, along with other promising methods such as siRNA and antisense therapy.[5] In general, a lower dosage of drugs is sufficient when targeting mRNA over proteins, mitigating the effects of drug toxicity (one mRNA gives rise to many copies of the protein). Recently, viral RNA has also been targeted, as reported by scientists from the Scripps Research Institute.[6]

DNA structural elements have also gained attention as attractive targets for drugs, such as Cisplatin, which crosslinks with DNA bases and induces cell death in various types of cancers,[7] intercalating agents (e.g., Doxorubicin), alkylating agents (e.g., Mechlorethamine), and more recently G-quadruplex binders. The binding and stabilization of the latter by small molecules have been shown to prevent hyperactive telomerases from repairing the ends of chromosomes in cancer cells.[8] This inhibition prevents cancer cells from being immortal, and is a promising area of drug discovery.

Nucleic acids, in contrast to proteins, possess shallow, water-permeable binding pockets. In addition, phosphate groups of nucleic acids and their corresponding countercations (e.g., Mg^{2+},

K^+) are highly charged and induce polarization upon both nearby water molecules and functional groups of drugs, significantly affecting the nucleic acid-drug complex stability. Modeling of these highly polarized molecules has been identified as a source of error during the simulation of biomacromolecules.[9, 10] Past molecular dynamics (MD) simulations of RNA have suggested that some positions near the nucleic acid have long retention times for water molecules.[11] Thus, water molecules are expected to play a critical role in nucleic acid-drug complex formation as illustrated in Figure 2.1.

Several medicinal chemistry approaches have been envisioned to develop protein binding drugs. Amongst these, structure-based drug design has become an indispensable strategy for medicinal chemists. In fact, docking libraries of small molecules to proteins was instrumental in discovering potential drugs and several docking programs have been reported since DOCK was first reported and released.

While docking methods for proteins have made significant advances, the current computational drug discovery methodologies targeting nucleic acid are inadequate. While docking libraries of small molecules to protein crystal structures has been instrumental in the discovery of potential drugs, nucleic acids have been largely overlooked by developers of docking programs with a few exceptions (e.g., RiboDock,[12] Mordor[13] and rDOCK[14]). In other situations, programs, initially developed for proteins, were retrained to work with RNA (ICM,[15] DOCK6 [16, 17]).[18] Due to the previously stated biochemical differences between nucleic acids and proteins, accurate placement of water molecules in the binding site is often crucial for successful *in silico* docking of ligands to nucleic acids. For a recent review on the field of docking to nucleic acids and the associated challenges, readers are referred to our recent review.[18]

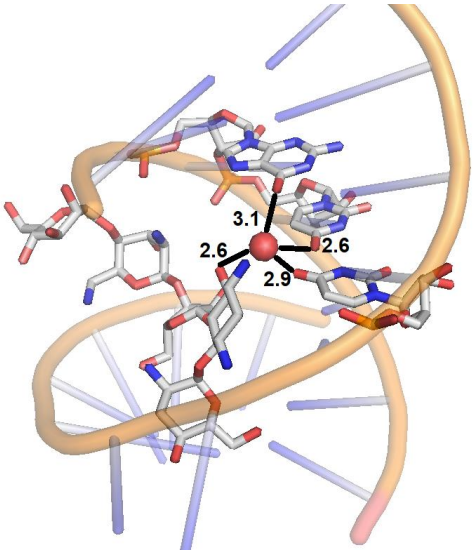


Figure 2.1. Water molecule (red) interacting with U19, G27, U28 and Lividomycin, in the binding site of 16S-rRNA (PDB: 2ESJ[19]).

Throughout the development of our docking program, FITTED,[20-23] conserved water molecules were found to be critical for optimal docking accuracy,[24] which was corroborated by other groups.[25] When we became interested in the docking of small molecules to nucleic acids,[26] proper treatment of water molecules was also found to significantly improve docking performances.[27] In the past, the critical role of water molecules and the prevalence of proteins as drug targets prompted the development of several methods to predict the position and orientations of water molecules at the interface of protein-ligand complexes and/or to evaluate their free energy of binding. Examples include WaterMap, which relies on MD simulations of solvated proteins;[28, 29] JAWS, which estimates the free energy for each water molecule from Monte Carlo simulations;[30] statistical methods such as Szmap[31] and 3D-RISM;[32, 33], docking water molecules to binding sites (e.g., WaterDock relying on AutoDock Vina[34]) and empirical methods such as AQUARIUS which predicts solvent sites within a protein by mapping each amino acid to a data set of crystal structures, and others as was recently reviewed.[35] In addition, methods and case studies for distinguishing between bound and displaced waters in the context of different ligands have also been reported.[36] However, to the best of our knowledge, these methods were developed primarily for proteins and never applied to nucleic acids.

Recently, a report on the use of a 3D-RISM-derived method illustrated the challenges of placing water molecules in nucleic acids;[37] it had an accuracy (of water placement within 1 Å of

crystallographic water molecules) below 40%, although testing was carried out on only a handful of structures. Herein, we present a methodology and available software for placing water molecules in the binding site of nucleic acid-ligand complexes, using a hybrid scoring function composed of statistical survey of water molecules in existing PDBs and a dedicated force field, developed specifically for water placement in nucleic acids.

2.3 Theory and Implementations

Challenges. At the outset of this project, we identified key challenges to address: (1) placing an explicit (3-atom) water molecule would require not only identification of its location but also description of its orientation; (2) the binding free energy of water molecules is a complex combination of entropy and enthalpy; (3) hydrogen bonds (H-bonds) are directional and sensitive to polarization; (4) water molecules could become highly polarized in such polar media, and electrostatic interactions would therefore be inaccurately computed using static point-charges; (5) the developed method should be available to medicinal chemists, fully automated, and user-friendly. In order to solve these challenges, a new strategy for water placement (Figure 2.2) in nucleic acids was developed, which will be addressed sequentially in the following sections.

Selection of a Suitable Water Model: United Field Water Molecules – Challenge #1. To address the first challenge, we have previously developed particle waters (PWs) which model the hydrogen bond donor (HBD) and acceptor (HBA) properties of water molecules in a single bead hence reducing the problem to only bead placement. The placement of PWs was originally implemented into PREPARE, a program previously developed to convert PDB files into useable structures (adding and optimizing hydrogen positions and water orientation).[24, 38] In this early version, interactions between proteins and PWs were modeled using a Lennard-Jones potential which was trained to account for all the interactions (ie. electrostatics, van der Waals, and hydrogen bonds). However, this implementation was poor in accuracy (see results and discussion) and we have fully redesigned these particles to take into account additional physical phenomena, such as directionality of hydrogen bonding and polarization of water molecules. Additionally, these PWs were completely redesigned for nucleic acids.

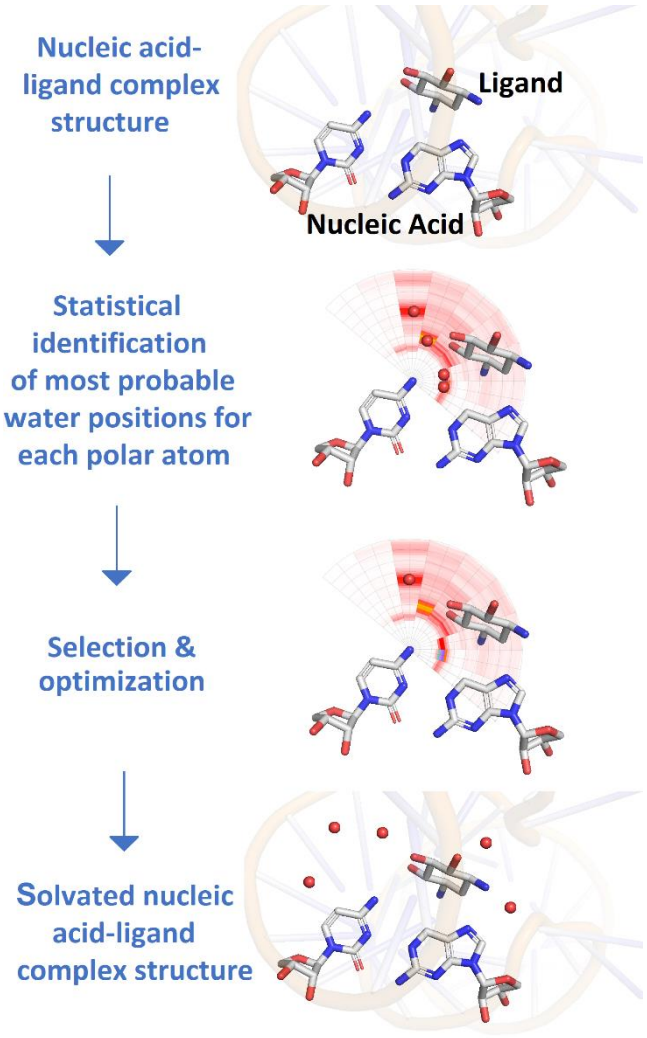


Figure 2. 2. Method flowchart designed for SPLASH'EM.

Statistics as a Scoring Method for Water Placement – Challenge #2. To address the difficult task of water placement in nucleic acids, statistics and molecular mechanics (MM) methods were used in combination. Statistics information could not only be used to place water molecules, but also to assign a probability to each position in space, which could be converted into a free energy of binding. As shown in Figure 2.2, statistics could provide key information on the likely positions of water molecules near bases, phosphates, sugars, and ligands. These locations of high occurrences would then be targeted to position water molecules.

Thus, our first task was to collect structural information on preferred water molecule positions. Statistical analysis around each of the nucleotides was carried out in over 4,100 crystal structures obtained from the Protein Data Bank (PDB), for a total of ca. 5.8M water molecules.

Water distribution around each H-bond acceptor/donor was plotted and summarized into a series of graphs (Figure 2.3). These results agreed well with a past computational study of the solvation of RNA using MD simulations.[39] This water distribution data was used as a starting point for our new program SPLASH'EM (Solvation Potential Laid around Statistical Hydration on Entire Macromolecules).

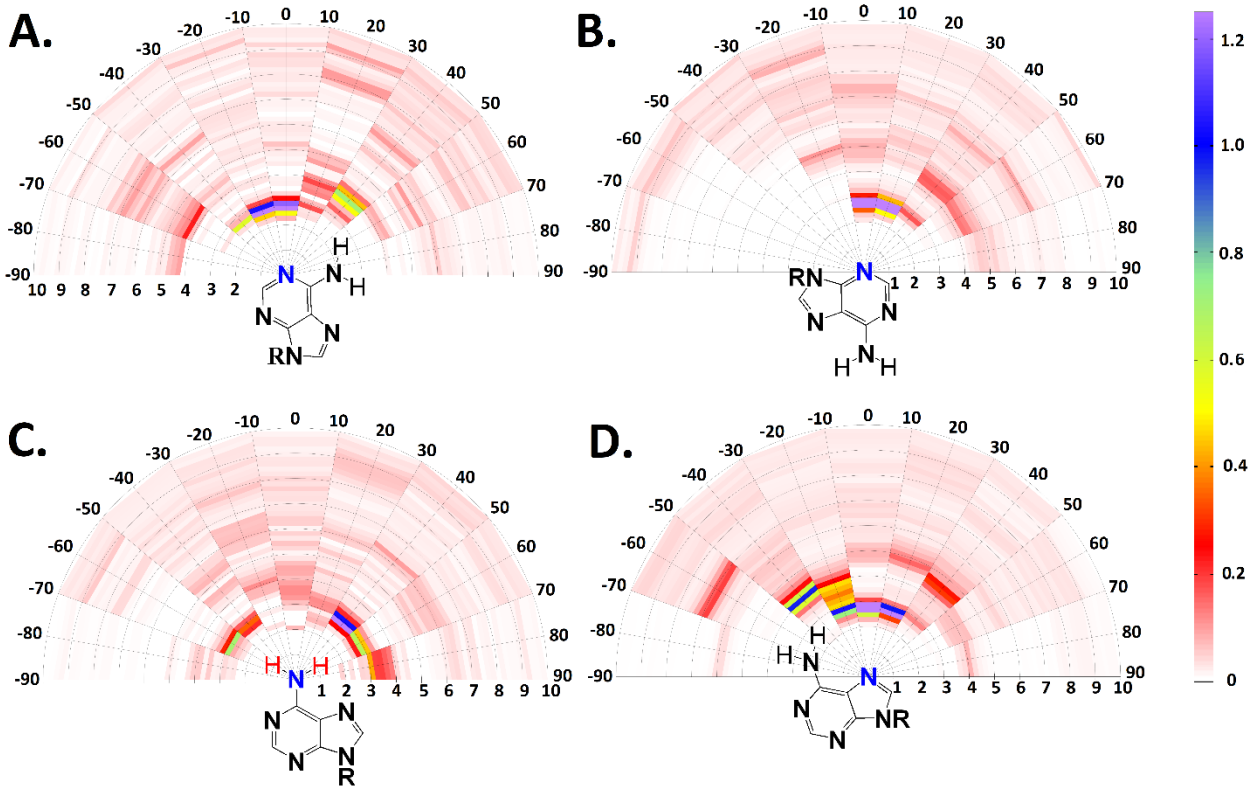


Figure 2.3. Statistics information collected on the occurrences of water molecules near: Adenine N1 (A), N3 (B), N6 (C), and N7 (D) at various angles (-90° to 90°) and distances (0 to 10 Å along the x-axis) within the plane of the base. Statistics were similarly collected outside the plane at different "pitch" angles. In each voxel, frequencies are indicated as a total percentage of waters collected. Statistics for other hydrogen bond acceptors/ donors at various pitch angles could be found in the supplementary information

As water molecules are often H-bonded to more than one residue (Figure 2.1), estimating the free energy of water molecules was achieved through the combination of the free energy distributions from each nearby polar atom (although we are aware that entropy is not additive).

Particle Water Force Field as a Scoring Method for Water Placement – Challenges #3 and #4. While the PDB-derived statistics of water positions give an approximation of the likely locations of water molecules, it was believed that refinement may be needed. This refinement could be performed by using gradient descent optimization in order to position the water molecules at regions of energy minima. Thus, the free energy score obtained from statistics was complemented with MM, the latter of which precludes any clashes, considers any other interactions with the nucleic acids, and enables an energy optimization necessary to refine PW positions. A new function and corresponding parameters (force field, or FF) for computing the nucleic acid-PW and ligand-PW interaction energies were developed for use in energy evaluation and optimization of PWs including directionality of hydrogen bonding (challenges #3) and polarization (challenge #4).

Finding a Suitable Function and Parameters for Modelling Hydrogen Bonding between Nucleic Acids and PWs. Functions (Equation 2.1) and parameters were trained to reproduce the energy profile of the interaction between nucleic acids and water molecules at the MP2/6-31+G(d,p) level of theory, as illustrated in Figure 2.4 and further discussed in the Appendix A.

Equation 2.1. Hydrogen Potential Function of PWFF.

$$E_P = \frac{A}{r^6} - \frac{B}{r^5} - \frac{C}{r^3}$$

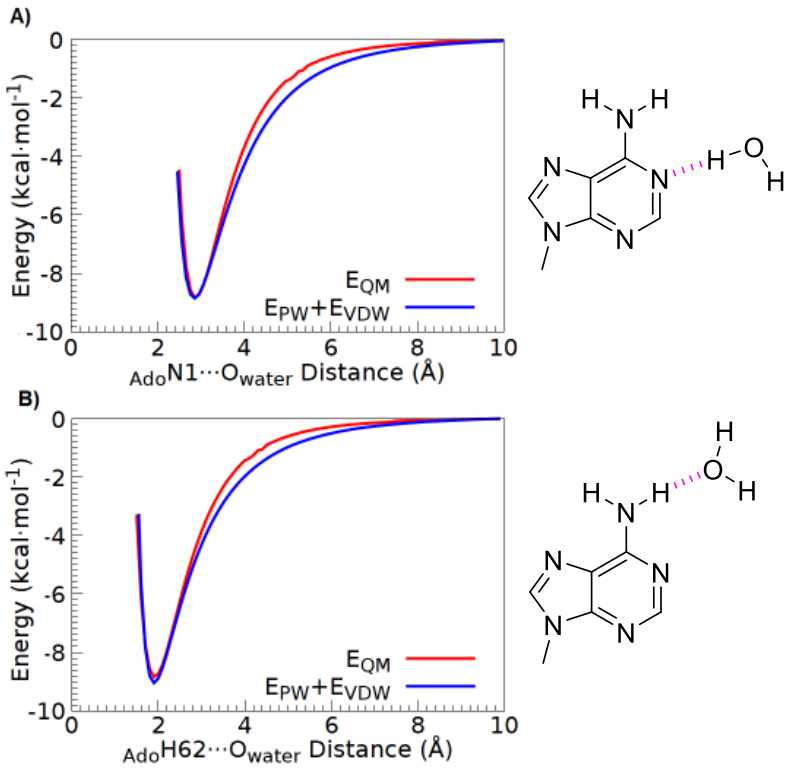


Figure 2.4. Hydrogen bonding potential in QM (red) as compared to the developed PWFF (blue) for a linear distance scan for Ado-N1 (A) and Ado-H62 (B). The corresponding chemical structures are shown on the right, with distances labeled in dashed lines (magenta).

This equation describes the hydrogen bonding and electrostatic energies between two atoms (E_p), as a function of their separation distance (r). A, B, and C are specific parameters to describe the strength of hydrogen bonding interaction between pairs of atoms (eg. $_{Ado}N1\cdots O_{water}$). This FF was named the Particle Water Force Field (PWFF). In PWFF, dispersion and Pauli repulsion terms were obtained from Lennard-Jones potential of Amber ff95. As PWs are neutral, the hydrogen bonding term implicitly accounted for electrostatic interactions. A Lennard-Jones 6-5/3 term was found to best reproduce this energy profile obtained through QM calculations (Equation 2.1).

Methodology Used to Develop the Particle Water Force Field. Optimizations were performed on entire nucleic acid fragments (ie. adenine, guanine, cytosine, thymine, uracil, ribose, deoxyribose, phosphate) in QM while the model for water molecules was derived from TIP3P. To develop the PWFF, H-bonding energy profiles between nucleic acid fragments and water molecule were calculated using the previously described QM- and MM-levels of theory (Figure 2.5). To

derive the H-bonding potential, the difference between the QM energy profile and the van der Waals MM energy profile was used (Equation 2.2). After attempting various Lennard-Jones potentials, including the well-known 12-10, 12-8, and many others, a 6-5/3 term was found to best reproduce the H-bond potential.

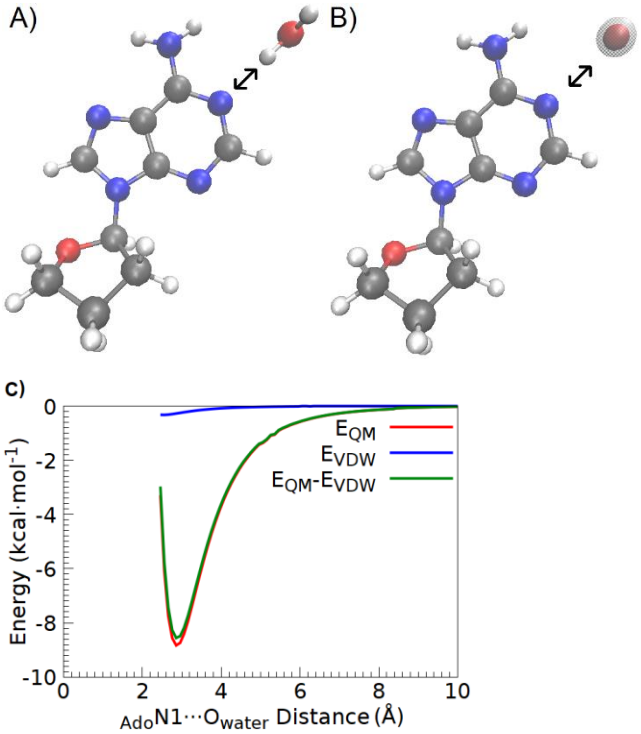


Figure 2.5. Model used for evaluating the energy at A) MP2/6-31+G (d,p) and B) MM, with united atom representation of water molecule. Arrows depict the path taken during the potential energy scan. C) Energy as a function of distance, depicting the enthalpic energy calculated in QM (red), van der Waals energy (blue), calculated using Amber ff95, and the difference of the former and latter (green), which is the hydrogen bonding potential.

Equation 2.2. Equation used for fitting and parametrizing PWFF.

$$E_{H-bond\ (PW)} = E_{QM} - E_{MM:\ LJ\ 12-6}$$

Incorporating Directionality of Hydrogen Bonding into Particle Water Force Field – Challenge #3. In the past, attempts to include directionality of H-bond terms in some force fields were reported,[40-42] although no clear improvements in performance were noted and these terms

were subsequently discarded. However, these previous developments were largely tailored towards MD simulations and not towards water placement methods or *in silico* docking studies. To this end, the angular dependence of H-bonding with respect to simplified nucleic acid subunits were investigated by scanning water positions with respect to these subunits (Figure 2.6).

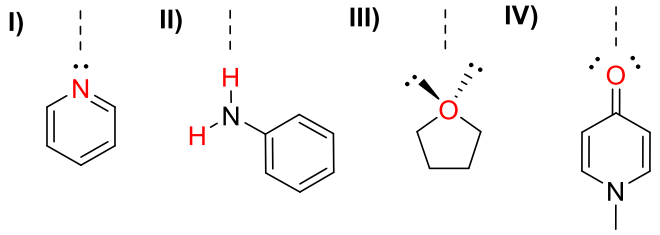


Figure 2.6. Simplified subunits used to probe the relationship between enthalpy and angles. I) pyridine, II) aniline, III) tetrahydrofuran, and IV) 1-methylpyridin-4(1H)-one are shown. Polar atoms of interest are in red, while angle bisectors are depicted by dashed lines.

An angular contribution function (Figure 2.7A) was next developed to account for the directionality of the H-bonding, penalizing deviations of H-bonding from ideal positions. For H-bond acceptors and donors, the ideal angle was designated at the positions of the electron lone pair(s) and directly opposite the R-H bond, respectively. An example of the H-bonding potentials as a function of angles in both QM and the developed PWFF potentials for various chemical fragments are shown (Figure 2.7B-E).

The angular contribution function utilizes a normal distribution-like function, which is always bound between 0 and 1. At ideal angles, the angular contribution function yielded a value close to 1, which instates the full strength of the hydrogen bond. In contrast, when far from the ideal angle, the hydrogen bonding is essentially turned off by a value close to 0. The augmentation of PWFF with angular contribution function (PWFFa), although not perfect, reproduced the QM energy profile for these molecules, even at non-ideal H-bonding geometries. Equation 2.3, along with Table 2.1 describes the PWFFa function form, along with its various associated parameters, the latter of which are the same as those used for PWFF.

Equation 2.3. The hydrogen bonding potential for PWFFa.

$$E_P = \begin{cases} \underbrace{\left(\frac{A}{r^6} - \frac{B}{r^5} - \frac{C}{r^3}\right)}_{distance\ contribution} \cdot \underbrace{\left(e^{-(\alpha\theta)^{2n}}\right)}_{angle\ contribution} \\ \underbrace{\left(\frac{A}{r^6} - \frac{B}{r^5} - \frac{C}{r^3}\right)}_{distance\ contribution}, \quad if\ r < r_{threshold} \end{cases}$$

More specifically in equation 2.3, the hydrogen bonding energy, E_p, is now also expressed as a function of θ which is the angular deviation from "ideal" hydrogen bonding geometry defined above. Two other parameters, α and n help to define the curvature of the well and how quickly the well depth changes as a function of θ. The ideal hydrogen bonding geometry for hydrogen bond donors and hydrogen bond acceptor was designated as the line connecting the donor and donor antecedent atoms and line connecting the electron lone pair and the acceptor atom, respectively.
A potential ambiguity with this angular contribution function could arise if a PW was within steric clashing distance, but at a non-ideal angle. In order to prevent this from occurring, a threshold distance was given for different atoms. For hydrogen atoms, this value was 1.5 Å while all other heteroatoms were assigned a value of 2.5 Å. At distances less than these values, the full steric clash would be restored.

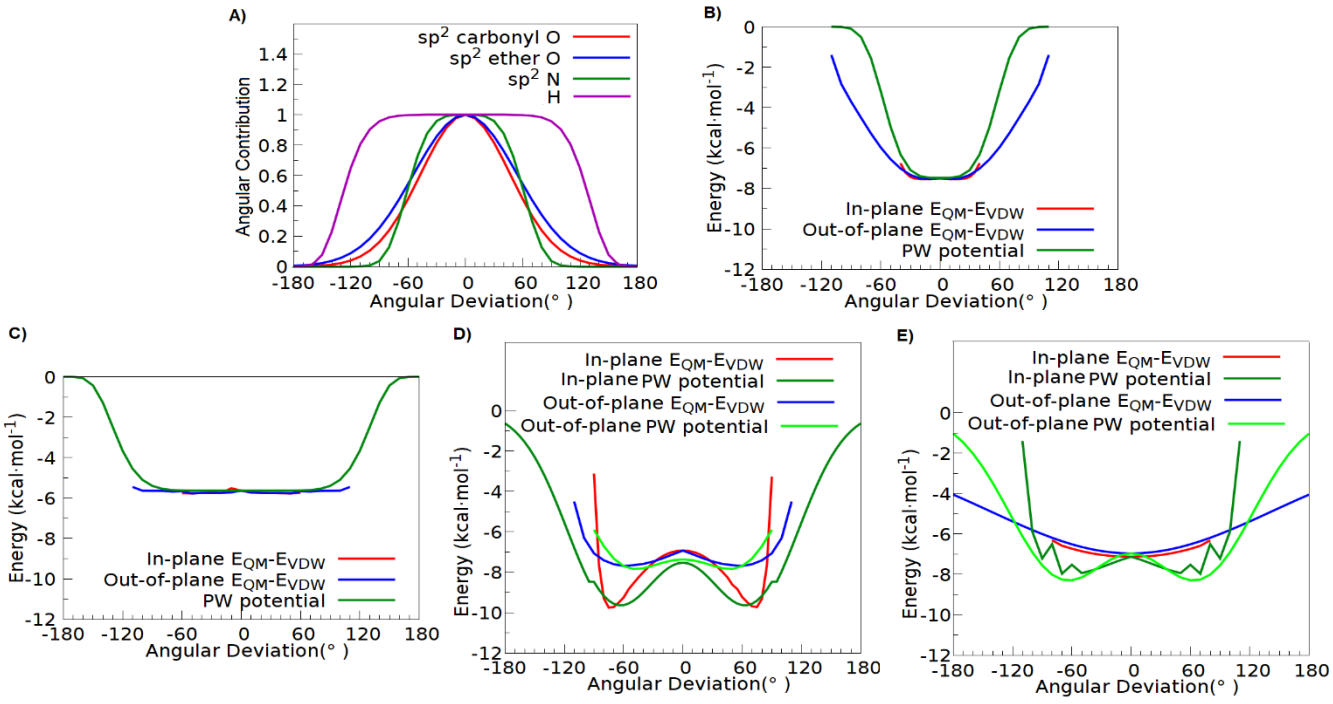


Figure 2.7. (A) Angular contribution function implemented into PWFF for various atom types. (B-E) Energy profile as a function of deviation from the angle bisector, obtained by E_{QM}-E_{VDW} for both in-plane and out-of-plane angles, with respect to the ring, as compared to the developed PW potential, (B) pyridine, (C) aniline, (D) tetrahydrofuran and (E) 1-methylpyridin-4(1H)-one are shown. For acceptors with more than one lone pair (ie. tetrahydrofuran and 1-methylpyridin-4(1H)-one), both in-plane and out-of-plane PW potential are shown.

Incorporating Polarization of Hydrogen Bonding into Particle Water Force Field – Challenge #4. Polarization of water molecules in nucleic acids occurs more frequently than in proteins due to the anionic charge of the phosphate backbone and presence of countercations such as Mg^{2+} and K^{+}. Polarization has been found to lead to artifacts during MD simulations of proteins,[9, 10] and methods to address this problem have been proposed.[43, 44] However, it remains to be addressed for the more polar nucleic acids. From our quantum mechanics (QM) calculations, it was observed that a water molecule bridged by two polar atoms experienced an overall potential unequal to the sum of the two individual potentials (Figure 2.8).

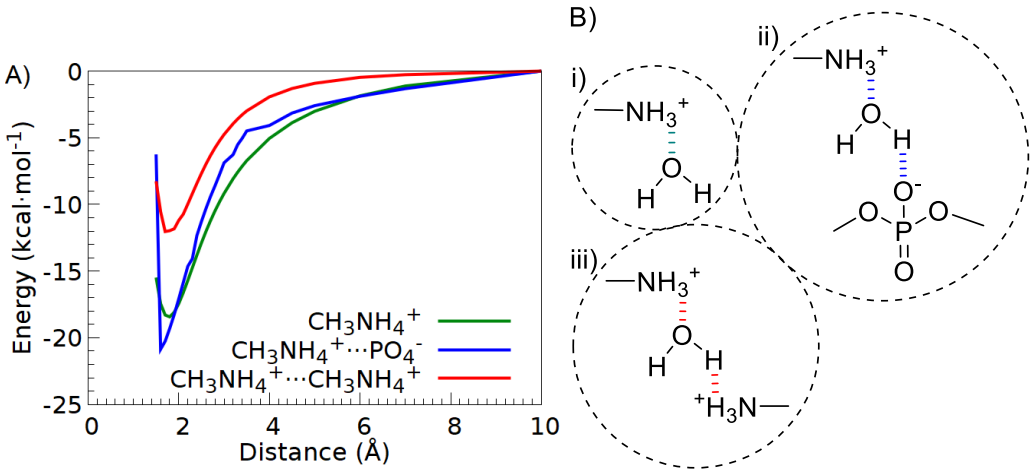


Figure 2.8. A) H-Bond potentials between $CH_3NH_4^+$ and H_2O (in QM) when the latter is: not part of a bridge (green), in a bridged to $(CH_3)_2PO_4^-$(blue), and in a bridge to another $CH_3NH_4^+$ (red). B) Corresponding schematics are shown in (i-iii), respectively.

At each water position, the two greatest hydrogen bonding energies were selected, and polarization factors were assigned to these two interactions. Five distinct classes of polarization were created: 1) strongly stabilizing, 2) strongly destabilizing, 3) weakly stabilizing, 4) weakly destabilizing, and 5) unchanged. A polarization factor was introduced for each of these cases. The first case involved the interaction of oppositely-charged polar atoms, and was given a polarization factor of 1+β. Similarly, the second case involved like-charged polar atoms interacting, and was given a polarization factor 1–β. The third and fourth cases involved an uncharged polar atom and a charged polar atom. In both cases, the uncharged atom was given a polarization factor while the charged was not. More specifically, the third case involved a HBA-HBD+ or HBD-HBA- pair and was given a polarization factor of 1+α while the third case involved a HBD-HBD+ or HBA-HBA- pairs and was given a polarization factor of 1-α. Finally, the polarization between uncharged groups were expected to be minimal and no polarization was given. Different sets of factors were evaluated by varying both α and β parameters between 0.0~0.9 at 0.05 increments with the criteria that $\alpha \leq \beta$ and the optimal values were found to be $\alpha = 0.2$ and $\beta = 0$ (Table 2.2). This polarizable force field was named polarizable PWFF with angle contribution function (pPWFFa), which is shown in equation 2.4.

Equation 2.4. Hydrogen bonding potential for pPWFFa.

$$E_P = \begin{cases} \underbrace{\left(\frac{A}{r^6} - \frac{B}{r^5} - \frac{C}{r^3}\right)}_{\substack{distance \\ contribution}} \cdot \underbrace{\left(e^{-(\alpha\theta)^{2n}}\right)}_{\substack{angle \\ contribution}} \cdot \underbrace{f_{pol}}_{\substack{polarization \\ correction}} \\ \underbrace{\left(\frac{A}{r^6} - \frac{B}{r^5} - \frac{C}{r^3}\right)}_{\substack{distance \\ contribution}}, \quad if\ r < r_{threshold} \end{cases}$$

Equation 2.4 expresses the hydrogen bonding potential, E_p, as a function of previously defined parameters in equation 2.2 and 2.3. In addition, a polarization factor, f_{pol}, was introduced to modulate the strengths of hydrogen bonding, depending on the five different types of neighboring environments as discussed above.

Water Placement Strategy. In our newly developed fully automated program (challenge #5), SPLASH'EM, water molecules are initially placed in the centers of the top 50 most populated grid points for each polar atom based on the previously collected statistical information obtained (Figure 2.2). At this point, each PW was assigned a binding free energy and force field energy based on its position relative to all polar atoms in the complex. PWs sterically clashing with the receptor and ligand were removed, along with PWs not located in the binding site. As our focus was on nucleic acid-ligand complexes, only bridging water molecules were considered (i.e. waters interacting with both the nucleic acid and a ligand). Finally, PWs are ranked energetically (force field and/or free energy), and placed from most to least stable, with the criteria that a PW does not clash with a previously placed PW and is not located on the water-exposed surface.

Validation Dataset for Water Placement. As of December 2018, there were over 4,000 nucleic acid crystal structures available in the PDB. Of these, approximately 300 nucleic acids contain ligands and at least one water molecule. To compile a validation set for water placement, a further refinement of the nucleic acid crystal structures was performed. A filtering step by a threshold resolution of 2.6 Å was performed on these previously obtained structures, resulting in 91 PDBs, which contained a total of 12,955 water molecules. However, many of these water molecules were found in bulk water or away from the ligand binding site, and thus unimportant for medicinal chemistry applications. Only bridging waters, less than 3.5 Å away from both a nucleic acid polar atom and a ligand polar atom were kept, leading to 398 water molecules. Finally, in order to ensure that these water molecules, present in the original PDB structures were

substantiated by electron density, the corresponding electron density data, containing the 2F0-Fc map was used as experimental evidence. To carry out this procedure, EDIA[36] calculation was performed on each crystal structure. EDIA calculates the experimental support for an atom by taking into account the electron density within its van der Waals radius, and allows comparison between different structures by normalization. Water molecules with EDIA score of less than 0.24 were discarded, as suggested by Rarey and co-workers, since these waters are less supported by electron density.[26, 36] The final compiled validation set was made up of 91 crystal structures, containing a total of 230 water molecules.

Table 2.1. PW hydrogen bonding parameters.

Particle Water-Polar Atom Interactions		Distance Contribution			Angle Contribution	
		A	B		α	n
Adenine	N1	18310	7535	0	0.015	1
	N3	27792	11522	0	0.015	1
	H61	297	0	76	0.01	4
	H62	297	0	76	0.01	4
	N7	19187	7814	0	0.015	1
Guanine	H1	1431	886	0	0.01	4
	H21	163	0	48	0.01	4
	H22	163	0	48	0.01	4
	N3	14986	6170	0	0.015	1
	O6	25510	10500	0	0.015	2
	N7	20596	8008	0	0.015	1
Cytosine	O2	25590	10568	0	0.015	2
	N3	20043	8262	0	0.015	1
	H41	364	0	99	0.01	4
	H42	364	0	99	0.01	4
Thymine/ Uracil	O2	17551	7230	0	0.015	2
	H3	800	574	0	0.01	4
	O4	17311	7521	0	0.015	2

Deoxyribose	O4’	17311	7521	0	0.015	2
Ribose	O2’	14300	6128	0	0.015	2
	HO2’	300	0	85	0.01	4
	O4’	18899	8090	0	0.015	2
Phosphate	O3’	22280	8740	0	0.015	2
	O5’	22280	8740	0	0.015	2
	OP1	5000	0	525	0.019	1
	OP2	5000	0	525	0.019	1
Pyridine	N	22770	9419	0	0.015	1
Methylamine	HN	631	0	231	0.01	4
Methanol	HO	1490	0	965	0.01	4
	OH	16690	7020	0	0.015	2
Acetylamide	O	17810	7145	0	0.015	2
	HN	320	0	89.1	0.01	4
Formate	O	5383	0	615	0.015	2

Table 2.2. Polarization factors assigned to H-bond potentials.

Moving	Stationary	Polarization Factor
Weak Donor	Weak Donor	1
Weak Donor	Weak Acceptor	1
Weak Donor	Strong Donor	1 - α
Weak Donor	Strong Acceptor	1 + α
Weak Acceptor	Weak Donor	1
Weak Acceptor	Weak Acceptor	1
Weak Acceptor	Strong Donor	1 + α
Weak Acceptor	Strong Acceptor	1 - α
Strong Donor	Weak Donor	1
Strong Donor	Weak Acceptor	1
Strong Donor	Strong Donor	1- β
Strong Donor	Strong Acceptor	1+ β

Strong Acceptor	Weak Donor	1
Strong Acceptor	Weak Acceptor	1
Strong Acceptor	Strong Donor	1+ β
Strong Acceptor	Strong Acceptor	1 - β

2.4 Results and Discussion

Assessing the Accuracy of Placed Water Molecules. Assessing the accuracy of water placement on nucleic acids is difficult primarily due to its dependence on X-ray crystal structures of nucleic acids. The structure is a static representation of a biological molecule in a non-physiological environment. Some waters may not be resolved due to their high mobility and low residence times at certain positions. Consequently, the placement of a water molecule at these locations should not necessarily be considered wrong. Inherent to X-ray crystallography are potential biases which may be introduced during the refinement process. This is especially true if the crystal structure is low in resolution. In fact, in a few cases, the placement of water molecule would have been satisfied in more than one way. For example, during the model construction of the crystal structure of a DNA dodecamer complexed with propamidine (PDB ID: 102D[45]), a single water molecule (O36) was placed in the region of high electron density, which yielded an EDIA score of 0.59 (Figure 2.9). Alternatively, two water molecules could have been placed adjacent to the crystallographic water location at distances of approximately 1.7 Å and 1.6 Å away. At these two alternate locations, there is also support from electron density with EDIA scores of 0.24 and 0.30. In this case, the X-ray diffraction pattern caused by these two water molecules would have superposed and created the maximum electron density signal at the location of the crystallographic water. In physiological conditions, the movement of water molecules would probably have allowed its presence at both locations. Consequently, this presents a challenge in scoring of placed water molecules since these alternative positions are not considered successful by a first criterion for placement success (PW within 1Å from crystallographic water). Using a different second criterion (EDIA ≥ 0.24), these water molecules are considered successful.

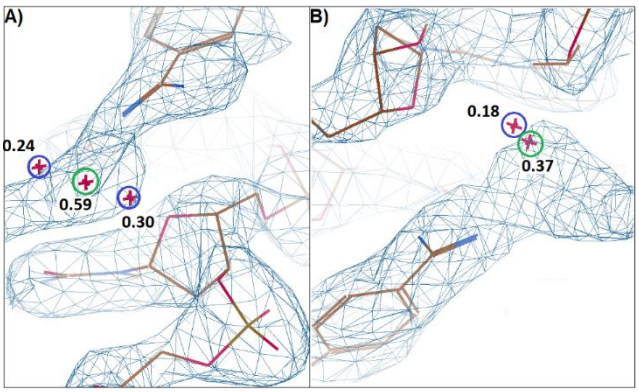


Figure 2.9. Crystallographic water position (green circle) compared to possible alternative water positions (blue circles) overlaid with the 2F0-Fc electron density map. EDIA score is given adjacent to each water. (A). Placed waters are supported by electron density but are situated away (distances of 1.6 and 1.7 Å) from the crystallographic water. (B). Placed waters are not supported by electron density but are in close proximity to the crystallographic water (distance of 1.0 Å).

However, a disadvantage with the latter arises when water molecules which are in close proximity to a crystallographic water (less than 1 Å) is considered unsuccessful by electron density, as illustrated by another bridging water molecule (Figure 2.9B). In this case, the crystallographic water and the placed water had an EDIA score of 0.37 and 0.18, respectively. As a result, the placed water was not scored as successful. However, due to their close spatial proximity (distance of 1 Å apart), the placed water may still be useful in providing important information for *in silico* docking. As a result, neither of the criterion could satisfactorily measure the accuracy or precision. Even worse, the lack of an electron does not necessarily preclude the presence of water. As an example, a ligand bound to 2'-Deoxyguanosine riboswitch (PDB: 3SKI[46]) was found to have two identical binding sites (Figure 2.10). In spite of this, one binding site lacked electron density for water molecules while the other showed the presence of two water molecules.

Consequently, until a robust scoring methodology could be envisioned, it is difficult to assess the true accuracy of a water placement methodology. Ultimately however, the accuracy of water placement methodologies depends on its utility towards a given problem (e.g., docking, MD, biocatalysis). Due to our interest in using water placement for docking, the accuracy of the developed water placement methodology was given with the first aforementioned criteria.

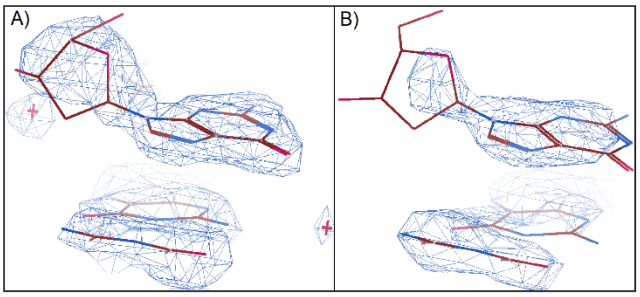


Figure 2.10. 2'-Deoxyguanosine bound to each binding site of the dimeric 2'-Deoxyguanosine riboswitch (PDB: 3SKI). Shown in A) and B) are the two co-crystallized ligands in each binding site, and their associated 2F0-Fc electron densities.

Accuracy of Various Placement Methods. In testing the newly developed water placement method, the list of 91 crystal structures from the PDB was included in the testing set, which contained a total of 230 bridging water molecules (Appendix A). Testing was first carried out with random water placement as a reference, which correctly positioned a PW within 1 Å of experimentally observed water molecule, in nearly 1 of every 5 cases (Table 2.3). Additionally, water placement using our previous protocol, PREPARE,[38] (developed for proteins), was tested along with the scoring methods developed herein.

Table 2.3. PW placement accuracy of current developments

Placement method	Scoring method	Accuracy[a]
Random	N/A	19%
Prepare[38]	PREPARE[38]	30%
Statistics	Statistics	51%
Statistics	PWFF	51%
Statistics	PWFFa	56%
Statistics	pPWFFa	60%
Statistics	Splash'Em	62%

[a] Number of crystallographic water with a PW within 1 Å.

Each distinct scoring method (free energy from statistics or FF energy) was tested individually and in combination to see how it would impact the accuracy of the methodology. In total, five different

scoring criteria are reported, including: statistics, PWFF, PWFFa, pPWFFa, and SPLASH'EM, which incorporates all of the above.

Accuracy of PREPARE. Water placement using PREPARE[38] gave an accuracy of 30%. Based on recent evidences, it has been suggested that the water network surrounding ligands differ based on its identity, as was the case of aminoglycosides binding to RNA.[27] However, the water placement method in PREPARE did not take the influence of the ligand into consideration. To rectify this, each new developed placement methodology took the ligand into consideration.

Accuracy of Statistics as Water Placement Methods. Statistics as a scoring criterion has many advantages, including the implicit incorporation of polarization, directionality of H-bonding, and free energies. Using this method for water placement gave a significantly improved accuracy of 51%. The additivity of free energies is a challenge in this approach, as the relationship for sum of entropy has not been fully elucidated. For this reason, it is difficult to predict the relative stability between two potential water positions. Additionally, the lack of available ligand statistics forced us to use approximate free energies, which used similar nucleic acid atom types, leading to possible errors in free energies.

Accuracy of Particle Water Force Field as Water Placement Method. The three variations of the developed Particle Water Force Field (PWFF, PWFFa, and pPWFFa) yielded accuracies of 51%, 56%, and 60%, respectively. We were very pleased that with each additional chemical principle incorporated into the force field (ie. angles and polarization), significant increases in accuracy were observed.

The use of PWFFa over PWFF eliminated weak H-bond interactions (Figure 2.11A), which were at large angular deviations from the ideal geometry. For example, during the placement of a water molecules in the binding site of an RNA riboswitch-SAM - ligand complex (125FJC[47]), PWFF placed a water molecule 2.3 Å away from the crystallographic position (Figure 2.11A). At this position, it was within H-bonding distances to two polar atoms: ribose-O2' and Ado-N3. Chemically however, the geometry at this position did not allow for a strong hydrogen bonding to ribose-O2', which was not taken into account by PWFF. On the other hand, PWFFa and pPWFFa were able to take directionality of H-bonding into consideration and greatly penalize this interaction. Consequently, an adjacent location, 0.4 Å away from the crystallographic water, was correctly chosen (Figure 2.11).

In spite of the successes of the angular correction function in PWFFa, there were still a number of water molecules which were incorrectly predicted—one reason being that it was unable to account for the nuance behaviors of the free energy surface, including polarization. pPWFFa, in contrast to PWFFa, identified destabilized interactions due to polarization (Figure 2.11B). In other cases, pPWFFa also identified stabilized interactions (ie. water bridged by a strong donor and a strong acceptor).

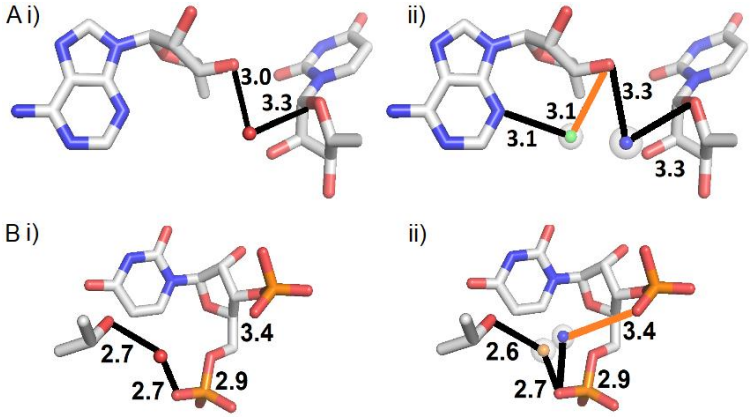


Figure 2.11. Comparison of (A) (i) crystallographic water position in red (ii) to water molecules placed by PWFF in green and PWFFa in blue. Comparison of (B) (i) crystallographic position in red to (ii) to water molecules placed by PWFFa in blue and pPWFFa in orange. H-bonds are denoted by black lines while weak H-bonds are denoted by orange lines.

In another example, during the placement of water molecules of an RNA-Paromomycin complex (Figure 2.11B; PDB 2PWT[48]), PWFF incorrectly selected the water position which emphasized interactions with two phosphate groups, which had the deepest innate H-bonding potentials. However, this location was expected to be destabilized since both groups are strong hydrogen bond acceptors. Consequently, the crystallographic water molecule was found away from this position, and instead H-bonded to a nearby hydroxyl hydrogen of the modified Paromomycin ligand although still H-bonded to one of the phosphate groups. This crystallographic water position was correctly predicted by pPWFFa, but not PWFFa.

Originally, it was expected that the hydroxyl hydrogen of the modified Paromomycin would increase its hydrogen bonding potential in the presence of a strong acceptor (phosphate group). However, during the search for a suitable polarization factor, negligible changes in accuracy was observed for incorporating polarization of weak HBA and HBD. As a result, polarization for these types of interactions were not included in pPWFFa. More specifically, it was

found that polarization factors of $\alpha = 0.2$, $\beta = 0$ yielded optimal accuracies for water placement. This meant that the polarization factor for strongly stabilizing, strongly destabilizing, weakly stabilizing, and weakly destabilizing was 1.2, 0.8, 1, and 1, respectively (Table 2.2). The fact that $\beta = 0$ suggests that weakly stabilizing and destabilizing interactions are in general less sensitive to polarization. Since pPWFFa does not take into account the distance-dependent nature of polarization, no polarization effect for this type of interaction was observed using our method. However, subtle polarization is still expected in these types of interactions at close distances between the water molecule and HBA/HBD. In the future, a more encompassing polarization scheme could be devised, such as varying the degree of polarization, based on the distances between water and HBA/HBD. However, the polarization scheme described in this method was a proof-of-principle showing that its consideration is important when attempting to accurately predict water molecule positions. Overall, in both cases (ie. moving from PWFF to PWFFa and PWFF to pPWFFa), the more advanced FF correctly identified the crystallographic water positions.

Accuracy of a Hybrid Scoring Method for Water Placement. We next envisioned that the two energy evaluation methods could be combined into a single metric, with the hopes of further increasing the accuracy of water placement. More specifically, a large number of overlapping PWs were placed, scored with pPWFFa, and two alternative populations of non-clashing sets of PWs extracted. These two distinct populations were subsequently combined, and the approximate free energy (from statistics), was used to select the final list of PWs. This method produced an accuracy of 62% on the selected 230 crystal water molecules, which was the final version implemented into SPLASH'EM itself integrated into our drug discovery platform FORECASTER (challenge #5).[38] More interestingly, decomposing the accuracy of this placement method into five equal bins of size 46, according to the B-factor of the PDB water molecules showed that SPLASH'EM was able to more accurately predict water molecules with lower B-factor than those with a higher B-factor (Figure SA5). In fact, the prediction for those water molecules with B-factors less than 22 was 70% compared to those with B-factors higher than 42 was just 52%. Although this is still a small dataset, it shows that SPLASH'EM is able to more accurately predict water molecules which are of higher quality or are less mobile.

SPLASH'EM employs a similar method to hybrid scoring functions developed for molecular docking, where two or more methods are used together in hopes of more accurately identifying the

correct binding mode of a ligand in the receptor. In the case of water placement, the combination of approximate free energy and force field energy gave an increased performance over either of these methods used alone.

2.5 Conclusion

In conclusion, we showed that the accuracy of water placement in nucleic acid structures is difficult to obtain due to the insufficient number of nucleic acid-ligand 3D structures. In addition, due to poor resolution, subjectivity during model construction, and a lack of electron density in some regions, crystal structures may not contain all of the desired information for validating the accuracy of a placement method. Ultimately, the accuracy of a given water placement method rests with its utility in a given problem.

We succeeded in developing SPLASH'EM, a program to identify and place water molecules (in the form of PWs) in nucleic acids on the order of seconds. This method was able to correctly identify as many as 62% of water molecules in nucleic acids, laying the foundation for water placement for these macromolecules. SPLASH'EM utilizes a hybrid scoring function composed of statistical data of water occurrences for each polar atom of nucleic acids and ligands; and a newly developed pPWFFa FF to describe H-bonding of polar atoms and water molecules. Associated parameters for the FF were developed, which is novel in that it adds H-bond directionality and polarization to our previously developed united field water model, Particle Waters. pPWFFa is expected to be useful with *in silico* docking of nucleic acids when incorporated into FITTED. Overall, SPLASH'EM has been shown, at the time of publication, to be the most accurate water placement tool developed for nucleic acids and available for use by chemists.

2.5 References

1. Wang, M. L.; Yu, Y. Y.; Liang, C.; Lu, A. P.; Zhang, G., Recent Advances in Developing Small Molecules Targeting Nucleic Acid. *Int. J. Mol. Sci.* **2016**, 17, 24.

2. Cross, R., The RNA drug hunters Academics, biotech start-ups, and big pharma companies are designing small molecules that target RNA. *Chemical & Engineering News* **2017**, 95, 16-18.

3. Howe, J. A.; Wang, H.; Fischmann, T. O.; Balibar, C. J.; Xiao, L.; Galgoci, A. M.; Malinverni, J. C.; Mayhood, T.; Villafania, A.; Nahvi, A.; Murgolo, N.; Barbieri, C. M.; Mann, P. A.; Carr, D.; Xia, E.; Zuck, P.; Riley, D.; Painter, R. E.; Walker, S. S.; Sherborne, B.; de Jesus, R.;

Pan, W. D.; Plotkin, M. A.; Wu, J.; Rindgen, D.; Cummings, J.; Garlisi, C. G.; Zhang, R. M.; Sheth, P. R.; Gill, C. J.; Tang, H. F.; Roemer, T., Selective small-molecule inhibition of an RNA structural element. *Nature* **2015**, 526, 672-677.

4. Palacino, J.; Swalley, S. E.; Song, C.; Cheung, A. K.; Shu, L.; Zhang, X. L.; Van Hoosear, M.; Shin, Y.; Chin, D. N.; Keller, C. G.; Beibel, M.; Renaud, N. A.; Smith, T. M.; Salcius, M.; Shi, X. Y.; Hild, M.; Servais, R.; Jain, M.; Deng, L.; Bullock, C.; McLellan, M.; Schuierer, S.; Murphy, L.; Blommers, M. J. J.; Blaustein, C.; Berenshteyn, F.; Lacoste, A.; Thomas, J. R.; Roma, G.; Michaud, G. A.; Tseng, B. S.; Porter, J. A.; Myer, V. E.; Tallarico, J. A.; Hamann, L. G.; Curtis, D.; Fishman, M. C.; Dietrich, W. F.; Dales, N. A.; Sivasankaran, R., SMN2 splice modulators enhance U1-pre-mRNA association and rescue SMA mice. *Nat Chem Biol* **2015**, 11, 511-+.

5. Howe, J. A.; Wang, H.; Fischmann, T. O.; Balibar, C. J.; Xiao, L.; Galgoci, A. M.; Malinverni, J. C.; Mayhood, T.; Villafania, A.; Nahvi, A.; Murgolo, N.; Barbieri, C. M.; Mann, P. A.; Carr, D.; Xia, E.; Zuck, P.; Riley, D.; Painter, R. E.; Walker, S. S.; Sherborne, B.; de Jesus, R.; Pan, W.; Plotkin, M. A.; Wu, J.; Rindgen, D.; Cummings, J.; Garlisi, C. G.; Zhang, R.; Sheth, P. R.; Gill, C. J.; Tang, H.; Roemer, T., Selective small-molecule inhibition of an RNA structural element. *Nature* **2015**, 526, 672-677.

6. Childs-Disney, J. L.; Tran, T.; Vummidi, B. R.; Velagapudi, S. P.; Haniff, H. S.; Matsumoto, Y.; Crynen, G.; Southern, M. R.; Biswas, A.; Wang, Z.-F.; Tellinghuisen, T. L.; Disney, M. D., A Massively Parallel Selection of Small Molecule-RNA Motif Binding Partners Informs Design of an Antiviral from Sequence. *Chem* **2018**, 4, 2384-2404.

7. Dasari, S.; Tchounwou, P. B., Cisplatin in cancer therapy: molecular mechanisms of action. *Eur. J. Pharmacol.* **2014**, 0, 364-378.

8. Collie, G. W.; Parkinson, G. N., The application of DNA and RNA G-quadruplexes to therapeutic medicines. *Chem. Soc. Rev.* **2011**, 40, 5867-92.

9. Meng, E. C.; Cieplak, P.; Caldwell, J. W.; Kollman, P. A., Accurate Solvation Free-Energies of Acetate and Methylammonium Ions Calculated with a Polarizable Water Model. *J. Am. Chem. Soc.* **1994**, 116, 12061-12062.

10. Guallar, V.; Jarzecki, A. A.; Friesner, R. A.; Spiro, T. G., Modeling of ligation-induced helix/loop displacements in myoglobin: Toward an understanding of hemoglobin allostery. *J. Am. Chem. Soc.* **2006**, 128, 5427-5435.

11. Auffinger, P.; Westhof, E., RNA solvation: A molecular dynamics simulation perspective. *Biopolymers* **2001**, 56, 266-274.

12. Morley, S. D.; Afshar, M., Validation of an empirical RNA-ligand scoring function for fast flexible docking using RiboDock®. *J. Comput.-Aided Mol. Des.* **2004**, 18, 189-208.

13. Guilbert, C.; James, T. L., Docking to RNA via Root-Mean-Square-Deviation-Driven Energy Minimization with Flexible Ligands and Flexible Targets. *J. Chem. Inf. Model.* **2008**, 48, 1257-1268

14. Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D., *PLoS Comp. Biol.* **2014**, 10, e1003571.

15. Stelzer, A. C.; Frank, A. T.; Kratz, J. D.; Swanson, M. D.; Gonzalez-Hernandez, M. J.; Lee, J.; Andricioaei, I.; Markovitz, D. M.; Al-Hashimi, H. M., Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble. *Nat. Chem. Biol.* **2011**, 7, 553.

16. Lang, P. T.; Brozell, S. R.; Mukherjee, S.; Pettersen, E. F.; Meng, E. C.; Thomas, V.; Rizzo, R. C.; Case, D. A.; James, T. L.; Kuntz, I. D., DOCK 6: Combining techniques to model RNA–small molecule complexes. *RNA* **2009**, 15, 1219-1230.

17. Philips, A.; Milanowska, K.; Łach, G.; Bujnicki, J. M., LigandRNA: computational predictor of RNA–ligand interactions. *RNA* **2013**, 19, 1605-1616.

18. Luo, J.; Wei, W.; Waldispuhl, J.; Moitessier, N., Challenges and current status of computational methods for docking small molecules to nucleic acids. *Eur. J. Med. Chem.* **2019**, 168, 414-425.

19. Francois, B.; Russell, R. J. M.; Murray, J. B.; Aboul-ela, F.; Masquida, B.; Vicens, Q.; Westhof, E., Crystal structures of complexes between aminoglycosides and decoding A site oligonucleotides: role of the number of rings and positive charges in the specific binding leading to miscoding. *Nucleic Acids Res.* **2005**, 33, 5677-5690.

20. Moitessier, N.; Pottel, J.; Therrien, E.; Englebienne, P.; Liu, Z.; Tomberg, A.; Corbeil, C. R., Medicinal Chemistry Projects Requiring Imaginative Structure-Based Drug Design Methods. *Acc. Chem. Res.* **2016**, 49, 1646-1657.

21. Corbeil, C. R.; Englebienne, P.; Moitessier, N., Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J. Chem. Inf. Model.* **2007**, 47, 435-449

22. Lawandi, J.; Toumieux, S.; Seyer, V.; Campbell, P.; Thielges, S.; Juillerat-Jeanneret, L.; Moitessier, N., Constrained Peptidomimetics Reveal Detailed Geometric Requirements of Covalent Prolyl Oligopeptidase Inhibitors. *J. Med. Chem.* **2009**, 52, 6672-6684

23. Pottel, J.; Therrien, E.; Gleason, J. L.; Moitessier, N., Docking ligands into flexible and solvated macromolecules. 6. Development and application to the docking of HDACs and other zinc metalloenzymes inhibitors. *J. Chem. Inf. Model.* **2014**, 54, 254-265.

24. Therrien, E.; Weill, N.; Tomberg, A.; Corbeil, C. R.; Lee, D.; Moitessier, N., Docking Ligands into Flexible and Solvated Macromolecules. 7. Impact of Protein Flexibility and Water Molecules on Docking-Based Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2014**, 54, 3198-3210.

25. Garcia-Sosa, A. T.; Mancera, R. L., Free Energy Calculations of Mutations Involving a Tightly Bound Water Molecule and Ligand Substitutions in a Ligand-Protein Complex. *Mol Inform* **2010**, 29, 589-600.

26. Kieltyka, R.; Englebienne, P.; Fakhoury, J.; Autexier, C.; Moitessier, N.; Sleiman, H. F., A platinum supramolecular square as an effective G-quadruplex binder and telomerase inhibitor. *J. Am. Chem. Soc.* **2008**, 130, 10040-10041

27. Moitessier, N.; Westhof, E.; Hanessian, S., Docking of aminoglycosides to hydrated and flexible RNA. *J. Med. Chem.* **2006**, 49, 1023-1033

28. Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A., Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc. Natl Acad. Sci. USA* **2007**, 104, 808-813.

29. Wei, W.; Gauld, J. W.; Monard, G., Pretransfer Editing in Threonyl-tRNA Synthetase: Roles of Differential Solvent Accessibility and Intermediate Stabilization. *ACS Catalysis* **2017**, 7, 3102-3112.

30. Michel, J.; Tirado-Rives, J.; Jorgensen, W. L., Prediction of the Water Content in Protein Binding Sites. *J. Phys. Chem. B* **2009**, 113, 13337-13346.

31. Bayden, A. S.; Moustakas, D. T.; Joseph-McCarthy, D.; Lamb, M. L., Evaluating Free Energies of Binding and Conservation of Crystallographic Waters Using SZMAP. *J. Chem. Inf. Model.* **2015**, 55, 1552-1565.

32. Kovalenko, A.; Hirata, F., Three-dimensional density profiles of water in contact with a solute of arbitrary shape: a RISM approach. *Chem. Phys. Lett.* **1998**, 290, 237-244.

33. Kovalenko, A.; Hirata, F., Self-consistent description of a metal–water interface by the Kohn–Sham density functional theory and the three-dimensional reference interaction site model. *J. Chem. Phys.* **1999**, 110, 10095-10112.

34. Ross, G. A.; Morris, G. M.; Biggin, P. C., Rapid and Accurate Prediction and Scoring of Water Molecules in Protein Binding Sites. *PLoS ONE* **2012**, 7, e32036.

35. Nittinger, E.; Flachsenberg, F.; Bietz, S.; Lange, G.; Klein, R.; Rarey, M., Placement of Water Molecules in Protein Structures: From Large-Scale Evaluations to Single-Case Examples. *J. Chem. Inf. Model.* **2018**, 58, 1625-1637.

36. Garcia-Sosa, A. T.; Mancera, R. L.; Dean, P. M., WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes. *J. Mol. Model.* **2003**, 9, 172-82.

37. Giambaşu, G. M.; Case, D. A.; York, D. M., Predicting Site-Binding Modes of Ions and Water to Nucleic Acids Using Molecular Solvation Theory. *J. Am. Chem. Soc.* **2019**.

38. Therrien, E.; Englebienne, P.; Arrowsmith, A. G.; Mendoza-Sanchez, R.; Corbeil, C. R.; Weill, N.; Campagna-Slater, V.; Moitessier, N., Integrating medicinal chemistry, organic/combinatorial chemistry, and computational chemistry for the discovery of selective estrogen receptor modulatorswith FORECASTER, a novel platform for drug discovery. *J. Chem. Inf. Model.* **2012**, 52, 210-224.

39. Auffinger, P.; Westhof, E., Hydration of RNA base pairs. *Journal of Biomolecular Structure & Dynamics* **1998**, 16, 693-707.

40. Lii, J. H.; Allinger, N. L., Directional Hydrogen-Bonding in the Mm3 Force-Field .1. *J. Phys. Org. Chem.* **1994**, 7, 591-609.

41. Mayo, S. L.; Olafson, B. D.; Goddard, W. A., Dreiding - a Generic Force-Field for Molecular Simulations. *J. Phys. Chem.* **1990**, 94, 8897-8909.

42. 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.

43. Farahvash, A.; Leontyev, I.; Stuchebrukhov, A., Dynamic and Electronic Polarization Corrections to the Dielectric Constant of Water. *J. Phys. Chem. A* **2018**, 122, 9243-9250.

44. Das, A. K.; Demerdash, O. N.; Head-Gordon, T., Improvements to the AMOEBA Force Field by Introducing Anisotropic Atomic Polarizability of the Water Molecule. *J. Chem. Theory Comput.* **2018**, 14, 6722-6733.

45. Nunn, C. M.; Neidle, S., Sequence-Dependent Drug-Binding to the Minor-Groove of DNA - Crystal-Structure of the DNA Dodecamer D(Cgcaaatttgcg)(2) Complexed with Propamidine. *J. Med. Chem.* **1995**, 38, 2317-2325.

46. Pikovskaya, O.; Polonskaia, A.; Patel, D. J.; Serganov, A., Structural principles of nucleoside selectivity in a 2 '-deoxyguanosine riboswitch. *Nat Chem Biol* **2011**, 7, 748-755.

47. Huang, L.; Wang, J.; Lilley, D. M., A critical base pair in k-turns determines the conformational class adopted, and correlates with biological function. *Nucleic Acids Res.* **2016**, 44, 5390-8.

48. Kondo, J.; Pachamuthu, K.; Francois, B.; Szychowski, J.; Hanessian, S.; Westhof, E., Crystal structure of the bacterial ribosomal decoding site complexed with a synthetic doubly functionalized paromomycin derivative: a new specific binding mode to an a-minor motif enhances in vitro antibacterial activity. *Chemmedchem* **2007**, 2, 1631-1638.

3 Torsional Energy Barriers of Biaryls could be Predicted by Electron-richness/deficiency of Aromatic Rings; Advancement of Molecular Mechanics toward Atom-Type Independence

3.1 Preface

As mentioned in **Chapter 1**, force fields are used for many different applications in drug discovery, from MD simulations to vHTS. Although nucleic acids and proteins are well-parametrized by existing force fields, other less commonly encountered organic molecules are often assigned inaccurate parameters. This is due to practical limitations as the size of the chemical space and desired parameters are too large. As a result, parameters are frequently transferred to "similar" molecules. This has a negative impact on the quality of the force field, and results in a loss of accuracy when applied to MD simulations and vHTS. To solve this problem, a conceptually new method for computing parameters on-the-fly was developed, based on applying organic chemistry principles. This effort continues the philosophy of previous works conducted by Liu et al (2016, 2018) and Champion et al (2019), and applies it to biaryl molecules. The reason for first focusing on biaryl systems was due to its importance and abundance in pharmaceuticals. With this method, more accurate vHTS and docking could be performed.

This chapter is based on work from: Wei, W. L.; Champion, C.; Liu, Z. M.; Barigye, S. J.; Labute, P.; Moitessier, N., Torsional Energy Barriers of Biaryls Could Be Predicted by Electron Richness/Deficiency of Aromatic Rings; Advancement of Molecular Mechanics toward Atom-Type Independence. J. Chem. Inf. Model 2019, 59, 4764-4777.

Author contributions: I performed the calculations, analysis, and write-up of the manuscript. Zhaomin Liu, Stephen Barigye, Candide Champion, Paul Labute, Prof. Nicolas Moitessier, and I designed the described methods.

3.2 Introduction

Biaryls. Biaryls are a class of conjugated molecules comprised of two aromatic systems, connected by a single bond. These compounds are of particular interest to the pharmaceutical field as they are found in many natural products and pharmaceuticals.[1] In fact, many essential drugs such as atorvastatin, celecoxib, nifedipine, rosuvastatin, valsartan, and others contain biaryl moieties (Figure 3.1). Many natural products containing biaryl fragments, extracted from plants and other organisms, have frequently been shown to possess biological activity.[2] For example, licopyranocoumarin extracted from *Xi-bei licorice* plant, was found to inhibit the cytopathic activity of HIV.[3] These types of molecules are ubiquitous in nature and are important bioactive pharmacophores. Besides their utility in pharmaceuticals, they are also well adapted for use in building polymers, sensors, and transition metal catalysts.[1]

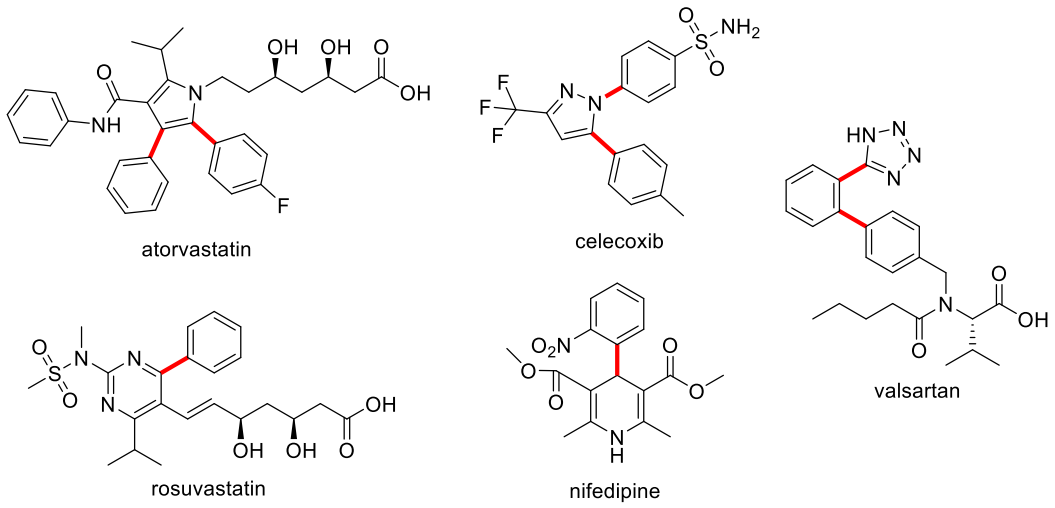


Figure 3.1. Commercially available drug molecules that possess biaryl fragments, with the torsion bonds of interest labelled in red.

Biaryls and molecular mechanics. Despite biaryls being promising pharmacophores and drug scaffolds, existing molecular mechanics (MM) force fields (FF) are not well-adapted to predict their conformational energy landscapes, apart from a few well-parametrized cases.[4, 5] In particular, the torsion parameters of biaryls are lacking for most compounds, due to the diversity of biaryls in drug-like molecules (Figure 3.1).[4] The accuracy of *in silico* docking methods directly rely on a robust torsional energy term for binding affinity calculations between the receptor and the ligand to predict the correct binding mode.[6] Consequently, this presents an interesting but difficult problem for structure-based drug design. Over the past decade, *in silico* docking methods

have proven to be an indispensable tool in the field of drug discovery, allowing millions of compounds to be screened quickly and cheaply.[7] Virtual high-throughput screening (HTS) using docking has allowed an enrichment of active compounds for subsequent synthetic or biological testing efforts.[8-10]

Besides its utility in HTS, a robust FF would also greatly improve the accuracy of molecular dynamics (MD) simulations, where it could be employed to study conformational changes of molecules over time.[11] This methodology has been used in various fields, such as enzyme catalysis,[11] pharmaceutical research,[12] and energy conversion.[13]

Biaryl torsional energy profiles. Torsion energy for biaryl molecules are difficult to predict.[14] While experiments and quantum mechanical (QM) calculations have shed light on the energy profiles and preferred geometries of individual compounds on a case-by-case basis, there have been no holistic, collective approach to studying these biaryls and predicting their torsion barriers. Past work on OPLS FF,[15-17] by Jorgensen and coworkers, empirically fit the torsion parameters to match the QM profiles for thirty-three biaryl compounds most frequently found in pharmaceuticals.[5] Although this allowed the energy profiles of these aforementioned thirty-three compounds to be predicted accurately, the number of possible biaryl compounds (and diversely functionalized biaryls) are several orders of magnitudes higher and may be in the millions. At this rate, it is both impractical and impossible to parametrize each biaryl torsional barrier, individually. In addition, the torsion profile of these biaryls were found to vary drastically and cannot be transferred from one compound to another without a loss of accuracy. Consequently, separate parameters must be developed for each biaryl molecule. For example, while the torsion barrier of biphenyl was approximately 2.1 kcal·mol^{-1}, 2-(1H-pyrrol-1-yl)pyrimidine had a much greater barrier of 11 kcal·mol^{-1}.[5] As a result, MM FF torsional parameters for most biaryl compounds are still inadequately parametrized,[4] and there still exists an unfilled need to accurately model biaryl torsion energy barriers.

In a separate study on biaryls, Stern et al., as a part of *The Open Forcefield Consortium*, found that the Wiberg bond order was an good descriptor of the torsional barrier strength.[18] This descriptor attempts to quantify the degree of bonding between connected atoms.[19] However, calculating the Wiberg bond order requires computationally expensive QM methods. Consequently, it may face challenges when applied to a drug discovery project, which must routinely screen thousands of drug candidates, containing distinct biaryl fragments.

In MM FFs, torsion is one of the bonded terms, along with bond and angle stretching (equation 3.1).[20-22] In combination with non-bonded terms (eg, electrostatics and van der Waals interactions), the linear combination of these terms give rise to the total MM energy, which determines the preferred geometry of molecules, such as *cis*/*trans* propensity. Each of these terms in equation 3.1 are modelled by different potential energy functions. In most FFs, including the General Amber force field (GAFF) 2,[23, 24] a simple elastic potential is used for bond and angle terms, while Coulomb's potential is used for electrostatic interactions, and Lennard-Jones 12-6 potential is employed for van der Waals and steric repulsion. The truncated Fourier series is most commonly employed (equation 3.2) for torsions. Although some FF use up to four terms within this series, only the first two (ie. $i = 1,2$) have been identified as chemically meaningful.[25, 26] The second term, V_2, describes the unmodulated amplitude of torsional energy barrier. In biaryls, V_2 is describes the strength of conjugation. This means that increasing values of V_2 increases the relative energy at dihedral angles of 0˚ and ±180˚ with respect to ±90˚. Effectively, an increase in V_2 decreases the overall strength of conjugation in biaryl systems. On the other hand, the first term describes the *cis*/*trans* preference of a dihedral angle, with V_1 parameter denoting the energy difference between *cis* and *trans* geometry. For biaryl systems, an increasing V_1 increases the relative energy of the *cis* geometry (0˚) with respect to the *trans* geometry (±180˚). This V_1 energy term further modulates and adds upon the V_2 energy term, along with other higher order terms. The third and fourth (ie. $i = 3,4$) are said to be correction terms. It should be noted that $V_{1\text{-}4}$ parameters are divided by 2 to allow these properties to be identified quickly. The γ describes the phase-shift while θ is the torsional angle.

Equation 3.1. Potential energy function of standard MM FFs.

$$E_{MM} = \underbrace{E_{bond} + E_{ang} + E_{tor}}_{bonded} + \underbrace{E_{ele} + E_{vdw}}_{non-bonded}$$

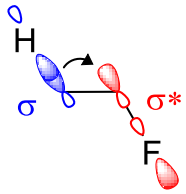
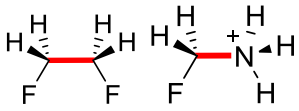
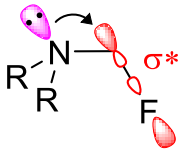
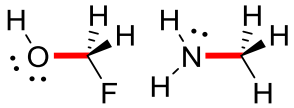
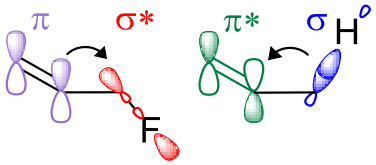
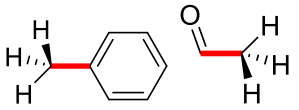
Equation 3.2. The torsional energy function of MM FFs.

$$E_{tor} = \sum_{i=1}^{4} \frac{V_n}{2} [1 + \cos n(\theta - \gamma)]$$

Atom-type independent FFs. In the recent past, several groups, including ours, began developing predictive MM FFs, which are atom type-independent.[25-28] Briefly, an atom type is a concept in MM, which allows parameters developed for one atom to be transferred to another by virtue of similarity of chemical environment, element identity, hybridization, and connectivity. As stated previously, the use of atom types imposes a limit on the accuracy of the FFs due to the issue of transferability. In fact, torsions suffer from one of the worst transferability within bonded interactions,[14, 29] which prompted us to develop torsion terms without depending on atom types. Other groups developed methods, based on QM calculations, to automate and accelerate the process of deriving atom-type parameters.[29-34] These include recent developments such as QUBEKit,[29] QMDFF,[30] GAAMP,[32] ffTK,[33] Paramfit32,[34] and Parmscan33.[31] These methods are well suited for use in MD simulations and free energy perturbations of a few specific receptor-ligand interactions. However, as stated previously, QM calculations are computationally expensive, which may limit its utility for HTS. Consequently, for drug discovery and possibly other fields, the process of deriving torsion parameters needs to be fast, without relying on QM calculations.

To accomplish this task, inspiration was drawn from chemical principles and knowledge accumulated by organic chemists. In the past, our first efforts were directed towards generating torsion parameters on-the-fly for molecules containing simple σ-bonds (eg. haloalkanes, alkylammonium). During that project, it was found that the strength of $\sigma \rightarrow \sigma^*$ hyperconjugation played a major role in determining the height of the torsional energy barrier.[6] It was also shown that this energy barrier could be quantitatively predicted based on the electronegativity of atoms within the dihedral angle. This lead to the development of H-TEQ 1.4 (Hyperconjugation for Torsional Energy Quantification 1.4), a standalone program based on the quantification of chemical principles to derive torsion parameters on-the-fly (Table 3.1). After this initial success, a similar strategy was used to predict the torsion barriers of molecules containing electron lone pairs in the central atom of the dihedral angle (eg. methanol, methylamine); thereby quantifying $\eta \rightarrow \sigma^*$ hyperconjugation.[25] This effort lead to the development of H-TEQ 2.[25] In our last report, we began quantifying torsion interactions between benzylic and allylic bonds adjacent to conjugated systems, by looking at $\pi \rightarrow \sigma^*$ and $\sigma \rightarrow \pi^*$ hyperconjugation, which lead to the development of H-TEQ 3.0.[35]

Table 3.1. Previous Developments of H-TEQ.

H-TEQ	Supported Torsional Bonds	Hyperconjugation Type	Example Molecules
1.4[6]	Saturated alkanes without adjacent electron lone pairs		
2.0[25]	All saturated alkanes		
3.0[35]	Single bonds adjacent to a double bond without adjacent electron lone pairs		

3.3 Understanding Chemical Origins

Qualitative predictions. Many physical and chemical properties have been employed by organic chemists to qualitatively explain observations of conformational preferences and reaction outcomes for decades. For example, hyperconjugation could explain gauche and anomeric effects,[36] while aromaticity gives a rationale for the superior stability of benzene compared to hexatriene.[37] In addition, hard and soft acids and bases (HSAB) theory has been used to account for the speeds of reaction and stability of formed products.[38] Unfortunately, despite the abundance of organic chemistry knowledge, many of these interactions remain qualitative. By understanding the underlying chemical principles, we aimed to quantify these interactions so that they could be applied to FFs.

Brief Overview of Hyperconjugation. Torsional energy profiles of molecules are primarily influenced by the effects of hyperconjugation, conjugation, sterics, and electrostatics.[4,25, 26, 39, 40] Although most existing FFs contain the two latter terms, the strengths of the two former terms have not been previously explored nor quantified.

In the context of hyperconjugation, two major factors modulate the strengths of stabilization: spatial overlap and molecular orbital energy match between the donor bonding and acceptor

antibonding orbitals.[41, 42] Spatial overlap dictates that hyperconjugation is maximized when the acceptor and donor bonds are properly aligned and overlap sufficiently to allow donation of electron density from the filled to the unfilled orbital. Taking fluoroethane as an example, the donation of $\sigma_{C\text{-}H} \rightarrow \sigma^*_{C\text{-}F}$ hyperconjugation is maximum when the $\theta_{H\text{-}C\text{-}C\text{-}F}$ angle is 180˚ (i.e. C-H pointing opposite in direction to C-F).[36] On the other hand, at 0˚, this donation is much weaker. Furthermore, when these orbitals are orthogonal, donation is virtually impossible due to having little or no overlap (Figure 3.2). This requirement for orbital overlap is true for all types of hyperconjugation. The orbital energy match of the donor and acceptor orbitals also plays a role in the strength of this interaction. Since the antibonding, unfilled orbitals are higher in energy than the bonding, filled orbitals, the former must be energetically accessible for donation of electron density to occur (Figure 3.3). Hence, a smaller energy gap increases the strength of hyperconjugation. In general, electronegative and electropositive atoms are known to lower and increase the overall energy levels of molecular orbitals, respectively.[37, 41, 43-45] Consequently, for stronger hyperconjugation interactions to occur, a more electronegative atom should comprise the acceptor orbital, while a more electropositive atom should be incorporated in the donor orbital.

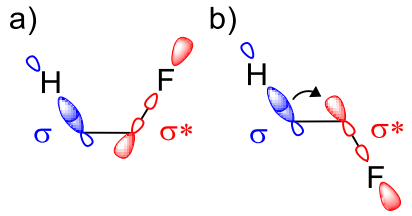


Figure 3.2. Two distinct conformations of fluoroethane: (a) $\theta_{H\text{-}C\text{-}C\text{-}F} = 180^\circ$ and (b) $\theta_{H\text{-}C\text{-}C\text{-}F} = 0^\circ$, are shown along with bonding (blue) and antibonding (red) orbitals participating in $\sigma_{C\text{-}H} \rightarrow \sigma^*_{C\text{-}F}$ hyperconjugation depicted. For simplicity, not all bonds nor orbitals are shown.

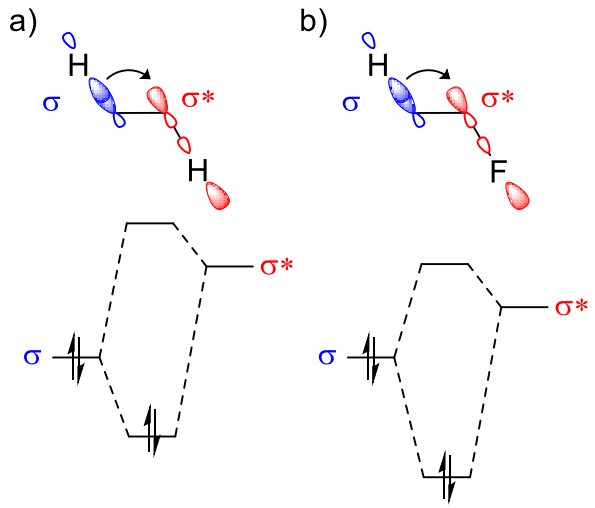


Figure 3.3. σ → σ* hyperconjugation in a) ethane and b) fluoroethane with selected orbitals (bonding orbital is shown in blue, while antibonding orbital is shown in red) and energies shown. In both cases, the σ-bond orbital energy level is the same, while the σ*-antibonding orbital energy for fluoroethane is lower than ethane. This causes a greater amount of hyperconjugative stabilization for fluoroethane compared to ethane.

Chemical Origins of Hyperconjugation and Conjugation in Biaryl Systems. Biaryl molecules are comprised of two aromatic cycles connected by a single bond. Consequently, various types of hyperconjugation effects exist, as exemplified by 2-(pyridin-2-yl)oxazole (Figure 3.4). The various orbital interactions of biaryl molecules affecting the central torsion, include σ → σ*, π → σ*, σ →π* hyperconjugation, and π → π* conjugation.[37, 46] In biaryl systems, the σ → σ* hyperconjugation (Figure 3.4b) and π → π* conjugation (Figure 3.4d) are in-phase and combined constructively. These interactions are maximum when the biaryl system is planar. In this orientation, the spatial overlap between σ and σ* orbitals and also the π-orbitals of both aromatic rings are at a maximum. Similarly, π → σ* and σ → π* hyperconjugation (Figure 3.4c) are also in-phase but are maximum when the two rings are orthogonal in orientation. As a result, the former, planar-favoring interactions and the latter, orthogonal-favoring interactions combine destructively, and compete with each other to modulate the overall strength of the biaryl torsional barrier.

It should also be noted that due to the geometric orientation of these orbitals, it is expected that the π and σ*-spatial overlap (Figure 3.4e) would be greater than the σ and π* spatial overlap (Figure 3.4d), leading to a greater stabilization from the former.

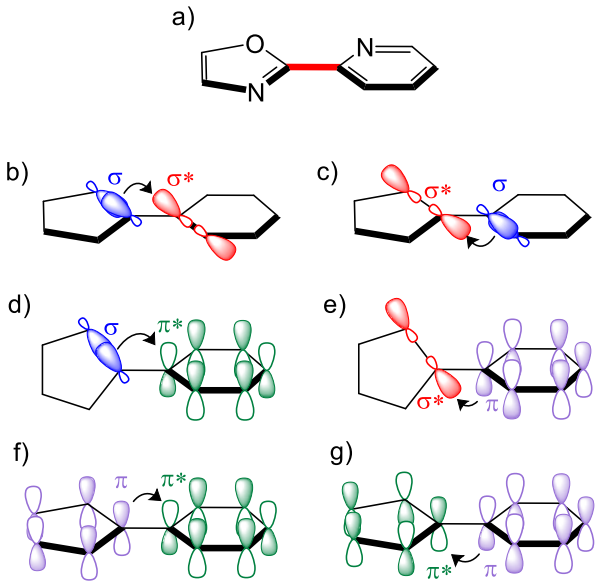


Figure 3.4. (a) Chemical structure of 2-(pyridin-2-yl)oxazole (with the torsional bond of interest highlighted in red) is shown with hyperconjugation of orbitals involving: (b and c) σ → σ*, (d) σ →π*, and (e) π → σ *, and (f and g) π → π* conjugation. For simplicity, a maximum of one σ-, σ*-, π-, and π*-orbitals are shown in each subfigure.

In biaryl systems, stabilization due to π-orbitals are derived from π → π* conjugation.[37] Unlike simple σ- and σ*-orbitals, the π- and π*-orbitals of aromatic systems are arranged in a more complex manner. In fact, Hückel molecular orbital theory predicted that the molecular orbitals of classical 5- and 6-membered aromatic cycles possess degenerate energy levels (eg. benzene and cyclopentadienyl anion).[37] Consequently, more than one possible π → π* transition is present between the two rings, and often in either direction. For example, in a biphenyl molecule (and any other biaryls composed of two 6-membered rings), there are 18 theoretically possible conjugation interactions (Figure 3.5a). For 2-phenylthiophene and any other biaryls composed of a 5- and a 6-membered aromatic cycle possesses 15 theoretically possible conjugation interactions (Figure 3.5b). Likewise, 2,2'-bithiophene, along with other biaryls with two 5-membered aromatic rings, has 12 theoretically possible transitions (Figure 3.5c). These interactions involve the conjugation interaction involving the donation of electron density from an occupied molecular orbital of one ring to an unoccupied molecular orbital of the adjacent ring. In all of these three cases however, the strongest conjugation interactions come from transitions that occur between donor and acceptor orbitals with the smallest energy gaps. Consequently, $\psi_2 \rightarrow \psi_5^*$, $\psi_2 \rightarrow \psi_6^*$, $\psi_3 \rightarrow \psi_5^*$, and $\psi_3 \rightarrow \psi_6^*$ transitions are most favourable to occur (Figure 3.5). Despite knowing the theoretical basis

from an FMO perspective, it is unclear how the substitution of one atom for a more electronegative one would affect the overall strength of conjugation. This is complicated by the number of possible transitions in conjugated systems, which will be explored in this manuscript.

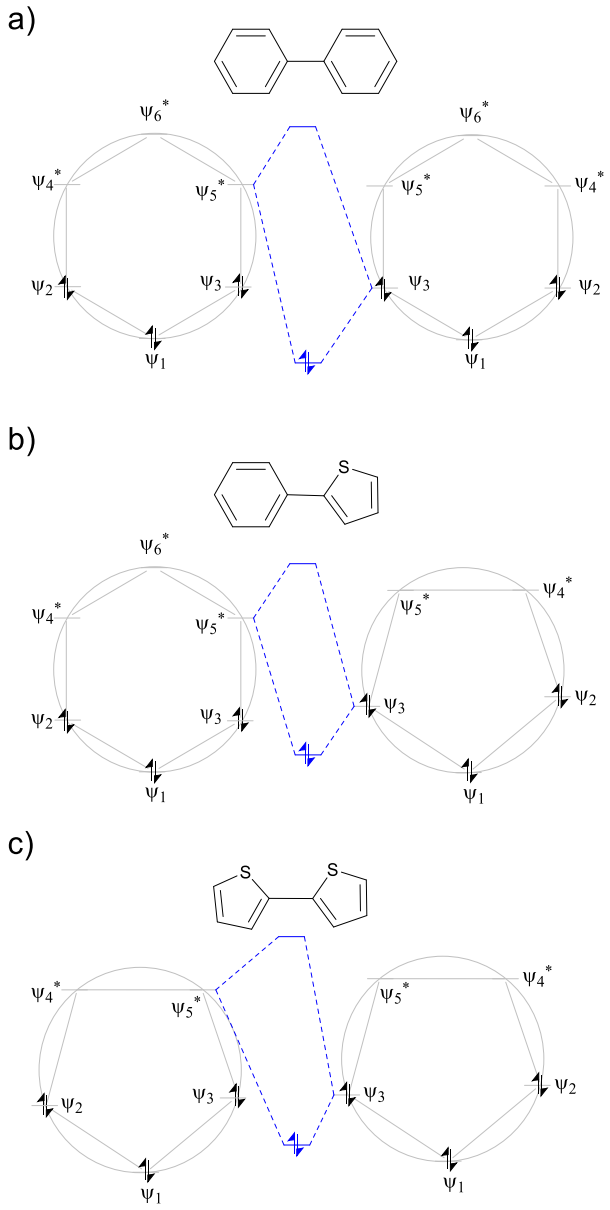


Figure 3.5. A qualitative molecular orbital diagram of: a) biphenyl molecule, b) 2-phenylthiophene, and c) 2,2'-bithiophene are shown, along with a major interaction, involving ψ3 → ψ5* conjugation. The right and left rings act as the donor and acceptor, respectively in this figure. For simplicity, not all possible transitions are shown.

3.4 Computational Methods

Construction of Development Set. 131 biaryl molecules of various types were included in the development set (Figure 3.6). The composition of this set included biaryls containing both 5- and 6- membered rings (ie., 5:5, 5:6 and 6:6 biaryls), as well as neutrally and positively charged central atoms. A suitable starting conformation was obtained for each molecule by performing a global optimization at the MP2/6-311++G** level of theory using GAMESS-US.[47] Subsequently, a constrained QM optimization was performed whereby the dihedral angle of interest was varied sequentially between -180˚ and 180˚ at 10˚ intervals. This resulted in a total of 36 conformations for each biaryl molecule. Single point energies for each conformation was then evaluated using MP2/6-311++G** and GAFF2. AM1-BCC charges were assigned to molecules by using the default protocol implemented in Antechamber for each torsional conformation.[48] These charges were subsequently averaged for each atom throughout the torsional angle scan to preclude discontinuity in the electrostatic energy profile nor to favor any one particular geometry.

To obtain the isolated torsional energy, the van der Waals and electrostatics components of GAFF2 energies were subtracted from the total energy as computed by MP2, which included 1-4 non-bonded interactions. A Fourier regression was performed to obtain $V_{1\text{-}3}$ parameters (equation 3.2) for each molecule in the development set. A variety of different chemical descriptors, including electron-richness/deficiency, central atom bond lengths, and electronegativity, were assigned to each molecule to observe their effects on the torsional barrier. The descriptor(s) which could best reproduce the torsion energy was chosen to formulate our method, which was subsequently incorporated into H-TEQ 4.0, our standalone program. To first evaluate whether these rules could reproduce the QM energy profiles of torsional rotation, H-TEQ 4.0 was first tested on the development set. The total MM energy was calculated by taking the GAFF2 energies for all terms in Equation 3.1, except for the torsion energy, which was computed by H-TEQ 4.0. Root-mean-square error (RMSE) was calculated for H-TEQ 4.0 and GAFF2 with the QM energy profile as a reference.

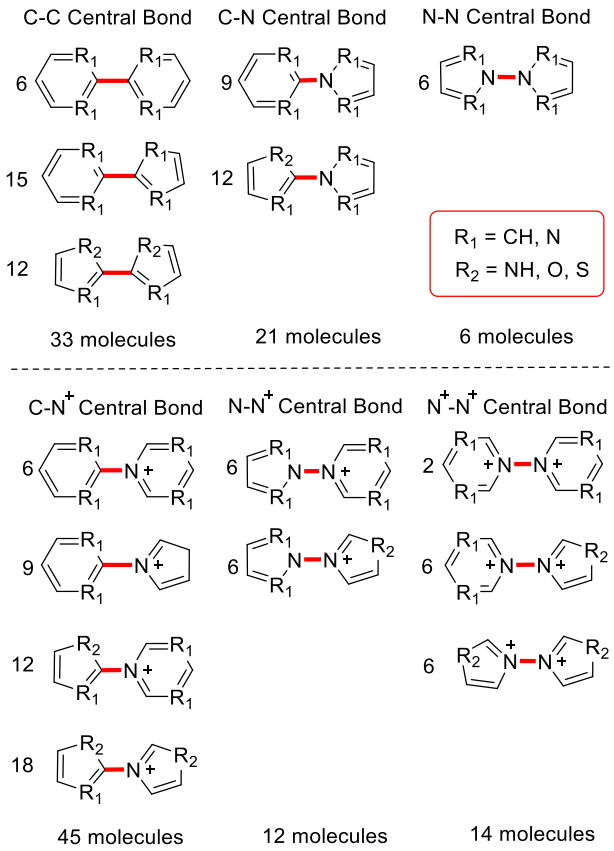


Figure 3.6. 131 biaryl molecules used for obtaining the torsion energy within the development set. The number on the left of each scaffold indicates the number of molecules contained in that category.

Construction of the Validation Set. 100 molecules were selected to be a part of the validation set. The composition of this set included drug-like biaryls used by Jorgenson and coworkers (20 molecules),[5] the MMFF94 set (32 molecules),[49] and a variety of molecules chosen from a previous Cytochrome P450 set (48 molecules).[50] In order to ensure the robustness of the developed method, the validation set was constructed so that it would have no overlapping molecules with the training set. From these three sources of molecules, 100 drug-like molecules were randomly selected and tested. These molecules also included those which were highly substituted, bicyclic, and tricyclic. The full validation set of molecules are shown in Figure 3.7.

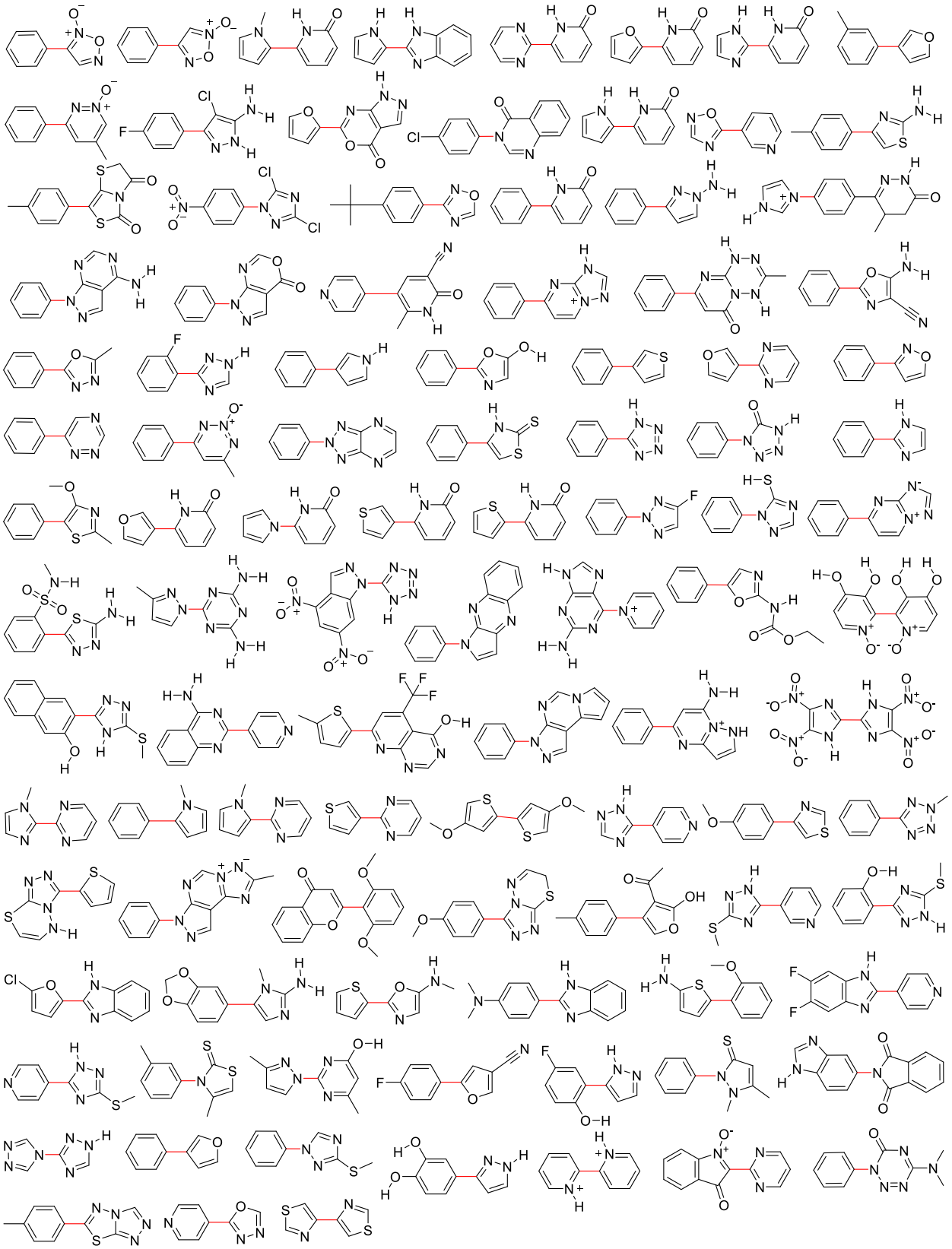


Figure 3.7. 100 molecules of the compiled validation set. The bonds of interest are shown in red.

3.5 Results and Discussion

Chemical Factors Modulating the Strength of Conjugation in Biaryl Molecules. It is known that the degree of π-electron delocalization in aromatic systems is influenced both by the number of available π-electrons and the electronegativity of the atoms comprising the π-system. For example, 5-member rings, such as pyrroles, are generally more electron-rich than 6-member rings, such as pyridine.[51-53] This is due to the fact that in the former, six π-electrons are distributed over five atoms as opposed to the latter where the same number of π-electrons are distributed over six atoms. As a result, it leads to a higher π-electron density in the former. In addition, it is also known that aromatic molecules with less electronegative atoms, such as pyrroles are more electron-rich than molecules with greater electronegative atoms such as furans. This is due to the propensity of more electronegative atoms to withhold π-electron density. To take both of these factors into consideration and to measure the electron-richness/deficiency of aromatic systems, a new electronegativity parameter, π-electronegativity (χ_π) was devised as shown in equation 3.3. Furthermore, two other descriptors, $\chi_{\pi tot}$, and $\Delta\chi_\pi$ were derived from equation 3.3, as shown in equation 3.4 and 3.5, respectively.

Equation 3.3. Equation for calculating the π-electronegativity of a single aryl group.

$$\chi_\pi = \frac{\sum_{i=0}^{n} \chi_i}{n(\pi)}$$

Equation 3.4. Equation for calculating the total π-electronegativity of a biaryl molecule.

$$\chi_{\pi tot} = \chi_{\pi\ Group2} + \chi_{\pi\ Group1}$$

Equation 3.5. Equation for calculating the π-electronegativity difference within a biaryl molecule

$$\Delta\chi_\pi = \chi_{\pi\ Group2} - \chi_{\pi\ Group1}$$

In this equation, χ_i is the electronegativity of atoms comprising the conjugated system and $n(\pi)$ is the number of π electrons of this system.

Overall π-electronegativity Modulates Strength of Conjugation. Since biaryl molecules are composed of two interconnected aromatic moieties, the χ_π of both cyclic systems were computed. In all tested cases, it was found that the π-electronegativity of the entire molecule, $\chi_{\pi tot}$, correlated well with V_2, within each category of molecules with differing central bonds (Figure 3.8). More specifically, it was found that an increase in the $\chi_{\pi tot}$ correlated with an increased magnitude of V_2. Since large χ_π suggests π-electron deficiency, these results indicated that biaryls which were more π-electron deficient had a greater degree of conjugative stabilization. Although additional calculations and experiments may be required, the increase in $\chi_{\pi tot}$ may decrease the energy gap between the ψ_2 or ψ_3 bonding and ψ_4* or ψ_5* antibonding molecular orbitals of biaryl systems (Figure 3.5). In effect, this would decrease the energy barrier required for electron donation from the bonding to the antibonding orbitals, leading to increased conjugation strength. This relationship between $\chi_{\pi tot}$ and V_2 was true within each category of biaryls tested, although the associated slope and y-intercept of the linear relationship and accuracy of the regression differed. For C-C, C-N, and N-N central bonds, squared correlation coefficients (R^2) of 0.48, 0.72, and 0.19 were found, respectively. For positively charged central bonds, C-N^+, N-N^+, and N^+-N^+, R^2 of 0.72, 0.65, and 0.60 were observed, respectively.

Interestingly, the slopes of the linear fit across each group varied slightly. In general, it was observed that biaryls with neutral central bonds (ie. C-C and C-N) had greater stabilization, and therefore lower V_2 than those with charged central atoms (ie. C-N^+ and N-N^+), despite having the same $\chi_{\pi tot}$ (Figure 3.8). Furthermore, both these aforementioned groups had greater conjugative stabilization when compared to those with large charge-charge repulsion within the central bond (N-N and N^+-N^+). Biaryls, comprising of N-N central bond (eg. 1,1'-bipyrrole), contain two adjacent lone pairs, which repel strongly when planar, as was also found in a previous study.[54] However, another explanation may be that a greater $\pi \rightarrow \sigma^*$ and/or $\sigma \rightarrow \pi^*$ hyperconjugation are present in these biaryls, although the exact mechanism of action is not clearly known.

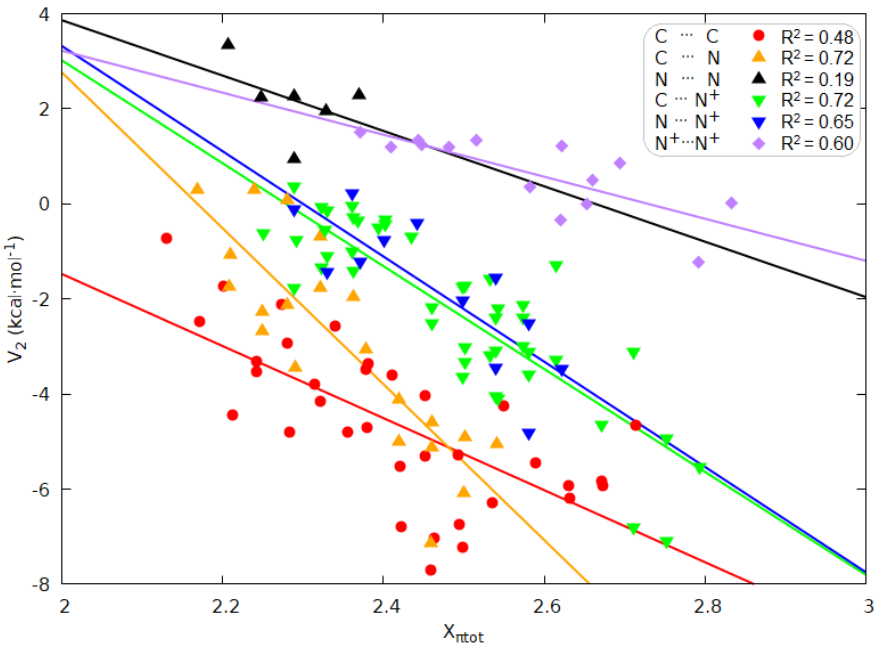


Figure 3.8. V_2 of 131 biaryl molecules were plotted against the total π-electron density ($\chi_{\pi tot}$), and categorized into six distinct groups, which differed by the atomic identity of the central bond. Smaller V_2 signifies stronger conjugation strength. The R^2 for each linear regression is also shown in the legends.

Difference in π-electronegativity between the Two Aromatic Rings Modulates Strength of Conjugation. In addition, the difference in π-electronegativity, $\Delta\chi_\pi$, between the two aromatic groups of biaryl molecules was investigated. Intriguingly, in some cases, an increase in $\Delta\chi_\pi$ increased V_2 while in others, a decrease was observed (Figure 3.9). Only for N^+-N^+ central bonds, was there no observable trend, which will be mentioned and explained later in this section. For biaryls with uncharged central bonds (ie. C-C, C-N, and N-N), increases in $\Delta\chi_\pi$ decreased V_2, leading to an increase in the strength of conjugative stabilization. In fact, when V_2 was plotted linearly against $(\Delta\chi_\pi)^2$, the squared correlation coefficients of 0.28, 0.69, and 0.73 were observed for biaryls containing C-C, C-N, and N-N central bonds, respectively (Figure 3.9A). For biaryls containing charged central bonds, trends were also found although they seemed to be influenced by another factor—the number of atoms comprising the aromatic moiety.

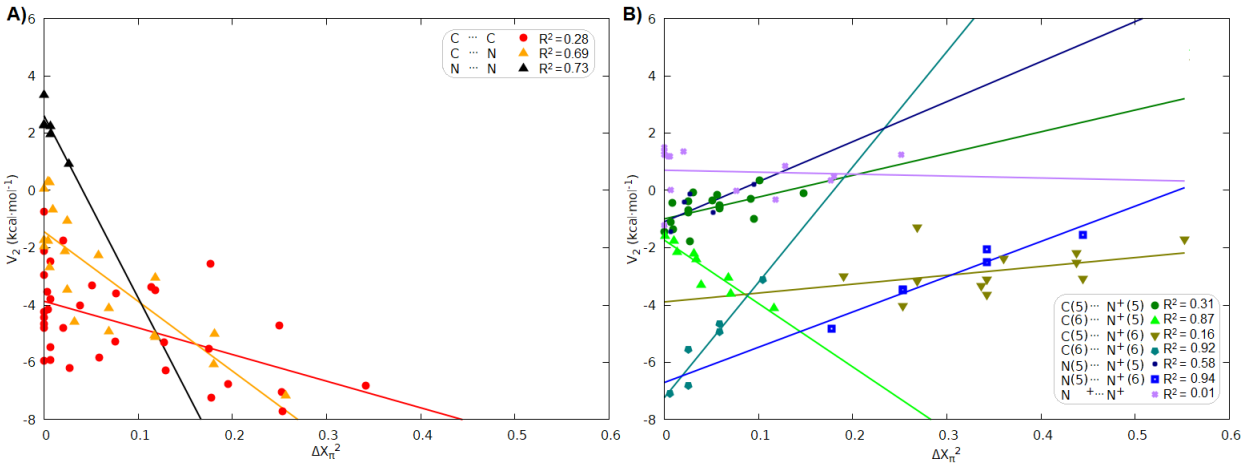


Figure 3.9. V_2 of 131 biaryl molecules were plotted against the squared difference of π-electron density between the two connected aromatic moieties $(\Delta\chi_\pi)^2$. Biaryls were separated into: A) neutral central bonds: C-C, C-N, and N-N, and B) positively charged central bonds: C-N+, N-N+, and N+-N+. C-N+ and N-N+ was further grouped by the identity of the central atom and number of atoms on each side of the central bond. Smaller V_2 signifies stronger conjugation strength. The R^2 for each linear regression is also shown in the legends.

For biaryls containing C-N^+ central bonds, an increase in $(\Delta\chi_\pi)^2$ decreased V_2 (Figure 3.9B), for those comprised of a 6-membered aromatic moiety with carbon central atom and a 5-membered aromatic moiety with a N^+ central atom (ie. C(5)-N^+(6)). All other biaryls with C-N^+ central bonds saw an increase in V_2 as $(\Delta\chi_\pi)^2$ increased.

Similarly, an increase in $(\Delta\chi_\pi)^2$ increased in V_2 for all N-N^+ central bond biaryls, regardless of the number of atoms in each ring. However, the trends for both N(5)-N^+(5) (R^2 = 0.58) differed from N(5)-N^+(6) (R^2 = 0.94) such that the slope and y-intercept differed. Although no relationship seemed to exist for N^+-N^+ central bonds, it was hypothesized that such a trend would be observed if it were sorted into N^+(5)-N^+(5), N^+(5)-N^+(6), and N^+(6)-N^+(6) central bonds. However, since these biaryls are rarely found in pharmaceuticals, the existing number of molecules was deemed sufficient for this current study, and not pursued further.

The relationship between $\Delta\chi_\pi$ and V_2 may also be caused by the lowering of the energy gap between adjacent bonding and antibonding orbitals of biaryl systems (Figure 3.5), leading to a more stable conjugation. In general, a greater $\Delta\chi_\pi$ in a biaryl system would suggest that the aromatic ring with the lower χ_π would act as the donor, while the aromatic ring with the higher χ_π

would act as the acceptor. However, this relationship did not hold for all central bonds, possibly due to more complex molecular orbital interactions present in certain biaryls.

Substituent Effects on the Torsional Profile of Biaryls. Substituent effects of biaryl molecules on its barrier to conjugation was tested on various molecules, which were categorized into two types of substituents: inductive and conjugated. To test the effects of the inductive ligands on biaryl molecules, mono- and di-substituted biaryls were tested. More specifically, various substituents were placed at various positions on the biaryl molecules, as shown in Figure 3.10. In total, 15 different scaffolds and 124 substituted biaryls were tested. Although some di-substituted biaryls lead to a change in its torsion energy by as much as 1.5 kcal·mol^{-1}, most substituents had negligible effects. In fact, the average change of 8.12% was observed from the original, unsubstituted biaryl molecule with 86% of substituted biaryls showing less than 15% change in V_2 (Table 3.2).

To further illustrate this point, fluoryl, chloryl, methyl, and trifluoromethyl groups were placed in the meta- and para-positions of the 1,1'-biphenyl molecule, individually and in combination. These results further collaborated that inductive ligands have minor effects on the V_2 of biphenyl molecules. In fact, the differences in V_2 of substituted 1,1'-biphenyl to that of the unsubstituted never exceeded 0.55 kcal·mol^{-1} (Table 3.2). Surprisingly, the V_2 term of 2,2'-difluoro-1,1'-biphenyl and 3,3'-difluoro-1,1'-biphenyl only differed by 0.02 and 0.12 kcal·mol^{-1}, respectively, when compared to 1,1'-biphenyl, in spite of being substituted by two highly electronegative atoms. Similar negligible increases in V_2 were observed with 4-(trifluoromethyl)-1,1'-biphenyl, which had an increase of only 0.12 kcal·mol^{-1} over the unsubstituted biphenyl. This may be due to the fact that σ-orbitals of the inductive substituent and the highly delocalized π-orbitals of the biaryl system may not interact very well due to their energetic differences.

Conjugative substituents were further evaluated using a set of 1,1'-biphenyl molecules. Amine, dimethylamine, nitro, and alcohol groups were added to the para-position by themselves and in combination. The results indicate that these substituents also had little effect on the overall torsional energy barrier to rotation, despite having π-electrons which could readily interact with the π-system. In fact, the highest difference in V_2 in this class of molecules was *N*,*N*-dimethyl-4'-nitro-[1,1'-biphenyl]-4-amine, which saw an increase of 0.57 kcal·mol^{-1} over the unsubstituted biphenyl (Table 3.2, Figure 3.11). As shown at the bottom of Figure 3.11, this increase can be

explained by the combination of a strong electron-donating group and a strong electron withdrawing group stabilizing a resonance structure leading to a strong double bond character of the central bond. However, this was somewhat surprisingly as it was expected that this combination would increase the torsional barrier more substantially. This further confirmed that conjugative substituents had little effect on the torsional energy barrier of biphenyl.

Overall, both types of substituents: inductive and conjugated, had minor effects on the torsion profile of the biphenyl molecule which was comparable to the level of error found in MM FFs, and could potentially be caused by errors in the electrostatics or van der Waals energies. We understand that other biaryls may behave slightly differently in the context of substitution, and expect some modulation of the overall torsion energy profile. However, the presence of substituted ligands did not greatly impact the torsional energy profile of biaryls. Consequently, these effects were not further pursued nor considered to build our predictive model for estimating the torsional profiles of biaryl molecules. Only the principle aromatic rings of biaryl, directly adjacent to the σ-bond, was considered.

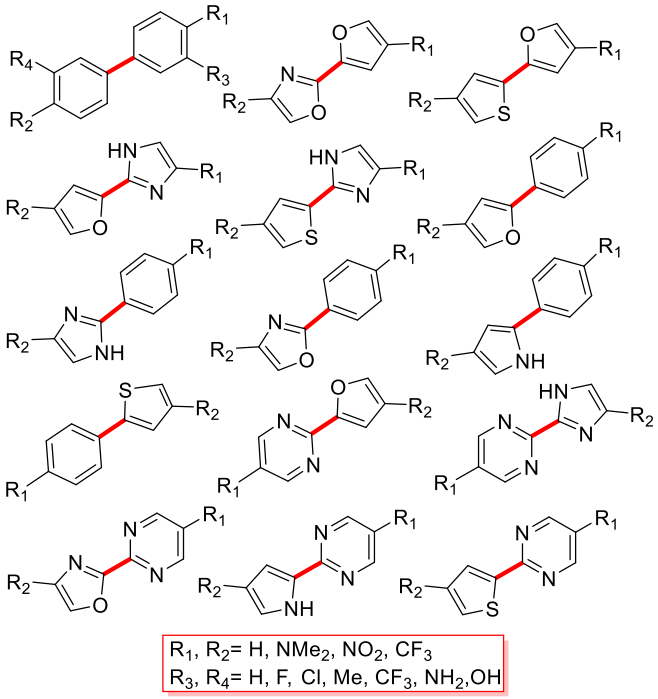


Figure 3.10. Biaryls used to probe the effects of substituents on its torsional energy barrier. The torsional bonds of interest are depicted in red.

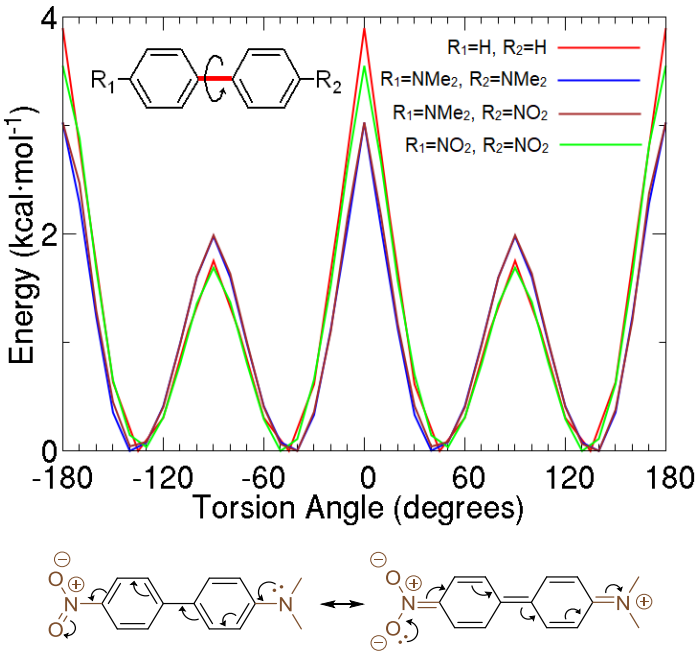


Figure 3.11. The torsional energy to rotation of biphenyl (red), and biphenyl substituted with strong electron-withdrawing and electron donating groups as computed by MP2/6-311++G(d,p) level of theory.

Table 3.2. V_2 for Various Substituted 1,1'-Biphenyl Molecules

Biaryl Molecules	V_2 (kcal·mol^{-1})
1,1'-biphenyl	-4.25
Inductive Substituents	
3-fluoro-1,1'-biphenyl	-4.29
3-chloro-1,1'-biphenyl	-4.56
4-fluoro-1,1'-biphenyl	-4.26
4-chloro-1,1'-biphenyl	-4.64
4-methyl-1,1'-biphenyl	-4.27
4-(trifluoromethyl)-1,1'-biphenyl	-4.37
3,3'-difluoro-1,1'-biphenyl	-4.37
3,3'-dichloro-1,1'-biphenyl	-4.80
3,4'-difluoro-1,1'-biphenyl	-4.31

4,4'-difluoro-1,1'-biphenyl	-4.23
4,4'-dichloro-1,1'-biphenyl	-4.69
4,4'-dimethyl-1,1'-biphenyl	-4.23
Conjugated Substituents	
[1,1'-biphenyl]-4-amine	-4.39
[1,1'-biphenyl]-4-ol	-4.38
benzidine	-4.52
[1,1'-biphenyl]-4,4'-diol	-4.46
N4,N4,N4',N4'-tetramethyl-[1,1'-biphenyl]-4,4'-diamine	-4.28
N,N-dimethyl-4'-nitro-[1,1'-biphenyl]-4-amine	-4.82
4,4'-dinitro-1,1'-biphenyl	-4.79

***Cis/Trans* Preference of Conjugation.** The V_1 term, responsible for the cis/trans preference of a torsion term, was also investigated for the 131 biaryls in the development set. 81% of these molecules (106 out of 131 molecules) had a V_1 term between the ranges of -1 to +1 kcal·mol^{-1}, suggesting that V_1 terms do not vary greatly for the majority of molecules. For the other 19% of biaryls, no discernable trends were observed when plotted against various chemical properties.

From a chemical perspective, *cis*/*trans* preference is dictated by various chemical effects, including: electrostatics, van der Waals, and σ → σ* hyperconjugation. According to the obtained results, the strength of the V_1 energy term was found to be minor compared to electrostatics and van der Waals interactions. This is reflected in the energy profiles of the tested biaryl molecules. For example, 2,2'-bipyridine had combined electrostatics and van der Waals energy which varied over a range of 10.65 kcal·mol^{-1} (Figure 3.12). In contrast, the V_1 energy term, representing the σ → σ* hyperconjugation, varied over a much smaller range of 1.81 kcal·mol^{-1}. For comparison, the V_2, term varied over a range of 5.83 kcal·mol^{-1}. These results suggest that the *cis*/*trans* preference in biaryls are predominately a result of van der Waals and electrostatic interactions, as reported previously.[4] In addition, although these latter terms are widely used within MM FFs, it is known that they also suffer from various sources of errors and simplifications.[14] For example,

electrostatics in most common FFs cannot be polarized, while van der Waals interactions often have a steeper steric wall than in reality. As a result, in addition to hyperconjugation effects, the obtained V_1 terms may also include correction terms for the errors found in non-bonded interactions of FFs. Consequently, for this reason, a method to predict the V_1 variable was not included in the developed method.

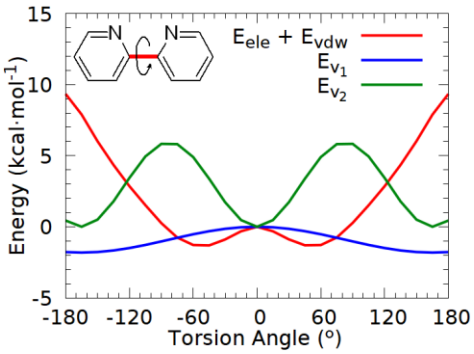


Figure 3.12. The energy terms extracted from a torsion scan of 2,2'-bipyridine. The red represents the sum of the GAFF2 van der Waals and electrostatics terms. The blue and green are V_1 and V_2, respectively, which were obtained from the torsion term as computed by $E_{QM} - (E_{vdw} + E_{ele})$.

Development of H-TEQ 4.0 Rules for Biaryl Molecules. With these observations in hand, rules were developed for all investigated biaryl molecules. In order to improve the accuracy of the method, separate rules were devised for each distinct central bond biaryls. V_2 was predicted based on a linear combination of the overall π-electronegativity and difference of π-electronegativity between the two aromatic groups, based on the equation 3.4, and its associated parameters (Table 3.3).

Equation 3.6. Equation for calculating V2 based on electron-richness/deficiency.

$$V_2 = A \cdot \chi_{\pi tot} + B \cdot (\Delta\chi_{\pi})^2 + C$$

Table 3.3. Associated Parameters to Reproduce V_2 for Various Categories of Biaryls

Central Atoms	A	B	C
C-C	-7.13	-7.75	13.21
C-N	-10.43	-13.37	22.38
N-N	-5.74	-63.81	15.75
C(5)-N^+(5)	-10.82	0	24.66
C(5)-N^+(6)		0	24.66
C(6)-N^+(5)		-20.46	25.67
C(6)-N^+(6)		34.36	22.60
N(5)-N^+(5)	-11.07	15.13	25.05
N(5)-N^+(6)		10.01	22.33
N^+-N^+	-5.109	0	14.22

In this equation, A and B are parameters for the weights of contribution of $\chi_{\pi tot}$ and $(\Delta\chi_\pi)^2$, respectively. These values were optimized through a multivariable regression to reproduce V_2 together; consequently, they were not the same parameters as those found in the previous section, when they were used separately. C represents the inherent strength of conjugative stabilization for a particular type of biaryl molecule. The value of these associated parameters (ie. A, B, C) were different, depending on the atomic identity of the central bond. It should also be noted that in the case of biaryls with both C-N^+ and N-N^+ central bonds, B and C parameters were also different depending on the number of atoms in each aromatic ring. Consequently, biaryls with C-N^+ central bonds were further separated into four categories of molecules: C(5)-N^+(5), C(6)-N^+(5), C(5)-N^+(6), and C(6)-N^+(6), and N-N^+ central bonds into 2 groups: N(5)-N^+(5) and N(5)-N^+(6).

Accuracy of H-TEQ 4.0 on Training Set- A First Validation. Prior to obtaining the accuracy of the developed method on the validation set, the accuracy of H-TEQ 4.0 was first tested

on molecules found within our development set. An average RMSE of 0.71 kcal·mol^{-1} was obtained for H-TEQ, as compared to 4.83 kcal·mol^{-1} for GAFF2 (Figure 3.13) for the 131 biaryl molecules. These results indicated the robustness of the developed method. More interesting, H-TEQ 4.0 predicted a sharp peak at an RMSE of 0.5 kcal·mol^{-1}. Overall, all of the H-TEQ 4.0 predicted torsional parameters fell within an RMSE of 2.5 kcal·mol^{-1}. In contrast, GAFF2 had a broader distribution containing several maximums, with peaks found at RMSEs of 0.7, 3.7, 6.7, and 7.9 kcal·mol^{-1}. In fact, there were RMSEs of up to 19 kcal·mol^{-1}. These results suggested that while GAFF2 had been parameterized for some molecules, corresponding to the peak of RMSE at 0.7 kcal·mol^{-1}, it performed poorly for other molecules which were not parameterized. These failures may also results from a poor atom type assignment by antechamber. Within the compiled training set of 131 biaryls, GAFF2 predicted 36% of the torsional energy profiles below an RMSE of 2.5 kcal·mol^{-1}.

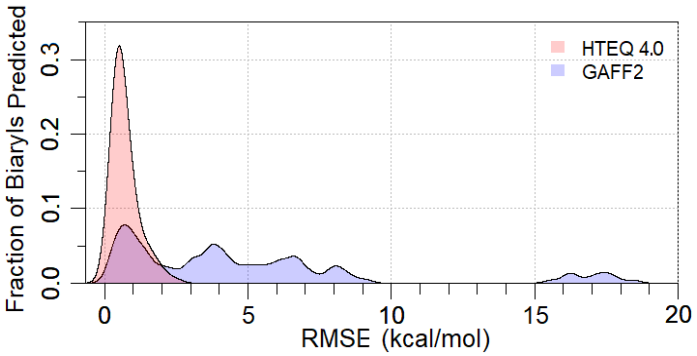


Figure 3.13. A smoothed histogram indicates the number of MM energy profile predictions made by H-TEQ 4.0 (pink) and GAFF2 (blue) within the training set for 131 biaryls. The RMSEs of these predictions are shown on the x-axis. The QM energy profile was used as the reference, with values closer to 0 kcal·mol^{-1} representing more accurate predictions.

Validation of H-TEQ on a Diverse Set of Drug-like Molecules. To test the accuracy of the newly developed H-TEQ 4.0, it was applied to the newly compiled set of 100 biaryl molecules (Figure 3.7). The results showed that the developed method had a higher overall accuracy when compared to GAFF2 (Figure 3.14) with mean RMSEs of 0.95 compared to 3.88 kcal·mol^{-1}, respectively, with reference to the QM torsional profile. In fact, these results indicated that the statistical mode of the RMSE distribution for H-TEQ 4.0 (0.55 kcal·mol^{-1}) was lower than GAFF2 (0.80 kcal·mol^{-1}).

In rare cases, GAFF2 performed better than H-TEQ4 for some common biaryls, such as 6-phenylpyridin-2(1H)-one (**1**), which had an RMSE of 0.17 compared to 0.97 kcal·mol^{-1}, respectively (Figure 3.15a). This is likely due to GAFF2 having been parametrized extensively for these molecules. When the torsion of a slightly more complex molecule, such as 6-(1H-imidazol-2-yl)pyridin-2(1H)-one (**2**), was computed however, GAFF2 drastically overpredicted the torsional barrier by approximately 25 kcal·mol^{-1}, with an RMSE of 14.15 kcal·mol^{-1} (Figure 3.15b). This particular biaryl had never been parametrized by GAFF2, leading to an erroneous use of a generic double bond torsion parameter, in place of a single bond parameter. In fact, 4,4'-bithiazole (**3**) and 6-(furan-2-yl)pyrazolo[3,4-d][1,3]oxazin-4(1H)-one (**5**) also had similar parameter transferability issues (Figure 3.15c and e), leading to an RMSE of 16.95 and 14.17 kcal·mol^{-1}, respectively when computed by GAFF2. H-TEQ 4.0, being atom type-independent, predicted a more accurate torsional barrier for these molecules, with RMSEs of 0.80, 1.10, and 0.41 kcal·mol^{-1} for **2, 3,** and **5**, respectively.

Despite more accurately predicting the height of the torsion barrier associated with conjugation for molecule **1**, **2**, and **3**, some parts of the H-TEQ 4.0 curve also deviated slightly from that of the QM profile. This might be due to the fact that the van der Waals term in GAFF2, used as a part of the H-TEQ 4.0 energy could not capture the nuance orbital interactions at close distances.[14] When in the planar geometry, molecule **1** had close steric clash between its two pairs of hydrogen atoms, leading to an energetically high barrier at 0˚ and ±180˚ (Figure 3.15a). This effect was not reproduced by H-TEQ.4.0 In fact, the GAFF2 van der Waals energy, when used with H-TEQ 4.0, seemed to underpredict this barrier. It should be noted that when GAFF2 was used independently for **1**, it introduced an additional V_1 term to offset this error, thereby likely "patching" the torsional energy profile. Similarly, molecule **2** also had a pair of sterically clashing hydrogen atoms, which lead to small deviations in its H-TEQ 4.0 profile from that of the QM at dihedral angles of approximately ±180˚. Its van der Waals energy was also underpredicted by 2.5 kcal·mol^{-1}. In contrast, the van der Waals energy of molecule **3** was overpredicted by 2 kcal·mol^{-1} at approximately ±140˚ dihedral angles when its hydrogen atoms were in close proximity to each other. For biaryls without a pair of sterically clashing hydrogens, such as 6-(3-methyl-1H-pyrazol-1-yl)-1,3,5-triazine-2,4-diamine (**4**) and **5**, H-TEQ 4.0 was able to predict the QM torsional profile with still greater accuracy, with an RMSE of 0.37 and 0.41 kcal·mol^{-1}, respectively (Figure 3.15d and e). For reference, GAFF2 predicted an RMSE of 1.75 and 14.17 kcal·mol^{-1}, respectively. The

problem of GAFF2 van der Waals seemed to extend indiscriminately to all molecules (ie. **1**, **2**, **3**, **6**, **7**, **8**, and **9**) possessing sterically clashing atoms in the ortho position with respect to the single bond connecting the two rings (Figure 3.15). Consequently, to further improve the accuracy of FFs, there is a need for more accurate models to describe van der Waal energies in the context of FFs.

The success of H-TEQ 4.0 on molecule **4**, **5,** and 7-(3,4-dimethylphenyl)-2-methyl-7H-pyrazolo[4,3-e][1,2,4]triazolo[1,5-c]pyrimidin-4-ium-3-ide (**9**) validated our hypothesis that both inductive and conjugated ligands had little or no effect on the torsion energy barrier of biaryls (Figure 3.15d, e, and i). In fact, **4** was meta-substituted by two conjugated -NH_2 and one inductive –CH_3 group. Similarly, **5** contained a bicyclic molecule and a carbonyl group. Lastly, **9** contained three methyl groups and a tricyclic aromatic moiety. In all three cases, H-TEQ 4.0 correctly computed the biaryl torsional barriers by only considering atoms within the primary aromatic rings. In fact, the RMSE of **9** was 0.29 kcal·mol^{-1} as predicted by H-TEQ 4.0, compared to 1.44 kcal·mol^{-1} for GAFF2.

The torsional energy profile of 3-(*m*-tolyl)furan (**7**) had an RMSE of 0.89 kcal·mol^{-1} as predicted by H-TEQ 4.0. The barrier to rotation at a dihedral angle of ±90˚ was overpredicted by approximately 0.8 kcal·mol^{-1}. This might be due to the fact that the developed method was trained on a similar molecule, 2-phenylfuran. Neglecting the presence of the methyl group on the former, the two are structural isomers. In spite of this fact, the torsional barrier was slightly different, suggesting that V_2 may be dependent on the position of the heteroatom within the aromatic ring. The reason for this is unclear, although it could be due to differences in the electron density of the biaryl system. Another reason for this might be due to the differences in the orbital energy of the σ- and σ*-orbitals adjacent to the single bond connecting the two aromatic rings. More specifically, 2-phenylfuran possesses a C-O σ*-antibonding orbital, which is a better hyperconjugation electron acceptor than the C-C σ*-orbital found in 3-phenylfuran. This phenomenon could be explored more conclusively in the future by using techniques such as energy decomposition analysis.[55, 56]
The scaffold present in the molecule, 6-isopropyl-3-phenyl-1,2,3,5-tetrazin-4(3H)-one (**8**), along with many other molecules in the validation set, was not part of the training set. In spite of this, H-TEQ 4.0 successfully computed the biaryl torsional barrier, V_2, using electronegativity of atoms comprising the ring. The RMSE for the torsion profile computed by H-TEQ 4.0 and GAFF2 was 0.67 and 2.53 kcal·mol^{-1}, respectively.

In order to also compare the accuracy of H-TEQ 4.0 to GAFF2 for those which were parametrized for the latter, a subset of biaryls below an RMSE of 5 kcal·mol^{-1} was extracted (with respect to GAFF2). Interestingly, despite only selecting for those which performed relatively well for GAFF2, the accuracy of both methods increased. In this subset, a RMSE of 0.79 kcal·mol^{-1} was observed compared to 1.14 kcal·mol^{-1} for H-TEQ 4.0 and GAFF2, respectively. Even for well parametrized biaryl molecules, H-TEQ 4.0 was better able to predict their torsional profiles. This further corroborates the idea that other energy terms in the FF, especially van der Waals and electrostatics, could be the source of error in certain molecules. Indeed, there has been ongoing interest in the FF community to improve the accuracy of these non-bonded interactions.[57]

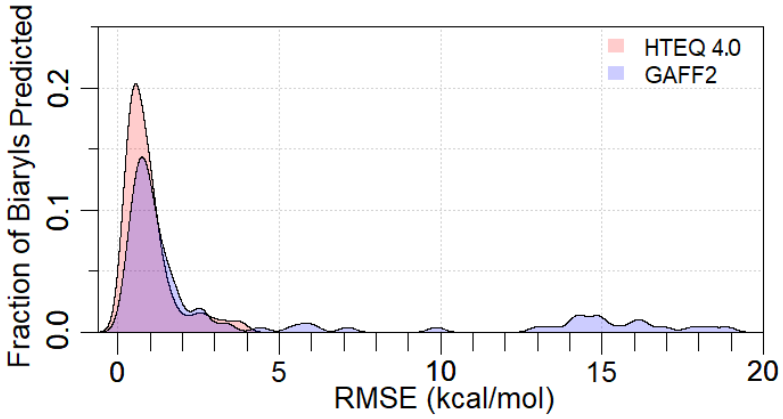


Figure 3.14. A smoothed histogram indicates the number of MM energy profile predictions made by H-TEQ 4.0 (pink) and GAFF2 (blue) within the validation set for 100 biaryls. The RMSEs of these predictions are shown on the x-axis. The QM energy profile was used as the reference, with values closer to 0 kcal·mol^{-1} representing more accurate predictions.

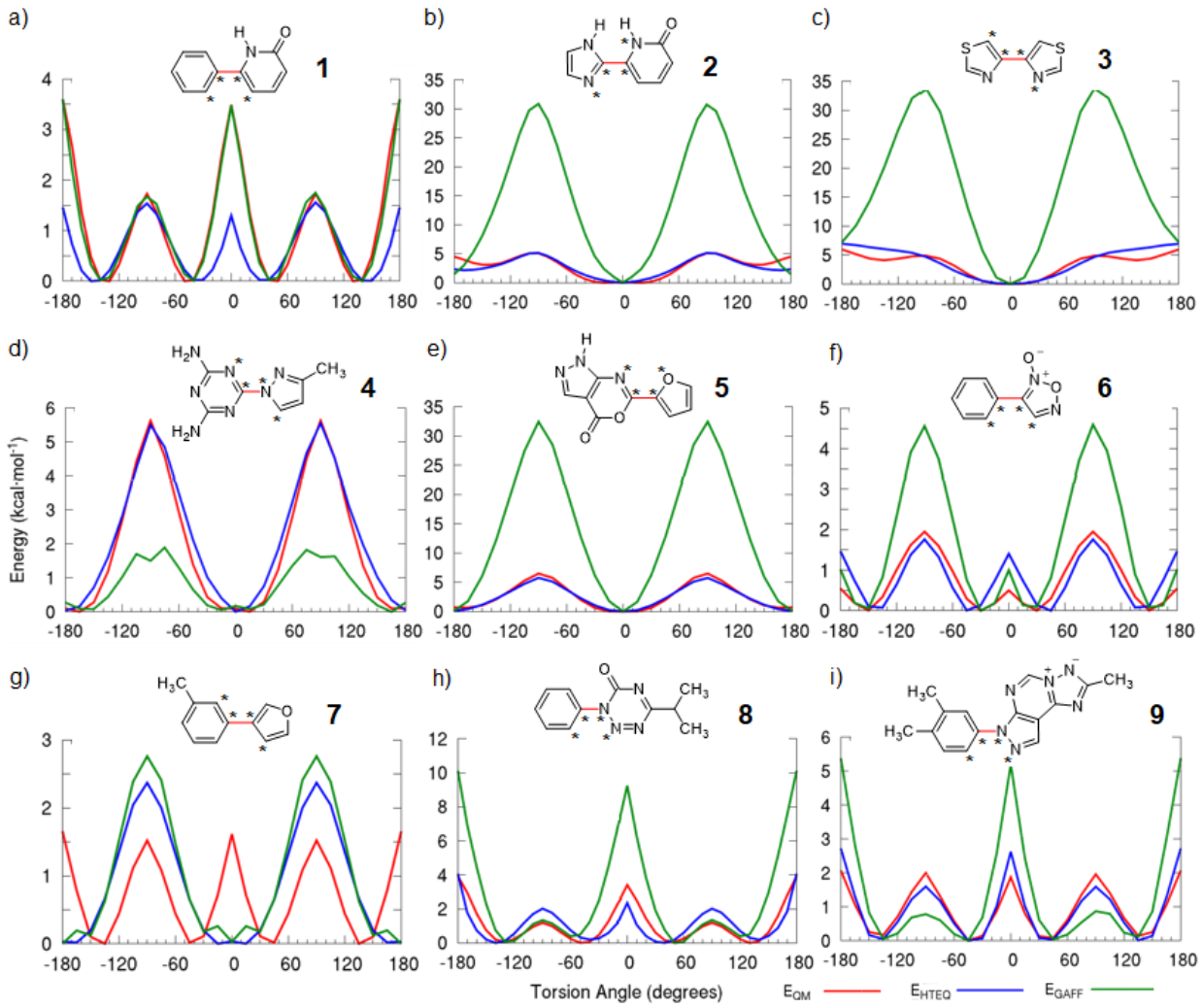


Figure 3.15. The torsion profiles of nine representative biaryl molecules (a-f). For each molecule, the torsion profiles as calculated by QM profile (red), HTEQ 4.0 (blue), and GAFF2 (green) are shown. The reference dihedral angle at 0˚ is marked by a series of four asterisks (*).

3.6 Conclusion

Biaryl systems are important pharmacophores, which are abundant in nature and existing drugs. Consequently, they are promising for the development of future pharmaceuticals. Unfortunately, existing FFs have difficulties in accurately reproducing their torsional profiles, due to the reliance on atom types. This limits its applicability towards virtual screening, using *in silico* docking. For example, despite GAFF2 performing well for several well-parameterized molecules, many were assigned inaccurate torsion parameters which deviated significantly from the QM torsional energy profile, resulting in RMSEs of over 15 kcal·mol^{-1}.

Using well-established organic chemistry-based principles, such as conjugation and hyperconjugation, as a theoretical basis, we have developed an atom-type independent method to predict the torsional energies of biaryls. In fact, by understanding the underlying chemistry, we rationalized that the torsional energy of biaryl molecules were composed of electrostatic, van der Waals, and various hyperconjugation interactions. More specifically, the latter could be decomposed into: $\sigma \rightarrow \sigma^*$, $\pi \rightarrow \sigma^*$, $\sigma \rightarrow \pi^*$ hyperconjugation, and $\pi \rightarrow \pi^*$ conjugation. Through a development set of 131 biaryl molecules, it was found that the strength of torsional barrier was directly proportional to the total electron-richness of the aromatic system, $\chi_{\pi tot}$. In addition, it was also found to be related to the difference in electron-richness between the two aromatic rings.
When the developed method, H-TEQ 4.0, was applied to a validation set of 100 biaryl systems, it outperformed GAFF2 in two crucial aspects. Firstly, GAFF2 suffered from transferability problems arising from atom type incompatibility issues. In fact, GAFF2 atom types were missing for several biaryl scaffolds. On the other hand, the H-TEQ 4.0, being a predictive method, more accurately computed the torsion parameters. H-TEQ 4.0 was able to solve the transferability issue associated with biaryl torsion barriers. Secondly, even for well-parametrized molecules, H-TEQ 4.0 achieved a higher accuracy than GAFF2. This proof-of-principle validation suggested that atom type-independent FFs could potentially solve the issue with transferability of atom types and improve the overall accuracy.

Despite training on GAFF2, this method could be easily applied to other FFs by following the same protocol highlighted in this manuscript. This could be done by taking the QM energy of the molecule at each dihedral rotation and subtracting the various other MM terms (i.e. bond, angle, and non-bonded terms) of the particular FF. Refitting using the same descriptors may be necessary, leading to slightly different parameters depending on the FF used.

In the future, the chemical rationale for the strength of the torsional barrier could be studied in more detail, perhaps using high level calculations or experimental methods. In addition, research should also focus on the development of more accurate non-bonded terms in FFs, such as electrostatics and van der Waals as these were shown to be poor in the current study. Consequently, there are still many ongoing developments in FFs.[57-61] Finally, in the future, H-TEQ 4.0 should be further extended to be applicable to all dihedral angles.

3.7 References

1. Bringmann, G.; Walter, R.; Weirich, R., The Directed Synthesis of Biaryl Compounds: Modern Concepts and Strategies. *Angew. Chem. Int. Ed.* **1990**, *29*, 977-991.

2. Cayla, N. S.; Dagne, B. A.; Wu, Y.; Lu, Y.; Rodriguez, L.; Davies, D. L.; Gross, E. R.; Heifets, B. D.; Davies, M. F.; MacIver, M. B.; Bertaccini, E. J., A Newly Developed Anesthetic Based on a Unique Chemical Core. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, 201822076.

3. Gill, M. Aromatic Compounds. In *The Chemistry of Natural Products*, Thomson, R. H., Ed.; Springer Netherlands: Dordrecht, 1993, pp 60-105.

4. Sanfeliciano, S. M. G.; Schaus, J. M., Rapid Assessment of Conformational Preferences in Biaryl and Aryl Carbonyl Fragments. *Plos One* **2018**, *13*.

5. Dahlgren, M. K.; Schyman, P.; Tirado-Rives, J.; Jorgensen, W. L., Characterization of Biaryl Torsional Energetics and Its Treatment in Opls All-Atom Force Fields. *J. Chem. Inf. Model* **2013**, *53*, 1191-1199.

6. 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.

7. Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R., Towards the Development of Universal, Fast and Highly Accurate Docking/Scoring Methods: A Long Way to Go. *Br. J. Pharmacol.* **2008**, *153*, S7-S26.

8. Mariaule, G.; De Cesco, S.; Airaghi, F.; Kurian, J.; Schiavini, P.; Rocheleau, S.; Huskic, I.; Auclair, K.; Mittermaier, A.; Moitessier, N., 3-Oxo-Hexahydro-1h-Isoindole-4-Carboxylic Acid as a Drug Chiral Bicyclic Scaffold: Structure-Based Design and Preparation of Conformationally Constrained Covalent and Noncovalent Prolyl Oligopeptidase Inhibitors. *J. Med. Chem.* **2016**, *59*, 4221-4234.

9. London, N.; Miller, R. M.; Krishnan, S.; Uchida, K.; Irwin, J. J.; Eidam, O.; Gibold, L.; Cimermancic, P.; Bonnet, R.; Shoichet, B. K.; Taunton, J., Covalent Docking of Large Libraries for the Discovery of Chemical Probes. *Nat. Chem. Biol* **2014**, *10*, 1066-+.

10. Bensinger, D.; Stubba, D.; Cremer, A.; Kohl, V.; Wassmer, T.; Stuckert, J.; Engemann, V.; Stegmaier, K.; Schmitz, K.; Schmidt, B., Virtual Screening Identifies Irreversible Fms-Like Tyrosine Kinase 3 Inhibitors with Activity toward Resistance-Conferring Mutations. *J. Med. Chem.* **2019**, *62*, 2428-2446.

11. Wei, W.; Gauld, J. W.; Monard, G., Pretransfer Editing in Threonyl-Trna Synthetase: Roles of Differential Solvent Accessibility and Intermediate Stabilization. *ACS Catal.* **2017**, *7*, 3102-3112.

12. Liu, K.; Kokubo, H., Exploring the Stability of Ligand Binding Modes to Proteins by Molecular Dynamics Simulations: A Cross-Docking Study. *J. Chem. Inf. Model* **2017**, *57*, 2514-2522.

13. Reddy, S. Y.; Kuppa, V. K., Molecular Dynamics Simulations of Organic Photovoltaic Materials: Structure and Dynamics of Oligothiophene. *J. Phys. Chem. C* **2012**, *116*, 14873-14882.

14. Leach, A. R. 4.05 - Ligand-Based Approaches: Core Molecular Modeling. In *Comprehensive Medicinal Chemistry Ii*, Taylor, J. B.; Triggle, D. J., Eds.; Elsevier: Oxford, 2007, pp 87-118.

15. 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.

16. 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.

17. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L., Evaluation and Reparametrization of the Opls-Aa Force Field for Proteins Via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **2001**, *105*, 6474-6487.

18. Stern, C.; G A Smith, D.; Chodera, J., *Is the Force with Us? Generating Chemically Relevant Data for Model Fitting*. 2019.

19. Wiberg, K. B., Application of Pople-Santry-Segal Cndo Method to Cyclopropylcarbinyl and Cyclobutyl Cation and to Bicyclobutane. *Tetrahedron Lett.* **1968**, *24*, 1083-+.

20. Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M., Charmm: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545-1614.

21. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26*, 1668-1688.

22. 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.

23. 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 (Vol 117, Pg 5179, 1995). *J. Am. Chem. Soc.* **1996**, *118*, 2309-2309.

24. 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.

25. 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.

26. Liu, Z.; 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.

27. Mobley, D. L.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Slochower, D. R.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K., Escaping

Atom Types in Force Fields Using Direct Chemical Perception. *J. Chem. Theory Comput.* **2018**, *14*, 6076-6092.

28. Schmidt, J. R.; Yu, K.; McDaniel, J. G., Transferable Next-Generation Force Fields from Simple Liquids to Complex Materials. *Acc. Chem. Res.* **2015**, *48*, 548-556.

29. Grimme, S., A General Quantum Mechanically Derived Force Field (Qmdff) for Molecules and Condensed Phase Simulations. *J. Chem. Theory Comput.* **2014**, *10*, 4497-4514.

30. Horton, J. T.; Allen, A. E. A.; Dodda, L. S.; Cole, D. J., Qubekit: Automating the Derivation of Force Field Parameters from Quantum Mechanics. *J. Chem. Inf. Model* **2019**, *59*, 1366-1381.

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

32. 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.

33. 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.

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

35. Champion, C. B., S; Wei, W; Liu Z; Labute P; Moitessier, N, H-Teq3.0: Modeling the Conformational Energy of Benzylic, Allylic, and Other Bonds Adjacent to Conjugated System [Manuscript in Preparation]. *J. Chem. Inf. Model*.

36. Goodman, L.; Gu, H. B.; Pophristic, V., Gauche Effect in 1,2-Difluoroethane. Hyperconjugation, Bent Bonds, Steric Repulsion. *J. Phys. Chem. A* **2005**, *109*, 1223-1229.

37. Fleming, I. Molecular Orbital Theory. In *Molecular Orbitals and Organic Chemical Reactions*; 2010, pp 1-67.

38. Mayr, H.; Breugst, M.; Ofial, A. R., Farewell to the Hsab Treatment of Ambident Reactivity. *Angew. Chem. Int. Ed.* **2011**, *50*, 6470-6505.

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

40. Mo, Y. R.; Gao, J. L., Theoretical Analysis of the Rotational Barrier of Ethane. *Acc. Chem. Res.* **2007**, *40*, 113-119.

41. Houk, K. N., Frontier Molecular Orbital Theory of Cycloaddition Reactions. *Acc. Chem. Res.* **1975**, *8*, 361-369.

42. Fernandez, I.; Frenking, G., Hyperconjugative Stabilization in Alkyl Carbocations: Direct Estimate of the Beta-Effect of Group-14 Elements. *J. Phys. Chem. A* **2007**, *111*, 8028-8035.

43. O'Hagan, D., Understanding Organofluorine Chemistry. An Introduction to the C-F Bond. *Chem. Soc. Rev.* **2008**, *37*, 308-319.

44. Smith, J. G.; Ho, I., Effect of Electronegative Substituents on the Reductive Dimerization of Schiff Bases. Formation of Vicinal Dianions. *J. Org. Chem.* **1973**, *38*, 2776-2779.

45. Mo, Y. R.; Jiao, H. J.; Schleyer, P. V., Hyperconjugation Effect in Substituted Methyl Boranes: An Orbital Deletion Procedure Analysis. *J. Org. Chem.* **2004**, *69*, 3493-3499.

46. Cappel, D.; Tullmann, S.; Krapp, A.; Frenking, G., Direct Estimate of the Conjugative and Hyperconjugative Stabilization in Diynes, Dienes, and Related Compounds. *Angew. Chem. Int. Ed.* **2005**, *44*, 3617-3620.

47. 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.

48. 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.

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

50. 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.

51. Katritzky, A. R. 2.3 - Structure of Five-Membered Rings with One Heteroatom**Based on Chapter 3.01 of 'Comprehensive Heterocyclic Chemistry', by C. W. Bird and G. W. H. Cheeseman, Queen Elizabeth College, University of London. In *Handbook of Heterocyclic Chemistry*, Katritzky, A. R., Ed.; Pergamon: Amsterdam, 1985, pp 53-86.

52. Katritzky, A. R. 2.4 - Structure of Five-Membered Rings with Two or More Heteroatoms**Chapter 4.01 of 'Comprehensive Heterocyclic Chemistry', by A. R. Katritzky, University of Florida, and J. M. Lagowski, the University of Texas at Austin. In *Handbook of Heterocyclic Chemistry*, Katritzky, A. R., Ed.; Pergamon: Amsterdam, 1985, pp 87-123.

53. Katritzky, A. R. 2.2 - Structure of Six-Membered Rings**Adapted from Chapter 2.01 of 'Comprehensive Heterocyclic Chemistry', by A. Mckillop and A. J. Boulton, University of East Anglia. In *Handbook of Heterocyclic Chemistry*, Katritzky, A. R., Ed.; Pergamon: Amsterdam, 1985, pp 23-51.

54. Dey, S. K.; Lightner, D. A., 1,1 '-Bipyrroles: Synthesis and Stereochemistry. *J. Org. Chem.* **2007**, *72*, 9395-9397.

55. Zhao, L. L.; von Hopffgarten, M.; Andrada, D. M.; Frenking, G., Energy Decomposition Analysis. *Wires. Comput. Mol. Sci* **2018**, *8*.

56. Mo, Y.; Bao, P.; Gao, J., Energy Decomposition Analysis Based on a Block-Localized Wavefunction and Multistate Density Functional Theory. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6760-6775.

57. 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.

58. Rusu, V. H.; Bachmann, S.; van Gunsteren, W. F., Gromos Polarisable Model for Acetone. *Mol. Phys.* **2016**, *114*, 845-854.

59. Qi, R.; Wang, L. P.; Wang, Q. T.; Pande, V. S.; Ren, P. Y., United Polarizable Multipole Water Model for Molecular Mechanics Simulation. *J. Chem. Phys.* **2015**, *143*.

60. Wang, J. M.; Cieplak, P.; Li, J.; Cai, Q.; Hsieh, M. J.; Luo, R.; Duan, Y., Development of Polarizable Models for Molecular Mechanical Calculations. 4. Van Der Waals Parametrization. *J. Phys. Chem. B* **2012**, *116*, 7088-7101.

61. Wei, W.; Luo, J.; Waldispühl, J.; Moitessier, N., Predicting Positions of Bridging Water Molecules in Nucleic Acid–Ligand Complexes. *J. Chem. Inf. Model* **2019**, *59*, 2941-2951.

4 Use of Extended-Hückel Descriptors for Rapid and Accurate Predictions of Conjugated Torsional Energy Barriers

4.1 Preface

In **Chapter 3**, a method to derive force field parameters of biaryl molecules was devised. Although pertinent to drug discovery and modelling of these organic molecules, it was strictly restricted to biaryls. To expand the utility and scope of this method to all conjugated molecules, other classes of molecules were explored. Again, organic chemistry principles and quantum chemistry methods were used to develop a method capable of producing parameters on-the-fly. To our delight, a simple and general method to predict parameters for all conjugated molecules was found and presented below.

This chapter is based on work from: Wei, W.; Champion, C.; Barigye, S. J.; Liu, Z.; Labute, P.; Moitessier, N., Use of Extended-Hückel Descriptors for Rapid and Accurate Predictions of Conjugated Torsional Energy Barriers. J. Chem. Inf. Model 2020. ASAP.

Author contributions: I performed the calculations, analysis, and write-up of the manuscript. Zhaomin Liu, Stephen Barigye, Candide Champion, Paul Labute, Prof. Nicolas Moitessier, and I designed the described methods.

4.2 Introduction

Conjugated Molecules in Pharmaceuticals. Conjugated chemical moieties are found in many druglike molecules and in biological systems.[1-3] Many well-known drugs contain these fragments, including aspirin, celecoxib, pyrantel, and midazolam, which are pharmaceuticals listed in the World Health Organization (WHO) Model Lists of Essential Medicines (Figure 4.1).[4] In fact, 90% of all currently approved small molecule drugs contain conjugated moieties.[5] Out of these molecules, 46% contain at least one single bond bridging conjugated moieties. In biological systems, conjugated moieties are found in, but not limited to, amino acids (eg. peptide bonds, glutamic acid, and tyrosine); nucleic acids (eg. adenine and cytosine), lipids (eg. phosphoinositol and triglycerides) and various essential cofactors (eg. nicotinamide adenine dinucleotide and pyridoxal phosphate).

Chemically, the p-orbitals of the conjugated fragments form π-bonds, which are delocalized over a part of the molecule. This causes that portion to adopt a more planar and rigid geometry than saturated moieties. In many cases, conjugated molecules are involved in unique chemical interactions, including the ability to participate in π- π stacking, π-cation, and π-anionic interactions. [6-8]

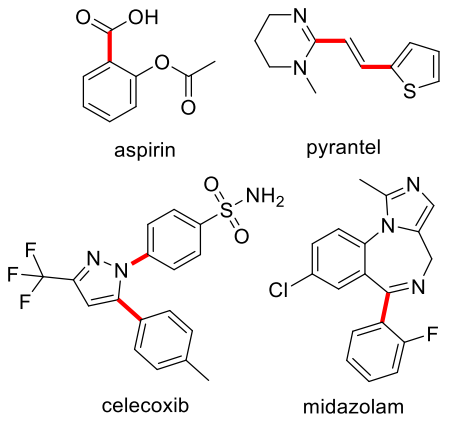


Figure 4.1. Commercially available drug molecules, which possess extensively conjugated moieties. The torsional bond of interest is highlighted in red.

Conjugated Molecules and Molecular Mechanics. Despite the importance of conjugated functional groups in pharmaceuticals, existing molecular mechanics (MM) methods are not well-adapted to describe their conformational energy landscapes with the exception of a limited number of well-parametrized cases (e.g., butadiene and biphenyl).[9-11] More specifically, torsional energy

barriers of conjugated molecules are inaccurate for most small molecules in current FFs. This is unfortunate because virtual high-throughput screening (vHTS) is dependent upon an accurate MM FFs for binding affinity calculations between the target and the drug candidate.[12] In the recent past, vHTS has allowed millions of drug molecules to be screened quickly and inexpensively to produce enriched libraries of compounds. The reliance of vHTS on poor FFs led to errors in affinity calculations, hence increased the number of false positives and negatives. An improvement in the underlying FF would, therefore, enhance the accuracy and enrichment capabilities of vHTS for subsequent biological testing, accelerating the initial drug discovery process.

In addition, molecular dynamics (MD) simulations, employing free energy perturbation (FEP) techniques, are also often used during this initial drug discovery phase.[13, 14] Despite the first use of FEP techniques several decades ago, it has only recently emerged as a computationally tractable method in drug discovery efforts.[15] FEP evaluates the relative binding free energies of analogous ligands using alchemical transformations, which performs mutations to the ligands through a step-wise process. Through ligand annihilation, the absolute free energy could also be calculated, by mapping all free energy values to the unbound state of the receptor.[16] During FEP calculations, non-equilibrium states are occasionally reached, and inaccurate torsional parameters could produce erroneous binding free energies. Consequently, an improved FF could enhance the accuracies of free energy predictions and have increased utility for future drug discovery efforts.

Torsional Energy Profiles of Conjugated Molecules. In conjugated and other molecules, torsional energy barriers, along with nonbonded interactions, are among the most difficult to predict with current FFs.[9, 11] Torsional barriers vary greatly depending on their identity. For example, while the barrier to rotation of the torsional bond associated with 1H-pyrrole-2-carbaldehyde was approximately 11 kcal·mol^{-1}, pyrimidine-2-carbaldehyde had a barrier of 7 kcal·mol^{-1}. Upon introduction of an -O^{-} group a ring position on the former, yielding 5-formyl-1H-pyrrol-3-olate, this barrier increases to 18 kcal·mol^{-1} (this work). The chemical space has previously been estimated to be as diverse as 10^{60} distinct molecules.[17, 18] Due to its size, FFs were never designed to cover all torsion parameters in their entirety, which would be an impossible task. Rather, FFs aimed to parametrize representative torsions, whose parameters, it was hoped, could be transferred to chemically similar but unparametrized molecules. Unfortunately, however, this strategy was found to be flawed due to the unpredictability of torsional energy barriers and the complexity of the underlying chemical interactions. As a result, accurate torsional energy barriers

have traditionally been obtained computationally using QM torsion scans, due to a scarcity of available experimental data.[19] This effectively meant that for each torsion in a molecule, approximately two dozen QM calculations had to be performed. This approach was prohibitively expensive and could not be used to cover the entire chemical space. Consequently, a more predictive and adaptable method, based on understanding the chemical interactions governing these torsions, was necessary.

To this end, we recently developed H-TEQ 4.0 (Hyperconjugation for Torsional Energy Quantification 4.0),[9] which could accurately predict the torsional energy barriers of biaryl molecules (e.g. biphenyl) using a molecular metric: the electron richness of the aromatic rings. This method was an effective solution to predict biaryl torsional energy profiles. However, its use is restricted to this class of molecules. Consequently, we sought to expand and broaden the coverage of this method. A general method to predict the torsional energy barrier of all conjugated molecules (e.g. butadiene, benzaldehyde, and biphenyl) with a high degree of accuracy was envisioned.

MM FFs are empirical equations, and corresponding sets of parameters, based on classical mechanics to reproduce the QM energies and thermodynamic properties of interest. The total energy in MM could be decomposed into its constituent potential energy functions (equation 4.1).[20-24] The van der Waals interactions and steric repulsion terms are usually approximated by the Lennard-Jones 12-6 potential, while electrostatics are described using Coulomb's law based on atomic partial charges. Previously, hydrogen bonding had been modelled in Amber and other FFs using a modified Lennard-Jones 12-10 potential. Together, these interactions make up the nonbonded interactions. The bonded terms, on the other hand, include bond and angle stretching, a torsion term and an out-of-plane energy term. The second and third terms are usually modelled by a harmonic spring function while the latter is described by the truncated Fourier series (equation 4.2).

In the context of conjugated molecules, only the first two terms of equation 4.2 have been cited to be chemically meaningful.[9] The first term in this series (n=1), describes the conformational preferences of the conjugated molecule (i.e. *cis*/*trans*) in the absence of nonbonded interactions. The second term in the Fourier series (n=2), gives an idea of the overall barrier height and strength of conjugation. In this report, a negative V_1 term signifies a preference of the molecule towards a

cis geometry (i.e. preference towards 0˚ over ±180˚ geometry) while a negative V_2 term signifies its preference for the 0˚ and ±180˚ over the ±90˚ geometry (*vice-versa* for positive V_1 and V_2 terms).

Equation 4.1. Potential energy function of standard MM FFs.

$$E_{MM} = \underbrace{E_{bond} + E_{ang} + E_{tor}}_{bonded} + \underbrace{E_{ele} + E_{vdw} + E_{H-bond}}_{non-bonded}$$

Equation 4.2. Potential energy function of standard torsions in MM FFs.

$$E_{tor} = \sum_{i=1}^{4} \frac{V_n}{2}[1 + \cos n(\theta - \gamma)]$$

Atom Type Independent MM FFs. Central to traditional MM FFs, is the concept of atom types. An atom type is a specific element associated with a set of parameters in a molecule in the context of its connectivity, hybridization, and chemical environment. Chemically "similar" atoms in a molecule are grouped together and given the same set of parameters.[23] For example, a given FF may parametrize the C=C-C=C torsion of a butadiene. Upon encountering a different molecule, such as 1,1-difluorohexa-1,3,5-triene, the torsional parameter of butadiene may be used, despite their possessing unidentical chemical environments (Figure 4.2). As mentioned earlier, atom typing has conventionally been necessary due to the large size of the chemical space. However, this has resulted in a loss of accuracy since molecules grouped together may intrinsically possess very different conformational preferences. In addition, this grouping of chemically similar torsional parameters is also challenging during atom typing.

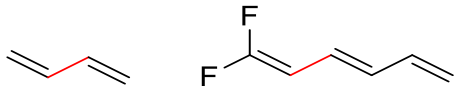


Figure 4.2. The torsional bond of interest (in red) of hexatriene (left) compared to that of 1,1-difluorohexa-1,3,5-triene (right) are shown.

To overcome this problem, research groups have utilized three main strategies. The first strategy was to attempt to cover as many functional groups and molecules as possible through large-scale parametrization and assume transferability of the developed parameters. This is seen

over the years by active atom typing by Parm@Frost FF spanning almost two decades[25]; and the OPLS FF.[10, 24, 26, 27] Despite these efforts, the limited transferability of the produced parameters limited its utility for non-parametrized molecules. The second strategy include those methods which aim to automate the process of developing torsional parameters, thereby simplifying the process and reducing the associated human labor costs. This includes methods and FFs, such as QUBEKit,[28] QMDFF,[29] GAAMP,[30] ffTK,[31] Paramfit32,[32] and Parmscan33.[33] While these methods are well suited to simulate the binding energy of a few receptor/ligand examples, these methodologies may face difficulty in the context of vHTS due to the high costs associated with performing QM torsional scans. The third methodology involves the development of MM FF without reliance on atom-types. Notably, this strategy has been pursued by Gerber et al,[34] Mobley et al, .[35, 36] and our own research group in the past.[9, 18, 37, 38] It has the advantage of being generally applicable to all molecules within a predefined scope; however, the difficulty of this approach lies in being able to identify the underlying chemical principles and their subsequent quantification. It should be noted that machine-learning algorithms are promising tools for torsional parameter prediction.[39] However, to the best of our knowledge, its accuracy has not yet been proven, perhaps due to the requirement for a large training set of torsional parameters which are computationally expensive to obtain.

Our efforts toward developing an atom-type free MM FF led us to first focus on the torsions of saturated molecules. During this project, it was found that the $\sigma \rightarrow \sigma^*$ hyperconjugation was the predominant interaction affecting their torsions. A chemistry knowledge-based method was developed to estimate the strengths of these interactions (H-TEQ 1.4).[38] After this initial success, we focused on torsions in saturated molecules involving lone pair electrons on the central atoms, thereby quantifying $n \rightarrow \sigma^*$ hyperconjugation (H-TEQ 2).[18] Next, a method to predict $\pi \rightarrow \sigma^*$ and $\sigma \rightarrow \pi^*$ hyperconjugation was developed to predict the torsional interaction of conjugated bonds with adjacent single bonds (H-TEQ 3.0).[37] Lastly, as mentioned previously, we began looking at the torsional energy barriers of conjugated molecules by first focusing on biaryls thereby quantifying $\pi \rightarrow \pi^*$ conjugation for a subset of molecules (H-TEQ 4.0).[9]

Hyperconjugation as a Main Factor Affecting Torsion Potential. Hyperconjugation has been used by organic chemists for decades to explain the conformational preferences and relative stabilities of organic molecules, hypothesizing the existence of interactions such as gauche and anomeric effects.[40,41] For example, 1,2-difluoroethane prefers the *gauche* conformation over that

of the *anti* due to the greater relative strength of $\sigma_{C\text{-}H} \rightarrow \sigma^*_{C\text{-}F}$ hyperconjugation compared to that of $\sigma_{C\text{-}H} \rightarrow \sigma^*_{C\text{-}H}$ (Figure 4.3). This facilitates the two fluorine atoms to be gauche to each other despite their electrostatic repulsion. Similarly, anomeric effect could be observed in D-glycopyranose, in which the α- is favored over the β-conformation (Figure 4.4). This could be rationalized by the former allowing a stronger $n \rightarrow \sigma_{C\text{-}OH}^*$ hyperconjugation to occur, whereas the latter is positioned to facilitate the weaker $n \rightarrow \sigma_{C\text{-}H}^*$ hyperconjugation. In general, more electronegative atoms are better hyperconjugation acceptors due to a net lowering in orbital energies. In fact, match in orbital energies between the donor and acceptor orbitals is one of the two main factors affecting the strength of hyperconjugation.[42] The other factor, spatial overlap between the donor and acceptor orbitals, imposes a physical restriction for interaction (Figure 4.5). Overall, frontier molecular orbital theory dictates that hyperconjugation and conjugation interactions, together, lead to interactions which result in a hybridized occupied molecular orbital (MO) of lower energy (and also a hybridized unoccupied orbital of higher energy) from the donor and acceptor orbitals.[43]

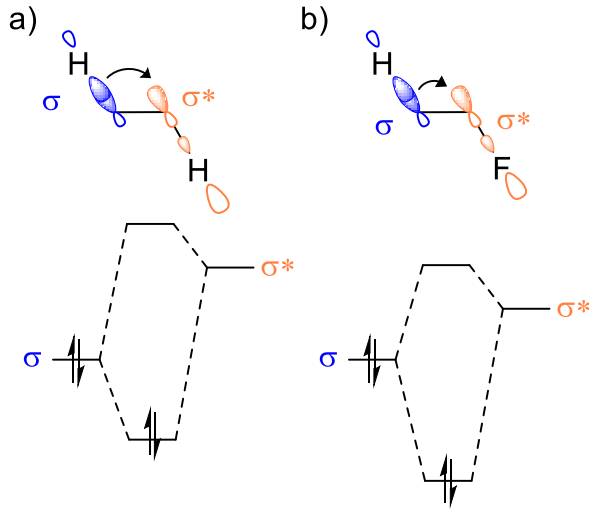


Figure 4.3. A qualitative depiction of hyperconjugation of involving a) $\sigma_{C\text{-}H} \rightarrow \sigma^*_{C\text{-}H}$ and b) $\sigma_{C\text{-}H} \rightarrow \sigma^*_{C\text{-}F}$. The more electronegative fluorine has an overall lower lying σ*-antibonding orbital than hydrogen. Consequently, b) is stabilized to a greater degree than a).

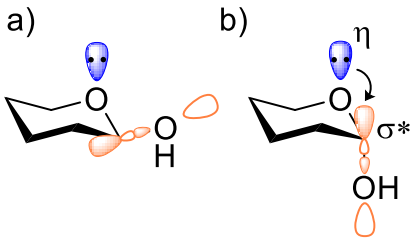


Figure 4.4. Hyperconjugation involving n → σ*C-OH interaction in the two ring conformations of D-glycopyranose. In a), the hydroxyl antibonding orbital is not properly positioned to accept the electron density donation from the lone pair. In b), the hydroxyl group is axial, leading to significant orbital overlap between the lone pair electrons on the ring oxygen and the hydroxyl oxygen antibonding orbital. This gives noticeable stabilization in energy. For simplicity, only the atoms of interest are shown.

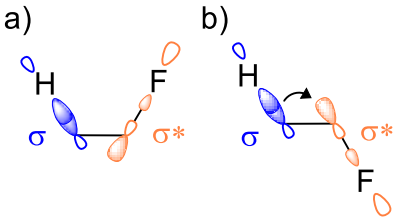


Figure 4.5. Hyperconjugation involving $\sigma_{C\text{-}H} \rightarrow \sigma^*_{C\text{-}F}$ in 1,2-diflouroethane is shown at two different torsion angles. In a), there is very little overlap between the two orbitals, leading to weaker overlap, while in b), there is significant orbital overlap, leading to noticeable stabilization in energy. For simplicity, only the atoms of interest are shown.

Conjugated molecules, unlike their saturated counterparts, are more complex since they possess π-orbitals, in addition to σ-orbitals. For example, the torsion preference in oxazole-2-carbaldehyde (Figure 4.6a-g) is governed by numerous σ → σ*, π → σ* and σ → π* hyperconjugation, and π → π* conjugation interactions. The orientations of the orbitals allow the σ → σ* and π → π* hyperconjugation interactions to combine constructively. Similarly, π → σ* and σ → π* hyperconjugation interactions are also constructive. However, these two groups combine destructively since the former favors the planar and while the latter favors the orthogonal conformation. This analysis is true for all conjugated molecules.

Due to the ability of electrons to flow freely in delocalized π-orbitals, torsions of conjugated molecules may be especially susceptible to the effects of even remote substituents.[44] It should be noted that although only two conjugation interactions are shown (Figure 4.6f and g), many are theoretically possible. For oxazole-2-carbaldehyde, a total of 5 conjugation interactions are possible (Figure 4.6h). The number and types of interactions may differ for other molecules. Generally, the dominant conjugation interactions should be composed of donor and acceptor orbitals which have the smallest energy gaps and the greatest spatial overlap. For oxazole-2-carbaldehyde, this should be the conjugation interaction depicted in Figure 4.6h.

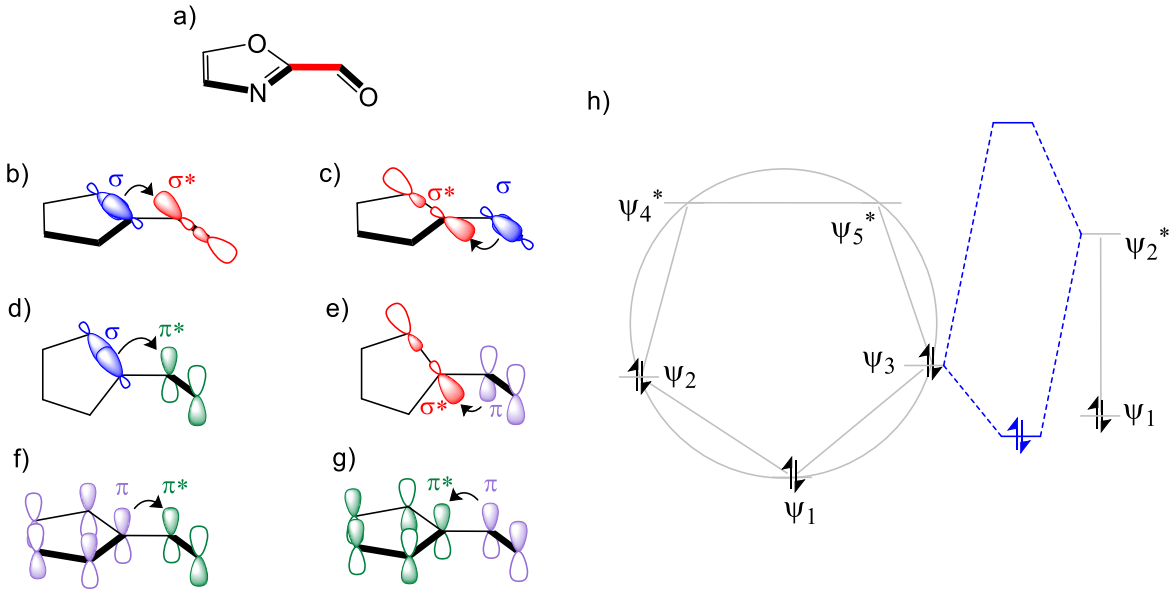


Figure 4.6. (a) Chemical structure of oxazole-2-carbaldehyde (with the torsional bond of interest highlighted in red) is shown with hyperconjugation of orbitals involving: (b and c) σ → σ*, (d) σ →π*, and (e) π → σ *, and (f and g) π → π* conjugation. For simplicity, a maximum of one σ-, σ*-, π-, and π*-molecular orbitals are shown in each subfigure. h) A qualitative molecular orbital diagram of oxazole-2-carbaldehyde is shown, along with the major conjugation interaction. In this case, the conjugation interaction involving the smallest energy overlap is depicted.

4.3 Computational Methods

Construction of a Development Set. A total of 684 conjugated molecules of various types were included in the development set. Its composition included various acyclic, aryl, and biaryl molecules (Figure 4.7) of which the latter were obtained from a prior study.[9] In order to ensure

that the developed method would be transferrable to torsions found in various conjugated systems, an wide range of molecules were considered, including strongly electron-withdrawing (EWG) and electron-donating groups (EDG). Furthermore, extensively conjugated and charged systems were also included.

For each molecule, a suitable starting conformation was obtained by performing a global optimization at the MP2/6-311+G** level of theory using GAMESS-US.[45] Subsequently, a series of constrained QM optimizations were performed whereby the dihedral angle of interest was varied sequentially between -180˚ and 180˚ at 10˚ or 15˚ intervals. Single point energies for each conformation were then evaluated using GAFF2,[46] MAB,[34] and MMFF94.[47] It should be noted that AM1-BCC charges[48] were calculated for GAFF2, using the Antechamber module, while MAB and MMFF94 used their native charging schemes. In order to obtain the $\sigma \rightarrow \sigma^*$ hyperconjugation energy, Natural Bonding Orbital (NBO) analysis was also performed at the HF/6-311+G** level of theory on the MP2 optimized structures.[49] To obtain the isolated torsional energy, the van der Waals and electrostatics components of each FF method were individually subtracted from the total energy as computed by MP2. This resulted in three distinct torsional energy profiles, each belonging to the previously mentioned FFs. For each molecule in the development set, Fourier transform was applied to these isolated torsional energies of each FF and $\sigma \rightarrow \sigma^*$ hyperconjugation energies to obtain the $V_{1\text{-}3}$ parameters (Eq. 3.2).

A variety of different molecular descriptors were chosen based on what is known to rigidify a torsion and tested to predict the strengths of torsional barriers, based on rationalizing how they would impact and rigidify them. These descriptors ranged from simple atomic properties such as atomic electronegativities to π-bond order obtained by Extended-Hückel theory (EHT). EHT properties were calculated using an implementation of this well-known method in Molecular Operating Environment (MOE).[50-52] In fact, the implementation of this Hückel method and its various parameters are identical to that of Gerber et al.[34] It has the advantage that it is coordinate-free and uses a self-consistent field theory to quickly optimize parameters, including off-diagonal elements. From EHT, the descriptor(s) which could best reproduce the torsional energy was chosen to formulate our method, which was subsequently incorporated into H-TEQ 4.5, which was itself incorporated into MOE.[53]

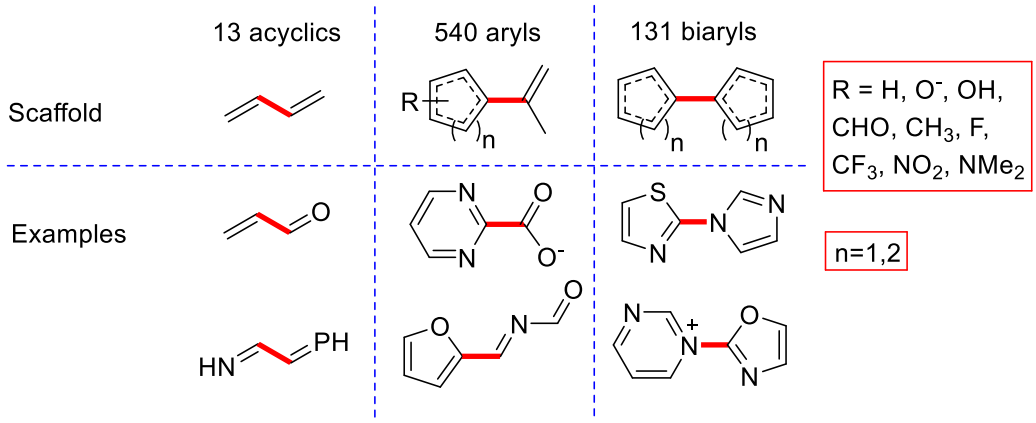


Figure 4.7. Various Molecules included in the development set, categorized into three distinct groups: acyclic, aryl, and biaryl conjugated molecules. The torsional bond of interest is highlighted in red.

Construction of the Validation Set. A total of 200 conjugated molecules were included in the validation set, including: 30 acyclic, 70 aryl, and 100 biaryl molecules, which were randomly selected. The biaryls were obtained from a previously constructed set and are not shown in Figure 4.8,[9] while the rest of the compounds were obtained directly from the MMFF94 validation set or truncated in order to limit the number of peripheral torsional bonds.[47] In order to ensure the robustness of H-TEQ 4.5, the validation set was constructed to contain a wide variety of dissimilar druglike molecules, while having no overlapping molecules with the development set. For comparative purposes, the full MM energy profile of H-TEQ 4.5, MMFF94, GAFF2, and MAB were compared to that of the QM for each molecule. The total MM energy was calculated by taking the FF energies for all terms in Equation 4.1, except for the torsion energy, which was computed by H-TEQ 4.5. This was done for GAFF2, MAB, and MMFF94. Root-mean-square error (RMSE) was calculated for H-TEQ 4.5 and each of the previously mentioned FFs, with the QM energy profile as a reference. When comparing between the QM and MM torsional energy profiles, a vertical translation was applied to each profile so that their average values were overlaid.

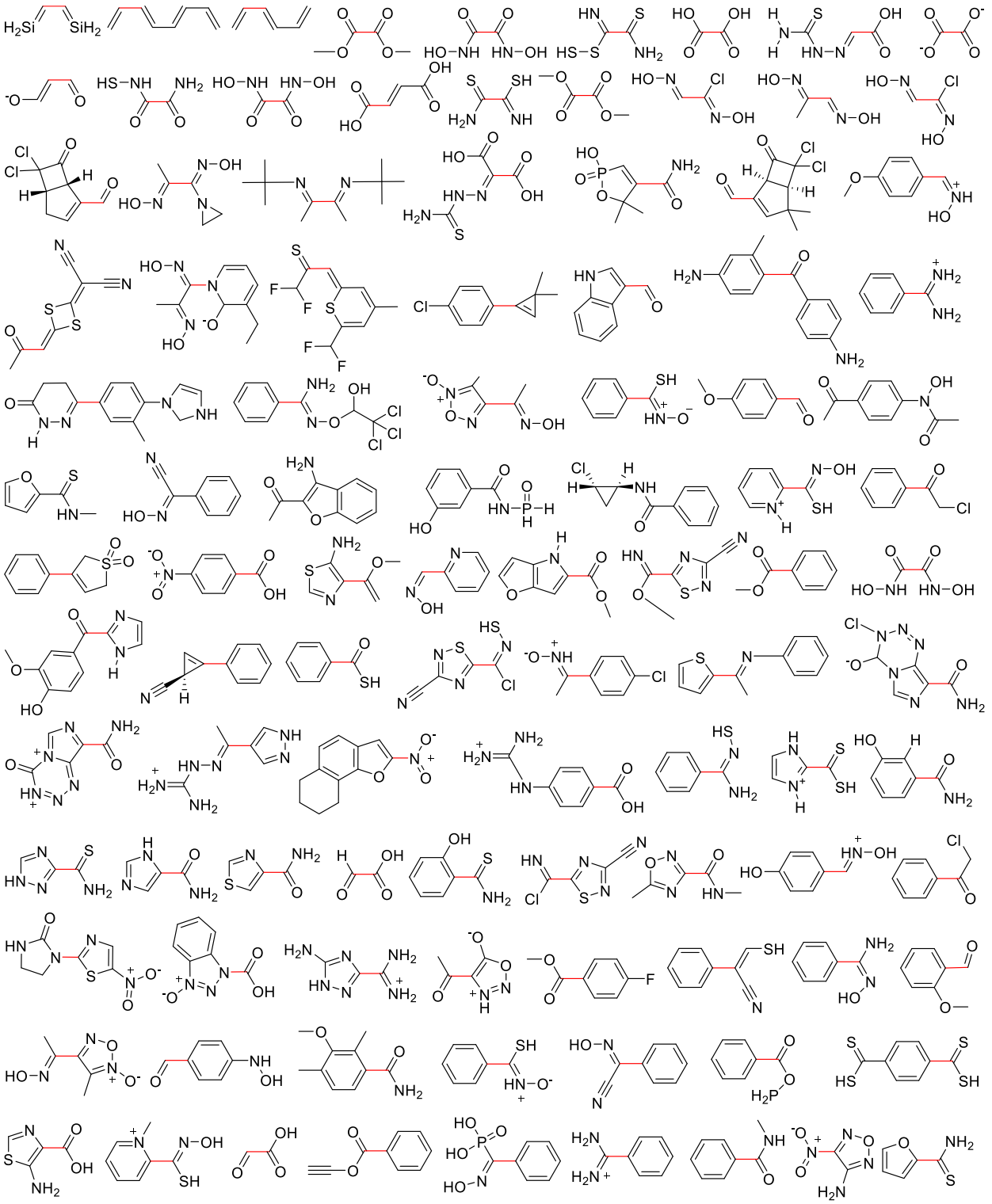


Figure 4.8. One hundred new molecules of the compiled validation set. Only molecules containing conjugated aryl and linear moieties are shown. The torsional bonds of interest are shown in red.

Subunit π-orbital Analysis. To analyze the π-orbitals in a subset of unsubstituted conjugated molecules and understand the direction of the conjugation donation, our test molecules were submitted for subunit π-orbital analysis, which involved breaking the molecule at the torsional bond of interest and capping each side with hydrogen atoms (Figure 4.9). Subsequently, both sides of the subunits were optimized at the CCSD(T)/cc-pVTZ level of theory. Thereafter, the energy levels of all π-MOs of both subunits were compared to determine the donor and acceptor MOs and the direction of the dominant conjugation interaction of the original molecule (Figure C1~28). The latter was determined by the smallest energy gap between π- and π*-MOs across the subunits. In total, 22 subunits were calculated (Section C1).

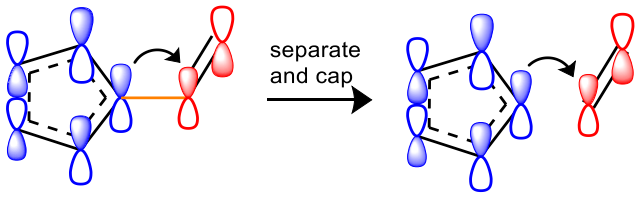


Figure 4.9. Schematic showing the subunit π-orbital analysis process. The conjugated molecule is separated at the torsional bond of interest and capped with hydrogen atoms. Subsequently, the π-MOs of both subunits are compared to determine the donor and acceptor side of the original molecule.

4.4 Results and Discussion

Qualitative Observations. Within the development set of molecules, it was found that the chemical nature of the bound subunit directly influenced V_2 for both conjugated aryls (Figure 4.10a) and acyclic molecules (Figure 4.10b). For conjugated aryls, it was found that aldehyde and thioaldehyde had lower V_2 than imine and phosphethene, which in turn were lower than vinyl and methylenesilane, when connected to the aromatic ring via their carbon atoms (Figure C29). By performing the subunit π-orbital analysis, it was found that in most cases, the ψ_3 π-MO of the aromatic ring acted as the donor, while the ψ_2 π*-MO of the double bond was the acceptor. In addition, moving from vinyl to imine to aldehyde resulted in a sequential lowering of the energy of the ψ_2 π*-MO, which is consistent with the qualitative effects of electronegativity on σ-MOs. This might explain the observed trends of conjugation strength (Figure 4.10a). Similarly, moving from methylenesilane to phosphethene to thioaldehyde also resulted in a sequential lowering of the ψ_2 π*-MO and an increase in conjugation strength. However, this trend does not hold when

these row 2 and 3 elements were combined, possibly due to differences in the spatial overlap of the π-MOs. Consequently, other factors may play a role in determining the conjugation strength. A chemical descriptor, which captured this information well, was the number of electron lone pairs on the β-atom (Figure 4.10a). Other possible factors for consideration included the amount of orbital overlap between the MOs, bond distances, and other hyperconjugation effects.

For conjugated acyclic molecules, containing C-C central bonds, the smallest V_2 seemed to occur when the difference in the number of the electron lone pairs on the β-atoms was greater (Figure 4.10b). This might result from aldehyde and thioaldehyde being good π*-acceptors while vinyl and methylenesilanes are good π-donors. It is believed that a combination of a strong donor and a strong acceptor had the greatest π-MO orbital stabilization, which is the source of torsional energy barrier. This is illustrated by the fact that acrylaldehyde (-6.52 kcal·mol^{-1}) had a smaller V_2 than both butadiene (-4.33 kcal·mol^{-1}) and oxalaldehyde (-3.88 kcal·mol^{-1}).

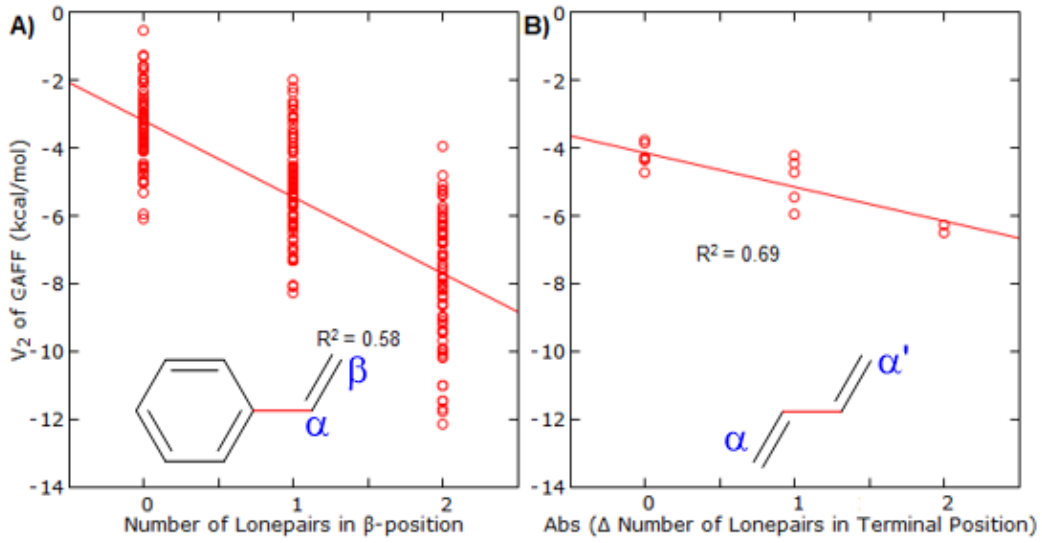


Figure 4.10. a) V_2 of 219 unsubstituted, conjugated aryls plotted against the number of electron lone pairs on the β-atom with respect to the ring. b) V_2 of 13 unsubstituted conjugated acyclic molecules, plotted against the absolute difference of electron lone pairs on the terminal position (ie. α- and α'-atom to the central atoms). A representative chemical structure is also depicted, with the labels for each atomic position of interest.

***V_2* Conjugation Strength in All Molecules are Correlated with Extended-Hückel Theory Descriptors.** A predictive and transferrable method to predict the torsional barrier of any conjugated molecules was desired. At the outset of this project, the torsion term profile and the

corresponding V_2 values were obtained from the difference between QM energy and, alternatively, of GAFF2, MMFF94 and MAB nonbonded energies. V_2 values were then plotted against EHT π-bond order (π-BO), which resulted in an R^2 of 0.56 (Figure 4.11a), 0.32, and 0.54 (for each FF respectively) across all 684 conjugated molecules in the development set. EHT is advantageous over other atomic and molecular properties as it considers the entire π-framework and could adapt to changes in the σ-substituents as well. In addition, it has given qualitatively accurate results in the past and is computationally inexpensive.[51]

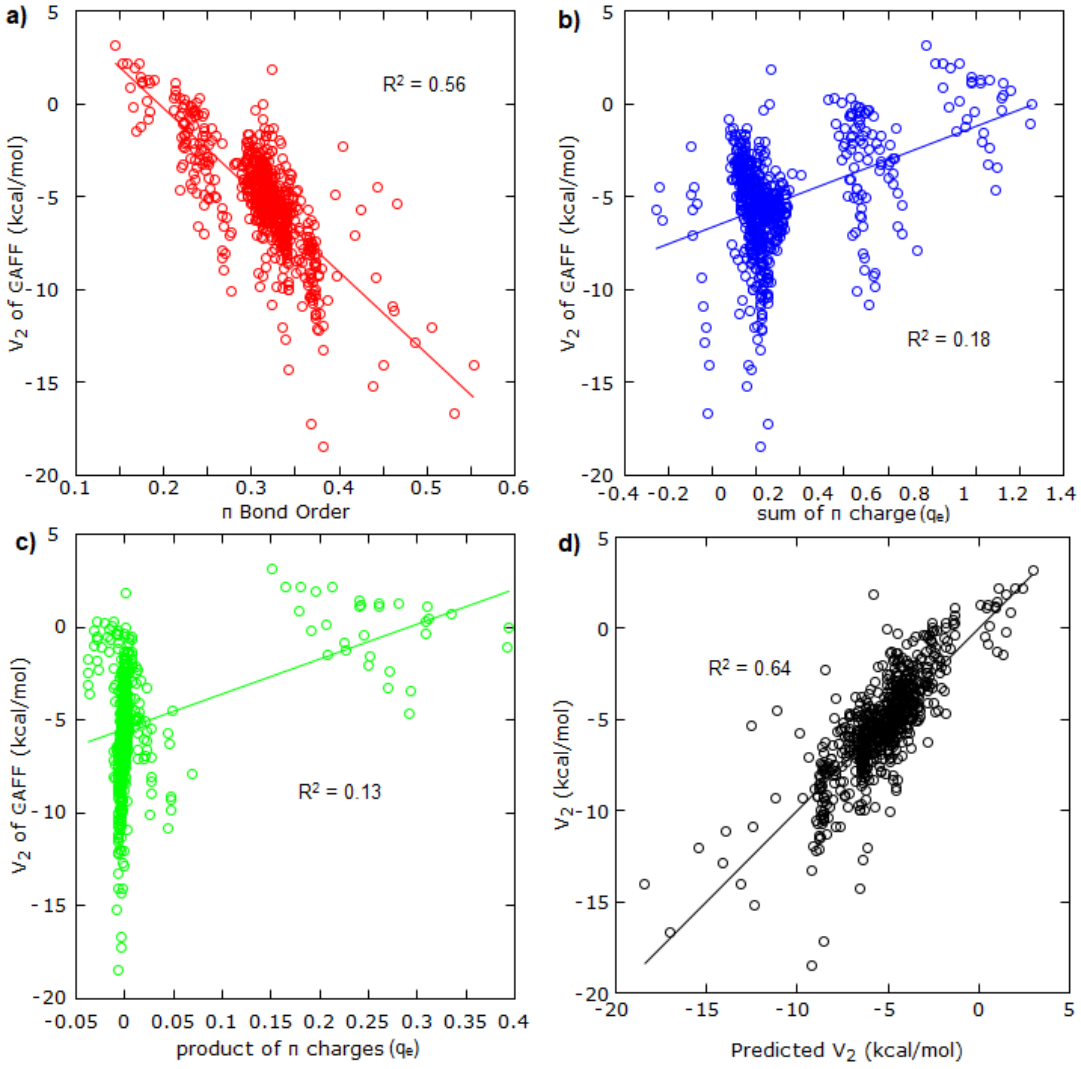


Figure 4.11. (a-c) Observed trends between various Extended Hückel Theory properties and V_2 strengths of the difference between QM and nonbonded interactions obtained from GAFF (similar trends and accuracies were obtained with MAB and MMFF94). In each subplot, V_2 is plotted against a) π-bond order, b) sum of atomic π-charges and c) product of atomic π-charges of the

central atoms. In total, 684 various conjugated molecules were taken into consideration. d) Correlation between the predicted and measured V_2, based on these descriptors (a-c).

A greater π-BO results from a greater electron density across the torsion of interest. Consequently, this resulted in increased rigidity and a greater conjugation strength. Using this sole metric, 18 (out of 684) molecules possessed greater than 2.5 standard deviations away from the predicted trendline, half of which contained an oxyanion. Another six molecules had central atoms containing nitrogen while two other molecules contained hydrogen bonding across the torsion of interest.

Next, π-charges of the central atoms involved in the torsion of interest were plotted against V_2. The π-charges (electron density of the π system) are based on the population of π-electrons (MO coefficients) residing on atoms in a molecule, which when combined with σ-charges, result in the total charge on each atom. In this manner, four distinct populations were observed. These populations could be grouped according to the average sum of atomic charges: -0.25, 0.0, 0.50, and 1.0 q_e (Figure 4.11b). The first of these populations at -0.25 q_e was composed of molecules with a negatively charged oxygen atom γ to the ring. This might be because of π-electron donation from the hydroxylate to the central atom, which strongly decreased its charge, and increased the torsional barrier Figure 4.12a). The second cluster, around 0.0 q_e, consisted of 588 molecules of various types. The third and fourth populations at 0.50 q_e and 1.0 q_e were made up of molecules having one and two non-carbon central atoms in the ring(s), respectively (ie. N: or N^+). N: or N^+, when incorporated into the ring, are generally more electron deficient than its neutral, elemental counterparts. For example, 1-vinyl-1H-pyrrole could delocalize the nitrogen π-electron lone pairs into the ring, which increases its π-partial charge, leading to resonance forms with positive charge on the nitrogen atom (Figure 4.12b). Similarly, the N^+ of 1-vinylpyridin-1-ium possesses a formal positive charge. In general, it was observed that a greater negative charge on the two central atoms increased the strength of conjugation, due to greater electron density and hence rigidity. Since both π-BO and π-charges are calculated from π-electron coefficients/occupancies, an increase in the correlation coefficient on the central atoms, generally led to a lower V_2 (higher absolute value) and higher conjugation strengths.

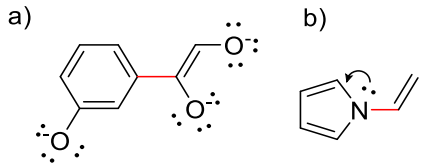


Figure 4.12. a) The donation of electron density of various oxyanions into the central atoms decreased their π-charges. b) The donation of the nitrogen electron lone pair away from the central nitrogen into the ring increased its net π-charge. The torsion of interest is in red.

The relationship between V_2 and sum of π-charges did not appear to be entirely additive. Consequently, a product of π-charges on the two central atoms helped to augment its description of V_2 (Figure 4.12c), by describing the polarity of the central bond. A larger charge polarity difference between the central bond (ie. central bond atoms with +1 and -1 charge) had a lower V_2 and therefore greater conjugation strength than a molecule with similar charges (ie. central bond atoms with +1 and +1 charge). This makes sense as a larger polarity across the torsional bond of interest could be a measure of acceptor/donor strength. When the product of π-charges was used in combination with the other two EHT descriptors, π-bond order and sum of atomic π-charges (Figure 4.12d), an equation to predict the V_2 strength was devised (Equation 4.3).

Equation 4.3. Function to predict V_2 based on various EHT descriptors.

$$V_2 = -62.24 \cdot BO_\pi - 5.96 \cdot (q_{\pi A} + q_{\pi B}) + 12.37 \cdot q_{\pi A} \cdot q_{\pi B} + 14.76$$

In this equation, BO_π denotes the π-bond order across the torsion of interest, while $q_{\pi A}$ and $q_{\pi B}$ are the π-atomic charges on the central atoms, A and B, respectively. It should be mentioned that this equation only considers the effects within the π-framework (ie. π → π* conjugation). Consequently, some of the other effects, σ → σ*, π → σ* and σ → π* hyperconjugation are missing and could be added to improve its accuracy. While σ → σ* hyperconjugation could be and was isolated from NBO calculations, π → σ* and σ → π* hyperconjugation cannot be accurately obtained from NBO due to its tendency to exaggerate the localization of π-orbitals (eg. partitioning the π-system of benzene into three equally localized π-orbitals, instead of treating them together).[49] Consequently, isolating π → σ* and σ → π* orbital interactions are challenging without an appropriate energy decomposition analysis methodology. Until such methods are developed, any torsion prediction would be restricted within the π-framework and σ-framework, respectively.

During the development of these methods, NBO analysis was used to sum up the 8 different σ → σ* hyperconjugation interaction energies during torsion rotation. It was found that this overall V_2 of σ → σ* hyperconjugation obtained from NBO calculations correlated with the overall V_2 of QM energy profile with an R^2 of 0.13 (Figure C30). This correlation is poor, possibly because it is a minor interaction and due to the stronger presence of other interactions, such as nonbonded interactions and π → σ* and σ → π* hyperconjugation effects, which made the σ → σ* hyperconjugation effects not observable. However, when the overall V_2 of σ → σ* was compared to that obtained from the torsional profile of the difference between QM and MM nonbonded energy, an even smaller correlation was unexpectedly observed (Figure C31). This could most likely be attributed to the poor quality of the nonbonded energy terms in the FFs, especially at close distances. This will be further explained in the next section. Fortunately, our obtained results suggest that π → π* conjugation is the dominant torsional interaction and much stronger than the other hyperconjugation interactions and nonbonded interactions. Thus, modelling this energetic stabilization allows a sufficiently accurate prediction of torsional barriers. It should also be mentioned that the higher correlation coefficient of H-TEQ 4.5, when used in combination with AM1-bcc charges and GAFF van der Waals terms, subsequently led to its adoption over those of other FFs.

***V_1* Values of Current Force Fields may be Contaminated with Inaccuracies in the Electrostatics and van der Waals Terms.** In conjugated molecules, σ → σ* hyperconjugation, along with electrostatics and van der Waals interactions, are the main interactions responsible for determining their overall *cis*/*trans* preferences. Consequently, taking the relative energy difference as calculated by QM ($\Delta E_{QM:cis\text{-}trans}$) between the cis and trans geometry and subtracting out the nonbonded interactions ($\Delta E_{NB:cis\text{-}trans}$) should produce the sum of all σ → σ* hyperconjugation energy ($\Delta E_{\sigma\rightarrow\sigma^*:cis\text{-}trans}$). However, when the *$V_1$ values* obtained using the former method for 684 molecules were plotted against the V_1 values of the latter as extracted from NBO, no correlation was observed for any of the FFs (Figure C32).

Although this is unexpected, it might be explained by the inaccuracies of current nonbonded interactions in FFs. Due to the empirical nature of existing FFs, nonbonded interactions are known to be inaccurate, especially when these interactions occur at close distances.[9, 54] In particular, the use of empirical and approximate scaling factors for 1-4 van der Waals and electrostatics in GAFF2 and other FFs is especially problematic when attempting to develop

chemically meaningful torsion parameters. The assumption that various long-range interactions behave similarly, albeit weaker, at short distances is inherently flawed. In addition, there is no agreed-upon consensus for separating nonbonded interactions from that of torsions.[55] Consequently, the results indicate that the obtained V_1 values using different MM methods are contaminated with errors from nonbonded interactions. In the future, in order to incorporate $\sigma \rightarrow \sigma^*$ hyperconjugation in FFs, a more accurate description of nonbonded interactions, especially at close distances (particularly 1-4 nonbonded interactions), should be first addressed.

Despite these challenges, our results also indicated that the *cis*/*trans* preference of 78.1% of all molecules in the training set were correctly predicted without explicitly incorporating a V_1 term in the torsion parameters using GAFF2 nonbonded parameters. More specifically, this meant that QM and GAFF2 nonbonded terms (and by extension H-TEQ 4.5) predicted the same global minimum conformations. It should also be noted that an accuracy of 45.7% and 42.8% were obtained for MMFF94 and MAB for predicting this global minimum conformation, respectively. In addition, 66.6%, 51.1%, and 51.5% of molecules possessed a magnitude of V_1 (the difference between that of QM and MM nonbonded energies) of less than 1.0 kcal·mol^{-1} using GAFF2, MMFF94, and MAB, respectively.

Validation of H-TEQ 4.5 on a Diverse Set of Druglike Molecules. As an unbiased validation of the developed method, H-TEQ 4.5 was tested on 200 diverse druglike molecules and compared to other FFs. On these molecules, our developed method achieved an average RMSE of 1.01 kcal·mol^{-1}, with a sharp peak signifying the mode at 0.54 kcal·mol^{-1} (Figure 4.13). More specifically, H-TEQ 4.5 successfully predicted 63.5%, 90.0%, and 98.0% of molecules below an RMSE of 1.0, 2.0, and 3.0 kcal·mol^{-1}, respectively. The remaining 2.0% of the molecules in the validation set were predicted with high RMSEs between 3.0 and 4.0 kcal·mol^{-1}. For MMFF94, MAB, and GAFF2, average RMSEs of 1.50 1.77, and 3.49 kcal·mol^{-1} were obtained respectively. Both MMFF94, GAFF2, and MAB performed well for most molecules in the validation set as could be seen in Figure 4.13 by the relatively sharp peak at around 0.55 kcal·mol^{-1}. However, they suffered from high RMSEs for many other molecules, as evidenced by the trailing "tail" region of the smoothed histogram. In fact, the torsional errors of molecules in the validation set were predicted by MMFF94, MAB, and GAFF2 to be as high as 15.23, 10.48, and 17.40 kcal·mol^{-1} RMSE, respectively for some molecules. This could be attributed to the lack of suitable torsion parameters in MMFF94 and GAFF2 due to their poor transferability. For example, MMFF94

predicted the incorrect energy minima at the torsion angles of ±90° for (S)-6-(4-(1H-pyrrol-3-yl)phenyl)-5-methyl-4,5-dihydropyridazin-3(2H)-one and 2-amino-2-thioxoethane(dithioperoxo)imidic (Figure 4.14c and i, respectively), leading to RMSEs of 2.32 and 1.65 kcal·mol^{-1} compared to 0.51 and 1.53 kcal·mol^{-1} for H-TEQ 4.5. While GAFF2 predicted the correct minima for these two molecules, it dramatically overpredicted the strength of their torsional energy barrier, leading to RMSEs of 3.01 and 5.24, respectively (Figure C33).

Despite MAB being an atom type-independent method, its poorer performance could be due to a limited number of molecules in its training set. Overall, H-TEQ 4.5 was shown to have notably smaller errors as compared to the other tested FFs for conjugated torsional energy barriers.

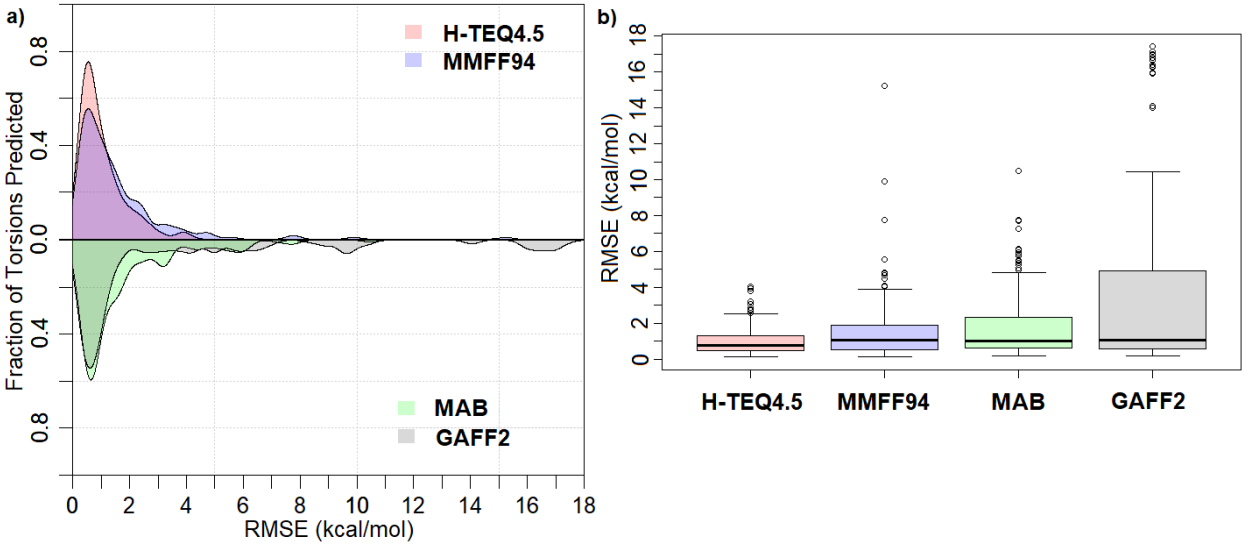


Figure 4.13. a) A smoothed histogram and b) a boxplot indicating the number of MM energy profile predictions made by H-TEQ 4.5 (pink), MMFF94 (blue), MAB (green), and GAFF2 (grey) for the 200 molecules in the validation set. The QM energy profile was used as the reference, with RMSE values closer to 0.0 kcal·mol^{-1} representing more accurate predictions.

Forty-two molecules were predicted by H-TEQ 4.5 to have RMSEs of greater than 1.5 kcal·mol^{-1}. The sources of these errors could be attributed to two main factors: the lack of a V_1 model (which accounted for 24 molecules) and imperfections in the V_2 model (which accounted for 18 molecules). This categorization was performed by plotting the RMSEs of these 42 molecules against the magnitude of the V_1 obtained from the difference of QM and GAFF2 nonbonded energies (Figure C34). It should be clarified that these V_1 terms are the ideal values required for a perfect torsional energy barrier. Molecules exhibiting a V_1 magnitude greater than 1.0 kcal·mol^{-1} were classified as V_1 deficient while those less than this cut-off was categorized as V_2 imperfect.

Despite the importance of V_1 terms for conjugated molecules in FFs, they were found to be contaminated with nonbonded interactions, as discussed in an earlier section. For example, it could be observed that the ±180° geometry of 2-amino-2-thioxoethane(dithioperoxo)imidic acid (Figure 4.14i) is favored compared to that of the 0° conformation. There are two potential sources of interactions to explain this behavior. The first interaction disfavoring the ±180° geometry could be a strong electrostatic repulsion between the electron lone pairs of the thialdehyde sulphur and imine nitrogen group, which was not captured by MM. This is a known limitation of type I FF, lacking explicit electron lone pairs. Consequently, accurate nonbonded energies, along with a robust σ → σ* hyperconjugation V_1 model in the FF, would enhance the accuracy of the FF. As another example, the underestimation of the electrostatic at the ±180° conformation of [2,2'-bipyridine]-1,1'-diium (between the two positively charged nitrogen atoms in the 1-4 position) could also be responsible for the inaccurate prediction of its torsion profile (Figure 4.4.14d). Interestingly, an increase in the 1-4 electrostatic and van der Waals energy by a factor of approximately 6.5 when using H-TEQ 4.5 allowed it to match the QM energy profile for this molecule (Figure C35). It should be noted that for this molecule, MMFF94 predicted a more robust torsional profile than H-TEQ 4.5. This is because MMFF94 used a different charging scheme than that used by H-TEQ 4.5 and GAFF2. Due to these problems, the next step in FF development should be to focus on and improve the treatment of electrostatics interactions which could be applicable to all small druglike molecules.

On the other hand, 19 out of the 200 molecules in the validation set possessed imperfect V_2 values. More specifically, it was found that H-TEQ 4.5 performed poorer on molecules containing significantly charged centers (Figure 4.14d and e). This could be due to various errors, including limitations in the EHT method, the neglect of σ → σ*, σ → π*, and π → σ* hyperconjugation interactions, and poor nonbonded interactions of the GAFF2. In order to include the effects of other hyperconjugation interactions, an energy decomposition methodology (EDA), appropriate for treating π-systems, should first be devised. While NBO provides a decomposition of each of these energies, its treatment of conjugated systems is not suitable for these purposes, as mentioned previously. Although other EDA methods exist, such as local energy decomposition (LED),[56] natural energy decomposition (NEDA),[57] absolutely localized molecular orbitals (ALMO-EDA),[58] and block-localized wavefunction energy decomposition analysis (BLW-EDA),[59] these methods are currently unable to further decompose to the specific types of hyperconjugation. These methods

provide only an amalgamated charge-transfer term, which combines all hyperconjugation and conjugation interactions into a single stabilization energy. However, despite the drawbacks of these methods, H-TEQ 4.5 predicted the torsional energy profiles of most molecules accurately, by focusing on the dominant $\pi \rightarrow \pi^*$ conjugation interaction (Figure 4.14a-c,f,h).

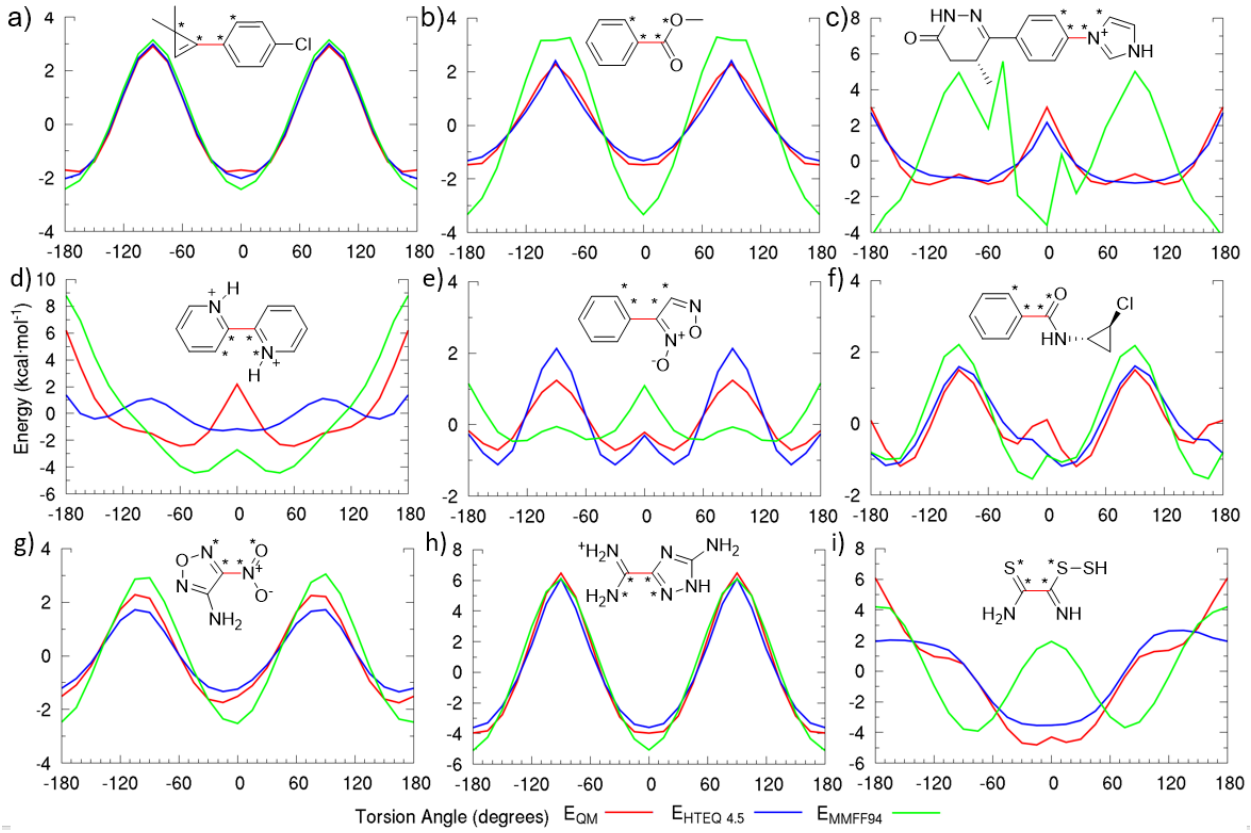


Figure 4.14. Nine representative structures and their corresponding torsional energy scans as calculated by MP2/6-311+G** (red), H-TEQ 4.5 (blue), and MMFF94 (green) for: a) 1-chloro-4-(3,3-dimethylcycloprop-1-en-1-yl)benzene, b) methyl benzoate, c) (R)-3-(4-(4-methyl-6-oxo-1,4,5,6-tetrahydropyridazin-3-yl)phenyl)-1H-imidazol-3-ium, d) [2,2'-bipyridine]-1,1'-diium, e) 3-phenyl-1,2,5-oxadiazole 2-oxide, f) N-((1S,2S)-2-chlorocyclopropyl)benzamide, g) 4-nitro-1,2,5-oxadiazol-3-amine, h) amino(5-amino-1H-1,2,4-triazol-3-yl)methaniminium, and i) 2-amino-2-thioxoethane(dithioperoxo)imidic acid. The reference torsional bond of interest is marked in red. The reference angle at 0° is marked by a series of four asterisks (*).

4.5 Conclusions

Over the recent years, vHTS has become increasingly important and a staple tool for drug discovery, allowing an enrichment of active compounds in the preliminary phase.[60] The reliance of vHTS on a robust and accurate FF has led to numerous efforts in FF development, spanning many decades.[10, 20, 24, 26, 27] Unfortunately, however, the accuracy and transferability of these FFs are still poor for many druglike molecules. This is especially true for conjugated species due to the unpredictability of their torsions despite their abundance in nature. In fact, many of these conjugated moieties are important pharmacophores. In this paper, we showed that the torsional energy parameters of common FFs were inaccurate for conjugated molecules. Conventionally, FFs have relied on a set of pre-parametrized atom types, which possessed poor transferability between molecules. In fact, despite accurately predicting the torsional energy profiles for few selected druglike molecules, GAFF2 and MMFF94 were found to struggle for others. In addition, the original MAB may also have been poor in accuracy due to the limitated number of molecules in its training set and a lack of large-scale validation. In order to improve the accuracy of the FF, a conceptually novel method to predict torsional barriers was envisioned. The developed method, H-TEQ 4.5, relies on an understanding of the molecular interactions governing torsional energy barriers, allowing the FF to adapt and tailor parameters suited for new molecules.

To this end, EHT was used to calculate the π-BO, sum of π-charges, and product of π-charges, which correlated well with the torsional energy barriers of conjugated molecules. These EHT obtained properties were chosen to capture and describe the π-electron coefficients across the torsional bond of interest. Using these three descriptors, the accuracies of H-TEQ 4.5, MMFF94, GAFF2, and MAB were tested on 200 torsions of conjugated molecules. This yielded an average RMSE of 1.01, 1.50, 3.49, and 1.77 kcal·mol^{-1}, respectively. The higher accuracy of H-TEQ 4.5 compared to the rest of these FFs showed that atom-type independent FFs are promising as an accurate and computationally tractable method to assign torsions in MMs. It should be noted that, in contrast to QM methods used to derive parameters, HTEQ 4.5 derives parameters within a fraction of a millisecond (ie. approximately 0.25 ms for a biphenyl molecule on a single core of an Intel Core i7 6700HQ CPU), allowing it to be applied to vHTS of millions of molecules.

In the future, it is envisioned that more advanced and chemically perceptive FFs could be developed. Future efforts and emphasis in FF developments should be placed on the accurate treatment of electrostatics and van der Waals interactions, especially for those at close ranges (1-

4) interactions. Until such interactions are robustly modelled in FFs, further improvements to predicting torsional parameters would be very difficult.

4.6 References

1. Bringmann, G.; Walter, R.; Weirich, R., The Directed Synthesis of Biaryl Compounds: Modern Concepts and Strategies. *Angew. Chem. Int. Ed.* **1990**, *29*, 977-991.

2. Cayla, N. S.; Dagne, B. A.; Wu, Y.; Lu, Y.; Rodriguez, L.; Davies, D. L.; Gross, E. R.; Heifets, B. D.; Davies, M. F.; MacIver, M. B.; Bertaccini, E. J., A Newly Developed Anesthetic Based on a Unique Chemical Core. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, 201822076.

3. Ertl, P., An Algorithm to Identify Functional Groups in Organic Molecules. *J Cheminform* **2017**, *9*, 36.

4. World Health, O. *Who Model List of Essential Medicines, 20th List (March 2017, Amended August 2017)*; World Health Organization: Geneva, 2017, 2017.

5. Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M., Drugbank 5.0: A Major Update to the Drugbank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074-D1082.

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

7. Gallivan, J. P.; Dougherty, D. A., Cation-Π Interactions in Structural Biology. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9459-9464.

8. Schottel, B. L.; Chifotides, H. T.; Dunbar, K. R., Anion-Pi Interactions. *Chem. Soc. Rev.* **2008**, *37*, 68-83.

9. Wei, W. L.; Champion, C.; Liu, Z. M.; Barigye, S. J.; Labute, P.; Moitessier, N., Torsional Energy Barriers of Biaryls Could Be Predicted by Electron Richness/Deficiency of Aromatic Rings; Advancement of Molecular Mechanics toward Atom-Type Independence. *J. Chem. Inf. Model* **2019**, *59*, 4764-4777.

10. Dahlgren, M. K.; Schyman, P.; Tirado-Rives, J.; Jorgensen, W. L., Characterization of Biaryl Torsional Energetics and Its Treatment in Opls All-Atom Force Fields. *J. Chem. Inf. Model* **2013**, *53*, 1191-1199.

11. Sellers, B. D.; James, N. C.; Gobbi, A., A Comparison of Quantum and Molecular Mechanical Methods to Estimate Strain Energy in Druglike Fragments. *J. Chem. Inf. Model* **2017**, *57*, 1265-1275.

12. Corbeil, C. R.; Englebienne, P.; Yannopoulos, C. G.; Chan, L.; Das, S. K.; Bilimoria, D.; L'Heureux, L.; Moitessier, N., Docking Ligands into Flexible and Solvated Macromolecules. 2. Development and Application of Fitted 1.5 to the Virtual Screening of Potential Hcv Polymerase Inhibitors. *J. Chem. Inf. Model* **2008**, *48*, 902-909.

13. Cournia, Z.; Allen, B.; Sherman, W., Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **2017**, *57*, 2911-2937.

14. Jorgensen, W. L.; Bollini, M.; Thakur, V. V.; Domaoal, R. A.; Spasov, K. A.; Anderson, K. S., Efficient Discovery of Potent Anti-Hiv Agents Targeting the Tyr181cys Variant of Hiv Reverse Transcriptase. *J. Am. Chem. Soc.* **2011**, *133*, 15686-15696.

15. Jorgensen, W. L.; Ravimohan, C., Monte Carlo Simulation of Differences in Free Energies of Hydration. *J. Chem. Phys.* **1985**, *83*, 3050-3054.

16. Heinzelmann, G.; Baştuğ, T.; Kuyucak, S., Free Energy Simulations of Ligand Binding to the Aspartate Transporter Gltph. *Biophys. J.* **2011**, *101*, 2380-2388.

17. Reymond, J. L., The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48*, 722-730.

18. 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.

19. Allen, A. E. A.; Robertson, M. J.; Payne, M. C.; Cole, D. J., Development and Validation of the Quantum Mechanical Bespoke Protein Force Field. *ACS Omega* **2019**, *4*, 14537-14550.

20. Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M., Charmm: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545-1614.

21. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26*, 1668-1688.

22. 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.

23. 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 (Vol 117, Pg 5179, 1995). *Journal of the American Chemical Society* **1996**, *118*, 2309-2309.

24. 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.

25. http://www.ccl.net/cca/data/parm_at_Frosst/ (March 3),

26. Jorgensen, W. L.; McDonald, N. A., Development of an All-Atom Force Field for Heterocycles. Properties of Liquid Pyridine and Diazenes. *Theochem-J Mol Struc* **1998**, *424*, 145-155.

27. 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.

28. Horton, J. T.; Allen, A. E. A.; Dodda, L. S.; Cole, D. J., Qubekit: Automating the Derivation of Force Field Parameters from Quantum Mechanics. *J. Chem. Inf. Model* **2019**, *59*, 1366-1381.

29. Grimme, S., A General Quantum Mechanically Derived Force Field (QMDFF) for Molecules and Condensed Phase Simulations. *J. Chem. Theory Comput.* **2014**, *10*, 4497-4514.

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

34. Gerber, P. R.; Muller, K., Mab, a Generally Applicable Molecular-Force Field for Structure Modeling in Medicinal Chemistry. *J. Comput. Aided Mol. Des.* **1995**, *9*, 251-268.

35. 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.

36. Mobley, D. L.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Slochower, D. R.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K., Escaping Atom Types in Force Fields Using Direct Chemical Perception. *J. Chem. Theory Comput.* **2018**, *14*, 6076-6092.

37. Champion, C.; Barigye, S. J.; Wei, W.; Liu, Z.; Labute, P.; Moitessier, N., Atom Type Independent Modeling of the Conformational Energy of Benzylic, Allylic, and Other Bonds Adjacent to Conjugated Systems. *J. Chem. Inf. Model* **2019**.

38. Liu, Z.; 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.

39. Smith, J. S.; Roitberg, A. E.; Isayev, O., Transforming Computational Drug Discovery with Machine Learning and AI. *ACS Med. Chem. Lett.* **2018**, *9*, 1065-1069.

40. Mullins, J., Hyperconjugation: A More Coherent Approach. *J. Chem. Educ.* **2012**, *89*, 834-836.

41. Plavec, J.; Tong, W.; Chattopadhyaya, J., How Do the Gauche and Anomeric Effects Drive the Pseudorotational Equilibrium of the Pentofuranose Moiety of Nucleosides? *J. Am. Chem. Soc.* **1993**, *115*, 9734-9746.

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. Barth, A., Status and Future Developments of Reaction Databases and Online Retrieval Systems. *Journal of Chemical Information and Computer Sciences* **1990**, *30*, 384-393

44. Fernández, I.; Frenking, G., Correlation between Hammett Substituent Constants and Directly Calculated Π-Conjugation Strength. *J. Org. Chem.* **2006**, *71*, 2251-2256.

45. 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.

46. 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.

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

48. 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.

49. E. D. Glendening, J., K. Badenhoop, A. E. Reed, J. E. Carpenter, J. A. Bohmann, C. M. Morales, P. Karafiloglou, C. R. Landis, and F. Weinhold *Nbo 7.0*, Theoretical Chemistry Institute, University of Wisconsin, Madison, 2018.

50. Chen, Z., Graphical Representation of Hückel Molecular Orbitals. *J. Chem. Educ.* **2020**, *97*, 448-456.

51. Hoffmann, R., An Extended Hückel Theory. I. Hydrocarbons. *Journal Chem. Phys.* **1963**, *39*, 1397-1412.

52. Kier, L. B., *Molecular Orbital Theory in Drug Research*. Academic Press: New York,, 1971; p xi, 258 p.

53. *Molecular Operating Environment (Moe) 2019.0102*, Chemical Computing Group ULC: Montréal, QC, Canada, 2020.

54. 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.

55. Monticelli, L.; Tieleman, D. P., Force Fields for Classical Molecular Dynamics. *Methods Mol. Biol.* **2013**, *924*, 197-213.

56. Schneider, W. B.; Bistoni, G.; Sparta, M.; Saitow, M.; Riplinger, C.; Auer, A. A.; Neese, F., Decomposition of Intermolecular Interaction Energies within the Local Pair Natural Orbital Coupled Cluster Framework. *J. Chem. Theory Comput.* **2016**, *12*, 4778-4792.

57. Glendening, E. D., Natural Energy Decomposition Analysis: Explicit Evaluation of Electrostatic and Polarization Effects with Application to Aqueous Clusters of Alkali Metal Cations and Neutrals. *J. Am. Chem. Soc.* **1996**, *118*, 2473-2482.

58. Thirman, J.; Head-Gordon, M., Efficient Implementation of Energy Decomposition Analysis for Second-Order Møller–Plesset Perturbation Theory and Application to Anion−Π Interactions. *J. Phys. Chem. A* **2017**, *121*, 717-728.

59. Mo, Y.; Bao, P.; Gao, J., Energy Decomposition Analysis Based on a Block-Localized Wavefunction and Multistate Density Functional Theory. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6760-6775.

60. Kontoyianni, M., Docking and Virtual Screening in Drug Discovery. *Methods Mol. Biol.* **2017**, *1647*, 255-266.

5 Influence of Molecular Mechanics Torsional Parameters on Docking Accuracies

5.1 Preface

The development of a conceptually novel type of molecular mechanics force field for small organic molecules was undertaken in the previous two chapters. In **Chapter 3 and 4**, parameters could be generated on-the-fly, based on atomic and molecular descriptors based on the degree of conjugation observed in molecules. In conjunction to previous developments conducted by Liu et al (2016 and 2018) and Champion et al (2019), there is almost complete coverage of the chemical space (apart from conjugated molecules containing electron lone pairs). As a result, it was decided that the current implementation of this novel force field be tested for practical applications during *in silico* docking of both proteins and nucleic acids. This chapter presents the results of these findings and compares our developed force field against other commonly used force fields in the literature. This chapter serves to benchmark where we are in terms of small molecule force field developments and would be helpful to guide future research directions.

5.2 Introduction

The binding affinity of pharmaceuticals to receptor macromolecules has important implications biochemically within living organisms. Increased binding affinity of drugs to their biological targets would improve the potency and effectiveness of treatments, and require less dosage, diminishing off-target effects and toxicity. The ability to predict the free energy of binding has been the subject of many studies and have inspired developments in experimental and computational techniques. Experimental determination of binding affinities could be achieved by measuring the kinetics and thermodynamics of binding (e.g. K_i's, which could be converted to binding energies), through methods such as isothermal titration calorimetry (ITC) and others.[1] However, these methods are time-intensive, and require the synthesis of the drug candidates and the expression and purification of the target macromolecule.

To accelerate and aid this process, computational chemistry methods have been developed to quickly and inexpensively screen drug compounds for a given protein or nucleic acid receptor. The most successful of these techniques are molecular docking methods, which are useful for predicting the binding mode and their associated energies, although with limited accuracy.[2-5] Docking methods employ various global search algorithms to rapidly identify (ie. conformational sampling) and score (ie. binding affinity evaluation) all ligand conformations within a static receptor structure. Due to the static nature of this method, configuration and solvation entropy changes of the ligand must be estimated using various means. This technique is computationally inexpensive and has been previously used to screen billions of compounds. Molecular docking is especially useful during the drug discovery and lead optimization stages of medicinal chemistry projects. Its recent application to screening 1.3 billion compounds for screening COVID-19 inhibitors is a testament to its wide adoption and past successes in enriching a compound library for potential actives.[4]

Molecular dynamics (MD) simulations are also useful in predicting the binding affinity, if the correct binding mode is known. MD simulation explores the time evolution of biomolecules and gives insights into receptor-ligand interactions and possible conformational changes.[6] A technique, known as free energy perturbation (FEP) method, could be applied during an MD simulation to calculate the free energy differences between ligands through alchemical transformations.[7] In FEP calculations, the free energy difference between two ligands, ligand A and B, would be computed. This is carried out by first performing an MD simulation on ligand A.

Subsequently, its parameters are slowly mutated into that of ligand B, in small incremental steps, solving the Zwanzig equation at each iteration. After the completion of the transformation, the free energy difference is obtained. This process would be repeated for other, chemically similar ligands. Ligand annihilation is also possible to calculate the absolute free energy of ligand binding. Unlike molecular docking, FEP is dynamic. Consequently, conformational entropy would be more accurately computed. Despite the dynamic nature of FEP, both methods are less reliable when the binding site is highly flexible or changes significantly upon ligand binding.[8, 9]

Although widely used, both docking and FEP rely on an energy evaluation method for determining the conformational preferences and strengths of interactions between the receptor and the ligand. To this end, classical molecular mechanics (MM) force fields (FFs) have been the method of choice, owing to its speed, despite trade-offs in accuracy.[10-13] MM FFs are based on Newtonian classical mechanics, which treat all atoms in a molecule as balls connected by springs (ie. potentials). As a result, quantum mechanical (QM) behaviors are approximated by a series of empirical equations, which when combined attempts to reproduce the total QM energy of the system. These empirical equations are classified according to bonded and nonbonded interactions (equation 5.1) in classical force fields. The former includes bond stretch, angle stretch, and torsional rotation, while the latter contains electrostatic and van der Waals interactions. During receptor-ligand binding events, bond and angle stretching terms are not expected to change significantly. Consequently, binding affinities and preferred conformations are predominately determined by torsions, electrostatics and van der Waals interactions.

Equation 5.1. Functional form found in type I MM FFs.

$$E_{MM} = \sum_{bond} k_b(r - r_0)^2 + \sum_{ang} k_\theta(\psi - \psi_0)^2 + \sum_{tor} \sum_{n=1}^{4} \frac{V_n}{2}[1 + \cos(n\theta - \theta_p)] + \sum_{ele} \frac{kq_1q_2}{r} + \sum_{vdw} \left(\frac{A}{r^{12}} - \frac{B}{r^6}\right)$$

In the equation above, the bond stretching term is composed of a spring constant and equilibrium distance of k_b and r_0, respectively. Similarly, the angle term is composed of a spring constant and equilibrium angle, k_θ and ψ_0, respectively. Torsion terms uses a truncated Fourier series, where V_n denotes the height, n denotes the periodicity, and θ_p denotes the phase-shift. The coulombic term,

is composed of the coulomb's constant, k and distance of separation, r. The van der Waals is composed of A and B, which are parameters to modulate the steepness of the steric and attractive forces due to dispersion, respectively. In widely used FFs, such as Amber and OPLS, electrostatic charges are obtained using the (restrained) electrostatic potential (RESP; ESP) method.[14, 15] Using QM calculations, classical atomic partial charges are assigned based on a series of test probes surrounding the ligand molecule. Due to the higher computational costs associated with running Hartree-Fock (HF) calculations, AM1-BCC charges were developed. The latter was designed to closely reproduce RESP charges through semiempirical calculations.[16] This approach makes the charges accurate for intermolecular interactions. However, despite the successes of these methods, the performance of AM1-BCC and RESP charges may suffer at close distances, especially in 1-4 interactions, as was alluded to previously.[17, 18] In addition, both charging schemes are conformation-dependent, which may bias a specific conformation during rotamer search or MD simulation. As a result, more advanced polarizable FFs have been envisioned and developed, although their accuracies are not yet proven. The CHARMM FF is unique in that it optimizes partial charges of atoms based on QM interactions with water.[19]

Other charge models, such as bond charge increment and empirical rules, taking electronegativity into consideration, are more routinely used in small molecule FFs, such as MMFF94 and MAB.[20, 21] These latter charging schemes are conformation-independent and do not require QM calculations. While both charge methods are designed to reproduce ESPs, they may suffer from lower accuracies due to their approximations and lack of reliance on QM calculations.

Despite Morse potential being well-suited to reproduce van der Waals interactions, its high computational costs have led to the adoption of other potentials, such as Lennard-Jones (LJ) and Buckingham functions.[22] Although LJ is used extensively in Amber FFs, CHARMM, OPLS, and others for simulating proteins, nucleic acids, and small molecules, its steep steric wall is unsuitable for atoms in close contact. To alleviate this issue, MMFF94 uses a buffered LJ 14-7 potential while MM2 and MM3 FFs use the Buckingham potential. Traditionally, hydrogen bonding had been a separate potential in some FFs, making use of a LJ 12-10 term.[23] However, most FFs, including Amber and CHARMM, have abandoned an explicit hydrogen bonding term. Rather, its chemical effects have been rolled into that of van der Waals and electrostatics, although recent evidences have suggested that a distinct hydrogen bonding potential may improve the overall accuracy of FFs.[24]

Lastly, torsional energy barriers are also expected to contribute to conformational preferences of certain molecules.[17, 18, 25, 26] Chemically, the degree of hyperconjugation of electron density from filled to vacant molecular orbitals (MOs) are believed to play a major role in determining the strength and orientation of torsions. These are exemplified in classical cases such as the gauche and anomeric effects (Figure 5.1). In the past, incorporating hyperconjugation into classical MM FFs, using empirical organic chemistry principles and extended-Hückel theory (EHT) descriptors, was shown to improve the prediction of torsional parameters.[17, 18, 25, 26] These methods, named H-TEQ (Hyperconjugation for Torsional Energy Quantification), developed in our laboratory, has recently been improved in terms of accuracy and transferability. Although quantification of torsions containing $\eta \rightarrow \pi^*$ interactions are still currently in development (e.g. aniline and 2-methoxypyridine), we sought to test the performance of existing H-TEQ (versions 1.0 to 4.5) methods towards molecular docking and computing conformational energies. The former would evaluate its usefulness towards an important method in drug discovery and medicinal chemistry. The latter would test the gas-phase accuracy of torsion parameters obtained by H-TEQ.

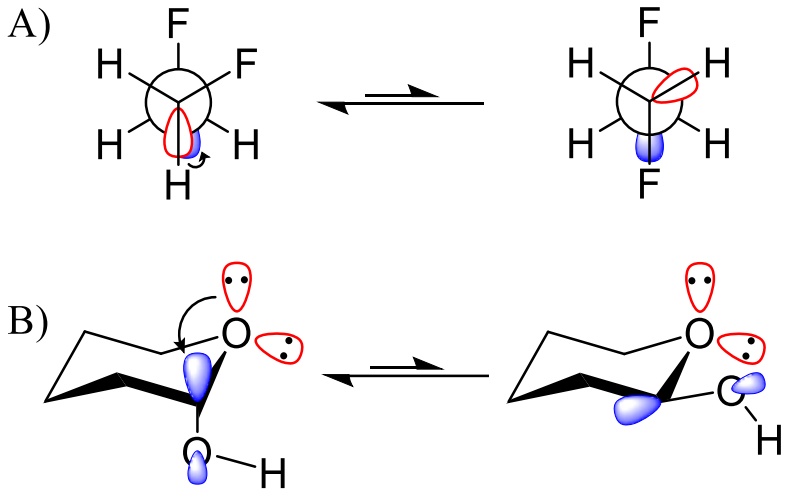


Figure 5.1. Hyperconjugation interactions are responsible for the observed conformational preferences in small molecules. A) The $\sigma_H \rightarrow \sigma_F^*$ hyperconjugation causes the gauche effect in 1,2-difluoroethane. B) Similarly, $\eta \rightarrow \sigma_O^*$ causes the anomeric effect in (R)-tetrahydro-2H-pyran-2-ol. In both cases, donation of electron density from occupied (red) to unoccupied (blue) orbitals occur.

5.3 Computational Methods

Self-docking. All *in silico* docking was performed using our program FITTED. Self-docking was carried out on a curated set of structures obtained from the Protein Data Bank (PDB). Self-docking is a method which gauges the accuracy of the docking protocol. This is performed by extracting the bound ligand and redocking into the binding site of the receptor. The root-mean square deviation (RMSD) of the docked pose was subsequently compared with that of the original ligand. In self-docking, an RMSD of less than 2.0 Å denotes a successful docking, which corresponds approximately to the distance of a strong hydrogen bond. Although this metrics has limitations, it is believed to be appropriate to evaluate differences in performance of different docking conditions on the same sets. The composition of this set included 178 nucleic acids and 289 proteins, which were co-crystallaized with small molecule ligands. The latter belonged to a high-quality set of structures, recommended by Astex.[27]

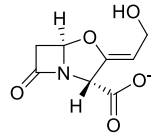
For nucleic acid docking, ligands which were nearly, but not exactly symmetric were removed from the dataset. Since X-ray structures are generated from electron density, these ligands possesses similar electron densities in more than one orientation. Consequently, docking of the ligands yields significantly high RMSDs if docked in the reverse direction. These ligands were especially prevalent in DNA minor groove binders.

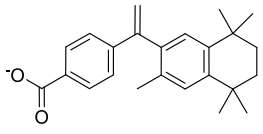
For each PDB structure, three independent docking trials were carried out with different FFs: GAFF, GAFF with H-TEQ torsional parameters (henceforth called $GAFF_{Tor:H\text{-}TEQ}$) and GAFF with all torsions set to zero (henceforth called $GAFF_{Tor:0}$). For the small number of missing dihedral parameters still under development, GAFF parameters were used instead. In each docking trial, 10 independent runs were performed, where the most energetically favored pose was extracted. To preclude the possibility of the search algorithm favoring any specific torsion model by chance or due to other inherent biases, the top pose of each FF was scored by the other two. Finally, the accuracies of each FF were compiled, analyzed, and compared. The set of protein and nucleic acid PDB structures are found in Appendix D1 and D2, respectively.

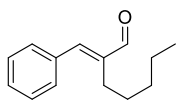
Conformational Energy Analysis. 51 drug molecules were randomly selected from the DrugBank database, which possessed torsional parameters covered by existing H-TEQ methods (Figure 5.2). Emphasis were placed on drug molecules with several rotatable torsions. SMILES strings were used to generate the 3D structures of pharmaceutical drugs within Molecular

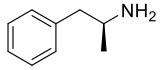
Operating Environment (MOE). The Scientific Vector Language (SVL) was used to randomly generate 20 rotamers, which were not sterically clashing. A steric clash was defined as two atoms within 2.0 Å, excluding atoms directly and indirectly bonded (via a shared, third atom). These rotamers were not subject to optimization, as the goal was to test the accuracy of torsional parameters at reproducing non-equilibrium geometries. The latter are often encountered during MD simulations and, to a lesser degree, during *in silico* docking.

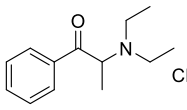
These generated conformations were then subjected to both QM and MM single point calculations. The former was performed at the MP2/6-311+G** level of theory using GAMESS-US, while the latter used MMFF94 and MAB FFs as available in MOE and GAFF2 as obtained using Antechamber. Partial charges were assigned according to the default charging scheme for each FF. For MMFF94 and MAB, bond charge increments was used, while GAFF, AM1-BCC charges were applied. H-TEQ parameters were generated by an SVL program, which replaced the torsional component of GAFF2 and MMFF94. The relative energy of each QM and MM method was obtained by taking the raw energy values and subtracting the mean energy of the 20 conformations. The QM relative energies were directly compared to that of the MM to calculate the absolute errors of the latter.

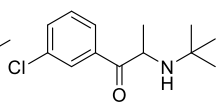

clavulanic acid

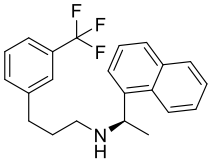

bexarotene

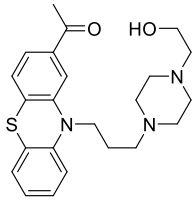

α amyl cinnamic aldehyde

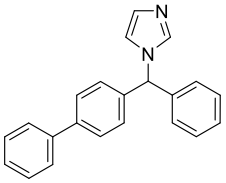

dextronmphetamine

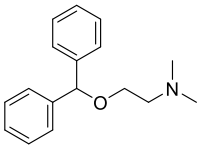

diethylpropion

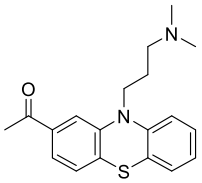

bupropion

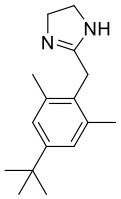

cinacalcet

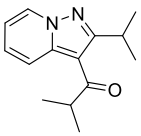

acetophenazine

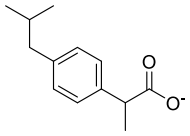

bifonazole

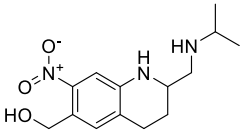

diphenhydramine

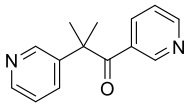

acepromazine

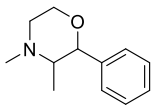

xylometazoline

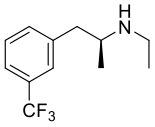

ibudilast

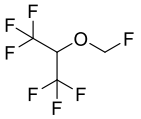

ibuprofen

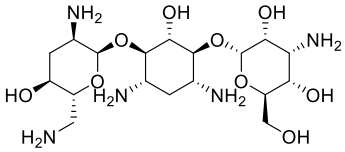

oxamniquine

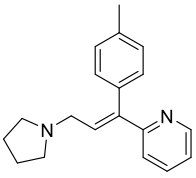

metyrapone

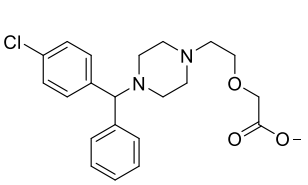

phendimetrazine

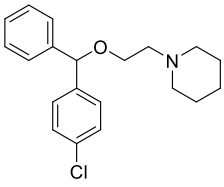

dexfenfluramine

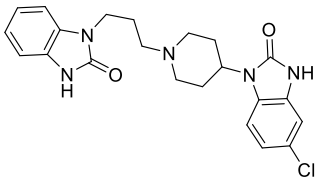

sevoflurane

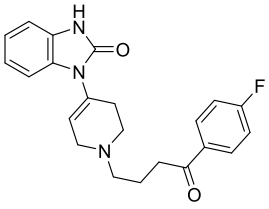

tobramycin

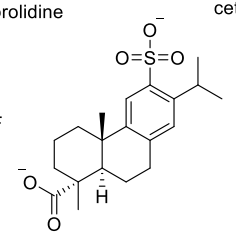

triprolidine

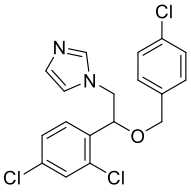

cetirizine

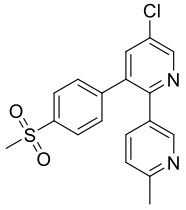

cloperastine

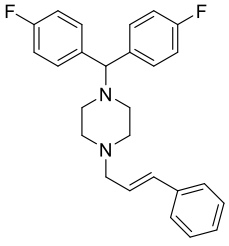

domperidone

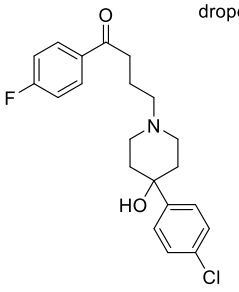

droperidol

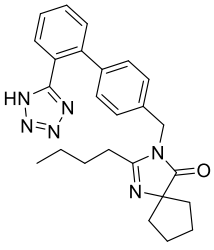

ecabet

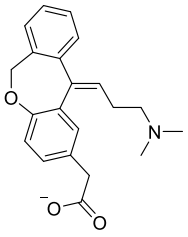

econazole


etoricoxib

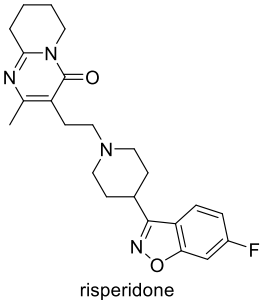

flunarizine

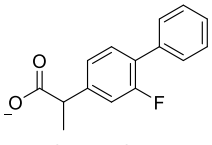

haloperidol

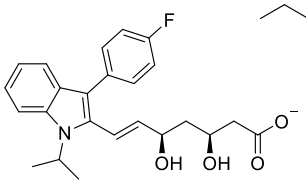

irbesartan

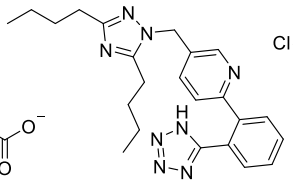

olopatadine

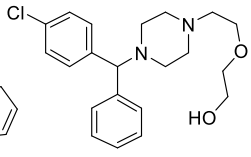

pentolinium

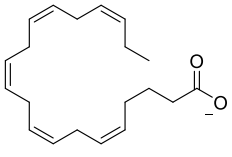

risperidone


flurbiprofen


fluvastatin


forasartan


hydroxyzine


icosapent

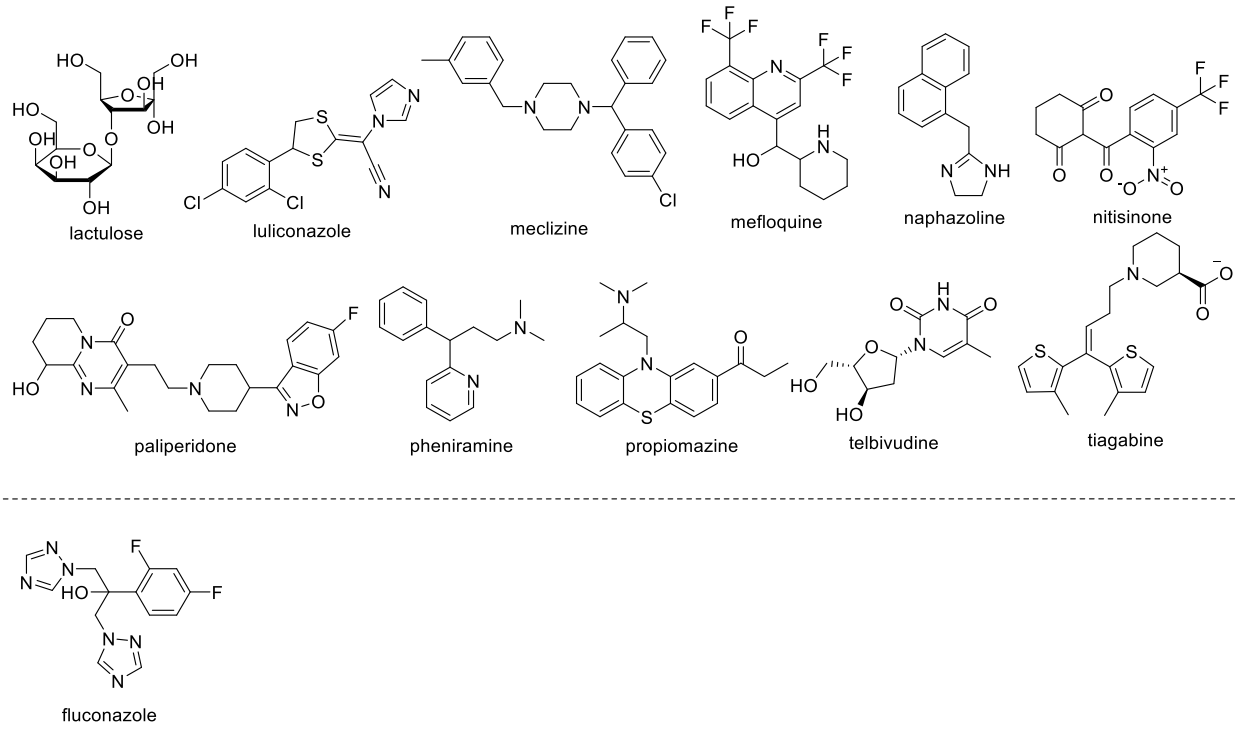


Figure 5.2. 51 drug molecules were subjected to conformational energy analysis. The 50 drugs above and 1 drug below the dotted line represents those with and without steric clashes in MM, respectively.

5.4 Results and Discussion

Protein Docking. Self-docking on the Astex set of 289 protein-ligand complexes using GAFF with various torsion parameters showed that it was crucial for docking. In fact, an accuracy of 71.3%, 71.6%, and 65.7% for GAFF, $GAFF_{Tor:\ H\text{-}TEQ}$, and $GAFF_{Tor:0}$ was achieved, respectively (Figure 5.3). The use of torsion terms on the ligands, regardless of GAFF or H-TEQ, was found to improve the self-docking accuracy by greater than 5%. Unfortunately, the use of H-TEQ torsion parameters in place of that of GAFF did not significantly affect the accuracy of docking. This is likely attributed to the similar accuracy of H-TEQ and GAFF torsions, which was later observed in gas-phase QM and MM calculations (below).

While most structures (270) remained unchanged when docked using GAFF or $GAFF_{Tor:H\text{-}TEQ}$, 19 structures did change. 10 out of the 289 structures recorded improved accuracies when docked using $GAFF_{Tor:H\text{-}TEQ}$ over GAFF, while 9 out of the 289 structures obtained better

accuracies when docked using GAFF over $GAFF_{Tor:\ H\text{-}TEQ}$. In the past, it had been shown that H-TEQ 4.5 performed significantly better than GAFF torsional energy barriers. Conversely, peptide-like ligands seemed to perform better when using GAFF over $GAFF_{Tor:\ H\text{-}TEQ}$, perhaps due to significant parametrization efforts during the development of the AMBER for peptides and proteins.

Despite the change in accuracy between GAFF and $GAFF_{Tor:\ H\text{-}TEQ}$ being very small, the difference between them and $GAFF_{Tor:0}$ was more significant. For example, 64 (or 22.1% of the) dockings improved when GAFF and $GAFF_{Tor:\ H\text{-}TEQ}$ was used in place of $GAFF_{Tor:0}$. Conversely, 44 (or 15.2% of the) structures observed an increase in accuracy when transitioning in the reverse direction. This suggests that although GAFF and H-TEQ torsion parameters are acceptable for many molecules, there exists a smaller number of ligands molecules for which torsion parameters are inaccurately parametrized. For this latter class of molecules, having no torsion parameters was more beneficial possibly due to two potential reasons. One reason is because a combination of van der Waals, hydrogen bonding, and electrostatics was better able to describe their conformational preferences. The other reason may be due to the poor and erroneous assignment of torsion parameters, which resulted in an incorrect conformation.

In general, molecules which performed better using torsions assigned by GAFF and H-TEQ seemed to include many charged groups (Figure 5.4a-c). Those molecules which exhibited poorer accuracies by using these torsional parameters had fewer charged groups (Figure 5.4d-f). Despite this, many exceptions were identified and it was not clear how and why torsional parameters impacted the accuracy of docking differently for different molecules. Nevertheless, in most cases, there was a clear advantage to using torsional parameters generated by GAFF and H-TEQ for docking.

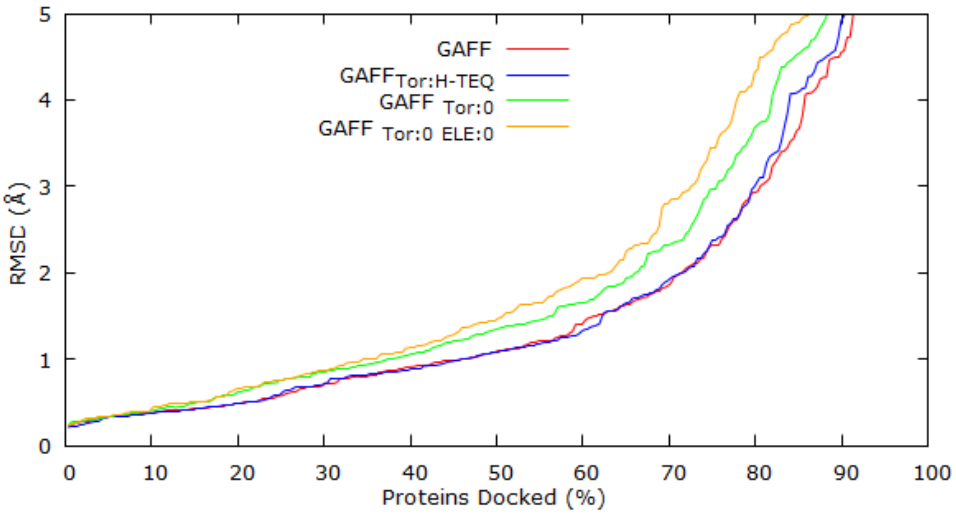


Figure 5.3. The accuracy of protein self-docking on 289 protein-ligand complexes were performed with GAFF (red), $GAFF_{Tor:\ H\text{-}TEQ}$ (blue), $GAFF_{Tor:0}$ (green), and $GAFF_{Tor:0\ ELE:0}$ (orange) shown.

In addition, the importance of electrostatic interactions was also investigated. All electrostatic interactions within the atoms of the ligand and between the protein and ligand was removed during docking of all 289 complexes. This is henceforth referred to as $GAFF_{Tor:0;\ Ele:0}$. An accuracy of 63.0% was obtained by this method ($GAFF_{Tor:0\ Ele:0}$), which was a reduction of 3% compared to $GAFF_{Tor:0}$. This meant that electrostatic interactions do play a role, albeit a small one, in the docking of protein-ligand complexes. The other interactions, such as hydrogen bonding, and especially van der Waals interactions, are believed to be the most important factors for during docking and for receptor-ligand complementarity. These findings suggest that current electrostatic interactions and torsional parameters each contribute approximately 5% and 3%, respectively, to the accuracy of protein docking.

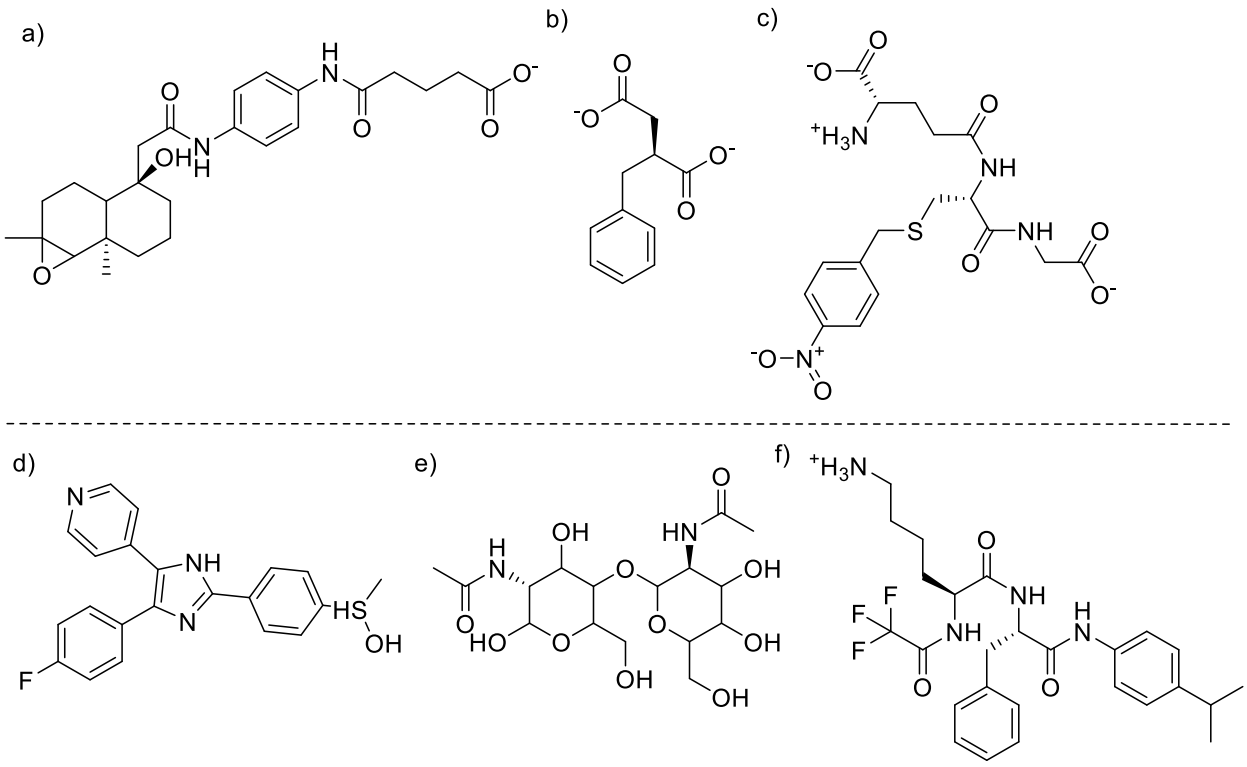


Figure 5.4. Examples of protein binders which were more accurately docked using: GAFF and $GAFF_{Tor:H\text{-}TEQ}$ (a-c) and $GAFF_{Tor:0}$ (d-f).

Nucleic Acid Docking. In contrast to protein-ligand complexes, self-docking on a curated set of 178 nucleic acid-ligand complexes showed that torsions did not affect its accuracy. In fact, nearly identical accuracies of 53.4%, 55.6%, and 55.6% were observed at 2 Å for GAFF, $GAFF_{Tor:H\text{-}TEQ}$, and $GAFF_{Tor:0}$, respectively (Figure 5.5). Nucleic acids possess polyanionic phosphate charges, and Mg^{2+} and K^{+} counter-cations. Consequently, more dominant electrostatic interactions are present in these macromolecules, which are lacking in proteins. In addition, the former also possesses more numerous hydrogen bond acceptors and donors compared with the latter, which are essential for many ligands binding. It was found that in nucleic acid docking, the combination of hydrogen bonding and van der Waals interactions adequately described most systems, without the need for torsion parameters nor electrostatic interactions. In fact, the 95.5% of all structures remained unchanged in accuracy between $GAFF_{Tor:H\text{-}TEQ}$ and $GAFF_{Tor:0}$.

To understand the importance of electrostatic interactions in nucleic acid docking, $GAFF_{Tor:0;\ Ele:0}$ was also used. Interestingly, setting all electrostatic interactions to zero resulted in

only a slight decrease in the accuracy of nucleic acid docking (53.9%). This meant that current electrostatic charging schemes provide no benefit to the docking accuracy. Due to the complexity of charge distribution and polarity of nucleic acids described earlier, a different charging method may be needed to describe the partial charges of ligands bond to nucleic acids. For example, our previous findings indicate that the phosphate charges had to be neutralized to implicitly account for the presence of Mg^{2+} cations.

Another possible reason that electrostatic interactions may not play a significant role in FITTED might be due to the use of the LJ 12-10 potential for hydrogen bonding. Hydrogen bonding is expected to play a large role for nucleic acid docking, as they improve the enthalpic stability of binding.

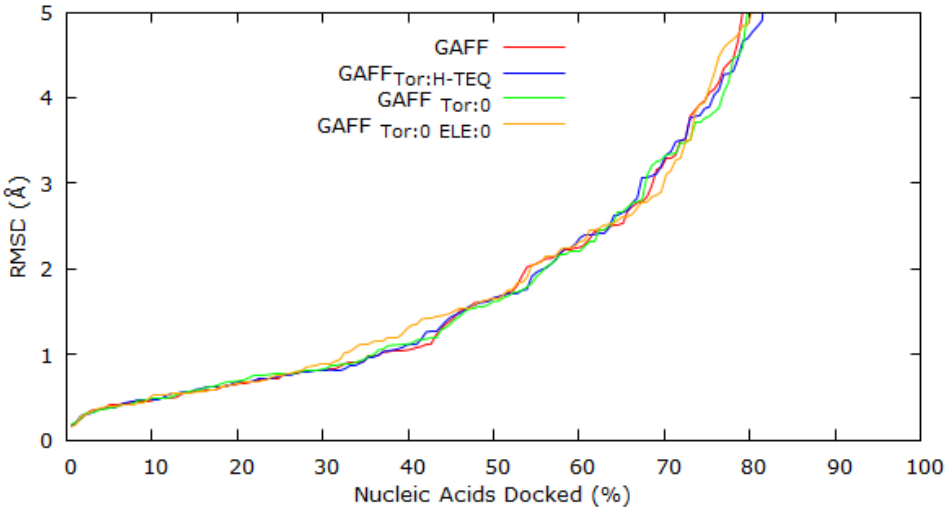


Figure 5.5. The accuracy of nucleic acid self-docking on 178 nucleic acid-ligand complexes were performed with GAFF (red), $GAFF_{Tor:\ H\text{-}TEQ}$ (blue), $GAFF_{Tor:0}$ (green), and $GAFF_{Tor:0\ Ele:0}$ (orange) shown.

Conformational Energy Analysis. To understand why $GAFF_{Tor:H\text{-}TEQ}$ did not outperform the original GAFF FF during docking, conformational analysis in gas-phase was performed. This analysis revealed that the original MMFF94, GAFF2, and MAB FFs achieved average accuracies of 2.11, 3.74, and 3.80 kcal·mol^{-1} in terms of mean absolute errors (MAE), respectively, compared to the reference QM energies. The superior performance of MMFF94 over both GAFF2 and MAB is explained in part by its use of a phased LJ 14-7 potential for van der Waals interactions. The phased nature of this potential, to the right, is more forgiving at close distances. In contrast, GAFF2

and MAB FFs both employed a LJ 12-6 term, which overpredicted these interactions at close distances. For example, the relative energies of the pharmaceutical, fluconazole, was dramatically overpredicted by GAFF2 and MAB, when the two rings were in proximity (Figure 5.6). This resulted in MAEs of 66.2 and 66.0, respectively. For this molecule, neither QM nor MMFF94 registered a steric clash.

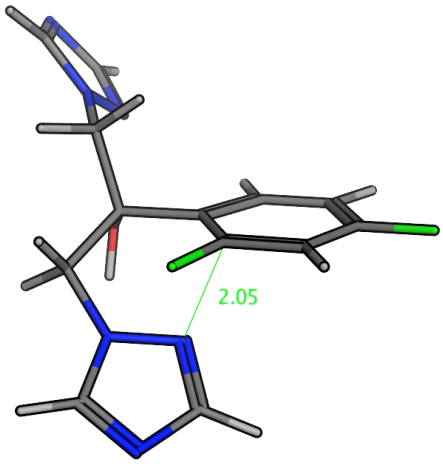


Figure 5.6. A conformation of fluconazole, which registered a steric clash using GAFF2 LJ 12-6 potential, but not QM and MMFF94 phased LJ 14-7 potential.

Unexpectedly, GAFF2$_{\text{Tor: H-TEQ}}$ and GAFF2 had similar accuracies, resulting in a MAE of 4.36 and 4.32, respectively. This was surprising as previous benchmarks of H-TEQ showed that it possessed higher accuracies than both MMFF94 and GAFF2. In particular, the accuracy of H-TEQ 4.0 and 4.5 for conjugated torsions was significantly better than either FFs. In addition, the accuracy of H-TEQ 1.0, 2.0, and 3.0 were similar or slightly better than that of MMFF94.

Their similar accuracies are most likely caused by the large number of nonbonded interactions, especially van der Waals interactions, at close distances. In fact, GAFF, GAFF$_{\text{Tor:H-TEQ}}$ and MAB share identical van der Waals functional form and parametrization, which is known to be very steep at close distances. A large correlation between the MAE of these previously mentioned FFs was found to that of GAFF2 VDW energies, alone (Table 5.1 and Figure 5.7). The correlation of GAFF2 VDW and MMFF94 was significantly lower. As mentioned previously MMFF94 uses a phased LJ 14-7 potential and different parametrization, which achieved greater accuracy.

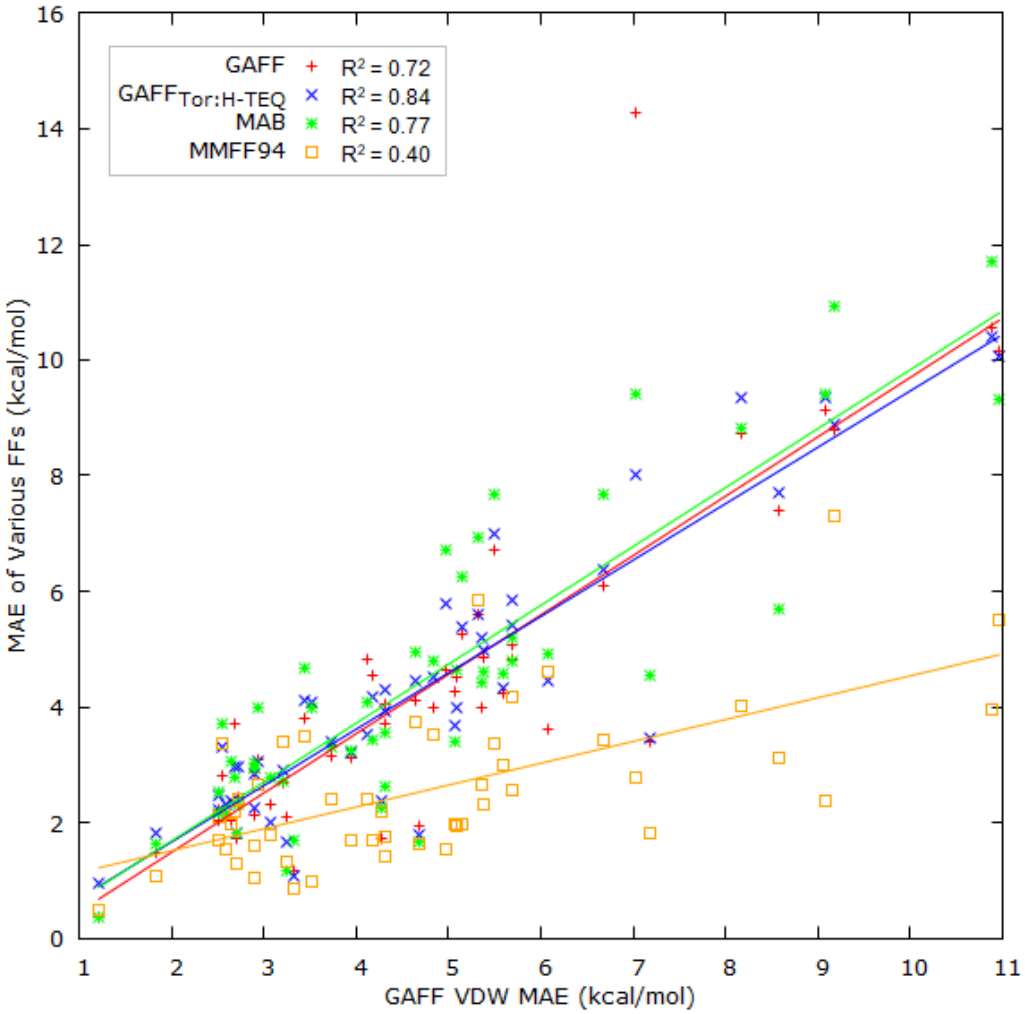


Figure 5.7. Correlation of MAE between GAFF2 VDW and various other methods. The squared correlation coefficient (R^2) are also shown.

As the tested pharmaceutical compounds are large, many more nonbonded interaction pairs were present compared to previous validations.[17, 18] It is also possible that electrostatic interactions of these methods were responsible for producing high MAEs. While restrained electrostatic potential (RESP) and AM1-BCC charges attempt to reproduce electrostatic interactions at larger distances, it is known that they are not well-suited to describe these nuance interactions at close distances. At close distances, quantum effects, such as polarization, multipole expansion, and charge-transfer effects predominate, which is difficult to capture using a simple coulombic potential. This is especially true in 1-4 and 1-5 electrostatic interactions, where there exists a component of a bonded interaction, in addition to electrostatic interactions. Traditionally, FFs did

not make a clear distinction between interaction energies belonging to torsional and electrostatic interactions. Consequently, torsional energy barriers of larger pharmaceuticals may have been more prone to contamination from several inaccurate electrostatic interactions. For smaller molecules, GAFF2$_{\text{Tor: H-TEQ}}$ was previously found to have much closer agreement with QM calculations, and superior accuracy to GAFF2. Consequently, FF development should first focus on devising a robust charge method and more accurate van der Waals functional form and parametrization prior to making any more progress in torsional parameters. These include the special treatment and a method to accurately reproduce 1-4, 1-5, and other close-range interactions.

It should be mentioned that we opted not to use H-TEQ torsion parameters with the MMFF94 FF, due to the self-consistency of that FF. More specifically, the torsion parameters of MMFF94 has previously been proposed to work only with its nonbonded interactions (ie. phased 14-7 LJ and MMFF94 charges). It also has more complex cross-terms, which are not found in GAFF2 and MAB.

Table 5.1. MAE (in kcal·mol^{-1}) of the energies of 50 common pharmaceuticals calculated using various MM methods compared to the reference QM. Only drugs without extensive steric clashes were included. Bold indicates most accurate MM method.

Pharmaceutical	GAFF2	GAFF2$_{\text{Tor:H-TEQ}}$	MAB	MMFF94	GAFF2$_{\text{VDW}}$
acepromazine	4.81	5.42	4.79	2.57	5.69
acetophenazine	3.97	4.50	4.79	3.53	4.84
α amyl cinnamic aldehyde	4.82	3.51	4.07	2.40	4.13
bexarotene	1.72	2.37	2.25	2.18	4.28
bifonazole	2.02	2.23	2.17	2.09	2.52
bupropion	5.27	5.37	6.25	1.97	5.15
cetirizine	5.59	5.61	6.92	5.86	5.32
cinacalcet	3.16	3.39	3.29	2.41	3.74
clavulanicacid	3.98	4.08	4.00	0.99	3.53
cloperastine	2.02	2.38	3.06	1.98	2.64
dexfenfluramine	2.11	2.24	2.93	1.59	2.91

dextronmphetamine	1.47	1.83	1.62	1.06	1.83
diethylpropion	4.84	4.99	4.62	2.31	5.39
diphenhydramine	2.31	2.01	2.78	1.77	3.08
domperidone	3.12	3.21	3.24	1.68	3.95
droperidol	4.05	4.28	2.61	1.40	4.32
ecabet	4.64	5.79	6.70	1.53	4.97
econazole	8.79	8.89	10.93	7.31	9.17
etoricoxib	3.61	4.44	4.93	4.61	6.07
flunarizine	2.43	2.95	2.31	2.41	2.73
flurbiprofen	2.02	2.31	2.16	1.55	2.58
fluvastatin	14.28	8.00	9.42	2.78	7.02
forasartan	8.72	9.35	8.83	4.01	8.17
haloperidol	4.25	3.67	3.39	1.97	5.07
hydroxyzine	3.10	3.06	3.98	2.64	2.94
ibudilast	3.70	3.92	3.56	1.75	4.32
ibuprofen	2.23	2.47	2.52	1.69	2.52
icosapent	3.71	2.95	2.77	2.20	2.69
irbesartan	6.70	6.98	7.67	3.35	5.49
lactulose	4.00	5.19	4.41	2.66	5.36
luliconazole	5.07	5.86	5.20	4.18	5.70
meclizine	2.69	2.91	2.70	3.39	3.22
mefloquine	3.81	4.12	4.66	3.50	3.44
metyrapone	4.22	4.32	4.57	2.98	5.59
naphazoline	2.88	2.85	3.02	1.04	2.91
nitisinone	10.56	10.41	11.70	3.95	10.88
olopatadine	4.54	4.17	3.44	1.70	4.18
oxamniquine	9.14	9.35	9.41	2.38	9.09

paliperidone	1.93	1.79	1.66	1.63	4.69
pentolinium	1.16	1.08	1.69	0.86	3.33
phendimetrazine	0.36	0.94	0.37	0.48	1.21
pheniramine	2.82	3.30	3.70	3.35	2.56
propiomazine	4.11	4.46	4.95	3.75	4.65
risperidone	1.73	1.80	1.81	1.28	2.71
sevoflurane	6.10	6.37	7.67	3.42	6.68
telbivudine	3.40	3.45	4.56	1.81	7.18
tiagabine	10.16	10.06	9.32	5.51	10.97
tobramycin	7.38	7.70	5.68	3.13	8.58
triprolidine	2.09	1.65	1.17	1.32	3.25
xylometazoline	4.50	3.98	4.63	1.93	5.09
Average	**4.32**	**4.36**	**4.50**	**2.56**	**4.77**

5.5 Conclusion

In conclusion, torsional parameters of FFs do have impact on the accuracy of docking. This was seen in the improved accuracy of GAFF and $GAFF_{Tor:H\text{-}TEQ}$ over that of $GAFF_{Tor:0}$ in protein docking. In nucleic acid docking, all three previously mentioned FFs performed with a similar degree of accuracy. Nucleic acids are especially difficult and tricky for docking, due to the greater number of charged species and counterions present. In fact, the current charging scheme and functional form used by FITTED, electronegativity equalization, was found to be poor for nucleic acids, especially at close distances.[28] In the future, ligand partial charges may need to be modified to implicitly account for polarity of the nucleic acid binding site, which could improve the accuracy of docking. For both protein and nucleic acid docking, electrostatic interactions contributed very little to the overall docking accuracy. The amalgamated term consisting of van der Waals and hydrogen bonding interactions played a greater role in mediating binding of ligands during docking.

Unexpectedly, this study also found that GAFF and $GAFF_{Tor:H\text{-}TEQ}$ had similar accuracies for both protein and nucleic acid docking. Upon performing gas-phase calculations, it was found that both methods performed equally poorly, with MAEs of 3.74 and 3.88, respectively. This was puzzling as H-TEQ torsional parameters, compatible with GAFF, was previously shown to be significantly more accurate. However, this previous validation was performed on smaller druglike molecules with a limited number of torsions. In addition, the experiment also consisted of rotating one torsion at a time. However, upon encountering real drug molecules and generating random conformations, the nonbonded interactions, especially charges, may have contaminated the overall conformational energies, leading to inaccuracies. Through this study, it has been realized that a method to address nonbonded interactions is desperately needed, especially those at close distances. For example, GAFF currently indiscriminately lowers the electrostatic interaction between 1-4 atoms by a constant factor. However, this is erroneous as there is no evidence to suggest that 1-4 electrostatic interactions are dampened by a predetermined constant. Work on a more robust electrostatic charge method is required. In addition, the steric wall of LJ 12-6 is too steep, and a selection of a softer function for future docking would be promising.

Furthermore, docking of face centered and T-shaped π-π stacking is currently achieved by a combination of electrostatic and van der Waals in type I FFs. Consequently, it does not take into consideration the chemistry and its multipole charge distribution, which may be important for ligand binding. In addition, pharmaceuticals often contain halogens, which possess sigma holes. For example, despite being mostly negatively charged, CH_3F molecule contains a sigma hole of positive charge on the fluorine atom, which could be used for binding.

5.6 References

1. Sahu, D.; Bastidas, M.; Lawrence, C. W.; Noid, W. G.; Showalter, S. A. Chapter Two - Assessing Coupled Protein Folding and Binding through Temperature-Dependent Isothermal Titration Calorimetry. In *Methods Enzymol.*, Feig, A. L., Ed.; Academic Press: 2016; Vol. 567, pp 23-45.

2. Corbeil, C. R.; Englebienne, P.; Moitessier, N., Docking Ligands into Flexible and Solvated Macromolecules. 1. Development and Validation of Fitted 1.0. *J. Chem. Inf. Model* **2007**, *47*, 435-449.

3. Corbeil, C. R.; Englebienne, P.; Yannopoulos, C. G.; Chan, L.; Das, S. K.; Bilimoria, D.; L'Heureux, L.; Moitessier, N., Docking Ligands into Flexible and Solvated Macromolecules. 2. Development and Application of Fitted 1.5 to the Virtual Screening of Potential Hcv Polymerase Inhibitors. *J. Chem. Inf. Model* **2008**, *48*, 902-909.

4. Ton, A.-T.; Gentile, F.; Hsing, M.; Ban, F.; Cherkasov, A., Rapid Identification of Potential Inhibitors of Sars-Cov-2 Main Protease by Deep Docking of 1.3 Billion Compounds. *Molecular Informatics n/a*.

5. Allen, W. J.; Balius, T. E.; Mukherjee, S.; Brozell, S. R.; Moustakas, D. T.; Lang, P. T.; Case, D. A.; Kuntz, I. D.; Rizzo, R. C., Dock 6: Impact of New Features and Current Docking Performance. *J. Comput. Chem.* **2015**, *36*, 1132-1156.

6. Durrant, J. D.; McCammon, J. A., Molecular Dynamics Simulations and Drug Discovery. *BMC Biology* **2011**, *9*, 71.

7. Chen, J.; Wang, X.; Pang, L.; Zhang, J. Z. H.; Zhu, T., Effect of Mutations on Binding of Ligands to Guanine Riboswitch Probed by Free Energy Perturbation and Molecular Dynamics Simulations. *Nucleic Acids Res.* **2019**, *47*, 6618-6631.

8. Wallraven, K.; Holmelin, F. L.; Glas, A.; Hennig, S.; Frolov, A. I.; Grossmann, T. N., Adapting Free Energy Perturbation Simulations for Large Macrocyclic Ligands: How to Dissect Contributions from Direct Binding and Free Ligand Flexibility. *Chemical Science* **2020**, *11*, 2269-2276.

9. Woo, H.-J.; Roux, B., Calculation of Absolute Protein–Ligand Binding Free Energy from Computer Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825.

10. 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.

11. D.A. Case, K. B., I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz,R Harris, S. Izadi, K. Kasava- jhala, A. Kovalenko, R. Krasny, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F. Pan, S. Pantano, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, L. Wilson, R.M. Wolf, X. Wu, D.M. York and P.A. Kollman, Amber 2020. *University of California, San Francisco* **2020**.

12. 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.

13. 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.

14. 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. *Journal of the American Chemical Society* **1996**, *118*, 11225-11236.

15. Corbino, K. A.; Barrick, J. E.; Lim, J.; Welz, R.; Tucker, B. J.; Puskarz, I.; Mandal, M.; Rudnick, N. D.; Breaker, R. R., Evidence for a Second Class of S-Adenosylmethionine Riboswitches and Other Regulatory Rna Motifs in Alpha-Proteobacteria. *Genome Biol* **2005**, *6*.

16. 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.

17. Wei, W. L.; Champion, C.; Liu, Z. M.; Barigye, S. J.; Labute, P.; Moitessier, N., Torsional Energy Barriers of Biaryls Could Be Predicted by Electron Richness/Deficiency of Aromatic Rings; Advancement of Molecular Mechanics toward Atom-Type Independence. *J. Chem. Inf. Model* **2019**, *59*, 4764-4777.

18. Wei, W.; Champion, C.; Barigye, S. J.; Liu, Z.; Labute, P.; Moitessier, N., Use of Extended-Hückel Descriptors for Rapid and Accurate Predictions of Conjugated Torsional Energy Barriers. *J. Chem. Inf. Model* **2020**.

19. 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.

20. Bush, B. L.; Bayly, C. I.; Halgren, T. A., Consensus Bond-Charge Increments Fitted to Electrostatic Potential or Field of Many Compounds: Application to Mmff94 Training Set. *J. Comput. Chem.* **1999**, *20*, 1495-1516.

21. Gerber, P. R.; Müller, K., Mab, a Generally Applicable Molecular Force Field for Structure Modelling in Medicinal Chemistry. *J. Comput. Aided Mol. Des.* **1995**, *9*, 251-268.

22. Yang, L.; Sun, L.; Deng, W.-Q., Combination Rules for Morse-Based Van Der Waals Force Fields. *The Journal of Physical Chemistry A* **2018**, *122*, 1672-1677.

23. 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 (Vol 117, Pg 5179, 1995). *Journal of the American Chemical Society* **1996**, *118*, 2309-2309.

24. Kührová, P.; Mlýnský, V.; Zgarbová, M.; Krepl, M.; Bussi, G.; Best, R. B.; Otyepka, M.; Šponer, J.; Banáš, P., Improving the Performance of the Amber Rna Force Field by Tuning the Hydrogen-Bonding Interactions. *J. Chem. Theory Comput.* **2019**, *15*, 3288-3305.

25. 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.

26. Champion, C.; Barigye, S. J.; Wei, W.; Liu, Z.; Labute, P.; Moitessier, N., Atom Type Independent Modeling of the Conformational Energy of Benzylic, Allylic, and Other Bonds Adjacent to Conjugated Systems. *J. Chem. Inf. Model* **2019**.

27. Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W., Diverse, High-Quality Test Set for the Validation of Protein−Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726-741.

28. Mortier, W. J.; Ghosh, S. K.; Shankar, S., Electronegativity-Equalization Method for the Calculation of Atomic Charges in Molecules. *Journal of the American Chemical Society* **1986**, *108*, 4315-4320.

6 Conclusion and Closing Remarks

6.1 Conclusion

The field of computational chemistry has grown considerably in the past few years and decades. Despite these progresses, computational chemistry is still growing rapidly as it is a relatively young field with promises of new discoveries and developments. It has an exciting future as further improvements in hardware would allow the use of more advanced and accurate simulations, which were previously intractable. The work presented in this thesis has focused primarily on improving the accuracy of *in silico* docking and MD simulations for the purposes of discovering potentially new drug candidates and their stabilities. As such, these computational methods had to be inexpensive. For this reason, MM was the method of choice rather than more expensive QM calculations. To this end, new methodologies docking and FFs were envisioned and developed.

In chapter 1, an overview of MM FFs for simulating nucleic acids was discussed. Briefly, it summarized structural and biological roles of nucleic acids. Historic developments and current state-of-the-art MM methods to treat and describe them were also discussed. In addition, it listed several hallmark computational studies, which were carried out by others in the past. The role of computational chemistry in SBDD was highlighted, with examples of the different binding modes of existing pharmaceuticals with nucleic acids and promise of the field. As common, type I MM FFs, used for docking and MDs, are simplistic in nature, future developments to improve the accuracy of MM FFs were suggested, including the treatment of hydrogen bonding, π-π stacking, and noncanonical nucleotides.

In chapter 2, key problems with regards to docking of small molecules to nucleic acids was identified. This was attributed to a lack of dependable water molecules in the binding site of nucleic acids during docking as each ligand was expected and shown to possess different water network within the same receptor macromolecule. In fact, this work builds upon the cross-docking work, started by Moitessier et al (2006), where docking accuracy to aminoglycosides was impacted significantly by the different water networks in the ligand binding site.[1] By considering a collection of water positions, they were able to improve docking to these molecules. However, at the time, water positions were manually performed. At the start of the project, no water placement protocol specific for nucleic acid-ligand complexes was available. Consequently, SPLASH'EM was

developed.[2] It used statistics collected from thousands of nucleic acid structures, containing water molecules to identify regions of high-water occupancy. In conjunction with a specialized FF, developed specifically for the placement of water molecules, it achieved the highest accuracy known to-date. In the future, this water placement technique should be incorporated during docking, which could improve its overall accuracy. This would be useful vHTS of potential nucleic acid binders, along with a development of an appropriate scoring function. Subsequently, organic synthesis and biological testing could be carried out on promising nucleic acid binders.

In chapter 3 and 4, the underlying MM FFs were scrutinized to see if improvements could be made. This work was built on previous work conducted by Moitessier and coworkers.[3-5] Torsions of molecules were a combination of steric and hyperconjugation effects. While steric effects were already modelled in FFs, hyperconjugation was not. By extracting the hyperconjugation and conjugation terms of torsions, we proved that torsional energy barriers of small molecules could be quantified and more accurately predicted. Due to the difficulty of this strategy, torsions were classified into different chemical groups. In this thesis, conjugated torsions were quantified.

Torsional energy barriers of biaryl and other conjugated molecules were analyzed. Existing torsional energy barriers for these molecules were found to be poor in popular MM FFs, including GAFF2, MAB, and MMFF94. This could be attributed to the use of atom typing in GAFF2 and MMFF94, which has its limitations in terms of transferability and assignment of torsion parameters. To circumvent this problem, we developed H-TEQ 4.0 for biaryl molecules (chapter 3), using simple molecular descriptors (ie. electron-richness/deficiency).[6] This method achieved an RMSE of 0.95 kcal/mol, which significantly outperformed GAFF2 on our compiled validation set, containing 100 small molecules. Our efforts to first focus on biaryls was attributed to its frequent appearances in pharmaceuticals and bioactive compounds. Subsequently, we focused on all conjugated molecules (chapter 4) and used bond orders and π-charges obtained from extended-Hückel descriptors to predict their torsional energy barriers. To our delight, this general method, H-TEQ 4.5, worked well for all conjugated torsions, including biaryls.[7] Again, testing on the validation set of 200 molecules, we obtained more accurate results (1.01 kcal·mol^{-1}) than other popular FFs, including GAFF2, MMFF94, and MAB (3.49, 1.50, and 1.77 kcal·mol^{-1}, respectively).

In chapter 5, we tested the ability of H-TEQ to improve the accuracy of protein and nucleic acid docking. In total, approximately 450 dockings were performed. Unfortunately, H-TEQ had similar performance to GAFF. This was unexpected as previous results indicated that H-TEQ had superior performance compared to GAFF2. Upon further analysis and inspection, it was hypothesized that nonbonded interactions were contaminating the conformational energies of these molecules. Consequently, gas-phase energies of 50 frequently used pharmaceuticals were compared using MMFF94, GAFF2, MAB, and $GAFF2_{Tor:H\text{-}TEQ}$. From these results, it was realized that gas-phase energies of the GAFF2, MAB, and $GAFF2_{Tor:H\text{-}TEQ}$ were similar owing to using the same LJ 12-6 potential for van der Waals interactions.[8, 9] MMFF94 was distinct amongst these as it used a phased LJ 14-7, which was much more accurate.[10] In addition, electrostatic potential at close distances was also identified as being potentially poor, as it currently uses a constant scaling factor for 1-4 interactions.[11] In addition, 1-5 interactions are also expected to behave differently than longer range interactions. This realization with regards to the low quality of the nonbonded interactions, especially at close distances, has given ideas for future projects. These include focusing on addressing the proximal nonbonded interactions, first, prior to further developing additional torsion models. Once a robust nonbonded scheme is ready, previous H-TEQ developments could be reoptimized.

Traditional use of a limited set of torsion parameters cannot be used to cover the entire chemical space during vHTS, due to its shear size. As a result, conventional FFs for small molecules have often emphasized specific molecular groups within the chemical space frequently encountered by existing pharmaceuticals. Since many have suggested that the chemical space of existing drug molecules do not evenly represent all the possible small molecule pharmaceuticals, the use of atom types, designed for existing drugs may hinder the discovery of novel drugs. Using an atom type independent method, such as H-TEQ would allow drug discovery scientists to break free from this cycle.

6.2 Closing Remarks

As alluded to previously, computational chemistry is a relatively young field. The use of calculations has complemented experimental studies, where performing the latter would have been

prohibitively expensive or sometimes impossible. It would be quite exciting, for example, to simulate, using MD simulations, an entire eukaryotic cell, or a collection of cells. If such a feat could be achieved, this would lead to a scientific revolution in cellular engineering and medical sciences. It would offer limitless possibilities, in terms of applications and benefits, such as exploring the effects of therapeutics from a cellular and wholistic perspective. Such simulations would allow the interactions of pharmaceuticals with cellular targets to be identified, such as off-target effects, degradation sites, and other interactions. The field of computational chemistry has a bright future, and would no doubt continue to contribute to scientific knowledge for the foreseeable future.

6.3 References

1. Moitessier, N.; Westhof, E.; Hanessian, S., Docking of Aminoglycosides to Hydrated and Flexible Rna. *J. Med. Chem.* **2006**, *49*, 1023-1033.
2. Wei, W.; Luo, J.; Waldispühl, J.; Moitessier, N., Predicting Positions of Bridging Water Molecules in Nucleic Acid–Ligand Complexes. *J. Chem. Inf. Model* **2019**, *59*, 2941-2951.
3. Liu, Z.; 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.
4. 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.
5. Champion, C.; Barigye, S. J.; Wei, W.; Liu, Z.; Labute, P.; Moitessier, N., Atom Type Independent Modeling of the Conformational Energy of Benzylic, Allylic, and Other Bonds Adjacent to Conjugated Systems. *J. Chem. Inf. Model* **2019**.
6. Wei, W. L.; Champion, C.; Liu, Z. M.; Barigye, S. J.; Labute, P.; Moitessier, N., Torsional Energy Barriers of Biaryls Could Be Predicted by Electron Richness/Deficiency of Aromatic Rings; Advancement of Molecular Mechanics toward Atom-Type Independence. *J. Chem. Inf. Model* **2019**, *59*, 4764-4777.

7. Wei, W.; Champion, C.; Barigye, S. J.; Liu, Z.; Labute, P.; Moitessier, N., Use of Extended-Hückel Descriptors for Rapid and Accurate Predictions of Conjugated Torsional Energy Barriers. *J. Chem. Inf. Model* **2020**.

8. 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.

9. Gerber, P. R.; Muller, K., Mab, a Generally Applicable Molecular-Force Field for Structure Modeling in Medicinal Chemistry. *J. Comput. Aided Mol. Des.* **1995**, *9*, 251-268.

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

11. 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 (Vol 117, Pg 5179, 1995). *Journal of the American Chemical Society* **1996**, *118*, 2309-2309.

Appendix A:
Supplementary Information for "Predicting Positions of Bridging Water Molecules in Nucleic Acid-Ligand Complexes"

A1 From Statistics to Binding Free Energy.

As water molecules are often hydrogen bonded to more than one residue, estimating the free energy of water molecule at each position was achieved through the combination of the free energy distributions from each nearby polar atom. For example, the bridging water molecule, depicted in Figure 2.1, is influenced by the hydrogen bonding of several polar atoms, including two Lividomycin ammonium groups, guanine-O6, guanine-N2, and cytosine–O2. Consequently, the free energy values of water distribution from each individual atom were added, where they overlapped in space, to derive an overall free energy distribution for the entire NA-ligand complex. Although entropy is not additive, the sum of statistical information was used as an estimate for the binding free energy of water molecules.

It should be noted that statistical data of water distribution for ligands were not available, and thus, the most structurally similar NA functional group statistics was used in its place with some modifications pertaining to symmetry. For example, the statistics for deoxyribose-O4 of NAs was first symmetrized along the in-plane surface and used in place of ether oxygens of ligands. This methodology was used to approximate the free energy of different ligand polar atoms.

A2 Hydrogen Bonding Potentials of Various Polar Atoms with Water

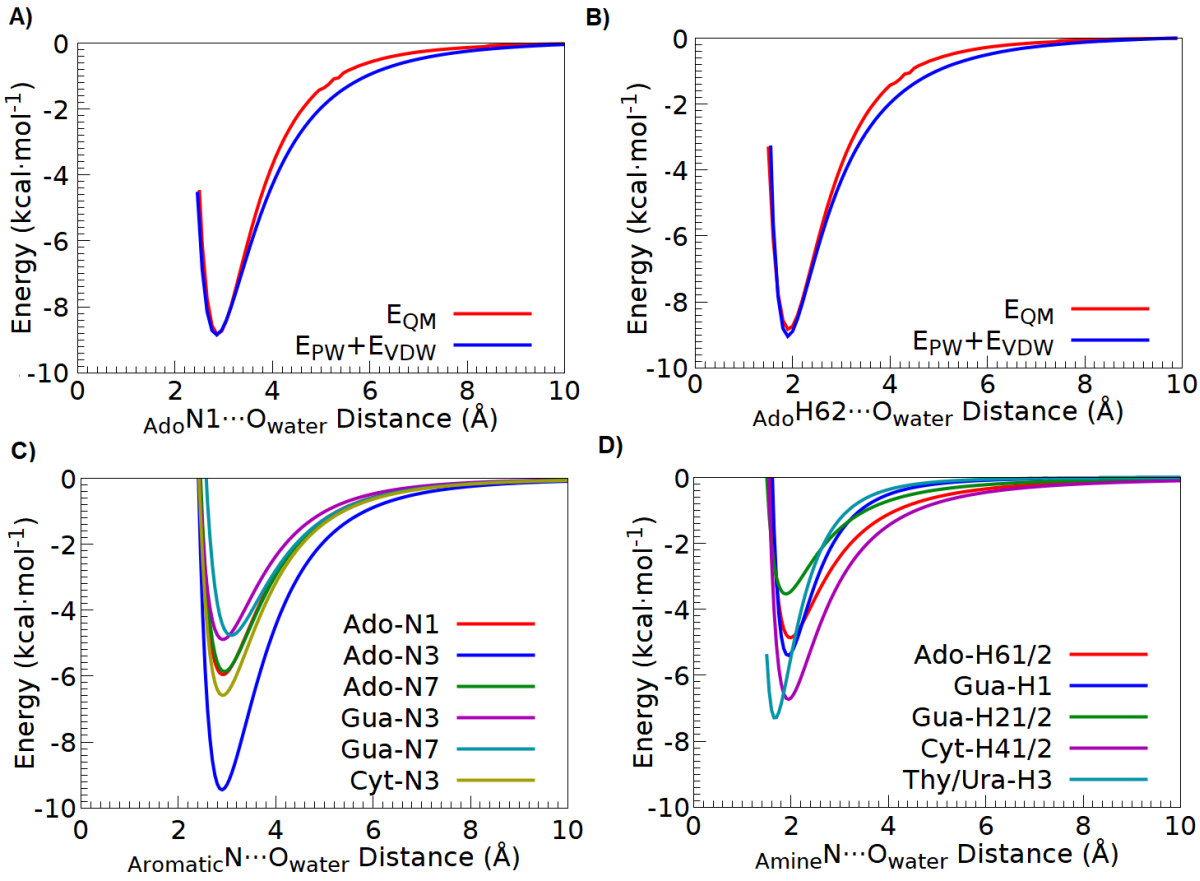


Figure A1. Hydrogen bonding potential in QM (blue) as compared to the developed MM FF (red) for a linear distance scan along the angle bisector of Ado-N1(A) and Ado-H62 (B). The MM potential is sum of the 6-5-3 Lennard-Jones hydrogen bonding potential and van der Waals 12-6 Lennard-Jones potential. Functional form of the developed 6-5-3 Lennard-Jones potential for sp^2 aromatic nitrogen (C) and amine hydrogen (D) used to describe hydrogen bonding potential between polar atoms and water.

A3 Obtaining an Angle Term

In order to accurately model the relationship between enthalpy and angles, explicit electron lone pairs (lp) were needed. These electron lp were at angles corresponding to or near the minimum(s) on the PES, of these simplified NA subunits, in accordance with the VSEPR theory. For sp^2

oxygen, the lp electrons were placed in the plane of the base, at 70° on either side of the bisector. However, there were two additional smaller minima at 70° on both PES, directly out-of-the-plane. Consequently, dummy lp were also placed at these positions, which were assigned weaker potentials, in order to better reproduce the QM PES. In the case of sp^3 oxygens, the lp's were placed at 70° on both sides of the bisector in the directions directly out-of-the plane. For anionic oxygens, three lp's were placed at 109.5° from the P-O bond, in a tetrahedral manner that would maximize the angle between the bonded atom and the three O-lp angles. For sp^2 nitrogens, the lp was placed in the plane of the two bonded atoms, at the bisector.

Due to the differences in chemical nature of different atoms, slightly different potentials were selected for different atom types. A normal distribution-like potential was employed at the location of the lp electrons, which allowed a better reproduction of the PES of simplified NA subunits (Equation. 2.3). The angular contribution function has the advantageous property that it is always bound between 0 and 1, and is a continuous function, allowing optimizations to be performed easily. In order to prevent a favorable steric clash between a PW and a polar atom from occurring (ie. at close distances but large angular deviations), cutoff distances of 2.5 Å and 1.5 Å were implemented for hydrogen bond acceptors and donors, respectively (Equation 2.3). Any distances closer than these threshold values between the polar atom and nucleic acid would result in an angular contribution of 1, restoring the full steric clash potential. As an example, the devised potential was able to reproduce the QM energy profile for pyridine (Figure 2.7B) and 1-methylpyridin-4(1H)-one (Figure 2.7C). It should be noted that where the in-plane and out-of-plane angular energy profiles differed, the in-plane angular PES was selected as water molecules are less likely to occupy a position intercalated between two nucleotide bases. The angle contribution function of simplified subunits was applied to chemically similar NA and ligand polar atoms.

A4 Removal of Solvent Exposed Waters

NAs have extremely shallow binding pockets, compared to proteins. Consequently, there are many water molecules near its surface, which are affected by bulk waters and/or adjacent crystallographic subunits. Additionally, these waters do not mediate direct contact between the

nucleic acid and ligand and are often unimportant for docking. A routine was developed to identify and remove those.

In order to remove these water molecules during water placement, the exterior surfaces of nucleic acids were mapped by rolling a spherical probe of different sizes along its surface (Figure A2). All placed waters within the regions in contact with the probe were considered bulk water, and thus removed. The optimal probe radius of 10 Å was selected by testing various sizes, which removed approximately a quarter of originally placed water molecules.

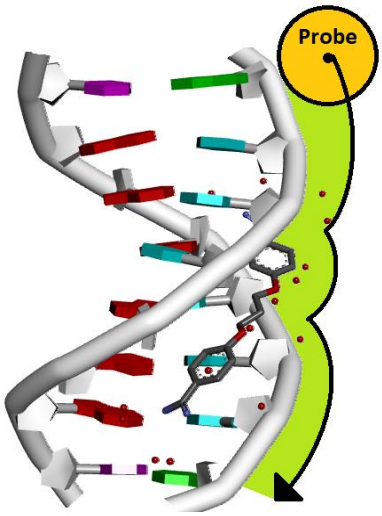


Figure A2. Mapping of bulk waters by rolling a spherical probe along the surface of the nucleic acid. Placed waters in the light green region were determined to be bulk solvents, and consequently removed.

A5 Validation Dataset for Water Placement

The compiled dataset used in this study for water placement is composed of 91 PDBs (listed below) selected using the workflow in Figure A3.

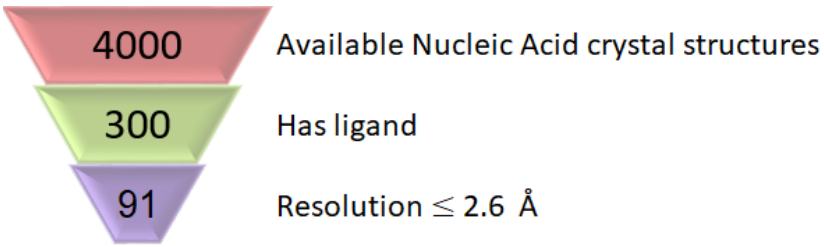


Figure A3. Testing set for water placement.

102D, 109D, 166D, 195D, 1D30, 1D43, 1D45, 1D64, 1FUF, 1J7T, 1JTL, 1L1H, 1LEX, 1M6F, 1NAB, 1O0K, 1PRP, 1VZK, 1Z8V, 298D, 2B0K, 2B3E, 2B57, 2BEE, 2EES, 2EET, 2EEU, 2EEV, 2ET8, 2FCX, 2FD0, 2G9C, 2GVR, 2I2I, 2I5A, 2NLM, 2OE5, 2OE8, 2PWT, 2XNW, 2XNZ, 2XO0, 2XO1, 2Z74, 2Z75, 302D, 303D, 311D, 358D, 360D, 3AJK, 3C5D, 3CDM, 3D2X, 3DIL, 3DS7, 3ERU, 3EUM, 3FO4, 3FO6, 3G4M, 3GAO, 3GES, 3GOT, 3LA5, 3OIE, 3Q3Z, 3S4P, 3SD3, 3SKI, 3T5E, 3UYH, 403D, 432D, 443D, 453D, 473D, 4AGZ, 4FE5, 4FEP, 4FXM, 4LVX, 4LVY, 4LW0, 4O5X, 4TZX, 5D99, 5FJC, 5NDH, 5O69, 6BNA

A6 Initial Placement of Water Molecules

The initial placement of water molecules was scored in order to ensure that there was a sufficient coverage of water molecules at the beginning of the placement protocol. To this end, the first scoring criteria was used (see methods). It was found that placing waters in the top 5, 10, 15, and 50 most statistically populated positions for each nucleic acid and ligand polar atom, allowed a population of waters with an overall coverage of 70%, 80%, 88%, and 99%, respectively (Figure A4). Consequently, 50 voxels were chosen due to its excellent initial water population and coverage. As a result, the accuracy of our method will be primarily relying on the accuracy of the scoring rather than initial placement.

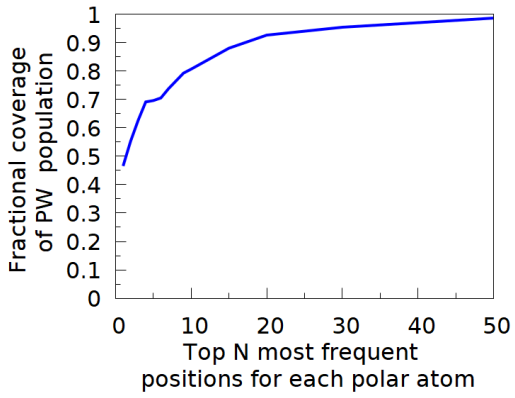


Figure A4. The top *N* most frequently populated position(s) around each individual polar atom of nucleic acid and ligand was used to place waters. This initial population of particle waters contained those which correctly predicted the crystallographic positions and also those which incorrectly predicted their positions. The coverage refers to the number of correctly predicted

crystallographic waters by each population, irrespective of the number of incorrectly predicted water molecules.

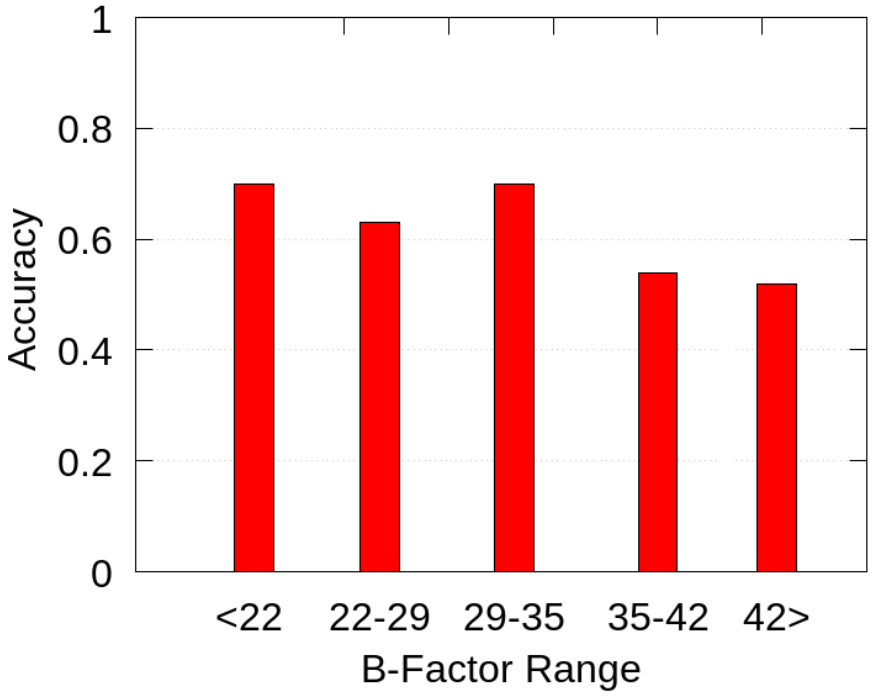


Figure A5. The accuracy of SPLASH'EM water placement method as decomposed into five different bins of equal sizes (46 PDB waters).

Appendix B:

Supplementary Information for "Torsional Energy Barriers of Biaryls could be Predicted by Electron-richness/deficiency of Aromatic Rings; Advancement of Molecular Mechanics toward Atom-Type Independence"

B1 Overall π-electronegativity Modulates Strength of Conjugation

The slopes of the linear fit across each group varied slightly, with C-C, C-N, N-N, C-N+, N-N+, and N+-N+ central atoms having slopes of -7.58, -13.37, -5.82, -10.82, -11.07, and -4.42, respectively. More noticeably, the y-intercept of these molecules differed as well with values of 13.69, 22.38, 15.51, 24.66, 25.46, and 12.06, respectively.

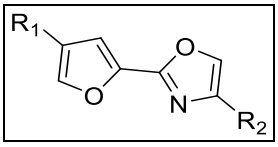
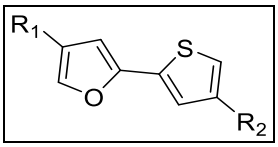
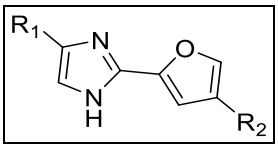
B2 Strength of Conjugation

For C-C, C-N, and N-N central bonds, the slopes of -9.30, -24.32, and -63.58 were found, respectively. Similarly, the y-intercept of these central bonds differed with values of -3.87, -1.44, and 2.61, respectively.

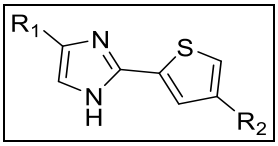
For biaryls containing C-N^+ central bonds, an increase in $(\Delta\chi_\pi)^2$ decreased V_2 (Figure 3.9B), for those comprised of a 6-membered aromatic moiety with carbon central atom and a 5-membered aromatic moiety with a N^+ central atom (ie. C(5)-N^+(6)). The slope and y-intercept corresponding to this was -22.11 and -1.74, respectively ($R^2 = 0.69$). All other biaryls with C-N^+ central bonds saw an increase in V_2 as $(\Delta\chi_\pi)^2$ increased. More specifically, C(5)-N^+(5), C(6)-N^+(5), and C(6)-N^+(6) had slopes of 7.60, 3.09, 40.25, y-intercepts of -0.99, -3.89, and -7.23, and an R^2 of 0.31, 0.87, and 0.92, respectively.

Similarly, an increase in $(\Delta\chi_\pi)^2$ increased in V_2 for all N-N^+ central bond biaryls, regardless of the number of atoms in each ring. However, the trends for both N(5)-N^+(5) differed from N(5)-N^+(6) such that the slope and y-intercept of the former were 13.97 and -1.09, respectively ($R^2 = 0.58$) compared to the latter with 12.31 and -6.71 ($R^2 = 0.94$), respectively.

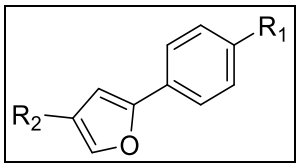
B3 Substituent Effects on Torsional Energy Barriers of Biaryls

R_1	R_2	Barrier Height (kcal/mol)	Difference from unsubstituted (kcal/mol)	Difference from unsubstituted
				
H	H	4.44	-	
CF_3	CF_3	4.51	0.07	2%
H	CF_3	4.86	0.43	10%
NMe_2	CF_3	5.08	0.64	14%
CF_3	H	4.23	-0.20	-5%
NMe_2	H	4.75	0.31	7%
CF_3	NMe_2	4.60	0.17	4%
H	NMe_2	5.04	0.61	14%
NMe_2	NMe_2	5.39	0.96	22%
				
H	H	2.29	-	
H	CF_3	2.08	-0.21	-9%
CF_3	CF_3	2.09	-0.20	-9%
NMe_2	CF_3	2.32	0.03	2%
CF_3	H	2.24	-0.05	-2%
NMe_2	H	2.53	0.24	11%
CF_3	NMe_2	2.35	0.06	3%
H	NMe_2	2.45	0.16	7%
NMe_2	NMe_2	3.14	0.85	37%
				
H	H	4.67	-	
H	CF_3	4.67	0.00	0%
CF_3	CF_3	5.11	0.44	9%
NMe_2	CF_3	5.17	0.49	11%
CF_3	H	5.33	0.65	14%
NMe_2	H	5.45	0.77	17%

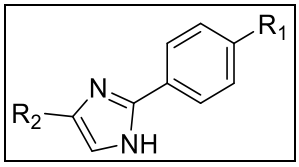
NMe_2	NMe_2	4.26	-0.41	-9%
H	NMe_2	5.23	0.56	12%
CF_3	NMe_2	5.95	1.27	27%



H	H	2.61	-	
H	CF_3	2.71	0.11	4%
NMe_2	CF_3	2.77	0.16	6%
CF_3	H	2.78	0.17	7%
CF_3	CF_3	2.88	0.28	11%
NMe_2	H	2.93	0.32	12%
CF_3	NMe_2	3.11	0.50	19%
H	NMe_2	3.17	0.56	21%
NMe_2	NMe_2	3.39	0.78	30%

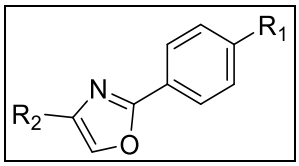


H	H	4.40	-	
NMe_2	CF_3	3.92	-0.48	-11%
H	CF_3	4.31	-0.10	-2%
CF_3	CF_3	4.57	0.17	4%
NMe_2	H	3.86	-0.54	-12%
CF_3	H	4.88	0.48	11%
NMe_2	NMe_2	4.31	-0.09	-2%
H	NMe_2	4.78	0.38	9%
CF_3	NMe_2	5.04	0.64	15%

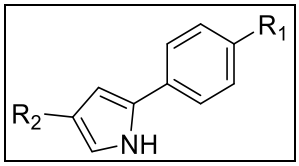


H	H	5.45	-	
NMe_2	H	4.42	-1.03	-19%
NMe_2	NMe_2	5.37	-0.08	-1%
NMe_2	CF_3	5.47	0.02	0%
H	CF_3	5.48	0.03	1%
CF_3	CF_3	5.57	0.12	2%

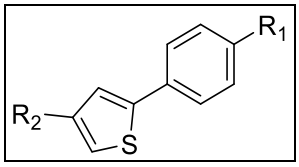
CF_3	H	5.87	0.42	8%
CF_3	NMe_2	6.04	0.59	11%



H	H	4.07	-	
NMe_2	H	4.03	-0.04	-1%
NMe_2	CF_3	4.12	0.05	1%
NMe_2	NMe_2	4.13	0.06	1%
H	NMe_2	4.18	0.11	3%
H	CF_3	4.21	0.14	4%
CF_3	H	4.30	0.23	6%
CF_3	CF_3	4.44	0.37	9%
CF_3	NMe_2	4.57	0.50	12%

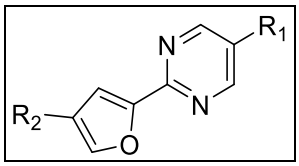


H	H	4.50	-	
NMe_2	H	4.02	-0.48	-11%
H	CF_3	4.05	-0.45	-10%
NMe_2	CF_3	4.13	-0.37	-8%
CF_3	CF_3	4.49	-0.01	0%
CF_3	H	5.10	0.61	14%
NMe_2	NMe_2	5.14	0.64	14%
H	NMe_2	5.33	0.84	19%
CF_3	NMe_2	5.80	1.31	29%

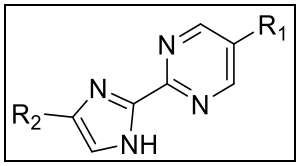


H	H	3.66	-	
NMe_2	H	3.54	-0.12	-3%
H	CF_3	3.60	-0.06	-2%
NMe_2	CF_3	3.62	-0.04	-1%
CF_3	CF_3	3.76	0.10	3%
NMe_2	NMe_2	3.92	0.26	7%
H	NMe_2	3.98	0.31	9%

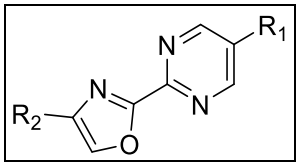
CF_3	H	4.10	0.44	12%
CF_3	NMe_2	4.30	0.64	17%



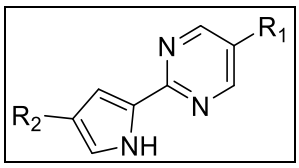
H	H	5.89	-	
NMe_2	CF_3	4.90	-0.99	-17%
H	CF_3	5.38	-0.51	-9%
NMe_2	H	5.52	-0.38	-6%
CF_3	H	6.90	1.01	17%
NMe_2	NMe_2	5.78	-0.12	-2%
H	NMe_2	6.25	0.36	6%
CF_3	NMe_2	7.16	1.27	22%



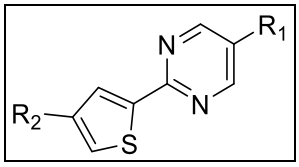
H	H	8.46	-	
H	CF_3	8.19	-0.27	-3%
NMe_2	CF_3	8.20	-0.25	-3%
NMe_2	H	8.24	-0.22	-3%
CF_3	H	9.21	0.76	9%
NMe_2	NMe_2	8.62	0.16	2%
H	NMe_2	9.01	0.56	7%
CF_3	NMe_2	9.90	1.45	17%



H	H	5.96	-	
NMe_2	CF_3	5.88	-0.08	-1%
H	CF_3	6.01	0.05	1%
CF_3	CF_3	6.46	0.50	8%
NMe_2	H	5.73	-0.22	-4%
CF_3	H	6.60	0.64	11%
NMe_2	NMe_2	6.16	0.20	3%
H	NMe_2	6.32	0.36	6%
CF_3	NMe_2	6.93	0.98	16%



H	H	9.18	-	
NMe_2	CF_3	7.87	-1.31	-14%
H	CF_3	8.34	-0.84	-9%
CF_3	CF_3	9.10	-0.08	-1%
NMe_2	H	8.52	-0.65	-7%
CF_3	H	10.38	1.21	13%
NMe_2	NMe_2	8.71	-0.47	-5%
H	NMe_2	9.45	0.27	3%
CF_3	NMe_2	10.65	1.48	16%



H	H	6.94	-	
NMe_2	CF_3	5.88	-1.06	-15%
H	CF_3	6.18	-0.76	-11%
CF_3	CF_3	6.62	-0.32	-5%
NMe_2	H	6.31	-0.63	-9%
CF_3	H	7.44	0.50	7%
NMe_2	NMe_2	6.64	-0.30	-4%
H	NMe_2	7.16	0.22	3%
CF_3	NMe_2	7.92	0.98	14%

B4 Accuracy of H-TEQ 4.0 and GAFF2 Compared to QM Torsional Energy Profile: Training Set- Number of Occurrences vs. RMSE.

RMSE (kcal/mol)	GAFF (count)	H-TEQ (count)
0.1	2	1
0.3	4	8
0.5	6	3
0.7	7	12
0.9	3	6
1.1	0	5
1.3	3	4
1.5	3	2
1.7	0	1
1.9	0	2
2.1	1	0
2.3	0	0
2.5	2	2
2.7	0	1
2.9	0	0
3.1	0	0
3.3	1	2
3.5	0	0
3.7	0	0
3.9	0	-
4.1	0	-
4.3	0	-
4.5	1	-
4.7	0	-
4.9	0	-
5.1	0	-
5.3	0	-
5.5	1	-
5.7	1	-
5.9	0	-
6.1	1	-
6.3	0	-
6.5	0	-
6.7	0	-
6.9	0	-
7.1	1	-
7.3	0	-

7.5	0	-
7.7	0	-
7.9	0	-
8.1	0	-
8.3	0	-
8.5	0	-
8.7	0	-
8.9	0	-
9.1	0	-
9.3	0	-
9.5	0	-
9.7	0	-
9.9	0	-
10.1	1	-
10.3	0	-
10.5	0	-
10.7	0	-
10.9	0	-
11.1	0	-
11.3	0	-
11.5	0	-
11.7	0	-
11.9	0	-
12.1	0	-
12.3	0	-
12.5	0	-
12.7	0	-
12.9	1	-
13.1	0	-
13.3	0	-
13.5	1	-
13.7	0	-
13.9	0	-
14.1	1	-
14.3	1	-
14.5	1	-
14.7	0	-
14.9	2	-
15.1	0	-
15.3	0	-
15.5	1	-
15.7	0	-
15.9	0	-

16.1	1	-
16.3	0	-
16.5	0	-
16.7	0	-
16.9	1	-
17.1	0	-
17.3	0	-
17.5	0	-
17.7	1	-
17.9	0	-

B5 Accuracy of H-TEQ 4.0 and GAFF2 Compared to QM Torsional Energy Profile: Validation Set- Number of Occurrences vs. RMSE.

RMSE (kcal/mol)	GAFF (count)	H-TEQ (count)
0.1	2	1
0.3	4	8
0.5	6	3
0.7	7	12
0.9	3	6
1.1	0	5
1.3	3	4
1.5	3	2
1.7	0	1
1.9	0	2
2.1	1	0
2.3	0	0
2.5	2	2
2.7	0	1
2.9	0	0
3.1	0	0
3.3	1	2
3.5	0	-
3.7	0	-
3.9	0	-
4.1	0	-
4.3	0	-
4.5	1	-
4.7	0	-

4.9	0	-
5.1	0	-
5.3	0	-
5.5	1	-
5.7	1	-
5.9	0	-
6.1	1	-
6.3	0	-
6.5	0	-
6.7	0	-
6.9	0	-
7.1	1	-
7.3	0	-
7.5	0	-
7.7	0	-
7.9	0	-
8.1	0	-
8.3	0	-
8.5	0	-
8.7	0	-
8.9	0	-
9.1	0	-
9.3	0	-
9.5	0	-
9.7	0	-
9.9	0	-
10.1	1	-
10.3	0	-
10.5	0	-
10.7	0	-
10.9	0	-
11.1	0	-
11.3	0	-
11.5	0	-
11.7	0	-
11.9	0	-
12.1	0	-
12.3	0	-
12.5	0	-
12.7	0	-
12.9	1	-
13.1	0	-
13.3	0	-

13.5	1	-
13.7	0	-
13.9	0	-
14.1	1	-
14.3	1	-
14.5	1	-
14.7	0	-
14.9	2	-
15.1	0	-
15.3	0	-
15.5	1	-
15.7	0	-
15.9	0	-
16.1	1	-
16.3	0	-
16.5	0	-
16.7	0	-
16.9	1	-
17.1	0	-
17.3	0	-
17.5	0	-
17.7	1	-
17.9	0	-

Appendix C:

Supplementary Information for "Use of Extended-Hückel Descriptors for Rapid and Accurate Predictions of Conjugated Torsional Energy Barriers"

C1 Results of Subunit π-orbital Analysis

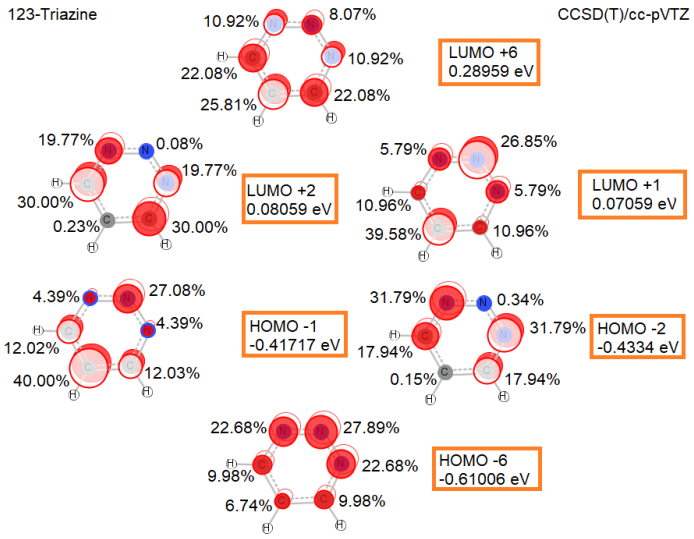


Figure C1. Orbital coefficient and molecular orbital energies of 1,2,3-triazine as optimized by CCSD(T)/cc-pVTZ.

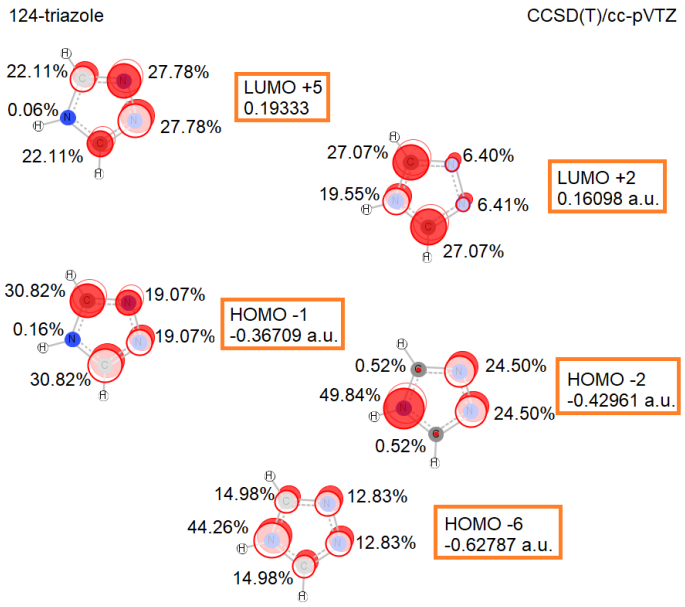


Figure C2. Orbital coefficient and molecular orbital energies of 4H-1,2,4-triazole as optimized by CCSD(T)/cc-pVTZ.

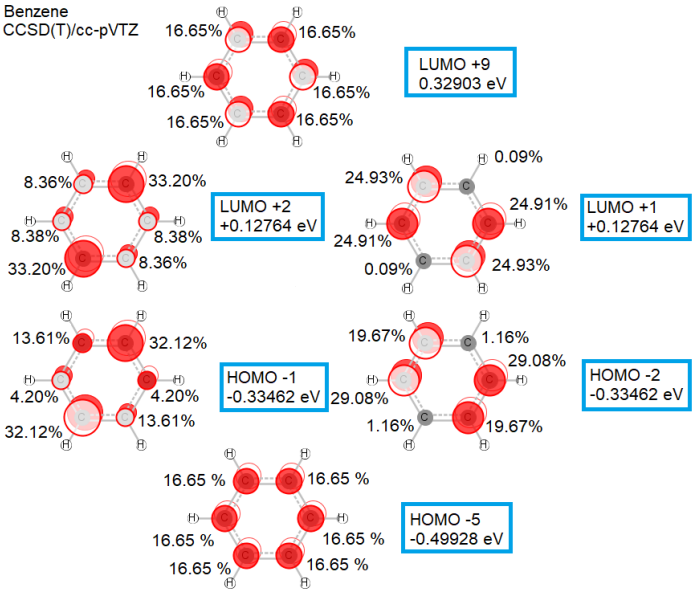


Figure C3. Orbital coefficient and molecular orbital energies of benzene as optimized by CCSD(T)/cc-pVTZ.

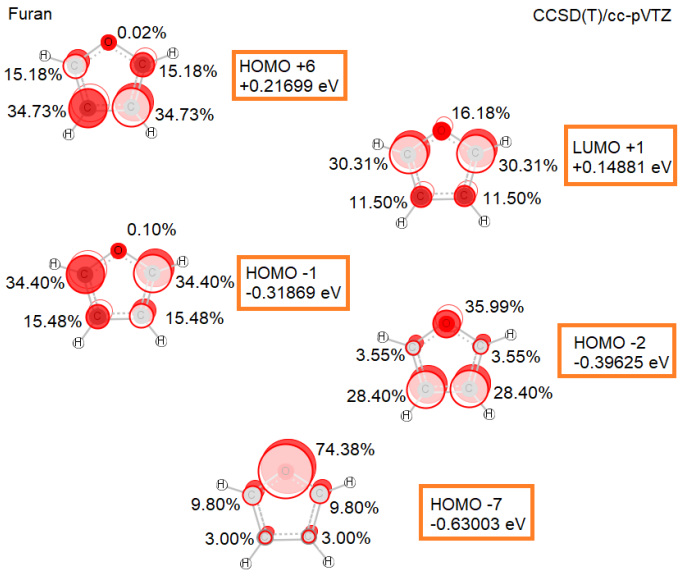


Figure C4. Orbital coefficient and molecular orbital energies of furan as optimized by CCSD(T)/cc-pVTZ.

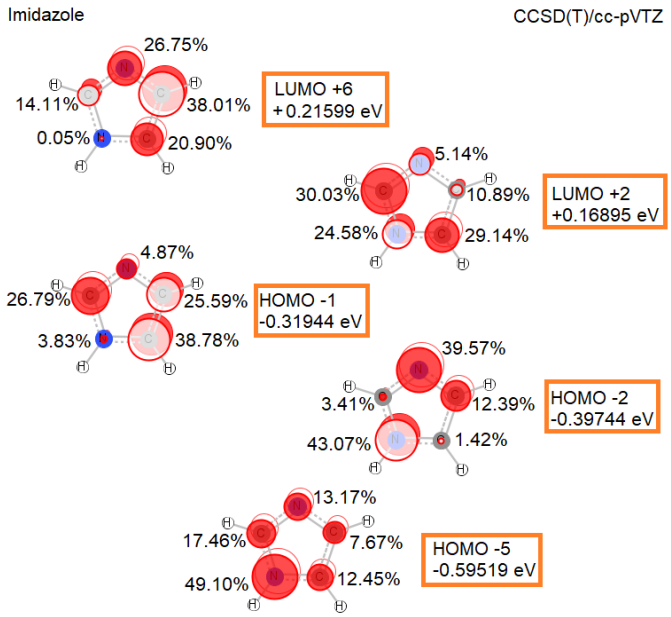


Figure C5. Orbital coefficient and molecular orbital energies of imidazole as optimized by CCSD(T)/cc-pVTZ.

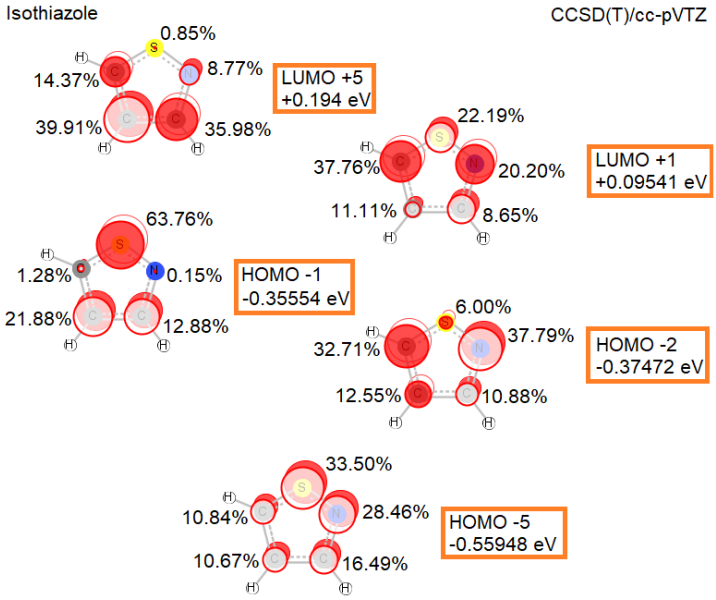


Figure C6. Orbital coefficient and molecular orbital energies of isothiazole as optimized by CCSD(T)/cc-pVTZ.

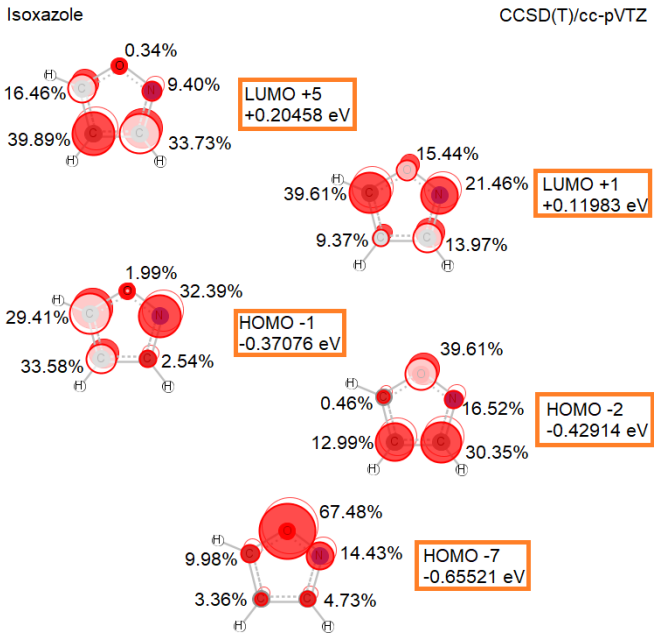


Figure C7. Orbital coefficient and molecular orbital energies of isoxazole as optimized by CCSD(T)/cc-pVTZ.

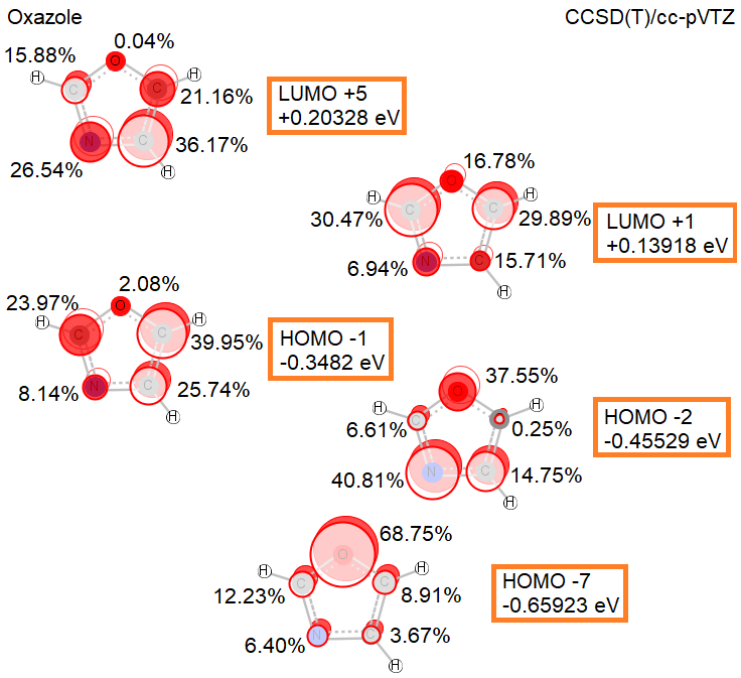


Figure C8. Orbital coefficient and molecular orbital energies of oxazole as optimized by CCSD(T)/cc-pVTZ.

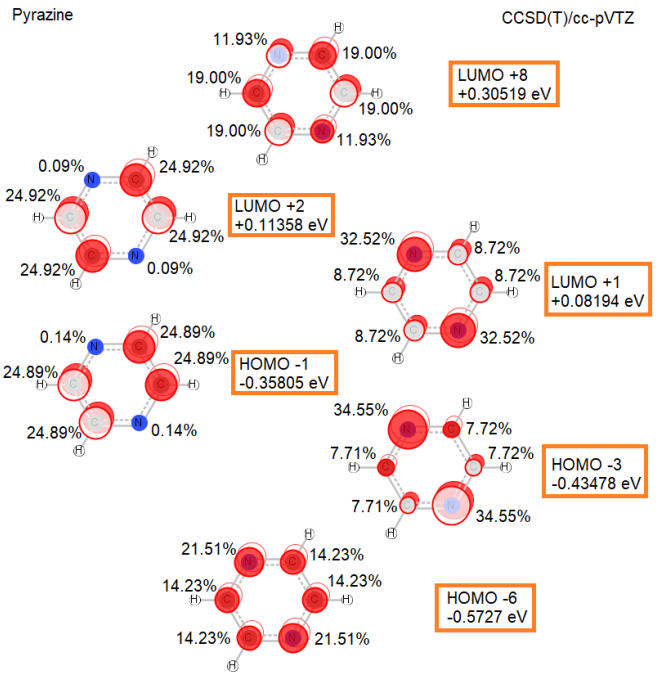


Figure C9. Orbital coefficient and molecular orbital energies of pyrazine as optimized by CCSD(T)/cc-pVTZ.

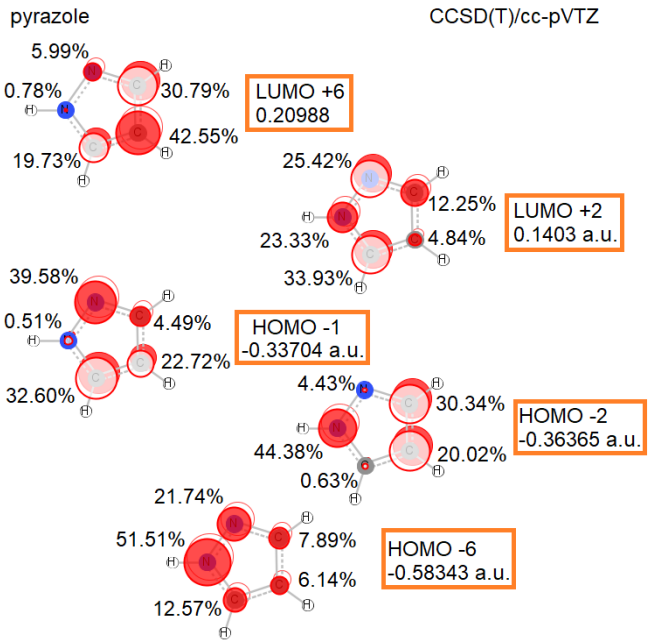


Figure C10. Orbital coefficient and molecular orbital energies of 1H-pyrazole as optimized by CCSD(T)/cc-pVTZ.

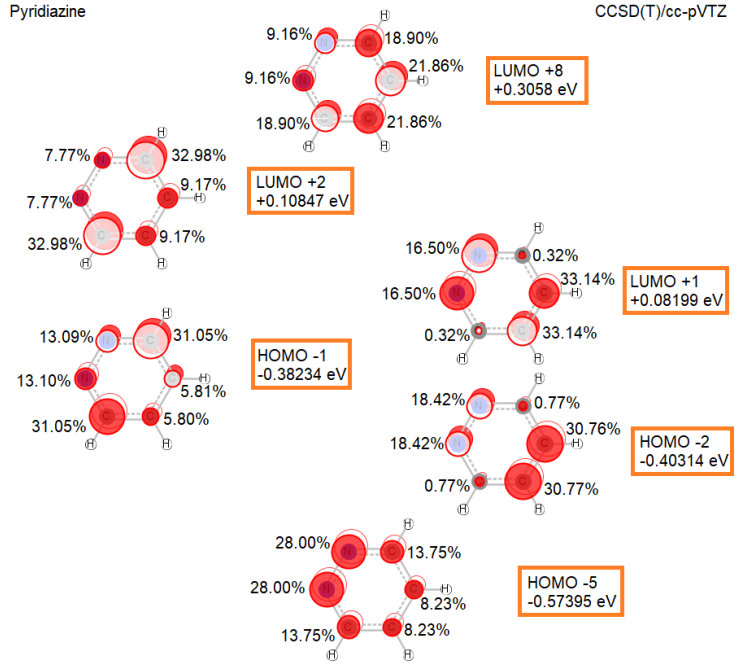


Figure C11. Orbital coefficient and molecular orbital energies of pyridazine as optimized by CCSD(T)/cc-pVTZ.

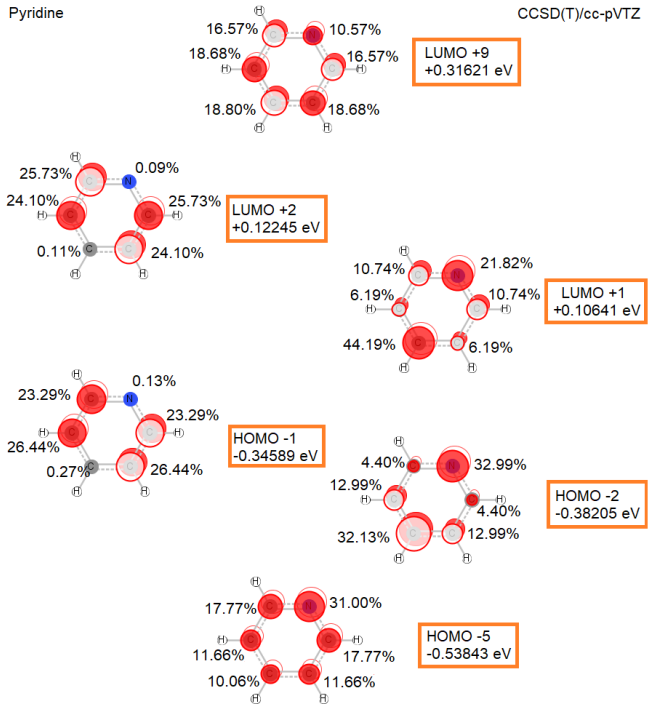


Figure C12. Orbital coefficient and molecular orbital energies of pyridine as optimized by CCSD(T)/cc-pVTZ.

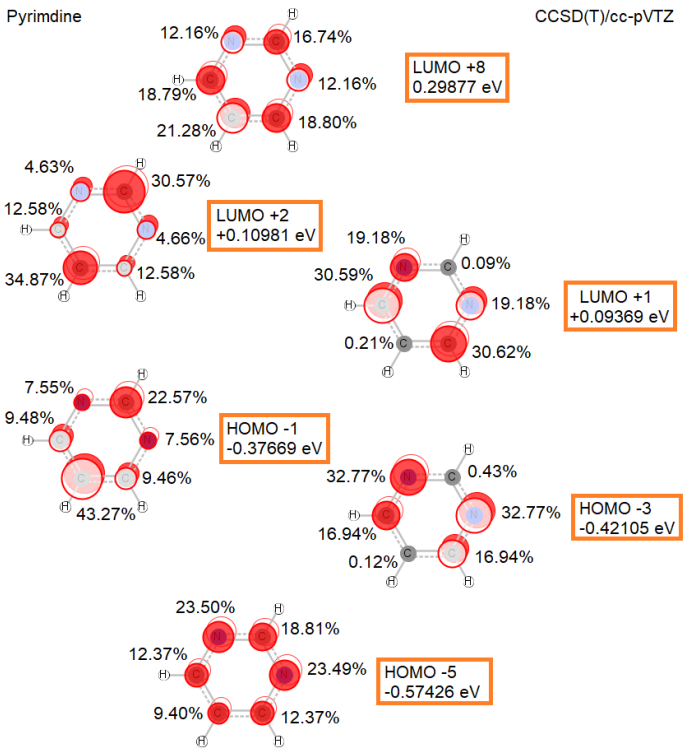


Figure C13. Orbital coefficient and molecular orbital energies of pyridmine as optimized by CCSD(T)/cc-pVTZ.

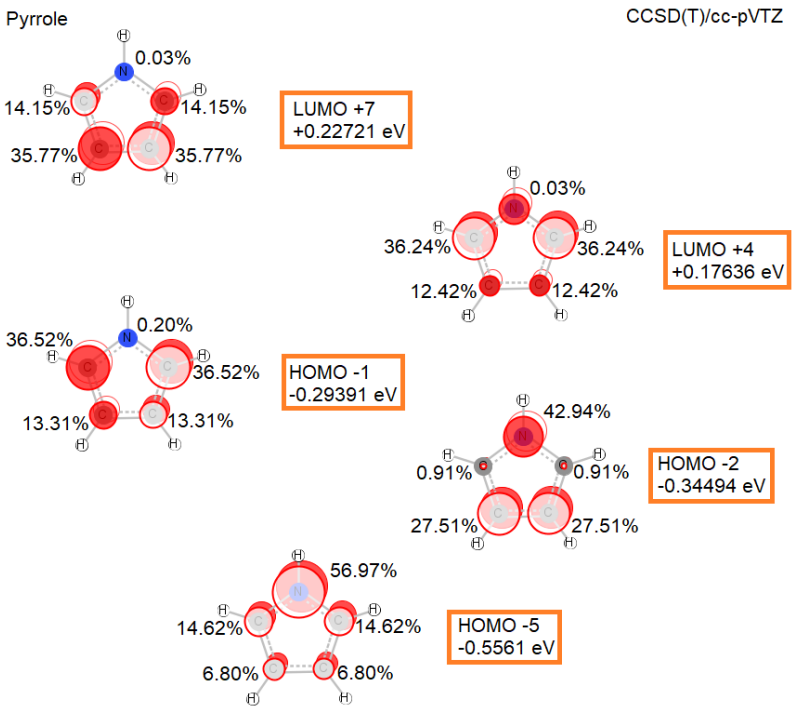


Figure C14. Orbital coefficient and molecular orbital energies of pyrrole as optimized by CCSD(T)/cc-pVTZ.

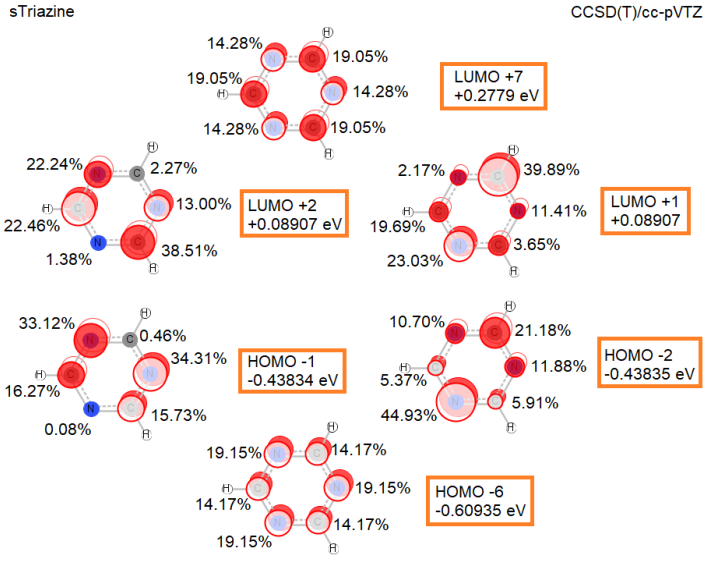


Figure C15. Orbital coefficient and molecular orbital energies of 1,3,5-triazine as optimized by CCSD(T)/cc-pVTZ.

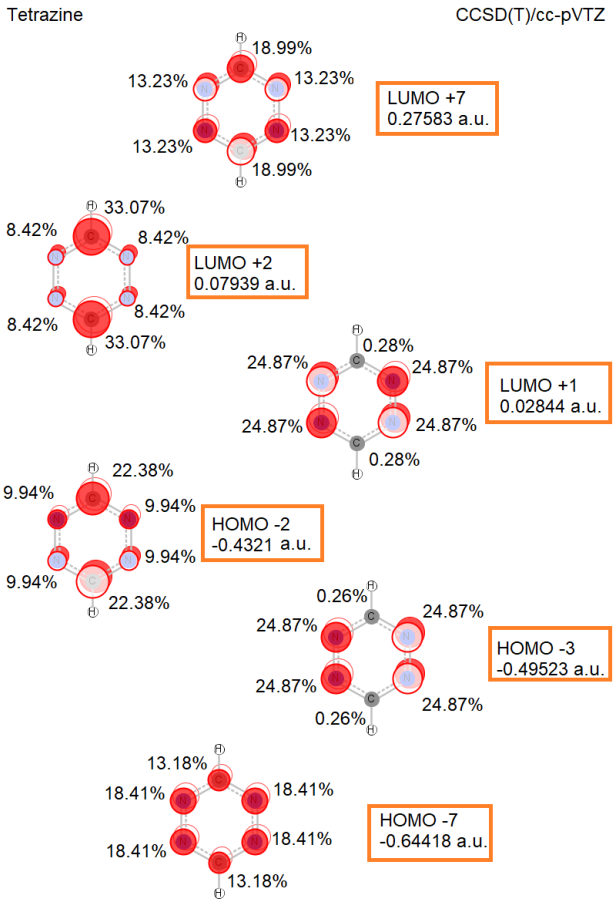


Figure C16. Orbital coefficient and molecular orbital energies of 1,2,4,5-tetrazine as optimized by CCSD(T)/cc-pVTZ.

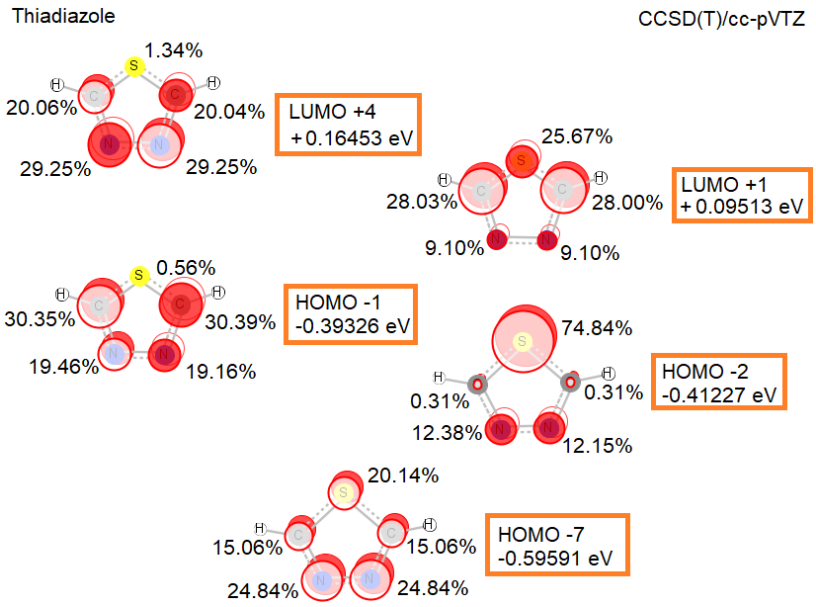


Figure C17. Orbital coefficient and molecular orbital energies of 1,3,4-thiadiazole as optimized by CCSD(T)/cc-pVTZ.

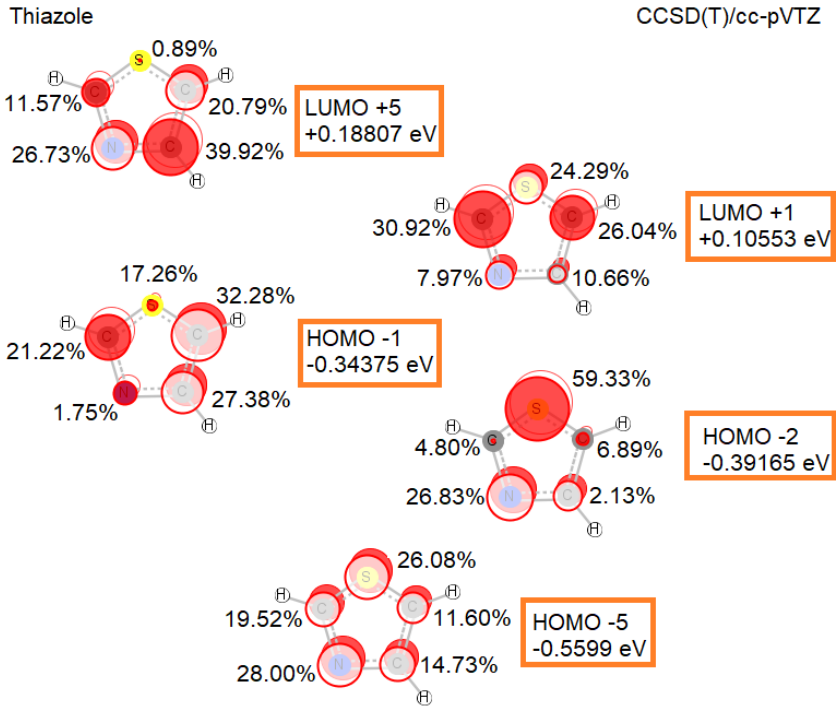


Figure C18. Orbital coefficient and molecular orbital energies of thiazole as optimized by CCSD(T)/cc-pVTZ.

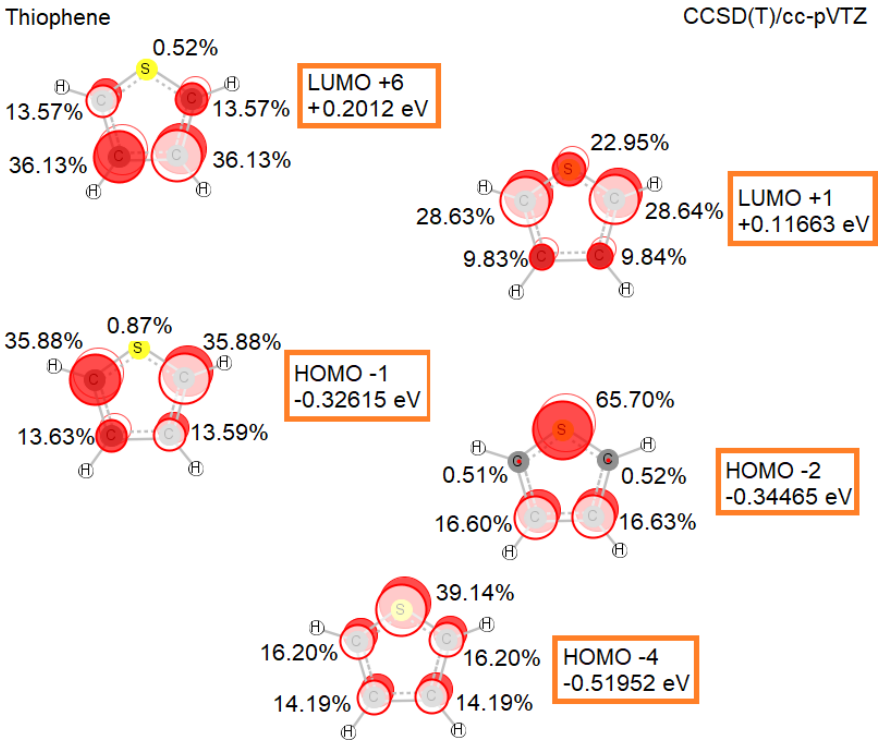


Figure C19. Orbital coefficient and molecular orbital energies of thiophene as optimized by CCSD(T)/cc-pVTZ.

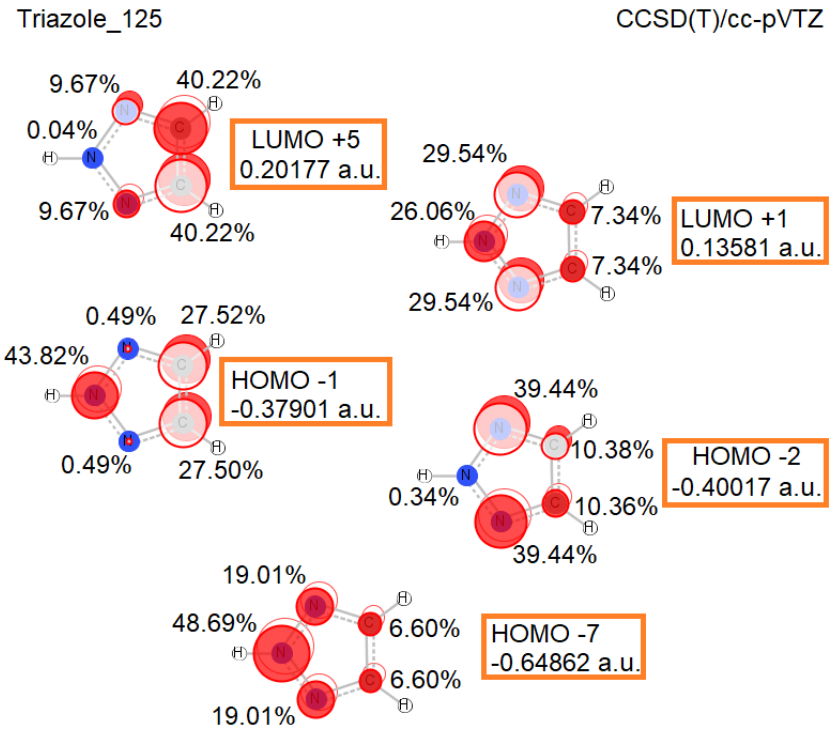


Figure C20. Orbital coefficient and molecular orbital energies of 2H-1,2,3-triazole as optimized by CCSD(T)/cc-pVTZ.

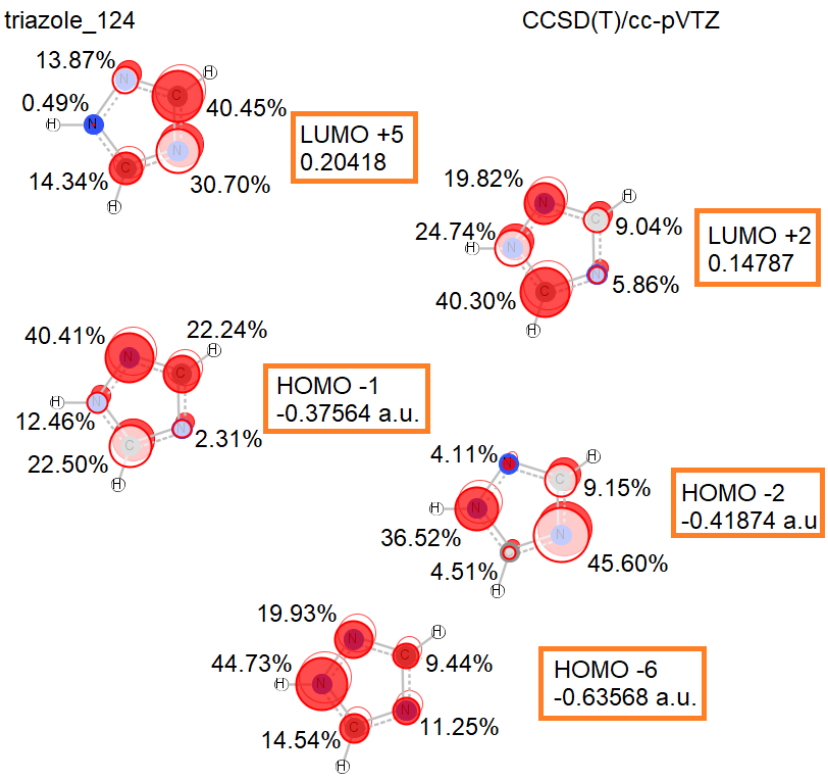


Figure C21. Orbital coefficient and molecular orbital energies of 1H-1,2,4-triazole as optimized by CCSD(T)/cc-pVTZ.

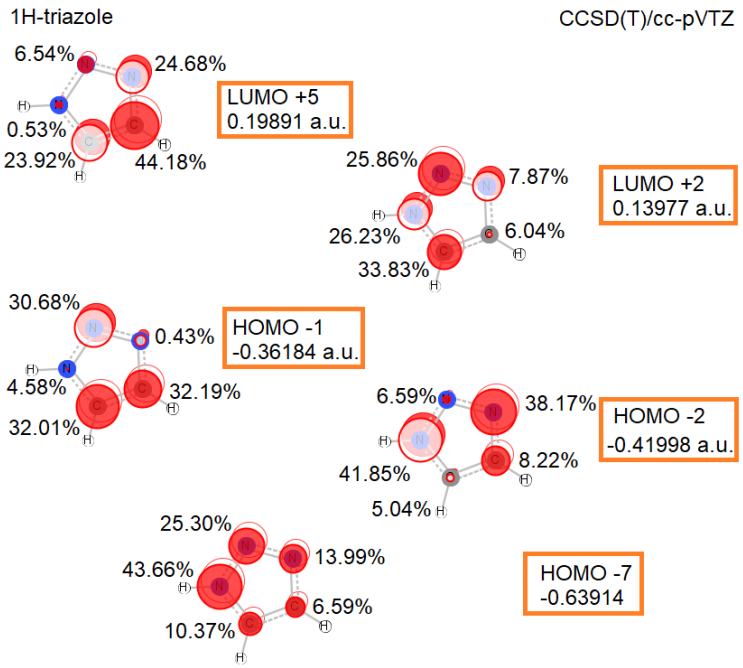


Figure C22. Orbital coefficient and molecular orbital energies of 1H-1,2,3-triazole as optimized by CCSD(T)/cc-pVTZ.

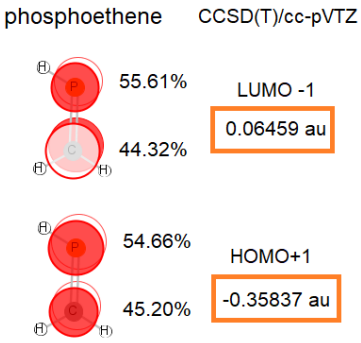


Figure C23. Orbital coefficient and molecular orbital energies of methylenephosphane as optimized by CCSD(T)/cc-pVTZ.

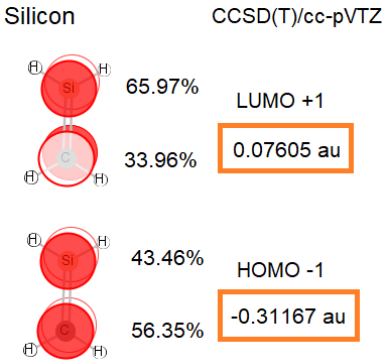


Figure C24. Orbital coefficient and molecular orbital energies of methylenesilane as optimized by CCSD(T)/cc-pVTZ.

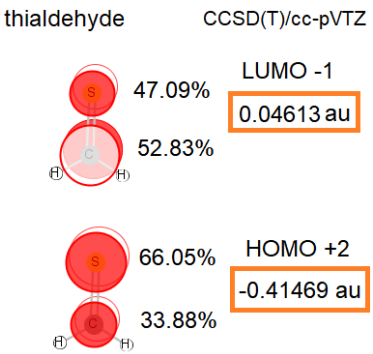


Figure C25. Orbital coefficient and molecular orbital energies of methanethial as optimized by CCSD(T)/cc-pVTZ.

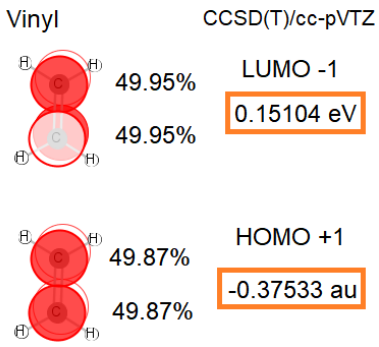


Figure C26. Orbital coefficient and molecular orbital energies of ethylene as optimized by CCSD(T)/cc-pVTZ.

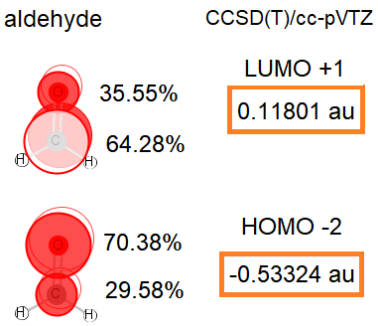


Figure C27. Orbital coefficient and molecular orbital energies of formaldehyde as optimized by CCSD(T)/cc-pVTZ.

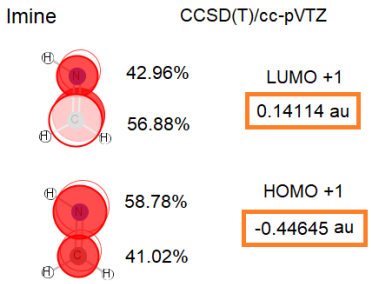


Figure C28. Orbital coefficient and molecular orbital energies of methanimine as optimized by CCSD(T)/cc-pVTZ.

C2 Qualitative Trends in V_2

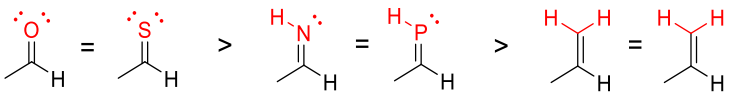


Figure C29. Observed trends in V_2. In general, V_2 increased from left to right, signifying a decrease in torsional energy barrier.

C3 V_2 of QM vs. V_2 from Sum of All σ → σ* hyperconjugation obtained from NBO

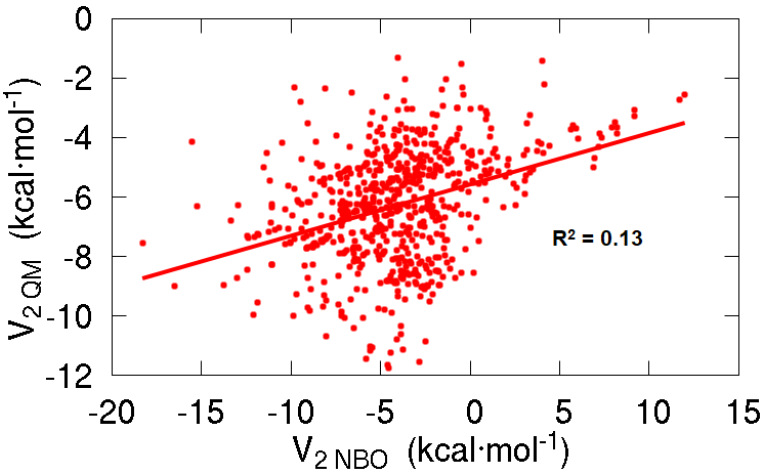


Figure C30. V_2 of QM vs. σ → σ* hyperconjugation obtained from NBO.

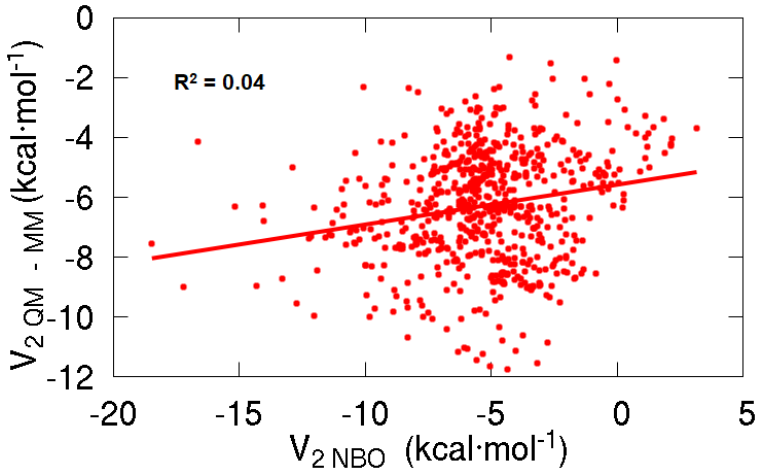


Figure C31. V_2 of QM-$MM_{nonbonded}$ vs. σ → σ* hyperconjugation obtained from NBO.

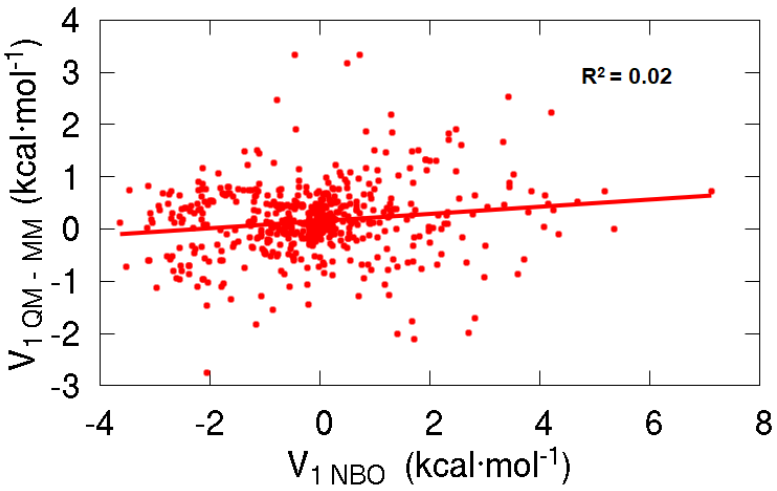


Figure C32. V_1 of QM-$MM_{nonbonded}$ of GAFF2 vs. $\sigma \rightarrow \sigma^*$ hyperconjugation obtained from NBO.

C4 Sample torsional profiles of GAFF2 vs. QM

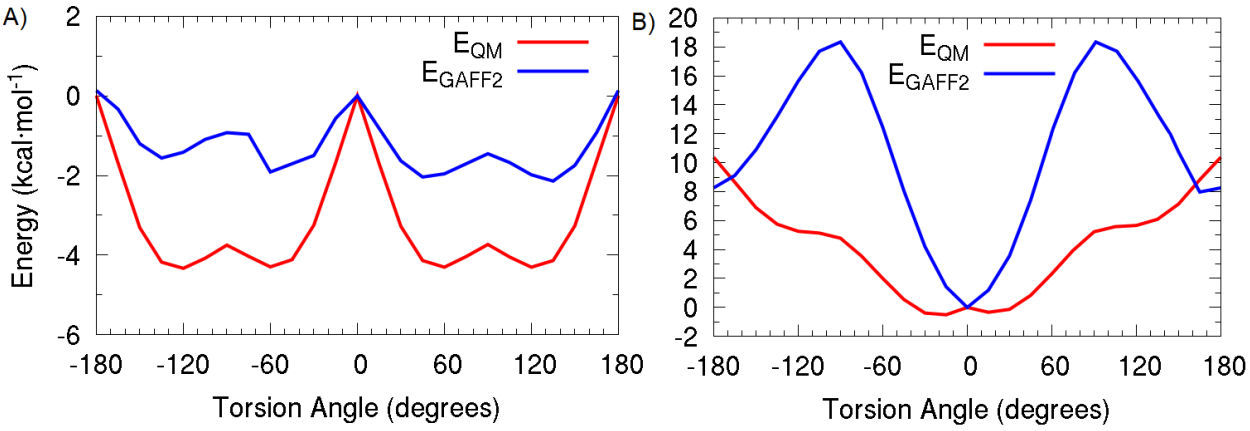


Figure C33. Comparison between the torsional energy profiles of a) (S)-6-(4-(1H-pyrrol-3-yl)phenyl)-5-methyl-4,5-dihydropyridazin-3(2H)-one and b) 2-amino-2-thioxoethane(dithioperoxo)imidic as predicted by QM (red) and GAFF2 (blue).

C5 Classification of Errors for H-TEQ 4.5

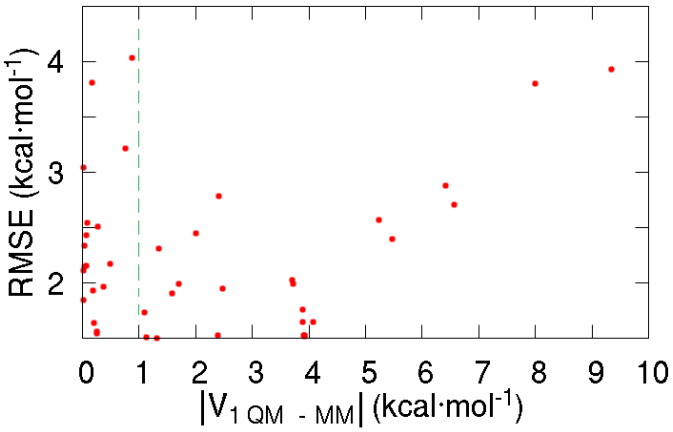


Figure C34. RMSE of H-TEQ 4.5 vs. magnitude of V_1 obtained from QM-$MM_{nonbonded}$ of GAFF2 are plotted for all molecules with RMSEs greater than 1.5 kcal·mol^{-1}. The green dotted line is used to classify molecules based on the performance of H-TEQ 4.5. Molecules to the right of this line were erroneous due to a missing V_1 term, while molecules to the left were inaccurate due to imperfections in the V_2 term.

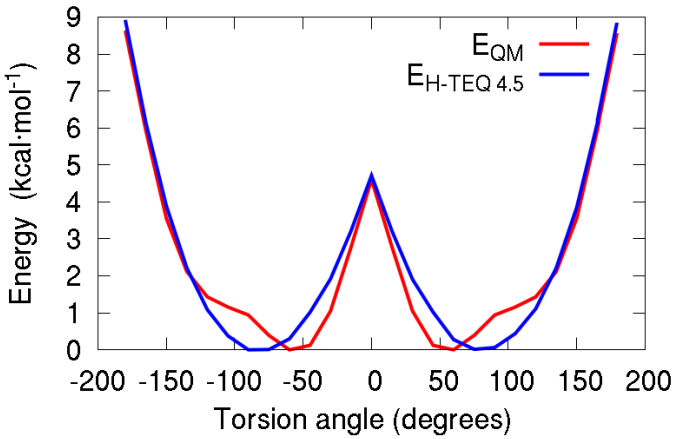


Figure C35. Torsional energy profile of [2,2'-bipyridine]-1,1'-diium, whereby the 1-4 electrostatics and 1-4 Van der Waals energies were multiplied by a factor of 6.5.

C6 Accuracy of H-TEQ 4.5, GAFF2, MAB, and MMFF94 Compared to QM Torsional Energy

Table C1. Profile. All RMSEs are reported in kcal/mol.

Validation Molecules in SMILES	rmse_GAFF	rmse_MMFF94	rmse_Amber14EHT	rmse_HTEQ
Fc1ccc(C(=O)OC)cc1	0.344863537	0.154829147	0.227455903	0.113673342
O=C(OC)c1ccccc1	0.407172736	0.192810682	0.331505911	0.128971991
O=[N+]([O-])c1ccc(C(=O)O)cc1	0.447923762	0.289627401	0.612737484	0.148542649
Sc1n(-c2ccccc2)ncn1	0.206896317	0.135775192	5.047025484	0.166322009
Clc1ccc(C=2C(C)(C)C=2)cc1	0.980088132	1.390741806	1.17657741	0.193881167
c1(-c2cscc2)ncccn1	0.254251535	0.400457761	0.272532323	0.2183828 5
POC(=O)c1ccccc1	0.179695015	0.250286722	0.166174208	0.220925019
C(C)(C)(C)c1ccc(-c2nocn2)cc1	0.765945434	0.495294328	0.784602703	0.237343477
[O-][n+]1onc(-c2ccccc2)c1	0.891915369	0.530263269	0.695705889	0.254843644
c1(-c2ccncc2)[nH]ncn1	0.3946580 4	0.145773332	0.178885311	0.2592316 2
Nn1nc(-c2ccccc2)cc1	0.612255275	0.429417808	0.474824347	0.2605001 1
O=C(OCC)Nc1oc(-c2ccccc2)cn1	3.342434299	2.187919348	4.945394308	0.264944846
c1(-c2ccccc2)[nH]ccn1	0.351645252	0.277053035	0.300250393	0.267731698
ClC(=N)c1snc(C#N)n1	16.999024 2	1.447254299	0.81957583	0.279652298
N#Cc1c(N)oc(-c2ccccc2)n1	0.2757170 6	0.269247608	0.344334383	0.280622876
S=C([S-])c1[n+H]cc[nH]1	6.426823532	1.086353772	0.686365501	0.284363792
N(C)(C)c1ccc(-c2[nH]c3c(n2)cccc3)cc1	0.319883973	0.361394285	0.389905221	0.294638819
N(C)(C)c1ccc(-c2[nH]c3c(n2)cccc3)cc1	0.319883973	0.361394285	0.389905221	0.294638819
c1(-c2cocc2)ncccn1	0.241620914	0.2593246	0.203649361	0.298308936
Fc1c(F)cc2nc(-c3ccncc3)[nH]c2c1	0.534642603	0.503259432	0.53541416	0.308054779
Cc1oc(-c2ccccc2)nn1	0.487584596	0.366689328	0.587190219	0.308210086

c1(-c2ccccc2)nocc1	0.840242092	0.594997343	0.594220952	0.309994979
O=C1OC(c2occc2)=Nc2[nH]ncc12	8.361419668	1.14230779	1.208980121	0.311575581
N#C[C@@H]1C(c2ccccc2)=C1	0.735711916	1.049729622	0.898374136	0.31390183
c1(-c2ccccc2)cscc1	0.728912843	0.510713555	0.277257562	0.332341562
c1(-c2ccccc2)c[nH]cc1	0.820221058	0.620152226	0.432019201	0.336825093
Clc1n(-c2ccc([N+](=O)[O-])cc2)nc(Cl)n1	0.276870319	0.423068034	5.74634828	0.34040993
O=[N+]([O-])c1c(N)non1	0.688673843	0.254960483	0.657841333	0.350166774
n1(-c2n[nH]cn2)cnnc1	4.965472552	1.779233018	7.773043606	0.353901909
O(C)c1ccc(-c2ncsc2)cc1	0.531137835	0.320566417	0.497811943	0.358602687
c1(-c2ccccc2)cocc1	0.993893024	0.687341371	0.507282969	0.369641395
Cn1nc(-c2ccccc2)nn1	0.597294353	0.407206492	0.723377091	0.37053435
S(C)c1[nH]nc(-c2c(O)cccc2)n1	0.250906985	1.051110751	1.078436321	0.371680097
Fc1ccc(-c2occ(C#N)c2)cc1	0.674049448	0.598179591	0.539634534	0.377983946
c1(-c2ccccc2)nn[n-]n1	0.194836684	0.21648341	0.285933085	0.378537011
Clc1c(N)[nH]nc1-c1ccc(F)cc1	0.790220647	0.660008141	0.410675684	0.392562327
Fc1c(-c2nc(C)[nH]n2)cccc1	0.658865545	0.573913657	1.119020109	0.393136179
Oc1c(O)ccc(-c2n[nH]cc2)c1	0.995554543	0.826558679	0.884083259	0.402846997
Nc1scc(-c2ccc(C)cc2)n1	1.270221813	1.189738897	1.302420857	0.403270511
S(C)c1nn(-c2ccccc2)cn1	0.259349993	0.261665052	5.772864449	0.405055368
O=C(O)c1ccc(NC(=[N+H2])N)cc1	0.617942831	0.298547157	0.841603392	0.423996302
O=C(N(O)c1ccc(C(=O)C)cc1)C	0.665055207	0.832184699	0.787602967	0.429292292
O=C1C(C#N)=CC(c2ccncc2)=C(C)N1	0.666169263	0.558813203	1.64994493	0.430955692
O=[N+]1C(c2ccccc2)=C[N-]O1	0.411906942	0.237240144	0.892644614	0.433201034
Cl[C@@H]1[C@@H](NC(=O)c2ccccc2)C1	0.772851753	0.420432722	0.767664508	0.433300628
Cc1cc(-c2cocc2)ccc1	1.164955451	0.833410049	0.660914797	0.437317537

O=C1N(c2ccccc2)N=NN1	0.445292869	0.377805968	5.833658353	0.44136999
S(C)c1[nH]c(-c2c(O)cc3c(c2)cccc3)nn1	0.916230461	1.436975995	1.725027077	0.454674296
ClCC(=O)c1ccccc1	0.869177294	0.736298884	0.414319731	0.455291294
ClCC(=O)c1ccccc1	0.869313273	0.736200402	0.414153141	0.455354487
O(C#C)C(=O)c1ccccc1	0.203621064	0.506133422	0.232804415	0.462667238
S=C(N)c1n[nH]cn1	6.115921274	0.823471341	1.556434929	0.463875873
O=[N+]([O-])c1sc(N2C(=O)NCC2)nc1	0.209443549	0.886954829	0.758534972	0.46388507
SC(=O)c1ccccc1	0.993669577	0.61042304	0.654307153	0.480653686
c1(-c2ccncc2)ocnn1	0.515933661	0.459219425	0.641663513	0.48232557
Cc1ccc(C=2Sc3n(N=2)cnn3)cc1	0.833535491	0.87011888	1.187898925	0.486382609
O=C(NC)c1ccccc1	0.476109801	0.371985101	0.788793178	0.489038415
[N+H2]=C(N)c1nc(N)[nH]n1	13.99997943	0.48165545	0.571153419	0.490363121
O=C1NN=C([C@@H](C)C1)c1ccc(-[n+]2c[nH]cc2)cc1	3.013785254	360.1144945	54.69633541	0.507815595
Cl/C(=N\S)/c1snc(C#N)n1	16.78617552	1.106180916	0.669622913	0.517212845
S(C)c1nc(-c2ccncc2)[nH]n1	0.512050383	0.279067266	0.472997124	0.518052485
Nc1nc(-c2ccncc2)nc2c1cccc2	0.79205113	0.669264377	0.324424341	0.526086854
O=C(C)c1ccc(N(O)C)cc1	0.818347395	0.835862257	0.924336237	0.526197963
O=Cc1ccc(OC)cc1	0.775660504	0.238308102	0.243621005	0.53111871
c1(-c2[nH]ccc2)[nH]c2c(n1)cccc2	8.962727057	2.210234605	1.793968243	0.534528341
NC=1n2nccc2N=C(c2ccccc2)C=1	0.297272439	0.512878175	0.491960756	0.537716923
Cn1c(-c2ccccc2)ccc1	0.363785644	0.309305444	0.514394691	0.538947192
Oc1oc(-c2ccccc2)nc1	1.156650051	1.140619944	1.358628232	0.550284841
O=S1(=O)CC(c2ccccc2)=CC1	0.594032339	1.321140712	0.72041424	0.55702091
O=C(N)c1cc(OC)ccc1	1.639688639	1.054532089	1.567611525	0.565228399
O=C1SC(c2ccc(C)cc2)=C2SCC(=O)N12	0.286590799	1.105531063	1.453369651	0.567995909

Fc1cc(c(O)cc1)-c1[nH]ncc1	0.643098648	1.094696457	1.577966185	0.570186147
O/N=C/c1ncccc1	0.401753191	1.685430664	1.758417492	0.570657489
c1(C2=Nc3n(ncn3)C=C2)ccccc1	0.37687632	0.447094316	0.54414095	0.575664148
[N+H](/N=C(\C)/c1c[nH]nc1)=C(N)N	15.95912707	0.715602074	0.607913176	0.5826757
S/C=C(\C#N)/c1ccccc1	0.543826242	0.383777303	1.255406862	0.58781319
O=C(NO)C(=O)NO	2.656581796	4.68244361	2.887852514	0.58871315
[PH2](=O)NC(=O)c1cc(O)ccc1	0.677139025	0.597400869	0.617722057	0.597798834
O=C1NC(c2ncccn2)=CC=C1	0.786519736	0.95405286	0.404229686	0.598714281
Cc1nn2c(n1)-c1c(n(-c3cc(C)c(C)cc3)nc1)N=C2	0.998520881	0.32501461	6.098805201	0.599865027
O(C)c1ccc(-c2n3N=CCSc3nn2)cc1	0.637368816	0.665774895	1.013953035	0.604940597
O=C(c1[nH]ccn1)c1cc(OC)c(O)cc1	4.982626743	3.568848262	3.388029366	0.612085474
O=C1N(c2ccccc2)N=NC(N(C)C)=N1	2.256407816	1.809982252	4.416287414	0.616600009
ClC(Cl)(Cl)[C@@H](O/N=C(\N)/c1ccccc1)O	0.477149112	0.461381142	0.988516474	0.620109875
O=C(N)c1cc(O)ccc1	1.688123064	1.087946805	1.598232788	0.621303013
O(C)c1c(-c2ccccc2)sc(C)n1	1.408092155	1.46101112	2.467400118	0.62991498
S=C1SC=C(c2ccccc2)N1	0.796860706	0.908144691	0.294396042	0.648371843
O=[N+]([O-])c1oc2c(c1)ccc1c2CCCC1	0.318577292	0.936272002	0.362476908	0.649699034
O=C(OC)c1[nH]c2c(occ2)c1	4.987426057	0.645053085	0.467969932	0.65060269
Cn1c(-c2ncccn2)ccc1	0.8294459	0.836348297	0.669695752	0.653810794
Nc1n(C)c(-c2cc3OCOc3cc2)cn1	0.280724226	0.345969382	0.734612091	0.664651701
O=[N+]1N=C(c2ccccc2)C=C(C)[N-]1	0.532160452	0.76277784	0.557819048	0.677774317
O=C1OC=Nc2n(-c3ccccc3)ncc12	0.572983056	0.319190535	6.110798041	0.687304563
S/N=C(\N)/c1ccccc1	0.420459954	0.385116156	0.800708485	0.688675062
O/N=C(\N)/c1ccccc1	0.57283676	0.452832303	1.140492437	0.701593384
S=C([S-])c1ccc(C(=S)[S-])cc1	1.090029883	1.142181393	1.658964476	0.704490771

O=C1NC(c2ccccc2)=CC=C1	0.609026077	0.54426994	0.632916481	0.709155785
Cn1c(-c2ncccn2)ncc1	0.875883	0.722238717	0.684548683	0.709764127
Oc1nc(-n2nc(C)cc2)nc(C)c1	1.164591799	1.3639808	5.279048025	0.715301037
O=[N+]1[N-]C(c2ccccc2)=CC(C)=C1	0.400510546	0.407843001	1.402721997	0.74080203
Cl/C(=N/O)/C=N/O	16.91566156	2.349280536	0.979549458	0.749004019
Cl/C(=N/O)/C=N/O	16.91566156	2.349280536	0.979549458	0.749004019
C(\C=C\C=C)=C/C=C	0.85513321	2.64970466	0.908171814	0.766019957
O=C(N)C=1C(C)(C)OP(=O)([O-])C=1	4.982418727	2.105387051	1.478794069	0.782194794
O=[N+]1C(C)=C(/C(=N/O)/C)[N-]O1	16.27834124	7.764517368	2.974362578	0.786451428
O=C(O)N1NN([O-])c2c1cccc2	1.958760892	0.519968756	3.09661365	0.788509069
c1(-c2cnccc2)oncn1	0.957721958	0.724770464	0.941390677	0.788678855
O/N=C(\C#N)/c1ccccc1	0.557481218	0.234403558	0.562344218	0.799424386
O/N=C(\C#N)/c1ccccc1	0.557579371	0.234427008	0.563459904	0.799641136
n1(-c2ccccc2)ncc-2c1N=Cn1c-2ccc1	0.890381763	0.609787455	5.478599672	0.806961973
O=C1N2NC(C)=NNC2=NC(c2ccccc2)=C1	0.504513444	0.452731293	0.702348586	0.808118936
C(=C/C=C)\C=C	16.65295544	2.616974256	0.602052701	0.809416327
O=[N+]1C(C)=C(/C(=N/O)/C)[N-]O1	16.2714052	7.763071509	2.974153326	0.811354128
O=C1NN=C(c2ccc(-[n+]3c[nH]cc3)cc2)CC1	1.070213229	0.391216726	0.407829188	0.859131201
Nc1ncnc2n(-c3ccccc3)ncc12	1.242664698	0.802832571	5.485027889	0.890439382
N(C)c1oc(-c2sccc2)nc1	9.582561403	0.891640079	0.594695221	0.894938489
Fc1nn(-c2ccccc2)nc1	0.841877525	0.449549624	5.930674289	0.901429558
FC(F)(F)c1c2c(O)ncnc2nc(-c2sccc2)c1	0.998664581	1.237172844	1.768643222	0.919079292
O=[N+]([O-])c1c2c(n(-c3[nH]nnn3)nc2)cc([N+](=O)[O-])c1	3.656196911	3.241984643	7.710520567	0.920940726
n1(-c2ccccc2)c2nc3c(nc2cc1)cccc3	1.244082339	0.857069039	4.476657456	0.937219297
[N+H2]=C(N)c1ccccc1	1.146660619	1.524732782	1.541792396	0.946272737

[N+H2]=C(N)c1ccccc1	1.146734444	1.524541337	1.541389836	0.946331607
O/N=C(/C=N\O)\C	16.40821068	3.895532294	1.024455966	0.953763108
O=C1NC(c2occc2)=CC=C1	9.394156558	1.075943397	0.612069026	0.969787512
O=C(C)c1c(N)c2c(o1)cccc2	4.048459035	1.01364652	1.465785164	0.994238513
O=C(OC)c1c(N)scn1	5.615405092	1.206551206	0.935919593	1.009430208
O=C(OC)c1c(N)scn1	5.615404861	1.206551272	0.935919471	1.009430259
N(=C(\C)/c1sccc1)\c1ccccc1	15.92943923	1.493004696	1.491888952	1.02431129
O=C1NC(c2[nH]ccn2)=CC=C1	8.950297247	0.476412579	0.408723964	1.024684387
ClC1(Cl)C(=O)[C@H]2C(C=O)=CC[C@@H]12	4.201344556	2.518197344	0.83669135	1.057299636
ClC1(Cl)C(=O)[C@H]2C(C=O)=CC[C@@H]12	4.203256513	2.519674415	0.837635723	1.05883304
Clc1ccc(N2C(=O)c3c(N=C2)cccc3)cc1	1.688468059	1.527898811	4.326919626	1.076505286
O=C1NC(n2cccc2)=CC=C1	2.847834231	0.724602231	4.799378073	1.080998147
O=C(OC)C(=O)OC	0.239053924	1.552515546	0.633197841	1.100582021
O=C(OC)C(=O)OC	0.239713842	1.553768849	0.633413431	1.101243503
c1(-c2ncsc2)ncsc1	9.9992402	1.650774258	1.384432733	1.106345822
O=Cc1c2c([nH]c1)cccc2	4.852541859	0.887365916	0.779459335	1.120192913
Nc1nc(N)nc(-n2nc(C)cc2)n1	1.694839619	1.892379706	4.46453309	1.125665981
O=C1NC(c2cscc2)=CC=C1	9.618758634	1.294450974	0.537982149	1.129609088
O=C1NC(c2[nH]ccc2)=CC=C1	8.727929795	0.867038328	0.47146354	1.143107299
O=C(O)C=O	0.447049227	2.376067097	0.831693711	1.171092769
O=C(O)C=O	0.447049227	2.376067097	0.831693711	1.171092769
O=Cc1c(OC)cccc1	1.270509866	1.852742833	1.724611866	1.192914005
O=C(NC)c1nc(C)on1	6.600266488	1.158147759	2.030572728	1.196469709
S(C)c1nc(-c2cnccc2)[nH]n1	1.234571568	0.788802538	0.817597552	1.21463407
O=C1NC(c2cocc2)=CC=C1	9.940534748	1.448461023	0.838195611	1.239497015

O=C(NO)C(=O)NO	3.207286318	4.768882604	3.274973491	1.261721055
ClN1C(=O)n2c(c(C(=O)N)nc2)N=N1	7.043913449	3.319620558	3.360477072	1.263161234
S=C(N/N=C/C(=O)O)N	4.099649539	2.570230182	2.027370062	1.276093362
O=C1N(/C(=N\O)/C(=N/O)/C)C=CC=C1C	14.09610146	1.745917831	3.792881411	1.280490936
O=C(N)c1ncn2C(=O)NN=Nc12	6.664229083	3.096224874	2.99513803	1.30809508
O(C)c1cc(-c2scc(OC)c2)sc1	10.36129774	1.909374132	1.021150961	1.31336871
O(C)c1c(-c2sc(N)cc2)cccc1	1.308432913	1.546838167	1.660249398	1.357204857
O[N+H]=Cc1ccc(O)cc1	2.480501448	4.031526076	4.046925697	1.386526811
n1(-c2ccccc2)nc2nccnc2n1	1.487185868	9.89092007	7.257485331	1.424827419
c1(-c2ccccc2)nc2[n+](nc[nH]2)cc1	0.810683934	1.479988209	1.528915707	1.452387284
c1(-c2sccc2)n2NC=CSc2nn1	9.583721058	1.65673613	2.006660771	1.465255195
O=C1NC(c2sccc2)=CC=C1	9.491107977	1.535901132	0.687525988	1.474810666
S=C(N)C(S)=N	5.899675969	2.689375898	3.14893269	1.505890879
[Si]=CC=[Si]	15.91496478	3.207132343	2.084531919	1.512801376
O=C(N)c1[nH]cnc1	5.819852721	1.352494606	1.528016448	1.516663502
S=C(NC)c1occc1	6.32595973	0.89389817	1.24136113	1.529276689
S=C(NC)c1occc1	6.325956376	0.893896435	1.241340553	1.529304975
S(S)C(=N)C(=S)N	5.240046682	1.646048796	3.25116104	1.529376349
O=C(c1ccc(N)cc1)c1ccc(N)cc1	0.290239784	0.325888073	0.423010004	1.542973895
S=C1N(N(C)C(C)=C1)c1ccccc1	1.172178099	1.049487503	2.568172072	1.561753953
O=C1N(c2cc3nc[nH]c3cc2)C(=O)c2c1cccc2	0.955914955	1.059718851	6.141028671	1.643871977
O(C(=N)c1snc(C#N)n1)C	17.40369382	1.206714187	1.894148267	1.649831142
O=C(N)c1ncsc1	6.844895027	2.26160993	2.369057418	1.652487947
P(=O)(O)([O-])/C(=N/O)/c1ccccc1	3.61557675	1.37681185	1.824296065	1.735869846
S=C(N)c1c(O)cccc1	2.810770273	0.926058984	0.84873544	1.759665986

O=C(O)C(=O)[O-]	0.703275492	1.8875669	2.330015446	1.845022171
O/N=C(/C(=N/O)/N1CC1)\C	16.4057333	3.652155847	3.206282371	1.906185814
O=C(O)/C=C\C(=O)[O-]	1.561052693	1.419842987	1.073441573	1.93309013
S(=O)(=O)(NC)c1c(-c2sc(N)nn2)cccc1	1.824740276	2.237849589	2.322529524	1.946352543
O(C)c1c(c(OC)ccc1)C=1Oc2c(C(=O)C=1)cccc2	2.491511945	2.251487198	0.575116797	1.963939228
S/C(=N\O)/c1[n+](C)cccc1	1.458674836	1.420836348	1.392278281	1.992489519
S=C(N/N=C(/C(=O)O)\C(=O)[O-])N	4.455088317	2.02865744	1.300165729	1.993838423
N(\C(C)(C)C)=C/C=N/C(C)(C)C	17.14388113	3.026200602	1.347794914	2.027900927
SC(=[N+H][O-])c1ccccc1	1.033746467	1.157291799	0.446161195	2.110112182
O=C1C(=[N+]([O-])c2c1cccc2)c1ncccn1	0.334964859	0.668817023	0.666434097	2.148536263
Clc1ccc(C(=[N+H][O-])C)cc1	0.520927674	1.752374727	0.314908701	2.15885882
SNC(=O)C(=O)N	2.524719639	4.078981821	2.623963116	2.169951114
O=CC=C[O-]	6.03222292	1.961878327	5.371951992	2.311397767
SC(=[N+H][O-])c1ccccc1	1.272634142	1.477962136	0.838812925	2.341721675
O=[N+]([O-])c1c([N+](=O)[O-])nc(-c2[nH]c([N+](=O)[O-])c([N+](=O)[O-])n2)[nH]1	9.675328079	3.726321731	2.521849161	2.401378994
O(C)c1ccc(C=[N+H]O)cc1	4.071111552	3.532083274	3.246781413	2.431191313
c1(-c2[n+H]cccc2)[n+H]cccc1	10.41808998	3.039460626	2.510410479	2.449873853
S=C1N(c2cc(C)ccc2)C(C)=CS1	0.833602575	1.075086401	2.670627756	2.505429411
Nc1nc(-[n+]2ccccc2)c2nc[nH]c2n1	3.226881949	3.222861339	3.226907389	2.54270296
Oc1c([n+]([O-])ccc1O)-c1[n+]([O-])ccc(O)c1O	3.791685341	4.463054833	1.949826238	2.570653374
Clc1oc(-c2[nH]c3c(n2)cccc3)cc1	9.772316524	2.060080786	1.643920426	2.706076484
O=C(C)C=1C(=O)O[N-][N+H]=1	4.910641836	2.015812059	3.187310322	2.781943551
O=C(NO)C(=O)NO	3.047866503	5.565789814	3.243840021	2.883199069
O=C([O-])C(=O)[O-]	2.342660371	2.407770562	3.353503912	3.041544207

O=C1C(=C([O-])C)C(c2ccc(C)cc2)=CO1	3.608700195	2.20820482	2.310078912	3.216573157
S=C(C(F)F)/C=C/1\SC(C(F)F)=CC(C)=C\1	5.680214676	15.23489302	10.47641122	3.799744663
S/C(=N\O)/c1[n+H]cccc1	0.680456055	2.466049315	2.209165292	3.813253068
O=C(/C=C/1\S/C(=C(/C#N)\C#N)/S\1)C	4.33514476	4.82396862	4.018519566	3.933370539
c1(-c2ccccc2)nncnc1	0.648257162	0.518028442	0.48212226	4.031166391

Table C2. Ideal V_1 and V_2 values for 200 druglike molecules using GAFF2, MAB, and MMFF94. All values are reported in kcal/mol.

mol	V1_GAFF	V1_MMFF94	V1_MAB	V2_GAFF	V2_MMFF94	V2_MAB
O=CC=C[O-]	1.340	5.421	6.537	-18.372	-18.839	-17.743
O=C(C)c1c([O-])on[n+H]1	-2.413	14.724	17.830	-8.455	-12.506	-11.692
O=C(/C=C/1\S/C(=C(/C#N)\C#N)/S\1)C	-9.338	-9.161	-10.003	-11.591	-13.766	-11.744
Clc1ccc(C(=[N+H][O-])C)cc1	0.061	0.111	0.055	-5.577	-8.861	-6.102
S=C(C(F)F)/C=C/1\SC(C(F)F)=CC(C)=C\1	-8.004	-21.716	-11.480	-11.319	-18.822	-13.125
Oc1c([n+]([O-])ccc1O)-c1[n+]([O-])ccc(O)c1O	5.235105	-84.7959372	-90.2249	-8.46915	-28.7286476	-21.2221
O=C(C)c1c(N)c2c(o1)cccc2	-1.677	1.274	0.299	-12.131	-14.044	-13.071
O=C1C(=[N+]([O-])c2c1cccc2)c1nccn1	-0.037	-0.096	-0.131	-4.673	-7.213	-7.857
O(C)c1ccc([C+H]NO)cc1	-0.067	-0.490	-0.570	-15.818	-18.248	-16.531
O[N+H]=Cc1ccc(O)cc1	0.797	0.130	-0.044	-12.012	-13.897	-11.760
S=C(N/N=C/C(=O)O)N	0.742	-0.173	0.386	-6.753	-7.080	-7.091
[N+H2]=C(N)c1nc(N)[nH]n1	-0.051	-0.369	-1.337	-10.721	-6.501	-8.301
O=C(OC)c1c(N)scn1	-1.760	-0.345	-0.254	-7.888	-9.703	-9.481
O=C(OC)c1c(N)scn1	-1.760	-0.345	-0.254	-7.888	-9.703	-9.481
S=C(N/N=C(\C(=O)O)/C(=O)[O-])N	-1.705	-21.124	-17.154	-9.269	-8.267	-10.372
SC(=[N+H][O-])c1ccccc1	-0.023	-0.033	-0.117	-2.576	-5.118	-1.066
SC(=[N+H][O-])c1ccccc1	-0.027	0.062	-0.019	-2.342	-7.697	-2.657
O=Cc1c2c([nH]c1)cccc2	3.014	1.744	1.191	-8.366	-9.725	-8.882
O/N=C(/C=N\O)\C	-0.421	-2.040	-3.679	-5.497	-3.621	-5.351

SC(=NO)c1[n+H]cccc1	-0.171	1.471	2.313	-5.393	-7.563	-5.101
S/C(=N\O)/c1[n+](C)cccc1	3.719	2.400	5.492	-3.293	-4.779	-1.735
O=C(C)c1ccc(N(O)C)cc1	0.416	-0.129	0.169	-6.349	-7.650	-6.227
O=C(OC)c1[nH]c2c(occ2)c1	-0.094	-0.578	-0.385	-8.996	-10.556	-9.428
N(=C(\C)/c1sccc1)\c1ccccc1	1.362	-4.076	-4.534	-7.273	-10.572	-7.802
C(\C=C\C=C)=C/C=C	1.040	0.367	0.968	-6.586	-7.821	-6.752
O=C1OC(c2occc2)=Nc2[nH]ncc12	-0.473	-2.123	-2.208	-6.777	-7.872	-7.089
ClN1C(=O)n2c(c(C(=O)N)nc2)N=N1	-2.602	-3.056	-2.393	-5.891	-5.517	-5.102
O=C(c1[nH]ccn1)c1cc(OC)c(O)cc1	0.841	6.730	5.669	-6.720	-6.370	-6.579
O=C(N(O)c1ccc(C(=O)C)cc1)C	0.036	0.208	0.129	-6.055	-7.845	-6.165
O=C1N(/C(=N\O)/C(=N/O)/C)C=CC=C1C	1.500	10.519	11.957	-6.073	-5.547	-6.614
[O-][n+]1c(-c2ccccc2)cno1	-0.038	0.005	0.015	-5.758	-5.284	-3.132
Cl/C(=N/O)/C=N/O	-0.171	-1.242	1.101	-4.966	-4.715	-4.892
Cl/C(=N/O)/C=N/O	-0.171	-1.242	1.101	-4.966	-4.715	-4.892
O=C(N)c1ncn2C(=O)NN=Nc12	-2.902	-6.469	0.975	-5.515	-5.616	-6.070
O=C(O)C=O	0.200	-0.310	0.055	-3.570	-3.901	-3.656
O=C(O)C=O	0.200	-0.310	0.055	-3.570	-3.901	-3.656
O=Cc1c(OC)cccc1	-3.365	-4.175	-5.002	-6.703	-7.469	-6.468
O/N=C(/C(=N/O)/N1CC1)\C	1.583	-1.325	-1.723	-6.967	-5.543	-7.061
O=Cc1ccc(OC)cc1	0.376	0.480	0.471	-7.891	-9.260	-8.049
Cl/C(=N\S)/c1snc(C#N)n1	-0.298	0.549	0.659	-5.243	-8.022	-5.425
Clc1oc(-c2[nH]c3c(n2)cccc3)cc1	6.567	3.827	4.428	-3.827	-4.591	-3.966
O=[N+]([O-])c1c([N+](=O)[O-])nc(-c2[nH]c([N+](=O)[O-])c([N+](=O)[O-])n2)[nH]1	5.480	6.573	6.113	-5.300	-4.794	-5.570
O=C1NC(c2[nH]ccn2)=CC=C1	-1.807	-1.481	-2.075	-3.962	-5.431	-4.561
ClC1(Cl)C(=O)[C@H]2C(C=O)=CC[C@@H]12	-2.624	-1.213	-0.531	-5.917	-6.863	-6.245
ClC1(Cl)C(=O)[C@H]2C(C=O)=CC[C@@H]12	-2.625	-1.201	-0.525	-5.918	-6.850	-6.246
O/N=C/c1ncccc1	-0.570	-1.922	-3.313	-5.101	-5.655	-4.987
O=C1NN=C(c2ccc(-[n+]3c[nH]cc3)cc2)CC1	0.127	0.217	0.190	-8.405	-9.535	-8.330
O=C1NC(c2occc2)=CC=C1	0.550	0.609	0.008	-3.358	-5.007	-3.719
O=C(c1ccc(N)cc1)c1ccc(N)cc1	-0.252	-0.567	-0.325	-2.603	-4.208	-2.687
[N+H2]=C(N)c1ccccc1	-0.001	0.010	-0.060	-5.207	-9.869	-2.705
[N+H2]=C(N)c1ccccc1	-0.001	0.010	-0.060	-5.207	-9.870	-2.705
c1(-c2[n+H]cccc2)[n+H]cccc1	-2.006	4.863	8.001	0.196	-7.527	1.661
O=C1NC(c2[nH]ccc2)=CC=C1	1.286	1.555	1.195	-2.733	-5.070	-3.206
Cn1c(-c2ncccn2)ncc1	0.125	0.134	0.162	-7.369	-5.693	-6.830
O=[N+H]C(=O)C(=O)[N+H]=O	-6.411	-8.030	-9.126	-3.372	-3.046	-2.316

ClCC(=O)c1ccccc1	0.059	0.115	0.120	-6.043	-8.086	-6.501
ClCC(=O)c1ccccc1	0.058	0.116	0.120	-6.043	-8.086	-6.501
O=C(O)[N+]1=N[N+](=O)c2c1cccc2	-0.812	0.918	0.751	-11.671	-12.421	-11.663
O=C1NC(c2sccc2)=CC=C1	0.656	1.648	0.786	-1.836	-4.648	-1.586
O=C(O)/C=C\C(=O)[O-]	-0.180	-0.886	-1.594	-1.750	-3.383	-5.209
O(C)c1c(c(OC)ccc1)C=1Oc2c(C(=O)C=1)cccc2	0.364	0.200	0.368	-8.044	-4.244	-8.358
O=C(N)c1[nH]cnc1	-3.924	-1.162	-1.024	-4.937	-6.159	-5.178
O=C(OC)C(=O)OC	0.028	-0.407	-0.096	-3.035	-3.635	-3.579
O=C(OC)C(=O)OC	0.033	-0.402	-0.088	-3.034	-3.632	-3.573
Cn1c(-c2ncccn2)ccc1	0.004	0.081	0.024	-6.699	-6.716	-6.801
O=C(N)c1ncsc1	-4.068	-2.927	-2.091	-4.894	-5.751	-5.106
N(C)c1oc(-c2sccc2)nc1	-0.446	-0.472	-0.787	-3.483	-5.331	-4.170
N(\C(C)(C)C)=C/C=N/C(C)(C)C	3.699	1.311	-1.042	-3.790	-4.332	-3.727
O=C(O)c1ccc(NC(=[N+H2])N)cc1	0.093	0.157	0.079	-4.704	-6.851	-6.802
O=C1NC(c2ncccn2)=CC=C1	0.072	0.100	0.029	-7.300	-8.283	-7.918
O=C(NC)c1nc(C)on1	2.796	1.373	1.063	-3.987	-4.260	-4.200
P(=O)(O)([O-])/C(=N/O)/c1ccccc1	-1.094	0.982	1.036	-8.150	-5.728	-10.408
FC(F)(F)c1c2c(O)ncnc2nc(-c2sccc2)c1	-2.208	-3.860	-4.321	-6.197	-9.410	-6.530
S(C)c1[nH]c(-c2c(O)cc3c(c2)cccc3)nn1	0.400	4.291	3.899	-5.307	-5.156	-5.195
c1(-c2[nH]ccc2)[nH]c2c(n1)cccc2	0.602	-2.152	-2.335	-3.786	-4.889	-4.130
O=C1SC(c2ccc(C)cc2)=C2SCC(=O)N12	-0.059	-0.010	-0.099	-3.444	-9.272	-2.453
O(C(=N)c1snc(C#N)n1)C	3.885	-1.062	0.407	-3.617	-4.855	-3.774
c1(-c2ccncc2)ocnn1	0.050	0.067	0.061	-4.136	-5.305	-4.376
S=C1N(N(C)C(C)=C1)c1ccccc1	0.260	0.435	0.257	0.757	-6.309	0.264
O=[N+]([O-])c1ccc(C(=O)O)cc1	0.079	0.142	0.117	-5.197	-6.952	-5.835
O=[N+]([O-])c1sc(N2C(=O)NCC2)nc1	0.007	0.133	0.132	-6.603	-8.819	-6.137
Fc1ccc(C(=O)OC)cc1	-0.036	-0.009	-0.023	-5.225	-7.505	-5.834
O=C(O)C(=O)[O-]	0.011	-0.547	-0.432	-0.454	0.800	0.698
O(C)c1ccc(-c2n3N=CCSc3nn2)cc1	0.229	0.195	0.208	-3.672	-4.471	-4.039
C(=C/C=C)\C=C	1.128	0.572	1.037	-5.277	-6.448	-5.429
N#Cc1c(N)oc(-c2ccccc2)n1	-0.036	-0.009	-0.033	-4.577	-5.967	-4.963
S(C)c1nc(-c2ccncc2)[nH]n1	0.003	0.004	0.009	-3.961	-5.705	-4.591
O=C(OC)c1ccccc1	-0.019	0.015	-0.002	-5.011	-7.083	-5.433
POC(=O)c1ccccc1	-0.024	0.028	-0.015	-5.736	-7.934	-6.209
S/N=C(\N)/c1ccccc1	0.046	0.059	0.032	-4.331	-7.171	-4.500
S(=O)(=O)(NC)c1c(-c2sc(N)nn2)cccc1	-2.480	-3.990	-3.075	-4.336	-4.705	-4.476

ClC(=N)c1snc(C#N)n1	-0.558	0.192	0.073	-4.766	-7.524	-5.048
c1(-c2ccncc2)[nH]ncn1	-0.019	0.006	-0.015	-2.823	-4.412	-3.530
O=C1NC(c2cscc2)=CC=C1	-0.451	-0.511	-0.520	-2.153	-5.199	-2.319
Oc1oc(-c2ccccc2)nc1	-0.004	0.029	0.023	-3.769	-5.528	-4.561
c1(-c2sccc2)n2NC=CSc2nn1	-1.109	-3.879	-3.264	-1.237	-2.639	-1.503
Cc1oc(-c2ccccc2)nn1	0.035	0.067	0.055	-4.217	-5.493	-4.560
c1(-c2cnccc2)oncn1	-0.023	-0.069	0.237	-2.872	-4.497	-3.429
N(C)(C)c1ccc(-c2[nH]c3c(n2)cccc3)cc1	-0.011	0.036	0.039	-4.562	-6.266	-4.874
N(C)(C)c1ccc(-c2[nH]c3c(n2)cccc3)cc1	-0.011	0.036	0.039	-4.562	-6.266	-4.874
c1(-c2ccccc2)nc2[n+](nc[nH]2)cc1	0.033	0.002	0.027	-8.639	-10.698	-8.302
SNC(=O)C(=O)N	-0.487	1.371	1.259	-1.707	-2.387	-1.733
Fc1c(F)cc2nc(-c3ccncc3)[nH]c2c1	-0.008	0.013	-0.031	-4.974	-6.746	-5.275
c1(-c2cscc2)ncccn1	0.014	0.017	0.010	-5.142	-7.297	-6.002
O=C(NO)C(=O)NO	-0.685	-1.414	-1.604	-5.057	-5.444	-5.223
O=C(NO)C(=O)NO	0.088	-0.678	-0.894	-4.281	-4.698	-4.503
S/C=C(\C#N)/c1ccccc1	0.073	0.172	0.097	-6.179	-9.569	-6.143
O=C1NC(n2cccc2)=CC=C1	-0.159	0.002	-0.261	-1.558	-4.400	-1.034
S=C([S-])c1[n+H]cc[nH]1	-0.012	0.092	0.035	-4.571	-12.132	-8.175
O/N=C(\C#N)/c1ccccc1	0.072	0.138	0.134	-6.860	-8.233	-6.657
O/N=C(\C#N)/c1ccccc1	0.071	0.137	0.134	-6.860	-8.231	-6.656
O=C(N)C=1C(C)(C)OP(=O)([O-])C=1	0.863	0.488	0.871	-3.004	-4.460	-2.515
O=[N+]1C(C)=C(/C(=N/O)/C)[N-]O1	1.382	7.316	6.109	-5.917	-7.249	-6.513
O=[N+]1C(C)=C(/C(=N/O)/C)[N-]O1	1.490	7.319	6.113	-5.920	-7.252	-6.516
O(C#C)C(=O)c1ccccc1	-0.060	-0.048	-0.074	-5.961	-8.041	-6.467
S=C(NC)c1occc1	-3.915	-2.327	-1.167	-4.908	-8.191	-5.277
S=C(NC)c1occc1	-3.915	-2.327	-1.167	-4.908	-8.191	-5.277
S(C)c1nc(-c2cnccc2)[nH]n1	-1.470	1.113	1.596	-2.058	-4.549	-3.326
Nc1nc(-c2ccncc2)nc2c1cccc2	0.020	-0.031	-0.073	-6.001	-7.260	-6.207
O=[N+]([O-])c1oc2c(c1)ccc1c2CCCC1	-0.019	0.163	0.194	-6.454	-7.468	-6.395
[N+H](/N=C(\C)/c1c[nH]nc1)=C(N)N	-0.457	-0.308	-0.324	-5.836	-7.623	-5.195
O=C1NC(c2ccccc2)=CC=C1	0.016	0.124	0.006	-2.940	-7.134	-3.250
c1(C2=Nc3n(ncn3)C=C2)ccccc1	-0.008	0.044	0.010	-5.914	-8.197	-5.803
O=C1NC(c2cocc2)=CC=C1	-1.245	-0.463	-0.407	-1.627	-4.150	-1.953
[PH2](=O)NC(=O)c1cc(O)ccc1	-0.889	-0.138	-0.130	-3.970	-6.990	-4.493
O(C)c1c(-c2sc(N)cc2)cccc1	-3.377	-3.308	-3.859	-6.093	-10.166	-6.319
Cl[C@@H]1[C@@H](NC(=O)c2ccccc2)C1	0.079	0.159	0.094	-3.617	-6.258	-3.905

O=C(NC)c1ccccc1	0.064	0.112	0.080	-3.397	-5.783	-3.493
c1(-c2ccccc2)nn[n-]n1	-0.001	0.020	0.006	-5.275	-8.285	-6.940
O=[N+]([O-])c1c(N)non1	-0.035	0.176	0.233	-5.328	-4.131	-2.914
O=C(N)c1cc(OC)ccc1	-0.670	-0.526	-0.373	-3.538	-5.806	-3.643
O=C(N)c1cc(O)ccc1	-0.757	-0.279	-0.262	-3.397	-5.650	-3.488
S=C(N)c1n[nH]cn1	0.893	-1.987	-1.689	-4.847	-5.917	-4.851
c1(-c2ccccc2)[nH]ccn1	-0.019	0.011	-0.025	-4.234	-6.088	-4.697
c1(-c2cocc2)ncccn1	0.043	0.011	0.008	-5.133	-6.850	-5.688
O=C(OCC)Nc1oc(-c2ccccc2)cn1	0.000	-0.040	-0.051	-3.756	-5.687	-4.012
SC(=O)c1ccccc1	-0.002	0.070	0.000	-3.541	-7.962	-3.716
O/N=C(\N)/c1ccccc1	0.037	0.039	0.020	-3.493	-5.652	-3.484
S=C1N(c2cc(C)ccc2)C(C)=CS1	-0.275	-0.622	-0.415	3.705	-2.841	3.280
ClC(Cl)(Cl)[C@@H](O/N=C(\N)/c1ccccc1)O	0.021	0.079	0.055	-3.480	-5.964	-3.525
Fc1ccc(-c2occ(C#N)c2)cc1	-0.010	-0.001	0.001	-3.398	-5.350	-3.593
O=C1C(C#N)=CC(c2ccncc2)=C(C)N1	-0.195	-0.142	-0.219	-3.770	-7.051	-3.842
NC=1n2nccc2N=C(c2ccccc2)C=1	0.033	0.091	0.104	-5.449	-8.090	-5.613
O=C1N2NC(C)=NNC2=NC(c2ccccc2)=C1	0.019	0.063	0.008	-6.235	-8.004	-5.655
Fc1cc(c(O)cc1)-c1[nH]ncc1	-1.300	-3.828	-2.832	-4.467	-5.564	-4.183
Nc1nc(N)nc(-n2nc(C)cc2)n1	-0.079	-0.127	-0.212	-6.748	-8.845	-8.223
O=C([O-])C(=O)[O-]	0.011	0.202	0.228	4.020	1.905	2.193
S=C1SC=C(c2ccccc2)N1	-0.004	0.047	-0.013	-2.582	-5.543	-2.717
Nc1nc(-[n+]2ccccc2)c2nc[nH]c2n1	0.082	-2.138	-0.852	-9.061	-9.436	-10.760
Nc1n(C)c(-c2cc3OCOc3cc2)cn1	0.300	0.389	0.479	-2.374	-5.798	-2.760
Oc1nc(-n2nc(C)cc2)nc(C)c1	-0.314	0.461	0.465	-5.446	-7.405	-6.944
Cc1ccc(C=2Sc3n(N=2)cnn3)cc1	-0.065	-0.026	-0.047	-3.346	-6.388	-2.976
O(C)c1c(-c2ccccc2)sc(C)n1	0.252	0.739	0.315	-5.009	-10.318	-5.319
Cn1c(-c2ccccc2)ccc1	0.086	0.232	0.092	-2.388	-5.867	-2.926
[Si]=CC=[Si]	1.132	-0.068	1.431	-3.499	-4.662	-3.762
S(C)c1[nH]nc(-c2c(O)cccc2)n1	-0.035	-1.867	-2.410	-4.532	-4.431	-4.998
O(C)c1cc(-c2scc(OC)c2)sc1	-0.721	-0.094	-0.332	-0.819	-4.300	-0.702
Fc1c(-c2nc(C)[nH]n2)cccc1	0.855	-0.762	-1.040	-3.493	-3.850	-3.889
Nn1nc(-c2ccccc2)cc1	0.001	-0.006	-0.025	-3.596	-5.566	-3.371
O=[N+]1N=C(c2ccccc2)C=C(C)[N-]1	0.021	0.055	0.030	-5.491	-8.762	-5.997
c1(-c2ccccc2)nncnc1	0.871	1.282	0.943	-4.465	-6.493	-4.755
C(C)(C)(C)c1ccc(-c2nocn2)cc1	0.015	0.082	0.023	-3.416	-4.963	-3.793
O=[N+]([O-])c1c2c(n(-c3[nH]nnn3)nc2)cc([N+](=O)[O-])c1	-1.429	6.622	5.918	-3.003	-0.783	0.207

Clc1ccc(C=2C(C)(C)C=2)cc1	0.010	0.031	0.008	-3.670	-4.108	-3.606
Cn1nc(-c2ccccc2)nn1	0.079	0.098	0.076	-4.034	-5.212	-4.048
S=C(N)c1c(O)cccc1	-3.886	0.792	0.918	-1.643	-6.577	-1.577
Clc1c(N)[nH]nc1-c1ccc(F)cc1	-0.005	0.016	0.026	-2.551	-5.269	-2.481
O=C1N(c2ccccc2)N=NC(N(C)C)=N1	0.417	0.081	0.381	-4.317	-4.565	-4.387
O=[N+]1[N-]C(c2ccccc2)=CC(C)=C1	0.008	0.044	0.045	-5.405	-7.009	-4.698
N#C[C@@H]1C(c2ccccc2)=C1	0.041	0.069	0.049	-4.044	-4.704	-4.201
O=S1(=O)CC(c2ccccc2)=CC1	-0.007	0.046	0.002	-4.133	-7.241	-4.667
c1(-c2ccccc2)nocc1	-0.005	0.029	0.001	-2.957	-5.235	-3.082
O=C1N(c2ccccc2)N=NN1	0.046	0.022	0.002	-4.222	-7.641	-4.753
S(S)C(=N)C(=S)N	-2.386	-6.697	-3.352	-2.063	-8.754	-2.529
S=C(N)C(S)=N	-1.312	-5.888	-3.996	-2.074	-7.211	-1.995
Oc1c(O)ccc(-c2n[nH]cc2)c1	-0.548	0.438	0.496	-3.410	-5.476	-3.231
n1(-c2ccccc2)c2nc3c(nc2cc1)cccc3	0.159	0.262	0.214	-4.418	-5.527	-3.699
Nc1scc(-c2ccc(C)cc2)n1	0.187	0.215	0.180	-3.538	-5.339	-3.780
Clc1n(-c2ccc([N+](=O)[O-])cc2)nc(Cl)n1	-0.060	0.054	0.108	-2.118	-3.666	-1.175
Clc1ccc(N2C(=O)c3c(N=C2)cccc3)cc1	0.400	0.484	0.051	-3.538	-9.610	-0.873
O(C)c1ccc(-c2ncsc2)cc1	0.288	0.322	0.327	-3.594	-5.403	-3.790
c1(-c2ccccc2)c[nH]cc1	-0.063	-0.008	-0.056	-2.641	-5.349	-2.628
c1(-c2ccccc2)cscc1	0.007	0.072	0.013	-2.763	-6.094	-2.855
O=C1C(=C([O-])C)C(c2ccc(C)cc2)=CO1	0.755	0.549	0.734	-6.819	-8.522	-7.357
Nc1ncnc2n(-c3ccccc3)ncc12	0.066	0.035	0.060	-4.823	-5.821	-4.544
[O-][n+]1onc(-c2ccccc2)c1	-0.004	0.029	0.009	-2.897	-3.922	-1.341
Cc1cc(-c2cocc2)ccc1	-0.008	0.033	-0.011	-2.207	-5.183	-2.542
c1(-c2ccccc2)cocc1	-0.022	0.033	-0.009	-2.202	-5.115	-2.532
S(C)c1nn(-c2ccccc2)cn1	0.037	0.066	0.042	-3.167	-5.285	-2.806
n1(-c2ccccc2)ncc-2c1N=Cn1c-2ccc1	0.052	0.123	0.111	-4.322	-4.906	-3.709
Cc1nn2c(n1)-c1c(n(-c3cc(C)c(C)cc3)nc1)N=C2	0.022	-0.048	-0.447	-3.880	-6.705	-4.036
Sc1n(-c2ccccc2)ncn1	0.016	0.039	0.056	-2.038	-5.425	-1.548
O=C1N(c2cc3nc[nH]c3cc2)C(=O)c2c1cccc2	0.199	0.165	0.168	-3.047	-2.946	-2.926
O=C1OC=Nc2n(-c3ccccc3)ncc12	0.105	0.018	0.082	-4.023	-4.668	-3.285
n1(-c2ccccc2)nc2nccnc2n1	0.007	0.048	0.037	-6.252	-6.662	-4.636
S=C([S-])c1ccc(C(=S)[S-])cc1	-0.052	0.067	-0.019	-4.050	-10.931	-2.829
O=C1NN=C([C@@H](C)C1)c1ccc(-[n+]2c[nH]cc2)cc1	0.346	-0.047	-0.229	-1.744	-4.462	-1.424
Fc1nn(-c2ccccc2)nc1	0.000	0.027	0.020	-4.648	-6.046	-4.080
c1(-c2ncsc2)ncsc1	2.005	1.854	2.453	-3.423	-4.196	-3.476

n1(-c2n[nH]cn2)cnnc1	0.028	-0.041	0.016	-0.458	-2.087	-0.787

Appendix D:

Supplementary Information for "Influence of Molecular Mechanics Torsional Parameters on Docking Accuracies"

D1 PDB IDs of Structures Used for Nucleic Acid Self-Docking

1AJU 1AM0 1ARJ 1BYJ 1EHT 1F1T 1F27 1FMN 1FUF 1KOC 1KOD 1L1H 1LC4 1LVJ 1MWL 1NBK-A 1NBK-B 1NTB 1NZM-B 1O0K-A 1O0K-B 1O15 1O9M 1RAW 1TOB 1UTS 1UUD 1UUI 1XPF 1YKV-A 1YKV-B 1YRJ 2AU4 2CKY 2ESI-A 2ESI-B 2ET3 2ET5-A 2ET5-B 2ET8 2F4S 2F4T 2F4U 2F4V 2FCX 2FCZ-A 2FCZ-B 2G5K 2G5Q 2GDI 2HOJ 2HOK 2HOM 2HOO 2HOP 2JWQ-A 2JWQ-B 2KXM 2L1V 2L7V-A 2L7V-B 2L8H 2L94 2LWK 2M4Q 2MG8 2MGN 2MIY 2MS6-A 2MS6-B 2N0J 2N6C-A 2N6C-B 2O3V 2O3X 2O3Y 2OE5 2OE8 2OEY 2QWY 2TOB 2XNZ 2XO1 2Z74 2Z75 3C3Z 3D2G 3D2V 3D2X 3DIG 3DIL 3DIQ 3DIR 3DJ0 3DS7 3E5C 3E5E 3E5F 3EM2 3EQW 3ES0 3ET8 3F2Q 3F2Y 3F30 3F4E 3F4H 3FO6 3GES 3GX2 3GX3 3GX5 3GX7 3K0J 3K1V 3NP6 3NPN 3NPQ 3NX5 3NZ7 3OWI 3OXE 3P49 3Q50 3R6R 3SD3 3SKI 3SKZ 3SLM 3SLQ 3SUH 3SUX 3TZR 3UD4 3V7E 453D 4AOB 4ERJ 4ERL 4F8U 4FE5-C 4FXM 4GPW 4GPX 4K31-A 4K31-B 4K32 4KQY 4L81 4LVV-A 4LVV-B 4LVX 4LVY-A 4LVY-B 4NYB 4NYG 4OQU 4P20 4P3S-A 4P3S-B 4PHY 4RZD 4TS2-A 4TS2-B 4WCQ 4ZC7-B 4ZNP 5BTP 5BWS 5BXK 5C45 5D99 5FJC 5FK1 5HBW 5KX9 5LWJ 5NEP 5TPY

D2 PDB IDs of Structures Used for Protein Self-Docking

1A07 1A0Q 1A1B 1A1E 1A28 1A42 1A4G 1A4K 1A4Q 1A6W 1A9U 1AAQ 1ABE 1ABF 1ACL 1ACM 1ACO 1AEC 1AJ7 1AKE 1AOE 1APT 1APU 1AQW 1ATL 1AZM 1B58 1B59 1B6N 1B9V 1BAF 1BBP 1BGO 1BL7 1BMA 1BMQ 1BYB 1BYG 1C12 1C1E 1C2T 1C5C 1C5X 1C83 1CBS 1CBX 1CDG 1CF8 1CIL 1CIN 1CKP 1CLE 1COM 1COY 1CPS 1CTR 1CTT 1CVU 1CX2 1D0L 1D3H 1D4P 1DBB 1DBJ 1DBM 1DD7 1DG5 1DHF 1DID 1DIE 1DMP 1DOG 1DR1 1DWB 1DWC 1DWD 1DY9 1EAP 1EBG 1EED 1EI1 1EJN 1ELA 1ELB 1ELC 1ELD 1ELE 1EOC 1EPB 1EPO 1ETA 1ETR 1ETS 1ETT 1ETZ 1F0R 1F0S 1F3D 1FAX 1FBL 1FEN 1FGI 1FIG 1FKG 1FKI 1FL3 1FLR 1FRP 1GHB 1GLP 1GLQ 1GPY 1HAK 1HDC 1HEF 1HFC 1HIV 1HOS 1HPV 1HRI 1HSB 1HSL 1HTF 1HTI 1HVR 1HYT 1IBG 1ICN 1IDA 1IGJ 1IVB 1IVC 1IVD 1IVE 1IVQ 1JAO 1JAP 1KEL 1KNO 1LAH 1LCP 1LDM 1LIC 1LKK 1LMO 1LNA 1LST 1LYB 1LYL 1MCQ 1MCR 1MDR 1ML1 1MLD 1MMB 1MMQ 1MNC 1MRG 1MRK 1MTS 1MTW 1MUP 1NCO 1NGP 1NIS 1NSD 1OKL 1OKM 1PBD 1PDZ 1PGP 1PHA 1PHD 1PHF 1PHG 1POC 1PPC 1PPH 1PPI 1PPL 1PSO 1PTV 1QBR 1QBT 1QBU 1QCF 1QH7 1QL7 1QPE 1QPQ 1RBP 1RDS 1RNE 1RNT 1ROB 1RT2 1SLN 1SLT 1SNC 1SRF 1SRG 1SRH 1SRJ 1STP 1TDB 1TKA 1TMN 1TNG 1TNH 1TNI 1TNL 1TPH 1TPP 1TRK 1TYL 1UKZ 1UVS 1UVT 1VGC 1VRH 1WAP 1XID 1XIE 1XKB 1YDR 1YDS 1YDT 1YEE 25C8 2AAD

2ACK 2ADA 2AK3 2CGR 2CHT 2CMD 2CPP 2CTC 2DBL 2FOX 2GBP 2H4N 2IFB 2LGS 2MCP 2MIP 2PCP 2PHH 2PK4 2PLV 2QWK 2R04 2R07 2SIM 2TMN 2YHX 2YPI 3CLA 3CPA 3ERD 3ERT 3GCH 3GPB 3HVT 3MTH 3NOS 3PGH 3PTB 3TPI 4AAH 4COX 4CTS 4DFR 4ER2 4EST 4FAB 4FBP 4LBD 4PHV 4TPI 5ABP 5ER1 5P2P 6ABP 6CPA 6RNT 7CPA 7TIM 8GCH