APPLICATION OF MACHINE LEARNING TO PROBLEMS IN COMPUTATIONAL CHEMISTRY AND BIOLOGY

By

Zhuoqin Yu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Chemistry—Doctor of Philosophy

2019

ABSTRACT

APPLICATION OF MACHINE LEARNING TO PROBLEMS IN COMPUTATIONAL CHEMISTRY AND BIOLOGY

By

Zhuoqin Yu

With the ever-increasing amounts of chemical and biological data, advancement of machine learning algorithms and computational power, machine learning techniques have started to play a more important role in computational chemistry and biology. We have implemented machine learning models to solve a range of problems from structure prediction, force field development to the prediction of drug molecule toxicity. Since protein chemical shift perturbations (CSPs) induced by ligand binding can be used to refine the structure of a protein-ligand complex I developed a regression model, called HECSP, to compute ligand induced CSPs of protons in a protein, which yielded correlation coefficients of 0.897 (^{1}HA), 0.971 (^{1}HN) and 0.945 (sidechain ^{1}H) with root-mean-square errors (RMSEs) of 0.151 (^{1}HA), 0.199 (^{1}HN) and 0.257 ppm (sidechain ^{1}H), respectively. Based on HECSP, we can further distinguish native ligand poses from decoys and refine protein-ligand complex structures by comparing predicted CSPs with observed values, which is realized with a scoring function (NMRScore_P). Other than HECSP, I have also developed a regression model (EZAFF) to determine force field parameters of 4-6 coordinated zinc containing systems. The reliability of the model has been tested on 6 metalloproteins and 6 organometallic compounds with different coordination spheres. Besides regression, another important part of machine learning are classification problems like the prediction of toxicity of small molecules. Based on the Tox21 dataset, I trained models to predict toxicity using both chemical descriptors and one-dimensional similarities as molecular features. These models cover support vector machine (SVM), random forest (RF) and deep neural network (DNN). AUC results

have showed the benefit of including similarities for both RF and DNN. The Highest AUC achieved on the test set is 0.879 by RF.

TO THE GLORY OF GOD
To my wonderful parents for their love and support.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God Almighty for giving me the opportunity, strength, knowledge and ability to undertake this Ph.D. study and to persevere to the end. Without his grace, I would not have gone this far.

The most precious gift from my Ph.D. study is that I found a mentor, a role model, Professor Kenneth M. Merz Jr. who is my advisor. I would like to express my sincere thanks to him for his patience, motivation, and wisdom. His guidance led me in all the seasons of research and writing of this thesis. I could not have imagined having a better mentor for my Ph.D. study.

My sincere thanks also go to the rest of my thesis committee: Professor Robert I. Cukier, Professor Katharine C. Hunt, and Professor Benjamin G. Levine, for their insightful comments and encouragement.

I would like to thank former and current lab mates. In particular, I thank Dr. Pengfei Li for collaborating with me on my research projects and guiding me and Dr. Nupur Bansal whose friendship and companion made my Ph.D. journey more memorable.

I would also like to thank iCER and HPCC facility at Michigan State University for providing me with the computational resources. I would also like to thank the chemistry staff for patiently taking care of tons of administrative problems and needs.

Endless thanks to my church family for supporting me spiritually throughout the years.

Last but not the least, I would like to thank my parents for always being there for me.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES xi

KEY TO ABBREVIATIONS xiv

CHAPTER 1 1
Introduction 1
1.1. Introduction to Machine Learning 2
1.2. Linear Regression 3
1.2.1. Linear Model Selection and Regularization 4
1.2.2. Ridge Regularization 4
1.2.3. Lasso Regularization 5
1.3. Random Forests 5
1.4. Support Vector Machines (SVMs) 8
1.5. Deep Neural Networks 10
1.5.1. Activation Functions 11
1.5.2. Dropout Regularization 12
1.6. Machine Learning in Drug Discovery 13
1.7. Deep Learning in Drug Discovery 14
REFERENCES 17

CHAPTER 2 21
Using Ligand Induced Protein Chemical Shift Perturbations to Determine Protein-ligand Structures 21
2.1. Abstract 22
2.2. Introduction 23
2.3. Methods 27
2.3.1 Preparation of the 1H Chemical Shift Perturbations Database 27
2.3.2. Ring Current Effects 29
2.3.3. Electric Field Effects 31
2.3.4. Hydrogen Bond Effects 32
2.3.5. Magnetic Anisotropic Group Contributions 33
2.3.6. Parameter Fitting 35
2.3.7. NMRScore_P and NMRScore_L 37
2.3.8. Application of Protein NMRScore_P to the *Apo*NCS-naphthoate ester complex 39
2.3.8.1. Scoring procedure for the structural ensemble. 39
2.3.8.2. Further refinement of the ligand solution structures. 39
2.3.8.3. Docking Procedure. 40
2.3.9. Ternary hIFABP-ketorolac-ANS complex determination through induced fit docking (IFD) and NMRScore_P. 40

2.4. Results and Discussion 41
2.4.1. Hydrogen bond term for the protein sidechain protons 41
2.4.2. Cutoff dependence for the electric field term 42
2.4.3. Charge model dependence 43
2.4.4. Leave-one-out-cross-validation analysis 44
2.4.5. Study on the apoNCS-naphthoate ester complex with NMRScore_P 45
2.4.5.1. The apoNCS-naphthoate ester complex ensemble 45
2.4.5.2. Further refinement of ligand solution structures 50
2.4.5.3. Native states and decoys. 52
2.4.6. The ternary hIFABP-ketorolac-ANS complex. 53
2.5. Conclusion 55
2.6. Acknowledgments 56
2.7. Supporting Information 57
REFERENCES 60

CHAPTER 3 69
The Extended Zinc AMBER Force Field (EZAFF) 69
3.1. Abstract 70
3.2. Introduction 70
3.3. Methods 74
3.3.1 Development and validation of the empirical scheme 74
3.3.1.1. Empirical scheme for deriving the bond and angle parameters 74
3.3.2 Benchmark evaluations of different MM and semi-empirical QM methods for modeling zinc-containing complexes 81
3.4. Results and Discussion 86
3.4.1 Validation of EZAFF 86
3.4.2. Benchmark assessment of different MM and semi-empirical QM methods for modeling zinc-containing systems 90
3.4.2.1. Energetic predictions 91
3.4.2.2. Structural Predictions 94
3.4.2.3. Gas phase MD simulations for CSD complexes with nonbonded parameter sets 97
3.5. Conclusions 99
3.6. Supporting Information 100
REFERENCES 102

CHAPTER 4 108
Deep Learning in Toxicity Prediction with One-Dimensional Similarity 108
4.1. Introduction 109
4.2. Method 111
4.2.1 Molecular Features 112
4.2.1.1. Chemical descriptors. 112
4.2.1.2. One-dimensional similarity. 112
4.2.2 Machine Learning 113
4.2.2.1. Support Vector Machines (SVM). 113
4.2.2.2. Random Forests (RF) 114
4.2.2.3. Deep Neural Networks (DNNs). 114

4.3. Results and Discussion 115
4.3.1. Benefit of One-Dimensional similarity 115
4.3.2. Comparison of DNN, RF and SVM 115
REFERENCES 117

CHAPTER 5 121
Conclusions and Future Outlook 121
REFERENCES 125

LIST OF TABLES

Table 2.1. List of proton atom types in HECSP. 28

Table 2.2. Local Atom Susceptibilities in the atomic axis frame in units of $4\pi\times10^{-12}$ m^3/mol 35

Table 2.3. Fitting statistics for the different parameter sets. 43

Table 2.4. Gasteiger_LOOCV NMRScore_Ps of the 44 experimentally determined apoNCS-naphthoate ester complex NMR models. 47

Table 2.5. List of protein proton types and corresponding parameters in Gasteiger_LS and Gasteiger_LOOCV parameter sets (except the ring current intensity factors). 57

Table 2.6. List of aromatic ring types and corresponding ring current intensity factors for the Gasteiger_LS and Gasteiger_LOOCV parameter sets. 57

Table 2.7. List of protein proton types and corresponding parameters in AM1-BCC_LS parameter set (except the ring current intensity factors). 58

Table 2.8. List of aromatic ring types and corresponding ring current intensity factors in the AM1-BCC_LS parameter set. 58

Table 3.1. Pearson's Correlation Coefficients, R^2 and Root Mean Square Errors (RMSEs) of three options of fitting to ZAFF bond stretching force constants for Zn-N, Zn-S, and Zn-O bond types. 75

Table 3.2. Twelve Zn-containing model systems considered in present study. 79

Table 3.3. RMSD values between EZAFF minimized, DFT optimized and CSD structures of six zinc containing organometallic compounds 88

Table 3.4. Mean unsigned errors (MUEs) of relative energies for 12 zinc complexes investigated 91

Table 3.5. RMSD values of the optimized structure by each method towards the DFT optimized geometry for 12 complexes 96

Table 3.6. RMSD values of the optimized structure by each method towards the crystal structure (PDB or CSD) for each of the 12 complexes 96

Table 3.7. MUEs and MEs of Zn-X distance values of the optimized structure by each method towards the crystal structure (PDB or CSD) and DFT optimized structure for 11 complexes (except 1Y9Q)..... 97

Table 3.8. Mean errors (MEs) of relative energies for 12 zinc complexes investigated..... 100

Table 3.9. RMSD values of the optimized structure by each method towards the DFT optimized geometry for 12 complexes..... 100

Table 3.10. RMSD values of the optimized structure by each method towards the crystal structure (PDB or CSD) for each of the 12 complexes..... 101

Table 4.1. AUC results for different learning methods and different input features for each task 116

Table 4.2. Average AUC results for different learning methods 116

LIST OF FIGURES

Figure 1.1. Confusion matrix 6

Figure 1.2. Schematic representation of a DNN 11

Figure 2.1. The ring current effect caused by an aromatic ring towards a target proton. The red sphere represents the target proton while the aromatic ring is shown in blue. The two smaller red dots are the center of the ring and the projection of the proton onto the plane of the aromatic ring. 31

Figure 2.2. The electric field effect caused by a polar atom. The red sphere represents a polar atom, which is the source of the electric field effect on the target proton shown in light grey. 31

Figure 2.3. (A). Workflow of ranking structures in NMR ensemble using NMRScore_P on the apoNCS-naphthoate ester complex (PDB: 1J5I). (B). Workflow of ligand structure refinement in the apoNCS-naphthoate ester complex (PDB: 1J5I) using NMRScore_P. The red dots in the scatterplots represent the original NMR models. (C). Workflow of ternary hIFABP-ketorolac-ANS complex structure determination with NMRScore_P. hIFABP is shown as a rosy brown ribbon and the locations that were observed to have significantly perturbed protons are mapped onto the hIFABP structure in orange (induced by ketorolac), blue (induced by ANS) or violet (induced by ketorolac and ANS binding at the same time). Ketorolac is shown as a blue stick whereas ANS is shown as a cyan stick. 38

Figure 2.4. Ligand structure 40

Figure 2.5. Correlation coefficients, RMSEs, and MUEs of the HECSP predictions for all proton CSPs along with the cutoff for electric field term. 42

Figure 2.6. Statistics of the predictions from the leave-one-out cross-validation of all protons in the corresponding complex structure. 44

Figure 2.7. Correlation between AF-QM/MM and HECSP calculated 1H CSPs over all the 44 protein-ligand solution structures in the ensemble (PDB:1J5I). 46

Figure 2.8. NMRScore_P and Rank for the 44 Experimentally Determined NMR Models (PDB:1J5I). The x axis shows the ranking of the solution structures predicted by corresponding parameter set of our method. (A). Comparison of NMRScore_P computed with Gasteiger_LOOCV and AF-QM/MM. (B). Comparison of NMRScore_P computed with Gasteiger_LS and AF-QM/MM. (C). Comparison of NMRScore_P computed with AM1-BCC_LS and AF-QM/MM. 48

Figure 2.9. Gasteiger_LOOCV NMRScore_P *vs* structural RMSD (Å) for corresponding models. The red dots represent the experimental NMR ligand structures (PDB: 1J5I) The blue dots represent the ligand conformers generated by rotating trihydroxy-cyclopentene moiety around two rotatable bonds. 49

Figure 2.10. NMR structures of ApoNCS-naphthoate ester complex (PDB 1J5I) and the refined ligand structures. The blue colored part is the experimentally determined ligand structure in all the small figures. The fragments in other colors demonstrate the refined trihydroxy-cyclopentene moiety. The numbers shown in each figure are the NMR model numbers. 51

Figure 2.11. (A). NMR structure of ApoNCS (PDB 1J5I model 6) together with refined ligand structure. It is the best ranked complex structure by NMRScore_P amongst all the NMR models and refined structures (as shown in yellow). (B). The gray counterpart is the best representative conformer in the experimental ensemble (PDB 1J5I model 1) 52

Figure 2.12. Gasteiger_LOOCV NMRScore_P vs structural RMSD (Å) of Glide docked poses. The red dots represent the experimental NMR ligand structures (PDB: 1J5I). The orange dots represent the docked poses. The blue dots represent the refined ligand structures. 53

Figure 2.13. NMRScore_Ps and rank for ternary hIFABP-ketorolac-ANS complex structures generated from IFD. The label for the x-axis is the original model number together with the pose number assigned by IFD. The top three structures are depicted with its ranking. hIFABP is shown as ribbon and the locations that were observed to have significantly perturbed protons are mapped onto the hIFABP structure in green. The red dots in the scatterplot represent the structures obtained by IFD of ANS into hIFABP-ketorolac complex, whereas, the blue dots represent the ones got by IFD of ketorolac into hIFABP-ANS complex. 55

Figure 2.14. hIFABP is shown as light gray ribbon and the locations that were observed to have significantly perturbed protons are mapped onto the hIFABP structure in green. The dim gray ligands in the center represent the best IDFScore poses of ketorolac and ANS. Whereas, the orange ligands are the best NMRScore_P poses. 59

Figure 2.15. Best NMRScore_P ranked ternary hIFABP-ketorolac-ANS complex structure generated from IFD of ketorolac into the hIFABP-ANS complex. 59

Figure 3.1. Fits of the ZAFF bond stretching force constants for the Zn-N, Zn-S, and Zn-O bond types. 75

Figure 3.2. Scatter plot of the bending force constant vs equilibrium bond angle for the Zn containing bond. As shown in the legend, X can be any atom and "X-Zn-X" doesn't require that the two Xs are the same atom type. 76

Figure 3.3. RMSD values of the protein backbone, metal binding site and binding atoms monitored along the 20 ns MD trajectory for each metalloprotein investigated (left), and the last snapshot from each trajectory (right). In the plots to the left, the black, red and blue curves represent backbone, metal binding site and binding atoms respectively. 85

Figure 3.4. Superimposition of the EZAFF optimized, DFT optimized, and CSD structures for 5 compounds taken from the CSD. These structures are shown in gray, green and yellow, respectively..... 87

Figure 3.5. Vibrational frequencies calculated based on DFT (blue) and the EZAFF model (green). In each plot, the normal modes are arranged in the order of their vibrational frequencies calculated by DFT as shown in the x-axis..... 89

Figure 3.6. B3LYP relative energies vs. heavy atom RMSDs over all 20 conformers for all the 12 test systems. The RMSDs were calculated with respect to the lowest energy conformer. The corresponding superposition is for the lowest single point energy conformer (gray) and the highest RMSD one (green or yellow)..... 90

Figure 3.7. Correlation of B3LYP relative energies vs. relative energies with different methods over all 20 conformers for all the 12 test systems. Relative energies were computed relative to the values of the conformers with the lowest B3LYP single point energy. The Pearson's correlation coefficients are given for each plot..... 94

Figure 3.8. Zn-X distance values from 1 ns gas-phase MD simulations of CSD complexes. The title for each plot includes the complex name and the nonbonded parameter set. Different colors represents different Zn-X interactions (ABOWOF: black, red, blue, green, yellow correspond to Zn-N and 4×Zn-O; AHOQIY: black, red, blue, green correspond to Zn-S, Zn-N, Zn-S and Zn-N; BEZKOH: black, red, blue, green all represent Zn-N; EGIXOH: black, red, blue, green, yellow, brown correspond to 2×Zn-O, 2×Zn-N and 2×Zn-S; KUBVOT: black, red, blue, green correspond to Zn-S, Zn-N, Zn-S and Zn-N; ZNTPBZ: black, red, blue, green all represent Zn-S)..... 98

Figure 4.1. Workflow overview..... 112

KEY TO ABBREVIATIONS

ML	machine learning
DL	deep learning
RF	Random Forest
SVM	Support Vector Machine
DNN	Deep Neural Network
CSP	chemical shift perturbation
CS	chemical shift
AIR	Ambiguous Intermolecular Restraints
HECSP	1H empirical chemical shift perturbation
SAR	structure-activity relationship
NMR	nuclear magnetic resonance
HSQC	heteronuclear single quantum coherence
NOE	nuclear Overhauser effect
apoNCS	apo-Neocarzinostatin
hIFABP	human intestinal fatty acid binding protein
RMSE	root-mean-square error
RMSD	root-mean-square deviation
SEM	standard error of mean
LS	least-squares
LOOCV	leave-one-out cross-validation
AF-QM/MM	automated fragmentation quantum mechanics/molecular mechanics

IFD	induced fit docking

CHAPTER 1

Introduction

1.1. Introduction to Machine Learning

Over the past decades machine learning has become one of the primary domains of information technology and statistical modeling. In recent years, a number of open source and high-quality software packages for implementing machine learning algorithms are available. Moreover, with the increasing amounts of data and the natural fit between machine learning and computational chemistry becomes apparent. In the case of cheminformatics and the pharmaceutical sciences, machine learning techniques are being used ever more widely.

Specifically, the emphasis of machine learning in computational chemistry is primarily concerned with supervised learning, which refers to predicting output variables (also known as dependent variables, response and labels) for a certain test case by learning a model from a training set with an associated response. The training set often contains chemical descriptors, observed physical or chemical properties as input variables (also known as independent variables, features and predictors) for different small molecules or biomolecules in a given dataset. The category of the problem is determined by the characteristics of the output variable. If output variables take on numerical values, these problems are defined as regression problems, whereas, for qualitative categorical output variables, they are defined as classification problems. The following parts in this chapter will discuss several machine learning and deep learning models and important details about each of them[1] and their applications in the field of drug discovery in particular. The second and third chapters are two examples of applying regression analysis to solve problems in protein-ligand complex structure prediction and zinc force field development. Whereas the fourth chapter is about using popular machine learning techniques including deep neural network in drug toxicity prediction.

1.2. Linear Regression

Linear regression is a very simple supervised learning model to predict a numerical response. Although it seems to be simplistic relative to more complex models, linear regression is still useful and widely used. It can help address some important questions: What are the predictors that help to explain the output variable? How well does the model fit the data? How accurate is the estimation of the effect of each feature and the predicted value?

Simple linear regression assumes that there is a linear relationship between X and Y which has the functional form of

$$Y = \beta_\circ + \beta_1 X + \epsilon. \tag{1.1}$$

Herein, $\beta_\circ$ is the intercept, β_1 is the slope and ϵ is the random error term, which is usually assumed to be independent of X. However, most problems involve multiple predictors. In these cases, the slope coefficient for each predictor is given. So that the functional form of multiple linear regression with p predictors is

$$Y = \beta_\circ + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \tag{1.2}$$

where X_j represents the jth predictor and β_j represents the coefficient of jth predictor. All the β_j are parameters of the multiple linear regression model. By minimizing the sum of square residuals (RSS), one can estimate the regression coefficients which are represented as $\hat{\beta}_\circ,\ \hat{\beta}_1, \ldots,\ \hat{\beta}_p$.

Once a model is fit, one can use common metrics to validate the model, including residual standard error (RSE), the fraction of variance explained (R^2) and the correlation between the response and the fitted model ($Cor(Y, \hat{Y})$). Besides these statistics, one should also plot the data to spot issues relating to nonlinearity of the response-predictor relationships, non-constant variance of the error

term, outliers and high-leverage points. In the case of non-linear relationships, the linear model can be extended to address this issue by including transformed versions of the predictors in the model (*e.g.* polynomial regression models).

1.2.1. Linear Model Selection and Regularization

As mentioned above, both simple and multiple linear regression models are fitted using least squares. However, there are some alternative fitting strategies which may have better prediction accuracy and interpretability. The first one is called subset selection, which refers to identifying a subset of features and then fitting a model based on the selected features using least squares. Another approach is regularization, which differs from linear regression in that regularization adds constraints to the coefficient estimates. Regularization can be further divided into ridge and lasso.

1.2.2. Ridge Regularization

In the least squares fitting procedure, the coefficients are estimated by minimizing RSS as the objective function. Whereas, a penalty term is added to RSS and the final objective function is given as:

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{1.3}$$

where λ is a parameter and helps equation 1.3 to balance between fitting the data well and keeping the coefficient estimates as closer to zero as possible. The optimal λ is usually determined by cross-validation studies.

In the setting of least squares estimation, especially when the number of predictors is close to the number of data points, the estimated parameters may have high variance, which means $\hat{\beta}_j$ will significantly change as the training data changes. Then ridge regression helps address this because it can decrease the variance by accepting a small increase in bias.

1.2.3. Lasso Regularization

One feature about ridge regularization is that all the coefficients will remain non-zero, so that all the features will be included in the trained model. This can be a disadvantage for ridge regression, especially for cases where the number of features is large. On the other hand, lasso regularization, can perform feature selection by setting coefficients to zero. The objective function for lasso regression is

$$\text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{1.4}$$

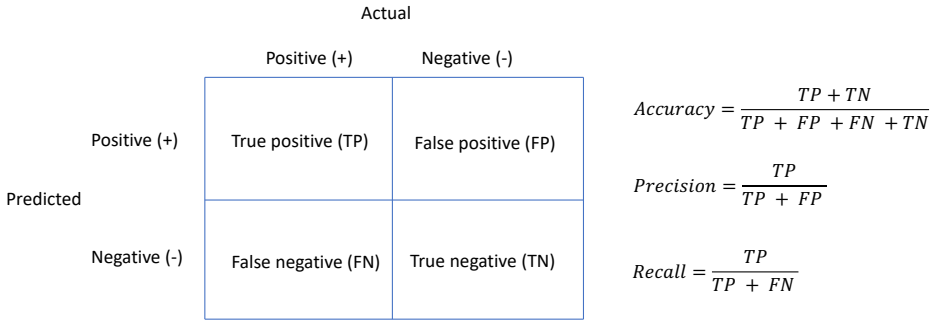
1.3. Random Forests

Random forest is an ensemble supervised learning approach built on the basis of a decision tree. Decision trees can be used in both regression and classification settings. We will consider classification problems herein.

In general, the common performance metrics for classification problems are accuracy, precision, recall and area under the curve (AUC). Usually, a classification model would output a probability of an instance being positive. Then a threshold is chosen to label the instance as positive or negative based on the predicted probability. Once the threshold is decided, you will have a classifier which results in a confusion matrix as shown in Figure 1.1. Accuracy, precision and recall are built on

the basis of the confusion matrix. Whereas, AUC is the area under the receiver operating characteristic curve and it is an estimate of the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance. An ideal model whose prediction is 100% correct has AUC of 1, whereas a model which fails to discriminate has a value of 0.5.

Figure 1.1. Confusion matrix



With the performance metrics defined, I can start explaining how classification models work. Basically, what a decision tree does is segment the feature space into several regions and predicts a new case by the region in which it falls. How are the several regions to be divided? It is done via growing a tree from the single root node, recursively splitting at each node and its subsequent child nodes and stopping when reaching a terminal node that has fewer data points than a predetermined number. The whole process of recursive binary splitting starts from the top of the tree and continues by continuously splitting feature space. Each split is determined by seeking the specific

predictor and associated cutting point with which the sum of the classification error rate in the two child nodes is minimized. Classification error rate is the fraction of the training observations in that child node that are not the majority class. In practice, there are two other metrics that are commonly used and preferred: the Gini index and cross-entropy:

$$\mathrm{G} = \sum_{k=1}^{K} \hat{p}_k \, (1 - \hat{p}_k) \tag{1.5}$$

and

$$\mathrm{D} = -\sum_{k=1}^{K} \hat{p}_k \, log\hat{p}_k, \tag{1.6}$$

where $\hat{p}_k$ is the fraction of training observations in the node that are from the class k. As for both metrics, when $\hat{p}_k$ approaches 0 or 1, the Gini index and cross-entropy both become smaller. So that the smaller G or D is, it means the split is of a better quality.

Decision trees have many advantages including they are easy to explain and to be displayed graphically. It is mimicking a logical human decision-making process. However, its predictive performance can be further improved by aggregating many decision tree models like random forest.

Random forest is built on a number of decision trees, with each decision tree trained on a random subset of training set and each split chosen among a random subset of predictors. Through this process, all of the trees will be decorrelated and have the chance to consider various features, so that the average prediction among the trees have smaller variance and is more reliable. To be specific, it is typical to consider the square root of the total number of predictors at each split.

Since random forests consist of a large number of trees, it is not reasonable to graphically display all the splitting criteria. So that model interpretability is sacrificed for better prediction performance. However, we can still explore the contribution of features by estimating their importance. In random forests, the total amount of Gini index reduction due to splitting associated with a given feature normalized by the number of trees can be used to measure variable importance. A larger value means more importance.

1.4. Support Vector Machines (SVMs)

Suppose the training data we have contains n observations with a p-dimensional feature space which means X is a n × P matrix. The labels of these observations are either +1 or -1. The ideal scenario is that there are many $(\mathrm{p}-1)$ dimensional hyperplanes that can perfectly separate these two classes. The question is which one is the best. What the maximal margin classifier does is that it uses the maximal margin hyperplane as the final separating hyperplane, where "margin" refers to the smallest distance from a training data point to the hyperplane. The resultant hyperplane maximizing the margin is defined as

$$\sum_{j=1}^{p} \beta_j^2 = 1 \tag{1.7}$$

$$\beta_\circ + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0. \tag{1.8}$$

Since $y_i \in (-1,1)$, then $y_i(\beta_\circ + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) > 0$ stands for the training data. For any new test observation $X^* = (X_1^*, X_2^*, \ldots X_p^*)^T$, the sign of $\beta_\circ + \beta_1 X_1^* + \beta_2 X_2^* + \cdots + \beta_p X_p^*$ determines its label. However, typically there is no "perfect" hyperplane, so in such cases, we need

to allow some misclassification in the training data. The support vector classifier is what is needed to do this task.

Support vector classifiers (SVCs) are also known as soft margin classifiers. It is allowed that some training data points can be on the wrong side of the margin or even on the wrong side of the hyperplane. The optimization is similar to a maximal margin classifier in that it also estimates parameters by maximizing the width of the margin (M) and with following restraints:

$$\sum_{j=1}^{p} \beta_j^2 = 1 \tag{1.9}$$

$$y_i(\beta_\circ + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) > M(1 - \epsilon_i) \tag{1.10}$$

$$\epsilon_i \geq 0 \text{ and } \sum_{i=1}^{n} \epsilon_i \leq C \tag{1.11}$$

Herein ϵ_i is a slack variable for each observation, which indicates where the ith data point is located relative to the margin and hyperplane. And C is a nonnegative "tuning" parameter, which sets a limit to the sum of all the slack variables. The larger the C, the wider the margin and more observations lie at the wrong side of the margin. "Support vectors" refers to those observations that are either on the margin or on the wrong side of the margin, since only these data points really have the power to determine the hyperplane.

Basically, what SVM does is it automatically transforms a linear decision boundary into a non-linear one. It is done using kernels that the SVM uses to enlarge the feature space on the basis of SVC. It turns out that the solution of SVC involves the inner products of the observations, which can be generalized as a kernel $K(x_i, x_{i'})$. So that the classifier function can be written as:

$$\beta_{\circ} + \sum_{\substack{support \\ vectors}} \alpha_i \, K(x_i, x). \tag{1.12}$$

where the polynomial kernel is given as:

$$K(x_i, x_{i'}) = (1 + \sum_j x_{ij} x_{i'j})^d \tag{1.13}$$

where d is the degree of the polynomial. Another popular kernel is the radial kernel, where

$$K(x_i, x_{i'}) = exp\left(-\gamma \sum_j \left(x_{ij} - x_{i'j}\right)^2\right). \tag{1.14}$$

γ is a positive constant.

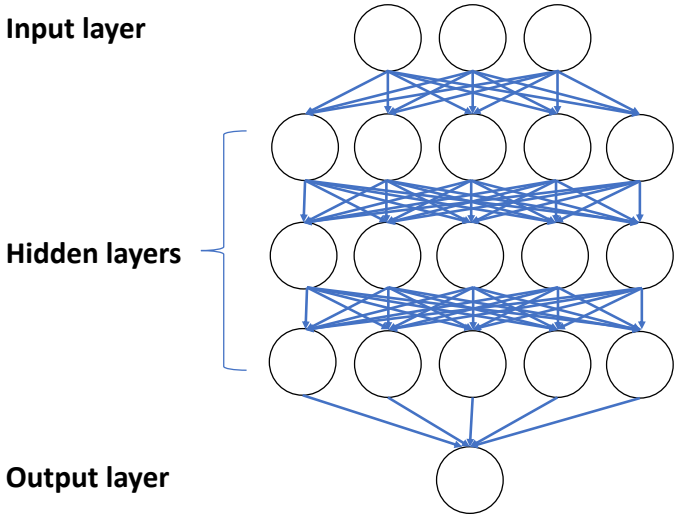
The advantages of the kernel in SVM are two-fold: firstly, it enlarges the feature space by changing the linear function into a non-linear one. The other is the computational efficiency of the kernel-based approach. We need only compute n(n-1)/2 pairs of $K(x_i, x_{i'})$ instead of literally projecting the features into a higher order space, which in some cases is impossible.

1.5. Deep Neural Networks

Deep neural networks (DNNs) are different from artificial neural networks (ANNs) by the depth, which refers to the number of hidden layers the data flows through. A DNN usually has multiple hidden layers, through which a large input vector can be mapped to a target output value or vector. Figure 1.2 provides a schematic representation of a DNN. A DNN is built with a series of layers of neurons (potentially thousands of them[2]) so that all possible facets of input information can be extracted.[3] Each neuron takes multiple activation values from the previous layer of neurons. To be specific, the activation value of a neuron is computed by subjecting the weighted sum of activation values of all neurons in the previous layer plus a bias term to an activation function. The activation

function determines whether a neuron should be activated, by which non-linearity is introduced into the output of a neuron. To be concise, the bias term in each hidden layer is omitted in Figure 1.2.

Figure 1.2. Schematic representation of a DNN



1.5.1. Activation Functions

There are three common activation functions in DNN: the sigmoid, tanh and rectified linear unit (ReLU). Sigmoid functions are non-linear, smooth function ($\mathrm{sigmoid}(x) = \frac{1}{1+e^{-x}}$). Since the curve is steep around x=0, it tends to bring the activation value close to either end of the range (zero or one), making a clear separation on predicted values. But there is a problem associated with it, which is vanishing gradients. It is because when y approaches either end as x gets further from zero the gradient gets smaller until it hits the floating point value limit. In such case, the DNN ends the learning process.

Another popular activation function is the tanh function ($\tanh(x) = \frac{2}{1+e^{-2x}} - 1$), which is a scaled sigmoid function with a range from -1 to 1 ($\tanh(x) = 2\text{sigmoid}(2x) - 1$). Making the tanh function similar to the sigmoid function in nature. The tanh function also suffers from a vanishing gradient.

In recent years, both the sigmoid and tanh functions have been superseded by the ReLU[4,5] ($\text{ReLU}(x) = \max(0, x)$). The ReLU function is zero when x<0 and linear when x>0. Due to its functional form, a few neurons in the DNN will not activate, so that the ReLU has the advantage of sparse activation unlike the sigmoid and tanh functions. However, it is not always a benefit, because if the activation value is zero for negative values, its gradient can be zero. This problem can cause several neurons to die and not respond. By using leaky ReLU, the horizontal part of the function would have a slight slope, so that dead neurons can recover during the learning process.

1.5.2. Dropout Regularization

Besides vanishing gradients, overfitting is also an issue in DNN. In DNN, the weights of neurons are trained for specific features and they tend to rely on the specialization of nearby neurons. This property can result in an overfit model is too biased towards the training data. Dropout is a regularization for DNN, which can be a remedy to overfitting.[6] What dropout does is that a certain proportion of neurons in the DNN architecture will be dropped out randomly in each epoch of the training. Since some neurons are ignored, other neurons will have to substitute for the dropped-out neurons in the prediction. Dropout regularization can result in a DNN less sensitive to the specific weights of the neurons making the model more generalized.

1.6. Machine Learning in Drug Discovery

Unlike physical models (quantum chemistry, molecular mechanics) which depend on explicit mathematical equations, machine learning techniques automates systems to learn from data and identify patterns with minimal human intervention. Also, machine learning can be easily scaled up to handle large datasets and require much less computational resources. Because of the complexity of biological systems and difficulty to identify all the relevant variables, machine learning is an alternative and has outperformed physical models.[7]

Machine learning has been increasingly applied in computational chemistry, specifically in the field of drug discovery.[8–11] One of the primary interests of researchers is to build the bridge between structural information and biological or chemical activities, which is referred to as structure-activity relationship (SAR). SAR can provide insights to optimize the binding affinity or other physiochemical properties of hit compounds discovered by screening. Traditionally, SAR was studied through cycles of time-consuming, expensive experiments. Quantitative structure-activity relationship (QSAR) can also be modeled with the help of machine learning. QSAR techniques have been utilized to predict how biological behavior changes with chemical modifications and to model the properties of drug molecules including toxicity, intermolecular interactions, and carcinogenesis.[12] The first application of machine learning using multivariate linear regression to QSAR modeling was carried out in the 1960s by Corwin Hansch[13]. To combat multicollinearity and the curse of high dimensionality in regression analysis, regularization and dimensionality reduction were carried out.[14–17] Due to the assumption of linearity of the underlying distribution, linear regression is sometimes not enough to tackle the complexity of QSAR tasks.

Support vector machines (SVMs) are another widely used method in various modeling problems in drug discovery, especially QSAR.[18,19] Basically, SVM solves these problems by mapping data sets into a high-dimension feature space with kernels and identifying a separating hyperplane, which maximizes the margin.

Besides SVMs, decision trees are a method of transparency and interpretability. In QSAR modeling, a molecular attribute is selected for each binary splitting, each leaf node resulting from a series of splitting represents a label in a classification problem. Decision trees have been utilized to model oral absorption properties and toxicity of drug molecules in recent years.[20,21]

It has been shown that standard decision trees can be improved by ensemble techniques like bagging and boosting. Random forest is an ensemble method built by applying the bagging idea on a number of decision trees, with each decision tree trained on a random subset of the training data and each split chosen among a random subset of input features. Random forest is a widely used machine learning technique, which has been implemented to classify bioactivity[22], toxicity[23], predict binding affinity[24] and identify human drug targets[25].

1.7. Deep Learning in Drug Discovery

As mentioned above, various machine learning approaches have been implemented in drug discovery. In the last decade, deep learning (DL) also has boomed in this field due to new solutions to the overfitting problem, algorithm development and improvements on contemporary computer hardware.

In the realm of drug discovery, notably in QSAR, the popular machine learning approaches discussed previously have been around for a long time. Among DL algorithms, the most straightforward one is the fully connected DNNs, which has been shown to perform slightly better than random forest on the Merck Kaggle challenge dataset.[26] It was learned through the challenge that DNNs can take in thousands of input variables without feature selection. The performance of DNN can be improved by hyper-parameter tuning including number of layers, number of nodes per layer and type of activation functions. Based on the winning algorithm used in the Merck challenge[26], Dahl et al.[27] further explored that multitask DNN gave a more effective performance than single task DNN. Herein, multitask means a model predicts multiple outputs at the same time so that a commonly shared feature extraction pipeline across different tasks is learned out of it. Similarly, Mayr et al.[28] won the Tox21 challenge with multitask DNN models which again demonstrate its advantage over single task DNNs. The advantages of multitask DNN models are rooted in the fact that they share multilabel information and utilize relations between tasks. These are especially important for task that have fewer training examples. There are other benchmark studies[26,29–31] showing that DNN models can easily achieve a better performance over some of the traditional machine learning approaches.

Besides benchmark studies of DNN, DNN has been used for toxicity prediction. Recently, Xu *et al*. did a toxicological study of drug-induced liver injury and trained DNNs that gave an AUC (area under the curve) of 0.955 and found out that undirected graph recursive neural networks method is an effective molecular encoding method.[32] It is suggested in this study that since DNNs are able to extract information on its own from the input variables, a good molecular encoding method like UGRNN may be even better than explicit molecular descriptors. Hughes *et al*. proposed a deep convolution network model to predict site of epoxidation (SOE) in drug molecules which is

another important application in toxicity modeling.[33] SOE is the site of a molecule that undergoes epoxidation with cytochrome P450s that results in toxic electrophilic reactive metabolites. With this model, it is possible to identify potential adverse effects related to reactive metabolites and modify the molecule to prevent epoxidation. More recently, a couple of studies were published using DNN models to tackle Tox21 challenge. The database consists of 12000 compounds and their assay test results on 12 different targets. As is mentioned above, Mayr et al.[28] developed the winning model called DeepTox which was a multitask DNN model. It outperformed single task models in 10 out of the 12 targets.

The impact of deep learning in drug discovery is not confined to QSAR and toxicity prediction. DNN models have been used to predict solubility of drug-like molecules by Baldi et al.[34] and to predict ADMET properties by Pande and coworkers[35]. Moreover, AtomNet[36], which is a deep leaning model was developed to predict new molecules with bioactivity for specific binding sites. Several benchmarks show that AtomNet performs better than some docking methods with a margin of 0.2 in the AUC score.

To conclude, deep learning is different from traditional machine learning techniques since it is implemented through a hierarchical cascade of nonlinear functions. Deep learning has significantly impacted the field of computer vision and speech recognition due to technological breakthroughs, and the growth of data and scientific computing power. Deep learning started to gain attention more recently in computational chemistry and biology and have already provided satisfying performance in many subfields of computational chemistry and biology over traditional machine learning algorithms.

REFERENCES

REFERENCES

1. James, G., Witten, D., Hastie, T., and Tibshirani, R. An Introduction to Statistical Learning.

2. Cireşan, D., Meier, U., and Schmidhuber, J. (2012) Multi-column Deep Neural Networks for Image Classification. *arXiv:1202.2745*.

3. Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2015) DeepTox: Toxicity Prediction using Deep Learning. *Front. Environ. Sci.*

4. Glorot, X., Bordes, A., and Bengio, Y. (2011) Deep sparse rectifier neural networks. *AISTATS '11 Proc. 14th Int. Conf. Artif. Intell. Stat.*

5. Nair, V., and Hinton, G. E. (2010) Rectified Linear Units Improve Restricted Boltzmann Machines. *Proc. 27th Int. Conf. Mach. Learn.*

6. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res. 15*, 1929–1958.

7. Hochreiter, S., Klambauer, G., and Rarey, M. (2018) Machine Learning in Drug Discovery. *J. Chem. Inf. Model. 58*, 1723–1724.

8. Lo, Y. C., Rensi, S. E., Torng, W., and Altman, R. B. (2018) Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today*.

9. Zhang, L., Tan, J., Han, D., and Zhu, H. (2017) From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discov. Today*.

10. Lavecchia, A. (2015) Machine-learning approaches in drug discovery: Methods and applications. *Drug Discov. Today*.

11. Lima, A. N., Philot, E. A., Trossini, G. H. G., Scott, L. P. B., Maltarollo, V. G., and Honorio, K. M. (2016) Use of machine learning approaches for novel drug discovery. *Expert Opin. Drug Discov.*

12. Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., Consonni, V., Kuz'min, V. E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., and Tropsha, A. (2014) QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem. 57*, 4977–5010.

13. Hansch, C., Maloney, P. P., Fujita, T., and Muir, R. M. (1962) Correlation of Biological

Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature 194*, 178–180.

14. Frank, L. E., and Friedman, J. H. (1993) A statistical view of some chemometrics regression tools. *Technometrics*.

15. Hoerl, A. E., and Kennard, R. W. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*.

16 Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., and Aziz, M. (2015) High-dimensional QSAR prediction of anticancer potency of imidazo[4,5-b]pyridine derivatives using adjusted adaptive LASSO. *J. Chemom. 29*, 547–556.

17. Rensi, S. E., and Altman, R. B. (2017) Shallow Representation Learning via Kernel PCA Improves QSAR Modelability. *J. Chem. Inf. Model.*

18. Nekoei, M., Mohammadhosseini, M., and Pourbasheer, E. (2015) QSAR study of VEGFR-2 inhibitors by using genetic algorithm-multiple linear regressions (GA-MLR) and genetic algorithm-support vector machine (GA-SVM): a comparative approach. *Med. Chem. Res. 24*, 3037–3046.

19. Liu, H. X., Zhang, R. S., Yao, X. J., Liu, M. C., Hu, Z. D., and Fan, B. T. (2003) QSAR study of ethyl 2-[(3-methyl-2,5-dioxo(3-pyrrolinyl))amino]-4-(trifluoromethyl) pyrimidine-5-carboxylate: An inhibitor of Ap-1 and NF-/κB mediated gene expression based on support vector machines. *J. Chem. Inf. Comput. Sci. 43*, 1288–1296.

20. Newby, D., Freitas, A. A., and Ghafourian, T. (2015) Decision trees to characterise the roles of permeability and solubility on the prediction of oral absorption. *Eur. J. Med. Chem.*

21. Gupta, S., Basant, N., and Singh, K. P. (2015) Estimating sensory irritation potency of volatile organic chemicals using QSARs based on decision tree methods for regulatory purpose. *Ecotoxicology*.

22. Singh, H., Singh, S., Singla, D., Agarwal, S. M., and Raghava, G. P. S. (2015) QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest. *Biol. Direct*.

23. Mistry, P., Neagu, D., Trundle, P. R., and Vessey, J. D. (2016) Using random forest and decision tree models for a new vehicle prediction approach in computational toxicology. *Soft Comput.*

24. Wang, Y., Guo, Y., Kuang, Q., Pu, X., Ji, Y., Zhang, Z., and Li, M. (2015) A comparative study of family-specific protein-ligand complex affinity prediction based on random forest approach. *J. Comput. Aided. Mol. Des.*

25. Kumari, P., Nath, A., and Chaube, R. (2015) Identification of human drug targets using

machine-learning algorithms. *Comput. Biol. Med.*

26. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., and Svetnik, V. (2015) Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model. 55*, 263–274.

27. Dahl, G. E., Jaitly, N., and Salakhutdinov, R. (2014) Multi-task Neural Networks for QSAR Predictions. *arXiv:1406.1231*.

28. Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016) DeepTox: Toxicity Prediction using Deep Learning. *Front. Environ. Sci. 3*, 80.

29. Rogers, D., and Hahn, M. (2010) Extended-Connectivity Fingerprints. *J. Chem. Inf. Model. 50*, 742–754.

30. Koutsoukas, A., Monaghan, K. J., Li, X., and Huan, J. (2017) Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *J. Cheminform. 9*, 42.

31. Lenselink, E. B., ten Dijke, N., Bongers, B., Papadatos, G., van Vlijmen, H. W. T., Kowalczyk, W., IJzerman, A. P., and van Westen, G. J. P. (2017) Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminform. 9*, 45.

32. Xu, Y., Dai, Z., Chen, F., Gao, S., Pei, J., and Lai, L. (2015) Deep Learning for Drug-Induced Liver Injury. *J. Chem. Inf. Model. 55*, 2085–2093.

33. Hughes, T. B., Miller, G. P., and Swamidass, S. J. (2015) Modeling Epoxidation of Drug-like Molecules with a Deep Machine Learning Network. *ACS Cent. Sci. 1*, 168–80.

34. Lusci, A., Pollastri, G., and Baldi, P. (2013) Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *J. Chem. Inf. Model. 53*, 1563–1575.

35. Kearnes, S., Goldman, B., and Pande, V. (2016) Modeling Industrial ADMET Data with Multitask Networks. *arXiv:1606.08793*.

36. Wallach, I., Dzamba, M., and Heifets, A. (2015) AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. *arXiv:1510.02855*.

CHAPTER 2

Using Ligand Induced Protein Chemical Shift Perturbations to Determine Protein-ligand Structures

† Reprinted (adapted) with permission from Yu, Z., Li, P. & Merz, K. M. Using Ligand-Induced Protein Chemical Shift Perturbations To Determine Protein–Ligand Structures. Biochemistry 56, 2349–2362 (2017).

2.1. Abstract

Protein chemical shift perturbations (CSPs), upon ligand binding, can be used to refine the structure of a protein-ligand complex by comparing experimental CSPs with calculated CSPs for any given set of structural coordinates. Herein we describe a fast and accurate methodology that opens up new opportunities to improve the quality of protein-ligand complexes using NMR based approaches by focusing on the effect of the ligand on the protein. The new computational approach, ^{1}H empirical chemical shift perturbation (HECSP), has been developed to rapidly calculate ligand binding induced ^{1}H CSPs in a protein. Given the dearth of experimental information by which a model could be derived we employed high-quality DFT computations using the automated fragmentation quantum mechanics/molecular mechanics (AF-QM/MM) approach to derive a database of ligand induced CSPs on a series of protein-ligand complexes. Overall, the empirical HECSP model yielded correlation coefficients between its predicted and DFT computed values of 0.897 (^{1}HA), 0.971 (^{1}HN) and 0.945 (sidechain ^{1}H) with root-mean-square errors (RMSEs) of 0.151 (^{1}HA), 0.199 (^{1}HN) and 0.257 ppm (sidechain ^{1}H), respectively. Using the HECSP model, we developed a scoring function (NMRScore_P). We describe two applications of NMRScore_P on two complex systems and demonstrate that the method can distinguish native ligand poses from decoys and refine protein-ligand complex structures. We provide further refined models for both complexes, which satisfy the observed ^{1}H CSPs in experiments. In conclusion, HECSP coupled with NMRScore_P provides an accurate and rapid platform by which protein-ligand complexes can be refined using NMR derived information.

2.2. Introduction

Chemical shift perturbation (CSP) represents the chemical shift change due to perturbation of the chemical environment and is a widely used technique to study protein-ligand interaction. The use of CSP information has attracted significant attention because of the introduction of the "SAR (structure-activity relationship) by NMR" method.[1] One of the advantages of CSP over other common techniques is that it is very sensitive to changes in the chemical environment and can be efficiently measured. Experimentally, the CSPs of a target protein can be recorded as a series of 2D heteronuclear single quantum coherence spectroscopy (HSQC) spectra via the titration of a ligand into a solution containing a ^{15}N-labelled protein, from which one can determine the binding site position or "pose" and the dissociation constant.[2] Moreover, no signal is observed if binding does not occur, which makes it a powerful tool for selecting hits via high-throughput screening.[3]

There have been a number of efforts made to improve the performance of *in silico* docking through the use of CSP information. Some have been implemented in a qualitative way by imposing a cutoff CSP value to determine significant changes[4, 5] and to then dock the ligand into the region with significant CSPs with largely ambiguous distance restraints[6-8]. HADDOCK is one of the most popular programs in this category, which first determines the binding interface (both active and passive residues are defined based on the magnitude of the observed CSP and solvent accessibility). Then the ligand is docked into the interface with so-called Ambiguous Intermolecular Restraints (AIRs) set up amongst the atoms in the active and passive residues. CSP values have also been incorporated into the BiGGER program [9] as a post-docking filter by Morelli *et al.*[10, 11]. AutoDockFilter[5] uses a CSP based scoring function to perform pose ranking after docking calculations. It is also possible to study binding modes and structure from CSP information. A

method was described using differential chemical shifts to map out the protein binding pocket and to determine the binding poses of a series of related ligands to a target protein.[12] Lugovskoy *et al.*[13] have further applied this method and successfully determined the structure-activity relationships of the BH3Is/Bcl-xL complexes. Even when related ligands do not share the same binding mode, by analyzing residue-wise CSP pattern, Riedinger *et al.*[14] determined binding mode "clusters" for isoindolinone inhibitors.

In light of the success of using chemical shifts to determine protein structure,[15-17] a number of empirical approaches have been developed to perform quantitative predictions of CSP values and to predict the structure of protein-ligand complexes. McCoy and Wyss using the aromatic ring effect coupled with Pople's equation to approximate the CSP induced by aromatic ring currents in a ligand they developed the "j-surface" method to locate aromatic rings.[2, 18] Cioffi *et al.* were able to perform protein-ligand structure refinement based on the correlation of experimental and semi-empirically calculated ^{1}HN CSP values.[19-21] Following the same strategy of comparing simulated and experimental CSPs, several attempts have been made in recent years to determine protein-ligand complex structure.[22-24] Quantum chemical[25-33], QM/MM methods[30, 34-39], the fragment based adjustable density matrix assembler (ADMA) method[40-42] and the fragment molecular orbital (FMO) method[43, 44] are all available for protein chemical shift calculation if protein is properly "parsed". Unlike empirical methods, which are confined by the training set, these theoretical methods can be readily applied to biomolecules containing ligands or other non-standard residues and directly extended to CSP calculation, but at an increased cost relative to empirical methods.

There are many empirical programs to efficiently calculate chemical shifts (CS) of proteins including ShiftS,[45-47] ShiftX,[48] ShiftX2,[49] Sparta+,[50, 51] CamShift,[52] PROSHIFT,[53, 54] SHIFTCALC,[55] *etc*. In this work, we proposed an empirical model (HECSP) to calculate ^{1}H NMR CSPs induced by ligand binding, which is an extension of previous approaches for protein chemical shift calculation (see equation 1)[45, 48]. Herein the conformation of the protein is rigid and the covalent connectivity of the protein is conserved during the binding process. Using this structural approximation, we calculate the ^{1}H CSPs inside the protein based on four nonlocal contributions (see equation 1):

$$\Delta\delta_H = \Delta\delta_{RC} + \Delta\delta_{EF} + \Delta\delta_{HB} + \Delta\delta_M \quad (1)$$

These four terms represent the contributions of the ring current, electric field, hydrogen bonding and magnetic anisotropy, respectively. The form of each term is described in the Methods section.

Although a considerable amount of experimentally determined CSPs and NMR structures are available, it is still difficult to extract the CSPs which are purely induced by ligand binding. This is because that the exact orientation or pose of the ligand is hard to match with the observed CSP values and the experimental CSPs arise due to the averaging over a number of factors including protein conformational changes.[56] In view of this, the target CSP values we used were calculated with the automated fragmentation QM/MM (AF-QM/MM) approach. As described earlier, using the B3LYP/6-31G** level of theory for the QM region and AMBER ff94 partial charges for the MM region, the AF-QM/MM approach gave root-mean-square errors (RMSEs) of less than 0.08 ppm and correlation coefficients higher than 0.95 with respect to ^{1}H NMR chemical shifts obtained by experiment.[34] The general parameterization process is given in the Methods section. Although

the current formulation (equation 1) is not physically complete, the results show that it accurately and rapidly predicts ^{1}H CSPs.

To further demonstrate the accuracy and utility of HECSP for protein ^{1}H CSP calculation, we applied this approach to two protein-ligand complexes: firstly, to an *apo*-Neocarzinostatin (*apo*NCS)-naphthoate ester complex (PDB: 1J5I). Excellent agreement between the AF-QM/MM and HECSP calculated ^{1}H CSPs were obtained over all the 44 protein-ligand solution structures in the ensemble (see discussion below). Wang *et al.* [57] developed a NMR scoring function based on ligand CSPs computed at the semiempirical level of theory that can readily rank protein-ligand complex. To study protein-protein systems, CS-HADDOCK[58] was developed to determine complex structures using CS-RMSD (RMSD between empirically calculated CSs and observed CSs) to score the docked complexes generated with HADDOCK CSP-AIRs. Similarly, we defined a scoring function NMRScore_P for protein-ligand systems, which is the RMSD between calculated protein ^{1}H CSPs and observed values. Based on HECSP calculated protein ^{1}H CSPs, NMRScore_P was able to rank the models in the structural ensemble and distinguish the native ligand pose from a set of decoys generated by Glide[59, 60]. We conclude that the HECSP derived protein ^{1}H CSPs used to form NMRScore_P is a good score function for the evaluation of protein-ligand complex structures. HECSP derived NMRScore_P can also be applied to complex structure determination between a protein and multiple ligands bound simultaneously. By simply adding the CSP contribution from each ligand, we can compute the overall CSP for the target protein protons induced by multiple ligand binding. We have explored this idea using HECSP derived NMRScore_P to determine the ternary human intestinal fatty acid binding protein (hIFABP)-

ketorolac-ANS complex structure. The best model given by the present NMRScore_P method agreed well with the available experimental observations.

2.3. Methods

2.3.1 Preparation of the ^{1}H Chemical Shift Perturbations Database

We selected 54 protein-ligand complexes from the PDBbind Database v.2013[61] core set for which high-resolution crystal structures were available. The protein in each complex was protonated by the H++ server[62] and then modeled by the AMBER ff99SB[63] force field (which uses the same charge set as the AMBER ff94 force field[64]). The ligand molecule was modeled by GAFF[65] with AM1-BCC partial charges. In order to remove bad contacts within the structure, The protein hydrogen atoms in each complex were minimized using the SANDER program from the AMBER 12 program suite.[66] After structural minimization, The AF-QM/MM approach[34] was used to compute the ^{1}H isotropic chemical shielding constants (σ) of the protein binding pocket in both the bound and unbound forms, and the difference between these two values was taken as the ^{1}H CSP ($\Delta\delta_H$) induced by ligand binding. The following two equations show the derivation of $\Delta\delta_H$ from $\sigma_{H(unbound)}$ and $\sigma_{H(bound)}$ based on the definition of the chemical shift.

$$\delta_H = \frac{\upsilon_H - \upsilon_{ref}}{\upsilon_{ref}} \times 10^6 = \frac{\sigma_{ref} - \sigma_H}{1 - \sigma_{ref}} \overset{\sigma_{ref} << 1}{=} \sigma_{ref} - \sigma_H \quad (2)$$

$$\Delta\delta_H = \delta_{H(bound)} - \delta_{H(unbound)} = \sigma_{ref} - \sigma_{H(bound)} - (\sigma_{ref} - \sigma_{H(unbound)}) = \sigma_{H(unbound)} - \sigma_{H(bound)} \quad (3)$$

Herein δ represents the chemical shift, υ stands for the absolute resonance frequency. In our AF-QM/MM calculations the QM region was described by the B3LYP/6-31G** level of theory while the MM region was described using AMBER ff94 partial charges.[34] For each protein-ligand complex system, there were 2N (where N is the number of residues in the binding pocket) parallel

calculations performed via treatment of each interacting residue as the center of the QM region, with and without ligand, respectively. We only collected the isotropic shielding constants of the protons in the center residue from each calculation. All calculations were carried out using the Gaussian 09 program.[67] A collection of all of the computed isotopic shielding constants and the complex structures is given as part of the SI.

Since it is the electron cloud surrounding the proton that serves to shield the proton from the external magnetic field, different molecular environments will lead to different (de)shieldings on the proton. Therefore, in order to accurately predict the CSPs for protein based protons induced by ligand binding to a given target protein, we categorized the protons from the backbone and side chain into 10 types with the side chain proton type assignment following the AMBER atom typing scheme.[68] The details of the proton type categorization in HECSP are given in Table 2.1.

Table 2.1. List of proton atom types in HECSP.

Major group	Type	Description
Backbone	HA	Alpha H in protein backbone
	HN	Amide H in protein backbone
Side chain	H	H attached to N
	HC	H attached to aliphatic carbon with no electron-withdrawing substituent
	H1	H attached to aliphatic carbon with one electron-withdrawing substituent
	Har	H attached to aromatic carbon
	H4	H attached to aromatic carbon with one electronegative neighbor
	H5	H attached to aromatic carbon with two electronegative neighbors
	HP	H attached to carbon directly bounded to formally positive atoms
	HO	H in alcohols and acids

[a] See references[64]

2.3.2. Ring Current Effects

The ring current effect can significantly influence protein protons close to the aromatic rings of the ligand. When buried in an external magnetic field, the π electrons of the aromatic system create an induced ring current, which generates an induced magnetic field. Right above or below the central part of the ring, the induced magnetic field is antiparallel to the external field, which increases the shielding (causing lower precession frequency) on the target protons. However, on the edge but beyond the plane of the ring, the induced field adds strength to the external field, resulting in dishielding (causing higher precession frequency) for the target protons in this area. Several models have been developed to represent the ring current effect (*e.g.,* Pople,[69] Johnson and Bovey,[70] and Haigh and Mallion[71]). We implemented the model from Haigh and Mallion into HECSP, which has performed well in previous studies.[45, 48, 72]

When computing the ring current contributions of the ligand aromatic rings to protein ^{1}H CSPs, HECSP first creates a list of aromatic rings in the ligand and then goes through the protein protons one by one. For each target proton, HECSP adds up the contribution from each ring whose ring center is within 6 Å of the current target proton. The total ring current contribution of the ligand to a particular ^{1}H CSP is calculated using equation 4:

$$\Delta\delta_{RC} = F\sum_{rings} GI \tag{4}$$

Herein F is a target-specific constant, G is the geometrical factor for a pair consisting of the ring and proton, and I is the ring current intensity, which represents the ratio of the intensity of an aromatic ring relative to that of a benzene ring (I =1 for benzene). F and I are the two dimensionless constants being determined via linear regression. G is further represented as:

$$G = \sum_{i,j} d_{ij} S_{ijO} \tag{5}$$

Where d_{ij} is expanded as:

$$d_{ij} = \frac{1}{r_i^3} + \frac{1}{r_j^3} \ . \tag{6}$$

Herein r_i and r_j are the distances from the target proton to the two adjacent ring atoms i and j respectively. S_{ijO} in equation 5 is the area of the triangle formed by the ith and jth ring atoms and the projection of the target proton onto the aromatic ring plane, which is shown as point O in Figure 2.1. S_{ijO} is positive if $\overrightarrow{Oi} \times \overrightarrow{Oj}$ is antiparallel to the normal vector of the ring ($\vec{n}$), and negative if it is parallel to the normal vector. The normal vector is calculated as the cross product of vectors pointing from the first ring atom to the second and the last one respectively. Hence, for each proton there will be a d_{ij} and S_{ijO} value for each edge (defined by two adjacent atoms – for a total of six in benzene) of the aromatic ring. The details of the ring current effect are illustrated in Figure 2.1.

Figure 2.1. The ring current effect caused by an aromatic ring towards a target proton. The red sphere represents the target proton while the aromatic ring is shown in blue. The two smaller red dots are the center of the ring and the projection of the proton onto the plane of the aromatic ring.

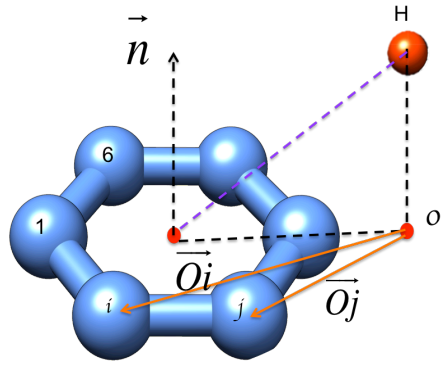
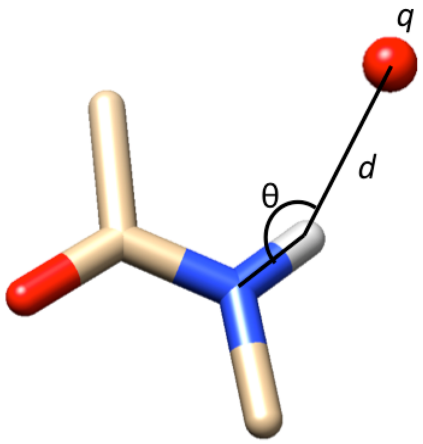


Figure 2.2. The electric field effect caused by a polar atom. The red sphere represents a polar atom, which is the source of the electric field effect on the target proton shown in light grey.



2.3.3. Electric Field Effects

The electric field effect is another important contribution to the ^{1}H CSP. This effect originates from distant polar groups polarizing the target X-H bond (where X is a heavy atom) and thereby

influencing the local shielding by increasing or decreasing the local electron density. The electric field effect was evaluated for all protein protons in HECSP, which adds up the effects caused by any polar ligand atom within 10 Å. We also observed that a rather large cutoff improves the fit quality (see results and discussion). This is consistent with the fact that the electrostatic interaction decays slowly with distance. Herein the "polar atoms" include each carbon or hydrogen atom that connects to at least one atom which is not carbon or hydrogen, and all other elements (besides carbon and hydrogen). According to the method of Buckingham,[73] the CSP of a particular proton is proportional to the local electric field projection on the X-H bond vector (see equation 7).

$$\Delta\delta_{EF} = kE(X-H) \tag{7}$$

Herein, k is a specific parameter which depends on the proton type. $E(X-H)$ is the sum of local electric fields induced by polar atoms which are evaluated as:

$$E(X-H) = \sum_i \frac{q_i \cos\theta_i}{d_i^2} \tag{8}$$

Herein q_i is the partial charge of a polar atom, θ_i is the angle formed by the polar atom-H-X and d_i^2 is square of the distance between the polar atom and the target proton. The electric field effect is depicted in Figure 2.2 We compared the performance of different charge models including the Gasteiger and AM1-BCC. Both charges were computed with antechamber in AmberTools15.[74] Based on our analysis we found that the Gasteiger charges gave slightly better computed results.

2.3.4. Hydrogen Bond Effects

Hydrogen bond interactions play an important role in protein-ligand binding. Early studies[48, 75, 76] showed that the hydrogen bond induced CSP for the amide and alpha proton (HN and HA) in a protein backbone can be modeled as:

$$\Delta\delta_{HB} = ar^{-3} + b \quad (9)$$

Where r is the hydrogen-acceptor distance and a and b are fitting parameters. Taking a cue from ShiftX[48], which does not include an explicit hydrogen bond term for side chain protons, we initially treated the CSPs of the side chain protons induced by nearby polar or charged atoms inside the ligand as an electrostatic effect (without the hydrogen bond term) and only employed the hydrogen bond term for the backbone protons. However, using this model, large errors occurred for the H and HO proton types during the fitting procedure. Therefore, we included the hydrogen bond term for H and HO atom types in the side chains which resulted in a concomitant decrease in the RMSE. Two steps were performed to detect hydrogen bonds between the protein binding site and its ligand. First, HECSP loops over all the ligand atoms and identifies potential hydrogen bond acceptors based on their SYBYL atom types, which were assigned using antechamber in AMBER.[66] The potential acceptor types encountered in our dataset are N.1, N.2, N.3, N.ar, O.3, O.2, O.co2, S.3, F, Cl and Br. Secondly, the following distance and angle criteria were applied to determine if a hydrogen bond was present: (1) both the hydrogen-acceptor and donor-acceptor distances are less than 3.5 Å; (2) donor-hydrogen-acceptor angle larger than 90°; and (3) only the bond with shortest hydrogen-acceptor distance is selected if there are multiple bonds fulfilling criteria (1) and (2).

2.3.5. Magnetic Anisotropic Group Contributions

The magnetic anisotropy of unsaturated groups in the ligand can perturb chemical shifts in proteins. HECSP employed McConnell's equation[77] and scaled it by a target specific constant C to compute the contribution of the anisotropic groups in the ligand to protein ^{1}H CSPs (see equation 10).

$$\Delta\delta_M = \frac{C}{3NR^3} \sum_{i=a,b,c} \chi_{ii}(3\cos^2\theta_i - 1) \quad (10)$$

Here, N is Avogadro's number; R is the distance between the target proton and the center of mass of the distant group; χ_{ii} is a component of the magnetic susceptibility tensor in the principal inertial axis system of that group; θ_i is the angle between the i^{th} principal axis and the vector from the group center to the target proton. Here the "group" is defined as a fragment which consists of several atoms that are connected by bonds other than single bonds.

In the present work, we used the χ_{ii} values of Flygare and co-workers (see Table 2.2).[78-80] They derived a collection of atom-based susceptibility components from gas-phase molecular Zeeman measurements of many molecules using a least-squares fitting strategy.[78] These localized atom susceptibilities can be summed up to estimate the susceptibility (χ_{aa}, χ_{bb} and χ_{cc}) of a molecular fragment by rotating the individual atom values from their atomic axis frame (see Table 2.2) into the principal inertial axis frame of the fragment using equation 11:[78-80]

$$\chi_{aa} = \chi_{xx}\cos^2\theta_{ax} + \chi_{yy}\cos^2\theta_{ay} + \chi_{zz}\cos^2\theta_{az} \quad (11\ a)$$

$$\chi_{bb} = \chi_{xx}\cos^2\theta_{bx} + \chi_{yy}\cos^2\theta_{by} + \chi_{zz}\cos^2\theta_{bz} \quad (11\ b)$$

$$\chi_{cc} = \chi_{xx}\cos^2\theta_{cx} + \chi_{yy}\cos^2\theta_{cy} + \chi_{zz}\cos^2\theta_{cz} \quad (11\ c)$$

where θ_{ax} is the angle formed by the principal inertial axis a and the atomic axis x.

Altogether, the full functional form of HECSP is:

$$\Delta\delta_H = F\sum_{rings} GI + k\sum_i \frac{q_i\cos\theta_i}{d_i^2} + ar^{-3} + b + \frac{C}{3NR^3}\sum_{i=a,b,c}\chi_{ii}(3\cos^2\theta_i - 1)$$

Table 2.2. Local Atom Susceptibilities in the atomic axis frame in units of $4\pi\times10^{-12}$ m^3/mol

y x	χ_{xx}	χ_{yy}	χ_{zz}
C=b	-3.64	-3.75	-7.33
—C≡a	-9.9	-7.4	-7.4
O=b	1.90	-1.29	-5.70
N—(planar) b	-13.82	-10.35	-6.13
N≡a	-9.5	-4.5	-4.5
S=a	4.7	-13.1	-23.0
P≡b	-24.1	-17.9	-17.9

[a] See reference [78, 79]

[b] See reference [80]

2.3.6. Parameter Fitting

Overall, our model has 5 linear parameters (F, k, a, b and C) for each of the atom types HA, HN, H and HO, 3 linear parameters (F, k and C) for each atom type of the side chain protons excluding H and HO, and 7 ring current intensity factor parameters (I) for tetrazole, imidazole, pyrazole, thiophene, oxazole, pyridine and pyrimidine (using I=1.0 for benzene).

Firstly, we performed a linear least-squares fit to parameterize the HECSP model. Initially, the fit was performed only for the HA atom type with 12 parameters (F, k, a, b, C for HA and 7 generalized ring current intensity factors). Next, 5 parameters (F, k, a, b, C) for the HN atom type were fit by fixing the 7 ring current intensity factors at the values obtained from the fit for HA and this was then repeated for the HO and H atom types as well. Then the same least-squares fitting

algorithm was employed to parameterize other types of side chain protons: with three parameters (F, k and C) determined per atom type. The same procedure was carried out with the ligand molecules employing the AM1-BCC and Gasteiger charges, respectively. Here the two sets of parameters are referred to as "AM1-BCC_LS" ("LS" stands for least-squares fitting) and "Gasteiger_LS" respectively. The Gasteiger_LS parameter set and the AM1-BCC_LS parameter set are shown in Supporting Information.

In addition to the linear least-squares fits, we performed leave-one-out cross-validation (LOOCV) analysis to assess the predictive ability of the model. We left out one structure and fitted the model using the remaining 53 structures with the same least-squares fitting strategy described above. Then the ^{1}H CSPs of the structure, which was left out, were predicted using the new parameter set. We defined a LOOCV estimate for each adjustable parameter p as $\bar{p}$, which was obtained as:

$$\bar{p} = \frac{1}{N}\sum_{i=1}^{N} p_i \tag{12}$$

Herein p_i is the least-squares estimate when the i^{th} complex structure is left out and we have N=54 in total. Then we further determined the uncertainties of these estimates by computing the standard errors of the mean (SEMs) using equation 13:

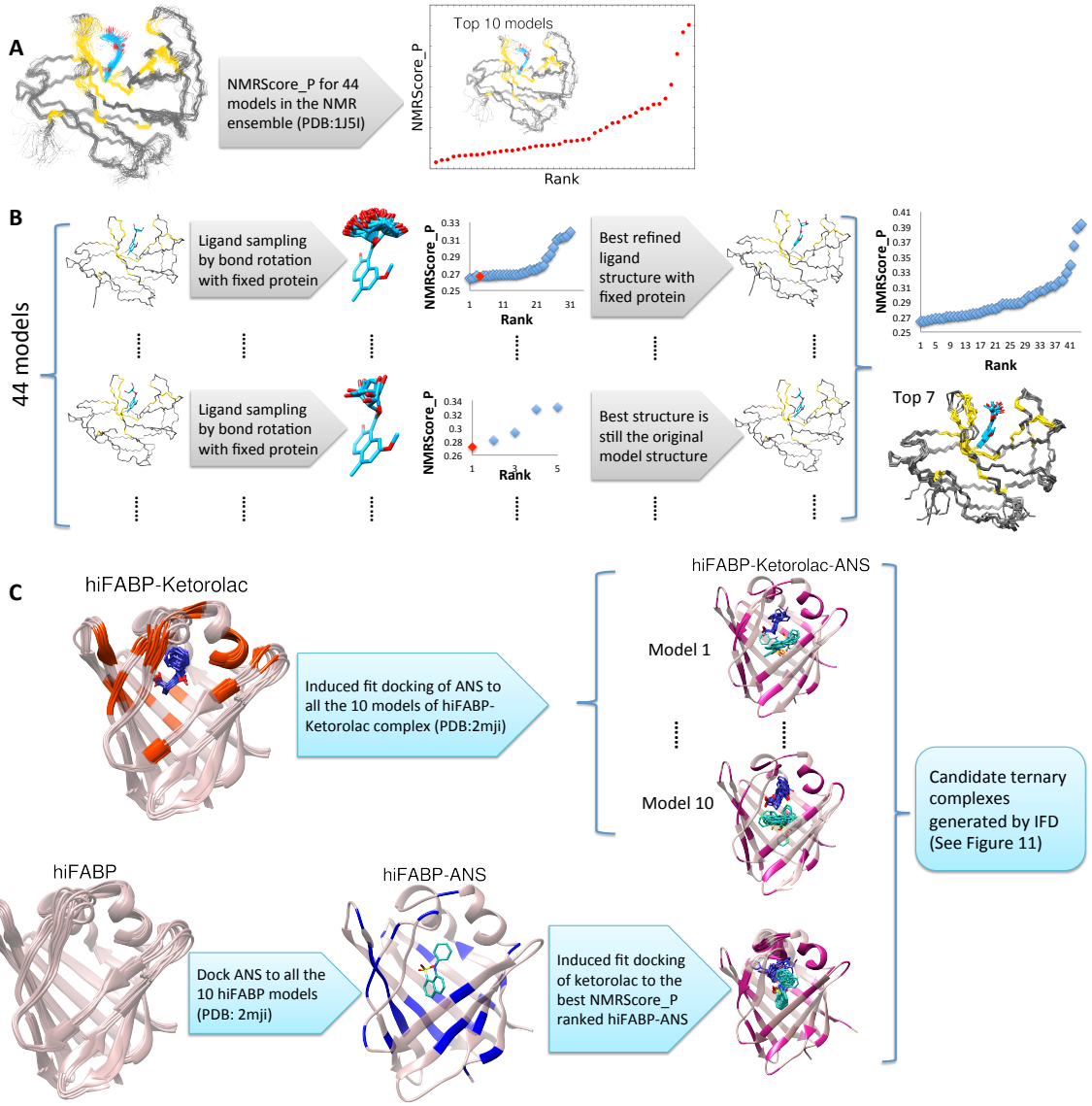
$$\hat{\sigma}_p = \sqrt{\frac{\sum_{i=1}^{N}(p_i - \bar{p})^2}{N(N-1)}} \tag{13}$$

All the LOOCV estimated parameters were collected and herein referred as the Gasteiger_LOOCV parameter set. The Gasteiger_LOOCV parameter set and the corresponding SEMs are shown in Supporting Information.

2.3.7. NMRScore_P and NMRScore_L

Previously, a NMR scoring function was developed based on semiempirical ligand CSPs induced by the protein environment (NMRScore_L)[57]. Similarly, we build another NMR scoring function (NMRScore_P), which is defined as the RMSD between calculated and experimental protein ^{1}H CSPs induced by ligand binding. We demonstrate the application of NMRScore_P on the following two systems and its workflow is illustrated in Figure 2.3.

Figure 2.3. (A). Workflow of ranking structures in NMR ensemble using NMRScore_P on the apoNCS-naphthoate ester complex (PDB: 1J5I). (B). Workflow of ligand structure refinement in the apoNCS-naphthoate ester complex (PDB: 1J5I) using NMRScore_P. The red dots in the scatterplots represent the original NMR models. (C). Workflow of ternary hIFABP-ketorolac-ANS complex structure determination with NMRScore_P. hIFABP is shown as a rosy brown ribbon and the locations that were observed to have significantly perturbed protons are mapped onto the hIFABP structure in orange (induced by ketorolac), blue (induced by ANS) or violet (induced by ketorolac and ANS binding at the same time). Ketorolac is shown as a blue stick whereas ANS is shown as a cyan stick.



2.3.8. Application of Protein NMRScore_P to the *Apo*NCS-naphthoate ester complex

2.3.8.1. Scoring procedure for the structural ensemble.

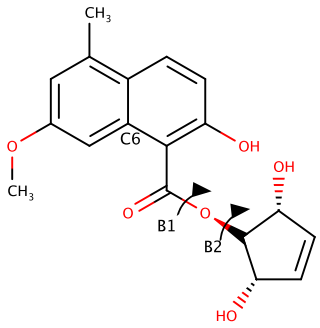
NMRScore_Ps were computed only for the residues that were significantly perturbed and within 10 Å of the ligand (see Figure 2.3 A). The significantly perturbed residues are the ones that have protons with CSPs greater than one standard deviation from the average CSP for each proton type. Since NMRScore_P is the ^{1}H CSP RMSD from experimental values, a lower score means a better NMR model. In order to verify the accuracy of NMRScore_P obtained with HECSP, we compared the performance of NMRScore_Ps with three parameter sets Gasteiger_LS, Gasteiger_LOOCV and AM1-BCC_LS and the NMRScore_Ps obtained with AF-QM/MM. Chemical shifts of the *apo*NCS-naphthoate ester complex and *apo*NCS are available in the BioMagResBank (BMRB accession number 5344 and 5343).

2.3.8.2. Further refinement of the ligand solution structures.

The solution structures of the *apo*NCS-naphthoate ester complex were originally determined from a set of distance and torsional angle restraints with intermolecular NOEs (nuclear Overhauser effect) available only between the four aromatic ring protons of the ligand and the receptor.[81] The superposition of 44 structures shows that the 2-hydroxy-7-methoxy-5-methyl-naphthalene-1-carboxylic acid fragment is well-defined, and binds deep into the pocket; however, the trihydroxy-cyclopentene moiety is highly flexible and points towards the opening of the pocket (see the structure in Figure 2.4). Herein, in order to further refine the trihydroxy-cyclopentene moiety, we generated poses by rotating around the designated rotatable bonds (B1 and B2 see Figure 2.4) with a 20° step size and the poses with either intra- or inter- molecular steric clashes were screened out.

Then we computed the NMRScore_Ps with Gasteiger-LOOCV (see Figure 2.3 B) to identify the best structure(s).

Figure 2.4. Ligand structure



2.3.8.3. Docking Procedure.

Glide was used to generate 50 different ligand decoys for each NMR model. The grid box was defined as a cube with an inner and outer edge of 10 and 30 Å and centered on the geometric center of the ligand pose from experiment. The top 10 poses were used to calculate the protein NMRScore_P. We then computed the NMRScore_P *versus* structural RMSD from the native ligand in the corresponding NMR model. The purpose here is to evaluate the ability of HECSP based NMRScore_P to distinguish the "native state" NMR model from decoy poses.

2.3.9. Ternary hIFABP-ketorolac-ANS complex determination through induced fit docking (IFD) and NMRScore_P.

IFD calculations were performed to generate ternary hIFABP-ketorolac-ANS complex candidates and Gasteiger_LOOCV based NMRScore_P calculations were then used to filter out the best

structures (see Figure 3 C). In one instance, we docked ANS into all ten models of the hIFABP-ketorolac complex NMR ensemble (PDB: 2MJI) using the IFD default protocol available in Maestro,[82] which allows receptor flexibility (herein hIFABP-ketorolac complex is the receptor). In the second instance, based on the observed CSP in the HSQC (heteronuclear single quantum coherence) spectra of hIFABP in the presence of ANS,[83] ANS was first docked into the lower part of the barrel of 10 *holo*-hIFABP models with Glide[59, 60] and the best hIFABP-ANS complex was screened out based on Gasteiger_LOOCV computed NMRScore_P. Then ketorolac was docked into the flexible receptor (hIFABP-ANS complex) with the same IFD protocol. Chemical shifts of the hIFABP in the ternary complex with ketorolac and ANS are available in the BioMagResBank (BMRB accession number 19727). Chemical shifts of *apo* hIFABP, hIFABP in complex with either ANS or ketorolac were provided by Professor Scanlon (private communication).

2.4. Results and Discussion

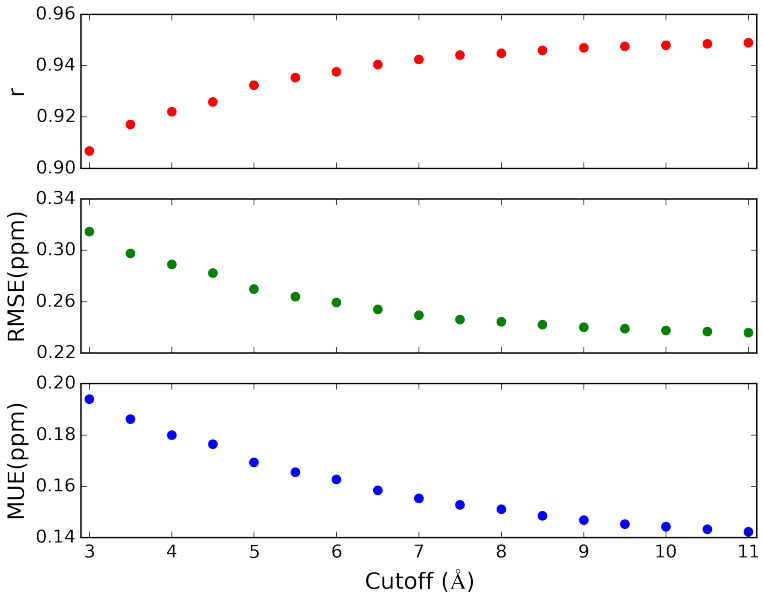
2.4.1. Hydrogen bond term for the protein sidechain protons

In a previous study[48], the intra-protein hydrogen bond effect on side chain polar protons were not explicitly described as an independent term, instead, all the CSPs caused by polar or charged atoms were generalized as the electric field effect for the side chain protons. Similarly, we made an assumption that the CSPs induced by hydrogen bond interactions between side chain polar protons and acceptor atoms in the ligand molecule could be treated as a part of the electric field effect. However, an independent hydrogen bond effect term was found to be crucial to predict the CSPs of side chain H and HO proton types. By adding an independent hydrogen bond term for H and HO proton types, both of the RMSEs decreased by ~0.2 ppm.

2.4.2. Cutoff dependence for the electric field term

We have explored fitting the HECSP model (equation 1) with different cutoffs for the electric field term. The correlation coefficients, RMSEs and MUEs between the predicted values and target ones for all the protein protons are given in Figure 2.5 with respect to the cutoff employed. We can see that the fitting quality improved considerably as a function of the cutoff value. This is consistent with the fact that electrostatic interactions decay as a functionof 1/R. A 10Å cutoff is adopted in HECSP because it provided excellent accuracy at a modest computational cost.

Figure 2.5. Correlation coefficients, RMSEs, and MUEs of the HECSP predictions for all proton CSPs along with the cutoff for electric field term.



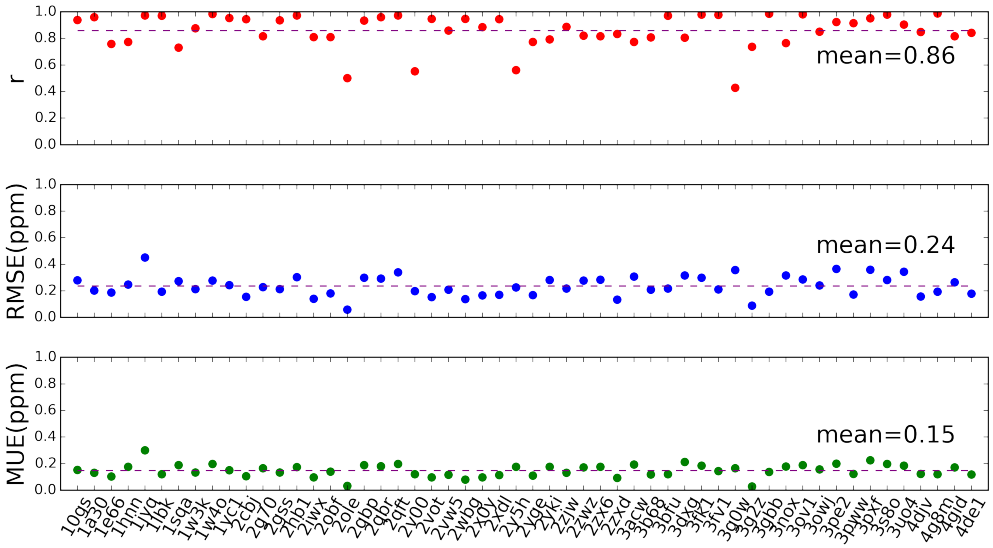
2.4.3. Charge model dependence

The charge model dependence of the HECSP approach was studied by comparing the fitting statistics of two different charge models for the ligand molecules. Table 2.3 illustrates the statistics of the AM1-BCC_LS, Gasteiger_LS, and Gasteiger_LOOCV parameter sets. As shown in Table 2.3, Gasteiger performs slightly better than AM1-BCC charge except for amide ^{1}H. With the Gasteiger_LS parameter set, HECSP yielded correlation coefficients pearson's r of 0.897, 0.971, 0.945 and 0.948 for the alpha ^{1}H, amide ^{1}H, side chain ^{1}H and all ^{1}H with RMSEs of 0.151, 0.199, 0.257 and 0.238 ppm, respectively. Meanwhile, the AM1-BCC charge model is also a good alternative and its performance is similar to the Gasteiger charge model and AM1-BCC_LS even performs better on amide ^{1}H than Gasteiger_LS. Overall it gave a RMSE for all ^{1}H CSPs of 0.244 ppm. Although there are noticeable gaps between the two charge models in predicting ^{1}H CSPs, in general both of their predictions correlate well with AF-QM/MM calculated values.

Table 2.3. Fitting statistics for the different parameter sets.

Data set		AM1-BCC_LS			Gasteiger_LS			Gasteiger_LOOCV		
Proton	No. of protons	r	RMSE (ppm)	MUE (ppm)	r	RMSE (ppm)	MUE (ppm)	r	RMSE (ppm)	MUE (ppm)
HA	1123	0.888	0.157	0.101	0.897	0.151	0.097	0.897	0.151	0.097
HN	1016	0.973	0.190	0.126	0.971	0.199	0.126	0.971	0.199	0.126
Side chain H	5844	0.941	0.265	0.164	0.945	0.257	0.157	0.945	0.257	0.157

Figure 2.6. Statistics of the predictions from the leave-one-out cross-validation of all protons in the corresponding complex structure.



2.4.4. Leave-one-out-cross-validation analysis

We performed the LOOCV analysis for the Gasteiger_LS parameter set to evaluate the reliability of the fitted parameters. Each protein-ligand system is left out in turn and the parameters were refitted based on the other 53 systems using the same least-squares fitting algorithm. One important outcome of the LOOCV procedure is to validate whether the proposed model can accurately predict ^{1}H CSPs in a new protein-ligand complex system. Each dot in Figure 2.6 represents the statistics for the corresponding system. The average of the correlation coefficients, RMSEs and MUEs are 0.86, 0.24 ppm and 0.15 ppm respectively. Although some fluctuations are seen for the correlation coefficient, the 54 RMSEs and MUEs in Figure 2.6 are consistently close to the mean values, which supports the predictive ability of our model.

The Gasteiger_LS parameter set and the Gasteiger_LOOCV parameter set with its SEMs are listed in Supporting Information. We can see that the Gasteiger_LOOCV parameter set is comparable to the Gasteiger_LS parameter set. And the SEMs are small in magnitude further validating the fitted parameters. Gasteiger_LS and Gasteiger_LOOCV parameter sets yield matching predictions for all the proton types, and they give identical correlation coefficients, RMSEs and MUEs when compared to the AF-QM/MM calculated values (see Table 2.3).

2.4.5. Study on the apoNCS-naphthoate ester complex with NMRScore_P

2.4.5.1. The apoNCS-naphthoate ester complex ensemble

To further validate our approach and show its ability to rank the solution structures of protein-ligand complexes, we have applied our approach to the *apo*NCS-naphthoate ester complex system. As shown in Figure 2.7, there is an excellent agreement between AF-QM/MM and HECSP calculated ^{1}H CSPs over all the 44 protein-ligand solution structures in the ensemble. The RMSEs corresponding to each parameter set are also displayed in the plot and range from 0.107 to 0.112 ppm.

We also computed the NMRScore_Ps using both AF-QM/MM and our approach. As shown in Figure 2.8, HECSP based NMRScore_Ps are comparable to the values obtained with AF-QM/MM and a similar ranking is predicted for all the NMR models. Both Gasteiger_LOOCV and Gasteiger_LS gave nearly the same order for the 44 structures, with only two swaps in the ordering for closely ranked structures (31, 35 and 9, 38). AM1-BCC_LS provided a different order but it was still in close accord with AF-QM/MM. Although some gaps are seen in the NMRScore_Ps in Figure 2.8, the absolute discrepancy is just ~ 0.05 ppm. Table 2.4 lists the NMRScore_Ps for the

44 NMR models using Gasteiger_LOOCV. The NMRScore_Ps were distributed between 0.266 and 0.401 ppm with a standard deviation of 0.031 ppm. NMRScore_Ps for the models numbered 6, 1, 3, 39, 2, 32, 31, 35, 29, 7, 36, 8, 9, 38, 23, 20, 22, 24, 14, 44, 19, 21, 34, 28, 16, 41, 30 and 42 (sorted from low to high) are below 0.297ppm, and, herein we consider them to be better representations of the native structure. Model 6 is the best one predicted by both Gasteiger_LS and Gasteiger_LOOCV and ranked second by AF-QM/MM NMRScore_P, which has an AF-QM/MM NMRScore_P of 0.276 ppm. The best model based on AF-QM/MM is model 24, whose AF-QM/MM NMRScore_P is 0.271 ppm.

Figure 2.7. Correlation between AF-QM/MM and HECSP calculated 1H CSPs over all the 44 protein-ligand solution structures in the ensemble (PDB:1J5I).

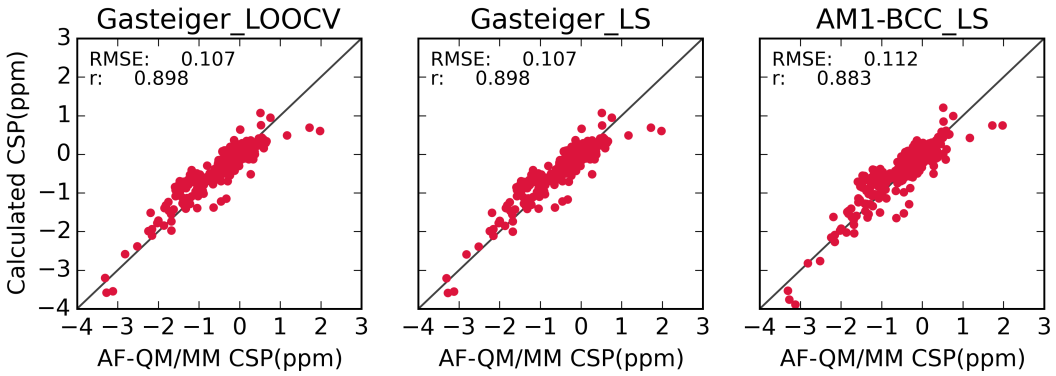


Table 2.4. Gasteiger_LOOCV NMRScore_Ps of the 44 experimentally determined apoNCS-naphthoate ester complex NMR models.

Rank	Model	NMRScore_P (ppm)	NMRScore_P after refinement	Rank	Model	NMRScore_P (ppm)	NMRScore_P after refinement
1	6	0.266	0.264	23	34	0.287	--
2	1	0.269	0.267	24	28	0.287	--
3	3	0.269	0.268	25	16	0.288	--
4	39	0.272	--	26	41	0.288	0.287
5	2	0.273	0.268	27	30	0.289	--
6	32	0.273	0.271	28	42	0.295	0.287
7	31	0.274	0.270	29	26	0.298	--
8	35	0.274	--	30	17	0.300	0.297
9	29	0.274	0.274	31	10	0.305	--
10	7	0.276	0.265	32	37	0.305	0.305
11	36	0.276	--	33	18	0.310	0.309
12	8	0.277	0.268	34	27	0.311	0.311
13	9	0.278	0.264	35	33	0.315	0.311
14	38	0.278	0.272	36	15	0.319	0.293
15	23	0.279	0.273	37	11	0.319	0.317
16	20	0.280	--	38	12	0.323	0.302
17	22	0.281	0.271	39	43	0.323	0.319
18	24	0.282	0.270	40	13	0.329	--
19	14	0.283	0.281	41	4	0.342	0.339
20	44	0.283	--	42	40	0.372	0.365
21	19	0.283	0.278	43	25	0.393	--
22	21	0.284	0.281	44	5	0.401	0.388

The models for which there were no better structures obtained after refinement were marked as --.

Figure 2.8. NMRScore_P and Rank for the 44 Experimentally Determined NMR Models (PDB:1J5I). The x axis shows the ranking of the solution structures predicted by corresponding parameter set of our method. (A). Comparison of NMRScore_P computed with Gasteiger_LOOCV and AF-QM/MM. (B). Comparison of NMRScore_P computed with Gasteiger_LS and AF-QM/MM. (C). Comparison of NMRScore_P computed with AM1-BCC_LS and AF-QM/MM.

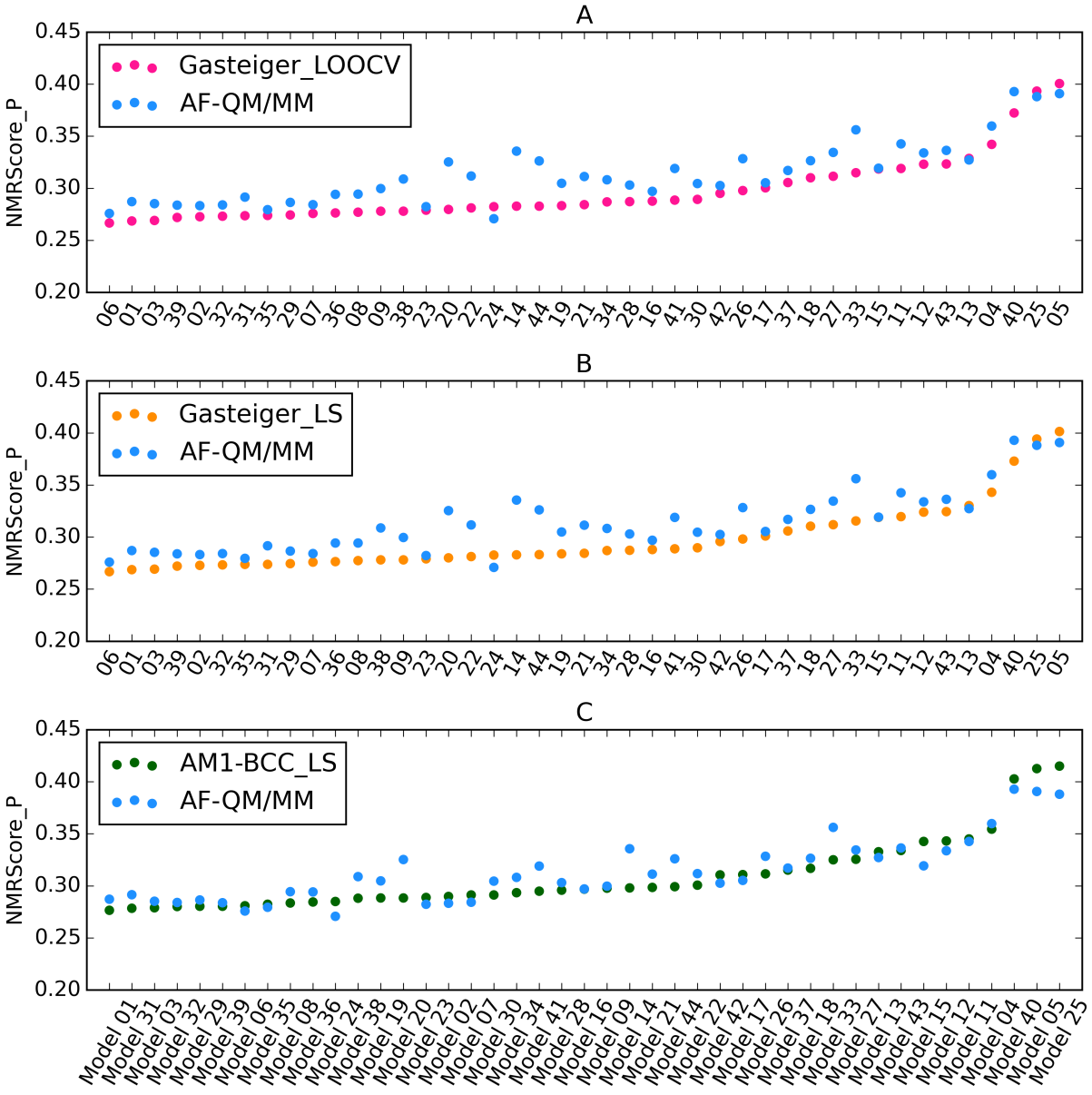
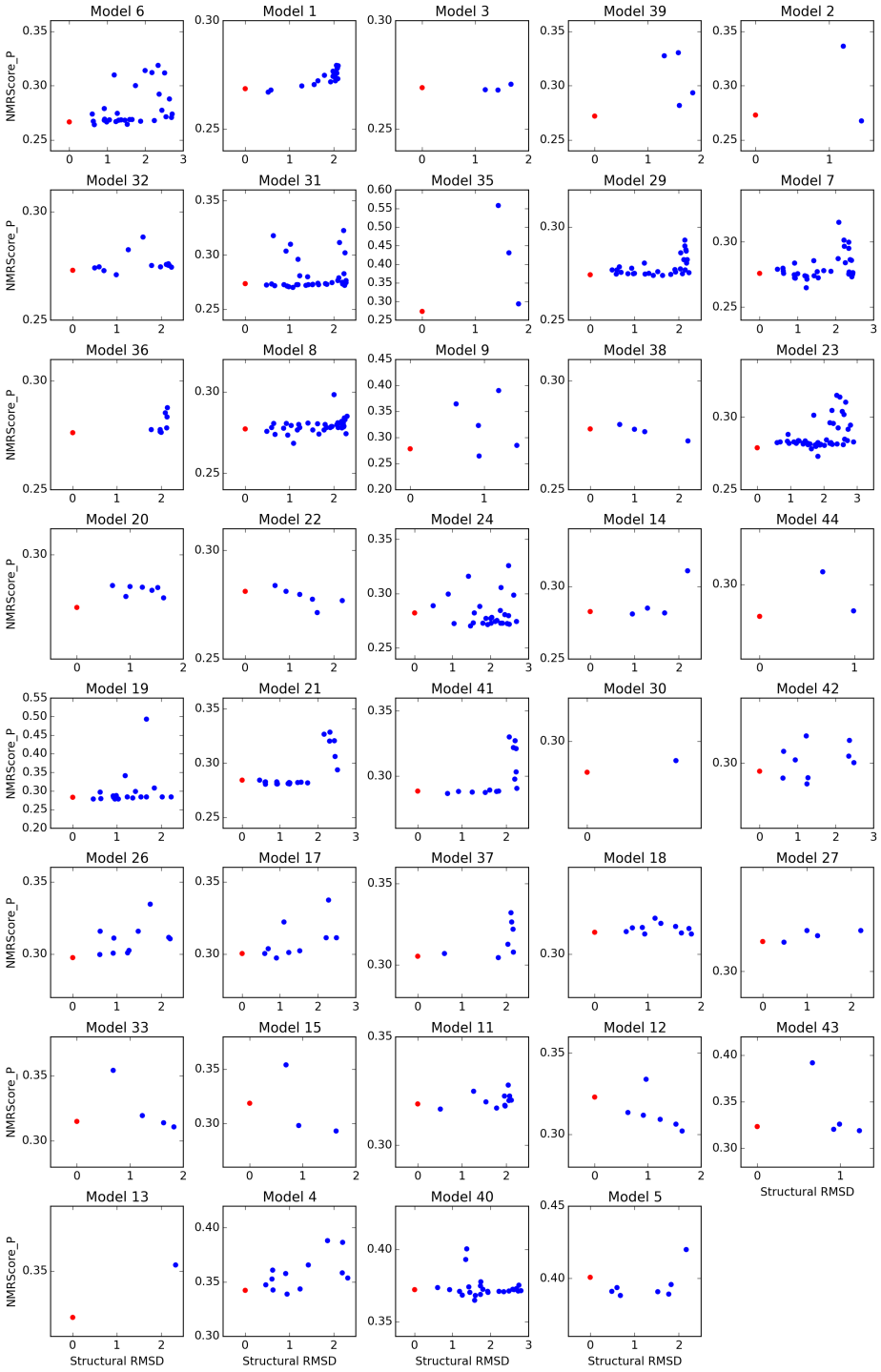


Figure 2.9. Gasteiger_LOOCV NMRScore_P *vs* structural RMSD (Å) for corresponding models. The red dots represent the experimental NMR ligand structures (PDB: 1J5I) The blue dots represent the ligand conformers generated by rotating trihydroxy-cyclopentene moiety around two rotatable bonds.



2.4.5.2. Further refinement of ligand solution structures

We also used Gasteiger_LOOCV NMRScore_P to further refine the ligand structures in all 44 models. In order to fulfill the intermolecular NOEs, the aromatic fragment was kept fixed and only the trihydroxy-cyclopentene moiety was sampled around two rotatable bonds. Each scatterplot in Figure 2.9 is the Gasteiger_LOOCV NMRScore_P *versus* the structural RMSD for the corresponding model. For good NMR models, we can see many conformations turn out to be around the best NMRScore_P in the scatterplots, which is consistent with the fact that the trihydroxy-cyclopentene moiety is highly flexible. The best NMRScore_Ps of the refined ligands are also listed in Table 2.4. Except for model 10, 16, 20, 25, 26, 28, 30, 34, 35, 36, 39, 41 and 44, there were ligand conformations generated with lower NMRScore_Ps than the corresponding experimental model counterparts, however, most of the enhancements were modest. Relatively big NMRScore_P improvements were observed for model 5, 7, 9, 12, 15 and 24 by 0.013, 0.011, 0.014, 0.021, 0.026 and 0.012 ppm. As shown in Figure 2.10, in all the 6 models, we observed that the refined orientation of the trihydroxy-cyclopentene moiety were rotated to a common orientation, which had a much smaller angle between the double bond in cyclopentene and the aromatic plane. It was also a general trend throughout all 44 models. The average angle between the double bond in cyclopentene and the aromatic plane was 62 degrees, whereas the average angle decreased to 38 degrees after refinement. After refinement, the best NMRScore_P we obtained for this system is 0.264 ppm for model 6 (See Figure 2.11, which shows the overlay of the experimentally determined apoNCS-naphthoate ester complex and the structure determined by NMRScore_P) and we also obtained a collection of complex structures that had NMRScore_Ps better than 0.270 ppm (for model 6, 9, 7, 1, 3, 2 and 8). Of our refined models, we consider models

6, 9, 7, 1, 3, 2 and 8 as being our best and most representative of the putative native state. The refined ensemble of *apo*NCS-naphthoate ester complexes is provided as a PDB file in the SI.

Figure 2.10. NMR structures of ApoNCS-naphthoate ester complex (PDB 1J5I) and the refined ligand structures. The blue colored part is the experimentally determined ligand structure in all the small figures. The fragments in other colors demonstrate the refined trihydroxy-cyclopentene moiety. The numbers shown in each figure are the NMR model numbers.

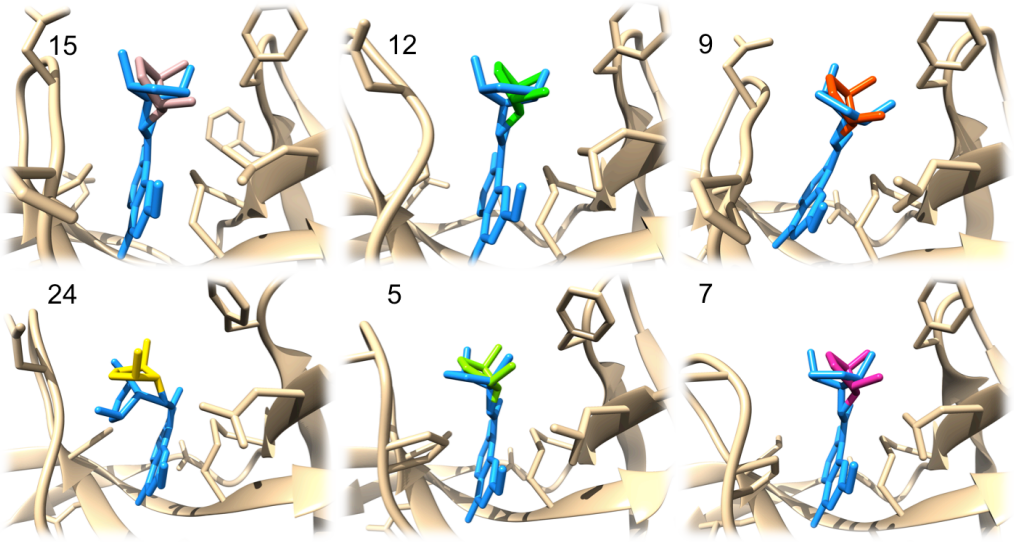
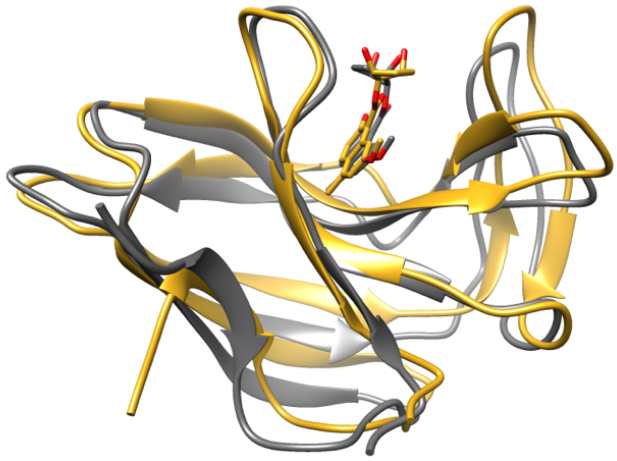


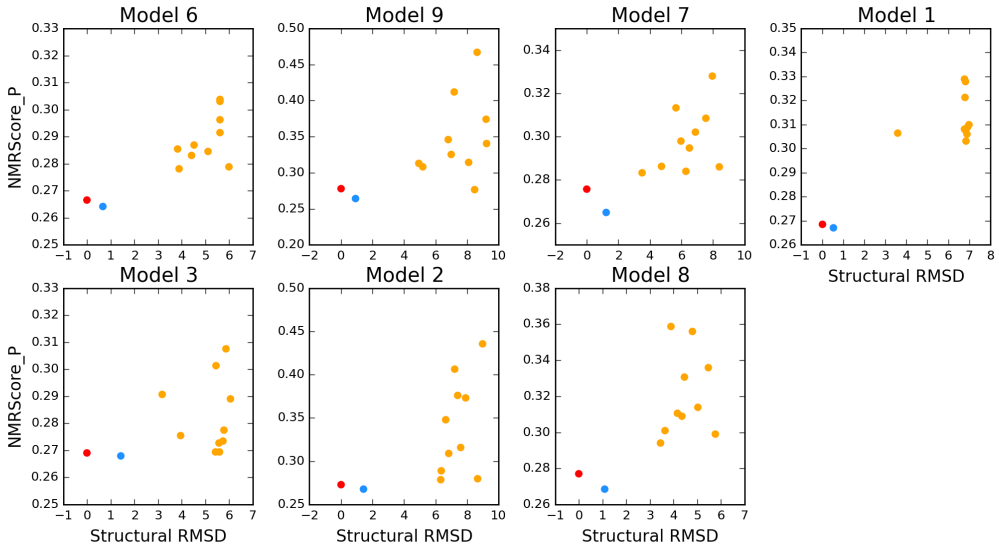
Figure 2.11. (A). NMR structure of ApoNCS (PDB 1J5I model 6) together with refined ligand structure. It is the best ranked complex structure by NMRScore_P amongst all the NMR models and refined structures (as shown in yellow). (B). The gray counterpart is the best representative conformer in the experimental ensemble (PDB 1J5I model 1)



2.4.5.3. Native states and decoys.

We used Glide to do rigid docking of the flexible ligand to the receptor structures for the models numbered 6, 9, 7, 1, 3, 2 and 8. Structural RMSDs of the docked ligand poses ranged from 3.2 to 9.2 Å with respect to the experimentally determined ligand structures (see Figure 2.12). For all the seven models, NMRScore_P successfully ranked the native state (refined ligand structure shown as blue dot in Figure 2.12) better than the top 10 poses generated using the Glide scoring function.

Figure 2.12. Gasteiger_LOOCV NMRScore_P vs structural RMSD (Å) of Glide docked poses. The red dots represent the experimental NMR ligand structures (PDB: 1J5I). The orange dots represent the docked poses. The blue dots represent the refined ligand structures.

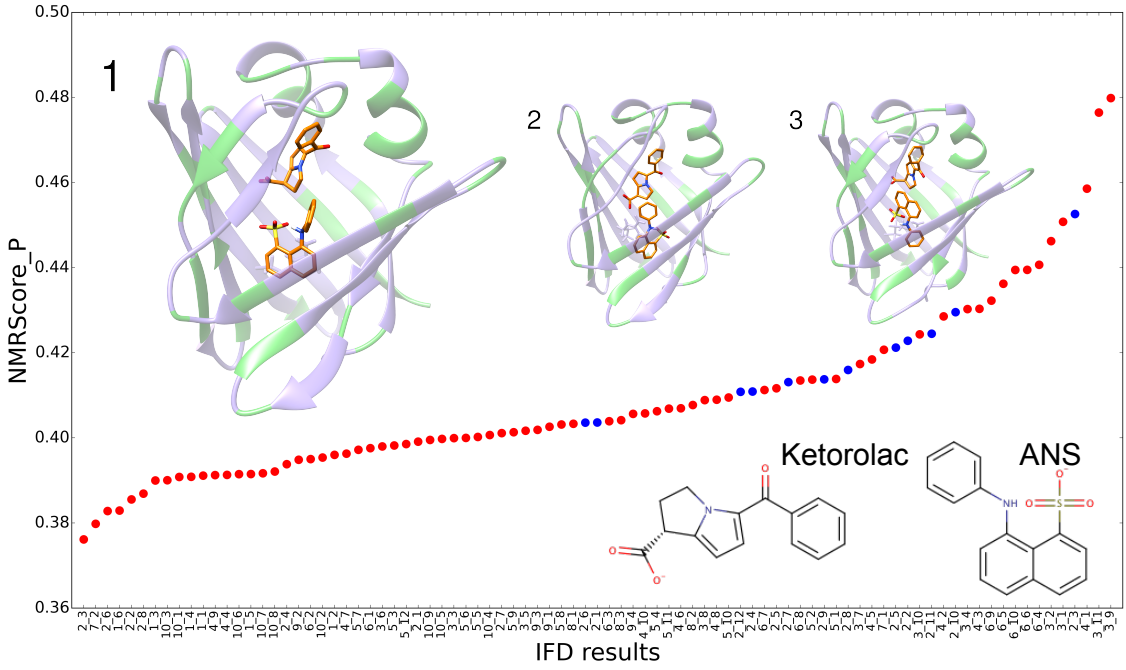


2.4.6. The ternary hIFABP-ketorolac-ANS complex.

The NMRScore_Ps of the 10 models of hIFABP-ketorolac complex in PDB 2MJI ranged from 0.142 to 0.146 ppm (only amide ^{1}H CSPs were available), making them all equally good representatives for this complex. After two independent workflows of IFD simulations (IFD of ANS into all 10 hIFABP-ketorolac models and IFD of ketorolac into the best NMRScore_P ranked hIFABP-ANS complex structure using Glide), a series of ternary complexes were generated. NMRScore_Ps were computed for all the generated ternary complexes based on residues that were significantly perturbed. The 3 models of the ternary complex having the best NMRScore_Ps are depicted in Figure 2.13 together with the rankings of all the poses from IFD simulation. The ternary complexes generated by IFD of ANS into the hIFABP-ketorolac complex are displayed as red dots in the scatterplot in Figure 2.13, whereas, the structures generated with the other workflow are displayed as blue dots. It turned out that the complexes generated from IFD of ANS into the

hIFABP-ketorolac models were better ranked by NMRScore_P (from 0.376 to 0.480 ppm) than the results of IFD of ketorolac into the hIFABP-ANS complex (from 0.404 to 0.453 ppm). The best model predicted by NMRScore_P is compared to the best IFDScore ranked structure (See Figure 2.14). (1) IFD of ANS into hIFABP-ketorolac. The 1st up to the 42nd ranked complex were obtained with the current IFD workflow. As the top 3 structures show, similar to the NMR-derived model for the hIFABP-ketorolac complex, ketorolac still binds in the "portal region" of hIFABP in the ternary complex. ANS binds at the bottom of hIFABP forming a hydrogen bond with Arg106. Unlike the model provided by Patil *et al.,* ANS adopts different binding conformations in our top 3 predictions: the naphthalene is always aligned with the length of the cavity instead of being orthogonal to it, which makes the ANS interact with residues deeper down in the cavity and partially explains the significant CSPs observed in this region. The ternary complex that Patil *et al.*[83] provided previously was generated from IFD simulation of ANS into the first model in the hIFABP-ketorolac NMR ensembles and selected by the IFD score. Their structure is similar to the 4th ranked structure by NMRScore_P, which is also happens to be the IFD result from the first model in the NMR ensemble. (2) IFD of ketorolac into hIFABP-ANS. The best NMRScore_P ranked ternary complex structure from this IFD workflow is, however, ranked as the 43rd over all models considered herein (See Figure 2.15). The corresponding NMRScore_P (0.404 ppm) shows that it is not the best fit to the experimentally observed CSPs. We can see that, there is no hydrogen bond formed between ANS and Arg106. The binding positions of ANS and ketorolac are also different in this model in that ANS is also binding in the upper part of the barrel so that pushing ketorolac slightly towards the top of the cavity. Since both ligands bind to the upper portion of the cavity, the origin of significant CSPs buried deeper into the pocket cannot be explained.

Figure 2.13. NMRScore_Ps and rank for ternary hIFABP-ketorolac-ANS complex structures generated from IFD. The label for the x-axis is the original model number together with the pose number assigned by IFD. The top three structures are depicted with its ranking. hIFABP is shown as ribbon and the locations that were observed to have significantly perturbed protons are mapped onto the hIFABP structure in green. The red dots in the scatterplot represent the structures obtained by IFD of ANS into hIFABP-ketorolac complex, whereas, the blue dots represent the ones got by IFD of ketorolac into hIFABP-ANS complex.



2.5. Conclusion

Considerable effort has been expended to develop empirical models to calculate intra-protein or protein-protein complex chemical shifts, all of which can generally well reproduce experimental values.[49] A notable gap though has been models that can reproduce chemical shifts changes in proteins as a result of the binding of a ligand. This technology gap is largely the result of limited experimental data that can be used to develop just such a model. Herein we developed the HECSP approach, which is designed to predict the CSPs of protein protons induced by ligand binding. In order to build the model we have built a data set using AF-QM/MM calculations on a selection of

protein-ligand complexes. The HECSP model consists of several empirical terms that address various effects that affect chemical shifts: the ring current effect, the electric field effect, the hydrogen bond effect and the magnetic anisotropy group contribution. Because of its empirical nature the model is very fast making it widely applicable. In general, HECSP computed values agreed well with the AF-QM/MM calculations for all the systems in our database. The resultant correlation coefficient is 0.948 and the RMSE is 0.238 ppm for ^{1}H CSPs when compared to the AF-QM/MM calculations.

The results of two studies on the *apo*NCS-naphthoate ester and hIFABP-ketorolac-ANS systems demonstrate that an NMRScore_P strategy for protein-ligand complexes, which is built upon HECSP, can be readily applied to solution NMR structures. In particular, we show the method can distinguish native ligand poses from decoys and refine protein-ligand complex structures. We provide further refined models for both complexes, which satisfy the observed ^{1}H CSPs in experiments.

2.6. Acknowledgments

We thank the National Institutes of Health (R44GM099411) for supporting the research described herein. The authors would like to thank high performance computing center (HPCC) at Michigan State University for providing computational resources. We would like to thank Dr. Dhruva K. Chakravorty for providing a bash script for preparing the AF-QM/MM input files.

2.7. Supporting Information

Table 2.5. List of protein proton types and corresponding parameters in Gasteiger_LS and Gasteiger_LOOCV parameter sets (except the ring current intensity factors).[a]

Type	Gasteiger_LS					Gasteiger_LOOCV									
	F	k	C	a	b	F	$\hat{\sigma}_F$	k	$\hat{\sigma}_k$	C	$\hat{\sigma}_C$	a	$\hat{\sigma}_a$	b	$\hat{\sigma}_b$
HA	22.416	12.399	1.310	12.721	-0.357	22.155	0.045	12.389	0.009	1.324	0.010	12.749	0.122	-0.357	0.007
HN	27.706	13.234	3.622	21.473	-1.213	27.624	0.042	13.287	0.014	3.703	0.018	21.519	0.058	-1.234	0.008
H	13.671	13.828	1.183	20.387	-0.905	13.690	0.119	13.829	0.023	1.229	0.035	20.379	0.034	-0.903	0.003
H1	17.749	13.672	2.138	--	--	17.404	0.083	13.657	0.009	2.235	0.021	--	--	--	--
H4	21.939	12.656	4.524	--	--	21.817	0.088	12.657	0.026	4.520	0.048	--	--	--	--
H5	12.563	14.232	4.216	--	--	12.771	0.228	14.263	0.062	4.123	0.038	--	--	--	--
HC	14.665	16.762	1.633	--	--	14.536	0.064	16.755	0.010	1.642	0.005	--	--	--	--
HP	29.683	18.547	1.692	--	--	29.074	0.328	18.560	0.045	1.649	0.009	--	--	--	--
Har	19.583	13.995	1.072	--	--	19.224	0.069	13.954	0.022	1.083	0.006	--	--	--	--
HO	20.781	15.639	1.246	18.439	-0.564	22.006	0.265	15.855	0.042	1.168	0.038	18.014	0.078	-0.533	0.008

[a]F is in units of ppm; k is in units of ppm•Å^2/e; C, a and b are unitless constant factors.

Table 2.6. List of aromatic ring types and corresponding ring current intensity factors for the Gasteiger_LS and Gasteiger_LOOCV parameter sets.[a]

Aromatic ring	Gasteiger_LS	Gasteiger_LOOCV	
	I	I	$\hat{\sigma}_I$
Tetrazole	1.020	1.020	0.002
Imidazole	1.024	1.021	0.002
Pyrazole	1.088	1.084	0.010
Thiophene	0.721	0.756	0.006
Oxazole	0.836	0.769	0.003
Pyridine	0.990	1.028	0.023
Pyrimidine	2.763	2.797	0.007

[a]Ring current intensity factors are unitless.

Table 2.7. List of protein proton types and corresponding parameters in AM1-BCC_LS parameter set (except the ring current intensity factors).[a]

Type	AM1-BCC_LS				
	F	k	C	a	b
HA	20.201	13.506	0.848	10.392	-0.231
HN	25.016	16.446	1.399	16.915	-1.082
H	15.108	15.108	0.590	15.303	-0.651
HC	14.109	18.232	1.237	--	--
H1	17.821	16.925	1.806	--	--
Har	18.455	13.790	0.664	--	--
H4	22.564	11.116	4.818	--	--
H5	11.329	13.003	1.188	--	--
HP	32.637	18.274	1.495	--	--
HO	19.353	20.000	1.689	15.224	-0.550

[a]F is in units of ppm; k is in units of ppm•$Å^2$/e; C, a and b are unitless constant factors.

Table 2.8. List of aromatic ring types and corresponding ring current intensity factors in the AM1-BCC_LS parameter set.[a]

Aromatic ring	AM1-BCC_LS
	I
Tetrazole	1.186
Imidazole	1.310
Pyrazole	0.939
Thiophene	0.864
Oxazole	1.111
Pyridine	1.076
Pyrimidine	2.593

[a]Ring current intensity factors are unitless.

Figure 2.14. hIFABP is shown as light gray ribbon and the locations that were observed to have significantly perturbed protons are mapped onto the hIFABP structure in green. The dim gray ligands in the center represent the best IDFScore poses of ketorolac and ANS. Whereas, the orange ligands are the best NMRScore_P poses.

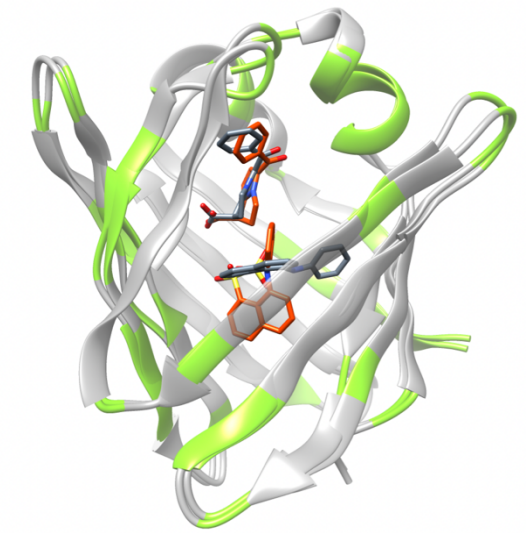
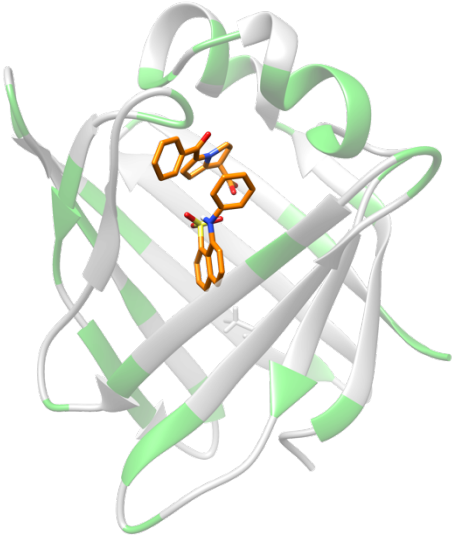


Figure 2.15. Best NMRScore_P ranked ternary hIFABP-ketorolac-ANS complex structure generated from IFD of ketorolac into the hIFABP-ANS complex.



REFERENCES

REFERENCES

1. Shuker, S. B., Hajduk, P. J., Meadows, R. P., and Fesik, S. W. (1996) Discovering high-affinity ligands for proteins: SAR by NMR, *Science*. *274*, 1531-1534.

2. McCoy, M. A., and Wyss, D. F. (2002) Spatial localization of ligand binding sites from electron current density surfaces calculated from NMR chemical shift perturbations, *J Am Chem Soc*. *124*, 11758-11763.

3. Williamson, M. P. (2013) Using chemical shift perturbation to characterise ligand binding, *Prog Nucl Magn Reson Spectrosc*. *73*, 1-16.

4. Krzeminski, M., Loth, K., Boelens, R., and Bonvin, A. M. J. J. (2010) SAMPLEX: Automatic mapping of perturbed and unperturbed regions of proteins and complexes, *BMC Bioinf*. *11*, 51-58.

5. Stark, J., and Powers, R. (2008) Rapid protein-ligand costructures using chemical shift perturbations, *J Am Chem Soc*. *130*, 535-545.

6. Dominguez, C., Boelens, R., and Bonvin, A. M. J. J. (2003) HADDOCK: A protein-protein docking approach based on biochemical or biophysical information, *J Am Chem Soc*. *125*, 1731-1737.

7. Schieborr, U., Vogtherr, M., Elshorst, B., Betz, M., Grimme, S., Pescatore, B., Langer, T., Saxena, K., and Schwalbe, H. (2005) How much NMR data is required to determine a protein-ligand complex structure?, *Chembiochem*. *6*, 1891-1898.

8. Clore, G. M., and Schwieters, C. D. (2003) Docking of protein-protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from H-1(N)/N-15 chemical shift mapping and backbone N-15-H-1 residual dipolar couplings using conjoined rigid body/torsion angle dynamics, *J Am Chem Soc*. *125*, 2902-2912.

9. Palma, P. N., Krippahl, L., Wampler, J. E., and Moura, J. J. G. (2000) BiGGER: A new (soft) docking algorithm for predicting protein interactions, *Proteins: Struct, Funct, Genet*. *39*, 372-384.

10. Morelli, X. J., Palma, P. N., Guerlesquin, F., and Rigby, A. C. (2001) A novel approach for assessing macromolecular complexes combining soft-docking calculations with NMR data, *Protein Sci*. *10*, 2131-2137.

11. Morelli, X., Dolla, A., Czjzek, M., Palma, P. N., Blasco, F., Krippahl, L., Moura, J. J. G., and Guerlesquin, F. (2000) Heteronuclear NMR and soft docking: An experimental

approach for a structural model of the cytochrome c(553)-ferredoxin complex, *Biochemistry*. *39*, 2530-2537.

12. Medek, A., Hajduk, P. J., Mack, J., and Fesik, S. W. (2000) The use of differential chemical shifts for determining the binding site location and orientation of protein-bound ligands, *J Am Chem Soc*. *122*, 1241-1242.

13. Lugovskoy, A. A., Degterev, A. I., Fahmy, A. F., Zhou, P., Gross, J. D., Yuan, J. Y., and Wagner, G. (2002) A novel approach for characterizing protein ligand complexes: Molecular basis for specificity of small-molecule Bcl-2 inhibitors, *J Am Chem Soc*. *124*, 1234-1240.

14. Riedinger, C., Endicott, J. A., Kemp, S. J., Smyth, L. A., Watson, A., Valeur, E., Golding, B. T., Griffin, R. J., Hardcastle, I. R., Noble, M. E., and McDonnell, J. M. (2008) Analysis of Chemical Shift Changes Reveals the Binding Modes of Isoindolinone Inhibitors of the MDM2-p53 Interaction, *J Am Chem Soc*. *130*, 16038-16044.

15. Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J. M., Liu, G. H., Eletsky, A., Wu, Y. B., Singarapu, K. K., Lemak, A., Ignatchenko, A., Arrowsmith, C. H., Szyperski, T., Montelione, G. T., Baker, D., and Bax, A. (2008) Consistent blind protein structure generation from NMR chemical shift data, *Proc Natl Acad Sci U S A*. *105*, 4685-4690.

16. Cavalli, A., Salvatella, X., Dobson, C. M., and Vendruscolo, M. (2007) Protein structure determination from NMR chemical shifts, *Proc Natl Acad Sci U S A*. *104*, 9615-9620.

17. Cavalli, A., Montalvao, R. W., and Vendruscolo, M. (2011) Using Chemical Shifts to Determine Structural Changes in Proteins upon Complex Formation, *J Phys Chem B*. *115*, 9491-9494.

18. McCoy, M. A., and Wyss, D. F. (2000) Alignment of weakly interacting molecules to protein surfaces using simulations of chemical shift perturbations, *J Biomol NMR*. *18*, 189-198.

19. Cioffi, M., Hunter, C. A., Packer, M. J., and Spitaleri, A. (2008) Determination of protein-ligand binding modes using complexation-induced changes in H-1 NMR chemical shift, *J Med Chem*. *51*, 2512-2517.

20. Cioffi, M., Hunter, C. A., Packer, M. J., Pandya, M. J., and Williamson, M. P. (2009) Use of quantitative (1)H NMR chemical shift changes for ligand docking into barnase, *J Biomol NMR*. *43*, 11-19.

21. Cioffi, M., Hunter, C. A., and Packer, M. J. (2008) Influence of conformational flexibility on complexation-induced changes in chemical shift in a neocarzinostatin protein - Ligand complex, *J Med Chem*. *51*, 4488-4495.

22. Gonzalez-Ruiz, D., and Gohlke, H. (2009) Steering Protein-Ligand Docking with Quantitative NMR Chemical Shift Perturbations, *J Chem Inf Model*. *49*, 2260-2271.

23. Aguirre, C., ten Brink, T., Cala, O., Guichou, J. F., and Krimm, I. (2014) Protein-ligand structure guided by backbone and side-chain proton chemical shift perturbations, *J Biomol NMR*. *60*, 147-156.

24. ten Brink, T., Aguirre, C., Exner, T. E., and Krimm, I. (2015) Performance of Protein-Ligand Docking with Simulated Chemical Shift Perturbations, *J Chem Inf Model*. *55*, 275-283.

25. Hartman, J. D., and Beran, G. J. O. (2014) Fragment-Based Electronic Structure Approach for Computing Nuclear Magnetic Resonance Chemical Shifts in Molecular Crystals, *J Chem Theory Comput*. *10*, 4862-4872.

26. Flaig, D., Maurer, M., Hanni, M., Braunger, K., Kick, L., Thubauville, M., and Ochsenfeld, C. (2014) Benchmarking Hydrogen and Carbon NMR Chemical Shifts at HF, DFT, and MP2 Levels, *J Chem Theory Comput*. *10*, 572-578.

27. Moon, S., and Case, D. A. (2006) A comparison of quantum chemical models for calculating NMR shielding parameters in peptides: Mixed basis set and ONIOM methods combined with a complete basis set extrapolation, *J Comput Chem*. *27*, 825-836.

28. Xu, X. P., and Case, D. A. (2002) Probing multiple effects on N-15, C-13 alpha, C-13 beta, and C-13 ' chemical shifts in peptides using density functional theory, *Biopolymers*. *65*, 408-423.

29. Arnold, W. D., and Oldfield, E. (2000) The chemical nature of hydrogen bonding in proteins via NMR: J-couplings, chemical shifts, and AIM theory, *J Am Chem Soc*. *122*, 12835-12841.

30. Scheurer, C., Skrynnikov, N. R., Lienin, S. F., Straus, S. K., Bruschweiler, R., and Ernst, R. R. (1999) Effects of dynamics and environment on N-15 chemical shielding anisotropy in proteins. A combination of density functional theory, molecular dynamics simulation, and NMR relaxation, *J Am Chem Soc*. *121*, 4242-4251.

31. Sitkoff, D., and Case, D. A. (1997) Density functional calculations of proton chemical shifts in model peptides, *J Am Chem Soc*. *119*, 12262-12273.

32. Dedios, A. C., Pearson, J. G., and Oldfield, E. (1993) Secondary and Tertiary Structural Effects on Protein Nmr Chemical-Shifts - an Abinitio Approach, *Science*. *260*, 1491-1496.

33. Dedios, A. C. (1993) Secondary and Tertiary Structural Effects on Protein Nmr Chemical-Shifts - an Ab-Initio Approach (Vol 260, Pg 1491, 1993), *Science*. *261*, 535-535.

34. He, X., Wang, B., and Merz, K. M., Jr. (2009) Protein NMR chemical shift calculations based on the automated fragmentation QM/MM approach, *J Phys Chem B*. *113*, 10380-10388.

35. Zhu, T., He, X., and Zhang, J. Z. H. (2012) Fragment density functional theory calculation of NMR chemical shifts for proteins with implicit solvation, *Phys Chem Chem Phys*. *14*, 7837-7845.

36. Zhu, T., Zhang, J. Z. H., and He, X. (2013) Automated Fragmentation QM/MM Calculation of Amide Proton Chemical Shifts in Proteins with Explicit Solvent Model, *J Chem Theory Comput*. *9*, 2104-2114.

37. Cui, Q., and Karplus, M. (2000) Molecular properties from combined QM/MM methods. 2. Chemical shifts in large molecules, *J Phys Chem B*. *104*, 3721-3743.

38. He, X., Zhu, T., Wang, X. W., Liu, J. F., and Zhang, J. Z. H. (2014) Fragment Quantum Mechanical Calculation of Proteins and Its Applications, *Acc Chem Res*. *47*, 2748-2757.

39. Swails, J., Zhu, T., He, X., and Case, D. A. (2015) AFNMR: automated fragmentation quantum mechanical calculation of NMR chemical shifts for biomolecules, *J Biomol NMR*. *63*, 125-139.

40. Dracinsky, M., Moller, H. M., and Exner, T. E. (2013) Conformational Sampling by Ab Initio Molecular Dynamics Simulations Improves NMR Chemical Shift Predictions, *J Chem Theory Comput*. *9*, 3806-3815.

41. Exner, T. E., Frank, A., Onila, I., and Moller, H. M. (2012) Toward the Quantum Chemical Calculation of NMR Chemical Shifts of Proteins. 3. Conformational Sampling and Explicit Solvents Model, *J Chem Theory Comput*. *8*, 4818-4827.

42. Frank, A., Onila, I., Moller, H. M., and Exner, T. E. (2011) Toward the quantum chemical calculation of nuclear magnetic resonance chemical shifts of proteins, *Proteins*. *79*, 2189-2202.

43. Gao, Q., Yokojima, S., Fedorov, D. G., Kitaura, K., Sakurai, M., and Nakamura, S. (2010) Fragment-Molecular-Orbital-Method-Based ab Initio NMR Chemical-Shift Calculations for Large Molecular Systems, *J Chem Theory Comput*. *6*, 1428-1444.

44. Gao, Q., Yokojima, S., Kohno, T., Ishida, T., Fedorov, D. G., Kitaura, K., Fujihira, M., and Nakamura, S. (2007) Ab initio NMR chemical shift calculations on proteins using fragment molecular orbitals with electrostatic environment, *Chem Phys Lett*. *445*, 331-339.

45. Osapay, K., and Case, D. A. (1991) A New Analysis of Proton Chemical-Shifts in Proteins, *J Am Chem Soc*. *113*, 9436-9444.

46. Xu, X. P., and Case, D. A. (2001) Automated prediction of (15)N, (13)C(alpha), (13)C(beta) and (13)C ' chemical shifts in proteins using a density functional database, *J Biomol NMR*. *21*, 321-333.

47. Moon, S., and Case, D. A. (2007) A new model for chemical shifts of amide hydrogens in proteins, *J Biomol NMR*. *38*, 139-150.

48. Neal, S., Nip, A. M., Zhang, H. Y., and Wishart, D. S. (2003) Rapid and accurate calculation of protein H-1, C-13 and N-15 chemical shifts, *J Biomol NMR*. *26*, 215-240.

49. Han, B., Liu, Y. F., Ginzinger, S. W., and Wishart, D. S. (2011) SHIFTX2: significantly improved protein chemical shift prediction, *J Biomol NMR*. *50*, 43-57.

50. Shen, Y., and Bax, A. (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology, *J Biomol NMR*. *38*, 289-302.

51. Shen, Y., and Bax, A. (2010) SPARTA plus : a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network, *J Biomol NMR*. *48*, 13-22.

52. Kohlhoff, K. J., Robustelli, P., Cavalli, A., Salvatella, X., and Vendruscolo, M. (2009) Fast and Accurate Predictions of Protein NMR Chemical Shifts from Interatomic Distances, *J Am Chem Soc*. *131*, 13894-13895.

53. Meiler, J. (2003) PROSHIFT: Protein chemical shift prediction using artificial neural networks, *J Biomol NMR*. *26*, 25-37.

54. Meiler, J., and Baker, D. (2003) Rapid protein fold determination using unassigned NMR data, *Proc Natl Acad Sci U S A*. *100*, 15404-15409.

55. Williamson, M. P., and Craven, C. J. (2009) Automated protein structure calculation from NMR data, *J Biomol NMR*. *43*, 131-143.

56. Robustelli, P., Stafford, K. A., and Palmer, A. G. (2012) Interpreting Protein Structural Dynamics from NMR Chemical Shifts, *J Am Chem Soc*. *134*, 6365-6374.

57. Wang, B., Westerhoff, L. M., and Merz, K. M. (2007) A critical assessment of the performance of protein-ligand scoring functions based on NMR chemical shift perturbations, *J Med Chem*. *50*, 5128-5134.

58. Stratmann, D., Boelens, R., and Bonvin, A. M. J. J. (2011) Quantitative use of chemical shifts for the modeling of protein complexes, *Proteins*. *79*, 2662-2670.

59. Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and

Shenkin, P. S. (2004) Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, *J Med Chem*. *47*, 1739-1749.

60. Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., and Banks, J. L. (2004) Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening, *J Med Chem*. *47*, 1750-1759.

61. Li, Y., Liu, Z., Li, J., Han, L., Liu, J., Zhao, Z., and Wang, R. (2014) Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set, *J Chem Inf Model*. *54*, 1700-1716.

62. Gordon, J. C., Myers, J. B., Folta, T., Shoja, V., Heath, L. S., and Onufriev, A. (2005) H++: a server for estimating pKas and adding missing hydrogens to macromolecules, *Nucleic Acids Res*. *33*, W368-W371.

63. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters, *Proteins*. *65*, 712-725.

64. 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., and Kollman, P. A. (1996) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995), *J Am Chem Soc*. *118*, 2309-2309.

65. Wang, J. M., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004) Development and testing of a general amber force field, *J Comput Chem*. *25*, 1157-1174.

66. D.A. Case, J. T. B., R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K.M. Merz, G. Monard, P. Needham, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, R. Salomon-Ferrer, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, D.M. York and P.A. Kollman (2015), AMBER 2015. *University of California, San Francisco*.

67. M. J. Frisch, G. W. T., H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A.

D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2009.

68. 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., and Kollman, P. A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules *J Am Chem Soc*. *117*, 5179-5197.

69. Pople, J. A. (1956) Proton Magnetic Resonance of Hydrocarbons, *J Chem Phys*. *24*, 1111-1111.

70. Johnson, C. E., and Bovey, F. A. (1958) Calculation of Nuclear Magnetic Resonance Spectra of Aromatic Hydrocarbons, *J Chem Phys*. *29*, 1012-1014.

71. Haigh, C. W., and Mallion, R. B. (1979) Ring Current Theories in Nuclear Magnetic-Resonance, *Prog Nucl Magn Reson Spectrosc*. *13*, 303-344.

72. Perkins, S. J., and Dwek, R. A. (1980) Comparisons of Ring-Current Shifts Calculated from the Crystal-Structure of Egg-White Lysozyme of Hen with the Proton Nuclear Magnetic-Resonance Spectrum of Lysozyme in Solution, *Biochemistry*. *19*, 245-258.

73. Buckingham, A. D. (1960) Chemical Shifts in the Nuclear Magnetic Resonance Spectra of Molecules Containing Polar Groups, *Can J Chem*. *38*, 300-307.

74. Wang, J. M., Wang, W., Kollman, P. A., and Case, D. A. (2006) Automatic atom type and bond type perception in molecular mechanical calculations, *J Mol Graphics Modell*. *25*, 247-260.

75. Wagner, G., Pardi, A., and Wüthrich, K. (1983) Hydrogen-Bond Length and H-1-Nmr Chemical-Shifts in Proteins, *J Am Chem Soc*. *105*, 5948-5949.

76. Wishart, D. S., Sykes, B. D., and Richards, F. M. (1991) Relationship between Nuclear-Magnetic-Resonance Chemical-Shift and Protein Secondary Structure, *J Mol Biol*. *222*, 311-333.

77. Mcconnell, H. M. (1957) Theory of Nuclear Magnetic Shielding in Molecules .1. Long-Range Dipolar Shielding of Protons, *J Chem Phys*. *27*, 226-229.

78. Flygare, W. H. (1974) Magnetic-Interactions in Molecules and an Analysis of Molecular Electronic Charge Distribution from Magnetic Parameters, *Chem Rev*. *74*, 653-687.

79. Schmalz, T. G., Norris, C. L., and Flygare, W. H. (1973) Localized Magnetic Susceptibility Anisotropies, *J Am Chem Soc*. *95*, 7961-7967.

80. Sutter, D. H., and Flygare, W. H. (1976) The molecular Zeeman effect, In *Bonding Structure* (Craig, D. P., Mellor, D. P., Gleiter, R., Gygax, R., Sutter, D. H., and Flygare, W. H., Eds.), pp 89-196, Springer Berlin Heidelberg, Berlin, Heidelberg.
81. Urbaniak, M. D., Muskett, F. W., Finucane, M. D., Caddick, S., and Woolfson, D. N. (2002) Solution structure of a novel chromoprotein derived from apo-Neocarzinostatin and a synthetic chromophore, *Biochemistry*. *41*, 11731-11739.

82. Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A., and Farid, R. (2006) Novel procedure for modeling ligand/receptor induced fit effects, *J Med Chem*. *49*, 534-553.

83. Patil, R., Laguerre, A., Wielens, J., Headey, S. J., Williams, M. L., Hughes, M. L. R., Mohanty, B., Porter, C. J. H., and Scanlon, M. J. (2014) Characterization of Two Distinct Modes of Drug Binding to Human Intestinal Fatty Acid Binding Protein, *ACS Chem Biol*. *9*, 2526-2534.

CHAPTER 3

The Extended Zinc AMBER Force Field (EZAFF)

† Reprinted (adapted) with permission from Yu, Z., Li, P. & Merz, K. M. Extended Zinc AMBER Force Field (EZAFF). J. Chem. Theory Comput. 14, 242–254 (2018).

3.1. Abstract

An empirical approach based on the previously developed zinc AMBER force field (ZAFF) is proposed for the determination of the parameters for bonds and angles involving zinc. We call it the extended ZAFF (EZAFF) model because the original ZAFF model was only formulated for 4-coordinated systems, while EZAFF additionally can tackle 5- and 6-coordinated systems. Tests were carried out for 6 metalloproteins and 6 organometallic compounds with different coordination spheres. Results validated the reliability of the current model to handle a variety of zinc containing complexes. Meanwhile, benchmark calculations were performed to assess the performance of 3 bonded molecular mechanics models (EZAFF, Seminario, and Z-matrix models), 4 nonbonded parameter sets (the HFE, IOD, CM and 12-6-4 models) and 4 semiempirical quantum mechanical methods (AM1, PM3, PM6 and SCC-DFTB methods) for simulating zinc containing systems. The obtained results indicate that, even with their increased computational cost, the semiempirical quantum methods only offered slightly better accuracy for the computation of relative energies and only afforded similar molecular geometries, when compared to the investigated molecular mechanics models.

3.2. Introduction

As the second most abundant metal in the human body and in other biological systems, zinc plays an essential role in various biological activities in the form of zinc-containing proteins[1]. The zinc sites in proteins can be divided into (1) catalytic sites like in alcohol dehydrogenase which breaks down alcohol, carbonic anhydrase which interconverts carbon dioxide and bicarbonate, and carboxypeptidase A which catalyzes peptide cleavage; and (2) structural sites like in zinc fingers, where zinc is critical for correct folding[2]. Given the importance of zinc metalloproteins, much

attention has been paid to force field development and there are a number of nonpolarizable and polarizable models extant.

Generally, there are three major types of nonpolarizable models to model zinc and its ligand sphere in proteins: the nonbonded model, the cationic dummy atom model and the bonded model. The nonbonded model is the simplest model among the three nonpolarizable models. Its potential only consists of the electrostatic and van der Waals (vdW) terms. The nonbonded model presented by Stote and Karplus[3] has been widely used due to its simplicity and efficiency. Babu and Lim parameterized vdW parameters for a set of divalent cations (including Zn^{2+}) which reproduced experimental relative hydration free energies, first-solvation-shell CNs and average ion-water distances at the same time.[4] Wu *et al*. introduced the short-long effective functions (SLEF) approach.[5] The method treated the short-range term with two parameters which were parameterized through force matching based on QM/MM MD simulations, while the long range term decays as 1/r. Results showed that the model could reproduce the different metal coordination modes from a number of crystal structures. However, it was found that the hydration free energy and ion-oxygen distance of the first solvation shell could not be simultaneously reproduced for the zinc ion when using the nonbonded model with the particle mesh Ewald (PME) method.[6] Li and Merz proposed a 12-6-4 LJ-type nonbonded model for M(II) ions to explicitly take into account the charge-induced dipole interaction.[7] Results demonstrated that the model reproduced the experimental hydration free energy, coordination number and ion-oxygen distance simultaneously. In order to improve the description of the electrostatic interactions between the zinc ion and other atoms in metalloproteins, Pang *et al*. developed the cationic dummy atom model for four-coordinated zinc, which places four covalently bonded dummy atoms around the zinc ion in a

tetrahedral geometry and evenly distributes the +2e charge over these dummy atoms. , The zinc center is only described as a van der Waals core and carries no charge.[8-9] In their study the tetra-coordination of the zinc site was well maintained during their simulations.

In the bonded model, harmonic bond and angle terms are incorporated into the potential energy function to describe the interactions between the metal ion and its ligating groups[10], which leads to the advantage of preserving the coordination environment over the course of the simulation. However, this feature can turn into a drawback in cases where the coordination changes around the metal ion during simulation. In the bonded model, the partial atomic charges are obtained from RESP fitting[11] or CMX models[12] typically resulting in a non-integer charge on the metal ion unlike the formal charge typically used in the nonbonded model. A common way to generate the bond and angle force constants is to derive them using *ab initio* or DFT vibrational analysis. One can either take the diagonal elements in the Hessian matrix with internal coordinates as the force constants or follow the Seminario method to derive these parameters from a Hessian matrix with Cartesian coordinates[13].

Besides these models, there are also polarizable models that have been developed for zinc containing complexes. For example, the SIBFA model developed by Gresh et al,[14-18] with polarization, charge transfer, penetration terms included, can give highly accurate ion-ligand interaction energies, which are comparable to those obtained using ab initio methods. Sakharov and Lim developed the CTPOL model for zinc containing proteins.[19] Their model incorporated polarization and charge transfer (through distance-dependent partial charges) effects for zinc and its ligating atoms, and reproduced the tetrahedral geometries of $Zn[Cys]_2[His]_2$ and $Zn[Cys]_4$ sites.

In the AMOEBA polarizable force field, the polarization effect is described via an induced dipole model and AMOEBA has been applied to the zinc-water system by Wu and coworkers.[20] They found that AMOEBA provided a robust estimation of the hydration free energy along with reasonable solvation structure and dynamics. Zhang *et al*. examined the AMOEBA force field for its ability to model zinc−containing proteins with the simulations yielding reasonable coordination structures and relative binding free energies.[21] Xiang and Ponder incorporated a valence bond model into the AMOEBA force field of Cu^{2+} and Zn^{2+}.[22] Considerable improvement was realized for the Cu^{2+} ion, while a trivial influence was observed for Zn^{2+} when applying the valence bond model. However, even higher accuracy was observed when using a polarizable model, even though nonpolarizable models are more widely used due to their advantage in functional simplicity and computational speed. For more extensive discussion on metal ion modeling see Li and Merz.[23]

The zinc AMBER Force Field (ZAFF) was developed by Peters *et al*[24] and It was specifically designed for simulating tetra-coordinated zinc sites with the bonded model. A number of validation tests were also performed with the results validating the reliability of ZAFF on a number of zinc containing systems, though later work showed deficiencies of the model for systems containing adjacent water molecules around the metal ion.[25] In the present study, we developed the extended ZAFF (EZAFF) models based on the empirical trends found in the ZAFF parameters and applied it to a broader range of systems (12 systems with a variety of coordination modes) to good effect. Meanwhile, benchmark calculations were performed amongst seven MM models (EZAFF, Z-matrix method, Seminario method, HFE, IOD, CM and 12-6-4) and four widely used semiempirical quantum models (AM1, PM3, PM6 and SCC-DFTB) for their accuracy on predicting the energetic and structural properties of zinc complexes. The obtained results indicate

that, even with their increased computational cost, the semiempirical quantum methods only offered slightly better accuracy for the computation of relative energies and only afforded similar molecular geometries, when compared to the molecular mechanics models.

3.3. Methods

3.3.1 Development and validation of the empirical scheme

3.3.1.1. Empirical scheme for deriving the bond and angle parameters

One intention of present research is to provide an empirical method to facilitate the force field parameterization of the bonded model for various zinc-containing systems. This idea originates from the parameter trends founded in the zinc AMBER force field (ZAFF).[24] ZAFF was developed by Peters *et al*.[24] based on the Seminario method.[13] The parameters employed in the present work are derived from the updated version of ZAFF of August 2011, which is part of the AMBER distribution ("$AMBERHOME/dat/mtkpp/ZAFF/201108/"). After checking the zinc related bond parameters, we found clear trends between the equilibrium bond lengths and bond force constants between Zn-N, Zn-O and Zn-S bonds (see Figure 3.1). In general, the force constant anti-correlates with the equilibrium bond length. Three fitting strategies (linear, quadratic and exponential) were tested to fit the data between the equilibrium bond lengths and bond force constants. The fitting results are shown in Table 3.1. We find that the exponential fitting gives the largest Pearson's r and R^2 and smallest RMSE values, outperforming the quadratic and linear fittings. Even though the improvement of exponential fitting over quadratic fitting is not dramatic, the former fitting curves have monotonic characteristics (compared to quadratic fitting which has minima), making it a much more favorable choice.

Figure 3.1. Fits of the ZAFF bond stretching force constants for the Zn-N, Zn-S, and Zn-O bond types.

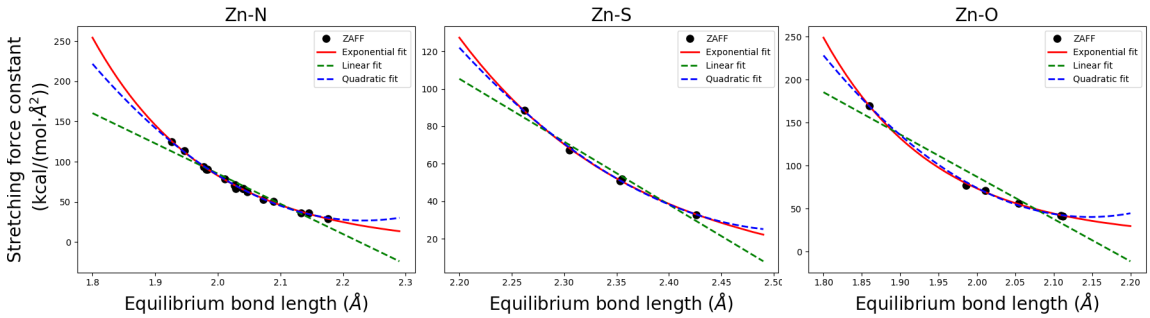


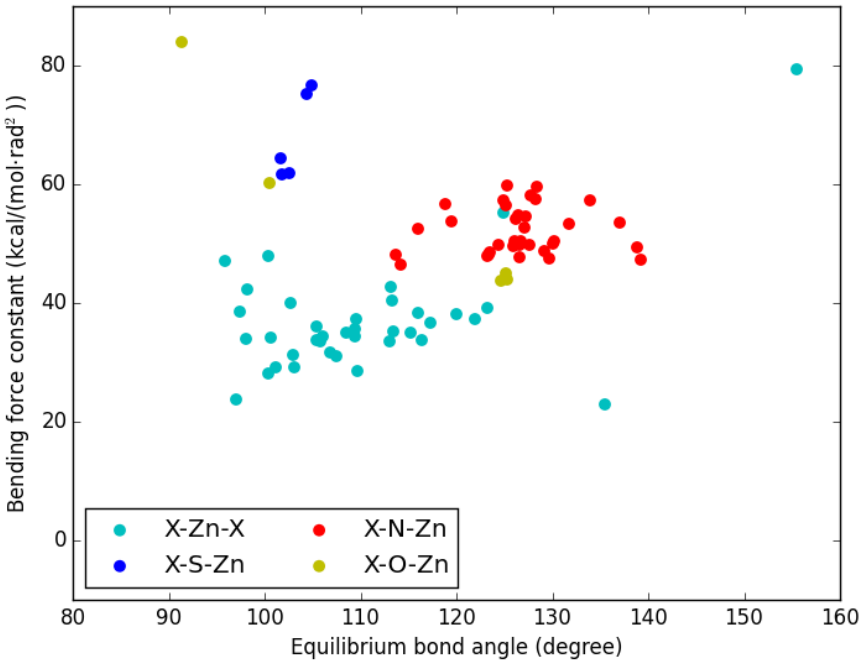
Table 3.1. Pearson's Correlation Coefficients, R^2 and Root Mean Square Errors (RMSEs) of three options of fitting to ZAFF bond stretching force constants for Zn-N, Zn-S, and Zn-O bond types.

Bond type	Pearson's r			R^2			RMSE		
	Exponential	Linear	Quadratic	Exponential	Linear	Quadratic	Exponential	Linear	Quadratic
Zn - N	0.9989	0.9766	0.9988	0.9978	0.9537	0.9976	1.2046	5.5060	1.2412
Zn - S	0.9993	0.9871	0.9990	0.9986	0.9744	0.9980	0.6898	2.9890	0.8356
Zn - O	0.9994	0.9665	0.9988	0.9989	0.9342	0.9976	1.4817	11.2213	2.1598

After checking the zinc related angle parameters in ZAFF, we found that, unlike the bond parameters, there is no obvious correlation (or anti-correlation) between the equilibrium angle values and angle force constants (see Figure 3.2). However, we found that the angle force constants in ZAFF have a pattern similar to that found in the general AMBER force field (GAFF). The authors of GAFF found that the A-B-C angles where both A and C are hydrogen atoms have force constants approximately 30-35 kcal/mol•rad^{-2}, with either A or C is a hydrogen atom the force constant is ~50 kcal/mol•rad^{-2}, and the remaining cases have force constants of ~70 kcal/mol•rad^{-2}[2.26]. For ZAFF we found that if the central atom B is zinc, the angle force constant is ~35

kcal/mol•rad^{-2}, and if zinc is a terminal atom in the angle and coordinated to an oxygen/nitrogen atom, the force constant is ~50 kcal/mol•rad^{-2}, while if zinc is terminal and coordinated to a sulfur atom, the force constant is ~70 kcal/mol•rad^{-2}. Hence in our empirical approach we assigned the angle force constants as 35, 50 and 70 kcal/mol•rad^{-2} respectively for the situations enumerated above. We note that this assignment protocol is consistent with the approach employed by GAFF.

Figure 3.2. Scatter plot of the bending force constant vs equilibrium bond angle for the Zn containing bond. As shown in the legend, X can be any atom and "X-Zn-X" doesn't require that the two Xs are the same atom type.



The above scheme has several advantages. First, it is straightforward and fast since it doesn't need any time-consuming QM calculations to derive the bond and angle related parameters. Moreover, it is more broadly applicable: it is applicable to systems for which QM methods may be

problematic (see an example below). Last but not least, as we show below, the accuracy of this scheme, especially for structure prediction, is remarkable.

We call this empirical scheme the extended ZAFF (EZAFF) model since it has broader applicability than ZAFF, which was designed for a number of four-coordinated zinc containing proteins. The EZAFF scheme has been added to the MCPB.py program[27] and is in AmberTools 15[28]. Herein we use MCPB.py to empirically assign the bonded and angle parameters involving zinc. During parameterization, the equilibrium bond lengths and angles involving zinc are calculated directly from the crystal structures except for Zn-X-H angles which are not in the ideal range (angle ± 5°), in which case the ideal angle is assigned as the equilibrium angle (the ideal value is determined based on the identity of X, e.g. 109.47° for four-coordinated X, 120° for three coordinated X). Next, the corresponding force constants were determined according to the EZAFF scheme. In light of the assumption of Hoops *et al*.[29] and the success of ZAFF,[24] the metal related torsion terms were treated as zero.

3.3.1.2. Partial charge parameters

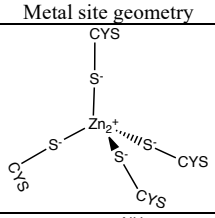
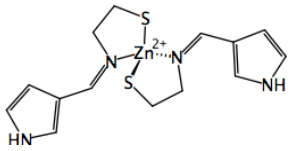
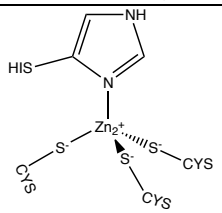
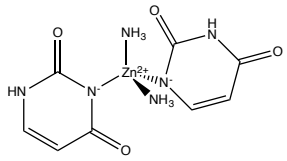
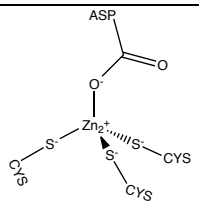
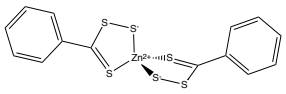
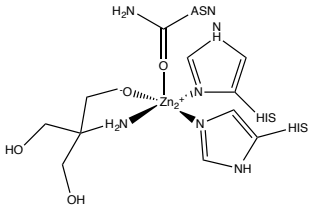
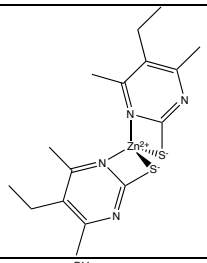
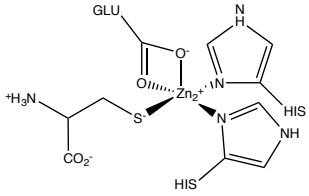
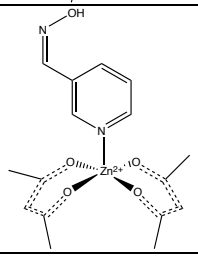
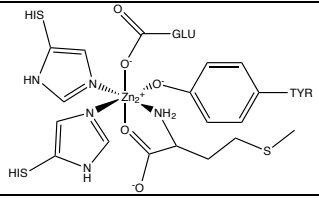
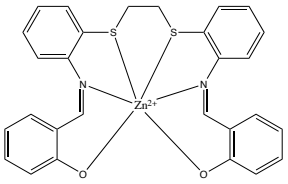
Partial charge parameters for the metal site (metal ion and the atoms in the metal coordinated residues) were obtained from RESP (restrained electrostatic potential)[11, 30-31] fits using the Merz-Singh-Kollman population analysis[32] based on QM calculation at the B3LYP/6-31G* level of theory. The VDW radius of the zinc ion was set at 1.375 Å for the RESP fits. This value was taken from the IOD parameter set (which reproduced the first solvation shell ion-oxygen distance) from Li et al.[6] For each of the metalloprotein systems the QM calculation was performed on the large model built by MCPB.py while for each of the organometallic systems the QM calculation was

performed on the entire structure. The ChgModB scheme in MCPB.py was used for the RESP fitting procedure, and the charges of backbone heavy atoms for metal site amino acid residues were restrained to the corresponding partial charges found in the AMBER ff14SB force field. This choice was made because the ChgModB scheme gave the best performance in ZAFF.[24] The partial charges of zinc ions in these complexes were found to be less than +2e, which is consistent with earlier work[24, 29] and an extended Born model proposed by Heinz and Suter[33].

3.3.1.3. Structure selection

In order to assess the present empirical model, tests were performed on 12 different systems (see Table 3.2), in which 6 are crystal structures of zinc-containing proteins from the protein data bank (PDB)[41] (the MESPEUS database[42] was utilized to facilitate the selection) and 5 are zinc containing organometallic structures taken from the Cambridge structural database (CSD)[43]. There is only one zinc ion in the binding pocket of all the structures. The resolution of the 6 metalloprotein structures were all ≤ 2.0 Å. These structures covered a variety of coordination modes in different chemical environments: the coordination number covers 4 to 6 and with different ligands involved. The assessment consisted of evaluating the force field reliability along with MD simulations for the metalloprotein systems and for the organometallic compounds we evaluated the ability of the force field to predict the molecular geometry and the corresponding vibrational frequencies. The results showed that the empirical method can be applied to a much wider range of zinc complexes than the original ZAFF model (see below).

Table 3.2. Twelve Zn-containing model systems considered in present study.

PDB Metalloprotein structures				CSD zinc ligand complex structures		
PDB code	Resolution	Charge	Metal site geometry	CSD entry	Charge	Geometry
1PZW[34]	2.0 Å	-2		AHOQIY[35]	0	
2AP1	1.9 Å	-1		BEZKOH[36]	0	
1P3J[33]	1.9 Å	-2		ZNTPBZ[37]	0	
1H4N[38]	2.0 Å	1		KUBVOT[39]	0	
1F57[23]	1.8 Å	0		ABOWOF[27]	0	
1Y9Q	1.9 Å	-1		EGIXOH[40]	0	

3.3.1.4. MD validations on the metalloproteins

MD simulations were performed on all 6 metalloprotein systems. Each system was solvated using a cubic box with box edges in each dimension of at least 10 Å away from any atom in the solute. An appropriate number of counter-ions were added to neutralize the system. The ff14SB force field and GAFF were used to model the amino acid residues and ligands, respectively. The metal site was modeled with the approach summarized above.

The system was minimized using a four-stage procedure to eliminate bad contacts: first, we minimized the solvent via imposition of harmonic positional restraints of 200.0 kcal/(mol•Å^2) on all atoms in the metalloprotein. Secondly, we imposed the same magnitude of restraint on all the non-hydrogen atoms in the metalloprotein, while in the third minimization step only the backbone heavy atom restraints remained. In each of these three stages, 2000 deepest descent steps followed by 1000 conjugate gradient steps were performed. Subsequently, the whole system was minimized using steepest descent for 5000 steps and conjugate gradient for 2000 steps. After the minimization, we equilibrated the entire system using a three-stage procedure. In the first stage, a heating step of 1 ns was performed to heat the system from 0 K to 298.15 K in the NVT ensemble. Afterwards 1 ns of NVT equilibration simulation was performed at 298.15 K. Subsequently a 1 ns NPT ensemble simulation was performed at 298.15 K and 1 atm to correct for the density of the system. Finally a production MD run of 20 ns was carried out using the NVT ensemble at 298.15K. Frames were saved every 10000 steps (10 ps) during the production run, providing 2000 frames for analysis for each metalloprotein. A 1 fs time-step was used during these MD simulations. The Particle Mesh Ewald (PME) method was utilized to handle the long-range electrostatic interactions. All bonds involving hydrogen atoms were constrained using SHAKE[44]. The Langevin thermostat was used

to control the temperature in these MD simulations. The root-mean-square deviation (RMSD) of the heavy atoms in the metal site and the RMSD of the backbone atoms of the entire protein were calculated based on these frames to assess the stability of the force field employed.

3.3.1.5. Validations of the CSD complexes

Besides assessing the performance of EZAFF in MD simulations of metalloproteins, we also evaluated its performance regarding the prediction of molecular geometries and normal mode frequencies of organometallic compounds. The RMSDs of the MM optimized structures with respect to the QM optimized structures (at the B3LYP/6-31G* level of theory) and CSD structures were calculated. These results are summarized in the results and discussion section below. We carried out normal mode frequency calculations for these organometallic compounds based on EZAFF using the NAB module in AMBER 14.[45] The results were compared to the QM calculated results at the B3LYP/6-31G* level of theory and are depicted below. All the B3LYP/6-31G* calculations were performed using the Gaussian 09 program.[46]

3.3.2 Benchmark evaluations of different MM and semi-empirical QM methods for modeling zinc-containing complexes

There are different methods used for deriving force field parameters for metal-containing systems. In the present work we performed benchmark calculations on three bonded MM models (EZAFF, Seminario[13] and Z-matrix approaches), four nonbonded parameter sets (HFE, IOD, CM and 12-6-4) and several popular semi-empirical molecular orbital methods (PM6, PM3, AM1, SCC-DFTB).

The benchmark calculations were performed on the systems shown in Table 3.2. For the 6 metalloprotein systems the small metal-containing active site models were built using MCPB.py as were the 6 organometallic systems (where the entire complex was the "small system"). The benchmark evaluations were performed to explore both the energetic and structural aspects of metal ion force field design. What we looked for in our benchmark calculations was how well these methods (7 MM methods and 4 semi-empirical quantum methods) reproduce QM (B3LYP/6-31G*) calculated relative conformational energies and optimized structures along with the available crystallographic structures.

The MCPB.py program was used to build the metal ion force field according to the EZAFF parameterization scheme (described above). This program was also used for the parameterization of the Seminario and Z-matrix methods based on QM calculations at the B3LYP/6-31G* level of theory. Unlike EZAFF, the equilibrium bond lengths and angle values were taken from the QM optimized structures coupled with the Seminario and Z-matrix methods. Except for the equilibrium values and force constants of the bonds and angles involving metal ions, the remaining parameters are the same for the EZAFF, Seminario and Z-matrix methods: torsion parameters involving metal ions were set to zero; partial charges were obtained from RESP fitting based on B3LYP/6-31G* calculations; the VDW parameter for the zinc ion was taken from Li et al.;[6] and the ligands were modeled using GAFF.

The single-point energy calculations and geometry optimization of the 7 MM methods were performed using the Sander program and the NAB module in AmberTools, except for the 12-6-4 model where the geometry minimization was carried out with OpenMM[40]. PM6, PM3, AM1 and

SCC-DFTB[29,30] calculations were performed using sqm in AmberTools. The benchmark QM calculations were carried out at the B3LYP/6-31G* level of theory. All the B3LYP/6-31G* calculations were performed using Gaussian 09[46].

3.3.2.1. Relative energies

In order to obtain a structure set to evaluate relative energies, we carried out gas-phase MD simulations employing the SCC-DFTB method to collect snapshots for the 12 complexes. First, the system was optimized with SCC-DFTB. Then a 10 ps simulation was carried out to heat the system from 0 K to 300 K. This was followed by a 5 ps simulation to further equilibrate the system. Finally the production MD run covered 15 ps with snapshots being stored every 750 steps (0.75 ps), providing 20 structures in total. A 1 fs time-step and the Langevin thermostat were used for these MD simulations.

Then all the methods were validated, including the QM reference calculations, by calculating single-point energies on each snapshot structure. We then calculated the QM based relative energies between each structure, providing 190 relative reference energy values for each system, for comparison. The mean error (ME) and mean unsigned error (MUE) of these values relative to the DFT calculated results were then calculated. The same approach was performed for all 12 systems yielding the overall ME and MUE for each method.

3.3.2.2. Structural prediction

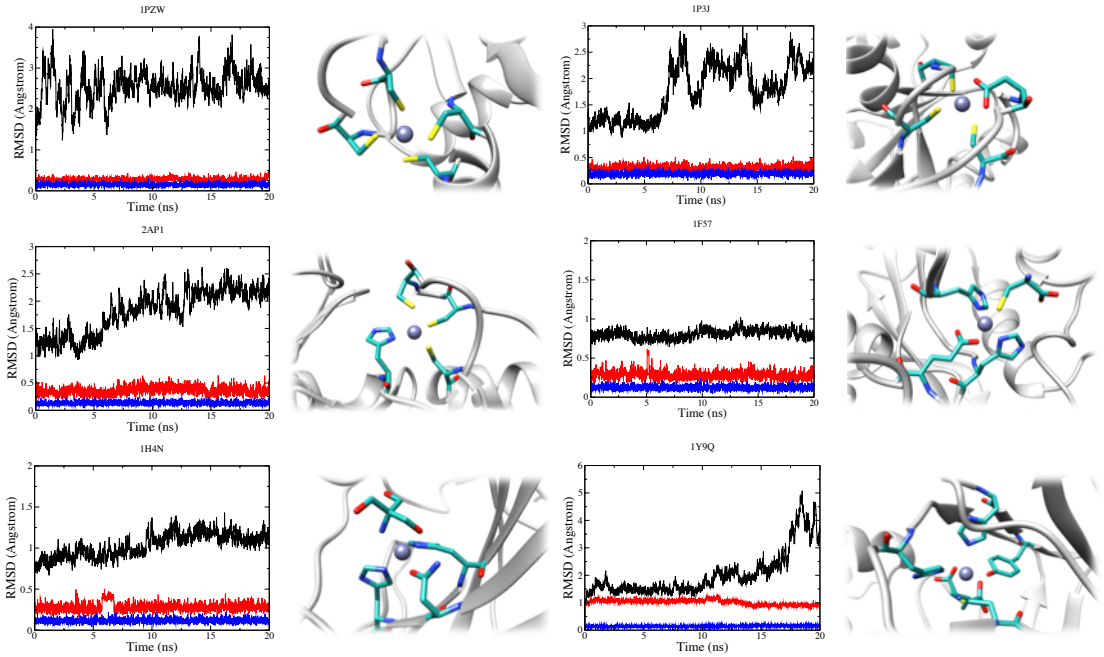
Besides assessing the ability of each method to accurately compute relative energies, we also evaluated their ability to predict structural data by assessing their minimized structures relative to

DFT optimized and the crystal structures. The criteria used herein were the RMSD values of the zinc ion together with its coordinating atoms and the RMSD values of all the heavy atoms in the system. Moreover, we also performed comparison of the Zn-ligand bond lengths in the optimized structures to the bond lengths given by both DFT and experiment.

3.3.2.3. Gas phase MD simulations for CSD complexes with nonbonded parameter sets

Four 1ns simulations using HFE, IOD, CM and 12-6-4 were carried out for each of the six CSD complexes systems. First, the system was minimized and then three iterations of 1 ns of heating and 1 ns of equilibration were run to bring the system from 0 K to 300 K. Finally the production MD run covered 1 ns. A 1 fs time-step and the Langevin thermostat were used for these MD simulations.

Figure 3.3. RMSD values of the protein backbone, metal binding site and binding atoms monitored along the 20 ns MD trajectory for each metalloprotein investigated (left), and the last snapshot from each trajectory (right). In the plots to the left, the black, red and blue curves represent backbone, metal binding site and binding atoms respectively.



3.4. Results and Discussion

3.4.1 Validation of EZAFF

3.4.1.1. Zinc Metalloproteins

EZAFF was tested on 6 zinc metalloproteins (see Table 3.2) in explicit water using MD simulations. RMSD values for the protein backbone, metal binding site (the zinc ion plus the heavy atoms in the ligating residues) and the directly bound atoms (the central zinc ion plus the ligating atoms) against relevant PDB structures were calculated for each MD trajectory. These results are plotted in Figure 3 where we find that the geometries of all the six metal binding centers were stable throughout the 20 ns of MD simulation with average RMSD values for the binding atoms of ~0.25 Å. The average RMSD values of the metal binding site were slightly larger (~0.5 Å) in comparison to those of the binding atoms. Based on these results, we concluded that EZAFF works well (at least for these systems) along with the AMBER ff14SB force field and GAFF to model zinc containing metalloproteins.

3.4.1.2. Zinc organometallics

We also tested EZAFF's ability to model zinc containing organometallic compounds. For each of the 6 CSD complexes shown in Table 3.2, we calculated all possible permutations of the RMSD values between the EZAFF minimized structure, the B3LYP/6-31G* optimized structure and the CSD structure. The results are summarized in Table 3.3. Superpositions of the three structures for each of the 6 systems are illustrated in Figure 3.4. For the coordinates of the zinc and its ligating atoms, the RMSD values with respect to the CSD structures are < 0.15 Å for all the 6 complexes and the RMSD values towards the QM optimized structures are < 0.56 Å. It was not unexpected that EZAFF would work better in reproducing CSD geometries over QM because it assigns

equilibrium bond lengths and angles involving the metal ion based on CSD structures. The RMSD values become larger when including the coordinates of all non-hydrogen atoms in the calculation. This may, in part, be due to crystal packing effects in structures obtained from the CSD. Nonetheless, these RMSD values are quite reasonable, showing the capabilities of EZAFF coupled with GAFF. Based on these results, we conclude that EZAFF performed well overall in reproducing CSD structures for all 6 systems.

Figure 3.4. Superimposition of the EZAFF optimized, DFT optimized, and CSD structures for 5 compounds taken from the CSD. These structures are shown in gray, green and yellow, respectively.

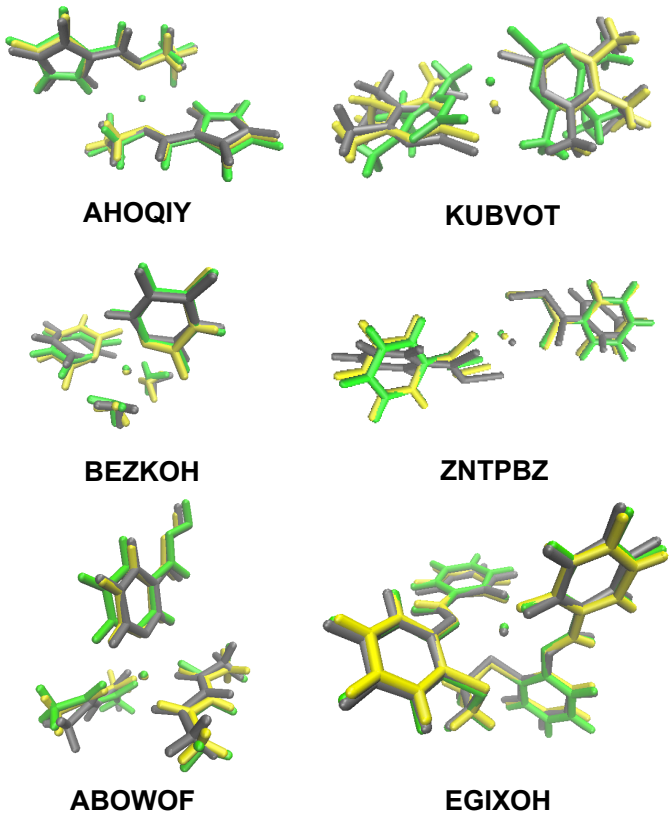


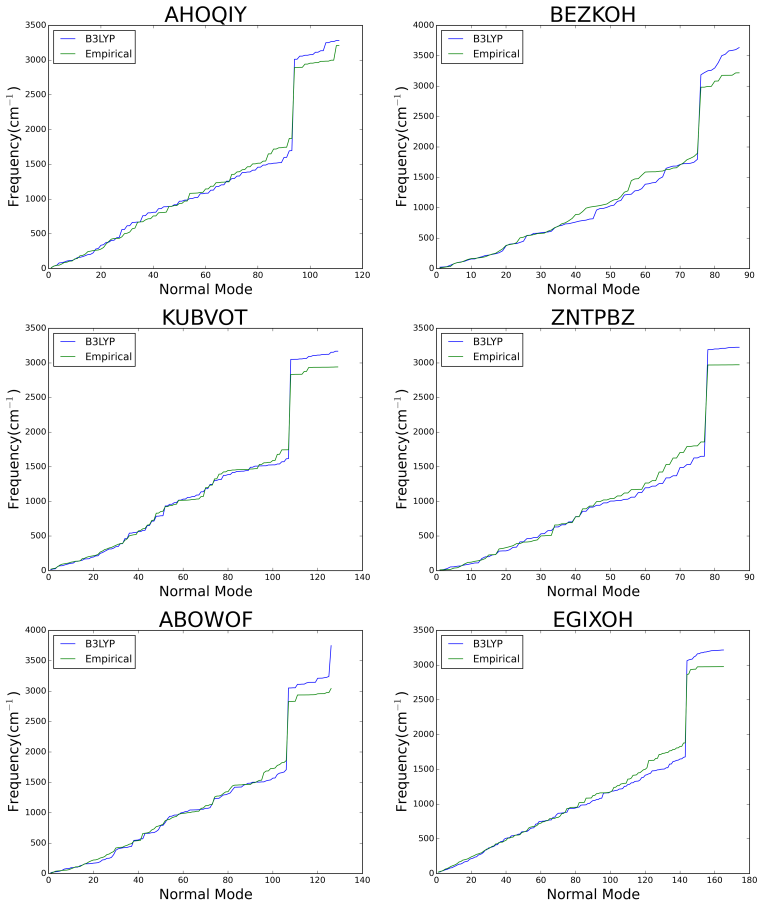
Table 3.3. RMSD values between EZAFF minimized, DFT optimized and CSD structures of six zinc containing organometallic compounds

Model systems	RMSD (Å)[a]			RMSD (Å)[b]		
	MM vs QM	MM vs CSD	CSD vs QM	MM vs QM	MM vs CSD	CSD vs QM
AHOQIY	0.110	0.088	0.094	0.306	0.198	0.195
BEZKOH	0.145	0.150	0.113	0.458	0.838	0.453
KUBVOT	0.557	0.131	0.534	1.222	0.440	1.212
ZNTPBZ	0.379	0.102	0.356	1.394	1.155	0.482
ABOWOF	0.228	0.077	0.216	0.545	0.344	0.438
EGIXOH	0.175	0.119	0.130	0.252	0.247	0.275

[a] RMSD values were computed for the coordinates of zinc and its directly ligated atoms.
[b] RMSD values were computed for the coordinates of all non-hydrogen atoms.

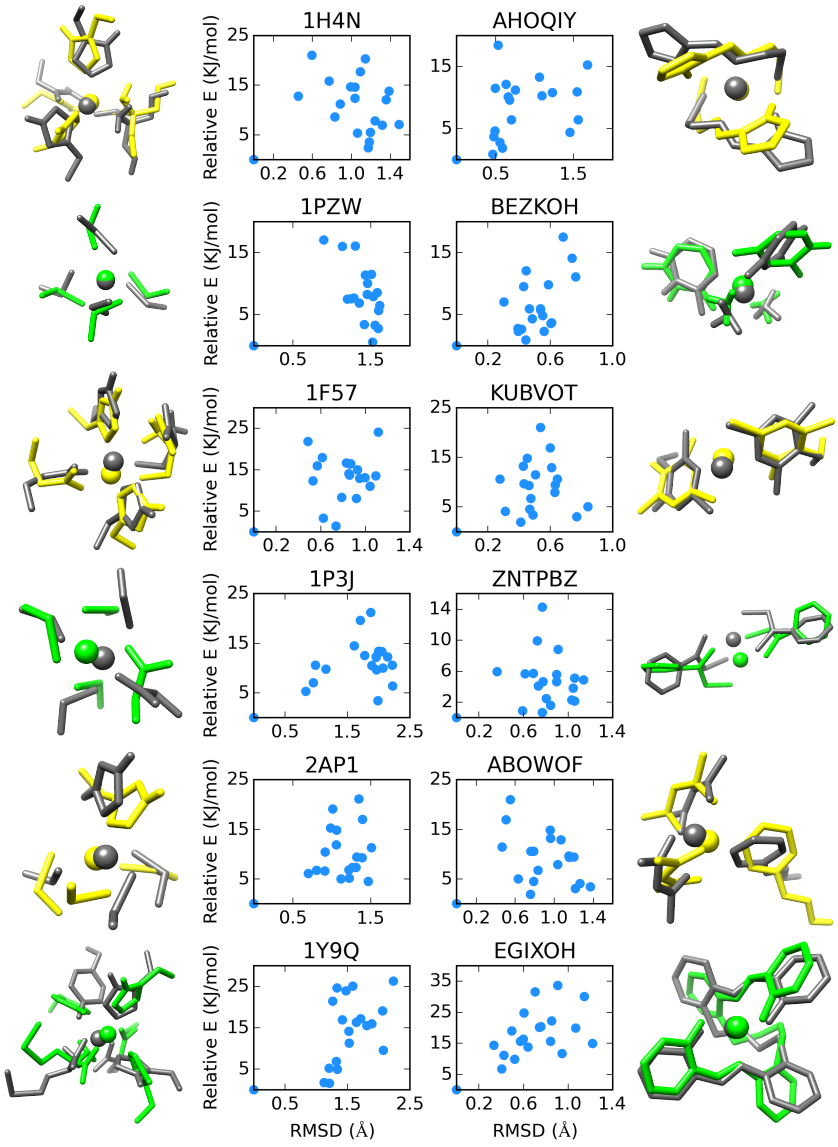
Besides structural predictions, accurately predicting vibrational frequencies is another way to evaluate the quality of a force field. The vibrational frequencies of the 6 complexes were calculated based on the EZAFF approach and at B3LYP/6-31G* level of theory respectively and their comparisons are shown in Figure 3.5.

Figure 3.5. Vibrational frequencies calculated based on DFT (blue) and the EZAFF model (green). In each plot, the normal modes are arranged in the order of their vibrational frequencies calculated by DFT as shown in the x-axis.



It can be seen that, excepting discrepancies in some of the high frequency modes which are located in the > 3000 cm^{-1} range, most of the vibrational frequencies generated by the EZAFF model match reasonably well with those generated by DFT calculations. Similar patterns were found by Lin and Wang.[47] They noted that the discrepancies in the high frequency range were mainly due to the stretching force constants of the C-H, N-H, and S–H bonds in GAFF[26] whose influence is negligible when applying SHAKE[44] to constrain the X-H bonds during simulation.

Figure 3.6. B3LYP relative energies vs. heavy atom RMSDs over all 20 conformers for all the 12 test systems. The RMSDs were calculated with respect to the lowest energy conformer. The corresponding superposition is for the lowest single point energy conformer (gray) and the highest RMSD one (green or yellow).



3.4.2. Benchmark assessment of different MM and semi-empirical QM methods for modeling zinc-containing systems

As described in the Method section, 3 bonded MM methods (EZAFF, Seminario and Z-matrix methods), 4 nonbonded parameter sets (HFE, IOD, CM and 12-6-4) and 4 popular semi-empirical

molecular orbital methods (AM1, PM3, PM6 and SCC-DFTB) were evaluated against energetic and structural B3LYP/6-31G* derived benchmark quantities. We also provided a detailed comparison amongst the 4 nonbonded models based on how they performed on gas phase MD simulations of the 6 CSD test systems.

Table 3.4. Mean unsigned errors (MUEs) of relative energies for 12 zinc complexes investigated

method	MUE[a] (kcal/mol)												
	1PZW	2AP1	1H4N	1P3J	1F57	1Y9Q[b]	AHOQIY	BEZKOH	ZNTPBZ	KUBVOT	ABOWOF	EGIXOH	All
PM6	3.88	4.86	4.66	3.75	8.11	4.68	5.14	4.13	2.58	4.25	8.71	6.21	5.08
PM3	3.75	3.81	5.51	4.94	8.06	8.84	3.78	6.54	2.13	4.33	10.52	6.64	5.74
AM1	5.26	3.34	6.60	4.08	9.19	7.21	3.65	6.74	4.14	4.72	12.36	5.46	6.06
SCC-DFTB	2.89	3.65	4.24	3.33	4.49	7.37	3.50	2.28	2.55	2.79	9.61	5.65	4.36
EZAFF	10.30	7.34	18.71	15.44	25.03	16.34	9.10	6.89	9.43	17.07	12.58	22.93	14.26
Seminario	5.38	7.37	10.84	7.12	39.25	--	7.17	8.30	5.31	10.00	17.09	10.15	11.64[c]
Z-matrix	4.46	5.37	15.68	6.15	21.75	--	5.86	8.16	4.63	4.48	11.40	7.22	8.65[c]
HFE	6.19	6.38	13.51	7.64	19.58	16.45	10.16	14.05	5.81	7.20	13.61	8.90	10.79
IOD	4.80	6.14	12.60	6.25	19.96	14.58	9.02	13.25	4.71	6.63	14.33	5.96	9.85
CM	5.87	6.23	13.10	7.28	19.55	15.89	9.73	13.45	5.55	6.88	13.57	8.08	10.43
12-6-4	5.89	7.42	13.23	6.62	21.37	14.56	9.69	14.78	6.30	8.02	14.57	6.27	10.73

a MUEs were computed against B3LYP/6-31G* values.
b One Zn-N bond broke after DFT optimization, so that Seminario and Z-matrix are not applicable.
c MUEs were computed over all systems except 1Y9Q.

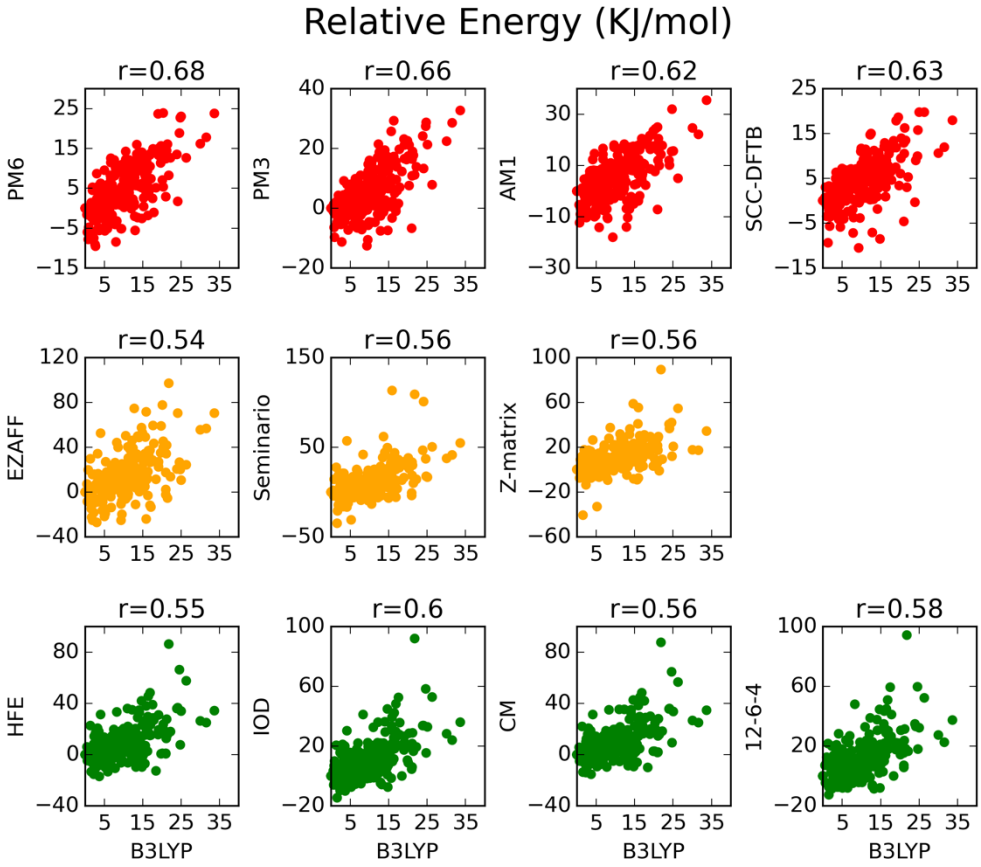
3.4.2.1. Energetic predictions

As described in the Methods section, 20 snapshots were taken for each system. DFT single point energy calculations were carried out for each of them. The scatterplots in Figure 3.6 show the B3LYP relative energies along with the corresponding structural RMSDs (with respect to the lowest energy point). Based on the superpositions in Figure 3.6 of the lowest energy conformer and the largest RMSD conformer, we can see that the metal ion coordination environment is maintained throughout the simulations. First, we compared the relative energies predicted by the 7 MM methods and 4 semi-empirical molecular orbital methods to the DFT benchmark values.

The performance of each model was evaluated by the mean unsigned errors (MUEs) and mean errors (MEs) against DFT derived results. Table 3.4 lists the MUEs for the relative energies for each system. The corresponding MEs are shown in Table 3.8. Generally, the semi-empirical molecular orbital methods outperform the MM models. SCC-DFTB gives the best predictions for the relative energies for 8 systems except 2AP1, 1Y9Q, ZNTPBZ and EGIXOH. The MUEs from Table 3.4 show the order SCC-DFTB, PM6, PM3, AM1, and then the Z-matrix and the IOD nonbonded parameter set (from best to worst). Finally, the remaining nonbonded parameter sets, Seminario's and ZAFF perform the worst. However, the difference in the performance of methods in the same class is small. Among semi-empirical molecular orbital methods, PM6 and PM3 give similar results that are subtly better than AM1 with difference of ~1 kcal/mol. Within the 4 nonbonded MM models, the IOD parameter set is the best option since it provides the lowest MUEs in 9 out of 12 of the test systems. This is followed by the CM and then the 12-6-4 models and which one of the two is best is case dependent. Finally, the HFE parameter set is a bit off from the other nonbonded models, which maybe isn't surprising since this model makes the largest structural compromise to get the energies right. The discrepancies are comparatively bigger for the bonded MM models: Z-matrix method yielded an overall MUE 3 kcal/mol less than the Seminario method, and Seminario outperforms EZAFF by less than 3 kcal/mol. Moreover, the performances between the three categories of methods are not significantly different. For example, the difference between SCC-DFTB, which is the best semi-empirical molecular orbital method, and the best bonded MM methods was ~4 kcal/mol, while Z-matrix slightly outperformed the IOD nonbonded model by just ~1 kcal/mol. Meanwhile, the MEs (see Table 3.8) show different methods offer similar systematic errors, with nonbonded models offering consistently slightly higher MEs. We also plotted the correlation between relative energies computed with B3LYP and other methods in

Figure 3.7. The data points in the plots represent relative energies of conformers with respect to the one having the lowest B3LYP single point energy for each test system. The computed Pearson's rs vary from 0.54 to 0.68. It is shown that semi-empirical methods are slightly more correlated to B3LYP values and render a similar range of relative energies. However, both bonded and nonbonded MM models give several relative energies out of range. Based on the above, we conclude that both bonded and nonbonded MM models offer slightly less accurate relative energies with respect to semiempirical quantum methods but with reduced computation cost. Moreover, with much less effort needed in parameterization, EZAFF, represents a good substitute for the Seminario and Z-matrix methods for relative energy calculations.

Figure 3.7. Correlation of B3LYP relative energies vs. relative energies with different methods over all 20 conformers for all the 12 test systems. Relative energies were computed relative to the values of the conformers with the lowest B3LYP single point energy. The Pearson's correlation coefficients are given for each plot.



3.4.2.2. Structural Predictions

We evaluated the structural models for each method by calculating the RMSDs between their minimized structures and the DFT optimized ones for the 11 applicable complexes (except 1Y9Q because one ZN-N bond was severed in the DFT optimization). All the RMSD values listed in Table 3.5 are based on the coordinates of the zinc ion and its ligating atoms. However, we find overall that it is hard to clearly distinguish between semi-empirical QM methods and bonded MM methods and their performance is case dependent. Overall, per Table 3.5, Seminario and PM6 are

slightly better overall since both of them have a higher probability of outperforming the other method and yield the lowest average RMSDs as shown in the last column of Table 3.5. When reproducing the DFT optimized structures, the Seminario method did the best job in 4 out of 12 complexes, while PM6, SCC-DFTB, PM3 and AM1 yielded the lowest RMSDs for 3, 2, 1 and 1 out of 12, respectively, which indicates a comparable ability of the bonded MM methods relative to the semi-empirical methods. Within the scope of the bonded MM methods, the Seminario method yielded a consistent performance overall and outperformed the Z-matrix and EZAFF method. However, both the Seminario and Z-matrix approaches have the same limitation in terms of the expense to build the model. Although the performance of nonbonded models are not up to semi-empirical QM methods and bonded models for most systems, they perform consistently well in reproducing DFT optimized structures. Amongst the 4 nonbonded models, IOD and 12-6-4 are the two best options, followed by CM and HFE.

Next, we compared the different methods in their ability to reproduce the crystal structure and their corresponding RMSDs are listed in Table 3.6. It is not surprising to find that EZAFF performs the best and gives the smallest RMSD values for 9 systems, since EZAFF assigns the equilibrium bond lengths and angle values based on the crystal structures. However, it is hard to rank order the semi-empirical QM methods, bonded models and nonbonded models since they all provide similar RMSD values for all test systems.

Table 3.5. RMSD values of the optimized structure by each method towards the DFT optimized geometry for 12 complexes

method	RMSD (Å)[a]												
	1PZW	2AP1	1H4N	1P3J	1F57	1Y9Q[b]	AHOQIY	BEZKOH	ZNTPBZ	KUBVOT	ABOWOF	EGIXOH	mean[d]
PM6	0.062	0.114	1.966	0.210	0.338[c]	--	0.081	0.050	0.044	0.045	0.083	0.242	0.294
PM3	0.067	0.107	2.124[c]	0.177	0.572	--	0.164	0.033	0.053	0.057	0.092	0.187	0.330
AM1	0.217	0.162	2.024	0.252	0.251	--	0.220	0.138	0.158	0.143	0.118	0.137	0.347
SCC-DFTB	0.071	0.116	1.960	0.059	0.835	--	0.076	0.049	0.071	0.047	0.092	0.233	0.328
EZAFF	0.260	0.241	1.987	0.285	0.205	--	0.110	0.145	0.379	0.557	0.228	0.175	0.416
Seminario	0.187	0.084	0.116	0.157	0.122	--	0.088	0.146	0.083	1.104	0.077	0.149	0.210
Z-matrix	0.197	0.904	1.070	1.288	0.400	--	0.086	0.346	0.244	0.374	0.915	0.456	0.571
HFE	0.514	0.453	0.912	0.477	0.396	--	0.429	0.320	0.437	0.455	0.958[c]	0.232	0.508
IOD	0.249	0.211	0.656	0.521	0.269	--	0.263	0.376	0.158	0.282	0.460	0.198	0.331
CM	0.411	0.356	0.789	0.375	0.308	--	0.344	0.436	0.334	0.304	0.469	0.205	0.394
12-6-4	0.237	0.189	0.666	0.473	0.255	--	0.299	0.348	0.218	0.423	0.373	0.246	0.339

a RMSD values were computed by considering the coordinates of zinc and the atoms in direct bonding with zinc.
b One Zn-N bond broke after DFT optimization.
c Coordination changed after minimization with corresponding method.
d The mean RMSD over all the systems except 1Y9Q.

Table 3.6. RMSD values of the optimized structure by each method towards the crystal structure (PDB or CSD) for each of the 12 complexes

method	RMSD (Å)[a]											
	1PZW	2AP1	1H4N	1P3J	1F57	1Y9Q[b]	AHOQIY	BEZKOH	ZNTPBZ	KUBVOT	ABOWOF	EGIXOH
PM6	0.172	0.272	0.822	0.337	0.429[c]	0.331	0.080	0.136	0.377	0.544	0.287	0.131
PM3	0.174	0.227	1.030[c]	0.255	0.422	2.193[c]	0.222	0.134	0.392	0.535	0.231	0.127
AM1	0.223	0.304	0.863	0.290	0.156	0.303	0.254	0.238	0.351	0.531	0.263	0.164
SCC-DFTB	0.190	0.277	0.896	0.183	0.781	0.283	0.137	0.140	0.390	0.531	0.176	0.281
EZAFF	0.105	0.083	0.117	0.239	0.078	0.124	0.088	0.150	0.102	0.131	0.077	0.119
Seminario	0.152	0.187	1.995	0.212	0.243	--	0.151	0.216	0.359	0.752	0.205	0.150
Z-matrix	0.131	0.844	2.045	0.190	0.541	--	0.092	0.390	0.375	0.782	0.826	0.487
HFE	0.407	0.445	0.875	0.470	0.406	2.010[c]	0.429	0.282	0.556	0.702	1.055[c]	0.232
IOD	0.190	0.284	1.959	0.464	0.188	0.130	0.285	0.343	0.411	0.773	0.270	0.140
CM	0.313	0.368	1.936	0.376	0.287	1.390[c]	0.354	0.386	0.490	0.645	0.295	0.149
12-6-4	0.190	0.272	1.974	0.427	0.205	0.390	0.337	0.319	0.329	0.253	0.190	0.144

a RMSD values were computed by considering the coordinates of zinc and the atoms in direct bonding with zinc.
b One Zn-N bond broke after DFT optimization, so that Seminario and Z-matrix are not applicable.
c Coordination changed after minimization with corresponding method.

The RMSD values listed in Table 3.9 and Table 3.10 are based on the coordinates of all heavy atoms. One will notice that the lowest all heavy atom RMSDs generated among the 7 molecular mechanics methods are similar to the lowest RMSDs from the 4 semi-empirical QM methods.

Therefore, we draw the conclusion that, when coupled with GAFF, the performance of both bonded and nonbonded models are comparable to semi-empirical QM methods.

The MUEs, MEs of Zn-X equilibrium bond lengths prodcued by semi-empirical molecular orbital methods, bonded MM models and nonbonded MM models are collected in Table 3.7. Again, as expected, EZAFF shows the best performance for crystal bond length prediction, followed by SCC-DFTB, PM6, Seminario, 12-6-4, AM1, IOD, Z-matrix, PM3, CM and HFE. For prediction of the DFT optimized Zn-X bond lengths, the performance ranking is SCC-DFTB, PM6, AM1, Seminario, Z-matirx, PM3, EZAFF, 12-6-4, IOD, CM and HFE.

Table 3.7. MUEs and MEs of Zn-X distance values of the optimized structure by each method towards the crystal structure (PDB or CSD) and DFT optimized structure for 11 complexes (except 1Y9Q)

method	crystal structure		DFT optimized structure	
	MUE (Å)	ME (Å)	MUE (Å)	ME (Å)
PM6	0.113	-0.015	0.085	-0.033
PM3	0.149	0.083	0.129	0.066
AM1	0.126	0.100	0.099	0.079
SCC-DFTB	0.104	-0.006	0.071	-0.023
EZAFF	0.092	0.036	0.133	0.019
Seminario	0.117	0.043	0.109	0.026
Z-matrix	0.139	0.081	0.120	0.064
HFE	0.452	-0.320	0.443	-0.337
IOD	0.127	-0.088	0.137	-0.105
CM	0.264	-0.255	0.278	-0.272
12-6-4	0.120	-0.078	0.136	-0.095

3.4.2.3. Gas phase MD simulations for CSD complexes with nonbonded parameter sets

Each one of the 6 CSD complexes were modeled with different nonbonded parameter sets and simulated in the gas-phase for 1 ns. Zn-X distance values were recorded along all the simulations

and plotted in Figure 3.8. It is shown that the complex structures modeled with either IOD or 12-6-4 parameter sets remain intact during the simulations, whereas, a few Zn-N and Zn-S interactions were broken when being modeled with HFE or CM, as seen in ABOWOF, BEZKOH and EGIXOH.

Figure 3.8. Zn-X distance values from 1 ns gas-phase MD simulations of CSD complexes. The title for each plot includes the complex name and the nonbonded parameter set. Different colors represents different Zn-X interactions (ABOWOF: black, red, blue, green, yellow correspond to Zn-N and 4×Zn-O; AHOQIY: black, red, blue, green correspond to Zn-S, Zn-N, Zn-S and Zn-N; BEZKOH: black, red, blue, green all represent Zn-N; EGIXOH: black, red, blue, green, yellow, brown correspond to 2×Zn-O, 2×Zn-N and 2×Zn-S; KUBVOT: black, red, blue, green correspond to Zn-S, Zn-N, Zn-S and Zn-N; ZNTPBZ: black, red, blue, green all represent Zn-S)

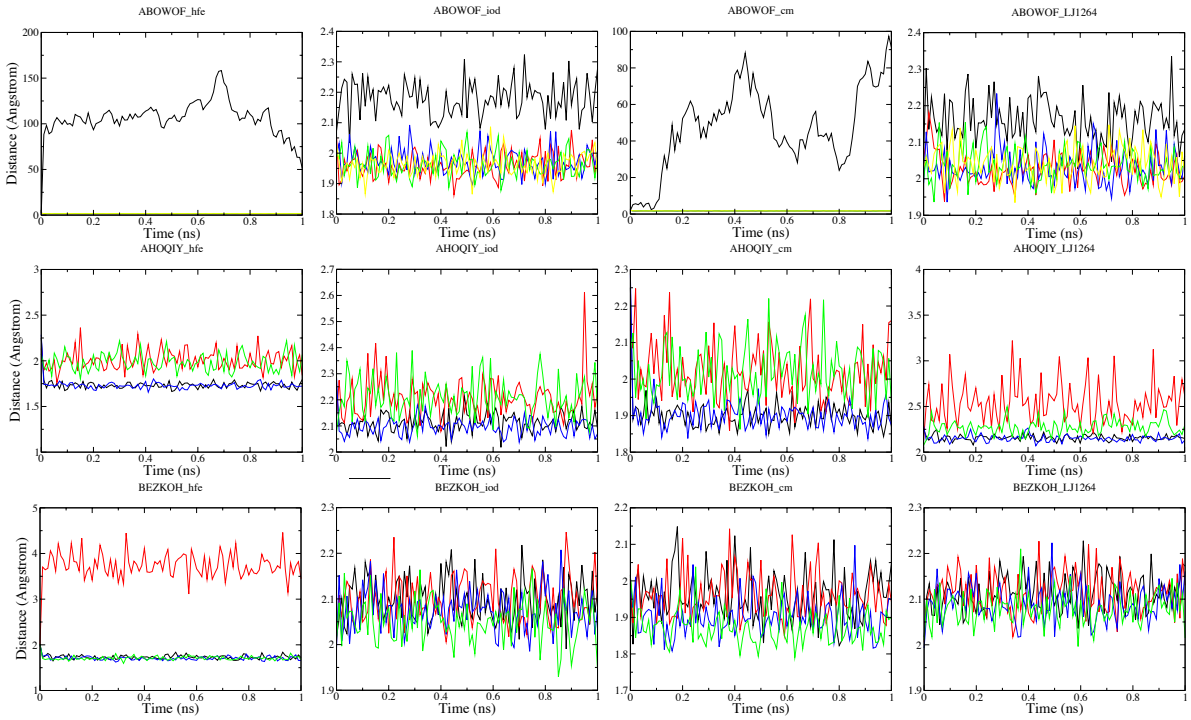
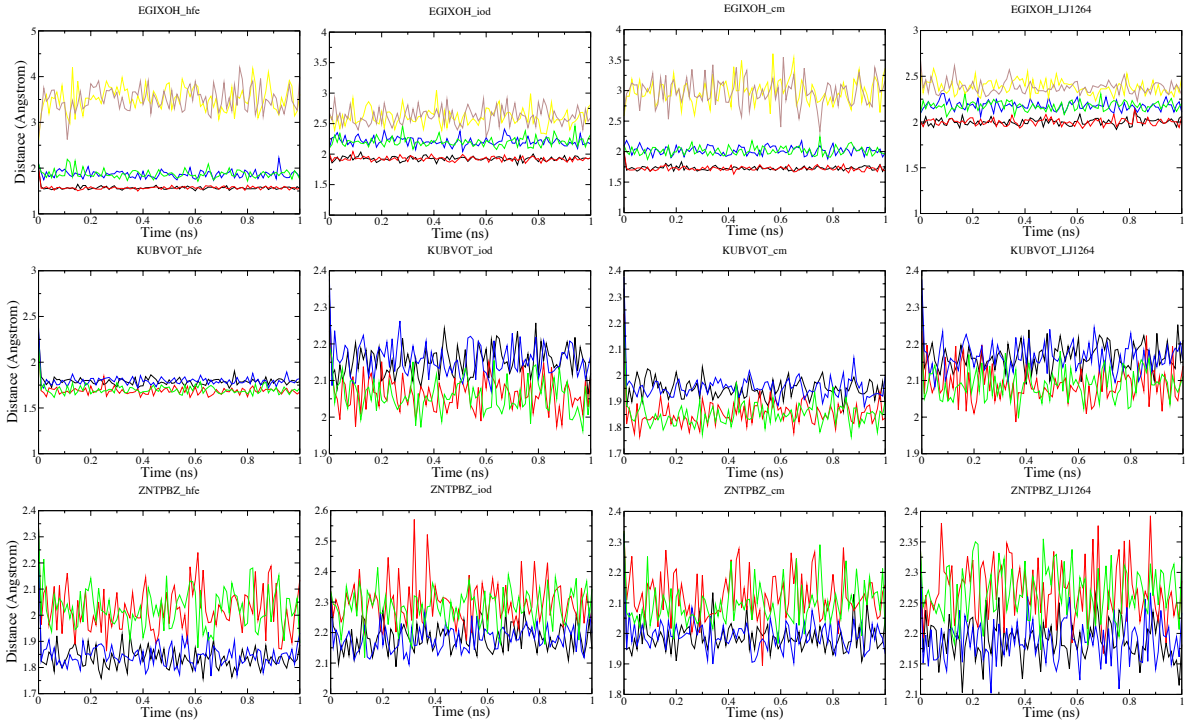


Figure 3.8. (cont'd)



3.5. Conclusions

In present work we developed the EZAFF method that is based on an empirical approach to determine the force field parameters for bonds and angles involving the zinc ion. Validations were performed on 6 metalloproteins and 6 organometallic compounds with various coordination environments. Results showed that EZAFF is reliable for simulating these systems. Meanwhile, we performed benchmark calculations on 3 bonded MM methods (the EZAFF, Seminario and Z-matrix methods), 4 nonbonded MM parameter sets (the HFE, IOD, CM and 12-6-4) and 4 semi-empirical molecular orbital methods (AM1, PM3, PM6 and SCC-DFTB methods). These benchmark calculations explored the performance of these methods to both reproduce structural data and relative energies. At a reduced computational cost, the MM models yield a comparable

performance to the more expensive semi-empirical models for modeling zinc-containing complexes.

3.6. Supporting Information

Table 3.8. Mean errors (MEs) of relative energies for 12 zinc complexes investigated

method	ME[a] (kcal/mol)												
	1PZW	2AP1	1H4N	1P3J	1F57	1Y9Q[b]	AHOQIY	BEZKOH	ZNTPBZ	KUBVOT	ABOWOF	EGIXOH	All
PM6	0.15	0.43	-0.84	-1.85	-0.69	-0.62	-0.07	-1.14	0.14	-0.48	-0.44	-0.13	-0.46
PM3	0.52	0.75	-1.07	1.61	-0.21	3.74	-0.67	-1.77	0.54	-0.99	-0.40	-0.85	0.10
AM1	0.13	1.48	0.46	0.98	-0.72	1.07	-0.16	-0.53	0.63	-1.04	-0.31	1.40	0.28
SCC-DFTB	0.42	0.03	0.41	1.29	0.15	-2.87	-0.17	-0.40	0.90	-0.53	-0.23	0.31	0.03
EZAFF	0.98	3.44	-2.05	7.17	-3.87	0.46	0.04	-3.05	3.53	3.34	-4.41	-3.89	0.14
Seminario	-0.21	2.88	-2.10	1.24	-3.01	--	-0.18	-3.38	2.06	4.61	-7.13	-0.43	-0.51[c]
Z-matrix	0.02	1.52	-1.95	-0.40	-5.49	--	-0.64	-2.22	1.16	1.58	-2.98	2.26	-0.65[c]
HFE	0.92	0.81	0.27	-4.18	-9.06	1.55	1.42	-3.24	-0.87	1.77	-4.23	0.08	-1.23
IOD	0.62	1.07	-0.36	-2.75	-8.78	1.81	0.24	-4.02	-0.14	0.61	-5.26	2.00	-1.25
CM	0.85	0.86	1.38	-3.87	-9.00	1.61	1.16	-3.41	-0.71	1.52	-4.45	0.48	-1.24
12-6-4	1.00	0.96	-0.21	-2.15	-8.63	3.19	0.12	-4.60	-0.49	0.63	-5.38	2.44	-1.09

a MEs were computed against B3LYP/6-31G* values.
b One Zn-N bond broke after DFT optimization, so that Seminario's and Z-matrix are not applicable.
c MEs were computed over all systems except 1Y9Q.

Table 3.9. RMSD values of the optimized structure by each method towards the DFT optimized geometry for 12 complexes

method	RMSD (Å)[a]											
	1PZW	2AP1	1H4N	1P3J	1F57	1Y9Q[b]	AHOQIY	BEZKOH	ZNTPBZ	KUBVOT	ABOWOF	EGIXOH
PM6	0.431	0.764	3.024	1.452	1.371[c]	--	0.772	0.743	0.271	0.083	0.090	0.563
PM3	0.743	0.548	3.880[c]	1.203	1.479	--	0.354	0.534	0.290	0.210	0.735	0.639
AM1	1.044	0.636	3.054	1.219	0.541	--	0.383	0.995	0.193	0.195	0.140	0.215
SCC-DFTB	0.659	0.652	2.936	0.453	2.170	--	0.724	0.697	0.108	0.087	0.183	0.295
EZAFF	1.074	0.571	3.446	0.721	0.686	--	0.306	0.838	1.394	1.222	0.545	0.252
Seminario	0.738	0.551	0.737	0.636	1.562	--	0.411	0.657	0.787	1.610	0.496	0.188
Z-matrix	0.813	1.089	1.297	1.314	1.304	--	0.286	0.932	1.150	0.575	1.124	0.984
HFE	1.557	1.198	1.705	1.551	0.605	--	1.525	0.592	0.709	0.439	1.209[c]	0.398
IOD	1.571	1.236	1.456	1.470	0.762	--	1.638	0.487	0.638	0.977	0.876	0.333
CM	1.444	1.280	1.600	1.874	0.698	--	1.589	0.610	0.653	0.402	0.844	0.366
12-6-4	0.751	0.671	1.326	1.000	0.631	--	1.728	0.445	1.745	1.067	0.622	0.317

a RMSD values were computed by considering the coordinates of all non-hydrogen atoms.
b One Zn-N bond broke after DFT optimization.
c Coordination changed after minimization with corresponding method.

Table 3.10. RMSD values of the optimized structure by each method towards the crystal structure (PDB or CSD) for each of the 12 complexes

method	RMSD (Å)[a]											
	1PZW	2AP1	1H4N	1P3J	1F57	1Y9Q[b]	AHOQIY	BEZKOH	ZNTPBZ	KUBVOT	ABOWOF	EGIXOH
PM6	0.547	0.706	1.579	1.769	1.423[c]	0.938	0.702	0.327	0.593	1.261	0.454	0.483
PM3	0.607	0.665	1.196[c]	1.523	1.066	2.135[c]	0.445	0.204	0.539	1.069	0.557	0.652
AM1	0.848	0.670	1.467	1.528	0.857	0.968	0.478	0.585	0.556	1.112	0.403	0.362
SCC-DFTB	0.886	0.689	1.479	0.590	1.932	1.711	0.694	0.294	0.507	1.237	0.546	0.362
EZAFF	1.139	0.689	0.848	0.483	0.864	1.323	0.198	0.458	1.155	0.440	0.344	0.247
Seminario	0.981	0.796	3.483	0.585	1.963	--	0.376	0.251	0.672	0.963	0.379	0.311
Z-matrix	1.034	1.026	3.611	1.199	1.718	--	0.187	0.531	1.018	0.995	0.912	1.162
HFE	1.540	1.448	3.683	1.653	0.947	2.070[c]	1.451	0.434	1.072	1.494	1.269[c]	0.296
IOD	1.565	1.512	3.763	1.295	0.955	1.906	1.570	0.547	1.046	2.122	0.484	0.229
CM	1.397	1.531	3.688	1.750	1.043	1.641[c]	1.518	0.743	1.031	1.530	0.460	0.232
12-6-4	0.492	0.617	3.699	0.616	0.663	1.077	1.664	0.559	1.614	0.252	0.241	0.173

a RMSD values were computed by considering the coordinates of all non-hydrogen atoms.
b One Zn-N bond broke after DFT optimization, so that Seminario and Z-matrix are not applicable.
c Coordination changed after minimization with corresponding method.

REFERENCES

REFERENCES

1. Coleman, J. E. (1992) Zinc Proteins - Enzymes, Storage Proteins, Transcription Factors, and Replication Proteins. *Annu. Rev. Biochem.* 897-946.

2. Anzellotti, A. I.; Farrell, N. P. (2008) Zinc metalloproteins as medicinal targets. *Chem. Soc. Rev. 8,* 1629-1651.

3. Stote, R. H.; Karplus, M. (1995) Zinc-Binding in Proteins and Solution - a Simple but Accurate Nonbonded Representation. *Proteins. 1,* 12-31.

4. Babu, C. S.; Lim, C. (2006) Empirical force fields for biologically active divalent metal cations in water. *J. Phys. Chem. A. 2,* 691-699.

5. Wu, R. B.; Lu, Z. Y.; Cao, Z. X.; Zhang, Y. K. (2011) A Transferable Nonbonded Pairwise Force Field to Model Zinc Interactions in Metalloproteins. *J. Chem. Theory Comput. 2,* 433-443.

6. Li, P. F.; Roberts, B. P.; Chakravorty, D. K.; Merz, K. M. (2013) Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for +2 Metal Cations in Explicit Solvent. *J. Chem. Theory Comput. 6,* 2733-2748.

7. Li, P. F.; Merz, K. M. (2014) Taking into Account the Ion-Induced Dipole Interaction in the Nonbonded Model of Ions. *J. Chem. Theory Comput. 1,* 289-297.

8. Pang, Y. P. (1999) Novel zinc protein molecular dynamics simulations: Steps toward antiangiogenesis for cancer treatment. *J. Mol. Model.* 1999, *10,* 196-202.

9. Pang, Y. P.; Xu, K.; El Yazal, J.; Prendergast, F. G. (2000) Successful molecular dynamics simulation of the zinc-bound farnesyltransferase using the cationic dummy atom approach, *Protein Sci. 12,* 2583-2583.

10. Vedani, A.; Huhta, D. W. (1990) A New Force-Field for Modeling Metalloproteins. *J. Am. Chem. Soc. 12,* 4759-4767.

11. Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. (1993) A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges - the Resp Model. *J. Phys. Chem. 40,* 10269-10280.

12. Li, J. B.; Zhu, T. H.; Cramer, C. J.; Truhlar, D. G. (1998) New class IV charge model for extracting accurate partial charges from wave functions. *J. Phys. Chem. A 10,* 1820-1831.

13. Seminario, J. M. (1996) Calculation of intramolecular force fields from second-derivative tensors. *Int. J. Quantum Chem. 7,* 1271-1277.

14. Gresh, N. (1995) Energetics of Zn2+ Binding to a Series of Biologically Relevant Ligands - a Molecular Mechanics Investigation Grounded on Ab-Initio Self-Consistent-Field Supermolecular Computations. *J. Comput. Chem. 7,* 856-882.

15. Tiraboschi, G.; Gresh, N.; Giessner-Prettre, C.; Pedersen, L. G.; Deerfield, D. W. (2000) Parallel ab initio and molecular mechanics investigation of polycoordinated Zn(II) complexes with model hard and soft ligands: Variations of binding energy and of its components with number and charges of ligands. *J. Comput. Chem. 12,* 1011-1039.

16. Tiraboschi, G.; Roques, B. P.; Gresh, N. (1999) Joint quantum chemical and polarizable molecular mechanics investigation of formate complexes with penta- and hexahydrated Zn(2+): Comparison between energetics of model bidentate, monodentate, and through-water Zn(2+) binding modes and evaluation of nonadditivity effects. *J. Comput. Chem. 13,* 1379-1390.

17. Gresh, N.; Derreumaux, P. (2003) Generating conformations for two zinc-binding sites of HIV-1 nucleocapsid protein from random conformations by a hierarchical procedure and polarizable force field. *J. Phys. Chem. B* 2003, *20,* 4862-4870.

18. Garmer, D. R.; Gresh, N.; Roques, B. P. (1998) Modeling of inhibitor-metalloenzyme interactions and selectivity using molecular mechanics grounded in quantum chemistry. *Proteins 1,* 42-60.

19. Sakharov, D. V.; Lim, C. (2005) Zn protein simulations including charge transfer and local polarization effects. *J. Am. Chem. Soc. 13,* 4921-4929.

20. Wu, J. C.; Piquemal, J. P.; Chaudret, R.; Reinhardt, P.; Ren, P. Y. (2010) Polarizable Molecular Dynamics Simulation of Zn(II) in Water Using the AMOEBA Force Field. *J. Chem. Theory Comput. 7,* 2059-2070.

21. Zhang, J. J.; Yang, W.; Piquemal, J. P.; Ren, P. Y. (2012) Modeling Structural Coordination and Ligand Binding in Zinc Proteins with a Polarizable Potential. *J Chem Theory Comput. 4,* 1314-1324.

22. Xiang, J. Y.; Ponder, J. W. (2013) A valence bond model for aqueous Cu(II) and Zn(II) ions in the AMOEBA polarizable force field. *J Comput Chem*. *9,* 739-749.

23. Li, P.; Merz, K. M. (2017) Metal Ion Modeling Using Classical Mechanics. *Chem. Rev. (Washington, DC, U. S.) 3,* 1564-1686.

24. Peters, M. B.; Yang, Y.; Wang, B.; Fusti-Molnar, L.; Weaver, M. N.; Merz, K. M., Jr. (2010) Structural Survey of Zinc Containing Proteins and the Development of the Zinc AMBER Force Field (ZAFF). *J. Chem. Theory Comput. 9,* 2935-2947.

25. Hu, L. H.; Ryde, U. (2011) Comparison of Methods to Obtain Force-Field Parameters for Metal Sites. *J. Chem. Theory Comput. 8,* 2452-2463.

26. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A.(2004) Development and testing of a general amber force field. *J. Comput. Chem. 9,* 1157-1174.

27. Konidaris, K. F.; Papi, R.; Katsoulakou, E.; Raptopoulou, C. P.; Kyriakidis, D. A.; Manessi-Zoupa, E. (2010) Synthesis, Crystal Structures, and DNA Binding Properties of Zinc(II) Complexes with 3-Pyridine Aldoxime. *Bioinorg. Chem. Appl.*

28. D.A. Case, J. T. B., R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K.M. Merz, G. Monard, P. Needham, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, R. Salomon-Ferrer, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, D.M. York and P.A. Kollman (2015), AMBER 2015

29. Hoops, S. C.; Anderson, K. W.; Merz, K. M. (1991) Force-Field Design for Metalloproteins. *J. Am. Chem. Soc. 22,* 8262-8270.

30. Besler, B. H.; Merz, K. M.; Kollman, P. A. (1990) Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem. 4,* 431-439.

31. Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. (1995) Application of the Multimolecule and Multiconformational Resp Methodology to Biopolymers - Charge Derivation for DNA, Rna, and Proteins. *J. Comput. Chem. 11,* 1357-1377.

32. Singh, U. C.; Kollman, P. A. (1984) An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem. 2,* 129-145.

33. Heinz, H.; Suter, U. W. (2004) Atomic charges for classical simulations of polar systems. *J Phys Chem B 47,* 18341-18352.

34. Jauch, R.; Bourenkov, G. P.; Chung, H. R.; Urlaub, H.; Reidt, U.; Jackle, H.; Wahl, M. C. (2003) The zinc finger-associated domain of the Drosophila transcription factor grauzone is a novel zinc-coordinating protein-protein interaction module. *Structure 11,* 1393-1402.

35. Wu, K. Y.; Hsieh, C. C.; Horng, Y. C. (2009) Mononuclear zinc(II) and mercury(II) complexes of Schiff bases derived from pyrrolealdehyde and cysteamine containing intramolecular NH center dot center dot center dot S hydrogen bonds. *J. Organomet. Chem. 13,* 2085-2091.

36. Escorihuela, I.; Falvello, L. R.; Tomas, M.; Urriolabeitia, E. P. (2004) Influence of noncovalent interactions on uracil tautomer selection: Coordination of both N1 and N3 uracilate to the same metal in the solid state. *Cryst. Growth Des. 4,* 655-657.

37. Bonamico, M.; Dessy, G.; Fares, V.; Scaramuz. L (1971) Crystal and Molecular Structures of Nickel(Ii) and Zinc(Ii) Bis(Trithioperoxybenzoates). *J. Chem. Soc. A 20,* 3191-&.

38. Lesburg, C. A.; Huang, C.; Christianson, D. W.; Fierke, C. A. (1997) Histidine --> carboxamide ligand substitutions in the zinc binding site of carbonic anhydrase II alter metal coordination geometry but retain catalytic activity. *Biochemistry 50,* 15780-15791.

39. Rodriguez, A.; Sousa-Pedrares, A.; Garcia-Vazquez, J. A.; Romero, J.; Sousa, A. (2009) Electrochemical synthesis and characterization of zinc(II) complexes with pyrimidine-2-thionato ligands and their adducts with N,N donors. *Polyhedron 11,* 2240-2248.

40. Rajsekhar, G.; Rao, C. P.; Saarenketo, P. K.; Kolehmainen, E.; Rissanen, K. (2002) C-S bond cleavage by cobalt: synthesis, characterization and crystal structure determination of 1,2-di-(o-salicylaldiminophenylthio)ethane and its Co(III) product with C-S bond cleaved fragments. *Inorg. Chem. Commun. 9,* 649-652.

41. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res. 1,* 235-242.

42. Hsin, K.; Sheng, Y.; Harding, M. M.; Taylor, P.; Walkinshaw, M. D. (2008) MESPEUS: a database of the geometry of metal sites in proteins. *J. Appl. Crystallogr.* 963-968.

43. Allen, F. H. (2002) The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.* 380-388.

44. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. (1977) Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys. 3,* 327-341.

45. Macke, T. J.; Case, D. A. (1998) Modeling unusual nucleic acid structures. *ACS Symp. Ser.* 379-393.

46. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. (2009) *Gaussian 09*, Gaussian, Inc.: Wallingford, CT, USA

47. Lin, F.; Wang, R. X. (2010) Systematic Derivation of AMBER Force Field Parameters Applicable to Zinc-Containing Systems. *J. Chem. Theory Comput. 6,* 1852-1870.

CHAPTER 4

Deep Learning in Toxicity Prediction with One-Dimensional Similarity

4.1. Introduction

Drug discovery and development is a process that requires a significant amount of time and resources to successfully see to completion. Hence, there is a great need for robust computational methods to accelerate and reduce the costs associated with the process. Many methods have emerged recently including data mining, machine learning[1] and deep learning[2, 3] to address various aspects of the drug design workflow.

Some traditional machine learning methods like Random Forest (RF) and Support Vector Machine (SVM) have been popular in predicting structure-activity relationships in drug discovery.[4, 5] They offer high prediction accuracy and are easy to implement. Machine learning methods describe a molecular system using a range of chemical descriptors or "features". The resultant vector of descriptors is input into the model to subsequently generate predictions about the properties of interest. Random Forest is an ensemble machine learning model built upon a collection of Decision Trees combined with bootstrap sampling and random feature selection. The output is determined by majority vote or the mean prediction of individual trees. Random Forest outperforms a single Decision Tree because it reduces variance and limits overfitting. Unlike common machine learning algorithm that reduce the dimensionality of a problem, SVM actually increases the dimension of the data space and builds a hyperplane in the transformed feature space using kernel functions and margin maximizing techniques. There are also several other related methods that are commonly used including k-nearest neighbors[6, 7], naïve Bayes classifier[8] and artificial neural networks[9].

Deep learning is a high-level machine learning algorithm, which is built on the basis of artificial neural networks with many layers of neurons. It is also called a deep neural network (DNN). A DNN can be considered as a process of data abstraction and transformation, or more specifically, a function mapping the input vector to an output vector. A layer in DNN is composed of up to thousands of neurons. Each neuron represents an abstract feature, which is activated by applying the activation function to the computed values of all the neurons in the previous layer. With an increased number of layers and neurons, higher levels of abstraction of the input features is done with the intent of teasing out more detailed information contained in the data relative to a shallow neural network.

Accurately predicting the toxicity of a compound prior to its synthesis and biological testing is one of the ways in which computations could save the pharmaceutical and biotechnology industries time and money ultimately by failing early in the process rather than in the more expensive later stages of the drug discovery workflow. Experimentally, the effects of chemicals on human health have been evaluated by both *in vivo* tests on animals and via high-throughput screening (HTS)[10]. However, the drawback of these techniques is for the former ethics concerns and for the later the cost. So that a growing number of researchers have been working on developing computational models to predict the toxicity of compounds using both traditional machine learning algorithms and Deep Learning.

The Tox21 10k compound library[11] is a well-known data set for building models for toxicity. The data set consists of the results generated for 12 different pathway assays.[12-16] Among the submissions of the Tox21 Challenge, many models have been built to predict toxicity and most of

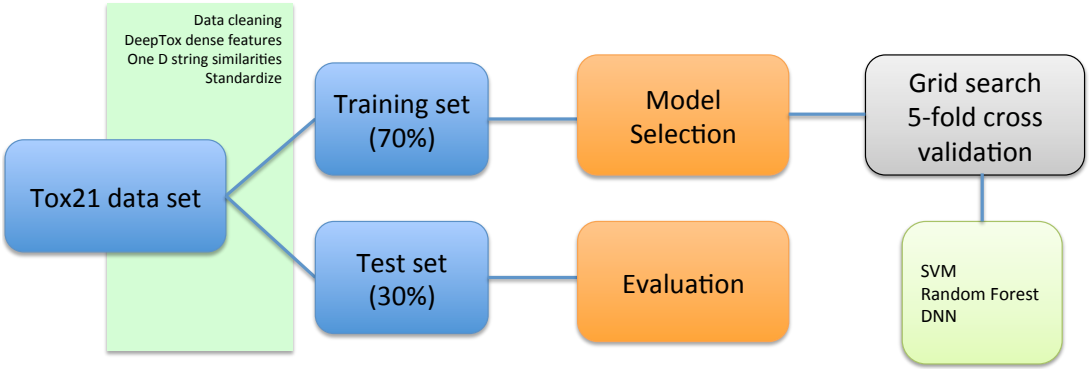
them are machine learning models, including support vector machine (SVM), Random Forest (RF) and Deep Learning.[17-25] The winning method, DeepTox, used a multi task Deep Learning model and performed the best in 6 of the assays.[22] DeepChem was later introduced to use one-shot learning in computational drug discovery including toxicity prediction, which required smaller amount of data.[26]

Herein, we present the models we built to predict toxicity using both chemical descriptors and one-dimensional similarities as molecular features. These models cover SVM, RF and DNN. We did thorough model selection with 5-fold cross validation and model evaluation on a test set. The quality metric we used is AUCs for these classifiers. We are able to outperform the top methods in the Tox21 challenge for most of the tasks. And the one-dimensional similarities proved to be good features to add on top of chemical descriptors to improve the predictive power of our machine learning model.

4.2. Method

An overview of the workflow is shown in Figure 4.1. We have used various machine learning models including SVM, Random Forest and Deep Learning to predict toxicity using the Tox21 dataset. The dataset contains the results from 12 nuclear receptor assays relevant to human toxicity. After data cleaning based on the structures, 8989 were kept out from more than 10k compounds. The criterion for structure selection is that there's no metal ions involved and the molecule can be represented by a one-dimensional representation, which is discussed below. The training and test sets were generated by randomly splitting the data set using a 7:3 ratio. The details of the molecular features and machine learning process are covered in subsequent sections.

Figure 4.1. Workflow overview



4.2.1 Molecular Features

4.2.1.1. Chemical descriptors.

The molecular features included in our machine learning models are a combination of chemical descriptors and one-dimensional similarities. DeepTox is the pipeline developed by the winning team of the Tox21 Challenge[27]. The chemical descriptors included in our model are static descriptors calculated with DeepTox. These include atom counts, various surface area values (polar, nonpolar, etc.), definitions of thousands of predefined toxicophores, Van der Waals volume, partial charges, etc..[28, 29]

4.2.1.2. One-dimensional similarity.

Apart from chemical descriptors, pair wise one-dimensional similarities were also included as model features. The similarity is computed as the normalized maximum overlap area of the one-dimensional representations of two molecules using the method developed by Dixon et al.[30] 1D representations are generated through projecting atoms in a molecule from 3-D onto a 1D primary axis followed by BFGS optimization. The primary axis is determined through principal component

analysis of the Gram matrix. Once the similarity matrix of the whole Tox21 data set was available, we included certain number of top similarity values as features compared to toxic compounds in the training set for each assay target. 4 sets of similarity values were compared: 5, 10, 15 or 20. For example, when the number is 5 we would pick for the molecule in question the largest 5 similarity values against the toxic compounds in the training set as new input variables appended to the chemical descriptors.

4.2.2 Machine Learning

Various classifiers have been built including SVM, random forest (RF) and Deep Neural Network (DNN) models. A grid search with 5-fold cross validation were carried out to optimize hyperparameter for each model. The selected models were then evaluated using the test set.

4.2.2.1. Support Vector Machines (SVM).

SVM is a robust method and widely used in modeling chemical properties.[5, 31, 32] SVM selects a kernel function to map the data points from the input space to a higher-dimensional space (feature space) where a separating hyperplane is defined. There are several parameters used to define a SVM model. Optimal parameters are not identical for each of the 12 tasks, so that hyperparameters were tuned through grid search over all combinations of kernel types including soft margin parameter C, class weight and Gamma specific to radial basis function (RBF) kernels. Best sets of hyperparameters were determined based on 5-fold cross validation. It was observed that the RBF kernel performs better for most of the cases.

4.2.2.2. Random Forests (RF).

Random Forest[33] classifier is an ensembled algorithm constructed by a collection of decision trees. A single decision tree is a weak estimator randomly, however, RF combines several weak estimators to form a stronger one. Similar to SVM, a grid search was done to find the optimal hyperparameters. The tuning parameters for RF are the number of trees in forest, the number of features considered when looking for the best split, the maximum depth of the tree and the minimum number of samples required to be at a leaf node.

4.2.2.3. Deep Neural Networks (DNNs).

The method we used was the standard fully connected multilayer neural network. For each task, the input layer takes an input vector including molecular features and similarity values. The hidden layers consist of 3 or 4 layers with rectified linear units (ReLU), which is by far the most popular activation function in DNN.[2] Since predicting toxicity is a binary classification problem, the output is either one for having a toxic effect or zero for being nontoxic. In the output layer of the DNN, softmax activation is applied to get an estimated probability. The two classes are mutually exclusive for each task, so that a loss function of average softmax cross entropy is used to measure the probability error. Learning minimizes this error with the Adam algorithm as the optimizer. Grid search and cross validation were implemented to select the best set of hyperparameters including the number of layers, number of hidden nodes, activation function, batch size, learning rate, dropout rate and class weight. It was discovered that the optimal setting for the DNN varies from task to task.

4.3. Results and Discussion

4.3.1. Benefit of One-Dimensional similarity

To investigate whether one-dimensional similarities improves the performance, I compared AUC results of various learning methods with DeepTox (DT) features only and DT plus one-dimensional similarities. Table 4.1 lists the resulting AUC values and indicates the best result for each task in bold font. If comparing all the combinations of features across three kinds of learning methods (SVM, DNN and RF), for 6 out of 12 tasks, DT plus similarities outperformed DT only. For 3 out of 12 tasks, they performed equally. For either DNN or RF, including similarities make AUC results better for 8 out of 12 tasks. Whereas, for SVM, similarities only improved the results for 4 assays. Table 4.2 shows the average AUC for each method. In general, similarities enhance the performance of DNN and RF models when predicting toxicity with DT static features.

4.3.2. Comparison of DNN, RF and SVM

I selected the best-performing models from each machine learning method through grid search 5-fold cross validations and evaluated them on the final test set. The methods I compared were DNN, SVM and RF. As shown in Table 4.1 and 4.2, RF is superior to the other two methods by achieving the best mean AUC on the test set and performing the best in 9 out of 12 assays.

To summarize, in this chapter, I have introduced one-dimensional similarities as new features to toxicity prediction and have tested the performance of three different machine learning methods. AUC results show the benefit of including similarities for both RF and DNN but not for SVM. Finally, I found it interesting that RF achieved the best performance out of the three approaches.

Table 4.1. AUC results for different learning methods and different input features for each task

Models	Features	Tasks											
		nr-ahr	nr-ar-lbd	nr-aromatase	nr-ar	nr-er-lbd	nr-er	nr-ppar-gamma	sr-are	sr-atad5	sr-hse	sr-mmp	sr-p53
DNN	DT	0.922	0.900	0.815	0.837	0.866	0.808	0.798	0.831	0.903	**0.806**	0.913	0.896
	DT+ Top 5 similarities	0.922	0.884	0.818	0.831	0.862	0.806	0.810	0.825	0.898	0.763	0.911	0.902
	DT+ Top 10 similarities	0.919	0.889	0.833	0.828	0.870	0.792	0.823	0.815	0.895	0.782	0.914	0.897
	DT+ Top 15 similarities	0.918	0.903	0.827	0.836	0.865	0.792	0.813	0.828	0.905	0.778	0.910	0.893
	DT+ Top 20 similarities	0.914	0.904	0.817	0.801	0.874	0.797	0.807	0.826	0.894	0.777	0.909	0.900
RF	DT	**0.931**	**0.924**	0.874	0.875	0.878	0.809	0.829	0.860	0.915	**0.806**	**0.928**	0.900
	DT+ Top 5 similarities	0.927	**0.924**	**0.882**	**0.878**	0.880	0.806	**0.837**	**0.861**	0.913	0.800	0.925	0.899
	DT+ Top 10 similarities	0.928	0.906	0.872	0.875	0.876	0.804	0.836	0.860	0.909	0.799	0.925	0.903
	DT+ Top 15 similarities	0.926	0.910	0.872	0.876	0.880	0.807	0.830	0.855	0.906	0.800	0.925	0.896
	DT+ Top 20 similarities	0.929	0.911	0.873	0.874	0.876	0.804	0.832	0.857	**0.915**	0.801	0.924	0.900
SVM	DT	0.919	0.886	0.866	0.851	**0.892**	0.816	0.816	0.836	0.882	0.787	0.915	0.910
	DT+ Top 5 similarities	0.917	0.884	0.857	0.855	0.890	0.817	0.815	0.834	0.881	0.783	0.915	0.913
	DT+ Top 10 similarities	0.917	0.884	0.856	0.855	0.891	0.817	0.815	0.833	0.879	0.781	0.915	**0.914**
	DT+ Top 15 similarities	0.917	0.884	0.857	0.855	0.891	**0.818**	0.815	0.834	0.876	0.782	0.914	0.913
	DT+ Top 20 similarities	0.917	0.885	0.857	0.855	0.891	0.817	0.812	0.834	0.876	0.780	0.914	0.911

DT stands for Static chemical descriptors calculated with DeepTox.

Table 4.2. Average AUC results for different learning methods

Models	Features	Average AUC
DNN	DT	0.858
	DT + similarities	0.861
RF	DT	0.877
	DT + similarities	0.879
SVM	DT	0.865
	DT + similarities	0.864

The average AUCs are calculated over best AUC of each task for "DT + similarities"

REFERENCES

REFERENCES

1. Mitchell, J. B. O. (2014) Machine learning methods in chemoinformatics. *Wires Comput Mol Sci.* 4, 468-481.

2. LeCun, Y.; Bengio, Y.; Hinton, G. (2015) Deep learning. *Nature* 521, 436-444.

3. Ma, J. S.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. (2015) Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *J Chem Inf Model* 55, 263-274.

4. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. (2003) Random forest: A classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comp Sci* 43, 1947-1958.

5. Xu, Y.; Zomer, S.; Brereton, R. G. (2006) Support Vector Machines: A recent method for classification in chemometrics. *Crit Rev Anal Chem* 36, 177-188.

6. Ajmani, S.; Jadhav, K.; Kulkarni, S. A. (2006) Three-dimensional QSAR using the k-nearest neighbor method and its interpretation. *J Chem Inf Model* 46, 24-31.

7. Shen, M.; Xiao, Y. D.; Golbraikh, A.; Gombar, V. K.; Tropsha, A. (2003) Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. *Journal of medicinal chemistry* 46, 3013-3020.

8. Sun, H. M. (2005) A naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *Journal of medicinal chemistry* 48, 4031-4039.

9. Agatonovic-Kustrin, S.; Beresford, R. (2000) Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J Pharmaceut Biomed* 22, 717-727.

10. O'Brien, P. J.; Irwin, W.; Diaz, D.; Howard-Cofield, E.; Krejsa, C. M.; Slaughter, M. R.; Gao, B.; Kaludercic, N.; Angeline, A.; Bernardi, P.; Brain, P.; Hougham, C. (2006) High concordance of drug-induced human hepatotoxicity with in vitro cytotoxicity measured in a novel cell-based model using high content screening. *Arch Toxicol* 80, 580-604.

11. Huang, R. L.; Sakamuru, S.; Martin, M. T.; Reif, D. M.; Judson, R. S.; Houck, K. A.; Casey, W.; Hsieh, J. H.; Shockley, K. R.; Ceger, P.; Fostel, J.; Witt, K. L.; Tong, W. D.; Rotroff, D. M.; Zhao, T. G.; Shinn, P.; Simeonov, A.; Dix, D. J.; Austin, C. P.; Kavlock, R. J.; Tice, R. R.; Xia, M. H. (2014) Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. *Sci Rep-Uk* 4.

12. Bartkova, J.; Horejsi, Z.; Koed, K.; Kramer, A.; Tort, F.; Zieger, K.; Guldberg, P.; Sehested, M.; Nesland, J. M.; Lukas, C.; Orntoft, T.; Lukas, J.; Bartek, J. (2005) DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. *Nature* 434, 864-870.

13. Chawla, A.; Repa, J. J.; Evans, R. M.; Mangelsdorf, D. J. (2001) Nuclear receptors and lipid physiology: Opening the X-files. *Science* 294, 1866-1870.

14. Grun, F.; Blumberg, B. (2007) Perturbed nuclear receptor signaling by environmental obesogens as emerging factors in the obesity crisis. *Rev Endocr Metab Dis* 8, 161-171.

15. Jaeschke, H.; McGill, M. R.; Ramachandran, A. (2012) Oxidant stress, mitochondria, and cell death mechanisms in drug-induced liver injury: Lessons learned from acetaminophen hepatotoxicity. *Drug Metab Rev* 44, 88-106.

16. Labbe, G.; Pessayre, D.; Fromenty, B. (2008) Drug-induced liver injury through mitochondrial dysfunction: mechanisms and detection during preclinical safety studies. *Fund Clin Pharmacol* 22, 335-353.

17. Abdelaziz, A.; Spahn-Langguth, H.; Schramm, K.-W.; Tetko, I. V. (2016) Consensus Modeling for HTS Assays Using In silico Descriptors Calculates the Best Balanced Accuracy in Tox21 Challenge. *Frontiers in Environmental Science* 4.

18. Barta, G. (2016) Identifying Biological Pathway Interrupting Toxins Using Multi-Tree Ensembles. *Frontiers in Environmental Science* 4.

19. Capuzzi, S. J.; Politi, R.; Isayev, O.; Farag, S.; Tropsha, A. (2016) QSAR Modeling of Tox21 Challenge Stress Response and Nuclear Receptor Signaling Toxicity Assays. *Frontiers in Environmental Science* 4.

20. Drwal, M.; Siramshetty, V.; Banerjee, P.; Goede, A.; Preissner, R.; Dunkel, M. (2015) Molecular similarity-based predictions of the Tox21 screening outcome. *Frontiers in Environmental Science* 3.

21. Koutsoukas, A.; St. Amand, J.; Mishra, M.; Huan, J. (2016) Predictive Toxicology: Modeling Chemical Induced Toxicological Response Combining Circular Fingerprints with Random Forest and Support Vector Machine. *Frontiers in Environmental Science* 4.

22. Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. (2016) DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science* 3.

23. Ribay, K.; Kim, M. T.; Wang, W.; Pinolini, D.; Zhu, H. (2016) Predictive Modeling of Estrogen Receptor Binding Agents Using Advanced Cheminformatics Tools and Massive Public Data. *Frontiers in Environmental Science* 4.

24. Stefaniak, F. (2015) Prediction of Compounds Activity in Nuclear Receptor Signaling and Stress Pathway Assays Using Machine Learning Algorithms and Low-Dimensional Molecular Descriptors. *Frontiers in Environmental Science* 3.

25. Uesawa, Y. (2016) Rigorous Selection of Random Forest Models for Identifying Compounds that Activate Toxicity-Related Pathways. *Frontiers in Environmental Science* 4.

26. Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. (2017) Low Data Drug Discovery with One-Shot Learning. *Acs Central Sci* 3, 283-293.

27. Mayr, A.; Günter, K.; Thomas, U.; Sepp, H. (2016) DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science* 3, 17-31.

28. Cao, D. S.; Xu, Q. S.; Hu, Q. N.; Liang, Y. Z. (2013) ChemoPy: freely available python package for computational biology and chemoinformatics. *Bioinformatics* 29, 1092-1094.

29. Kazius, J.; McGuire, R.; Bursi, R. (2005)Derivation and validation of toxicophores for mutagenicity prediction. *Journal of medicinal chemistry* 48, 312-320.

30. Dixon, S. L.; Merz, K. M., Jr. (2001) One-dimensional molecular representations and similarity calculations: methodology and validation. *Journal of medicinal chemistry* 44, 3795-809.

31. Li, H. D.; Liang, Y. Z.; Xu, Q. S. (2009) Support vector machines and its applications in chemistry. *Chemometr Intell Lab* 95, 188-198.

32. Niazi, A.; Jameh-Bozorghi, S.; Nori-Shargh, D. (2008) Prediction of toxicity of nitrobenzenes using ab initio and least squares support vector machines. *J Hazard Mater* 151, 603-609.

33. Breiman, L. (2001) Random forests. *Mach Learn* 45, 5-32.

CHAPTER 5

Conclusions and Future Outlook

The objective of this dissertation was to implement data science techniques in computational chemistry/biology using building machine learning (ML) models to predict protein-ligand complex structure, develop zinc ion force fields and predict the toxicity of drug molecules.

The first chapter after the introduction in this thesis describes a fast and accurate methodology that uses protein chemical shift perturbations (CSP) to determine the structure of protein-ligand complexes[1]. The whole methodology consists of two parts. Firstly, I introduced a new regression model called HECSP to calculate ligand binding-induced proton CSPs in a protein. On the basis of HECSP, I further built a scoring function NMRScore_P which can rank the poses according to the discrepancy between computed and observed proton CSPs. Over all, HECSP coupled with NMRScore_P provides an accurate and rapid platform to refine protein-ligand complexes using NMR-derived information.

Besides using a regression model in the development of HECSP, I also demonstrated the application of regression models in force field development. Based on the zinc AMBER force field (ZAFF), I developed an empirical approach called extended ZAFF (EZAFF) to determine the parameters for bonds and angles involving zinc.[2] The advantage of EZAFF is that it can handle not only four-coordinated systems like ZAFF but also five- and six-coordinated systems. The reliability of EZAFF has been validated by tests on twelve different systems including metalloproteins and organometallic compounds with various coordination numbers. Further benchmark calculations were done as well to evaluate the relative performance of eleven methods for simulating zinc containing systems. It was discovered that the MM models could yield a comparable performance with a reduced computational cost compared to semiempirical models.

After building regression models for protein CSPs prediction and zinc ion force field development, we next applied ML and deep learning (DL) to predict the toxicity of small molecule compounds. I have presented the models I built in Chapter 4, which include SVM, RF and DNN. The dataset I used is from the Tox21 challenge. And the features included in my models are DeepTox[3] static features and inter-molecular similarities based on a one-dimensional molecular representations[4]. The study has demonstrated the benefit of adding similarities on top of static features in terms of toxicity prediction. It is interesting to note that RF provides the best performance with the highest AUC on test set.

Overall, the ultimate goal of my work is to implement ML in the field of computational chemistry and biology, where I have built several machine learning models ranging from basic linear regressions to DL models. It is shown that ML is indeed a strong alternative for physics-based models with enough high-quality data, especially for DL.

ML and DL has gained more attention and has achieved good performance in computational chemistry and biology with access to big data, improved algorithms and powerful computers. However, it is not a clear-cut decision to totally shift gears to ML and DL despite of all the previous success stories in this field. It is still a challenge to find the right representations of chemical or biological molecules which bear the causal relationship with the property observed in experiments.[5] Also, in order to build models with satisfying performance, the training dataset is the main factor. Another challenge for ML in chemistry is that the training data is usually coming from experiments, which means it is often sparse and unbalanced. A closer collaboration between

experimentalists and computational chemists is crucial to create data sets of good quality and better serve the machine learning community. To fully make the best use of ML and DL in computational chemistry the models and performance metrics need to be carefully tailored to specific problem. Even with these challenges, I am convinced that ML and DL will help drive the advancement in computational chemistry and biology.

REFERENCES

REFERENCES

1. Yu, Z., Li, P., and Merz, K. M. (2017) Using Ligand-Induced Protein Chemical Shift Perturbations To Determine Protein–Ligand Structures. *Biochemistry 56*, 2349–2362.

2. Yu, Z., Li, P., and Merz, K. M. (2018) Extended Zinc AMBER Force Field (EZAFF). *J. Chem. Theory Comput. 14*, 242–254.

3. Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016) DeepTox: Toxicity Prediction using Deep Learning. *Front. Environ. Sci. 3*, 80.

4. Dixon, S. L., and Merz, K. M. (2001) One-dimensional molecular representations and similarity calculations: methodology and validation. *J. Med. Chem. 44*, 3795–809.

5. Hochreiter, S., Klambauer, G., and Rarey, M. (2018) Machine Learning in Drug Discovery. *J. Chem. Inf. Model. 58*, 1723–1724.