Marquette University
e-Publications@Marquette

Dissertations (2009 -) Dissertations, Theses, and Professional Projects

Novel Numerical Models of Electrostatic Interactions and Their Application to S-Nitrosothiol Simulations

Maxim Vadimovich Ivanov
Marquette University

Recommended Citation

Ivanov, Maxim Vadimovich, "Novel Numerical Models of Electrostatic Interactions and Their Application to S-Nitrosothiol Simulations" (2016). *Dissertations (2009 -)*. Paper 641.
http://epublications.marquette.edu/dissertations_mu/641

NOVEL NUMERICAL MODELS OF ELECTROSTATIC INTERACTIONS AND THEIR APPLICATION TO S-NITROSOTHIOL SIMULATIONS

by
Maxim V. Ivanov, B.S.

A Dissertation submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy

Milwaukee, Wisconsin
May 2016

ABSTRACT
NOVEL NUMERICAL MODELS OF ELECTROSTATIC INTERACTIONS AND THEIR APPLICATION TO S-NITROSOTHIOL SIMULATIONS

Maxim V. Ivanov, B.S.

Marquette University, 2016

Atom-centered point charge model of the molecular electrostatics remains a major workhorse in the atomistic biomolecular simulations. However, this approximation fails to reproduce anisotropic features of the molecular electrostatic potential (MEP), and the existing methods of the charge derivation are often associated with the numerical instabilities. This work provides an in-depth analysis of these limitations and offers a novel approach to describe electrostatic interactions that paves the way toward efficient next-generation force fields.

By analyzing the charge fitting problem from first principles, as an example of the mathematical inverse problem, we show that the numerical instabilities of the charge-fitting problem arise due to the decreasing contribution from the higher multipole moments to the overall MEP. This insight suggests that if the point charges are arranged over the sphere using Lebedev quadrature, the resulting point charge model is able to exactly reproduce multipoles up to a given rank. At the same time, point charge values can be derived without fitting to the MEP, avoiding numerically unstable method of the charge derivation. This approach provides a systematic way to reproduce multipole moments up to any rank within the point charge approximation, which makes this model a computationally efficient analog of the multipolar expansion. Moreover, the proposed charged sphere model can be also used in the multi-site expansions with the expansion centers located at each atom in a molecule. This provides a natural approach to expand the traditional atom-centered point charge approximation to include higher-rank atomic multipoles and to account for the anisotropy of the MEP.

We applied the proposed charged sphere model to S-nitrosothiols (RSNOs)—a class of biomolecules that serves to store and transmit nitric oxide, a biologically important signaling molecule. We showed that when the atom-centered charged spheres are optimized together with the Lennard-Jones parameters, the resulting force field can accurately reproduce the anisotropic features of the intermolecular interactions that play a crucial role in the biological regulation of RSNO chemistry. Overall, the developed charge model is a promising approach that can be used in the biomolecular simulations and beyond, e.g. in the multipolar force fields for atomistic and coarse-grained simulations.

ACKNOWLEDGMENTS

Maxim V. Ivanov, B.S.

I thank my advisor, Dr. Qadir Timerghazin, for his exceptional mentorship and support throughout my graduate study. His dedication to my success and numerous advice were essential to the completion of this dissertation.

I also acknowledge my committee members, Dr. Dmitri Babikov, Dr. James Gardinier and Dr. Daniel Sem, for their suggestions and careful consideration of my research.

I thank my group members, Elena Ivanova, Dmitry Khomyakov, Matthew Flister and Marat Talipov. Special thanks to my group member, Elena Ivanova, who happened to be my wonderful wife. Her love and constant care was in the end what made this dissertation possible. I am also thankful to the friendship of Marat Talipov and the numerous inspiring conversions I had with him during these years.

Finally, I thank my family, my parents and brother, for their love and encouragement throughout my life.

TABLE OF CONTENTS

ACKNOWLEDGMENTS i

LIST OF TABLES vi

LIST OF FIGURES xvii

1 INTRODUCTION 1

- 1.1 Force Field Models of Intermolecular Interactions 1
- 1.2 S-Nitrosothiols and Their Biological Role 8
- 1.3 Objectives 14

2 OVERVIEW OF THEORETICAL AND NUMERICAL METHODS 15

  - 2.0.1 Spherical Harmonics 15
  - 2.0.2 Molecular Electrostatic Potential 17
  - 2.0.3 Multipolar Expansion of Electrostatic Potential 19
- 2.1 Methods of Matrix Decomposition 20
  - 2.1.1 Eigendecomposition 21
  - 2.1.2 Singular Value Decomposition 21
  - 2.1.3 Least Squares Approximation 24
- 2.2 Minimization algorithms 27
  - 2.2.1 Newton's Algorithms 27
  - 2.2.2 Evolutionary Algorithms 28

3 GENETIC ALGORITHM OPTIMIZATION OF POINT CHARGES IN FORCE FIELD DEVELOPMENT 32

3.1 Introduction 32

3.2 Details of Charge Fitting 35

3.2.1 Least Squares Fitting 35

3.2.2 Fitting with Genetic Algorithms 36

3.3 GA Charge Fitting for Small Molecules 38

3.4 Covariance Matrix Analysis 43

3.5 Rotation of the Optimization Coordinates 47

3.6 Large-Molecule Example 50

3.7 Variance of the Least Squares Solution and the Buried Atom Effect 52

3.8 Summary 55

3.9 Computational Details 56

4 REVEALING THE ILL-CONDITIONING OF THE CHARGE FITTING PROBLEM 58

4.1 Introduction 58

4.2 Point Charge Fitting as an Inverse Problem 61

4.3 The Two-Sphere Model 64

4.4 Analytical Point Charge Model 67

4.5 Lebedev vs. Atom-Centered Model: Numerical Example 70

4.6 Total-Charge Constraint 77

4.7 Summary 80

4.8 Computational Details 82

5 POINT CHARGES MEET ACCURACY OF MULTIPOLES 83

5.1 Introduction 83

5.2 Point-Charge Representation of the Multipolar Expansion 84

5.3 Numerical Examples of the Lebedev Charge Model 88

5.3.1 Modeling Single-Site Molecular Multipoles 88

5.4 Modeling Multi-Site Atomic Multipoles 90

5.5 Summary 94

5.6 Computational Details 94

6 APPLICATION OF THE MODEL TO S-NITROSOTHIOLS 96

6.1 Simultaneous Fitting of Several Force Field Terms for CysNO . . . 96

6.1.1 Bonded Terms: Equilibrium Bond lengths, Angles and Force Constants 96

6.1.2 Non-Bonded Terms: Point Charges and Lennard-Jones Parameters 97

6.2 Summary 103

6.3 Computational Details 103

7 GSNO SYNTHESIS AND NMR SPECTROSCOPY 105

8 CONCLUSIONS 109

BIBLIOGRAPHY 112

Appendices 123

A STATISTICAL DATA ON THE ELECTROSTATIC PROPERTIES OF THE MODEL 124

A.1 Ammonia 124

A.2 Bromomethane 124

A.3 Chloromethane 128

A.4 cis-MeSNO 130

A.5 Fluoromethane 133

A.6 Formamide 135

A.7 Furan 138

A.8 Imidazol 140

A.9 Methanesulfonamide 143

A.10 Methanesulfonic acid 146

A.11 Methanethiol 149

A.12 Methanol 151

A.13 Tetrazole 153

A.14 Thiazole 156

A.15 trans-MeSNO 159

A.16 Uracil 162

A.17 Water 165

LIST OF TABLES

3.1 Parameters used in the GA fitting of the MEP point charges . . . 37

3.2 Genetic operators used in the GA fitting of the MEP point charges 37

3.3 Average values and the standard deviations (in parenthesis) of the monopole and dipole moments computed from the GA-optimized point charges for CH_3X, CH_2X_2 (X = F, Cl), and CH_3O^- molecules along with the reference values from DFT calculations. 42

3.4 Charge fitting for two- and three-charge model molecules: average fitness scores with standard deviations (in parenthesis) for the GA optimizations (200 runs with 30 chromosomes per generation) using the point charge coordinates vs the coordinates defined by the LS-sum Hessian eigenvectors, along with the fitness scores of the reference ESP solutions; all units are in kcal/mol. 50

3.5 Charge fitting for 1-chlorobutane conformers: average fitness scores with standard deviations (in parenthesis) for the GA optimizations using two coordinate systems (point charges and Hessian eigenvectors), along with the fitness scores of the CMA-ES and ESP solutions; all units are in kcal/mol. 52

4.1 Effect of the numerical rank r (SVD in eq. 4.50) and the total-charge constraint on the values of atom-centered PCs of methanol and the RMSD (kcal/mol). 74

4.2 Effect of the rank r (eq. 4.47), degree n (eq. 4.46) and type of the charge constraint on the methanol multipole moments and the RMSD (kcal/mol) within the Lebedev grid PC model ($a = 2$ au, $n = 1, 2$ and $N = 6, 14$) with probe sphere S_R ($R = 8$ au, $T = 194$) and atom-centered PC model with vdW-type grid. . . . 76

4.3 Eigenvalues μ_i^2 and eigenvectors $\mathbf{u}_i$ of the LS Hessian matrix $\mathbf{H}$ in constraint-free case and with the Lagrange multiplier to constraint the total charge. 78

5.1 Averaged over the set of reference molecules $\langle a \rangle$, minimum a_{min} and maximum a_{max} values of the radius required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 88

5.2 Comparison of the multi-site Lebedev ($a = 0.5$ au and $n = 2$) and atom-centered point charge models to reproduce AEPs. Averaged statistical parameters are reported, see Appendix A for values in individual cases. All dimensional quantities are in kcal/mol. . . . 94

6.1 Force constants and equilibrium values fitted to relaxed PES scans 97

6.2 Optimized SN and NO force constants in MeSNO in case of separate and combined optimization of each conformer. 102

6.3 Optimized non-bonded force field parameters of -SNO group in case of separate and combined optimization of each conformer. . . 102

7.1 Predicted chemical shifts (in ppm) in CH_2 hydrogens of ethylSNO relative to the ethylSH at PBE0/pcS-2 level of theory using gauge-independent atomic orbital (GIAO) method. 106

A-B1 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 125

A-D1 Molecular multipole moments Q_{lm} of ammonia calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles. 125

A-F1 Comparison of ammonia QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 125

A-G1 Comparison of ammonia QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. . . . 126

A-B2 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 126

A-D2 Molecular multipole moments Q_{lm} of bromomethane calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles. 126

A-F2 Comparison of bromomethane QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 127

A-G2 Comparison of bromomethane QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. . . . 127

A-B3 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 128

A-D3 Molecular multipole moments Q_{lm} of chloromethane calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles. 128

A-F3 Comparison of chloromethane QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 129

A-G3 Comparison of chloromethane QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. . . . 129

A-B4 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 130

A-D4 Molecular multipole moments Q_{lm} of cis-mesno calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles. 130

A-F4 Comparison of cis-mesno QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 131

A-G4 Comparison of cis-mesno QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. . . . 132

A-H4 Atom-centered point charge values of cis-mesno fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule. 132

A-I4 Atomic coordinates of cis-mesno optimized at mp2/aug-cc-pVTZ level of theory. 132

A-B5 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 133

A-D5 Molecular multipole moments Q_{lm} of fluoromethane calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles. 133

A-F5 Comparison of fluoromethane QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 134

A-G5 Comparison of fluoromethane QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. . . . 134

A-B6 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 135

A-D6 Molecular multipole moments Q_{lm} of formamide calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles. 135

A-F6 Comparison of formamide QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 136

A-G6 Comparison of formamide QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. . . . 137

A-H6 Atom-centered point charge values of formamide fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule. 137

A-I6 Atomic coordinates of formamide optimized at mp2/aug-cc-pVTZ level of theory. 137

A-B7 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 138

A-D7 Molecular multipole moments Q_{lm} of furan calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles. 138

A-F7 Comparison of furan QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 139

A-G7 Comparison of furan QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. 139

A-B8 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 140

A-D8 Molecular multipole moments Q_{lm} of imidazol calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles. 140

A-F8 Comparison of imidazol QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 141

A-G8 Comparison of imidazol QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. 142

A-H8 Atom-centered point charge values of imidazol fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule. 142

A-I8 Atomic coordinates of imidazol optimized at mp2/aug-cc-pVTZ level of theory. 142

A-B9 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 143

A-D9 Molecular multipole moments Q_{lm} of methanesulfonamide calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles. 143

A-F9 Comparison of methanesulfonamide QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 144

A-G9 Comparison of methanesulfonamide QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. 145

A-H9 Atom-centered point charge values of methanesulfonamide fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule. 145

A-I9 Atomic coordinates of methanesulfonamide optimized at mp2/aug-cc-pVTZ level of theory. 145

A-B10 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 146

A-D10 Molecular multipole moments Q_{lm} of methanesulfonic acid calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles. 146

A-F10 Comparison of methanesulfonic acid QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 147

A-G10 Comparison of methanesulfonic acid QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. 148

A-H10 Atom-centered point charge values of methanesulfonic acid fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule. 148

A-I10 Atomic coordinates of methanesulfonic acid optimized at mp2/aug-cc-pVTZ level of theory. 148

A-B11 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 149

A-D11 Molecular multipole moments Q_{lm} of methanethiol calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles. 149

A-F11 Comparison of methanethiol QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 150

A-G11 Comparison of methanethiol QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. . . . 150

A-B12 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 151

A-D12 Molecular multipole moments Q_{lm} of methanol calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles. 151

A-F12 Comparison of methanol QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 152

A-G12 Comparison of methanol QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. . . . 152

A-B13 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 153

A-D13 Molecular multipole moments Q_{lm} of tetrazole calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles. 153

A-F13 Comparison of tetrazole QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 154

A-G13 Comparison of tetrazole QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. 155

A-H13 Atom-centered point charge values of tetrazole fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule. 155

A-I13 Atomic coordinates of tetrazole optimized at mp2/aug-cc-pVTZ level of theory. 155

A-B14 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 156

A-D14 Molecular multipole moments Q_{lm} of thiazole calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models (a = 0.5 au, n = 2) as compared with the QM multipoles. 156

A-F14 Comparison of thiazole QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 157

A-G14 Comparison of thiazole QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. 158

A-H14 Atom-centered point charge values of thiazole fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule. 158

A-I14 Atomic coordinates of thiazole optimized at mp2/aug-cc-pVTZ level of theory. 158

A-B15 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 159

A-D15 Molecular multipole moments Q_{lm} of trans-mesno calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models (a = 0.5 au, n = 2) as compared with the QM multipoles. 159

A-F15 Comparison of trans-mesno QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 160

A-G15 Comparison of trans-mesno QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. . . . 161

A-H15 Atom-centered point charge values of trans-mesno fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule. 161

A-I15 Atomic coordinates of trans-mesno optimized at mp2/aug-cc-pVTZ level of theory. 161

A-B16 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 162

A-D16 Molecular multipole moments Q_{lm} of uracil calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles. 162

A-F16 Comparison of uracil QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 163

A-G16 Comparison of uracil QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. 164

A-H16 Atom-centered point charge values of uracil fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule. 164

A-B17 Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD. 165

A-D17 Molecular multipole moments Q_{lm} of water calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles. 165

A-F17 Comparison of water QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol. 166

A-G17 Comparison of water QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule. 166

LIST OF FIGURES

1.1 Intermolecular energy as a function of the separation distance . . 2

1.2 Effect of σ-hole on the molecular electrostatic potential of chloromethane CH_3Cl and methanethiol CH_3SH. Formation of the covalent σ bond leaves a region of diminished negative charge on its non-involved side along the extension of the bond. Calculations were performed at MP2/aug-cc-pVTZ level of theory; the charge density isosurface was plotted at 0.002 au. 5

1.3 S-Nitrosation of a cysteine in a peptide/protein. 8

1.4 Structures of two MeSNO isomers: cis and trans 9

1.5 Resonance representation of RSNO electronic structure as a combination of covalent (**S**), zwitterionic (**D**) and ion pair (**I**) resonance structures. 9

1.6 Molecular electrostatic potential of MeSNO around sulfur (top left) and nitrogen and oxygen (bottom left) atoms and the chalcogen- and hydrogen-bonded complexes with sulfur (top right) and hydrogen-bonded complexes with nitrogen and oxygen (bottom right). Calculations were performed at PBE0/def2-TZVPPD level of theory; the charge density isosurface was plotted at 0.002 au. 10

1.7 Two pathways of the reaction between S-nitrosothiol (RSNO) and thiol in neutral (RSH) and anionic (RS^-) states. 12

2.1 Linear transformation A maps a vector in U to a vector in V. Vectors in V that are transformed from vectors in U define image im(A). Vectors in U that are transformed to a zero in V define a kernel ker($\mathbf{A}$). 25

2.2 A simplified scheme of the genetic algorithm procedure. 29

3.1 GA convergence with 20 chromosomes in the population (a) as compared to 50 chromosomes in the population (b). 39

3.2 Distributions of the GA-optimized charges for the model molecules with two symmetry-independent charges, obtained from 200 GA runs with 20 chromosomes in the population. Yellow dots indicate the solutions obtained with the ESP method. 40

3.3 Correlations between the GA-optimized charges for the two-charge model molecules obtained from 200 independent GA runs with 20 chromosomes in the population; all trend lines have correlation coefficient $R^2 = 1.00$. Yellow dots indicate the solutions obtained with the ESP method. 41

3.4 The coordinate system for the CH_3X and CH_2X_2 molecules. . . . 41

3.5 The correlation between the chloromethane point charges obtained from 200 independent GA runs shown in three dimensions (A) and as two-dimensional projections, i.e. pairwise correlations between charges (B). 43

3.6 Numerical equivalence of the eigenvectors of the covariance matrix for the results of 200 GA runs, the eigenvectors of the least-squares sum Hessian matrix, and the normalized vectors $\mathbf{u}_1$ and $\mathbf{u}_2$, on the example of water molecule. 45

3.7 Fitness function profiles for the two-charge model molecules: full profiles (3D plots) and the profiles along the zero total charge line (2D plots). Red dots show the solutions obtained from GA optimizations (200 runs), and the yellow dots indicate the ESP solutions. 46

3.8 The effect of coordinate rotation on the convergence of GA minimizations on the example of a simple model function f of two variables associated with highly different curvatures: the model function plotted in the original coordinate system (A) and in the coordinate system rotated by 45°(B); the average of f_{min} values obtained from 50 GA minimization runs (blue) and the corresponding standard deviations (red) vs the rotation angle θ. . . . 48

3.9 Five conformers of 1-Chlorobutane with carbon atom numbering. . 50

3.10 Bar-chart representation of the eigenvectors of the Hessian matrix for five conformers of 1-chlorobutane. 52

3.11 Standard deviations σ of the charges $\mathbf{q}$ and the corresponding coordinates defined by the LS-sum Hessian eigenvectors $\mathbf{n}$ obtained from the solutions of 200 GA performed in terms of the charge coordinates(left), and in terms of the LS-sum Hessian eigenvector coordinates (right). 53

4.1 Schematic representations of the probe S_R and charged S_a spheres in the continuous (A) and discrete (B) forms. Operators K (eq. 4.28) and matrix $\widetilde{\mathbf{K}}$ are represented schematically. 64

4.2 Number of Lebedev quadrature points N (red triangles) and dimension $d_n = (n+1)^2$ (green circles) as functions of the degree n 67

4.3 Cross-section representations of the quadratures used for two-sphere model (A) ($n = 1$, $N = 6$ and $t = 11$, $T = 194$ for spheres S_a and S_R, respectively) as compared with the traditional atom-centered model (B). Green circles correspond to the point charges; blue circles correspond to the reference grid points 70

4.4 PC representation of the charged sphere S_a using the Lebedev quadrature with $n = 1$ and $n = 2$ as compared with the geometry of methanol molecule. 70

4.5 RMSD as the function of the number N of PCs on the charged sphere S_a 71

4.6 Few selected multipole moments $Q_{lm}^{S_a}$ of the charged sphere S_a as the functions of the probe sphere radius R. 72

4.7 Normalized singular values μ_i/μ_1 obtained using the exact analytical expression eq. 4.30 (green circles) as compared with the numerical values obtained from SVD of the LS matrix for the two-sphere model (red stars) and for atom-centered model (black circles). Lebedev quadratures with $n = 1$, $N = 6$ and $t = 11$, $T = 194$ were used for the charged S_a ($a = 2$ au) and probe S_R ($R = 8$ au) spheres, respectively. 73

4.8 The orthonormal bases of the right singular vectors: basis of spherical harmonics $\widetilde{\mathbf{Y}}_{S_a}$ (A), basis from the numerical SVD of the LS matrix in two-sphere PC model (B), and atom-centered model (C). 74

4.9 Hessian eigenbases along with corresponding singular values in the constrained free case and with the total charge constraint by the elimination of one of the atoms. 79

5.1 Continuous charged sphere model centered at the origin of the molecular multipole expansion (A) and its point-charge (B) representation. 85

5.2 Effect of the sphere radius a on the RMSD between the multipolar expansion (eq. 5.2) and point-charge potential (eq. 5.11) in water. 89

5.3 Electrostatic potential of water in the plane of the molecule within single-ste Lebedev model with $n = 1, 2$ and $a = 0.5,\ 3.5$ (two right columns) as compared with the true multipole moment potentials (left column). 90

5.4 Electrostatic potential over the isosurface of constant charge density (0.002 au) of CH_3SH calculated with single-site Lebedev models ($a = 0.5$, $n = 1,\ 2,\ 3$). 91

5.5 Electrostatic potentials calculated using single-site Lebedev model ($a = 0.5$, $n = 1,\ 2,\ 3,\ 4$) of water and CH_3SH in the plane of the molecules. Contour levels: -100, -50, -25, -12, 0, 12, 25, 50, 100 kcal/mol. 91

5.6 Electrostatic potential over the isosurface of constant charge density (0.002 au) calculated using three charge models: quantum mechanical (left), multi-site Lebedev model (middle, $a = 0.5$, $n = 2$) and atom-centered point charge model (right). 92

5.7 Point charge representation of single-site and multi-site Lebedev models of CH_3SH. 92

5.8 Convergence of the CH_3SH electrostatic potential to the QM MEP within single-site (circles) and multi-site Lebedev (squares) models. RMSD and Pearson R^2 correlation coefficient are used to quantify the convergence. 93

6.1 QM vs. optimized FF potential energy scans along bonds in cis- (top) and trans-MeSNO (bottom). 98

6.2 QM vs. optimized FF potential energy scans along angles in cis- (top) and trans-MeSNO (bottom). 98

6.3 Representation of the the interaction energies between MeSNO (cis-MeSNO on the left and trans-MeSNO on the right) and ammonium ion NH_4^+. Position of each colored sphere corresponds to the position of the nitrogen in NH_4^+. The color of the sphere represents the strengths of the interaction: red for repulsion and blue for attraction. 99

6.4 Correlation between interaction energies calculated using PBE0/def2-TZVPPD (QM energies) and optimized force field (FF energies). . 100

6.5 PES scans between MeSNO (cis on the left and trans on the right) and acetate anion $MeCOO^-$ (top) and the intrinsic reaction coordinate (IRC) profile along the minimum energy path between three hydrogen bonded complexes of MeSNO (cis on the left and trans on the right) and ammonium ion NH_4^+ (bottom). IRC path is calculated at PBE0/def2-TZVPPD at the geometries calculated using PBE0/def2-SV(P)+d level of theory. 101

7.1 Structure of glutathione (GSH) and S-Nitrosoglutathione (GSNO). 105

7.2 UV-vis spectra GSNO at different times after mixing. 106

7.3 ^{1}H-^{1}H TOCSY spectra of 10mM GSH and 1 mM GSNO, both at pH 7.0 and room temperature. 107

A-A1 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 124

A-E1 Convergence of the RMSD between ammonia QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). . . . 125

A-A2 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 126

A-E2 Convergence of the RMSD between bromomethane QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). . 127

A-A3 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 128

A-E3 Convergence of the RMSD between chloromethane QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). . 129

A-A4 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 130

A-E4 Convergence of the RMSD between cis-mesno QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). . . . 131

A-A5 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 133

A-E5 Convergence of the RMSD between fluoromethane QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). . 134

A-A6 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 135

A-E6 Convergence of the RMSD between formamide QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). . . . 136

A-A7 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 138

A-E7 Convergence of the RMSD between furan QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). . . . 139

A-A8 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 140

A-E8 Convergence of the RMSD between imidazol QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). . . . 141

A-A9 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 143

A-E9 Convergence of the RMSD between methanesulfonamide QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). 144

A-A10 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 146

A-E10 Convergence of the RMSD between methanesulfonic acid QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). 147

A-A11 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 149

A-E11 Convergence of the RMSD between methanethiol QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). . 150

A-A12 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 151

A-E12 Convergence of the RMSD between methanol QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). . . . 152

A-A13 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 153

A-E13 Convergence of the RMSD between tetrazole QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). . . . 154

A-A14 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 156

A-E14 Convergence of the RMSD between thiazole QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). . . . 157

A-A15 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 159

A-E15 Convergence of the RMSD between trans-mesno QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). . 160

A-A16 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 162

A-E16 Convergence of the RMSD between uracil QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). . . . 163

A-A17 Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model. 165

A-E17 Convergence of the RMSD between water QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au). . . . 166

Chapter 1

Introduction

1.1 Force Field Models of Intermolecular Interactions

The importance of understanding and proper description of the forces between molecules cannot be overestimated.[1,2] The very existence of a liquid, solid or even biological systems is a direct consequence of these intermolecular interactions. Depending on the physical phenomenon behind a particular interaction, all intermolecular forces can be split into two major classes: short-range (usually repulsive) and long-range (usually attractive) forces.

At the short range, the molecular wavefunctions overlap significantly and the energy increases exponentially.[1,2] The repulsive behavior of the energy is determined by the antisymmetry of the wave-function with respect to the exchange of electrons and is called exchange interaction.

The long range interactions are usually classified into electrostatic, induction and dispersion.[1,2] Despite their seeming difference, all of these interactions follow the Coulomb's law of electrostatic interaction, either between static charge distributions of the molecules (that can be either attractive or repulsive) or between perturbed distributions of the molecular charge densities (that are strictly attractive). Induction effects arise from the distortion of one molecule's charge density in the electric field of the second, while dispersion interactions are purely quantum-mechanical in their origin and arise from the correlated motion of electrons in two molecules that gives rise to instantaneous multipole moments interacting with each other.

As a result of the short-range repulsion and long-range attraction, a typical interaction energy curve has a single minimum (Figure 1.1). At the distance R_0 the energy has its minimal value $-\epsilon$ and all forces are compensated in such way that the system is at its equilibrium. Smaller separation distance results in the exponential growth in energy with a negative slope (i.e. positive force or repulsion), while at larger distance the energy slowly increases with a positive slope (i.e. negative force or attraction). Due to the Coulombic nature of the interactions at the long range, the energy increases proportionally to the power function of the inverse distance R, i.e. $-1/R^n$, approaching zero at the infinity.

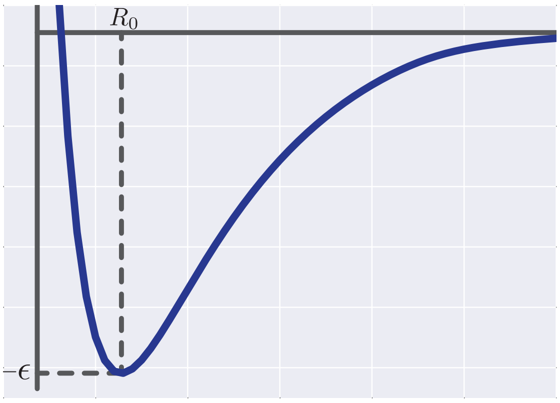


Figure 1.1: Intermolecular energy as a function of the separation distance

Using the simplified mathematical models of these interactions, the thermodynamic and kinetic properties of gases, liquids and even complex biological systems like proteins, nucleic acids, and lipids can be studied by sampling their conformational space via such simulation techniques as Monte Carlo or molecular dynamics (MD).[3–5] In MD simulations, particles coordinates $\mathbf{r}(t)$ are propagated using Newton's equations of motion, i.e.

$$-\frac{dU}{d\mathbf{r}} = m\frac{d^2\mathbf{r}}{dt^2} \tag{1.1}$$

where the potential energy $U(\mathbf{r})$ of the system is calculated using the mathematical representations of the intra and intermolecular:

$$U = U_{bonds} + U_{angles} + U_{torsions} + U_{Coulomb} + U_{LJ} \tag{1.2}$$

Together these intra and intermolecular potentials are usually called a molecular mechanics force field. The intramolecular portion of the potential energy U includes the Hooke's law to model bond U_{bonds} and angle U_{angles} vibrations, a periodic potential to model the torsional strain $U_{torsions}$, while intermolecular part is usually described by the Coulomb $U_{Coulomb}$ and Lennard-Jones U_{LJ} potentials. The former models the electrostatic interaction between static charge distributions and the latter is combination of short-range repulsion and long-range dispersion attraction. Although the exact form of the potential U depends on the actual implementation, most of the popular force fields, such as CHARMM,[6–8] AMBER,[9–11] GROMOS,[12] and OPLS[13,14] have very similar forms of the potential energy. More sophisticated force field libraries may also include addition terms, e.g. atomic polarizations, cross-terms, etc.

Among all terms in a force field, the Coulomb term of electrostatic interactions is among the most crucial terms for a proper description of proteins, nucleic acids, lipids, and other macromolecules, as well as for their interactions with solvent, ions, and other molecules.[15] Due to the properties of the Coulomb's law at the long range, the electrostatic interaction between static charge densities of two molecules can be accurately described by the interaction between multipole moments of each molecule (e.g. total charge, dipole, quadrupole, etc).[16] Furthermore, various partitioning/distribution schemes allows obtaining a multi-site multipolar expansion centered at each atom in a molecule.[16–19] Application of the atomic multipoles in the force fields to describe molecular electrostatic resulted in several multipolar force fields, such as AMOEBA, SIBFA, NEMO.[20–22]

However, inclusion of several multipole moments (usually up to quadrupole moment) per atom even in the modestly sized biomolecule quickly become the computational bottleneck. Up to this date, only a limited number of small systems have been studied using the multipolar force fields. Besides being computationally demanding, the implementation of multipole-multipole interactions in a simulation is non-trivial.[23–31] Firstly, since all multipolar components are given in a global coordinate system, it is necessary to transform them into a local coordinate system associated with each atom. Secondly, besides a regular force that arises due to the gradient in the energy, torques produced by every multipole need to be added to each atomic force; and finally, in order to use multipolar electrostatics with periodic boundary conditions methods that take into account long-range electrostatics within multipolar formalism (such as particle mesh Ewald) are required. Due to these reasons, only very few simulation packages support multipolar formalism thus prohibiting their widespread usage.

Therefore, the multi-site multipolar expansion expansion is usually truncated at the atomic monopole (charge), leading to much less computationally demanding approximation. In this approximation, the continuous charge density of a molecule is modeled by a set of atom-centered point charges, and electrostatic interaction between two molecules is simply modeled by a pairwise Coulomb's law between atomic charges from each molecule:[32–34]

$$U_{Coulomb} = \sum_{i>j} \frac{q_i q_j}{r_{ij}} \tag{1.3}$$

where q_i is the point charge at atom i and r_{ij} is the distance between atoms i and j. The atom-centered charges provide a clear chemically intuitive interpretation of the electrostatic properties, require a straightforward implementation and thus have been used in such force field libraries as AMBER, CHARMM, GROMOS, OPLS since the introduction of the molecular dynamics simulations.

The common approach to derive point charges in the force field development is to use the least squares fitting to the reference quantum mechanical molecular electrostatic potential Φ_{QM} over the N grid points in the solvent-accessible region of the molecule:[32–34]

$$\chi^2 = \sum_{i}^{N} \left[\Phi_i^{QM} - \sum_{j}^{M} \frac{q_j}{r_{ij}} \right]^2 , \tag{1.4}$$

where M is the number of point charges and r_{ij} is the distance between point charge j and grid point i.

Although robust numerical methods of solving linear least squares problems exist, the charges obtained with this method are very sensitive to even small perturbations in the problem setup.[35–38] These numerical instabilities are usually related to a large variation of the point charge values for atoms in the interior of the molecule. These buried atom charges (usually methyl and methylene carbons) can be dramatically changed due to trivial changes in the the reference grid sampling, spatial orientation of the molecule, and/or have inconsistent values across very similar molecules or even conformers of the same molecule.[39,40] In order to suppress these large variation of the charge value on the buried atoms, a restraining function is usually added to the least squares sum that prevents convergence to the large charge values by keeping them close to a predefined value, e.g. zero, or some other chemically reasonable value.[23,36,41–51] This, however, can negatively affect the molecular dipole moment and the overall quality of the fit.[38,52]

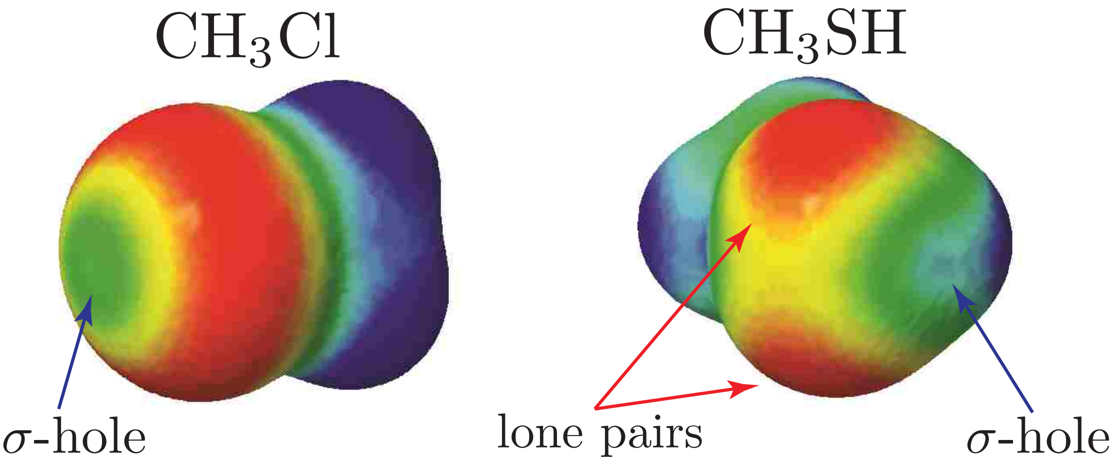


Figure 1.2: Effect of σ-hole on the molecular electrostatic potential of chloromethane CH_3Cl and methanethiol CH_3SH. Formation of the covalent σ bond leaves a region of diminished negative charge on its non-involved side along the extension of the bond. Calculations were performed at MP2/aug-cc-pVTZ level of theory; the charge density isosurface was plotted at 0.002 au.

Besides the numerical instabilities of the charge fitting problem, there are also issues with the point charge approximation itself: the isotropic nature of the single point-charge potential cannot describe anisotropic character of the true molecular electrostatic potential around each atom in a molecule. As a result, local atomic properties such as donor/acceptor features due to the lone pairs, σ-holes (Figure 1.2) and π-electrons are usually missed by the atom-centered point charge approximation.[53,54] Therefore, intermolecular interactions that involve a Lewis base B and hydrogen atom in a molecule HX (i.e. hydrogen bond $\mathrm{B}\cdots\mathrm{HX}$) or Lewis base B and a halogen/chalcogen atom Y in molecule YX (i.e. halogen/chalcogen bond $\mathrm{B}\cdots\mathrm{YX}$)[55–59] are significantly underestimated or even entirely missed by the atom-centered point charge model.

Nevertheless, despite the obvious lack of accuracy and numerical difficulties in the point charge derivations, the simplicity of the point charges drives scientific community to go beyond the atom-centered paradigm and use point charges to reproduce effects of higher (above monopole) atomic multipoles by proper placement of point charges out of the atomic center.[60–64]

However, none of the existing methods offers a systematic approach in optimizing the proper position of the off-center point charges as well as in the derivation of their values. Moreover, due to the numerical instabilities associated with the buried atom charges it is not clear how to alleviate these instabilities in the case of off-center charges, as their inclusion into the model produces even more buried centers.

Besides the electrostatic term, another crucial ingredient of any force field is the part that models the short-range repulsion and long-rand dispersion forces. These interactions are often combined in a single Lennard-Jones potential:

$$U_{LJ} = \sum_{i>j} \varepsilon_{ij} \left[\left(\frac{r_{ij}^*}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^*}{r_{ij}} \right)^{6} \right] \tag{1.5}$$

where ε_{ij} if the well depth and r_{ij}^* is the equilibrium van der Waals distance between atoms i and j.

The parameters ε_{ij} and r_{ij}^*, often referred to as van der Waals parameters, are usually obtained by fitting to reproduce experimental liquid properties, such as density and enthalpy of vaporization.[9,14] Then, assuming the transferability of the atomic properties these parameters are applied to the molecular systems other than liquids. Together with the atom-centered point charges fitted to reproduce electrostatic potential, vdW parameters constitute the non-bonded part of the molecular mechanics force field. This paradigm have been historically adopted by many simulation packages and the force field libraries and have been used throughout the scientific community. In this approach, the only source to verify the accuracy of the force field potentials and their parameters is to compare a simulation with the experimental data. This is based on the notion that the simulation can result in the correct macroscopic observables only if the microscopic parameters of the system are correct.

Unfortunately, in many cases and especially in the case of complex biomolecular systems, the amount of the high-quality spectroscopic and thermodynamic data can be limited to develop a robust methodology that could validate/adjust force field parameters. Even in the case when experimental data is available, such methodology would reflect the accuracy of the underlying force field only implicitly.

Only recently, the dramatic increase in computational power and development of accurate quantum chemistry methods allowed obtaining, with a relatively modest computational requirements, large amounts of high quality data that are often inaccessible to the experiment. For example, the potential energy surface of a protein residue interaction with its local environment can be now obtained using the density functional theory or even *ab initio* methods. This information is an important source of the reference data to fit the force field parameters and ensure a correct description of the microscopic properties.[28]

In order to take advantage of this reference data, a major reconsideration of the entire workflow in the force field parametrization process is required. For example, instead of a separate optimization of several force field terms, different non-bonded parameters (point charges, Lennard-Jones parameters, and, in the case of polarizable force fields, atomic polarizabilities) can be fitted simultaneously to extensive training sets of interaction energies. Then, even in the case of fixed point charges, fitting to the energy of interaction can implicitly include the polarization effects, thus improving overall quality of the force field.[65–71] However, such simultaneous force field fitting represents a technically challenging multi-objective optimization of the parameters of different physical nature and mathematical form. This is a complex minimization problem that requires a cautious approach as the search space is nonlinear and ill-defined. Even in a much simpler case of the linear least squares fitting of point charges to the reference electrostatic potential, the solution to the problem can be numerically unstable. Therefore, a simultaneous optimization of point charges along with Lennard-Jones parameters against a diverse training set would be even more challenging.

Although available force field libraries contain parameters for all standard amino acids, accurate force field parameters for non-standard residues might be missed. A representative example of the non-standard residue for which there is no accurate force field is S-nitrosocysteine, the most common biological S-nitrosothiol.

1.2 S-Nitrosothiols and Their Biological Role

Protein S-nitrosation—a covalent post-translational modification of the cysteine amino acid residue (Figure 1.3)—is involved in a signaling pathway of nitric oxide, an important cellular signaling molecule that plays role in many physiological and pathological processes.[72–74]

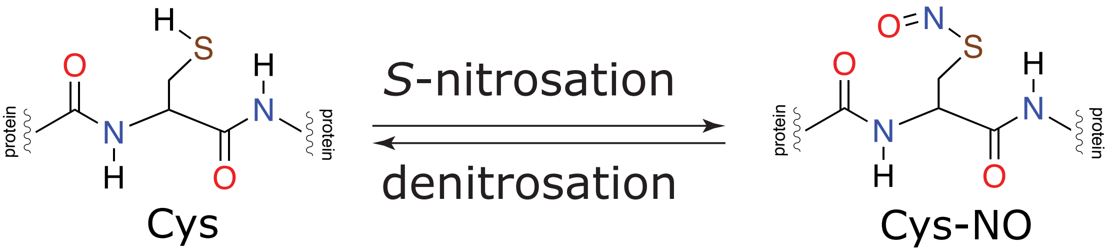


Figure 1.3: S-Nitrosation of a cysteine in a peptide/protein.

Cysteine-containing proteins as well as low molecular weight peptides, like glutathione (GSH) can be S-nitrosated and form S-nitrosated proteins (SNO proteins) and S-nitrosoglutathione (GSNO), respectively.[75,76] More than 1000 proteins have been already identified to undergo S-nitrosation in vivo across a wide variety of living organism.[77,78] S-nitrosation has been implicated in regulating enzymatic activity, protein-protein interaction, protein stability, and such signaling pathways as cell apoptosis and blood flow vasodilation.[79–81]

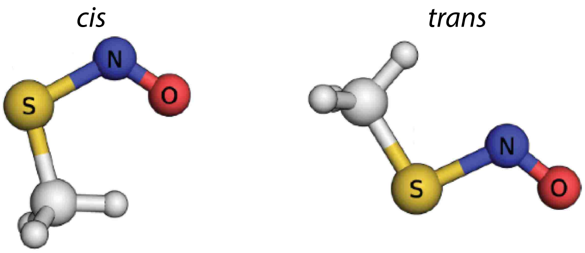


Figure 1.4: Structures of two MeSNO isomers: cis and trans

S-nitrosothiols exist in two isomeric forms, cis and trans (Figure 1.4), that are separated by an appreciable barrier ∼ 10 kcal/mol around the S-N bond rotation, suggesting the presence of a strong double bond character.[82–84] Nevertheless, kinetic experiments on RSNO decomposition have shown that the stability of RSNOs drastically depends on their substituents, pH, presence of metal and thiolate (RS^-) ions, all of which imply weak S-N bond.[72,85,86]

These unusual properties in RSNO can be rationalized by a combination of covalent (**S**), zwitterionic (**D**) and ion pair (**I**) resonance structures (Figure 1.5).[87–90] Coexistence of structures **D** and **I** with opposite S-N bond character and formal charge on sulfur (i.e. antagonistic structures) explains the planar geometry of -SNO group and tendency of RSNO towards decomposition. This

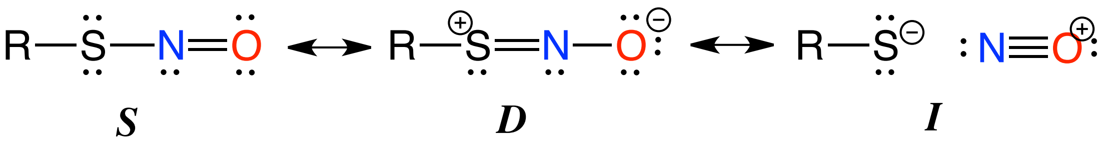


Figure 1.5: Resonance representation of RSNO electronic structure as a combination of covalent (**S**), zwitterionic (**D**) and ion pair (**I**) resonance structures.

also explains the stability of the RSNO complexes with charged species: for example, S-coordination to positive ion (e.g. Cu^+) favors structure **I** and thus leads to the destabilization of the S-nitrosothiol.[91,92] At the same time, N-coordination to positive ion (e.g. ${Ir_3}^+$) favors structure **D** and leads to the stabilization.[93]

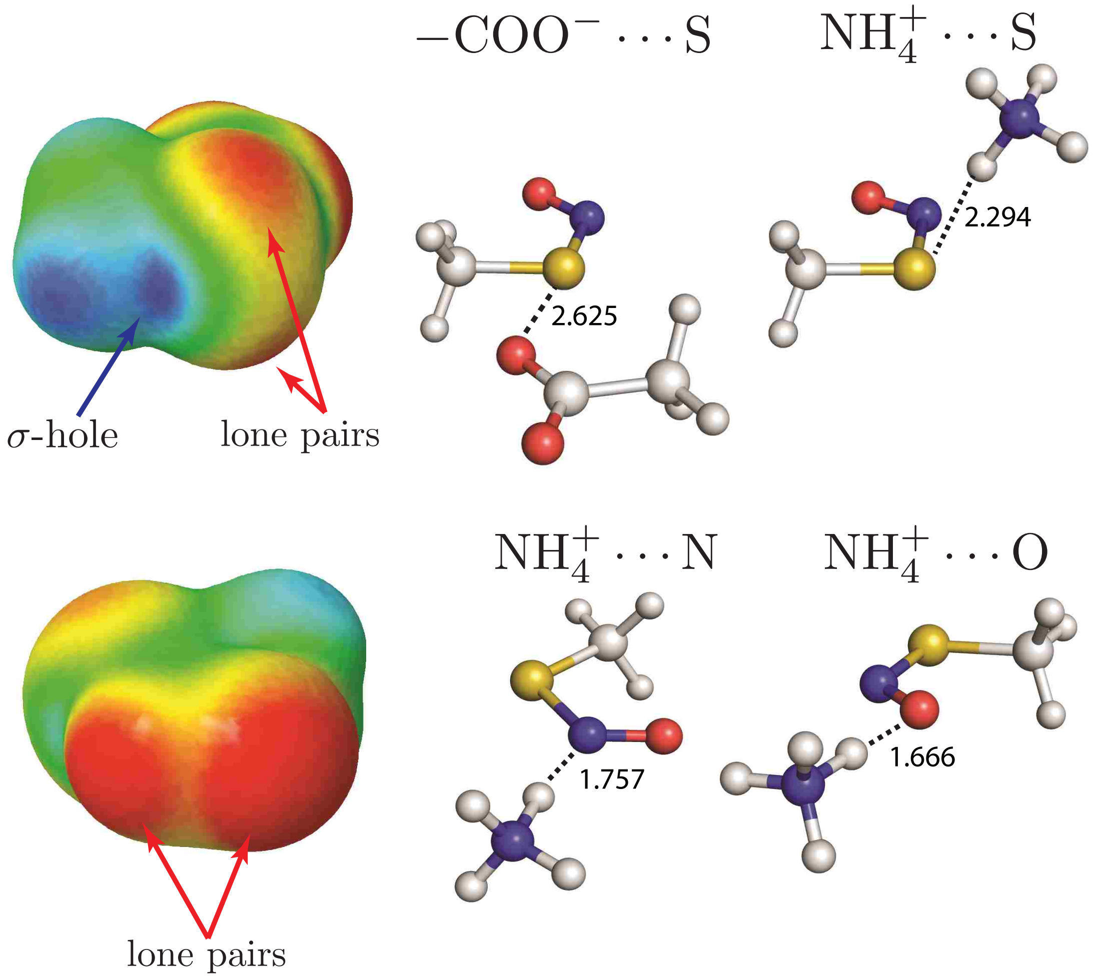


Figure 1.6: Molecular electrostatic potential of MeSNO around sulfur (top left) and nitrogen and oxygen (bottom left) atoms and the chalcogen- and hydrogen-bonded complexes with sulfur (top right) and hydrogen-bonded complexes with nitrogen and oxygen (bottom right). Calculations were performed at PBE0/def2-TZVPPD level of theory; the charge density isosurface was plotted at 0.002 au.

The unique electronic structure of RSNOs may also suggest how the –SNO group interacts with charged and polar environment of CysNO in a S-nitrosated protein. For example, it was shown computationally that protonated basic residues (Lys, Arg, His) form hydrogen-bonded complexes due to the presence of the lone pairs at each atom of the –SNO group (Figure 1.6).[89] Among three possible complexes, S-coordinated complex is the weakest, while N- and O-coordinated complex are similar in stability. At the same time, presence of the positively charged area along the extension of the S-N bond (i.e. σ-hole) stabilizes the coordination of the negatively charged residue (deprotonated glutamic or aspartic acid) at the sulfur of the –SNO group, resulting in the formation of the chalcogen-bonded complex (Figure 1.6).

Depending on the trade-off between the energy released upon the coordination and the strain caused by the deformation of the protein scaffold, formation of these complexes inside a real protein can induce conformational changes leading to the change in the protein activity. For example, Wang and coworkers proposed a possible mechanism how CysNO induces conformational change in apolipoprotein E3 (ApoE3).[94] They showed that CysNO112 could form hydrogen bonds and/or ion pairings with the charged Arg61 and Glu109 residues. Formation of these complexes can potentially kink the helix where Cys112 is attached to, inducing a large conformational change, leading to the loss of ApoE3 binding to the low-density lipoprotein (LDL) receptors. Decrease in the binding to LDL receptors is known to play a role in the development of Alzheimers disease.

While the –SNO group can induce the conformational change, the protein environment around the –SNO group can control its reactivity. It was shown computationally that the reactivity of the –SNO group may be changed when the charged residues coordinate sulfur, nitrogen or oxygen. For example, when $MeNH_3^+$ coordinates oxygen or nitrogen, the S-N bond shortens and RSNO is stabilized (contribution of resonance **D** structure increases), while coordination to sulfur atom weakens S-N bond and RSNO is destabilized (contribution of **I** structure increases). As a result, the tight balance between structures **D** and **I** controls the reactivity of RSNO and can promote either the reaction of trans-S-nitrosation—NO^+ transfer from one thiol to another—or S-thiolation—formation of the disulfide and HNO (Figure 1.7).[95–97] The importance of the precise control over the RSNO reactivity cannot be overestimated as the trans-S-nitrosation is a major pathway of selective protein S-nitrosation in vivo, while S-thiolation may lead to the S-glutathionylation, which is another post-translational modification of proteins, and production of a signaling agent nitroxyl HNO.[98–100]

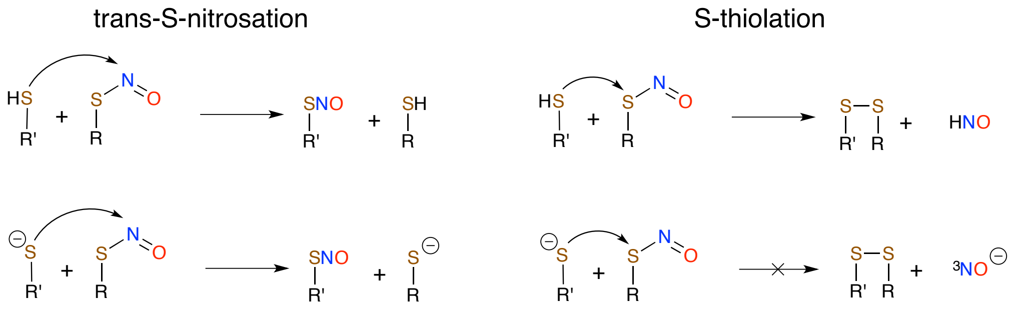


Figure 1.7: Two pathways of the reaction between S-nitrosothiol (RSNO) and thiol in neutral (RSH) and anionic (RS^-) states.

In the cases when CysNO is positioned between two oppositely charged residues, the –SNO group experiences appreciable external electric field. This field can be strong enough to induce the change in the electronic structure of the –SNO group and thus its reactivity.[88] Depending on the direction and strength of the external electric field either **D** structure is promoted, resulting in the stabilization of RSNO, or structure **I** is promoted, resulting in the destabilization of RSNO. This, in turn, determines the barrier for the possible reactions of RSNO with thiols: trans-S-nitrosation or S-thiolation. By changing the direction and strengths of the external field it was computationally shown that one of the two reaction barriers can proceed almost barrierlessly, while the other reaction became completely inhibited. This clearly demonstrates the possible catalytic effect on the reactivity of CysNO produced by the local charged environment.

To investigate if specific interactions of the –SNO group with basic and acidic amino acid residues are involved in biological processes, it is necessary to obtain 3-dimensional structures of the S-nitrosated proteins. Determination of the structures using high-resolution X-ray crystallography is challenging due to the

–SNO group instability and only very limited number of the crystal structures of SNO-proteins have been reported.[101–103] Nuclear magnetic resonance (NMR) spectroscopy, on the other hand, provides a convenient way of probing local environment of the species under the interest without disturbing it. However, since NMR spectroscopy does not explicitly provide 3D structure of a protein, its analysis is often complemented by the computer simulation, including molecular dynamics (MD) simulations. In this cases accurate force field parameters for the –SNO group are required. Unfortunately, only the most basic force field description of the CysNO residue is available, which is unable to describe the hydrogen- and chalcogen-bond interactions that are specific to the -SNO group.[104,105] The complex electronic structure of CysNO and its possible specific interactions with charged and polar residues require much more accurate description of its intermolecular interactions, which are mostly of electrostatic nature and the accurate description of the electrostatic potential is available using the multipolar force fields. However, due to the significant resources required to simulate even a modestly sized protein and also a lack of mature molecular dynamics packages that support multipolar force fields, a direct application of multipolar formalism to the –SNO group is not feasible at the moment.

Moreover, although an interaction of a Lewis base with a σ-hole is electrostatically driven and results in the formation of the halogen or chalcogen bond, the spatial orientation of the interacting species is also influenced by the induction, dispersion and exchange-repulsion terms.[106] Thus, the existing methodologies in the force field development *a priory* can not provide a set of parameters that would accurately reproduce these interactions: atom-centered point charge approximation fails to account for the anisotropy due to lone-pairs and σ-holes, while separate parametrization of Lennard-Jones parameters cannot fully account for the exchange-repulsion interactions that are specific to the –SNO group.

1.3 Objectives

Driven by the need to develop a force field description that is capable to describe the hydrogen- and chalcogen-bonded complexes in RSNOs, while being limited to the point charge approximation this works aims (1) to develop the point charge model with the accuracy comparable with the multipolar force fields and (2) to apply this approach to a model RSNO molecule. Due to the complexity of the problem, following concerns have to be taken into the account during the development of this model:

- Since inclusion of the off-center point charges in a force field parametrization implies a non-linear optimization, properties of several optimization algorithms have to be investigated. These algorithms may include the traditional gradient methods as well as stochastic algorithms such as evolutionary methods.

- In order to extend the atom-centered approximation into the charge model with any number of off-center charges, the origin of the numerical instabilities associated with the buried atoms has to be investigated.

- A general solution to the charge fitting problem implies its compatibility with a wide range of molecules. Thus, the model should be easily applied to any molecule.

- In the specific case of -SNO group, the developed charge model should be parametrized together with other force field terms, such as Lennard-Jones, in order to reproduce the energy of interaction between the group and other charged residues.

Chapter 2

Overview of Theoretical and Numerical Methods

In this Chapter we briefly overview the theoretical and numerical methods used in this work, which includes a concise introduction into the theory of electrostatic potential and its spherical harmonics expansion,[1,107] formal algebraic definitions of the numerical techniques (eigenvalue and singular value decompositions of a matrix, least squares approximation, and matrix ill-conditioning)[108] and a brief introduction into the minimization algorithms.

2.0.1 Spherical Harmonics

Spherical harmonics $Y_{lm}(\theta, \varphi)$ are functions defined on a unit sphere that found a widespread application in many fields of science, including electromagnetism, astronomy, fluid dynamics, etc. Spherical harmonics define the angular part of the solution $f(r, \theta, \varphi)$ to the Laplace equation, a second-order partial differential equation,

$$\nabla^2 f(r, \theta, \varphi) = 0,$$

$$f(r, \theta, \varphi) = R(r) Y_{lm}(\theta, \varphi). \tag{2.1}$$

where the angular part depends on azimuth angle θ and polar angle φ, the integer indices l and m ($m \leq |l|$) are referred to as the degree and order of spherical harmonic Y_{lm}, respectively and Laplace operator ∇^2 in spherical coordinates is defined as

$$\nabla^2 f = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial f}{\partial r}\right) - \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial}{\partial \theta}\right) - \frac{1}{r^2 \sin^2\theta}\frac{\partial^2}{\partial \varphi^2} \tag{2.2}$$

The angular part of the Laplace operator is also known as the angular momentum operator $\hat{\mathbf{L}}^2$. Then, spherical harmonics Y_{lm} are its eigenfunctions, i.e.

$$\hat{\mathbf{L}}^2 Y_{lm} = \hbar l(l+1) Y_{lm}, \tag{2.3}$$

$$\hat{\mathbf{L}}^2 = -\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) - \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}, \tag{2.4}$$

and thus form a complete orthogonal basis set with the orthogonality relation:

$$\int_S Y_{lm}(\theta,\varphi) Y^*_{lm}(\theta,\varphi) d\Omega = \delta_{ll'}\delta_{mm'}, \tag{2.5}$$

where $d\Omega = \sin\theta d\theta d\varphi$ is the differential solid angle in spherical coordinates and asterisk corresponds to the complex conjugation such that the phase factor $(-1)^m$ is maintained according to $Y^*_{lm} = (-1)^m Y_{l,-m}$.[1] The completeness property implies that any function of angles θ and φ can be represented as a linear combination of spherical harmonics:

$$f(\theta,\varphi) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} c_{lm} Y_{lm}(\theta,\varphi), \tag{2.6}$$

where c_{lm} are Fourier coefficients:

$$c_{lm} = \int_S f(\theta,\varphi) Y_{lm}(\theta,\varphi) d\Omega. \tag{2.7}$$

The properties of spherical harmonics Y_{lm} are largely defined by the associated Legendre polynomials P_{lm} as these polynomials directly appear in the analytical expression for spherical harmonics:

$$Y_{lm}(\theta,\varphi) = (-1)^m \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}} P_{lm}(\cos\theta) e^{im\varphi} \tag{2.8}$$

---

[1] In some problems renormalized spherical harmonics C_{lm} are used such that $\int_S C_{lm}(\theta,\varphi) C^*_{lm}(\theta,\varphi) d\Omega = \frac{4\pi}{2l+1}\delta_{ll'}\delta_{mm'}$

Associated Legendre polynomials are solutions to the general Legendre equation

$$\left(1-x^2\right)\frac{d^2P_{lm}}{dx^2}-2x\frac{dP_{lm}}{dx}+\left[l(l+1)-\frac{m^2}{1-x^2}\right]P_{lm}=0 \tag{2.9}$$

and are defined as:

$$P_{lm}(x)=\frac{1}{2^l l!}(1-x^2)^{m/2}\frac{d^{l+m}}{dx^{l+m}}(x^2-1)^l \tag{2.10}$$

where the integer indices l and m $(m \leq |l|)$ are referred to as the degree and order of the associated Legendre polynomials, respectively. When $m = 0$ these function correspond to Legendre polynomials $P_l(x)$ that can be defined as the coefficients in a Taylor series expansion of the generating function:

$$\frac{1}{\sqrt{1-2xt+t^2}}=\sum_{l=0}^{\infty}P_l(x)t^l, \tag{2.11}$$

where the function on the left side of the eq. 2.11 is the generating function of Legendre polynomials. This expansion plays a critical role in the multipolar expansion of the molecular electrostatic potential.

2.0.2 Molecular Electrostatic Potential

In 1785, the French physicist Charles-Augustin de Coulomb in a series of experiments showed that the magnitude of the electrostatic force between two point charges is directly proportional to the product of charge values and inversely proportional to the square of the distance between them. This force is directed along the straight line between charges, is attractive when charge are of opposite sign and repulsive if charges have the same sign:

$$\mathbf{F}=\frac{q_1q_2}{r^2}\hat{\mathbf{r}} \tag{2.12}$$

The electrostatic force is a result of interaction between two charges and it is useful to introduce a concept of electric field that is defined by one of the two charged species:

$$\mathbf{F} = q_1\mathbf{E}, \tag{2.13}$$

where electric field $\mathbf{E}$ is measured at the position of the charge q_1. In case of a continuous charge distribution the electric field $\mathbf{E}$ can be calculated by integrating the charge density $\rho(\mathbf{r})$:

$$\mathbf{E}(\mathbf{r}) = \int \rho(\mathbf{r}') \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} d\mathbf{r}' \tag{2.14}$$

Electric field is a vector field and requires three components in order to be fully defined. Since the vector factor in eq. 2.14 is the negative gradient of the inverse distance:

$$\frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} = -\boldsymbol{\nabla} \left(\frac{1}{|\mathbf{r} - \mathbf{r}'|} \right) \tag{2.15}$$

the vector field $\mathbf{E}$ can be uniquely defined by a scalar potential Φ with help of the gradient operation:

$$\mathbf{E} = -\boldsymbol{\nabla}\Phi. \tag{2.16}$$

The electrostatic potential $\Phi(\mathbf{r})$ is uniquely defined by the charge density $\rho(\mathbf{r})$ and has a physical interpretation of energy required to bring the unitary charge from infinity to the point $\mathbf{r}$ in the electrostatic field of the charge density $\rho(\mathbf{r})$.

Then, given the charge density of a molecule, the molecular electrostatic potential (MEP) can be computed as:

$$\Phi(\mathbf{r}) = \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3r', \tag{2.17}$$

A direct calculation of the integral eq. 5.1 is often impractical for most numerical applications, so different approximations are usually used instead.

2.0.3 Multipolar Expansion of Electrostatic Potential

Given the charge density of a molecule, its electrostatic potential $\Phi(\mathbf{r})$ is defined by the Coulomb law as

$$\Phi(\mathbf{r}) = \int_V \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' \tag{2.18}$$

where the source coordinate $\mathbf{r}' = (r', \theta', \varphi')$ is integrated over the volume V occupied by the molecular charge density $\rho(\mathbf{r}')$ and the observation vector $\mathbf{r} = (r, \theta, \varphi)$ is the the point where the electrostatic potential $\Phi(\mathbf{r})$ is calculated.

Let γ be an angle between vectors $\mathbf{r}$ and $\mathbf{r}'$, then according to the cosine theorem, the difference $|\mathbf{r} - \mathbf{r}'|$ can be expressed as

$$|\mathbf{r} - \mathbf{r}'| = \sqrt{r^2 + r'^2 - 2rr' \cos\gamma}. \tag{2.19}$$

Then, from the definition of the generating function of Legendre polynomials (eq. 2.11) it immediately follows that in the case when $r' < r$, the inverse distance can be expressed as:

$$\frac{1}{|\mathbf{r} - \mathbf{r}'|} = \frac{1}{r} \frac{1}{\sqrt{1 - 2r'/r \cos\gamma + (r'/r)^2}} = \frac{1}{r} \sum_{l=0}^{\infty} \left(\frac{r'}{r} \right)^l P_l(\cos\gamma) \tag{2.20}$$

In the case of $r' > r$, the r and r' can be interchanged in eq. 4.10. Then for convenience, the notation where $r_<$ is the smaller among r and r' and $r_>$ is the larger of the two is usually used. Using this notation,

$$\frac{1}{|\mathbf{r} - \mathbf{r}'|} = \sum_{l=0}^{\infty} \frac{r_<^l}{r_>^{l+1}} P_l(\cos\gamma). \tag{2.21}$$

With the help of the addition theorem that expands $P_l(\cos\gamma)$ into spherical harmonics:

$$P_l(\cos\gamma) = \frac{4\pi}{2l+1} \sum_{m=-l}^{l} Y_{lm}(\theta, \varphi) Y_{lm}(\theta', \varphi') \tag{2.22}$$

the inverse distance between two vectors can be also expanded so that the source $\mathbf{r}'$ and observation $\mathbf{r}$ vectors are separated:

$$\frac{1}{|\mathbf{r}-\mathbf{r}'|} = \sum_{l=0}^{\infty} \frac{4\pi}{2l+1} \frac{r_<^l}{r_>^{l+1}} \sum_{m=-l}^{l} Y_{lm}(\theta,\varphi) Y_{lm}(\theta',\varphi'). \tag{2.23}$$

This expression leads to the multipolar expansion of the molecular electrostatic potential:

$$\Phi(\mathbf{r}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \sqrt{\frac{4\pi}{2l+1}} r^{-l-1} Q_{lm} Y_{lm}(\theta,\varphi), \tag{2.24}$$

where Q_{lm} are multipole moments of a molecule:

$$Q_{lm} = \sqrt{\frac{4\pi}{2l+1}} \int_V r^l \rho(\mathbf{r}) Y_{lm}(\theta,\varphi) d\mathbf{r}. \tag{2.25}$$

For example, Q_{00} is the monopole, i.e. the total charge; Q_{1m} with $m = -1, 0, 1$ are three components of the dipole moment; Q_{2m} with $m = -2, -1, 0, 1, 2$ are five components of the quadrupole moment, etc.

2.1 Methods of Matrix Decomposition

In many numerical problems it is useful to decompose a matrix into a product of two or more matrices. Depending on the particular class of problems different decomposition techniques exist. For example, eigendecomposition can be useful when transformation to the diagonal form is required. Another useful technique is singular value decomposition (SVD) which can be applied to find solution of the least squares problem. In this section, we briefly overview the eigendecomposition, least squares approximation, and singular value decomposition.[108]

2.1.1 Eigendecomposition

Let U be n-dimensional inner product space, i.e. a space where the length of a vector is defined, and linear transformation $\tau : U \rightarrow U$ map any vector in U to another vector in the same space U. Then, a scalar λ is an eigenvalue of τ if there exist an eigenvector $u \in U$ associated with λ for which

$$\tau u = \lambda u. \tag{2.26}$$

Equivalently, in the matrix form $\mathbf{A}$ of the linear operator τ, λ is an eigenvalue of the matrix $\mathbf{A}$ if there exist an eigenvector $\mathbf{u}$ associated with λ for which

$$\mathbf{Au} = \lambda \mathbf{u} \tag{2.27}$$

The set of n eigenvectors forms an eigenbasis of orthogonal vectors $\{\mathbf{u}_i\}_i^n$ such that:

$$\mathbf{u}_i \cdot \mathbf{u}_j = \delta_{ij}, \tag{2.28}$$

where $\mathbf{u}_i \cdot \mathbf{u}_j$ is the dot product on space U, δ_{ij} is Kronecker symbol.

Given the set of n eigenvectors $\mathbf{u}_i$ with corresponding eigenvalues λ_i, matrix $\mathbf{A}$ can be factorized into:

$$\mathbf{A} = \mathbf{U \Lambda U}^*, \tag{2.29}$$

where $\mathbf{\Lambda}$ is diagonal matrix with the eigenvalues λ_i on the diagonal, $\mathbf{U}$ is unitary matrix with the columns consisting from eigenvectors $\mathbf{u}_i$, and $\mathbf{U}^*$ is the conjugate transpose of $\mathbf{U}$.

2.1.2 Singular Value Decomposition

Let U be n-dimensional inner product space, V be m-dimensional inner product space and linear operator τ be such that $\tau : U \rightarrow V$. Then, there exist orthonormal bases $\{u_i\}_{i=1}^n$ in U and $\{v_j\}_{j=1}^m$ in V such that:

$$\tau u_i = \begin{cases} s_i v_i, & i \leq r \\ 0, & i > r \end{cases} \tag{2.30}$$

and

$$\tau^* v_j = \begin{cases} s_j u_j, & j \leq r \\ 0, & j > r \end{cases} \tag{2.31}$$

where r is the rank of operator τ.

From eqs. 2.30-2.31 it immediately follows that $\{u_i\}_{i=1}^n$ and $\{v_i\}_{i=1}^m$ are eigenvectors of $\tau^*\tau$ and $\tau\tau^*$, respectively:

$$\tau^*\tau u_i = s_i^2 u_i, \tag{2.32}$$

$$\tau\tau^* v_j = s_j^2 v_j, \tag{2.33}$$

and s_i^2 are their eigenvalues, which are called singular values of operator τ. The vectors $\{u_i\}_{i=1}^n$ are called right singular vectors and $\{v_j\}_{j=1}^m$ are called left singular vectors of operator τ.

The matrix representation of the operator τ leads to the widely used singular value decomposition (SVD) of a matrix. Let $\mathbf{A}$ be $m \times n$ matrix that represents operator τ from eqs. 2.30-2.31. Then, changing the orthonormal bases from U to V gives:

$$\mathbf{A} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^* \tag{2.34}$$

where $\mathbf{U}$ is unitary matrix with the columns consisting from right singular vectors $\mathbf{u}_i$ and $\mathbf{U}^*$ is its conjugate transpose, matrix $\mathbf{V}$ is unitary matrix with the columns consisting from left singular vectors $\mathbf{v}_i$, and $\mathbf{\Sigma}$ is a diagonal matrix with singular values s_i on the diagonal:

$$\mathbf{U} = (\mathbf{u}_1 \ \mathbf{u}_2 \ \ldots \ \mathbf{u}_m) \tag{2.35}$$

$$\mathbf{V} = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n) \tag{2.36}$$

$$\mathbf{\Sigma} = \mathrm{diag}\,(s_1, s_2, \dots, s_r, 0, \dots, 0) \tag{2.37}$$

where r is the rank of matrix $\mathbf{A}$. As it is clear from eqs. 2.32 and 2.33, right and left singular vectors are also eigenevectors of $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$, respectively, while s_i^2 are their eigenvalues:

$$\mathbf{A}^*\mathbf{A}\mathbf{u}_i = s_i^2\mathbf{u}_i, \tag{2.38}$$

$$\mathbf{A}\mathbf{A}^*\mathbf{v}_j = s_j^2\mathbf{v}_j, \tag{2.39}$$

The expression in eq. 2.34 is called singular value decomposition of matrix $\mathbf{A}$ and has many practical applications, for example the calculation of pseudoinverse of $\mathbf{A}$, least squares fitting of data, analysis of the numerical stability of the solutions to the linear matrix equations, etc.

Pseudoinverse of a matrix

Singular value decomposition leads to a generalized version of the inverse of the rectangular matrix $\mathbf{A}$. Given the linear transformation $\tau : U \to V$, its inverse $\tau^+ : V \to U$ is defined by:

$$\tau^+ v_i = \begin{cases} \frac{1}{s_i} u_i, & i \le r \\ 0, & i > r \end{cases} \tag{2.40}$$

or, equivalently,

$$\tau^+ \tau u_i = \begin{cases} u_i, & i \le r \\ 0, & i > r \end{cases} \tag{2.41}$$

The transformation τ^+ is called the Moore-Penrose generalized inverse or pseudoinverse of τ. If $m = n = r$ the pseudoinverse τ^+ is equivalent to the inverse τ^{-1}, which in case of the matrix corresponds to the inverse of the square matrix $\mathbf{A}$.

In another words, the concept of pseudoinverse generalizes the matrix inverse to any rectangular matrix $\mathbf{A}$. For example, given $m \times n$ matrix $\mathbf{A}$ with singular value decomposition:

$$\mathbf{A} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^*, \tag{2.42}$$

its pseudoinverse is defined as

$$\mathbf{A}^+ = \mathbf{U}\mathbf{\Sigma}^+\mathbf{V}^*. \tag{2.43}$$

Then, the solution to the system of linear equation can be easily obtained as:

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{2.44}$$

$$\mathbf{x} = \mathbf{A}^+\mathbf{b} = \sum_i^r \frac{\mathbf{v} \cdot \mathbf{b}}{s_i}\mathbf{u}_i \tag{2.45}$$

where r is the rank of matrix $\mathbf{A}$ and the dot product defined as $\mathbf{v} \cdot \mathbf{b} = \sum_i^n v_i^* b_i$.

2.1.3 Least Squares Approximation

Consider a system of linear equations:

$$\mathbf{A}\mathbf{x} = \mathbf{v} \tag{2.46}$$

where $m \times n$ matrix $\mathbf{A}$ corresponds to the linear transformation $A : \mathbb{R}^n \to \mathbb{R}^m$ (Figure 2.1), $\mathbf{v} \in \mathbb{R}^n$ is a n-dimension vector and $\mathbf{x} \in \mathbb{R}^m$ is a m-dimensional vector of unknown parameters. This system has a solution if and only if $\mathbf{v} \in \mathrm{im}(\mathbf{A})$, where image of $\mathbf{A}$ is defined as (Figure 2.1)

$$\mathrm{im}(\mathbf{A}) = \{\mathbf{z} \in \mathbb{R}^m \mid \mathbf{x} \in \mathbb{R}^n,\ \mathbf{z} = \mathbf{A}\mathbf{x}\} \tag{2.47}$$

In case the system has no exact solution, i.e. when $\mathbf{v} \notin \mathrm{im}(\mathbf{A})$, the solution $\mathbf{x}$ that minimizes the difference between $\mathbf{A}\mathbf{x}$ and $\mathbf{v}$ is considered as the least-squares solution to the system of linear equations (eq. 2.46). In another

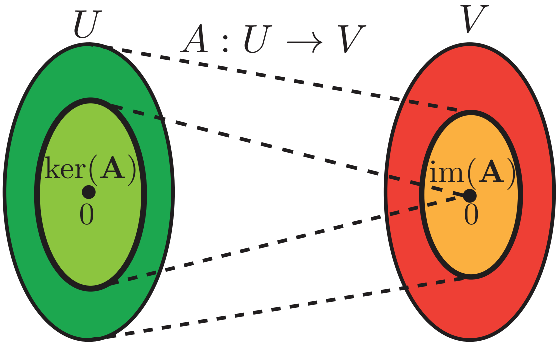


Figure 2.1: Linear transformation A maps a vector in U to a vector in V. Vectors in V that are transformed from vectors in U define image im(A). Vectors in U that are transformed to a zero in V define a kernel ker($\mathbf{A}$).

words, $\mathbf{x}$ is the solution if the vector $\mathbf{Ax}$ is the closest to $\mathbf{v}$, which is equivalent to

$$\mathbf{Ax} - \mathbf{v} \perp \operatorname{im}(\mathbf{A}). \tag{2.48}$$

Since $\operatorname{im}(\mathbf{A})^\intercal = \ker(\mathbf{A}^*)$, where kernel $\ker(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{Ax} = 0\}$ (Figure 2.1), then eq. 2.48 can be rewritten as

$$\mathbf{A}^* (\mathbf{Ax} - \mathbf{v}) = 0, \tag{2.49}$$

which results in a system of normal equations:

$$\mathbf{A}^*\mathbf{Ax} = \mathbf{A}^*\mathbf{v}. \tag{2.50}$$

Accordingly, solution to the normal equations (eq. 2.50) is equivalent to the solution of the original system of linear equations (eq. 2.46). Then, due to the property of pseudoinverse ($\mathbf{AA}^+ = 1$), the solution to the eq. 2.46 can be found using the pseudoinverse A^+ of $\mathbf{A}$:

$$\mathbf{x} = A^+\mathbf{v} = \sum_i^r \frac{\mathbf{v} \cdot \mathbf{b}}{s_i}\mathbf{u}_i, \tag{2.51}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are left and right singular vectors and s_i are singular values of matrix $\mathbf{A}$.

Ill-conditioned matrix

If the columns in the matrix $\mathbf{A}$ are (near) orthogonal then all singular values s_i have (almost) identical values indicating a well-conditioned matrix. However, often singular values of a matrix may vary in a wide range of values indicating an ill-conditioned matrix. To quantitatively measure the condition of matrix $\mathbf{A}$, the ratio between the largest s_{max} and smallest s_{min} singular values, also known as the condition number $\kappa(\mathbf{A})$, is usually used:

$$\kappa(\mathbf{A}) = \frac{s_{max}}{s_{min}} \tag{2.52}$$

The condition number measures how sensitive is the output to the changes in the input, often induced by the errors/noise in the input data. For example, the condition number associated with a linear equation $\mathbf{Ax} = \mathbf{b}$ indicates how strongly the solution $\mathbf{x}$ can change with respect to a change in $\mathbf{b}$. If the condition number is large, even a small error in $\mathbf{b}$ may lead to a significant change of the solution $\mathbf{x}$. On the other hand, if the condition number is close to unity, then the change in $\mathbf{x}$ will be comparable to the change in $\mathbf{b}$.

These numerical instabilities can be suppressed using different regularization techniques.[109–111] One way is to truncate the solution expansion of $\mathbf{A}$ (eq. 2.34) ignoring the contribution from the smallest singular values. However, since the singular values often tend to decay gradually to zero, it can be problematic to define an appropriate threshold.

In case of the least squares problem the conditioning of the problem can be improved using Tikhonov regularization[109] that adds a quadratic penalty function to the least squares sum so the solution with a smaller norm is preferred. This penalty function effectively increases the singular values of the matrix and improves the stability of the problem.

2.2 Minimization algorithms

2.2.1 Newton's Algorithms

Any kind of Newton's method exploits the quadratic approximation to the objective function within the vicinity of the minimum as suggested by a truncated Tailor's series expansion:

$$f(\mathbf{x}+\mathbf{d}) \approx f(\mathbf{x}) + \mathbf{g}(\mathbf{x})^{\intercal}\mathbf{d} + \frac{1}{2}\mathbf{d}^{\intercal}\mathbf{H}(\mathbf{x})\mathbf{d}, \tag{2.53}$$

where $\mathbf{x}$ is n-dimensional argument vector, $\mathbf{g}(\mathbf{x})$ is the gradient vector and $\mathbf{H}(x)$ is the Hessian matrix at a point $\mathbf{x}$.

The minimum of the $f(\mathbf{x})$ requires its derivative with respect to $\mathbf{x}$ to be equal to zero, resulting in the system of linear equations:

$$\mathbf{g}(\mathbf{x}) + \mathbf{H}(\mathbf{x})\mathbf{d} = 0 \tag{2.54}$$

which gives the Newton's direction $\mathbf{d}$ towards the minimum:

$$\mathbf{d} = -\mathbf{H}(x)^{-1}\mathbf{g}(x) \tag{2.55}$$

The algorithm starts from the exact calculation of the Hessian matrix $\mathbf{H}_0$ and the gradient vector $\mathbf{g}_0$ at the initial guess $\mathbf{x}_0$. Then the Newton direction d is computed, which defines the vector $\mathbf{x}_k$ for the next iteration:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d} \tag{2.56}$$

where the Hessian matrix $\mathbf{H}_k$ and gradient vector $\mathbf{g}_k$ are calculated again. This iterative procedure repeats until the solution $\mathbf{x}_k$ is not converged to a minimum $\mathbf{x}^*$ satisfying the $f'(\mathbf{x}) = 0$ within a predefined threshold.

However, the exact Newton direction is reliable only when the Hessian matrix positive definite and the the difference between the true objective function and its quadratic approximation is not too large.

The method requires exact calculation of $(n^2 + n)/2$ second-order partial derivatives of function $f(\mathbf{x})$ at each step (where n is the dimension of the vector $\mathbf{x}$) and can be computationally too demanding. In quasi-Newton's methods, instead of exact computation of the Hessian matrix, it is adjusted at each iteration and can be produced in different ways ranging from very simple to highly advanced techniques.

Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm

In the method of Broyden, Fletcher, Goldfarb, Shanno (BFGS) at each minization step k the Hessian matrix $\mathbf{H}_k$ is approximated by $\mathbf{B}_k$ using the updating formula, which converges to the true Hessian at the minimum:

$$\mathbf{B}_{k+1} = \mathbf{B}_k - \frac{\mathbf{B}_k \mathbf{s}_k \mathbf{s}_k^{\mathsf{T}} \mathbf{B}_k}{\mathbf{s}_k^{\mathsf{T}} \mathbf{B} \mathbf{s}_k} + \frac{\mathbf{y}_k \mathbf{y}_k^{\mathsf{T}}}{\mathbf{y}_k^{\mathsf{T}} \mathbf{s}_k}, \tag{2.57}$$

where $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$.

At the first iteration, $\mathbf{B}_0$ can be set to any symmetric positive definite matrix, for example, the identity matrix. The BFGS method converges superlinearly and has $O(n^2)$ complexity per iteration for n-dimensional argument vector $\mathbf{x}$.

2.2.2 Evolutionary Algorithms

An evolutionary algorithm (EA) is a class of minimization algorithms, inspired by the mechanisms of the biological evolution, such as reproduction, mutation, recombination, and selection.Each candidate solution to the optimization problem in the algorithm is called a chromosome or an individual. A set of chromosomes, called population, is evolving in a EA through the process of competition and controlled variation. For each chromosome in the population an associated score of a fitness function is evaluated to measure how well a chromosome is adapted to the environment, or in another words, how close the solution is to the minimum.

Although all evolutionary algorithms share the same principles of biological evolution to find a minimum of the function, they may differ in the details of implementation and the nature of the particular applied problem.

Genetic Algorithms

Genetic algorithms are among the most popular evolutionary algorithms, where chromosomes are usually represented in the binary code, although more recent implementations also use the real number representation.[112–114]

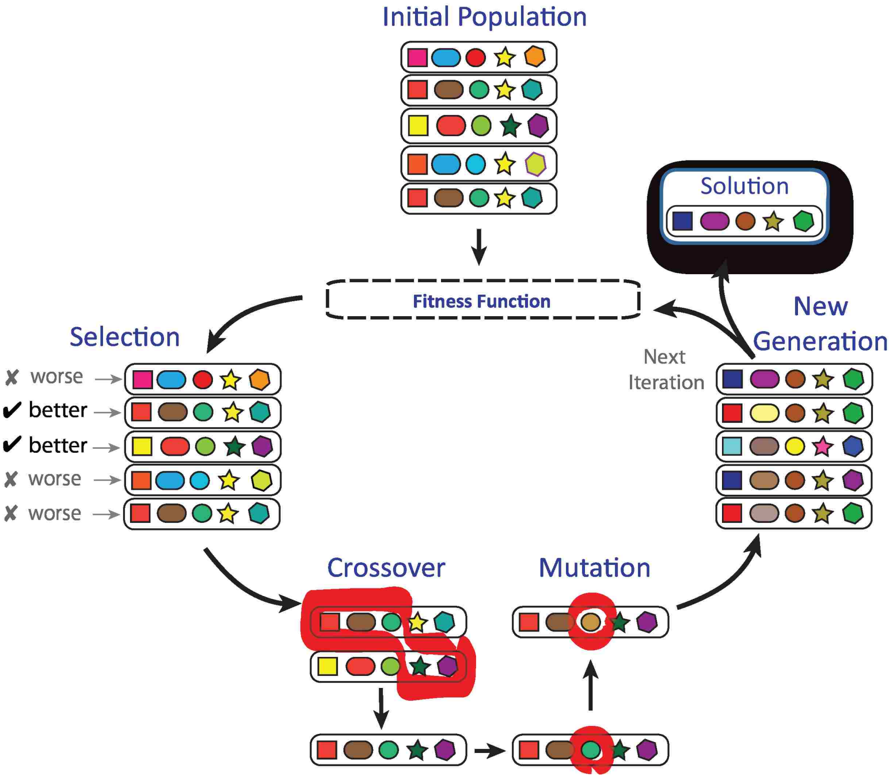

Initial Population
Fitness Function
Solution
Selection
✗ worse
✔ better
✔ better
✗ worse
✗ worse
Next Iteration
New Generation
Crossover
Mutation

Figure 2.2: A simplified scheme of the genetic algorithm procedure.

A GA starts with the initialization of the population by random generation of solutions followed by the calculation of their fitness function value (score). Then, selection operator chooses a pair of chromosomes from the population. Crossover operator breeds two chromosomes to produce an offspring, which is then added to the new generation. During each iteration, a mutation operator can mutate an offspring with a low probability. This is done to increase the diversity of solutions and avoid getting into a local minimum. When a new generation is created, score is calculated for each new chromosome, and then the next iteration starts with the selection of chromosomes for the new generation. This iterative procedure continues until the maximum number of generations is reached. A chromosome with the best fitness in a population is considered as a solution of the problem (Figure 2.2).

Covariance Matrix Adaptation Evolution Strategy(CMA-ES)

In self-adapted evolution strategy algorithms the population of new candidate solutions is sampled according to a multivariate normal distribution:[115,116]

$$\mathbf{x}_k^{(g+1)} \sim \mathbf{m}^{(g)} + \sigma^{(g)} \mathcal{N}\left(\mathbf{0}, \mathbf{C}^{(g)}\right) \text{ for } k = 1, \ldots, \lambda, \tag{2.58}$$

where $\sim$ denotes the same distribution on the left and right sides, $\mathcal{N}\left(\mathbf{0}, \mathbf{C}^{(g)}\right)$ is multivariate normal distribution with zero mean and covariance matrix $\mathbf{C}^{(g)}$, $\mathbf{x}_k^{(g+1)}$ is the k-th offspring from generation $g+1$, $\mathbf{m}^{(g)}$ is mean value of the search distribution at generation g, $\sigma^{(g)}$ is the overall standard deviation (also a step-size), λ is the number of individuals in the population.

In a single iteration of the algorithm, the mean $\mathbf{m}^{(g+1)}$, covariance matrix $\mathbf{C}^{(g+1)}$ and standard deviation $\sigma^{(g+1)}$ are calculated resulting in the self-adaptation of the solutions. The covariance matrix adaptation evolution strategy (CMA-ES) exploits a maximum-likelihood principle for the adaptation of the parameters of the search distribution. The mean $\mathbf{m}$ of the distribution is updated such that the likelihood of previously successful candidate solutions is maximized. The covariance matrix $\mathbf{C}$ of the distribution is incrementally updated such that the likelihood of previously successful search steps is increased.

The CMA-ES can be a good alternative method of function minimization in the cases when gradient methods, e.g. quasi-Newton methods (BFGS), fail due to a non-convex landscape with sharp bends, discontinuities, outliers, noise, and local optima. Calculation of the covariance matrix in the CMA-ES is analogous to the calculation of the inverse Hessian matrix in a quasi-Newton method. CMA-ES demonstrates an improved performance on ill-conditioned and/or non-separable problems by several orders of magnitude.[115,116]

Chapter 3

Genetic Algorithm Optimization of Point Charges in Force Field Development

3.1 Introduction

Molecular dynamics (MD) simulations is a powerful tool to study structure and function of biological macromolecules at the atomic level.[3–5] The accuracy of MD simulations is highly dependent on the molecular mechanics force field usedits functional form, as well as its empirical parameters. In traditional macromolecular all-atom force fields, the bonded parameters include equilibrium bond distances, bond and dihedral angles, along with the corresponding force constants and rotation barriers, while non-bonded interactions are typically described by atom-centered point charges and Lennard-Jones parameters. These bonded and non-bonded force field parameters are fitted against either experimental data or, more commonly, data obtained from electronic structure calculations. Generally, force field parameterization involves separate optimization of the bonded and non-bonded parameters, as it is common in parameterization of the classical force field models such as CHARMM,[6–8] AMBER,[9–11] GROMOS,[12] and OPLS,[13,14] as well as in more recent developments.[117–121] For instance, in parameterization of the non-bonded terms in the popular AMBER family of force fields,[9,122,123] the point charges are fitted to the reference molecular electrostatic potential (MEP) of the molecule, while Lennard-Jones parameters are fitted to reproduce the experimental bulk properties. However, simultaneous fitting of several parameters describing intermolecular interactions (point charges, Lennard-Jones parameters, and, in the case of polarizable force fields, atomic polarizabilities) may significantly improve the accuracy of force field description.[124,125] These simultaneous optimizations of different force field terms can take advantage of extensive training sets that can be easily generated using electronic structure calculations and may include data on the intermolecular interaction energies.[65–69] Moreover, in this approach the fitted interaction energy would implicitly include the polarization effects, even staying within the fixed point-charge force field framework.[70,71] However, such simultaneous force field fitting represents a technically challenging multi-objective optimization of the parameters of different physical nature. Among various optimization algorithms available for this purpose, evolutionary methods such as genetic algorithms (GAs) provide a powerful technique that can efficiently deal with complex and poorly understood search space.[112,113,126] GAs have been successfully used in force field development, including fitting of dihedral angle[127,128] and van der Waals parameters,[68,121] atomic polarizabilities,[118] parameterization of coarse-grained[129] and reactive force fields,[130,131] as well as applied in numerous ad hoc force field parameter optimizations.[132–135] Interestingly, although the assignment of the fixed point charges is a critical part of many force fields, the application of GAs and other evolutionary/stochastic optimization techniques to the MEP point-charge fitting has not been explored, to the best of our knowledge. The traditional approach for determining point charges in the force field development, usually referred to as the ESP (Electrostatic Potential) method,[32] is to fit the point charges against the reference quantum mechanical (QM) MEP Φ^{QM} by minimizing the sum of squared residuals calculated over the N point on a grid:

$$\chi^2 = \sum_{i}^{N} \left[\Phi^{QM}(\mathbf{R}_i) - \Phi^{PC}(\mathbf{R}_i)\right]^2 \tag{3.1}$$

where Φ^{PC} is the potential produced by the point charges:

$$\Phi^{PC}(\mathbf{R}) = \sum_{j}^{M} \frac{q_j}{|\mathbf{R} - \mathbf{r}_j|} \tag{3.2}$$

Examples of different implementations of this method include Merz-Kollman,[33,34] CHELP,[136] CHELPG,[137] which mainly differ by the choice of the reference grid.

These approaches typically employ Lagrange multipliers to impose constraint on the overall molecular charge and, sometimes, on the molecular dipole moment. Alternatively, the χ^2 function can be minimized directly using gradient-based methods with restraint on the total charge and dipole moment.[138] Although the atom-centered MEP-derived point charges provide a clear interpretation of the electrostatic properties and are computationally inexpensive, they can poorly reproduce the anisotropic electronic features (e.g. lone pairs, π-systems),[53,54] and also suffer from several technical difficulties. The optimized values of the point charges not only depend on the grid density and size, or the spatial orientation of the molecule relative to the Cartesian axes,[35,38,137,139,140] they also can be inconsistent even across very similar molecules, at odds with the fundamental chemical concept of the transferability of atomic properties. Not only the MEP-fitted charges for atoms of a common functional group in chemically similar molecules may be very different, the charges obtained for the conformers of the same molecule often vary by more than one electron unit. Stouch and Williams reported[39,40] that the disparate charges obtained for directly connected atoms in different conformers seem to linearly correlate with each other with high variation (~ 1.3 e) of the charge values on the interior, buried atoms (mostly aliphatic carbon atoms), while the exterior atoms (mostly hydrogens) vary in a much smaller range (~ 0.3 e). Later, the large variations of charge values have been rationalized by the low statistical contribution of the buried carbons to the overall electrostatic potential.[36] Furthermore, the ill-conditioned character of the MEP fitting problem seems to be exacerbated by the introduction of the total charge constraint using Lagrange multipliers that leads to the rank deficiency of the least-squares (LS) matrix.[37,38] The conformational dependence of the MEP-derived point charges has been significantly reduced in the Restrained Electrostatic Potential (RESP) method by Bayly et al.[36,141] that uses an external hyperbolic restraint to force the buried carbon atoms to have small point charges, thus decreasing the charge variations across different conformers. Although several alternative methods of charge derivation have been proposed,[37,136,140,142,143] restraining the charges of buried atoms to prevent the optimization from converging towards unreasonable values and/or to reduce conformational dependence of the charges became the most popular in force field development.[23,41–51] In most of these methods, besides a constraint on the total charge of the molecule, an additional restraining function is added to the LS sum to keep the buried atom charges close to some predefined values, despite its possible negative effect on the dipole moment values and the overall quality of MEP.[38,52] Considering the challenges presented by the relatively straightforward single-objective point charge fitting against the MEP, simultaneous optimization of point charges along with other force field parameters against a diverse training set could be expected to present more pitfalls. Therefore, in this chapter we investigate the performance of the GA techniques when applied to the MEP point charge fitting problem in a case of small model molecules with the emphasis on the convergence properties of the algorithm.

3.2 Details of Charge Fitting

3.2.1 Least Squares Fitting

In the ESP method the solution is obtained by minimizing the LS sum (eq. 3.1) that can be rewritten in a more compact algebraic form:

$$\chi^2 = |\mathbf{\Phi} - \mathbf{A}\mathbf{q}|^2 = |\mathbf{\Phi}|^2 + \mathbf{g}^\intercal \cdot \mathbf{q} + \mathbf{q}^\intercal \mathbf{H} \mathbf{q} \tag{3.3}$$

$$\mathbf{g} = -2\mathbf{A}^\intercal \mathbf{\Phi} \tag{3.4}$$

$$\mathbf{H} = \mathbf{A}^\intercal \mathbf{A} \tag{3.5}$$

where the vector $\mathbf{\Phi}$ consists from the reference electrostatic potential calculated at each point of the grid; $\mathbf{q}$ is a set of point charges; $\mathbf{A}$ is the LS matrix with the elements corresponding to the inverse distance $1/r_{ij}$ between point i of the grid and point charge j in the molecule; vector $\mathbf{g}$ and matrix $\mathbf{H}$ are gradient vector and Hessian matrix of the LS sum, correspondingly. Because of the quadratic dependence of the LS sum on the charge vector $\mathbf{q}$ the solution to the LS problem can be found by setting partial derivatives of χ^2 with respect to each point charge to zero, which results in the system of linear equations, known as normal equations:[144]

$$\mathbf{A}^{\intercal}\mathbf{A}\mathbf{q} = \mathbf{A}^{\intercal}\mathbf{\Phi} \tag{3.6}$$

where $\mathbf{q}$ is the solution to the problem which is further referred to as ESP charges and used as the reference to compare against the GA-optimized values. No additional constraints or restraints have been imposed to these charges, except for the atom equivalence due to the symmetry of the molecule.

3.2.2 Fitting with Genetic Algorithms

In the GA approach, each candidate solution is referred to as a chromosome or an individual. A set of chromosomes, called population, is evolving during a GA run through iterative application of genetic operators of selection, crossover and mutation.[112,113] Each chromosome in the population has an associated fitness function value, or a fitness score, that measures how close this candidate solution is to the desired optimum solution. The algorithm starts by randomly generating the initial population of the chromosomes, followed by evaluation of their fitness function values. These scores are then used to select chromosomes for further crossover and mutation that produce the next generation of the chromosomes. When the number of generation reaches maximum, the algorithm stops and the chromosome with the best fitness score in the final population is considered as solution to the optimization problem.The GA parameters used here for the point charge fitting against a reference MEP are given in Tables 3.2.2 and 3.2.2. Each chromosome encoded a set of atom-centered point charges either in a traditional binary or real number representation. We found that, as in several other cases,[114,145] the real-number coding requires smaller population size than the binary coding to achieve the results of the same quality. Therefore, the real-coded chromosomes were used throughout this work.

Table 3.1: Parameters used in the GA fitting of the MEP point charges

Parameter	Value
Maximum number of generations	100
Population size	20-200
Variable range	$[-1; 1]$

Table 3.2: Genetic operators used in the GA fitting of the MEP point charges

Operator	Binary-Coded	Real-Coded	Probability
Crossover	Two-point	BLX-α, $\alpha = 0.5$	0.90
Mutation	Flip bit	Random	0.03
Selection	Proportional selection		–

All point charges have been fitted within the -1 to +1 e range, with no additional restraints, unless stated otherwise. The root-mean square error (RMSE) was used as the fitness function:

$$f = \text{RMSE} = \sqrt{\frac{\sum_i^N \left[\Phi^{QM}(\mathbf{R}_i) - \Phi^{PC}(\mathbf{R}_i)\right]^2}{N}} = \sqrt{\frac{\chi^2}{N}} \tag{3.7}$$

Thus, the chromosome with the lowest fitness score in the last generation was considered as the solution being sought. RMSE has been chosen as the fitness function because of its clear statistical meaning–an average error per grid point; however, using either the RMSE or the LS sum χ^2 (eq. 3.1) as the fitness function in the GA optimizations gives very similar results. The average fitness score of a population $\langle f \rangle$ calculated at each generation was used to characterize the convergence of a single GA run, while the standard deviation σ_f was used to characterize how diverse or localized are the chromosomes in the population:

$$\langle f \rangle = \frac{1}{S} \sum_i^S f_i \tag{3.8}$$

$$\sigma_i = \sqrt{\frac{1}{S} \sum_i^S (f_i - \langle f \rangle)^2} \tag{3.9}$$

Due to the stochastic nature of the algorithm, several independent GA runs were used to assess the quality/scatter of the obtained solutions. In most cases several runs converged to a set of widely dispersed solutions. To understand the nature of this dispersion and reveal possible correlations between optimized parameters, we computed variance-covariance (or covariance) matrices $\mathbf{\Sigma}$ for each set of the obtained GA solutions. The diagonal elements of the covariance matrix contain the charges variances (eq. 3.10) and the off-diagonal elements contain the covariances between each pair of charges (eq. 3.11):

$$\mathrm{var}(q_j) = \frac{1}{N-1}\sum_i^N (q_{ij} - \langle q_j \rangle)^2 \tag{3.10}$$

$$\mathrm{cov}(q_j, q_k) = \frac{1}{N-1}\sum_i^N (q_{ij} - \langle q_j \rangle)(q_{ik} - \langle q_k \rangle) \tag{3.11}$$

where N is the number of GA runs, q_{ij} is the charge on atom j from ith GA run, $\langle q_j \rangle$ is charge on atom j averaged over all GA runs. Eigenvectors of the covariance matrix form an eigenbasis $\widetilde{\mathbf{\Sigma}}$ consisting from the orthonormal vectors $\mathbf{s}_i$ (principal components), along which the data is changing with the variance defined by the corresponding eigenvalue σ_i^2:

$$\mathbf{\Sigma}\mathbf{s}_i = \sigma_i^2 \mathbf{s}_i \tag{3.12}$$

$$\widetilde{\mathbf{\Sigma}} = (\mathbf{s}_1 \ \ldots \ \mathbf{s}_M) \tag{3.13}$$

where $\widetilde{\mathbf{\Sigma}}$ is the square matrix of size M, defined by the number of point charges; σ_i is standard deviation along eigenvector $\mathbf{s}_i$.

3.3 GA Charge Fitting for Small Molecules

First, we examine the performance of GAs for the MEP point charge fitting in a straightforward case of several small molecules with only two symmetry-independent charges, but vastly different electrostatic properties: water, ammonia, benzene, and methane. For these systems, a single GA run with a small population size (< 40 chromosomes) converges to a localized set of solutions within 25-50 generations, after which the population stabilizes with only small fluctuations of the charge values/fitness scores (Figure 3.1). Surprisingly, although all GA runs demonstrate robust convergence, independent runs converge to vastly different solutions for the same molecule (Figure 3.2). For instance, 200 GA runs for CH_4 produced solutions with charges on the carbon atom q_C varying from -0.99 to 0.95 e, while the charge on the hydrogen varied from -0.24 to 0.25 e. Similar scatter of the small-population GA-derived charge values is observed for other molecules. In the case of H_2O, NH_3, and CH_4 the charges of the central, "buried" atoms show much larger deviations than the hydrogen atom charges. Although highly dispersed, the GA solutions tend to cluster around the solutions that correspond to the charges derived with the ESP method, eq. 3.6 (shown as yellow dots in Figure 3.2). Increase of the population size decreases the scatter: GA runs with populations < 50 chromosomes yield solutions within ±0.01 e of the ESP values.

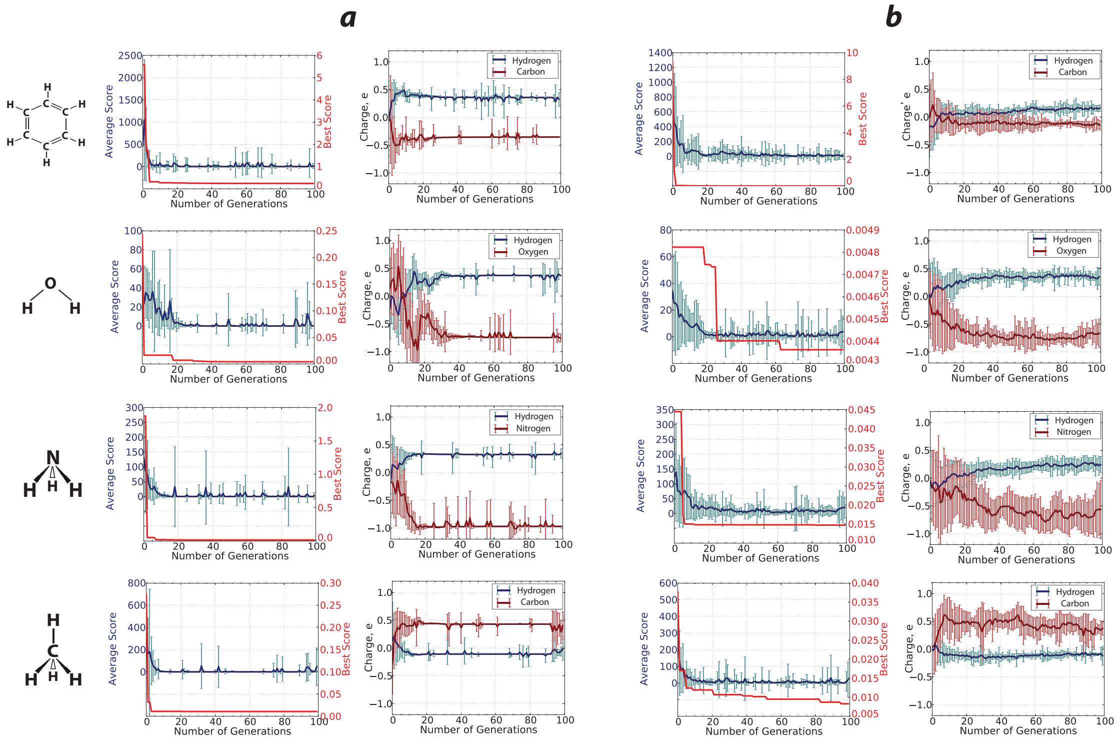


Figure 3.1: GA convergence with 20 chromosomes in the population (a) as compared to 50 chromosomes in the population (b).

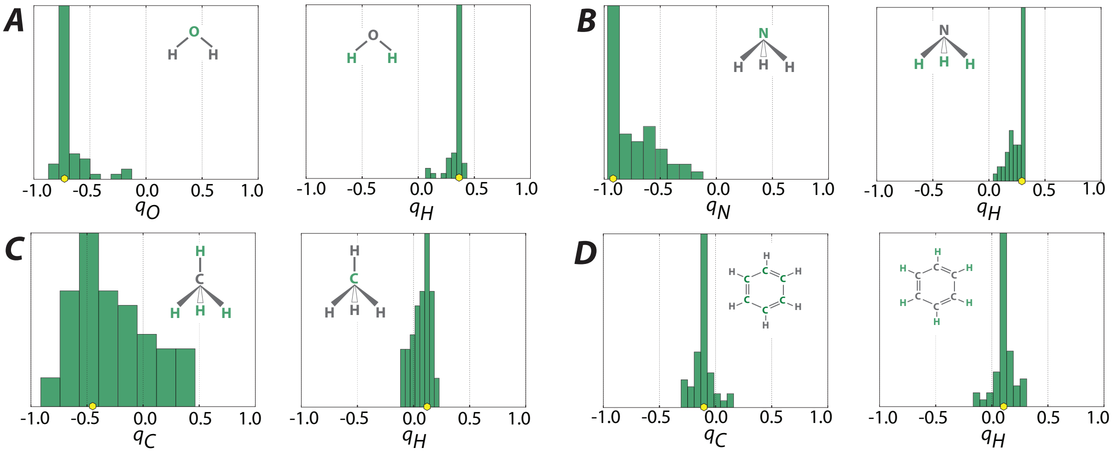


Figure 3.2: Distributions of the GA-optimized charges for the model molecules with two symmetry-independent charges, obtained from 200 GA runs with 20 chromosomes in the population. Yellow dots indicate the solutions obtained with the ESP method.

At first glance, these results simply suggest that MEP point charge fitting with GAs is highly inefficient and requires larger population sizes. It is, however, intriguing why the small-population GA runs quickly converge to non-optimal solutions that cannot be improved upon any further, even in hundreds of additional generations (premature convergence). In other words, what is the origin of these non-optimal solutions that trap small-population GA runs? Further investigation revealed that there is almost a perfect ($R^2 = 1.00$) linear correlation between the pairs of q_X (X = O, N, or C) and q_H values produced from different GA runs (Figure 3.3). For each correlation, the slopes correspond to the number of hydrogen atoms per atom X in the molecule, while the intercept correspond to the overall charge $Q = 0.0$ e of the molecule:

$$Q = n_x q_x + n_H q_H \tag{3.14}$$

$$q_X = -\frac{n_H}{n_X} q_H + \frac{1}{n_X} Q \tag{3.15}$$

where n_X is the number of X atoms, and n_H/n_X is the number of hydrogen atoms per atom X. Indeed, although the GA runs converge to dispersed solutions, the zero total charge is always reproduced, with standard deviation σ in the range from 0.001 to 0.01 e.

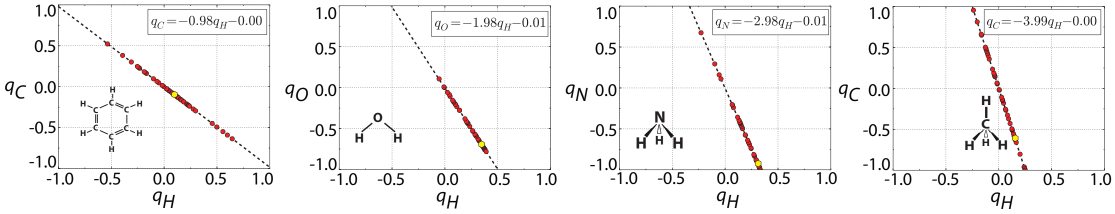


Figure 3.3: Correlations between the GA-optimized charges for the two-charge model molecules obtained from 200 independent GA runs with 20 chromosomes in the population; all trend lines have correlation coefficient $R^2 = 1.00$. Yellow dots indicate the solutions obtained with the ESP method.

We further investigated the GA-fitting performance for molecules with three symmetry-independent charges on the example of mono- and di-substituted methane derivatives CH_3X, X = F, Cl, O^-, and CH_2X_2, X = F, Cl. Similarly to the two-charge systems, multiple small-population GA runs (< 100 chromosomes) yield highly scattered solutions, which tend to cluster around the ESP values as the population sizes increase. However, only GA runs with > 100 chromosomes yield consistent results that match the ESP charges within ±0.01. The scatter is the largest in case of the charges on the carbon atoms q_C: e.g. 200 30 chromosome GA runs for CH_3Cl produce q_C values covering the entire -1 to +1 e range, while the charge on hydrogen and chlorine vary in much small ranges (-0.1 to 0.3 e and -0.3 to -0.1 e, respectively).

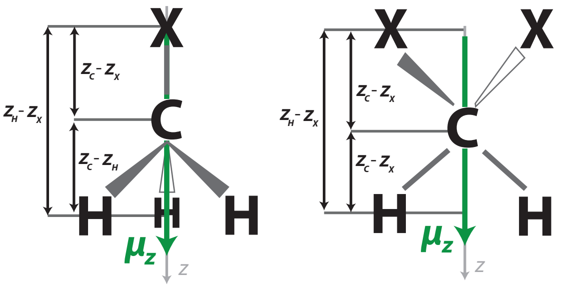


Figure 3.4: The coordinate system for the CH_3X and CH_2X_2 molecules.

Unlike the two-charge systems, the GA solutions for CH_3X, and CH_2X_2 not only reproduce the correct total charge, but also produce constant dipole

Table 3.3: Average values and the standard deviations (in parenthesis) of the monopole and dipole moments computed from the GA-optimized point charges for CH_3X, CH_2X_2 (X = F, Cl), and CH_3O^- molecules along with the reference values from DFT calculations.

Molecule	Monopole, au	Dipole, au	DFT dipole, au
CH_3F	0.002 (0.001)	0.782 (0.005)	0.771
CH_3Cl	0.000 (0.002)	0.827 (0.042)	0.794
CH_2F_2	-0.002 (0.002)	0.814 (0.087)	0.803
CH_2Cl_2	-0.001 (0.006)	0.712 (0.047)	0.667
CH_3O^-	-0.9674 (0.006)	0.847(0.018)	0.772

moment values, which are close to the reference DFT values (Table 3.3): the standard deviation σ is in 0.001 to 0.006 e range for the total charge and in 0.005 to 0.087 au range for the dipole moment. Thus, regardless of the population size, the GA-optimized point charges satisfy the eqs. 3.16 and 3.17 for the first two terms of the multipole expansion: the monopole/total charge and the dipole moment. These equations can be written as dot products between the charge vector $\mathbf{q}$ and the corresponding vector $\mathbf{u}_i$:

$$Q = n_X q_X + n_C q_C + n_H q_H = \mathbf{u}_1 \cdot \mathbf{q} \tag{3.16}$$

$$\mu_z = n_X z_X q_X + n_C z_C q_C + n_H z_H q_H = \mathbf{u}_2 \cdot \mathbf{q} \tag{3.17}$$

where n_A is the stoichiometric number of the atom A in the molecule, z_A is its coordinate along the z axis (oriented along the symmetry axis as shown in Figure 3.4), and q_A is its point charge. Geometrically, these equations define two planes with the vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ which are orthogonal to the corresponding plane. The GA solutions align along a three-dimensional line formed by the intersection of these two planes (Figure 3.5A) which is defined by the cross product vector $\mathbf{u}_3 = \mathbf{u}_1 \times \mathbf{u}_2$:

$$\mathbf{q} = \mathbf{q}_0 + t\mathbf{u}_3 \tag{3.18}$$

where t is a free parameter, the vector $\mathbf{q}_0$ is a set of point charges that satisfies eqs. 3.16 and 3.17. Projections of this three-dimensional line give three pairwise linear relationships between each pair of the atomic charges (Figure 3.5B): e.g., a projection on the (q_C, q_H) plane results in a linear correlation between q_C and q_H. These pairwise correlations can be derived using the geometric parameters (Figure 3.4) and dipole moment values:

$$q_C = -\frac{n_H}{n_C}\frac{z_H - z_X}{z_C - z_X}q_H + \frac{\mu_z - Qz_X}{z_C - z_X} \tag{3.19}$$

Importantly, there is a good numerical agreement between the correlations obtained analytically using the DFT dipole moments and from the linear fitting of the scattered GA solutions (Table A2 in the Appendix A). Thus, the linear relationships observed for the two- and three-independent charge systems arise because all GA solutions satisfy the constant total charge and (for the three-charge systems) the dipole moment requirements, while the higher multipole moments produced by these solutions are scattered.

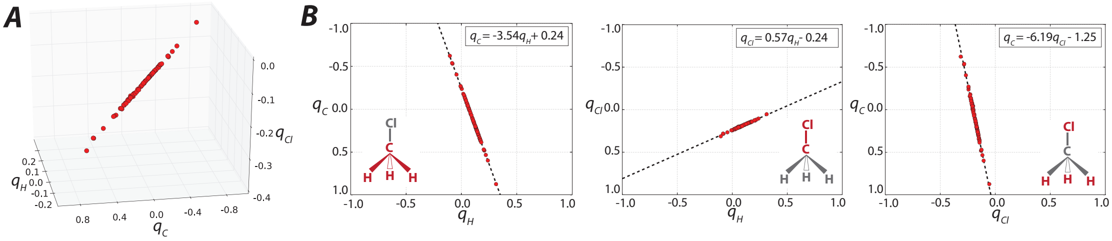


Figure 3.5: The correlation between the chloromethane point charges obtained from 200 independent GA runs shown in three dimensions (A) and as two-dimensional projections, i.e. pairwise correlations between charges (B).

3.4 Covariance Matrix Analysis

In the trivial case of the two- and three-independent charge systems, the scattered nature of the small-population GA-optimized point charges can be interpreted using a simple correlation analysis (Figures 3.3 and 3.5). However, understanding the results for larger, more realistic molecules would require more general approach, such as the analysis of the eigenvectors of the covariance matrix $\mathbf{\Sigma}$ computed for a set of GA solutions. We tested this approach by re-examining the small-population GA results for the two- and three-charged model systems discussed above. For the two-charge molecules, the covariance matrix diagonalization (Table A3 in the Appendix A) yields one vector with almost negligible variance/eigenvalue ($\sigma_1^2 < 10^5$) and one vector with much higher variance ($\sigma_2^2 \in [0.06; 0.19]$). The first vector $\mathbf{s}_1$, along which the data does not vary, numerically corresponds to the normalized vector $\mathbf{u_1}$ that defines the total charge and is determined by the stoichiometry of the molecule:

$$Q = n_X q_X + n_H q_H = \mathbf{u}_1 \cdot \mathbf{q} \tag{3.20}$$

where $\mathbf{u}_1 = (n_X \ n_H)$ and $\mathbf{q} = (q_X \ q_H)$. The second vector s_2, i.e. the vector along which the data shows a significant variation, numerically corresponds to a normalized vector $\mathbf{u} = (n_H \ \ -n_X)$, also determined by the stoichiometry. Thus, the eigenbasis of the covariance matrix $\widetilde{\mathbf{\Sigma}}$ can be represented as:

$$\widetilde{\mathbf{\Sigma}} = (\mathbf{s}_1 \ \mathbf{s}_2) = \left(\frac{\mathbf{u}_1}{|\mathbf{u}_1|} \ \frac{\mathbf{u}_2}{|\mathbf{u}_2|} \right) \tag{3.21}$$

The dramatic difference in the data variation along the two covariance eigenvectors suggests that the fitness function has very different curvatures along these two directions. This curvature of the fitness function can be examined explicitly by computing and diagonalizing its Hessian matrix, or, for simplicity, the Hessian of the LS sum H (eq. 3.5):

$$\mathbf{H}\mathbf{h}_i = \kappa_i \mathbf{h}_i \tag{3.22}$$

$$\widetilde{\mathbf{H}} = (\mathbf{h}_1 \ \ldots \ \mathbf{h}_M) \tag{3.23}$$

As can be seen from Figure 3.6 and Table A3 in the Appendix A, the Hessian eigenbases $\widetilde{\mathbf{H}}$ computed for all four two-charge molecules are numerically identical to the corresponding covariance matrix eigenbases $\widetilde{\mathbf{\Sigma}}$ and the basis of normalized vectors $\mathbf{u}_i$ in $\widetilde{\mathbf{U}}$:

$$\widetilde{\mathbf{\Sigma}} = \widetilde{\mathbf{H}} = \widetilde{\mathbf{U}} = \left(\frac{\mathbf{u}_1}{|\mathbf{u}_1|} \ \frac{\mathbf{u}_2}{|\mathbf{u}_2|} \right) \tag{3.24}$$

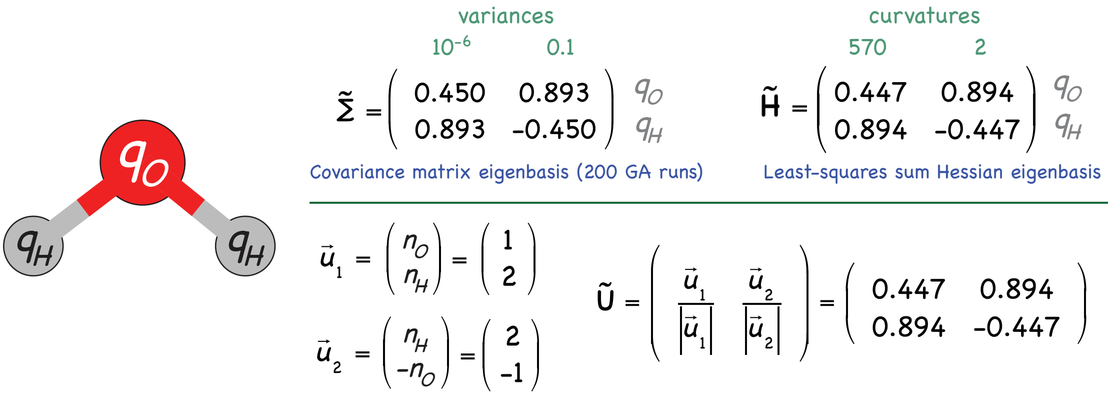

variances
10⁻⁶ 0.1
Σ̃ = (0.450 0.893; 0.893 -0.450) q_O q_H
Covariance matrix eigenbasis (200 GA runs)
curvatures
570 2
H̃ = (0.447 0.894; 0.894 -0.447) q_O q_H
Least-squares sum Hessian eigenbasis
q_O
q_H
q_H
u₁ = (n_O; n_H) = (1; 2)
u₂ = (n_H; -n_O) = (2; -1)
Ũ = (u₁/|u₁| u₂/|u₂|) = (0.447 0.894; 0.894 -0.447)

Figure 3.6: Numerical equivalence of the eigenvectors of the covariance matrix for the results of 200 GA runs, the eigenvectors of the least-squares sum Hessian matrix, and the normalized vectors $\mathbf{u}_1$ and $\mathbf{u}_2$, on the example of water molecule.

There is an inverse relationship between the eigenvalues of the fitness function/LS sum Hessian and the covariance matrices: the Hessian eigenvector $\mathbf{h}_2$ with near-zero eigenvalue/curvature corresponds to the covariance eigenvector $\mathbf{s}_2$ with a large variance; at the same time, the Hessian eigenvector $\mathbf{h}_1$ with a large curvature corresponds to the covariance eigenvector $\mathbf{s}_1$ with near-zero variance. The latter high-curvature/small-variance vector is also the vector that defines the total charge of the molecule, $\mathbf{u}_1$ (eq. 3.20). Thus, the linear correlations observed for the GA solutions (Figure 3.3) arise due to a high curvature of the fitness function with respect to the deviation of the total charge from the optimal value (zero for the studied molecules).

The fitness function plots indeed show a dramatic difference in the curvatures (Figure 3.7): when plotted against q_X and q_H, the fitness function has a characteristic V-like shape, with the line of zero total charge going through the bottom of the valley (eq. 3.20). As evident from the 3D plots, changing the central atom charge q_X from -1 to 1 e can result in up to 300-800 kcal/mol increase of the fitness function. At the same time, 2D profiles along the zero total charge line show 1-2 orders smaller variation of the fitness function values ($<$ 60 kcal/mol, note the difference in scales for the 3D and 2D plots in Figure 3.7). The actual minimum of the fitness function is determined by the next non-vanishing multipole moment indicated by the positions of the arrows in Figure 3.7.

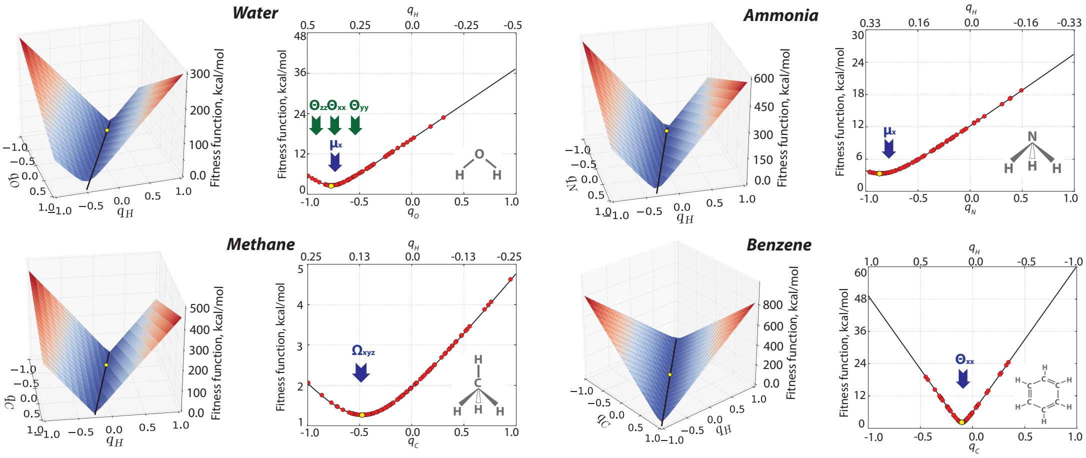


Figure 3.7: Fitness function profiles for the two-charge model molecules: full profiles (3D plots) and the profiles along the zero total charge line (2D plots). Red dots show the solutions obtained from GA optimizations (200 runs), and the yellow dots indicate the ESP solutions.

In case of the three-charge model molecules CH_3X and CH_2X_2, diagonalization of the covariance matrices $\mathbf{\Sigma}$ of the scattered GA solutions yields two vectors, $\mathbf{s}_1$ and $\mathbf{s}_2$, along which the variance is negligible ($\sigma^2_{1,2} < 10^{-5}$), and the third $\mathbf{s}_3$ with much larger variation of the data ($\sigma^2_3 \in [0.1; 0.2]$). As the GA solutions conserve both the total charge Q and the dipole moment μ_z, we can expect that the $\mathbf{s}_1$ and $\mathbf{s}_2$ vectors correspond to the vectors $\mathbf{u}_1 = (n_X \; n_C \; n_H)$ and $\mathbf{u}_2 = (n_X z_X \; n_C z_C \; n_H z_H)$, eqs. 3.16 and 3.17, in which case the third vector $\mathbf{s}_3$ should be collinear with the cross product $\mathbf{u}_3 = \mathbf{u}_1 \times \mathbf{u}_2$, along which the GA solutions are distributed. Unlike the $\mathbf{s}_1$ and $\mathbf{s}_2$ vectors, the $\mathbf{u}_1$ and $\mathbf{u}_2$ vectors are generally not orthogonal, but their orthogonality can be achieved by appropriately shifting the coordinate origin:

$$\mathbf{u}_1 \cdot \mathbf{u}_2 = 0 \tag{3.25}$$

$$n_X^2(z_X - z_0) + n_C^2(z_C - z_0) + n_H^2(z_H - z_0) = 0 \tag{3.26}$$

$$z_0 = \frac{n_X^2 z_X + n_C^2 z_C + n_H^2 z_H}{n_X^2 + n_C^2 + n_H^2} \tag{3.27}$$

where z_0 is the coordinate of the new origin along the z axis.

As expected, the set of the three orthogonal vectors $\mathbf{u}_i$:

$$\begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix} = \begin{pmatrix} n_X & n_X(z_X - z_0) & n_C n_H(z_C - z_H) \\ n_C & n_C(z_C - z_0) & n_H n_X(z_H - z_X) \\ n_H & n_H(z_H - z_0) & n_X n_C(z_C - z_C) \end{pmatrix} \tag{3.28}$$

numerically matches, after normalization, with the eigenbasis of the corresponding covariance matrix of the GA solutions $\mathbf{\Sigma}$ and the eigenbasis of the LS sum Hessian matrix $\mathbf{H}$ (Table A4 in the Appendix A):

$$\widetilde{\mathbf{\Sigma}} = \widetilde{\mathbf{H}} = \widetilde{\mathbf{U}} = \begin{pmatrix} \mathbf{u}_1/|\mathbf{u}_1| & \mathbf{u}_2/|\mathbf{u}_2| & \mathbf{u}_3/|\mathbf{u}_3| \end{pmatrix} \tag{3.29}$$

Thus, analysis of the covariance matrix provides a convenient and general method to understand the nature of the premature convergence of the small-population GA point charge optimizations that yields highly dispersed suboptimal solutions.

3.5 Rotation of the Optimization Coordinates

As it was shown, GA optimizations of point charges tend to quickly converge with respect to the leading terms of the multipole expansion associated with large curvature of the LS sum, but have difficulty navigating towards the minima along the other directions defined by the Hessian eigenvectors associated with small curvatures. Thus, the Hessian/covariance matrix eigenvectors provide a set of linearly independent, natural coordinates expressed as linear combinations of the point charge coordinates. The latter, on the other hand, represent a linearly dependent set of coordinates for the fitness function minimization problem. In fact, optimization in a rotated coordinate system is known to dramatically deteriorate the GA convergence.[146] This can be illustrated on the example of minimization of a simple function of two variables (Figure 3.8A) that has a low curvature along the x-axis and much higher curvature along the y-axis, resulting in a V-shaped surface similar to the fitness function of the two-charge systems (Figure 3.7). This model function does not present a problem for GA optimization in terms of the linearly independent parameters x and y, as written in Figure 3.8A: all GA runs quickly converge to the true minimum (zero standard deviation of the GA solutions). However, if the coordinate system is rotated by angle θ relative to the original axes (Figure 3.8B), the GA performance significantly deteriorates, as is evident from the increasing standard deviation, which reaches the maximum for $\theta = 45°$(Figure 3.8C).

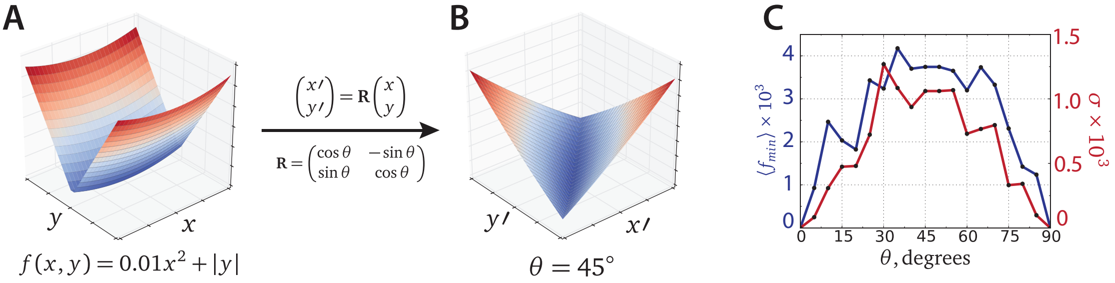


Figure 3.8: The effect of coordinate rotation on the convergence of GA minimizations on the example of a simple model function f of two variables associated with highly different curvatures: the model function plotted in the original coordinate system (A) and in the coordinate system rotated by 45°(B); the average of f_{min} values obtained from 50 GA minimization runs (blue) and the corresponding standard deviations (red) vs the rotation angle θ.

This effect can be understood in terms of the high selective pressure along the high-curvature component y. The first chromosome to reach the minimum along y, i.e the line at the bottom of the valley, will quickly dominate the entire GA population; any new chromosome that even slightly deviates in the high-curvature direction incurs high fitness penalty and is not propagated to the next generation. In the original non-rotated coordinate system, the population is free to explore various values of the low-curvature parameter x without straying away from the bottom of the valley along the coordinate y. However, in the case of a rotated coordinate system, the population would produce a viable offspring in the direction of the global minimum only if both linearly dependent variables x' and y' change in a precise way to stay at the bottom of the valley. Since this is a low-probability event for a small population, the population stops changing once it reaches the minimum along the high-curvature direction, even though it may be far from the minimum along the low-curvature direction.

This population stagnation/premature convergence of the GA optimizations in rotated coordinate systems can be overcome by using large populations and/or higher mutation rates, which can lead to a significant computational cost. A more appealing solution is to perform the optimization in linearly independent coordinates determined by the eigenbasis of the LS-sum Hessian $\widetilde{\mathbf{H}}$. In this case, the chromosomes encode a vector $\mathbf{n}$ of M real numbers–the optimization coordinates in the basis $\widetilde{\mathbf{H}}$, while the fitness function is still evaluated in terms of the point charges $\mathbf{q}$ (eq. 3.7) obtained using a linear transformation:

$$\mathbf{q} = \widetilde{\mathbf{H}}\mathbf{n} \tag{3.30}$$

We tested this approach for the same two- and three-charge model molecules discussed above. With other GA parameters kept unchanged, optimizations in the new coordinate system demonstrated much more robust convergence, as they require less than a half of the population size to achieve results of the same accuracy. For example, in the case of the three-charge CH_3X and CH_2X_2 molecules, 30 chromosomes were sufficient to converge to solutions that match the ESP charges within ±0.01 e, and to completely eliminate the linear correlations observed for the direct point charge optimizations (Table 3.5).

Thus, the efficiency of the point charge fitting using GAs can be dramatically improved by rotating the optimization coordinates using the eigenvectors of the LS-sum Hessian. As we already discussed, the covariance matrix of the GA solutions is numerically equivalent to the Hessian, and this useful property of the covariance matrices is utilized in some recently developed advanced evolutionary methods such as the covariance matrix adaptation evolution strategy (CMA-ES) approach.[115,116,147] Like other evolutionary strategy (ES) techniques, CMA-ES differs from less sophisticated classical GA methods in the implementation of the crossover and mutation operations; in some cases (CMA-ES included), new

Table 3.4: Charge fitting for two- and three-charge model molecules: average fitness scores with standard deviations (in parenthesis) for the GA optimizations (200 runs with 30 chromosomes per generation) using the point charge coordinates vs the coordinates defined by the LS-sum Hessian eigenvectors, along with the fitness scores of the reference ESP solutions; all units are in kcal/mol.

Molecule	Point-charge coordinate	Eigenvector coordinates	ESP
H_2O	2.91(1.02)	$2.66(5.34 \times 10^{-6})$	2.66
NH_3	3.89(1.06)	$3.34(1.06 \times 10^{-5})$	3.34
C_6H_6	2.82(1.83)	$2.15(1.50 \times 10^{-5})$	2.15
CH_4	1.66(0.57)	$1.27(1.30 \times 10^{-6})$	1.27
CH_3Cl	2.46(0.41)	$2.14(4.84 \times 10^{-2})$	2.14
CH_2Cl_2	2.79(0.42)	$2.46(1.95 \times 10^{-5})$	2.46
CH_3F	2.26(0.41)	$1.89(3.06 \times 10^{-5})$	1.89
CH_2F_2	2.35(0.64)	$1.84(2.17 \times 10^{-5})$	1.84
CH_3O^-	4.71(0.98)	$3.61(5.02 \times 10^{-5})$	3.61

candidate solutions/offspring are sampled from the multivariate normal distribution, rather than produced by the traditional crossover operator. However, the most important CMA-ES feature in the context of this discussion is that a new set of solutions is generated using an approximate covariance matrix, which is updated at every step of the optimization. In this respect, CMA-ES is highly reminiscent of the quasi-Newton optimization techniques that use an approximate Hessian matrix which is updated at every step.

3.6 Large-Molecule Example

Here, we tested the performance of the GA and CMA-ES methods for the point-charge fitting problem in the case of five conformers of 1-chlorobutane, a more realistic example than the two- and three-charge models discussed so far (Figure 3.9).

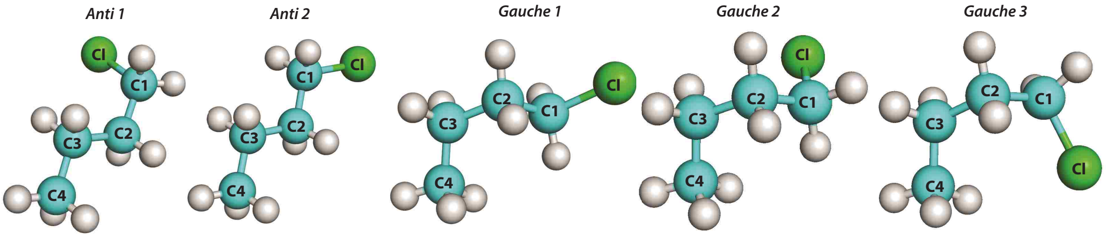


Figure 3.9: Five conformers of 1-Chlorobutane with carbon atom numbering.

In line with the assumptions made in the force field development, the hydrogen atoms within each methyl and methylene group were considered equivalent, giving nine point charge values overall to optimize for each conformer; the point charges were fitted separately for each conformer. In each case, 200 GA runs with population of 200 chromosomes produced highly scattered solutions with the average fitness score significantly higher than that of the reference ESP solutions (Table 3.6). However, just like in the case of the small models, the GA solutions consistently reproduce the total charge and the magnitude of the dipole moment; also, there is a very good correspondence between the eigenvectors of the covariance matrix of the GA solutions and the LS sum Hessian (Figure 3.10). The eigenvector that corresponds to the highest curvature (~ 3600) and the smallest variance ($\sim 10^{-6}$) corresponds to the total charge; it is identical for all conformers. While in the case of a large molecule such as 1-chlorobutane it is less straightforward to derive analytical expressions for the other high-curvature/low-variance eigenvectors, they seem to correspond to the leading multipole momentsthe correspondence which is especially clear for the second highest-curvature vector (curvature ~ 200, variance $\sim 10^{-5}$) that defines the main dipole moment component. As the curvature decreases, the physical interpretation of the associated eigenvectors becomes less clear, and the similarity between the eigenvectors calculated for different conformers decreases, reflecting different electrostatic properties of these conformers. The last four eigenvectors have curvatures in the 0.3 to 0.03 range and correspondingly large variances, $\sim 10^{-2} - 10^{-1}$. These low-curvature/high-variance coordinates have a small contribution to the overall MEP, do not seem to be associated with particular multipole moments, and primarily depend on the charges of the buried carbon atoms.

The GA optimizations in terms of the variables defined by the LS-sum Hessian eigenvectors yielded solutions with much better fitness scores (Table 3.6) and significantly decreased the scatter of the solutions (Figure 3.11). At the

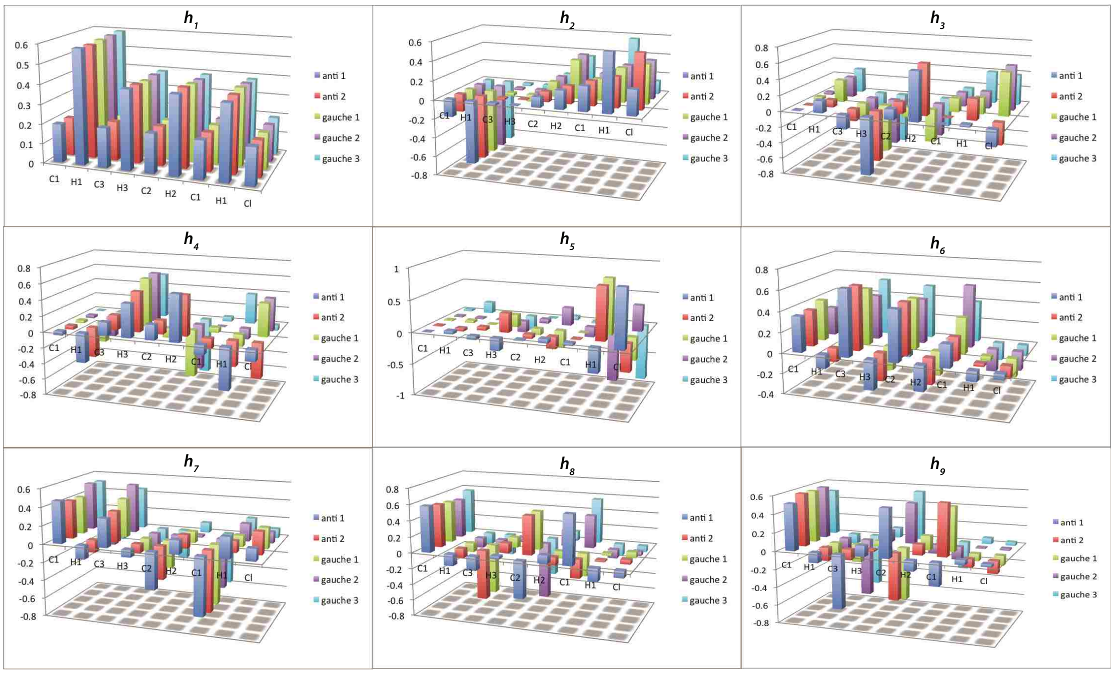


Figure 3.10: Bar-chart representation of the eigenvectors of the Hessian matrix for five conformers of 1-chlorobutane.

Table 3.5: Charge fitting for 1-chlorobutane conformers: average fitness scores with standard deviations (in parenthesis) for the GA optimizations using two coordinate systems (point charges and Hessian eigenvectors), along with the fitness scores of the CMA-ES and ESP solutions; all units are in kcal/mol.

Conformation	Point charges	Eigenvectors	CMA-ES	ESP
anti 1	3.05(0.43)	2.58(0.20)	2.09	2.09
anti 2	3.02(0.41)	2.57(0.19)	2.13	2.13
gauche 1	3.06(0.45)	2.61(0.21)	2.12	2.12
gauche 2	3.08(0.45)	2.58(0.19)	2.10	2.10
gauche 3	3.15(0.49)	2.62(0.18)	2.14	2.14

same time, multiple CMA-ES runs converged to the identical solutions, which are also equal within more than five decimal placesto the ESP values. The superb performance of CMA-ES method in this test case suggests that it could be a promising global-search evolutionary technique for force field development.

3.7 Variance of the Least Squares Solution and the Buried Atom Effect

Besides their importance for the application of evolutionary methods in the force field development, the insights into the severe convergence problems of the point charge fitting using classical GA methods can also be useful to revisit some of the well-known issues with the ESP method. The ESP charges can vary

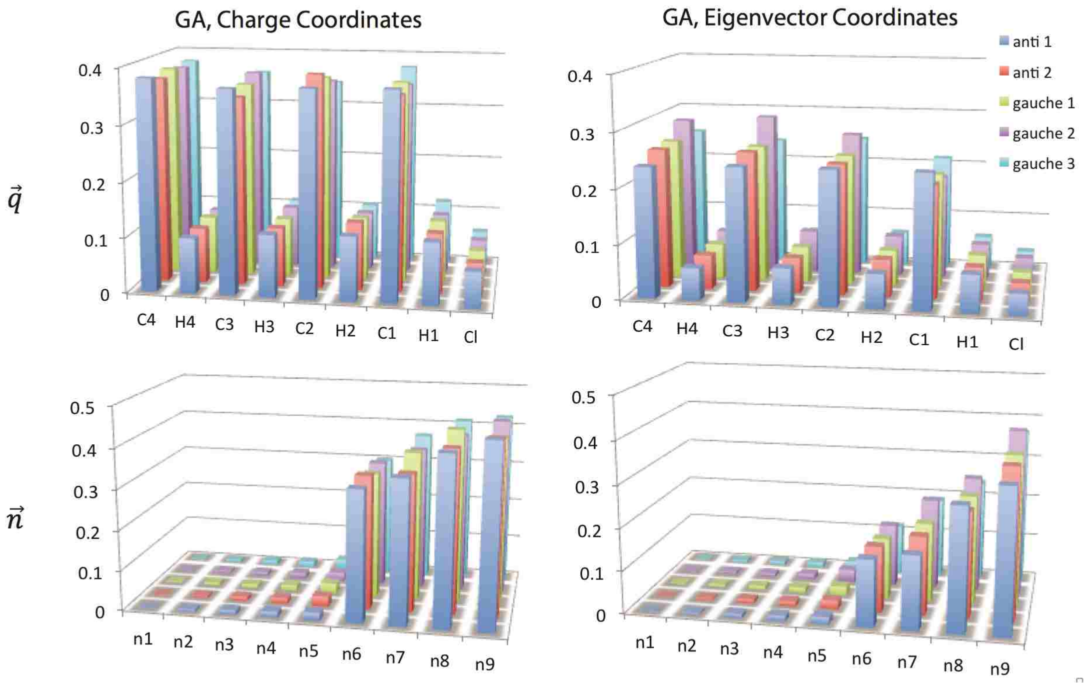


Figure 3.11: Standard deviations σ of the charges $\mathbf{q}$ and the corresponding coordinates defined by the LS-sum Hessian eigenvectors $\mathbf{n}$ obtained from the solutions of 200 GA performed in terms of the charge coordinates(left), and in terms of the LS-sum Hessian eigenvector coordinates (right).

depending on the grid setup, and often are highly inconsistent for even slightly different conformers of the same molecule; the variation is especially large for the carbon atoms of methyl and methylene groupsthe buried atom effect. These difficulties, commonly ascribed to the rank-deficient character of the LS matrix, can be understood in a new light once we recognize that the variation of the ESP solutions has the same underlying factors as the much larger scatter of the GA solutions. In fact, all LS fitting problems, not just the ESP, produce slightly different solutions from the LS matrices A that differ by the number of grid points, type of the grid, its density, etc. The covariance of these solutions, has been shown to be proportional to the inverse of the Hessian matrix:[111]

$$\operatorname{cov}(\mathbf{q}^{*}) \propto \mathbf{H}^{-1} = \left(\mathbf{A}^{\intercal}\mathbf{A}\right)^{-1} \tag{3.31}$$

Since a matrix inversion does not change the corresponding eigenvectors, this covariance matrix also shares the eigenbasis $\mathbf{U}$ with the covariance matrix of the

GA solutions. Thus, the variance/scatter of the ESP and GA solutions are related to the same fundamental properties of the LS-sum Hessian matrix, whose eigenvectors $\mathbf{h}_i$ define the natural, linearly independent coordinates for the MEP fitting problem. This provides a convenient framework to discuss the ill-conditioned nature of the ESP problem, and the buried atom effect associated with it. The numerical instabilities observed for the standard ESP implementations can be related to the LS-sum Hessian eigenvectors with the highest and the lowest curvatures. For any molecule, the first eigenvector defines the total charge coordinate and the curvature along this coordinate is orders of magnitude larger than the curvatures along other coordinates. Hence, a very strong total charge restraint is naturally built into the ESP problem. Nevertheless, most of the ESP implementations introduce an additional total charge constraint using Lagrange multipliers,a redundancy that leads to the known rank-deficiency of the resulting LS matrix. On the other hand, the vexing problem of the buried atoms arises as a natural consequence of the high-variance coordinates with curvatures many orders of magnitude smaller than the curvatures of the coordinates associated with the leading multipole moments. These low-curvature/high-variance coordinates have a small contribution to the MEP and do not significantly affect the overall fitness of a solution. Thus, several solutions can have very similar fitness scores because they have the same positions along the high-curvature coordinates, although their positions along the low-curvature coordinates could be quite different. Yet, these very similar solutions would appear very different when expressed in terms of the linearly dependent point-charge coordinates. Importantly, the lowest-curvature/highest-variation eigenvectors have the dominant contributions from the charges on the buried carbon atoms, as can be seen in the case of the CH_3X and CH_2X_2 molecules and the 1-chlorobutane conformers. As a result, these carbon atoms show the highest variation of the point charges, which is further amplified by the hydrogen/carbon stoichiometric ratios for the CH_3 and

CH_2 groups. The usual approach to prevent the wide variation of the ESP charges on buried the carbon atoms is to use additional restraints to keep these charges close to a predefined value, e.g. zero, or simply to constrain them to zero or some chemically reasonable value. This, however, can negatively affect the overall dipole moment values produced by the fitted point charges, as well as the overall quality of the fit; a better strategy may involve restraining or constraining the values along the low-curvature Hessian eigenmode coordinates.

3.8 Summary

In this chapter, motivated by the idea of using evolutionary approaches for the simultaneous optimizations of several types of force field parametersincluding point charges, we explored the performance of the genetic algorithm (GA) approach for a simpler problem of point-charge fitting against the reference molecular electrostatic potential (MEP). We find that unless unreasonably large population sizes are used, the GA optimizations produce highly scattered, but correlated, solutions. Analysis of the covariance matrices for these scattered sets of GA solutions revealed a remarkable correspondence between the covariance matrices and the fitness function Hessian matrix, which share the same set of the eigenvectors. This eigenbasis represents a linearly independent set of coordinates that are natural for the MEP point-charge fitting problem, unlike the linearly dependent point charge coordinates. Some of the Hessian/covariance matrix eigenvectors define the coordinates related to the leading terms of the multipole expansion (the total charge/monopole, dipole moment components); these coordinates are associated with high curvature of the fitness function and thus negligible variation of the GA solutions. On the other hand, other eigenvectors are associated with negligible fitness function curvatures and thus large variance. The huge disparity between the curvatures of the Hessian eigenvector coordinates causes premature convergence of the GA optimizations performed in terms of the linearly dependent point-charge coordinates, because of the high fitness penalty for even a slight deviation from the minimum along the high-curvature direction that effectively prevents the GA population from exploring the fitness profile along the low-curvature direction. This leads to a variety of GA solutions with highly scattered point charge values and moderately low, but not always optimal fitness scores. The severe scatter of the GA solutions can be seen as an exaggerated version of the well-known buried atom effect, the variation of the ESP charges of the buried carbon atoms observed for different grid setups and/or for different conformers. This effect arises from the coordinates defined by the low-curvature Hessian eigenvectors and the fact that the point charges are inappropriate, highly linearly dependent (and also redundant)96 coordinates for the MEP fitting problem. Thus, MEP fitting in coordinates defined by the fitness function/LS-sum Hessian eigenbasis is essential when using evolutionary methods. In this respect, the most promising approach is to take advantage of the correspondence between the eigenvectors of the covariance matrix of the solutions and the fitness function Hessian matrix, as it is done in advanced evolutionary techniques such as covariance matrix adaptation evolution strategy (CMA-ES). Besides not being proper quantum mechanically observed parameters, atom-centered point charges are not even proper variables for the classical MEP fitting problem. At the same time, the simplicity and efficiency of the point charge model ensures its continuing survival in the field of the biomolecular simulations, at least in the short term.[11,70,71,148] Thus, the insights revealed by the analysis of the GA performance for the point charge fitting problem could prove useful for the further development and optimization of the biomolecular force fields using evolutionary methods, as well as other optimization techniques.

3.9 Computational Details

All geometry optimizations were performed at the B3LYP/aug-cc-pVDZ level,[149–151] as implemented in Gaussian 09 package.[152] Reference MEPs were generated as cubic grids with linear density of 1.5 points/Å, followed by removal of the points outside 1.42.0 van der Waals radii range around each atom. This sampling procedure covers the solvent-accessible region of the molecule, in line with common charge fitting procedures.[36,37]

All charge-fitting procedures were implemented using Python programming language within fftoolbox and *genetica* modules with the source code available online. The *fftoolbox* module extracts molecular geometry and the reference electrostatic potential from the Gaussian cube file and performs calculation of the LS sum over the points in the grid. Besides the atom-centered point charges, *fftoolbox* also supports the optimization of the extra points placed out of the atomic centers. The ESP method is implemented as a part of *fftoolbox* with the normal equation solved using *numpy* library.[153] GA optimization routines are implemented in the *genetica* module using either binary and real-number chromosome representation. The point charge optimization can be performed in three coordinate systems: point charges, multipole moments, or in the eigenbasis of the LS-sum Hessian matrix. Besides a single-objective minimization, *genetica* also supports vector-valued FFs using Vector Evaluated GA (VEGA)an extension of the single-objective GA method to support multi-objective optimizations. Covariance Matrix Adaptation Evolution Strategy (CMA-ES) optimizations were performed using *cma* Python library;[115,116,147] in these optimizations, all values of the initial solution were set to zero and the initial standard deviation was set to 0.1. Covariance matrix calculations as well as all matrix eigendecompositions were performed using *numpy* library. Graphical representation of the results is supported by *matplotlib* library.[154]

Chapter 4

Revealing the Ill-Conditioning of the Charge Fitting Problem

4.1 Introduction

The atom-centered point charge (PC) model of molecular electrostatics has been a mainstay of biomolecular simulations for decades.[32,33,36,44,49,71,122,142,143,155–159] While chemically intuitive and straightforward in technical implementation, this model does not provide a sufficiently detailed description of the anisotropic features of the molecular electrostatic potential (MEP), such as lone pairs, π-systems, and σ-holes, etc. which are mostly governed by higher-order multipole terms.[53,54] These anisotropic effects, however, can be described within the PC approximation by moving beyond the atom-centered paradigm, i.e. by adding non-atom centered PCs/extended points.[60,61,70,160] Although increasing the number of PCs per atom improves the quality of the electrostatic model, it also can exacerbate well-known ill-conditioning and redundancy problems[37,38,140] of the PC fitting procedures, leading to numerically unstable solutions.[36,161,162]

These numerical instabilities are usually related to a large variation of the PC values for atoms in the interior of the molecule, so-called buried atom effect.[36,39,40,142] The buried atom (usually methyl and methylene carbons) charges can dramatically change due to trivial changes in the PC fitting problem (the probe grid sampling, spatial orientation of the molecule, etc.), and/or have inconsistent values across very similar molecules or even conformers of the same molecule.[35,137] As the inclusion of non-atom centered PCs into the model produces even more buried centers, it should also increase the numerical instabilities of the PC fitting problem.

In fact, these numerical problems are rooted in the mathematical nature of the PC derivation—the least squares (LS) fitting to the reference MEP:[32,33]

$$\chi^2(\mathbf{q}) = |\mathbf{\Phi} - \mathbf{A}\mathbf{q}|^2 = |\mathbf{\Phi}|^2 + \mathbf{g}^\intercal \cdot \mathbf{q} + \mathbf{q}^\intercal \mathbf{H}\mathbf{q}, \tag{4.1}$$

$$\mathbf{g} = -2\mathbf{A}^\intercal \mathbf{\Phi}, \tag{4.2}$$

$$\mathbf{H} = \mathbf{A}^\intercal \mathbf{A}, \tag{4.3}$$

where the LS sum χ^2 is the subject of minimization and the solution satisfies normal equations:[144]

$$\mathbf{A}^\intercal \mathbf{A}\mathbf{q} = \mathbf{A}^\intercal \mathbf{\Phi}. \tag{4.4}$$

Here, the elements of the LS matrix **A** correspond to the inverse distance $1/r_{ij}$ between the PC i and the grid point j; $\mathbf{\Phi}$ is T-dimensional vector of the reference values of MEP; $\mathbf{q}$ is N-dimensional vector of the PC values; $\mathbf{g}$ is the gradient of the function χ^2 at the origin ($\mathbf{q} = 0$); $\mathbf{H}$ is the Hessian matrix of LS sum χ^2.

While the ill-conditioning is common to many LS fitting problems,[163–165] numerical difficulties associated with PC fitting are further compounded by commonly used total charge constraint using Lagrange multiplier.[34,37,38,48]

One of the most widely used techniques to alleviate the numerical instabilities of PC fitting is to add artificial restraints to the PC values of the buried atoms.[36,41,43,49] Although this method can be extended to models with off-center PCs/extended points, one may wonder if it would be possible to overcome these difficulties in a more elegant way, based on better physical understanding of the problem.

For instance, an important insight can be gleaned from the eigendecomposition of the LS sum Hessian matrix (eq. 4.3):

$$\mathbf{H}\mathbf{u}_i = \kappa_i \mathbf{u}_i. \tag{4.5}$$

Indeed, the ill-conditioned nature of the LS matrix $\mathbf{A}$ can be related to the significant differences in the eigenvalues κ_i, i.e. the LS sum curvatures along the directions defined by the eigenvectors $\mathbf{u}_i$.[111,166] Because of the 2–3 order of magnitude variation of the κ_i values, different sets of PCs can produce essentially the same MEP, as these solutions have the same positions along the high-curvature directions, although the positions along the low-curvature directions could be quite different.[166] Importantly, the eigenvectors with the largest curvatures usually correspond to the total charge and dipole moment components of the molecule, while the lower-curvature eigenvectors do not seem to be associated with particular multipole moments.[166,167]

However, the exact physical origin of the correspondence between the large curvature eigenvectors and the first terms of the multipole expansion is unclear, along with the nature of the low-curvature eigenvectors. Particularly, it is not clear if the presence of the low-curvature modes of the $\mathbf{H}$ matrix and thus the ill-conditioning of the LS problem is solely because of the nature of the PC fitting problem, or due to some numerical factors, e.g. an incomplete sampling of the reference MEP grid.

To address these questions, in this chapter we revisit the PC fitting problem from the first principles. While the atom-centered PC model traces back to the intuitive chemical concept of the atomic charge, we consider a general PC model as a case of the inverse problem, where one seeks to recover the source charge distribution from its effect, i.e. electrostatic potential distribution. Based on the properties of the Coulomb law, we construct a best-case electrostatic model for which the inverse problem can be solved exactly, both in the continuous case, as well as in the case of a discrete (non-atom centered) PC approximation.

Using this model, we investigate the nature of the eigenvectors $\mathbf{u}_i$ and their eigenvalues κ_i, and dissect the factors responsible for the ill-conditioning of the LS fitting problem, and discuss how these insights can be used to improve and simplify the existing PC derivation procedures.

4.2 Point Charge Fitting as an Inverse Problem

A problem where given an effect (in this case the MEP Φ) defined in the region V_Φ, its cause (a charge distribution ρ) defined in the region V_ρ needs to be determined belongs to a general class of inverse problems and can be described by the Fredholm integral equation of the first kind:[168]

$$\int_{V_\rho} k(\mathbf{r},\mathbf{r}')\rho\left(\mathbf{r}'\right)d\mathbf{r}' = \Phi\left(\mathbf{r}\right), \tag{4.6}$$

where kernel $k(\mathbf{r},\mathbf{r}')$ specifies the evolution of the cause $\rho(\mathbf{r}')$ into the effect $\Phi(\mathbf{r})$, that in this case corresponds to the Coulomb law:

$$k(\mathbf{r},\mathbf{r}') = \frac{1}{|\mathbf{r}-\mathbf{r}'|}. \tag{4.7}$$

The integral equation can also be represented as an operator equation:

$$K\rho = \Phi, \tag{4.8}$$

where $K : U \to V$ is a linear operator defined on space $U = \text{range}(K^*) \in L^2$ of square integrable functions, and takes values in space $V = \text{range}(K) \in L^2$; $K^* : V \to U$ is adjoint of K. This equation can be solved exactly if and only if $\Phi \in V$. However, in general it is not the case, so a function ρ that minimizes the residual norm $|\Phi - K\rho|$ is considered as the LS solution and thus satisfies the normal equation:[108,168]

$$K^*K\rho = K^*\Phi. \tag{4.9}$$

This LS solution can be obtained as the linear combination of the basis vectors $u_i \in U$:[168]

$$\rho = K^\dagger\Phi = \sum_{i=1}^{\infty}\frac{\langle\Phi, v_i\rangle}{\mu_i}u_i, \tag{4.10}$$

where $K^\dagger$ is the Moore-Penrose inverse, μ_i is a singular value, v_i and u_i are left and right singular vectors, respectively and the inner product $\langle \Phi, v_i \rangle$ is defined as

$$\langle \Phi, v_i \rangle = \int_{V_\Phi} \Phi(\mathbf{r}) v_i(\mathbf{r}) d\mathbf{r} \tag{4.11}$$

The orthogonal bases $\{u_i\}_{i=1}^{\infty}$ and $\{v_i\}_{i=1}^{\infty}$ also form the eigenbases of K^*K and KK^* with eigenvalues μ_i^2:

$$K^*Ku_i = \mu_i^2 u_i, \tag{4.12}$$

$$KK^*v_i = \mu_i^2 v_i. \tag{4.13}$$

To obtain a numerical solution to the integral equation (eq. 4.6), the regions over which the MEP and charge distribution are defined are sampled using a numerical quadrature. Given N quadrature nodes over the charge distribution and T nodes over the MEP region the integral equation is transformed into a system of T linear equations:

$$\mathbf{Kq} = \mathbf{\Phi}, \tag{4.14}$$

where the $T \times N$ matrix $\mathbf{K}$ is identical to the LS matrix $\mathbf{A}$ from eq. 4.1 and contains the kernel elements k_{ij}, as this matrix originates from the kernel $k(\mathbf{r}, \mathbf{r}')$ in the integral equation (eq. 4.6). It will be further referred to as $\mathbf{K}$ in order to highlight its mathematical origin.

Then, the PC value at the node i is the product of the charge density ρ_i and the quadrature weight w_i:

$$q_i = \rho_i w_i \tag{4.15}$$

Since the number of the reference values T is usually larger than the number of the unknown PC values N, the system of linear equations is overdetermined. Then, a solution that minimizes the LS sum $\chi^2(\mathbf{q})$ (eq. 4.1) and satisfies normal equations (eq. 4.4) is considered as the numerical solution to the integral equation (eq. 4.6). This solution can be obtained using singular value decomposition (SVD) of matrix $\mathbf{K}$:[108,111,144]

$$\mathbf{q} = K^{\dagger}\mathbf{\Phi} = \sum_{i=1}^{r} \frac{\mathbf{\Phi} \cdot \mathbf{v}_i}{\mu_i}\mathbf{u}_i, \tag{4.16}$$

where $K^{\dagger}$ is the Moore-Penrose pseudoinverse; μ_i are singular values of matrix $\mathbf{K}$; vectors $\mathbf{v}_i$ and $\mathbf{u}_i$ are left and right singular vectors. If the rank r of matrix $\mathbf{K}$ is less than the dimension of $\mathbf{q}$ ($r < N$), then the matrix $\mathbf{K}$ is rank deficient.

Similarly to the continuous case (eqs. 4.12-4.13), the orthogonal bases $\{v_i\}_{i=0}^{r}$ and $\{u_i\}_{i=0}^{r}$ form eigenbases for $\mathbf{K}\mathbf{K}^{\intercal}$ and $\mathbf{K}^{\intercal}\mathbf{K}$:

$$\mathbf{K}\mathbf{K}^{\intercal}\mathbf{v}_i = \mu_i^2\mathbf{v}_i, \tag{4.17}$$

$$\mathbf{K}^{\intercal}\mathbf{K}\mathbf{u}_i = \mu_i^2\mathbf{u}_i, \tag{4.18}$$

where $\mathbf{K}^{\intercal}\mathbf{K}$ is also a Hessian matrix (eq. 4.5) and μ_i^2 is identical to its eigenvalue κ_i, which is the χ^2 curvature along the direction $\mathbf{u}_i$:[169]

$$\mu_i^2 = \kappa_i \tag{4.19}$$

In many LS problems, PC fitting included, the singular values vary in a wide range, revealing the underlying ill-conditioning.[37,38,163,164] As a singular value μ_i is a denominator in the LS solution (eq. 4.16), the smaller the singular value, the larger the effect of the corresponding singular vector $\mathbf{u}_i$ on the LS solution. Thus, even small variations along $\mathbf{u}_i$ with small singular value lead to a significant variations of the LS solution, although these variations do not lead to significant change in the quality of the fit χ^2.[166] To understand the origins of the ill-conditioning in PC fitting, we next consider a system for which the inverse electrostatic problem can be analytically solved.

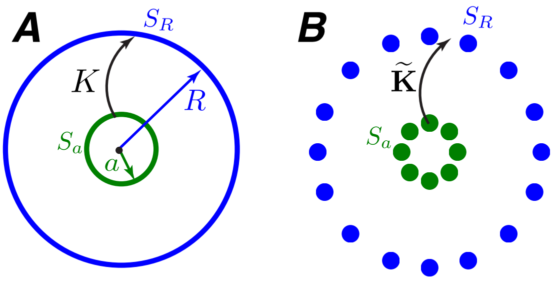


Figure 4.1: Schematic representations of the probe S_R and charged S_a spheres in the continuous (A) and discrete (B) forms. Operators K (eq. 4.28) and matrix $\widetilde{\mathbf{K}}$ are represented schematically.

4.3 The Two-Sphere Model

The Coulomb kernel (eq. 4.7) can be conveniently expanded in terms of spherical harmonics so the source $\mathbf{r}'$ and the observation $\mathbf{r}$ coordinates are separated but share the same origin:[1,107]

$$\begin{aligned} k(\mathbf{r},\mathbf{r}') &= \frac{1}{|\mathbf{r}-\mathbf{r}'|} \\ &= \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\frac{4\pi}{2l+1}\frac{r_<^l}{r_>^{l+1}}Y_{lm}(\widehat{\mathbf{r}'})Y_{lm}(\widehat{\mathbf{r}}), \end{aligned} \tag{4.20}$$

where $\widehat{\mathbf{r}}=\mathbf{r}/r$ denotes the unit vector defined by the polar φ and azimuthal θ angles; $r_<$ is the smaller and $r_>$ is the larger of r and r'; Y_{lm} are orthogonal real-value spherical harmonics:[1]

$$\int_S Y_{lm}(\widehat{\mathbf{r}})Y_{l'm'}(\widehat{\mathbf{r}})d\Omega = \delta_{ll'}\delta_{mm'}, \tag{4.21}$$

where $d\Omega$ is the differential of the solid angle.

Then, in the region beyond the divergence sphere where the charge density vanishes, the MEP can be expanded in a multipole series:[1,107]

$$\Phi(\mathbf{r}) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\sqrt{\frac{4\pi}{2l+1}}r^{-l-1}Q_{lm}^{mol}Y_{lm}(\widehat{\mathbf{r}}), \tag{4.22}$$

---

[1] For practical purpose, we use real-valued spherical harmonics, thus to compact the derivation and not to obscure the main idea of the work Y_{lm} with, $m<0$ ($m>0$) corresponds to Y_{lms} (Y_{lmc}) in standard notation[1]

where a molecular multipole moment Q_{lm}^{mol} is given by

$$Q_{lm}^{mol} = \sqrt{\frac{4\pi}{2l+1}} \int r^l \rho(\mathbf{r}) Y_{lm}(\widehat{\mathbf{r}}) d^3 r. \tag{4.23}$$

The form of the kernel expansion (eq.4.20) suggests that if the radii $r = R$ and $r' = a$ are fixed, the kernel $k(\mathbf{R}, \mathbf{a})$ can uniquely map a charge density over a spherical surface S_a to the corresponding potential $\Phi(\mathbf{R})$ on a sphere S_R and vice versa. Thus, for a probe sphere S_R with the radius R greater than the radius of divergence sphere, the MEP can be reproduced exactly by a sphere S_a with surface charge density $\sigma(\mathbf{a})$ such that the multipole moments of the sphere $Q_{lm}^{S_a}$ are equivalent to the multipole moments of the molecule Q_{lm}^{mol}:

$$Q_{lm}^{S_a} \equiv Q_{lm}^{mol}, \tag{4.24}$$

where multipole moments of the sphere are:

$$Q_{lm}^{S_a} = \sqrt{\frac{4\pi}{2l+1}} a^l \int_{S_a} \sigma(\mathbf{a}) Y_{lm}(\widehat{\mathbf{a}}) d\Omega. \tag{4.25}$$

In this case, the original integral eq. 4.6 is transformed into a surface integral equation:

$$\int_{S_a} k(\mathbf{R}, \mathbf{a}) \sigma(\mathbf{a}) d\Omega = \Phi(\mathbf{R}), \tag{4.26}$$

or, equivalently, in an operator form

$$K\sigma = \Phi, \tag{4.27}$$

where $K : L^2(S_a) \to L^2(S_R)$ is a compact infinite-rank operator (Figure 4.1A):

$$K\sigma = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \mu_l \langle \sigma, Y_{lm}^{S_a} \rangle Y_{lm}^{S_R}, \tag{4.28}$$

where subscripts S_a and S_R denote the spheres, on which the corresponding spherical harmonics are defined; the projection $\langle \sigma, Y_{lm}^{S_a} \rangle$ is the inner product on the $L^2(S_a)$ space:

$$\langle \sigma, Y_{lm}^{S_a} \rangle = \int_{S_a} \sigma(\mathbf{a}) Y_{lm}^{S_a}(\widehat{\mathbf{a}}) d\Omega \tag{4.29}$$

and for each degree l there is a singular value μ_l in the form of the distance-dependent factor from the MEP expansion (eq. 4.20):

$$\mu_l = \frac{4\pi}{2l+1} \frac{a^l}{R^{l+1}}. \tag{4.30}$$

Accordingly, the spherical harmonics $Y_{lm}^{S_R}$ and $Y_{lm}^{S_a}$ are left and right singular vectors and thus the eigenfunctions of the operators K^*K and KK^*, while the squares of the singular values μ_l are their eigenvalues (eqs. 4.12-4.13). Since the singular values μ_l and spherical harmonics $Y_{lm}^{S_a}$ and $Y_{lm}^{S_R}$ form a singular system of the operator K, the solution to integral equation (eq. 4.26) can be expressed as:

$$\sigma = K^\dagger \Phi = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \frac{\langle \Phi, Y_{lm}^{S_R} \rangle}{\mu_l} Y_{lm}^{S_a}. \tag{4.31}$$

According to the multipole expansion (eq. 4.22), the inner product $\langle \Phi, Y_{lm}^{S_R} \rangle$ depends on the radius R of the probe sphere and the multipole moments of the molecule:

$$\langle \Phi, Y_{lm}^{S_R} \rangle = \int_{S_R} \Phi(\mathbf{R}) Y_{lm}^{S_R}(\widehat{\mathbf{R}}) d\Omega = \sqrt{\frac{4\pi}{2l+1}} \frac{1}{R^{l+1}} Q_{lm}^{mol}. \tag{4.32}$$

The dependence on the radius R cancels out, so the charge density depends only on the radius a of the sphere S_a and the molecular multipole moments:

$$\sigma(\mathbf{a}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \sqrt{\frac{2l+1}{4\pi}} a^{-l} Y_{lm}^{S_a}(\widehat{\mathbf{a}}) Q_{lm}^{mol}, \tag{4.33}$$

and the charged sphere S_a exactly reproduces the MEP $\Phi(\mathbf{R})$.

4.4 Analytical Point Charge Model

We can construct an approximate discrete analog of the two-sphere model (eqs. 4.26-4.33, Figure 4.1) using a quadrature that exactly integrates spherical harmonics Y_{lm} over a sphere up to a given l (eqs. 4.29 and 4.32), e.g. the widely used[170–172] Lebedev quadrature,[173] that defines N quadrature nodes (Figure 4.4) with predetermined angular coordinates θ_i, φ_i, and integration weights w_i:

$$\int_S Y_{lm}(\theta, \varphi) d\Omega = \sum_i^N Y_{lm}(\theta_i, \varphi_i) w_i. \tag{4.34}$$

Then, given the surface charge density σ_i the corresponding point charge is:

$$q_i = \sigma_i w_i. \tag{4.35}$$

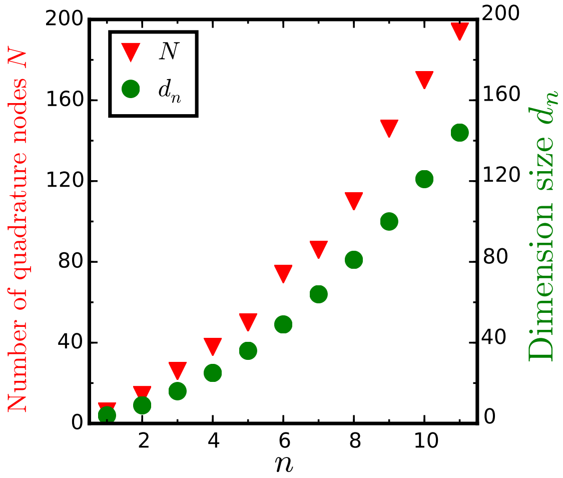


Figure 4.2: Number of Lebedev quadrature points N (red triangles) and dimension $d_n = (n + 1)^2$ (green circles) as functions of the degree n

Due to the orthogonality of the spherical harmonics Y_{lm} (eq. 4.21), the N-node Lebedev quadrature that exactly integrates spherical harmonics over the sphere S_a up to $l = 2n$

$$\sum_i^N Y_{lm}^{S_a}(\theta_i, \varphi_i) Y_{l'm'}^{S_a}(\theta_i, \varphi_i) w_i^{S_a} = \widetilde{\mathbf{Y}}_{lm}^{S_a} \cdot \widetilde{\mathbf{Y}}_{l'm'}^{S_a} = \delta_{ll'} \delta_{mm'}, \tag{4.36}$$

defines an orthonormal basis of dimension $d_n = (n+1)^2$:

$$\widetilde{\mathbf{Y}}_{S_a} = \{\ \widetilde{\mathbf{Y}}^{S_a}_{lm},\ -l \le m \le l\ \}_{l=0}^{n}, \tag{4.37}$$

where the $\widetilde{\mathbf{Y}}^{S_a}_{lm}$ vectors have N elements defined as:

$$\widetilde{Y}^{S_a}_{lmi} = Y_{lm}(\theta_i, \varphi_i)\sqrt{w_i^{S_a}}. \tag{4.38}$$

Similarly, the probe sphere S_R can be represented by a T-node Lebedev grid that integrates spherical harmonics up to $l = 2t$ and defines an orthogonal basis $\widetilde{\mathbf{Y}}_{S_R}$ of dimension $d_t = (t+1)^2$.

In this discrete representation, the operator K (eq. 4.28) then becomes a $T \times N$ matrix $\widetilde{\mathbf{K}}$:[174]

$$\widetilde{\mathbf{K}}\ \widetilde{\boldsymbol{\sigma}}\ = \widetilde{\boldsymbol{\Phi}}, \tag{4.39}$$

where the elements of $\widetilde{\mathbf{K}}$, $\widetilde{\boldsymbol{\sigma}}$, and $\widetilde{\boldsymbol{\Phi}}$ are:

$$\widetilde{K}_{ij} = \sqrt{w_i^{S_a} w_j^{S_R}}/r_{ij}, \tag{4.40}$$

$$\widetilde{\sigma}_i = \sigma_i \sqrt{w_i^{S_a}}, \quad \widetilde{\Phi}_j = \Phi_j \sqrt{w_j^{S_R}}. \tag{4.41}$$

Since usually the probe grid has more points than the source grid, i.e. $T > N$, the matrix equation (eq. 4.39) is a LS problem (eq. 4.1) that can be solved using SVD of the matrix $\widetilde{\mathbf{K}}$ (eq. 4.16), giving a discrete analog of eq. 4.33:

$$\widetilde{\boldsymbol{\sigma}} = \sum_{l=0}^{n} \sum_{m=-l}^{l} \frac{\widetilde{\boldsymbol{\Phi}} \cdot \widetilde{\mathbf{Y}}^{S_R}_{lm}}{\mu_l} \widetilde{\mathbf{Y}}^{S_a}_{lm}, \tag{4.42}$$

where $\widetilde{\mathbf{Y}}^{S_R}_{lm}$ and $\widetilde{\mathbf{Y}}^{S_a}_{lm}$ are left and right singular vectors, and the corresponding singular values μ_l are the same as in the continuous case (eq. 4.30).

Since we use the Lebedev quadrature, the dot product $\widetilde{\boldsymbol{\Phi}} \cdot \widetilde{\mathbf{Y}}^{S_R}_{lm}$ corresponds to exact numerical integration and gives a result identical with the continuous case (eq. 4.32):

$$\widetilde{\mathbf{\Phi}} \cdot \widetilde{\mathbf{Y}}^{S_R}_{lm} = \sum_{j=0}^{T} \Phi_j Y^{S_R}_{lmj} w_j = \sqrt{\frac{4\pi}{2l+1}} \frac{1}{R^{l+1}} Q^{mol}_{lm}, \tag{4.43}$$

so the solution to eq. 4.39 depends only on the radius a and the multipole moments Q^{mol}_{lm}:

$$\widetilde{\boldsymbol{\sigma}} = \sum_{l=0}^{n} \sum_{m=-l}^{l} \sqrt{\frac{2l+1}{4\pi}} a^{-l} Q^{mol}_{lm} \widetilde{\mathbf{Y}}^{S_a}_{lm}. \tag{4.44}$$

The corresponding PC values q_j can be obtained using the quadrature weights $w^{S_a}_j$:

$$q_i = \sigma_i w^{S_a}_i = \widetilde{\sigma}_i \sqrt{w^{S_a}_i}, \tag{4.45}$$

or, in a vector form:

$$\mathbf{q} = \sum_{l=0}^{n} \sum_{m=-l}^{l} \sqrt{\frac{2l+1}{4\pi}} a^{-l} Q^{mol}_{lm} \mathbf{Y}^{S_a}_{lm} \odot \mathbf{w}^{S_a}, \tag{4.46}$$

where $\mathbf{w}^{S_a}$ is the vector of the quadrature weights for the sphere S_a. Therefore, we can use Lebedev grid that shares the origin with a molecule to construct an analytical PC model that exactly reproduces molecular multipole values up to the degree n.

From this model, we can see that the ill-conditioning of the PC fitting due to the decay of the singular values is intrinsic to the inverse electrostatic problem, as the singular values μ_l decrease with increasing l (eq. 4.30). Indeed, the higher the multipole moment, the smaller its contribution to the overall electrostatic potential. Also, this contribution gets smaller as we move the probe further away from the source, and the singular values get smaller with the increasing radius of the probe sphere R, or decreasing radius of the source sphere a.

The ill-conditioning problems become even more severe as we switch from modeling the MEP using the Lebedev quadrature, which is the best suited to reproduce the molecular multipoles, to an irregular atom-centered quadrature, as shown on a numerical example below.

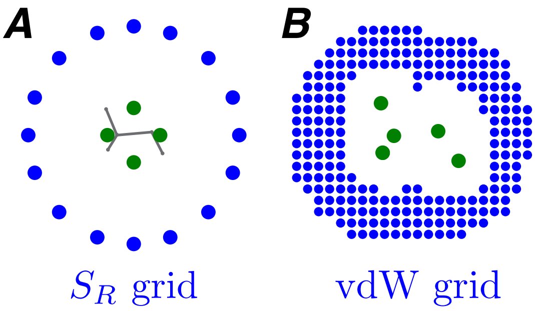


Figure 4.3: Cross-section representations of the quadratures used for two-sphere model (A) ($n = 1$, $N = 6$ and $t = 11$, $T = 194$ for spheres S_a and S_R, respectively) as compared with the traditional atom-centered model (B). Green circles correspond to the point charges; blue circles correspond to the reference grid points

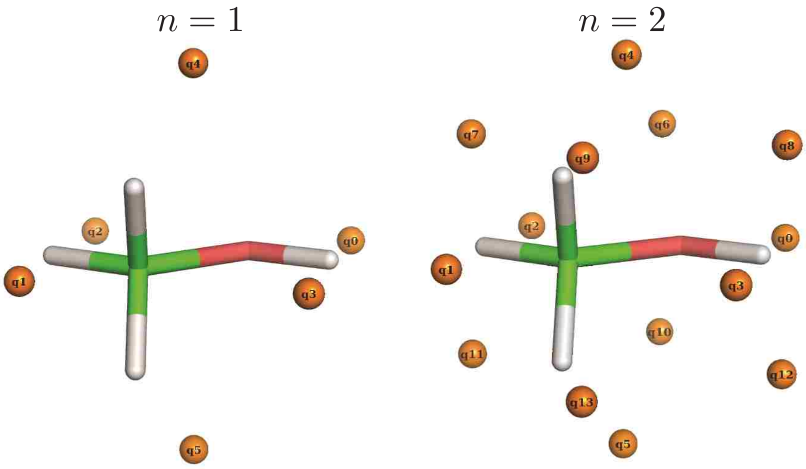


Figure 4.4: PC representation of the charged sphere S_a using the Lebedev quadrature with $n = 1$ and $n = 2$ as compared with the geometry of methanol molecule.

4.5 Lebedev vs. Atom-Centered Model: Numerical Example

First, we consider an electrostatic PC model of a methanol molecule with PCs placed at the nodes of the Lebedev quadrature over the sphere S_a ($a = 2$ au) (Figures 4.3A and 4.4). In this case, the PC values can be obtained analytically from the reference multipole moments (eq. 4.46) or by numerical fitting to the reference MEP over the probe sphere S_R ($R = 8$ au, $t = 11$, $T = 194$):

$$\widetilde{\boldsymbol{\sigma}} = \sum_{i=1}^{r} \frac{\widetilde{\boldsymbol{\Phi}} \cdot \widetilde{\mathbf{v}}_i}{\mu_i} \widetilde{\mathbf{u}}_i, \tag{4.47}$$

where the PC value can be found as $q_j = \widetilde{\sigma}_j \sqrt{w_j^{S_a}}$ and the maximum rank r is the number N of quadrature nodes/PCs over the sphere S_a. The quality of the fit is measured using the root mean square deviation (RMSD) calculated over the T nodes of the probe grid:

$$\mathrm{RMSD} = \sqrt{\frac{\chi^2}{T}}. \tag{4.48}$$

Naturally, the analytical PC values from eq. 4.46 exactly reproduce the molecular multipole moments up to the degree n defined by the quadrature (Table 4.2). For each degree l there are $2l + 1$ values of order m, so overall $(n + 1)^2$ multipole moments are reproduced, which matches the dimension d_n of the corresponding basis $\widetilde{\mathbf{Y}}_{S_a}$ (eq. 4.37). As the dimension d_n increases (i.e number of nodes T), more multipole moments are reproduced and the RMSD rapidly approaches zero (Figure 4.5).

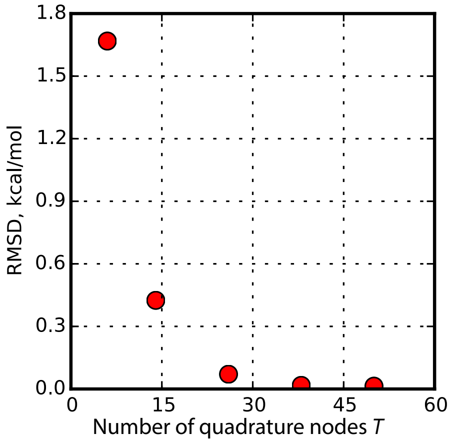


Figure 4.5: RMSD as the function of the number N of PCs on the charged sphere S_a

Since the dimension d_n does not match the number of quadrature nodes N (Figure 4.5),[175,176] we can obtain numerical solutions with eq. 4.47 that are equivalent to the analytical results (eq. 4.46) by setting the rank r to the dimension of the grid, $d_n = (n + 1)^2$ (Table 4.2). Note, that the slight differences in the resulting PC and multipole values obtained with the two methods arise due to the finite radius R of the probe sphere S_R used in the numerical approach. As R increases, the probe sphere S_R entirely encompasses the molecular charge density, and the multipole moments of the charged sphere $Q_{lm}^{S_a}$ converge to the true molecular multipole moments Q_{lm}^{mol} (Figure 4.5).

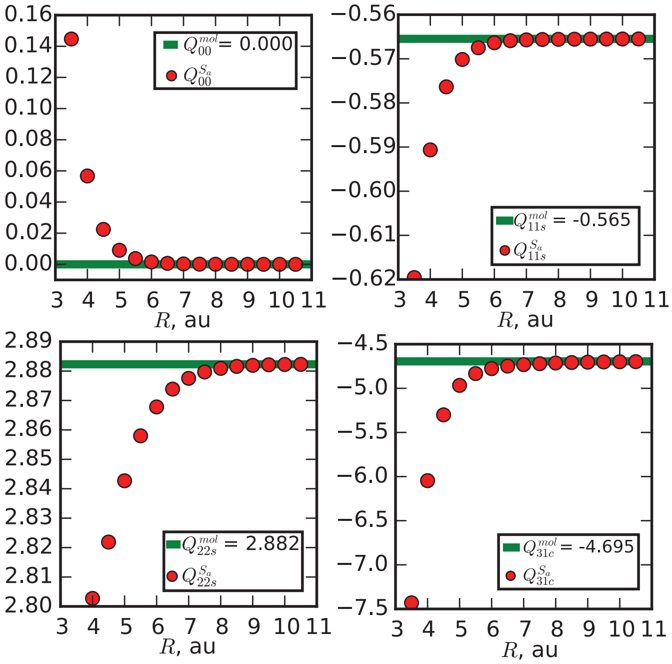


Figure 4.6: Few selected multipole moments $Q_{lm}^{S_a}$ of the charged sphere S_a as the functions of the probe sphere radius R.

As the first d_n multipole moments Q_{lm}^{mol} are reproduced by the PC model, the first d_n numerical singular values μ_i exactly match the radius-dependent part (eq. 4.30) from the inverse distance expansion (Figure 4.7), and the corresponding right singular vectors $\widetilde{\mathbf{u}}_i$ match the basis $\widetilde{\mathbf{Y}}_{S_a}$ (Figure 4.8):

$$\{\widetilde{\mathbf{u}}_i\}_{i=1}^{d_n} = \{\widetilde{\mathbf{Y}}_{lm}^{S_a},\ -l \leq m \leq l\}_{l=0}^{n}. \tag{4.49}$$

If we do not restrict the rank r to the dimension of the grid d_n, numerical SVD of the LS matrix $\widetilde{\mathbf{K}}$ (eq. 4.47) produces N singular vectors/values. While this slightly improves the RMSD (Table 4.2), the additional $N - d_n$ singular vectors cannot be described analytically (Figure 4.8), as they go beyond the dimension d_n of the corresponding basis $\widetilde{\mathbf{Y}}_{S_a}$. However, in the fortuitous case of the quadrature with $n = 1$ and $N = 6$, the remaining $6 - 4 = 2$ vectors resemble the basis vectors $\widetilde{\mathbf{Y}}_{2-2}$ and $\widetilde{\mathbf{Y}}_{2-1}$, so the corresponding quadrupole moments

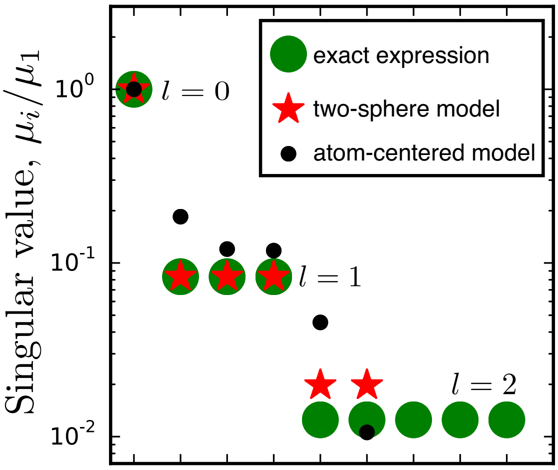


Figure 4.7: Normalized singular values μ_i/μ_1 obtained using the exact analytical expression eq. 4.30 (green circles) as compared with the numerical values obtained from SVD of the LS matrix for the two-sphere model (red stars) and for atom-centered model (black circles). Lebedev quadratures with $n = 1$, $N = 6$ and $t = 11$, $T = 194$ were used for the charged S_a ($a = 2$ au) and probe S_R ($R = 8$ au) spheres, respectively.

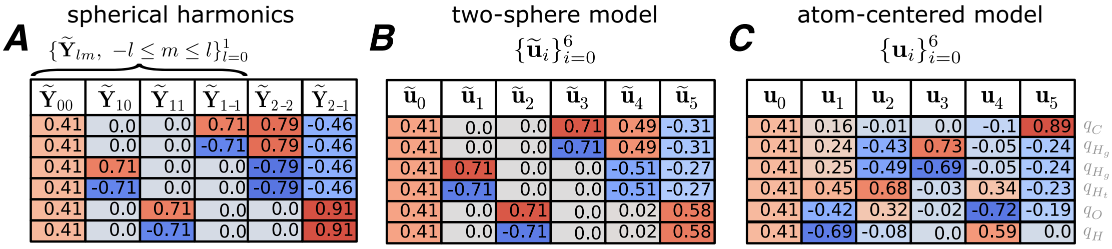
Q_{2-2} and Q_{2-1} are accurately reproduced, although the exact numerical integration of the spherical harmonics Y_{2-2} and Y_{2-1} is not provided by the 6-node Lebedev grid.

Now, we can use the insights from the best-case scenario spherical PC model based on the Lebedev quadrature (Figure 4.3A) to understand the traditional PC fitting problem with atom-centered charges and the probe grid that follows the solvent-accessible surface (vdW grid, Figure 4.3B). From the point of view of the inverse electrostatic model, the atom-centered PC fitting corresponds to a numerical solution using an irregular and suboptimal integration grid to represent the source charge distribution. This problem can be treated by SVD of the LS matrix $\mathbf{K}$:

$$\mathbf{q} = \sum_{i=1}^{r} \frac{\mathbf{\Phi} \cdot \mathbf{v}_i}{\mu_i} \mathbf{u}_i, \tag{4.50}$$

where the maximum value of rank r is the number of atoms in the molecule, i.e. $r = 6$ in the case of methanol.

We can see that even in this case the singular vector $\mathbf{u}_1$ with the largest singular value μ_1 corresponds the total charge (Figure 4.8), which is reproduced



A spherical harmonics

$\{\widetilde{\mathbf{Y}}_{lm},\ -l \le m \le l\}_{l=0}^{1}$

$\widetilde{\mathbf{Y}}_{00}$	$\widetilde{\mathbf{Y}}_{10}$	$\widetilde{\mathbf{Y}}_{11}$	$\widetilde{\mathbf{Y}}_{1\text{-}1}$	$\widetilde{\mathbf{Y}}_{2\text{-}2}$	$\widetilde{\mathbf{Y}}_{2\text{-}1}$
0.41	0.0	0.0	0.71	0.79	-0.46
0.41	0.0	0.0	-0.71	0.79	-0.46
0.41	0.71	0.0	0.0	-0.79	-0.46
0.41	-0.71	0.0	0.0	-0.79	-0.46
0.41	0.0	0.71	0.0	0.0	0.91
0.41	0.0	-0.71	0.0	0.0	0.91

B two-sphere model

$\{\widetilde{\mathbf{u}}_i\}_{i=0}^{6}$

$\widetilde{\mathbf{u}}_0$	$\widetilde{\mathbf{u}}_1$	$\widetilde{\mathbf{u}}_2$	$\widetilde{\mathbf{u}}_3$	$\widetilde{\mathbf{u}}_4$	$\widetilde{\mathbf{u}}_5$
0.41	0.0	0.0	0.71	0.49	-0.31
0.41	0.0	0.0	-0.71	0.49	-0.31
0.41	0.71	0.0	0.0	-0.51	-0.27
0.41	-0.71	0.0	0.0	-0.51	-0.27
0.41	0.0	0.71	0.0	0.02	0.58
0.41	0.0	-0.71	0.0	0.02	0.58

C atom-centered model

$\{\mathbf{u}_i\}_{i=0}^{6}$

$\mathbf{u}_0$	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$	
0.41	0.16	-0.01	0.0	-0.1	0.89	q_C
0.41	0.24	-0.43	0.73	-0.05	-0.24	q_{H_g}
0.41	0.25	-0.49	-0.69	-0.05	-0.24	q_{H_g}
0.41	0.45	0.68	-0.03	0.34	-0.23	q_{H_t}
0.41	-0.42	0.32	-0.02	-0.72	-0.19	q_O
0.41	-0.69	-0.08	0.0	0.59	0.0	q_H

Figure 4.8: The orthonormal bases of the right singular vectors: basis of spherical harmonics $\widetilde{\mathbf{Y}}_{S_a}$ (A), basis from the numerical SVD of the LS matrix in two-sphere PC model (B), and atom-centered model (C).

Table 4.1: Effect of the numerical rank r (SVD in eq. 4.50) and the total-charge constraint on the values of atom-centered PCs of methanol and the RMSD (kcal/mol).

	q_C	q_{H_g}	q_{H_t}	q_O	q_H	RMSD
SVD, $r = 6$	0.215	-0.018	0.048	-0.592	0.371	2.457
tSVD, $r = 5$	-0.058	0.056	0.118	-0.532	0.370	2.625
tSVD, $r = 4$	0.007	0.089	-0.101	-0.070	-0.010	8.729
Lagrange, $Q_0 = 0$	0.276	-0.035	0.030	-0.603	0.367	2.587
Elimination, $Q_0 = 0$	0.276	-0.035	0.030	-0.603	0.367	2.587
SVD, $Q_0 = 0$	0.214	-0.019	0.047	-0.593	0.370	2.597
Trivial, $Q_0 = 0$	0.214	-0.019	0.047	-0.593	0.370	2.597

with only a slight slight numerical deviation (< 0.01), a consequence of the molecular charge density spillover beyond the solvent-accessible surface defining the vdW grid.[167]

Although the other singular vectors do not exactly match the corresponding spherical harmonics, the $\mathbf{u}_2$–$\mathbf{u}_4$ vectors can be roughly related to the three components of the dipole moment (Figure 4.8), and the corresponding singular values are commensurate with the singular value μ_l ($l = 1$) obtained for the Lebedev grid model (Figure 4.7). The remaining singular values μ_5 and μ_6 are significantly distorted from the singular value μ_l ($l = 2$), so the components of the quadrupole moment are not reproduced as precisely as the the dipole moment components (Table 4.2).

Among all singular vectors $\{\mathbf{u}_i\}_{i=1}^{6}$, the singular vector $\mathbf{u}_6$ with the lowest singular value μ_6, which is 100 times smaller than μ_1, is dominated by the contribution from the methyl carbon atom (Figure 4.8). Since such small singular values cause numerical instabilities of the LS solution, once can use a regularization technique such as truncated SVD (tSVD) that reduces the rank r by removing the lowest-μ_i vector(s) from the SVD expansion.[111] Removal of $\mathbf{u}_6$ that decreases the rank to $r = 5$ leads to dramatic change in the methyl group charges—the carbon atom charge in particular, which drops from 0.22 to -0.06. Yet, these changes lead only to marginal changes in the the multipole moment and RMSD values, a typical example of the buried atom effect (Tables 4.1, 4.2). This suggests a natural way to impose a restraint on the buried atom charges without introducing a restraining function into the LS sum χ^2, an addition that can negatively affect the electrostatic properties of the PC model.[38,52]

Further removal of the singular vectors $\mathbf{u}_5$ and $\mathbf{u}_6$ (i.e. $r = 4$) leads to severe deterioration of the LS solution, as the corresponding multipole moment strongly deviate from the reference values and the RMSD significantly increases (Tables 4.1, 4.2). Thus, it appears that the tSVD approach should be applied only to the singular vectors that strongly depend on the buried atoms, an important point that will be discussed in detail elsewhere.

Table 4.2: Effect of the rank r (eq. 4.47), degree n (eq. 4.46) and type of the charge constraint on the methanol multipole moments and the RMSD (kcal/mol) within the Lebedev grid PC model ($a = 2$ au, $n = 1, 2$ and $N = 6, 14$) with probe sphere S_R ($R = 8$ au, $T = 194$) and atom-centered PC model with vdW-type grid.

PC model, probe grid	Details	Q_0	Q_{10}	Q_{11}	Q_{1-1}	Q_{20}	Q_{21}	Q_{2-1}	Q_{22}	Q_{2-2}	RMSD
$S_a(N = 6)$, $S_R(T = 194)$	eq. 4.46, $n = 1$	0.000	0.000	-0.325	-0.565	0.000	0.000	0.000	0.000	0.000	1.762
	SVD, $r = 4$	0.000	0.000	-0.323	-0.562	0.000	0.000	0.000	0.000	0.000	1.762
	SVD, $r = 6$	0.000	0.000	-0.323	-0.562	-0.881	0.000	0.000	-0.497	0.000	1.668
$S_a(N = 14)$, $S_R(T = 194)$	eq. 4.46, $n = 2$	0.000	0.000	-0.325	-0.565	-0.887	0.000	0.000	-0.503	2.882	0.559
	SVD, $r = 9$	0.000	0.000	-0.325	-0.566	-0.881	0.000	0.000	-0.497	2.865	0.559
	SVD, $r = 14$	0.000	0.000	-0.325	-0.566	-0.881	0.000	0.000	-0.497	2.865	0.425
atom-centered, vdW	SVD, $r = 6$	0.007	0.000	-0.335	-0.546	-1.020	0.001	0.001	0.110	2.761	2.457
	tSVD, $r = 5$	0.009	0.000	-0.338	-0.543	-1.060	-0.002	0.000	0.219	2.688	2.625
	tSVD, $r = 4$	0.003	0.001	-0.270	-0.377	0.574	-0.003	-0.003	0.225	-1.317	8.729
	Lagrange, $Q_0 = 0$	0.000	0.000	-0.332	-0.543	-0.989	0.001	0.001	0.073	2.739	2.587
	Elimination, $Q_0 = 0$	0.000	0.000	-0.332	-0.543	-0.989	0.001	0.001	0.073	2.739	2.587
	SVD, $Q_0 = 0$	0.000	0.000	-0.331	-0.544	-1.010	0.001	0.001	0.097	2.756	2.597
	Trivial, $Q_0 = 0$	0.000	0.000	-0.331	-0.544	-1.010	0.001	0.001	0.097	2.755	2.597
Reference		0.000	0.000	-0.325	-0.565	-0.887	0.000	0.000	-0.503	2.882	–

4.6 Total-Charge Constraint

Commonly used PC fitting approaches also modify the LS sum (eq. 4.1) by adding a Lagrange multiplier λ in order to constrain the total charge to the correct value:[34,37,48]

$$\chi^2(\mathbf{q}) = |\mathbf{\Phi} - \mathbf{K}\mathbf{q}|^2 + \lambda(\mathbf{1}^\intercal \cdot \mathbf{q} - Q_0), \tag{4.51}$$

which increases the dimension of the Hessian matrix $\mathbf{H} = \mathbf{K}^\intercal\mathbf{K}$ in the normal equation (eq. 4.4):

$$\begin{bmatrix} \mathbf{H} & \mathbf{1} \\ \mathbf{1}^\intercal & 0 \end{bmatrix} \begin{bmatrix} \mathbf{q} \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{K}^\intercal\mathbf{\Phi} \\ Q_0 \end{bmatrix}, \tag{4.52}$$

where $\mathbf{1}$ is an all-ones column-vector.

However, as we have seen, both in the case of the idealized Lebedev grid and the less-than-ideal atom-centered PC models, the Hessian eigenvector with the largest curvature corresponds to the total charge (Figure 4.8 and also Ref.[166]). Thus, in the case of the two-sphere PC fitting, the total charge is reproduced exactly ($Q_0 < 10^{-5}$), while in the atom-centered PC model the total charge only slightly deviates from the exact value due the close proximity of the vdW grid and slight distortion of the total-charge vector $\mathbf{u}_1$ from its analytical analog $\widetilde{\mathbf{Y}}_0$ ($Q_0 = 0.003$ for methanol, Table 4.2).

Addition of the Lagrange multiplier leads to an extra eigenvector $\mathbf{u}_7$ that appears in the eigenbasis of the Hessian matrix (Table 4.3). The curvature along this vector is the smallest in the magnitude ($\kappa_7 = -0.009$) and the vector itself primarily depends on the Lagrange multiplier λ, with only marginal contribution from the PC values. At the same time, remaining eigenvectors $\{\mathbf{u}\}_{i=1}^{6}$ preserve the structure of the original eigenbasis, with negligible contribution from the Lagrange multiplier λ (Table 4.3). Thus, application of the the total charge constraint in addition to already strong restraint (imposed by the eigenvector

$\mathbf{u}_1$) appears to be redundant. Moreover, addition of the Lagrange multiplier aggravates the rank deficiency of already ill-conditioned LS problem.[37,38]

Table 4.3: Eigenvalues μ_i^2 and eigenvectors $\mathbf{u}_i$ of the LS Hessian matrix $\mathbf{H}$ in constraint-free case and with the Lagrange multiplier to constraint the total charge.

	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$	$\mathbf{u}_6$	$\mathbf{u}_7$	atom
μ_i^2	732.3	25.0	10.6	10.2	1.5	0.1		
Constraint-free	0.412	0.158	-0.005	-0.002	0.101	0.891	–	q_C
	0.410	0.249	-0.486	0.688	0.051	-0.241	–	q_{H_g}
	0.410	0.239	-0.434	-0.725	0.052	-0.242	–	q_{H_g}
	0.409	0.447	0.682	0.023	-0.341	-0.226	–	q_{H_t}
	0.407	-0.416	0.324	0.015	0.720	-0.194	–	q_O
	0.401	-0.695	-0.080	0.001	-0.592	0.004	–	q_H
μ_i^2	732.3	25.0	10.6	10.2	1.5	0.1	-0.01	
Lagrange multiplier	0.412	-0.158	-0.005	-0.002	-0.101	0.889	0.066	q_C
	0.410	-0.249	-0.486	0.688	-0.051	-0.240	-0.019	q_{H_g}
	0.410	-0.239	-0.434	-0.725	-0.052	-0.241	-0.019	q_{H_g}
	0.409	-0.447	0.682	0.023	0.341	-0.225	-0.020	q_{H_t}
	0.407	0.416	0.324	0.015	-0.720	-0.194	-0.012	q_O
	0.401	0.695	-0.080	0.001	0.592	0.005	-0.005	q_H
	0.003	0.001	0.000	0.000	0.006	-0.075	0.997	λ

Alternatively, the total charge can be constrained by incorporating condition on the proper total charge directly into the LS sum,[32,140,142] by eliminating one of the charges and setting it to:

$$q_n = Q_0^{mol} - \sum_i^{N-1} q_i, \tag{4.53}$$

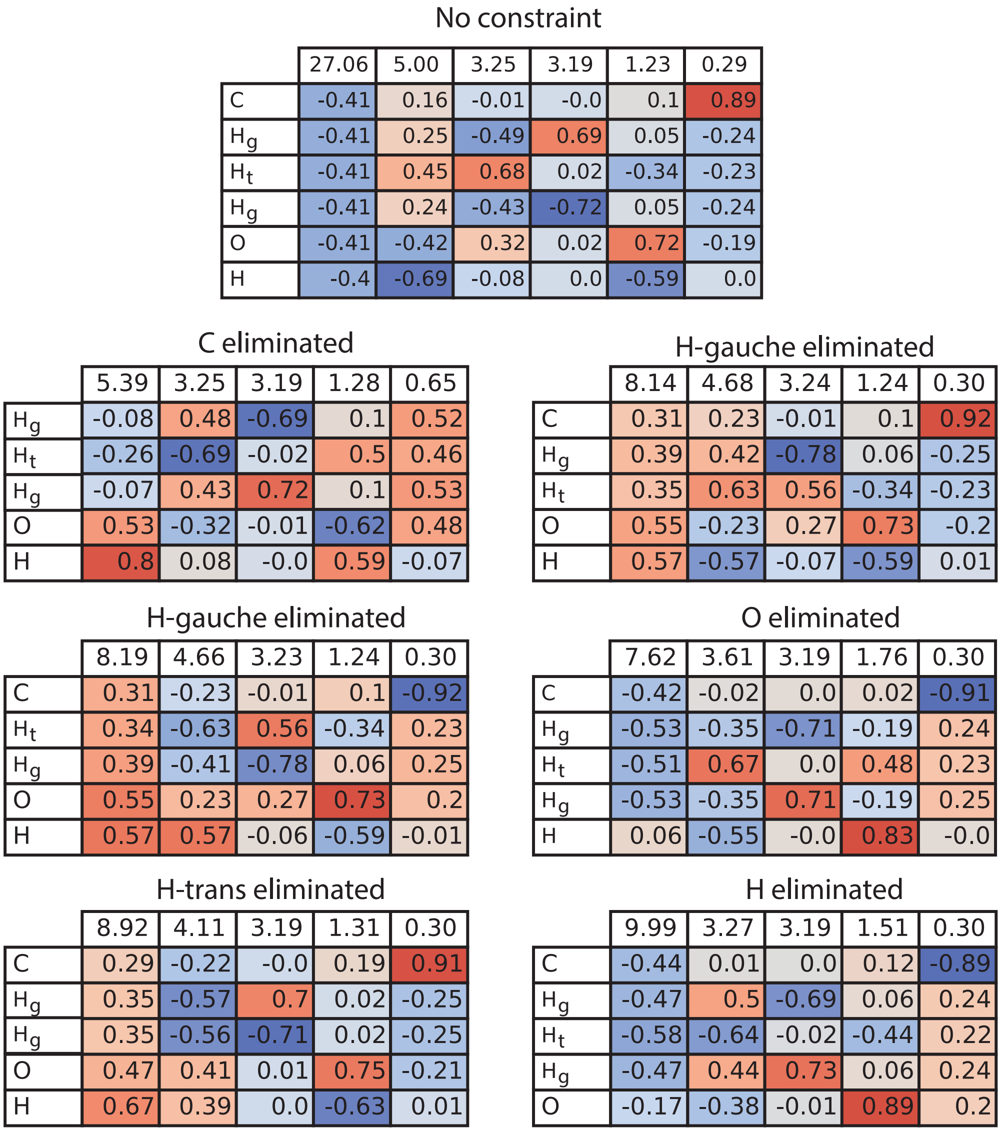
where n is the index of the eliminated charge. This reduces the dimension of the LS problem by one:

$$\chi^2(\mathbf{q}) = \sum_j^T \left[\Phi_j - \frac{Q_0^{mol}}{r_{nj}} - \sum_i^{N-1} \left(\frac{1}{r_{ij}} - \frac{1}{r_{nj}} \right) q_i \right]^2 \tag{4.54}$$

and modifies the elements of the Hessian matrix:

$$H_{km} = \sum_j^T \left(\frac{1}{r_{kj}} - \frac{1}{r_{nj}} \right) \left(\frac{1}{r_{mj}} - \frac{1}{r_{nj}} \right). \tag{4.55}$$

Although the solution obtained with this approach is numerically equivalent to the solution with Lagrange multiplier, regardless which atom has been eliminated (Elimination, $Q_0 = 0$ in Tables 4.1 and 4.2), the structure of the right singular vectors becomes disrupted, (Figure 4.6) which prevents the application of the truncated SVD to improve the numerical stability of the solution.



No constraint

	27.06	5.00	3.25	3.19	1.23	0.29
C	-0.41	0.16	-0.01	-0.0	0.1	0.89
H_g	-0.41	0.25	-0.49	0.69	0.05	-0.24
H_t	-0.41	0.45	0.68	0.02	-0.34	-0.23
H_g	-0.41	0.24	-0.43	-0.72	0.05	-0.24
O	-0.41	-0.42	0.32	0.02	0.72	-0.19
H	-0.4	-0.69	-0.08	0.0	-0.59	0.0

C eliminated

	5.39	3.25	3.19	1.28	0.65
H_g	-0.08	0.48	-0.69	0.1	0.52
H_t	-0.26	-0.69	-0.02	0.5	0.46
H_g	-0.07	0.43	0.72	0.1	0.53
O	0.53	-0.32	-0.01	-0.62	0.48
H	0.8	0.08	-0.0	0.59	-0.07

H-gauche eliminated

	8.14	4.68	3.24	1.24	0.30
C	0.31	0.23	-0.01	0.1	0.92
H_g	0.39	0.42	-0.78	0.06	-0.25
H_t	0.35	0.63	0.56	-0.34	-0.23
O	0.55	-0.23	0.27	0.73	-0.2
H	0.57	-0.57	-0.07	-0.59	0.01

H-gauche eliminated

	8.19	4.66	3.23	1.24	0.30
C	0.31	-0.23	-0.01	0.1	-0.92
H_t	0.34	-0.63	0.56	-0.34	0.23
H_g	0.39	-0.41	-0.78	0.06	0.25
O	0.55	0.23	0.27	0.73	0.2
H	0.57	0.57	-0.06	-0.59	-0.01

O eliminated

	7.62	3.61	3.19	1.76	0.30
C	-0.42	-0.02	0.0	0.02	-0.91
H_g	-0.53	-0.35	-0.71	-0.19	0.24
H_t	-0.51	0.67	0.0	0.48	0.23
H_g	-0.53	-0.35	0.71	-0.19	0.25
H	0.06	-0.55	-0.0	0.83	-0.0

H-trans eliminated

	8.92	4.11	3.19	1.31	0.30
C	0.29	-0.22	-0.0	0.19	0.91
H_g	0.35	-0.57	0.7	0.02	-0.25
H_g	0.35	-0.56	-0.71	0.02	-0.25
O	0.47	0.41	0.01	0.75	-0.21
H	0.67	0.39	0.0	-0.63	0.01

H eliminated

	9.99	3.27	3.19	1.51	0.30
C	-0.44	0.01	0.0	0.12	-0.89
H_g	-0.47	0.5	-0.69	0.06	0.24
H_t	-0.58	-0.64	-0.02	-0.44	0.22
H_g	-0.47	0.44	0.73	0.06	0.24
O	-0.17	-0.38	-0.01	0.89	0.2

Figure 4.9: Hessian eigenbases along with corresponding singular values in the constrained free case and with the total charge constraint by the elimination of one of the atoms.

Given that even for the atom-centered PC/vdW probe model the total charge value deviates only very slightly from the reference value, it should be possible to correct for this deviation without exacerbating the numerical instabilities of the

LS problem, e.g. using the total charge vector $\mathbf{u}_1$. To do that, we convert the SVD solution (eq. 4.50) to a system of linear equations:

$$\mathbf{q} = \sum_{i=0}^{r} \underbrace{\frac{\Phi \cdot \mathbf{v}_i}{\mu_i}}_{c_i} \mathbf{u}_i = \mathbf{U}\mathbf{c}, \tag{4.56}$$

$$\mathbf{U}^{\intercal}\mathbf{q} = \mathbf{c}. \tag{4.57}$$

Then, we replace $\mathbf{u}_1$ in $\mathbf{U}^{\intercal}$ by an all-ones vector $\mathbf{1}$, and set the corresponding coefficient c_1 in $\mathbf{c}$ to the exact value of the molecular total charge Q_0^{mol}:

$$\mathbf{U}_{Q_0}^{\intercal}\mathbf{q} = \mathbf{c}_{Q_0}, \tag{4.58}$$

where

$$\mathbf{U}_{Q_0}^{\intercal} = \begin{bmatrix} \mathbf{1} & \mathbf{u}_2 & \cdots & \mathbf{u}_N \end{bmatrix}^{\intercal}, \tag{4.59}$$

$$\mathbf{c}_{Q_0} = \begin{bmatrix} Q_0^{mol} & c_2 & \cdots & c_N \end{bmatrix}^{\intercal}. \tag{4.60}$$

This approach does not introduce any redundant constraints, preserves the electrostatic properties of the unconstrained solution, and results only in to minor changes in the PC values (SVD, $Q_0 = 0$ in Tables 4.1 and 4.2) and is compatible with truncated SVD. Also, the error in the total charge value is small enough and can be corrected by simply distributing the Q_0 error correction across the atomic charges; this trivial total charge correction gives result nearly identical to eq. 4.58 (Trivial, $Q_0 = 0$ in Tables 4.1 and 4.2).

4.7 Summary

To understand the origins of the ill-conditioning of the least-squares (LS) point charge (PC) fitting problem, we revisited the PC representation of the molecular electrostatic potential (MEP) from the first principles, as an example of the inverse problem.

Based on the properties of the Coulomb potential that can be expanded in terms of spherical harmonics, we introduce a model where the MEP of a molecule is exactly reproduced by a charged sphere that has the same multipole moments Q_{lm} as the molecule. Using Lebedev quadrature this continuous model is converted into a discrete PC model, where the PC values are evaluated analytically from the multipole moments Q_{lm} up to the maximum value determined by the quadrature.

In this context, the traditional atom-centered PC model can be viewed as an irregular numerical quadrature, poorly suited to reproduce the multipolar expansion of the MEP. As such, this quadrature only allows integration of the monopole and, approximately, dipole terms. The corresponding large-curvature—or 'stiff'[164,165]—Hessian eigenvectors $\mathbf{u}_i$ can still be related to the corresponding multipoles Q_{lm}. This explains previously observed correspondence between the highest-curvature Hessian eigenvectors and the total charge and the dipole moment components;[166,167] this correspondence quickly breaks down for the higher multipole moments.

This consideration then reveals the origins of the ill-conditioning of the PC fitting due to the presence of low-curvature—or 'sloppy'[164,165]—vectors $\mathbf{u}_i$. The intrinsic ill-conditioning arises even in the case of the ideal spherical model: since the higher-rank multipole moments Q_{lm} have smaller contribution to the MEP, the singular values μ_l decay as l increases. The ill-conditioning is further exacerbated in the numerical treatment of the Lebedev grid model because the number of PCs does not match the dimension of the basis formed by Lebedev quadrature. The remaining singular values/curvatures are even lower in magnitude and do not correspond to particular multipole moments Q_{lm}. The same rank-deficiency problems apply to the atom-centered PC grids. However, in that case most of the eigenvectors do not have a direct correspondence to the multipole moments, which leads to even wider spread-out of the singular values/curvatures.

These insights can suggest several ways to alleviate the ill-conditioning of the problem. For instance, the buried atom problem can be addressed by truncating the sloppy singular vectors with dominant contribution from these atom, instead of introducing additional restraining functions[36,41,43,49] that can negatively affect the overall electrostatic properties of the molecule.[38,52] Also, slight deviations of the total charge of the fitted PC solution can be fixed by adjusting the stiff total-charge vector $\mathbf{u}_1$ and the corresponding coordinate Q_0^{mol}, rather than introducing a Lagrange multiplier that increases the rank-deficiency of the Hessian matrix.[37,38]

The results presented here can help further application of the PC model in biomolecular simulations. Although the force fields using point charges may not be as accurate as the force fields that explicitly include multipoles and/or polarization effects, the simplicity and computational efficiency of the PC model has ensured its continued survival.[70] In fact, representation of multipoles using the Lebedev grid PC model can provide an alternative to the multipole moment expansion;[177] it also can be used to extend recently proposed Distributed Charge Model.[61,160]

4.8 Computational Details

MEP and multipole moments were calculated at the B3LYP/aug-cc-pVDZ level[149–151,178] as implemented in Q-Chem package.[179] For atom-centered PC fitting the reference MEP was generated as the cubic grid with linear density 2.8 points/Å, followed by the removal of the points outside of 1.0-2.0 van der Waals radii range around each atom (vdW grid). For the two-sphere PC model the Lebedev quadrature rules were used as implemented in PyQuante package.[180,181] Charge fitting procedures were implemented in the in-house developed *fftoolbox* Python library.[182] SVD was performed using *numpy* library.[153] Spherical harmonics were accessed from *scipy* library.[183]

Chapter 5

Point Charges Meet Accuracy of Multipoles

5.1 Introduction

Many molecular interactions of (bio)chemical importance are governed by subtle anisotropic features of the molecular electrostatic potential (MEP), such as lone pairs, σ-holes, π-systems.[15,57,184,185] In this regard, computationally efficient yet accurate modeling of electrostatic interactions is critical for reliable simulation of the biological macromolecules at the atomic level. Expansion of the electrostatic interaction into a series of the interacting atomic multipoles provides a rigorous framework to introduce anisotropic electrostatic features into a simulation.[16–19] Recent advancements in the development of the new force fields and methodologies bring closer to reality the routine application of the multipolar force fields to the systems in the size range of practical interest.[29–31] Yet, up to date only limited number of systems has been studied using the multipolar force fields.[24,186–188] Besides a significant computational cost, implementation of the multipole-multipole interactions is non-trivial as it requires definition of a local reference frame for each atom, calculation of forces and torques for each atomic multipole, and finally, implementation of the Particle-Mesh Ewald schemes[29–31] for simulating periodic boundary conditions.

Therefore, in the majority of existing force fields the multi-site expansion expansion is truncated at the monopole, leading to the computationally less demanding atom-centered point charge approximation.[32–34] However, isotropic nature of a single point charge potential is unable to describe the anisotropic character of the MEP around a given atom.[53,54,189] Moreover, numerical difficulties in the point charge derivations due to the ill-conditioned least squares charge fitting problem have resulted in the sophistications of the derivation by addition of the restraints/constraints.[36–38,49,140] Nevertheless, low computational requirements and relative simplicity in the implementation of the point charge potentials motivate to go beyond the atom-centered paradigm and use off-center point charges to reproduce effects of higher-rank (above monopole) atomic multipoles.[60–64] However, none of the existing methods offer a systematic approach in placing the off-center point charges as well as in the derivation of their values.

In the previous Chapters we showed that the atom-centered point charges correspond to an improper coordinate system/quadrature that exacerbates the numerical stability of the obtained solution due to the ill-conditioned nature of the inverse electrostatic problem.[166,190] In order to alleviate the numerical instability, point charges must be placed over the sphere, according to the Lebedev quadrature rule that exactly integrates spherical harmonics—this solves the inverse electrostatic problem analytically and allows one to obtain point charges directly, avoiding ill-conditioned least squares fitting.[190]

In this chapter, we demonstrate that the point charges placed at the nodes of a Lebedev quadrature can approximate the multipole moment expansion of MEP up to any given degree in a systematic fashion. Using Stone's distributed multipoles[16] we introduce the multi-site Lebedev model where the electrostatic potential of each atom is modeled by an atom-centered Lebedev sphere. On several examples we demonstrate that a simple point charge framework can describe anisotropy of the MEP with the accuracy previously thought to be achieved only by the multipolar force fields.

5.2 Point-Charge Representation of the Multipolar Expansion

The MEP can be computed exactly using the molecular charge density $\rho(\mathbf{r})$ from quantum mechanical (QM) calculations:

$$\Phi_{QM}(\mathbf{r}) = \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3r'. \tag{5.1}$$

At the long range, MEP can be expanded into the multipole series up to a degree n using the multipole moment values Q_{lm}:[1,107]

$$\Phi_n(\mathbf{r}) = \sum_{l=0}^{n} \sum_{m=-l}^{l} Q_{lm} r^{-l-1} Y_{lm}(\theta, \varphi), \tag{5.2}$$

where $\mathbf{r}$ is defined by the magnitude r, polar angle φ and azimuthal angle θ; $Y_{lm}(\theta, \varphi)$ are renormalized real-value spherical harmonics.[1]

As long as the molecular multipole moments are reproduced there is no concern about the way the charge density $\rho(\mathbf{r})$ is distributed in space. Therefore, it is indifferent to the MEP if we replace the molecular charge density $\rho(\mathbf{r})$ by the sphere of radius a with surface charge density $\sigma(\mathbf{a})$ centered at the origin of the molecular multipole expansion in such way that the multipole moments of the sphere are equivalent to the multipoles of the molecule:

$$Q_{lm} = \int_V r^l \rho(\mathbf{r}) Y_{lm}(\theta, \varphi) d\mathbf{r} = a^l \int_{S_a} \sigma(\mathbf{a}) Y_{lm}(\theta, \varphi) d\Omega. \tag{5.3}$$

where $d\Omega$ is the differential of the solid angle.

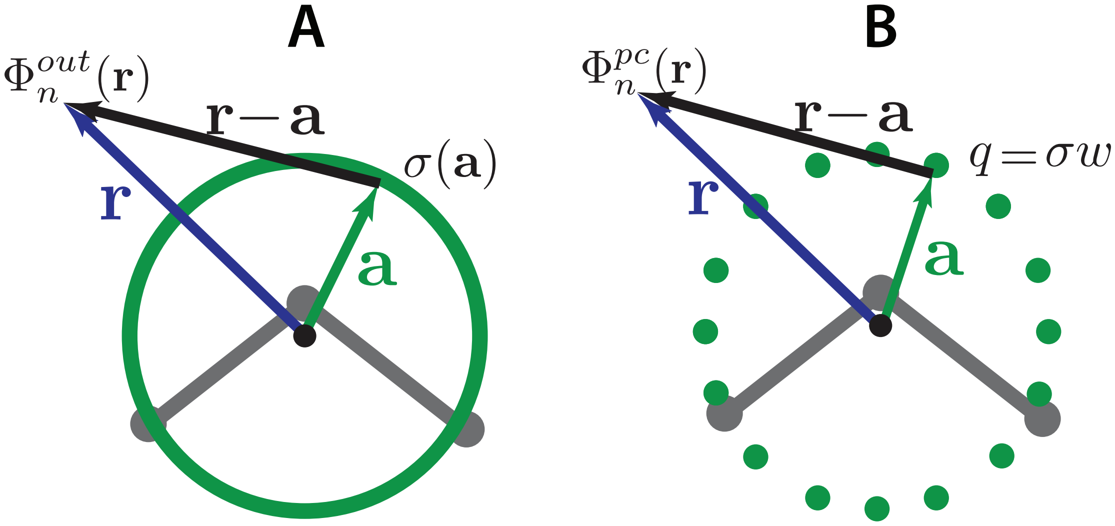


Figure 5.1: Continuous charged sphere model centered at the origin of the molecular multipole expansion (A) and its point-charge (B) representation.

---

[1] Here we use real-value spherical harmonics with the normalization $\int Y_{lm}(\theta, \varphi) Y_{l',m'}(\theta, \varphi) = \frac{4\pi}{2l+1} \delta_{ll'} \delta_{mm'}$

Then, the sphere splits space into two regions (Fig. 5.1): the region outside the sphere with potential Φ_{out} and the region inside the sphere with potential Φ_{in}. Depending on the relative magnitude of the source **a** and observation **r** vectors the multipole series expansions takes either its regular or inverse forms.[191] In the region outside the sphere, where $r > a$, the electrostatic potential Φ_{out} is equivalent to the multipole moment expansion of the MEP (eq. 5.2):

$$\Phi_n^{out}(\mathbf{r}) = \sum_{l=0}^{n} \sum_{m=-l}^{l} Q_{lm} r^{-l-1} Y_{lm}(\theta, \varphi), \tag{5.4}$$

where the regular multipole moments of the sphere Q_{lm} are given by eq. 5.3. At the same time, inside the sphere, where $r < a$, the multipole expansion takes its inverse form:

$$\Phi_n^{in}(\mathbf{r}) = \sum_{l=0}^{n} \sum_{m=-l}^{l} Q_{lm}^{I} r^{l} Y_{lm}(\theta, \varphi), \tag{5.5}$$

where the inverse multipole moments Q_{lm}^I are given by

$$Q_{lm}^{I} = a^{-l-1} \int \sigma(\mathbf{r}) Y_{lm}(\theta, \varphi) d\Omega. \tag{5.6}$$

Then, according to the Gauss's law, the discontinuity in the normal component of the electric field $\mathbf{E} = -\boldsymbol{\nabla}\Phi$ in crossing the sphere defines the charge density of its surface:

$$\frac{\partial \Phi_n^{in}}{\partial r} - \frac{\partial \Phi_n^{out}}{\partial r}\bigg|_{r=a} = \frac{4\pi}{a^2}\sigma_n, \tag{5.7}$$

Using the that fact the regular and inverse multipoles of the sphere are uniquely related $Q_{lm}/Q_{lm}^I = a^{2l+1}$, the surface charge density $\sigma(\mathbf{a})$ over the sphere can be expressed in terms of the multipoles Q_{lm} and radius a:

$$\sigma_n(\mathbf{a}) = \sum_{l=0}^{n} \sum_{m=-l}^{l} \frac{2l+1}{4\pi} a^{-l} Q_{lm} Y_{lm}(\theta, \varphi). \tag{5.8}$$

The derived charge density (eq. 5.8) also corresponds to the outer expansion introduced by Rogers (eq. 17 in Ref.[177]) and to the solution of the inverse electrostatic problem previously introduced in Chapter 4 (eq. 4.29).

We can now represent the charged sphere numerically using a spherical quadrature. For example, using the Lebedev-Laikov grid[173] that exactly integrates spherical harmonics up to $l = 2n$, the multipole moments of the sphere can be obtained computed up to degree n:[176]

$$Q_{lm} = a^l \sum_{i}^{N} \sigma_i Y_{lm}(\theta_i, \phi_i) w_i \tag{5.9}$$

where N is the number of the nodes in the quadrature, θ_i and φ_i are the angular coordinates and w_i is the integration weight at the node i.

Accordingly, using the discrete representation of the multipole moments (eq. 5.9), the expansion outside the sphere (eq. 5.4) can be represented in the discrete form:

$$\Phi_n(\mathbf{r}) = \sum_{i}^{N} \sigma_i w_i \sum_{l=0}^{n} \sum_{m=-l}^{l} \frac{a^l}{r^{l+1}} Y_{lm}(\theta_i, \varphi_i) Y_{lm}(\theta, \varphi). \tag{5.10}$$

For a sufficiently large separation between the sphere and an observation point $\mathbf{r}$, this expansion can be reduced to a point-charge potential Φ_n^{PC} (Figure 5.1):

$$\Phi_n(\mathbf{r}) \simeq \Phi_n^{PC} = \sum_{i}^{N} \frac{q_i}{|\mathbf{r} - \mathbf{a}_i|}, \tag{5.11}$$

where q_i is the point charge at the node i:

$$q_i = w_i \sum_{l=0}^{n} \sum_{m=-l}^{l} \frac{2l+1}{4\pi} a^{-l} Q_{lm} Y_{lm}(\theta_i, \varphi_i). \tag{5.12}$$

Thus, the point charges arranged over the sphere according to the Lebedev quadrature rule produce electrostatic potential that at large distances is numerically identical to the potential produced by the multipole moments.

Below, numerical examples provide a quantitative consideration of the proposed Lebedev charge model.

Table 5.1: Averaged over the set of reference molecules $\langle a \rangle$, minimum a_{min} and maximum a_{max} values of the radius required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	$\langle a \rangle$	a_{min}	a_{max}
1	0.8	0.5	1.4
2	1.2	0.7	2.0
3	1.4	0.9	2.1
4	1.9	1.0	2.8

5.3 Numerical Examples of the Lebedev Charge Model

5.3.1 Modeling Single-Site Molecular Multipoles

First, we tested if the Lebedev charge model (eq. 5.11) accurately reproduces the electrostatic potential from the molecular multipole expansion (eq. 5.2) within the solvent-accessible surface of the molecule. A set of 17 organic molecules was used for the reference calculations (Appendix A). To measure the quality of the approximation we used the root mean square deviation (RMSD) between the multipolar potential Φ_n and its point-charge analog Φ_{PC} for different ranks n over the M points in the van der Waals grid:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_i^M \left[\Phi_n(\mathbf{r}_i) - \Phi_n^{PC}(\mathbf{r}_i)\right]^2}{M}} \tag{5.13}$$

In all molecules that we considered, the RMSD approaches zero as the rank n increases and the radius of the sphere decreases (Figure 5.2 and Appendix A). For example, in order to achieve 0.05 kcal/mol difference in the RMSD in the expansion up to octopole moment, the radius of the sphere should be $a = 1.4$ au on average. (Table 5.1) The small-radius requirement tends to increase for smaller molecules and lower rank n, with the smallest value of 0.5 au for water molecule with rank $n = 1$ (Figure 5.2). According to eq. 5.9, the point charge values scale as a^{-l} and for extremely small spheres they approach infinity (if

$l > 0$). We verified that the radius of 0.5 au does not lead to unreasonably high charge values (with the highest absolute value of 18.0 e in uracil due to its large quadrupole moment, see Appendix A) and thus was used as the sphere's radius in all calculations throughout this chapter.

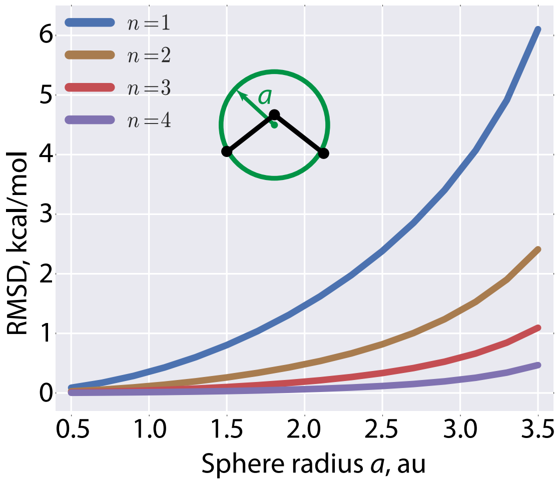


Figure 5.2: Effect of the sphere radius a on the RMSD between the multipolar expansion (eq. 5.2) and point-charge potential (eq. 5.11) in water.

Regardless the radius value, multipole moments of the sphere are constant and exactly match the molecular multipoles. As the Lebedev rule suggests, in order to reproduce the dipole moment of the molecule, six point charges in octahedral arrangement are required with two points along each dimension. Thus, in the case of water molecule, the Lebedev charge model with $n = 1$ is reduced to a trivial case where two non-zero point charges are separated by the distance $2a$ and the corresponding electrostatic point charge potential corresponds to the potential produced by a point dipole (Figure 5.3). As n increases, more points are added according to the quadrature rule (e.g. $N = 14$ points for $n = 2$, $N = 26$ points for $n = 3$, see Appendix A for details) and the electrostatic potential produced by the collection of point charges converges to the QM MEP (Figure 5.8 and Appendix A). While for small molecules the single-site molecular multipole expansion converges relatively fast, truncation at a higher degree n is required for larger molecules in order to properly describe local features of the potential around each atom (atomic electrostatic potential, AEP).[2] Nevertheless, in most cases, starting from $n = 2$ or 3 molecular expansion calculated using Lebedev charge model becomes superior than the MEP-fitted atom-centered point charges and such anisotropy features as lone pairs and σ-holes start to emerge (Figures 5.4 and 5.5 and Appendix A).

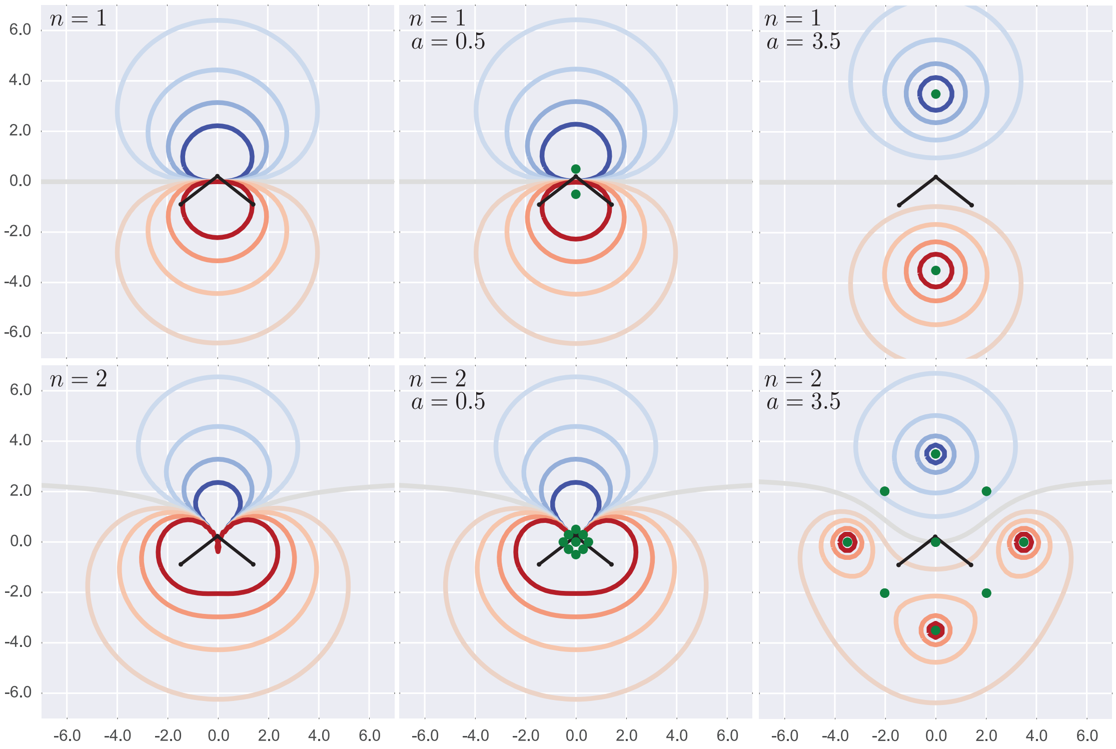


Figure 5.3: Electrostatic potential of water in the plane of the molecule within single-ste Lebedev model with $n = 1, 2$ and $a = 0.5,\ 3.5$ (two right columns) as compared with the true multipole moment potentials (left column).

5.4 Modeling Multi-Site Atomic Multipoles

In order to describe the electrostatic properties of larger molecules, multi-site expansion of the MEP centered at the nuclei positions is required. Application of various partitioning/distribution schemes allows one to obtain multi-site multipolar expansion beyond the monopole up to any rank.[1,16–19] However, in many existing force fields multipolar expansion is truncated at the monopole leading to the atom-centered approximation that fails completely in describing the AEP, especially around atoms such as nitrogen, oxygen, sulfur and halogens

---

[2] Atomic electrostatic potential (AEP) grid is obtained from the MEP grid by selecting the grid points that are the closest to the corresponding atom.

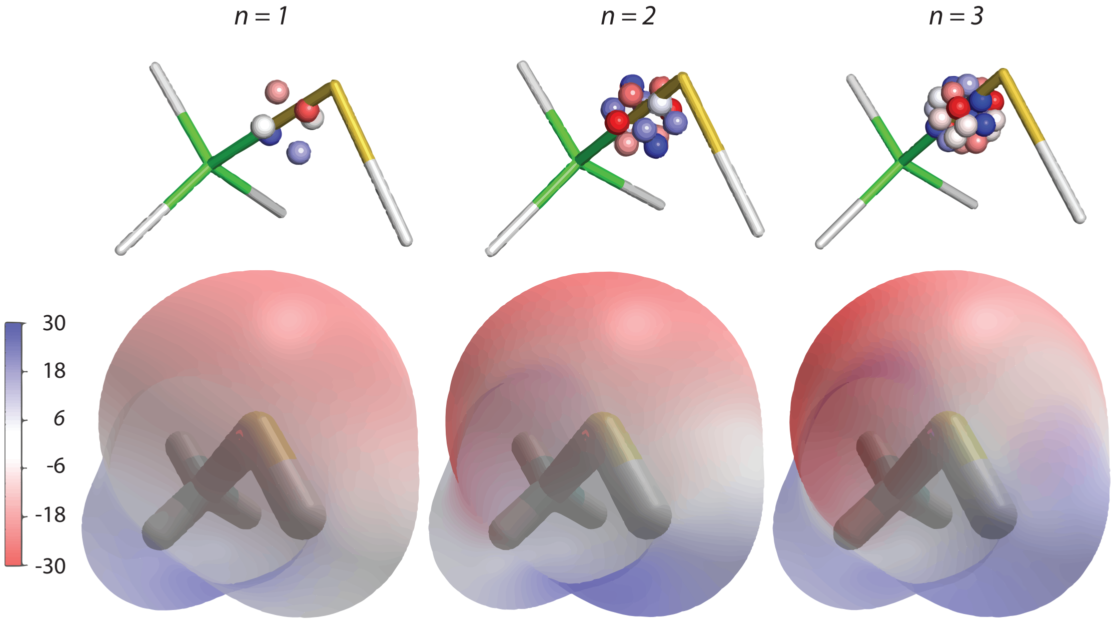


Figure 5.4: Electrostatic potential over the isosurface of constant charge density (0.002 au) of CH_3SH calculated with single-site Lebedev models ($a = 0.5$, $n = 1, 2, 3$).

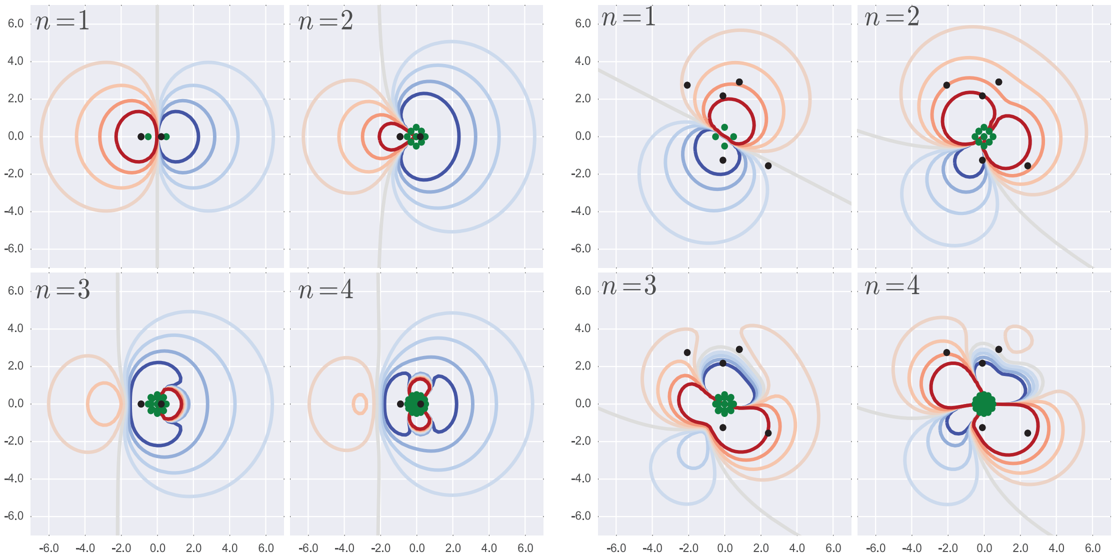


Figure 5.5: Electrostatic potentials calculated using single-site Lebedev model ($a = 0.5$, $n = 1, 2, 3, 4$) of water and CH_3SH in the plane of the molecules. Contour levels: -100, -50, -25, -12, 0, 12, 25, 50, 100 kcal/mol.

where such anisotropy features as lone pairs and σ-holes become dominant.[53,54] Quantitatively it can be seen by comparing the atom-centered point charge potential with the QM MEP over the solvent-accessible region of the molecule. The improper reproduction of the QM MEP is indicated low values of Pearson's correlation coefficient, high root mean square deviation (RMSD), high root mean absolute error (RMAE) and high maximum error between the point charge and QM potentials (Table 5.2 and Appendix A).[54] Among the worst cases is sulfur atom in CH_3SH molecule, which in atom-centered point charge model produces isotropic potential instead of displaying areas of negative potential above and below the plane of the molecule due to the lone pairs and a small area of the positive potential along the C-S bond in the plane of the molecule due to the σ-hole (Figure 5.6 and Appendix A).

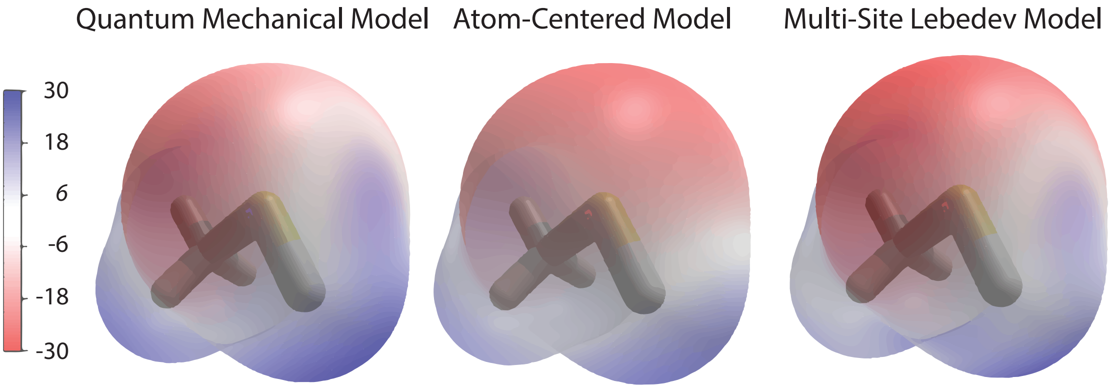

Quantum Mechanical Model
Atom-Centered Model
Multi-Site Lebedev Model
30
18
6
-6
-18
-30

Figure 5.6: Electrostatic potential over the isosurface of constant charge density (0.002 au) calculated using three charge models: quantum mechanical (left), multi-site Lebedev model (middle, $a = 0.5$, $n = 2$) and atom-centered point charge model (right).

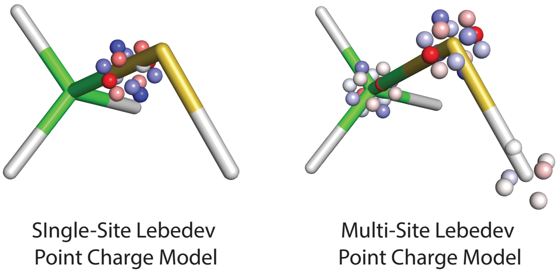

SIngle-Site Lebedev
Point Charge Model
Multi-Site Lebedev
Point Charge Model

Figure 5.7: Point charge representation of single-site and multi-site Lebedev models of CH_3SH.

Here, to include effects of higher atomic multipoles we used Stone's distributed multipole analysis (DMA)[16] and applied Lebedev charge model at each atomic center (Fig. 5.7). To avoid an excessive proliferation of the expansion centers, we removed the methyl hydrogens from the analysis and retained other types of hydrogens (e.g. hydrogens in amine, hydroxyl groups, etc) up to rank $n = 1$ such that the effects of the higher multipoles ($n > 1$) on these hydrogens are transferred to the neighboring atoms.[1]

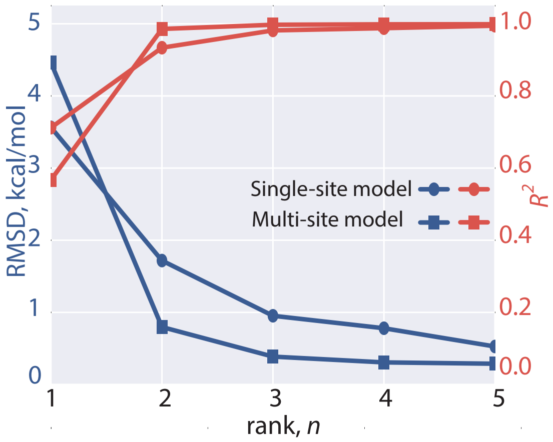


Figure 5.8: Convergence of the CH_3SH electrostatic potential to the QM MEP within single-site (circles) and multi-site Lebedev (squares) models. RMSD and Pearson R^2 correlation coefficient are used to quantify the convergence.

The multi-site Lebedev model reproduces all distributed multipoles as well as overall molecular multipole moments (Appendix A). It converges to the QM MEP faster than the single-site model with significantly improved description of the AEP around each atom (Figure 5.8, Appendix A). Also, most cases with $n > 2$ on all heavy atoms are significantly superior than the atom-centered model in the description of the overall MEP and local AEPs as indicated by the lowered RMSD, RMAE and max. error values and increased correlation coefficient (Table 5.2). Indeed, visual inspection of the MEP map on the isodensity surface of CH_3SH reveals the presence of the areas with negative potential above and below the plane of the molecule and the area of the positive potential along the C-S bond in the plane of the molecule (Figure 5.6).

Table 5.2: Comparison of the multi-site Lebedev ($a = 0.5$ au and $n = 2$) and atom-centered point charge models to reproduce AEPs. Averaged statistical parameters are reported, see Appendix A for values in individual cases. All dimensional quantities are in kcal/mol.

Parameter	Model	S	N	O	Br
RMSD	Atom-Centered	1.240	1.235	0.699	1.017
	Multi-Site Lebedev	0.382	0.433	0.313	0.316
RMAE	Atom-Centered	0.383	0.183	0.064	0.161
	Multi-Site Lebedev	0.085	0.130	0.028	0.049
Max. Error	Atom-Centered	3.682	3.505	2.218	3.498
	Multi-Site Lebedev	1.150	1.097	1.152	0.924
R^2	Atom-Centered	0.683	0.873	0.901	0.328
	Multi-Site Lebedev	0.980	0.990	0.986	0.933

5.5 Summary

The Lebedev charge model—a model where point charges are arranged over the sphere using the Lebedev quadrature rule—reproduces atomic and molecular multipoles and describes major local features of the MEP including the presence and directionality of the donor/acceptor features such as lone pairs and σ-holes. As compared to other methods where point charge values are derived directly from the multipole moments,[38,61,62,143] in the proposed model charge values can be obtained analytically without any fitting and/or solving systems of linear equations. The quality of the potential can be systematically improved within the point charge approximation, which makes this model a computationally more efficient numerical analog to the multipolar formalism. Finally, existing support of the off-center point charges in most simulation packages allows an immediate implementation of the model to achieve the multipolar quality within the point charge framework.

5.6 Computational Details

Geometry optimizations were performed at the MP2/aug-cc-pVTZ level[151] as implemented in Gaussian package. Single-site molecular multipole moments and distributed multipole moments were calculated using Stone's Generalized Distributed Multipole Analysis (GDMA) software. Methyl hydrogens atoms were removed from the analysis; all other hydrogens were retained up to $n = 1$. The van der Waals grid was generated with linear density of 2.8 points/Å, followed by the removal of the points outside of 1.66-2.2 van der Waals radii range around each atom. Atom-centered point charges were fitted to the quantum mechanical MEP over the vdW grid points using singular value decomposition (SVD) in *numpy* library as implemented in the in-house developed Python library *fftoolbox*.[182] Lebedev quadrature rules were used as implemented in PyQuante package.[180,181] Lebedev point charge models were implemented in the *fftoolbox* library. Spherical harmonics were accessed from *scipy* library.[183]

Chapter 6

Application of the Model to S-Nitrosothiols

6.1 Simultaneous Fitting of Several Force Field Terms for CysNO

In order to reproduce the effect of the charged residues on the properties of the –SNO group during the molecular dynamics simulations, the force field parameters must be able to properly describe the interaction between the –SNO group and the charged or polar protein residues. While the description of the electrostatic potential using the accurate multipolar force fields is computationally expensive, the Lebedev charge model introduced in the Chapters 4 and 5 can provide multipolar quality of the description within computationally inexpensive point charge approximation.

Although the interactions between a Lewis base and σ-hole are electrostatically driven, the spatial orientation of interacting species is largely due to the induction, dispersion and exchange-repulsion.[106] Thus, here we perform simultaneous fitting of electrostatic and Lennard-Jones terms, as well as some of the bonded terms. The reference interaction energies were obtained with MeSNO as a model for CysNO, while NH_4^+ and $MeCOO^-$ were used to model lysine residue and aspartic/glutamic acid, respectively.

6.1.1 Bonded Terms: Equilibrium Bond lengths, Angles and Force Constants

In AMBER force fields, bond and angle terms are usually described by the harmonic approximation:

$$U^{FF} = \sum_{bonds} k_r (r - r_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2, \tag{6.1}$$

where k_r and k_θ are the force constants and r_0 and r_θ are equilibrium bond lengths and angles. Due to the functional form of the harmonic potential, optimization of the bonded force field parameters was performed in two steps: first, the force constants were optimized with the equilibrium bond lengths and angles taken from the optimized geometries of cis- and trans-MeSNO, then the equilibrium bonds and angles were optimized with the force constants taken from the first step. The optimization of the parameters results in two minimization procedures where at the first step the least square sum is taken as the function of the force constants:

$$\chi^2(k_r, k_\theta) = \sum_i \left[U_i^{QM} - U_i^{FF}(k_r, k_\theta)\right]^2 \tag{6.2}$$

and at the second step, the least square sum is taken as the function of the equilibrium constants:

$$\chi^2(r_0, \theta_0) = \sum_i \left[U_i^{QM} - U_i^{FF}(r_0, \theta_0)\right]^2 \tag{6.3}$$

Here, the reference QM energies U^{QM} were obtained from the relaxed PES scan along C-S, S-N, N-O bonds and C-S-N, S-N-O angles of cis- and trans-MeSNO. The optimized force field parameters (Table 6.1) closely reproduce QM energy scans (Figures 6.1 and 6.2).

Table 6.1: Force constants and equilibrium values fitted to relaxed PES scans

	r_0, Å	k_r, kcal/(mol Å^2)
r_{CS}	1.793	211.191
r_{SN}	1.803	80.689
r_{NO}	1.184	733.739
	θ_0, degree	k_θ, kcal/(mol degree^2)
θ_{CSN}	98.166	0.0134
θ_{SNO}	116.862	0.0265

6.1.2 Non-Bonded Terms: Point Charges and Lennard-Jones Parameters

The reference grid of quantum mechanical (QM) interaction energies between MeSNO and ammonium ion NH_4^+ was obtained in Ref.[89] by density functional

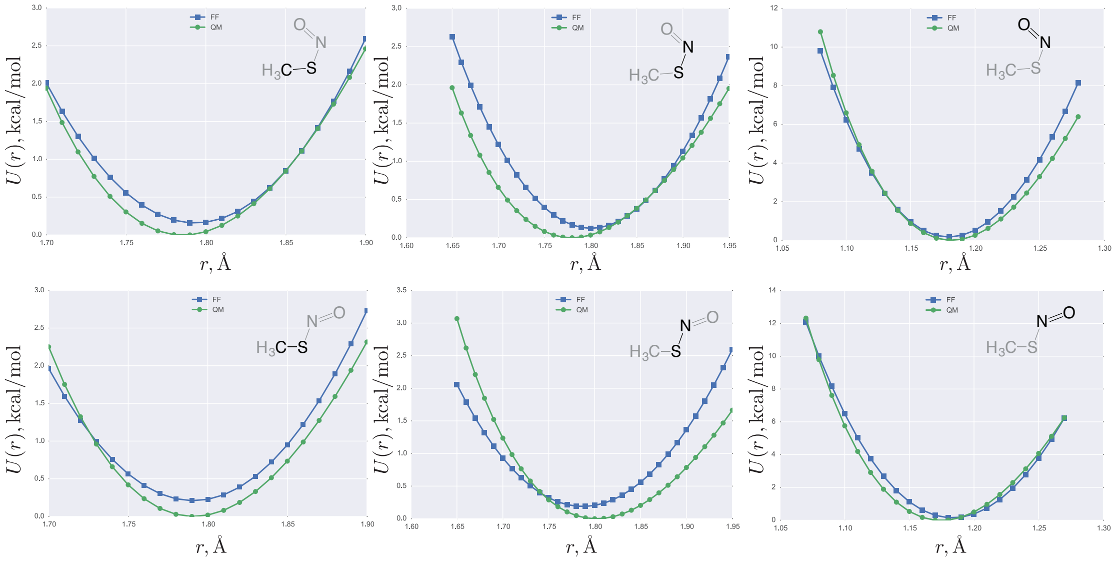


Figure 6.1: QM vs. optimized FF potential energy scans along bonds in cis- (top) and trans-MeSNO (bottom).

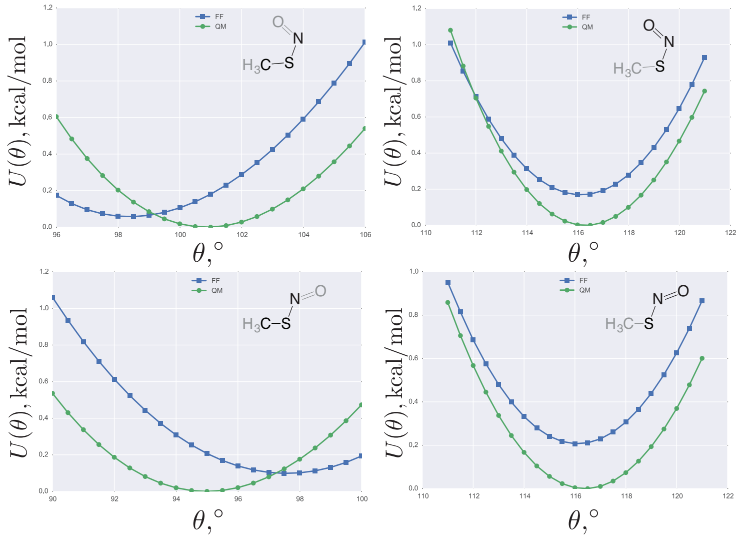


Figure 6.2: QM vs. optimized FF potential energy scans along angles in cis- (top) and trans-MeSNO (bottom).

theory (DFT) calculations at PBE0/def2-SV(P)+d level of theory. The reference grid was constructed by placing the probe ${NH_4}^+$ molecule at different positions over the solvent-accessible region of the MeSNO around C, S, and N atoms (Figure 6.3). For each grid point, a constrained optimization was performed with the probe fixed at the nitrogen atom.

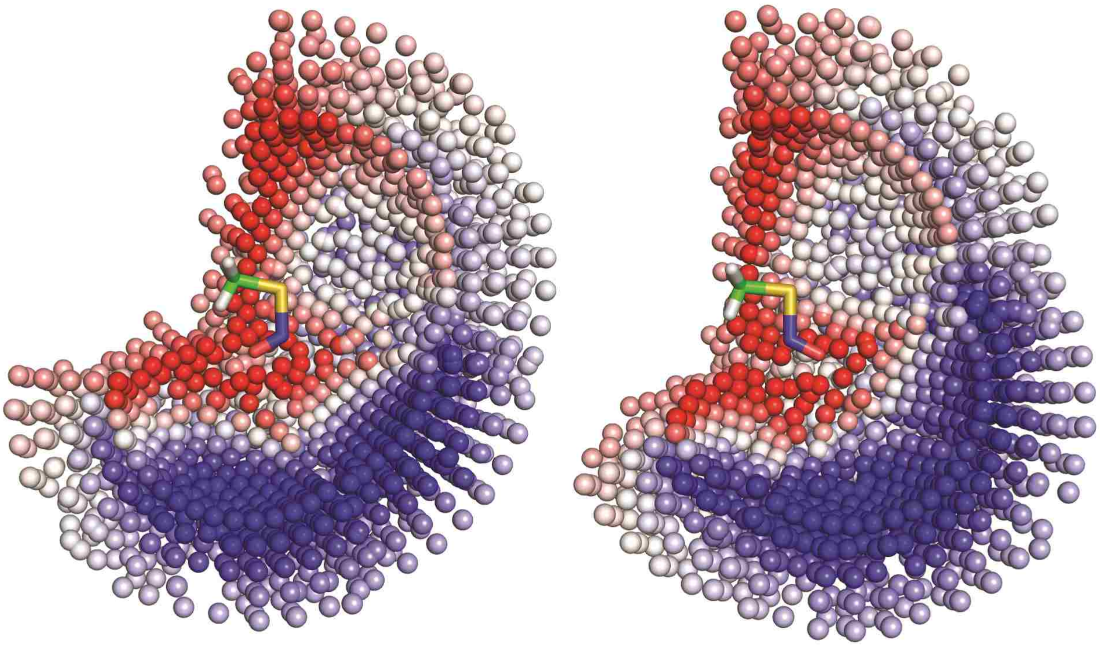


Figure 6.3: Representation of the the interaction energies between MeSNO (cis-MeSNO on the left and trans-MeSNO on the right) and ammonium ion ${NH_4}^+$. Position of each colored sphere corresponds to the position of the nitrogen in ${NH_4}^+$. The color of the sphere represents the strengths of the interaction: red for repulsion and blue for attraction.

Given the reference QM interaction energies, the point charges and Lennard-Jones parameters can be obtained by the non-linear least square fitting. To account for possible polarization effects due to the S-N and N-O bond terms, force constants k_{SN} and k_{NO} were also included in the fit in a way that the overall least square sum contains both, bonded and non-bonded force field terms:

$$\chi^2(Q_{lm}, \varepsilon, r^*, k_{SN}, k_{NO}) = \sum_i \left[U_i^{QM} - U_i^{non-bonded}(Q_{lm}, \varepsilon, r^*) - U_i^{bonded}(k_{SN}, k_{NO}) \right]^2 \tag{6.4}$$

where the bonded term is described by the harmonic potential of the S-N and N-O bonds:

$$U^{bonded}(k_{SN}, k_{NO}) = k_{SN}(r_{SN} - r_{0SN})^2 + k_{NO}(r_{NO} - r_{0NO})^2 \tag{6.5}$$

and the non-bonded term contains Coulomb and Lennard-Jones terms:

$$U^{non-bonded}(Q_{lm}, \varepsilon, r^*) = \sum_{i<j} \frac{q_i q_j}{r_{ij}} + \varepsilon_{ij} \left[\left(\frac{r^*_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r^*_{ij}}{r_{ij}} \right)^{6} \right] \tag{6.6}$$

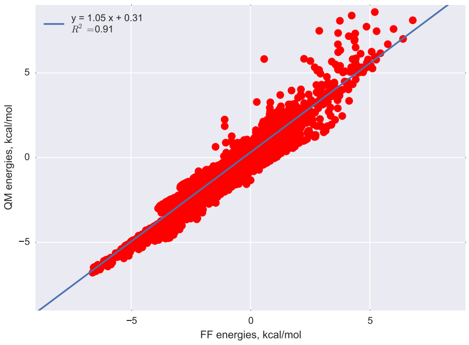


Figure 6.4: Correlation between interaction energies calculated using PBE0/def2-TZVPPD (QM energies) and optimized force field (FF energies).

Unlike in the traditional force fields where the point charges are placed at the atomic positions, here we place point charges according to the Lebedev quadrature rule around each atom in MeSNO. As it was shown in Chapter 5, such arrangement allows multipolar description of the electrostatic properties. While point charges over the sphere were used for the energy calculations in the least squares sum (eq. 6.4), the minimization was performed using atomic multipoles Q_{lm}. At each minimization step atomic multipoles were converted to the point charges using eq. 4.46:

$$q_i = \sum_{l=0}^{n} \sum_{m=-l}^{l} \sqrt{\frac{2l+1}{4\pi}} a^{-l} Q_{lm} Y_{lm}(\theta_i, \varphi_i) w_i. \tag{6.7}$$

where radius of the spheres was set to $a = 0.5$ au and multipole moments were included up to quadrupole ($n = 2$).

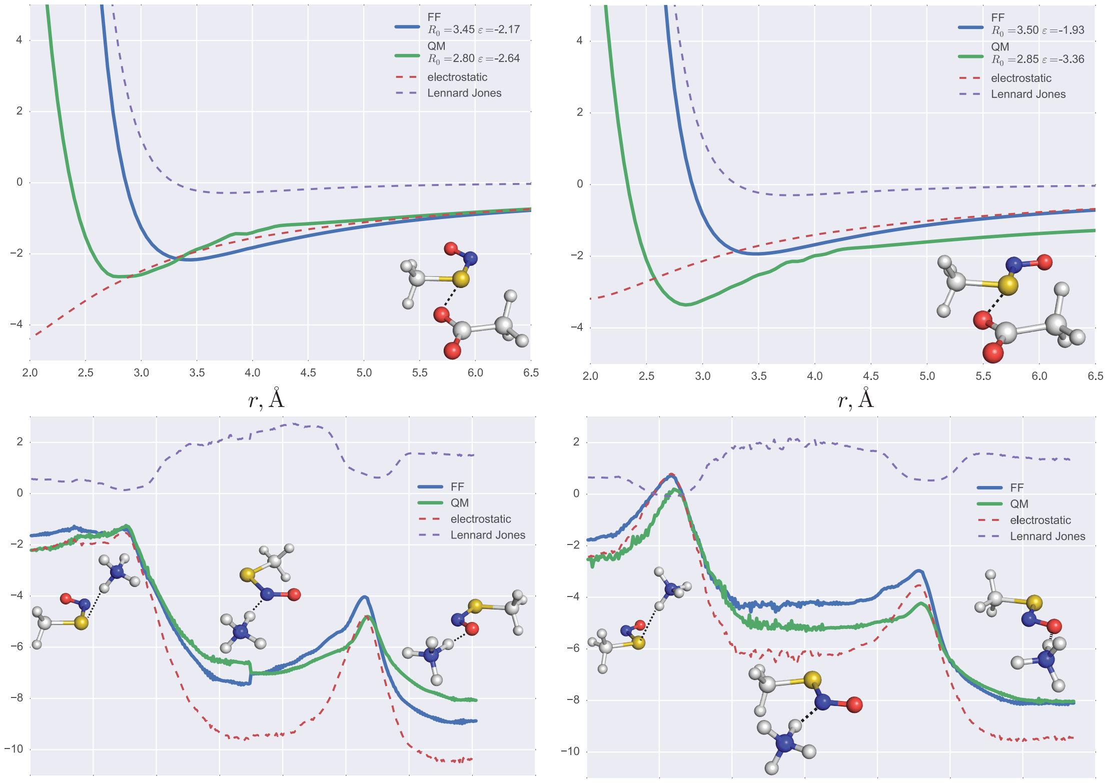


Figure 6.5: PES scans between MeSNO (cis on the left and trans on the right) and acetate anion $MeCOO^-$ (top) and the intrinsic reaction coordinate (IRC) profile along the minimum energy path between three hydrogen bonded complexes of MeSNO (cis on the left and trans on the right) and ammonium ion NH_4^+ (bottom). IRC path is calculated at PBE0/def2-TZVPPD at the geometries calculated using PBE0/def2-SV(P)+d level of theory.

The optimized point charge values, Lennard-Jones parameters and force constants (Tables 6.3 and 6.2) yeild interaction energy very closely matching the reference QM energies, with the root mean squared deviation (RMSD) of 0.64 kcal/mol and Pearson's correlation coefficient R^2 of 0.91. Especially good correlation was obtained for the attractive part of the PES, while the repulsion interactions resulted in a more scattered correlation due to the poor approximation of the short-range repulsion using the Lennard-Jones potential (Figure 6.4). To confirm that the optimized –SNO group force field corresponds to the physically meaningful parameters, the force field interaction energy was compared with the QM energies along the minimum energy path between three hydrogen-bonded complexes of MeSNO and ammonium ion NH_4^+ (Figure 6.5). The optimized force field reproduces stabilization of all three complexes as well the energy barrier between them. To verify that the sulfur atom force field parameters reproduce the formation of the chalcogen-bonded complex, force field interaction energy was compared against the QM PES scan along the $S \cdots O$ separation distance in the MeSNO $\cdots$ $MeCOO^-$ complex (Figure 6.4). The optimized force field underestimates the depth of the potential well by 0.5 kcal/mol for cis-MeSNO and by 1.43 kcal/mol for trans-MeSNO and overestimates the position of the minimum by 0.65 Åfor both conformers. This can be explained by the fact that force field parameters of the acetate anion were taken from the standard AMBER ff99SB library that are not optimized for this specific interaction.

Table 6.2: Optimized SN and NO force constants in MeSNO in case of separate and combined optimization of each conformer.

conformer	SN	NO
cis/tran	80.422	733.705
cis	80.739	733.693
trans	79.587	733.785

Table 6.3: Optimized non-bonded force field parameters of -SNO group in case of separate and combined optimization of each conformer.

atom	conformer	Q_{00}	Q_{11c}	Q_{11s}	Q_{22c}	Q_{22s}	Q_{20}	r_0	ε
O	cis/tran	-0.906	0.346	-0.296	-0.420	-0.002	0.097	1.961	0.002
	cis	-1.245	0.535	-0.611	-0.172	0.164	0.187	1.872	0.005
	trans	1.085	-0.262	1.641	-1.388	-0.587	-0.544	1.191	4.159
N	cis/tran	1.270	0.953	-0.896	-1.694	0.148	-1.128	2.239	0.002
	cis	1.500	0.993	-1.509	-1.979	-0.165	-1.253	2.081	0.005
	trans	-0.351	0.185	2.000	-1.104	1.996	-0.337	2.548	0.000
S	cis/tran	-0.478	0.264	0.428	-0.039	-0.017	-0.123	2.725	0.001
	cis	-0.369	0.023	0.386	0.506	0.591	-0.246	2.702	0.002
	trans	-0.848	0.328	1.216	1.325	-0.069	0.299	2.827	0.001

6.2 Summary

Here, in order to describe anisotropic character of the interaction between the –SNO group and the charged residues, atomic multipoles were fitted together with the Lennard-Jones parameters and S-N and N-O bonds force constants. To model the multipolar character of the electrostatic properties of the –SNO group, Lebedev charge model proposed in the Chapters 4 and 5 was used with the charged spheres centered at the atomic positions. We showed that the optimized –SNO group force field can accurately reproduce anisotropic interactions such as formation of hydrogen and chalcogen bonds. On the example of the interaction between MeSNO and the charged residue models, it is shown that the Lebedev charge model is a promising instrument to simulate specific interactions where the reproduction of anisotropy in the electrostatic potential is of the crucial importance.

6.3 Computational Details

Density functional theory (DFT) calculations were performed with the Gaussian 09 package[152] using Perdew-Burke-Ernzerhof hybrid functional (PBE0).[192,193] Double- and triple-ζ basis sets def2-SV(P) and def2-TZVPPD by Weigend and Ahlrichs[194] with diffuse functions by Rappoport and Furche[195] were obtained from the EMSL Basis Set Exchange Database.[196,197] The def2-SV(P) basis set was further augmented by a tight d function at the sulfur atom with $\zeta = 2.994$ and the resulting basis set is denoted as def2-SV(P)+d. Solvent effects were included using the implicit integral equation formalism polarizable continuum model (IEF-PCM)[198] with diethyl ether ($\varepsilon = 4.24$) parameters to mimic the protein environment.[199]

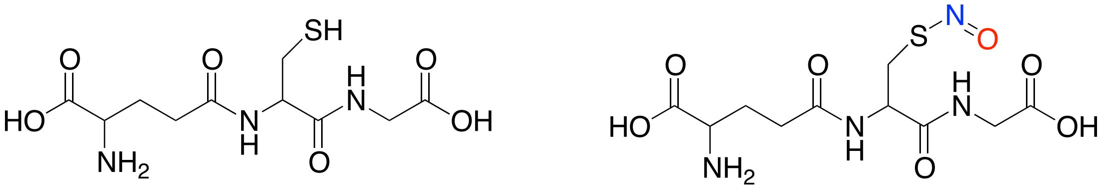
Sequential least squares programming (SLSQP) was used in the least-squares minimizations as implemented in *scipy* library.[153,183] SLSQP is a quasi-Newton method with a BFGS update of the $\mathbf{B}$ matrix (eq. 2.57) and can handle constraints and boundaries. In the force field optimization total charge of

MeSNO was constrained to the value of -0.1136 to be compatible with the CysNO residue, force constants and Lennard-Jones parameters were bounded within the range of positive values.

Chapter 7

GSNO Synthesis and NMR Spectroscopy

Glutathione (GSH), a tripeptide γ-glutamyl-cysteinyl-glycine, found to be S-nitrosated in vivo with formation of S-Nitrosoglutathione (GSNO, Figure 7.1). GSNO has been reported to be an integral part of the physiological function of nitric oxide.[75,76] Being the smallest biological S-nitrosothiol, GSNO is an ideal system to initiate the study of biological RSNOs using NMR spectroscopy.



Glutathione

S-Nitrosoglutathione

Figure 7.1: Structure of glutathione (GSH) and S-Nitrosoglutathione (GSNO).

GSNO can be easily synthesized from GSH and $NaNO_2$ under acidic conditions. The mechanism of GSNO formation, which is also true for any S-nitrosothiol, can be described by the following equations:[72,86]

$$\begin{gathered} \mathrm{NaNO_2 + H_2O \longrightarrow NaOH + HNO_2} \\ \mathrm{HNO_2 + H^+ \longrightarrow H_2NO_2{}^+} \\ \mathrm{GSH + H_2NO_2{}^+ \longrightarrow GSNO + H_2O + H^+} \end{gathered}$$

Formation of GSNO was verified by the appearance of pink colored solution and quantitatively by measuring light absorption at 335 nm (Figure 7.2) using $\varepsilon \sim 900\mathrm{M}^{-1}\mathrm{cm}^{-1}$. UV-vis spectrum has two characteristic peaks at 335 nm and 545 nm that are responsible for $n_{\mathrm{N}} \rightarrow \pi^*$ and $n_{\mathrm{O}} \rightarrow O\pi^*$ transition, respectively.

^{1}H-^{1}H TOCSY spectra of GSH and GSNO were provided by Dr. Sem's group from Concordia University, Wisconsin (Figure 7.3). The spectrum of GSH shows the peaks corresponding to cysteine and glutamate. The spectrum of GSNO is significantly different from the GSH spectrum. First, besides the peaks

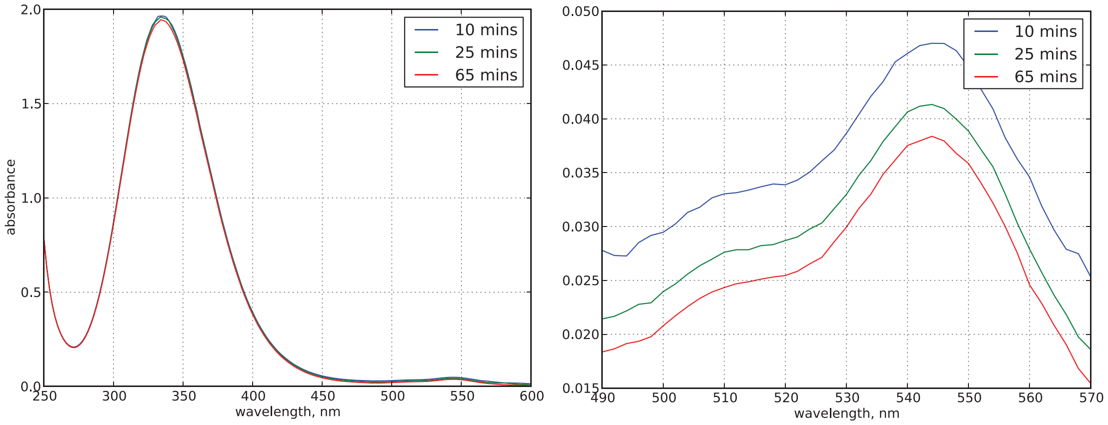


Figure 7.2: UV-vis spectra GSNO at different times after mixing.

Table 7.1: Predicted chemical shifts (in ppm) in CH_2 hydrogens of ethylSNO relative to the ethylSH at PBE0/pcS-2 level of theory using gauge-independent atomic orbital (GIAO) method.

conformation	calculations		experiment
cis	0.35	1.12	0.98-1.43
trans	3.90	2.01	

corresponding to GSNO there are also peaks from an unknown system, probably due to the presence of a contaminant.

Upon S-nitrosation, chemical shifts of H_α in cysteine are shifted downfield by 0.2 ppm, β protons of cysteine are split and shifted by 1.2 and 1.3 ppm. According to the literature, H_β of several RSNOs can be shifted downfield by 0.98-1.43 ppm, which is in agreement with the observed shifts for GSNO here. Quantum mechanical calculations of ethylSNO (a simple model for CysNO) predict H_β shifts in 0.35-3.90 range, depending on the conformation of ethylSNO (cis or trans), which is in relative agreement with the experimental findings.

The split in the cysteine β-protons suggests that the protons become non-equivalent upon S-nitrosation. Conformationally flexible GSH can lead to formation of stable conformation of GSNO where protonated glutamate amine is hydrogen-bonded to one of the atoms of the –SNO group. Krezel and Bal studied protonation macro-constants of the thiol and amine groups of GSH and revealed that several electrostatic self-interactions of GSH are possible and may be responsible for its structure and reactivity. One of the proposed interactions

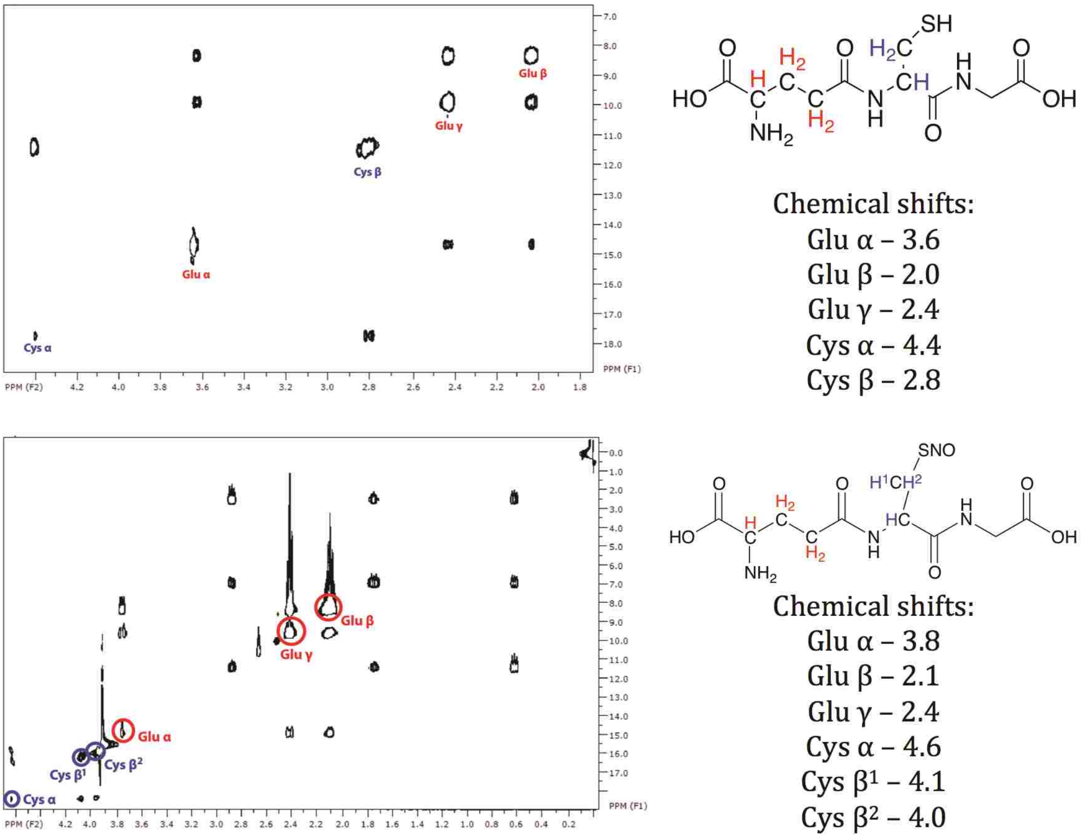


Figure 7.3: ^{1}H-^{1}H TOCSY spectra of 10mM GSH and 1 mM GSNO, both at pH 7.0 and room temperature.

is a direct interaction between protonated amine and deprotonated thiolate. By analogy, the identical interaction but with the –SNO group is possible in GSNO and probably is responsible for the surprising stability of GSNO, which is the most stable among all primary RSNOs, as well as it can explain its pink color—an exception in the series of red primary RSNOs. Possible self-stabilization of GSNO by protonated amine group is also supported by the resonance description of RSNO. If the positively charged ligand coordinates the –SNO group at nitrogen or oxygen, the double bond character of S-N bond increases, thus the stability of GSNO also increases. Besides, such coordination could simply protect the –SNO group from being decomposed by copper ions. However, these hypotheses require additional experimental validation.

Experimental and Computational Details

All reagents were purchased from Sigma-Aldrich. GSNO solution was prepared by mixing 40 mM GSH with 40 mM NaNO2 in 125 mM HCl, followed by a 10 min incubation period at room temperature in the dark. The stock GSNO solutions were prepared on the day of the experiment and kept on ice before use. The same samples of GSNO were used in UV-vis spectroscopy and ^{1}H-^{1}H total correlation spectroscopy (TOCSY). NMR experiments were performed on a 500 MHz Varian NMR System at 25 °C.

All calculations were performed using density functional theory (DFT) with the Gaussian 09 package[152] using Perdew-Burke-Ernzerhof hybrid functional (PBE0)[192,193] and polarization-consistent pcS-2 basis set by Jensen.[200] Isotropic shielding constant were calculated using the gauge-independent atomic orbital (GIAO) method.[201,202]

Chapter 8

Conclusions

Motivated by the limitations of the atom-centered point charge model currently used in the field of biomolecular simulations, this work provides in-depth analysis of the point charge approximation and offers a novel approach to describe electrostatic interactions. The proposed model can describe multipolar features of the molecular electrostatic potential (MEP) within computationally inexpensive point charge framework and paves the way toward efficient next-generation force fields.

Traditional atom-centered point charge approximation of the MEP not only fails to reproduce the complexity of the MEP but also is associated with numerical problems that arise during the least squares (LS) fitting. For example, slight changes in the setup of the charge fitting problem may significantly change the optimized charge values, especially in the case of buried methyl carbon atoms.

We show in Chapter 3 that this well-known effect becomes exacerbated in case of the genetic algorithm optimizations where several minimization runs converged to different but correlated solutions with the variations that are especially large for the buried atoms. Analysis of the covariance matrix for these scattered solutions revealed that the large variation of the optimized solutions is due to the wide range of the curvature values along the eigenvectors of the LS Hessian matrix. The solutions tend to be constrained along the eigenvectors with the largest curvatures and tend to be spread out along the eigenvectors with the smallest curvatures. Remarkably, the *stiff* large-curvature eigenvectors correspond to the first few multipole moments in the multipolar expansion of MEP (total charge, dipole moment), while the *sloppy* small-curvature eigenvectors largely depend on the buried atoms and do not bear any physical meaning.

In order to provide a physical interpretation of this observation, in Chapter 4 we considered the LS charge fitting problem from the first principles as an example of the inverse problem, opposed to the traditional view as being merely a statistical method of finding the best fit. Similarly to many other inverse problems, inverse electrostatic problem can be described by an integral equation, which in most cases can be solved only approximately using numerical techniques. However, we have shown that if the charge density is defined over a sphere and the reference electrostatic potential is defined over a larger outer sphere, the inverse electrostatic problem can be solved exactly. Availability of the exact solution provides an opportunity to demonstrate general properties of the charge fitting problem.

First, the exact solution reveals the origin of the underlying ill-conditioning of the charge fitting problem. Analysis of the singular values/vectors of the LS matrix reveals that the numerical instabilities associated with the LS point charge fitting are due to the decreasing contribution from higher multipoles to the overall electrostatic potential. The different sets of charges that yield the same first multipole moments and differ in the higher moments may equally well reproduce electrostatic potential of a molecule.

Second, analysis of the point charge LS problem suggests, that if the point charges are arranged over a sphere according to the Lebedev quadrature rule that exactly integrates spherical harmonics, the charge values can be obtained directly from multipole moments without fitting to the reference MEP. Importantly, such arrangement provides a systematic way to introduce any rank of multipole moments within the point charge approximation, which makes this model a computationally efficient analog to the multipolar formalism.

As an analog of the multipolar expansion, the Lebedev charge model can be also used in the multi-site expansions with expansion centers located at the positions of each atom in a molecule. In this respect, atom-centered Lebedev spheres provides a natural approach to expand the traditional atom-centered point charge approximation to include higher-rank multipoles. In Chapter 5 we demonstrated on a set of reference molecules that the atom-centered Lebedev spheres can reproduce MEP to the same accuracy as the multipole moments.

When the atomic multipoles are fitted together with the Lennard-Jones parameters, the resulting force field can accurately reproduce the anisotropic interactions such as hydrogen and chalcogen bonds. On the example of the interaction between MeSNO and the charged amino acid residue models, it is shown that the atom-centered charged spheres model is a promising instrument to simulate specific interactions where the reproduction of the anisotropy in the electrostatic potential is of the crucial importance.

Overall, the proposed Lebedev charge model can find its place in a variety of applications. For example, the model can be used in the development of the next-generation multipolar force fields for atomistic simulations. Since the point charge potentials are already used in the majority of the simulation packages, such implementation would require less technical difficulties as compared to the introduction of the actual multipolar formalism, especially in the case of the implementation of the boundary conditions. Besides atomistic simulations, the same approach can be applied to the coarse-grain simulations where charged spheres can model the electrostatic properties of groups of atoms, e.g. amino acid residues. Finally, a single charged sphere can be used to model electrostatic properties of small spherical molecules, e.g. drug candidates, solvent molecules, etc.

BIBLIOGRAPHY

(1) Stone, A., *Theory of intermolecular forces*; Oxford University Press Inc.: New York, NY, USA, 1996.

(2) Kaplan, I. G., *Intermolecular interactions: physical picture, computational methods and model potentials*; John Wiley Sons, Ltd: 2006.

(3) Sim, A. Y. L.; Minary, P.; Levitt, M. *Curr. Opin. Struct. Biol.* **2012**, *22*, 273–278.

(4) Kamerlin, S. C. L.; Vicatos, S.; Dryga, A.; Warshel, A. *Annu. Rev. Phys. Chem.* **2011**, *62*, 41–64.

(5) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

(6) Lopes, P. E. M.; Huang, J.; Shim, J.; Luo, Y.; Li, H.; Roux, B.; MacKerell, A. D. *J. Chem. Theory Comput.* **2013**, *9*, 5430–5449.

(7) MacKerell Jr., A. D.; Feig, M.; Brooks III., C. L. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

(8) MacKerell Jr., A. D. et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(9) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz Jr., K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(10) Wang, Z.-X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y. *J. Comput. Chem.* **2006**, *27*, 781–790.

(11) Cerutti, D. S.; Swope, W. C.; Rice, J. E.; Case, D. A. *J. Chem. Theory Comput.* **2014**, *10*, 4515–4534.

(12) Oostenbrink, C.; Soares, T. A.; van der Vegt, N. F. A.; van Gunsteren, W. F. *Eur. Biophys. J.* **2005**, *34*, 273–284.

(13) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(14) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.

(15) Cisneros, G. A.; Karttunen, M.; Ren, P.; Sagui, C. *Chem. Rev.* **2014**, *114*, 779–814.

(16) Stone, A. J. *J. Chem. Theory Comput.* **2005**, *1*, 1128–1132.

(17) Popelier, P. In *Intermolecular Forces and Clusters I*, Wales, D., Ed.; Structure and Bonding, Vol. 115; Springer Berlin Heidelberg: 2005, pp 1–56.

(18) Popelier, P. L. A.; Bremond, E. A. G. *Int. J. Quantum Chem.* **2009**, *109*, 2542–2553.

(19) Hirshfeld, F. English *Theor. Chim. Acta* **1977**, *44*, 129–138.

(20) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio Jr, R. A., et al. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.

(21) Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P. *J. Chem. Theory Comput.* **2007**, *3*, 1960–1986.

(22) Engkvist, O.; strand, P.-O.; Karlstrm, G. *J. Phys. Chem.* **1996**, *100*, 6950–6957.

(23) Bereau, T.; Kramer, C.; Monnard, F. W.; Nogueira, E. S.; Ward, T. R.; Meuwly, M. *J. Phys. Chem. B* **2013**, *117*, 5460–5471.

(24) Bereau, T.; Meuwly, M. *J. Phys. Chem. B* **2015**, *119*, 3034–3045.

(25) Bereau, T.; Kramer, C.; Meuwly, M. *J. Chem. Theory Comput.* **2013**, *9*, 5450–5459.

(26) Kramer, C.; Gedeck, P.; Meuwly, M. *J. Comput. Chem.* **2012**, *33*, 1673–88.

(27) Kramer, C.; Bereau, T.; Spinn, A.; Liedl, K. R.; Gedeck, P.; Meuwly, M. *J. Chem. Inf. Model.* **2013**, *53*, 3410–3417.

(28) Kramer, C.; Gedeck, P.; Meuwly, M. *J. Chem. Theory Comput.* **2013**, *9*, 1499–1511.

(29) Giese, T. J.; Panteva, M. T.; Chen, H.; York, D. M. *J. Chem. Theory Comput.* **2015**, *11*, 436–450.

(30) Giese, T. J.; Panteva, M. T.; Chen, H.; York, D. M. *J. Chem. Theory Comput.* **2015**, *11*, 451–461.

(31) Simmonett, A. C.; Pickard, F. C.; Schaefer, H. F.; Brooks, B. R. *J. Chem. Phys.* **2014**, *140* 184101, DOI: `http://dx.doi.org/10.1063/1.4873920`.

(32) Cox, S. R.; Williams, D. E. *J. Comput. Chem.* **1981**, *2*, 304–323.

(33) Singh, C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129–145.

(34) Besler, B. H.; Merz, K. M.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431–439.

(35) Woods, R. J.; Khalil, M.; Pell, W.; Moffat, S. H.; Smith, V. H. *J. Comput. Chem.* **1989**, *3*, 297–308.

(36) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(37) Francl, M. M.; Carey, C.; Chirlian, L. E.; Gange, D. M. *J. Comput. Chem.* **1996**, *17*, 367–383.

(38) Sigfridsson, E.; Ryde, U. *J. Comput. Chem.* **1998**, *19*, 377–395.

(39) Stouch, T. R.; Williams, D. E. *J. Comput. Chem.* **1992**, *13*, 622–632.

(40) Stouch, T. R.; Williams, D. E. *J. Comput. Chem.* **1993**, *14*, 858–866.

(41) Burger, S. K.; Schofield, J.; Ayers, P. W. *J. Phys. Chem. B* **2013**, *117*, 14960–14966.

(42) Arnautova, Y. A.; Jagielska, A.; Scheraga, H. A. *J. Phys. Chem. B* **2006**, *110*, 5025–5044.

(43) Zeng, J.; Duan, L. L.; Zhang, J. Z. H.; Mei, Y. *J. Comput. Chem.* **2013**, *34*, 847–853.

(44) Huang, L.; Roux, B. *J. Chem. Theory Comput.* **2013**, *9*, 3543–3556.

(45) Rai, B. K.; Bakken, G. A. *J. Comput. Chem.* **2013**, *34*, 1661–1671.

(46) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; Gonzalez-Outeirino, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2007**, *29*, 622–655.

(47) Seo, M.; Castillo, N.; Ganzynkowicz, R.; Daniels, C. R.; Woods, R. J.; Lowary, T. L.; Roy, P.-N. *J. Chem. Theory Comput.* **2008**, *4*, 184–191.

(48) Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1357–1377.

(49) Dupradeau, F.-Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7821–7839.

(50) Laio, A.; Gervasio, F. L.; VandeVondele, J.; Sulpizi, M.; Rothlisberger, U. *J. Phys. Chem. B* **2004**, *108*, 7963–7968.

(51) Graen, T.; Hoe, M.; Grubmu, H. *J. Chem. Theory Comput.* **2014**, *10*, 5505–5512.

(52) Vhringer-Martinez, E.; Verstraelen, T.; Ayers, P. W. *J. Phys. Chem. B* **2014**, *118*, 9871–9880.

(53) Cardamone, S.; Hughes, T. J.; Popelier, P. L. A. *Phys. Chem. Chem. Phys.* **2014**, *16*, 10367–10387.

(54) Kramer, C.; Spinn, A.; Liedl, K. R. *J. Chem. Theory Comput.* **2014**, *10*, 4488–4496.

(55) Politzer, P.; Lane, P.; Concha, M. C.; Ma, Y.; Murray, J. S. *J. Mol. Model.* **2007**, *13*, 305–311.

(56) Politzer, P.; Murray, J. S.; Concha, M. C. *J. Mol. Model.* **2007**, *13*, 643–650.

(57) Politzer, P.; Murray, J. S.; Clark, T. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7748.

(58) Politzer, P.; Murray, J. S. *Theor. Chem. Acc.* **2012**, *131*, 1114.

(59) Cavallo, G.; Metrangolo, P.; Milani, R.; Pilati, T.; Priimagi, A.; Resnati, G.; Terraneo, G. *Chem. Rev.* **2016**, *116*, 2478–2601.

(60) Karamertzanis, P. G.; Pantelides, C. C. *Mol. Simul.* **2004**, *30*, 413–436.

(61) Devereux, M.; Raghunathan, S.; Fedorov, D. G.; Meuwly, M. *J. Chem. Theory Comput.* **2014**, *10*, 4229–4241.

(62) Izadi, S.; Anandakrishnan, R.; Onufriev, A. V. **2014**.

(63) Anandakrishnan, R.; Baker, C.; Izadi, S.; Onufriev, A. V. *PLoS One* **Jan. 2013**, *8*, e67715.

(64) Duarte, F.; Bauer, P.; Barrozo, A.; Amrein, B. A.; Purg, M.; Aqvist, J.; Kamerlin, S. C. L. *J. Phys. Chem. B* **2014**, *118*, 4351–62.

(65) Li, W.; Grimme, S.; Krieg, H.; Mollmann, J.; Zhang, J. *J. Phys. Chem. C* **2012**, *116*, 8865–8871.

(66) Fischer, M.; Kuchta, B.; Firlej, L.; Hoffmann, F.; Froba, M. *J. Phys. Chem. C* **2010**, *114*, 19116–19126.

(67) McDaniel, J. G.; Yu, K.; Schmidt, J. R. *J. Phys. Chem. C* **2012**, *116*, 1892–1903.

(68) Chen, L.; Morrison, C. A.; Duren, T. *J. Phys. Chem. C* **2012**, *116*, 18899–18909.

(69) MacKerell Jr., A. D. *J. Comput. Chem.* **2004**, *25*, 1584–1604.

(70) Cerutti, D. S.; Rice, J. E.; Swope, W. C.; Case, D. A. *J. Phys. Chem. B* **Feb. 2013**, *117*, 2328–38.

(71) Gtz, A. W.; Bucher, D.; Lindert, S.; McCammon, J. A. *J. Chem. Theory Comput.* **Apr. 2014**, *10*, 1631–1637.

(72) Williams, D. L. H. *Acc. Chem. Res.* **1999**, *32*, 869–876.

(73) Smith, B. C.; Marletta, M. A. *Curr. Opin. Chem. Biol.* **2012**, *16*, Mechanisms Aesthetics Molecular imaging, 498–506.

(74) Broniowska, K. A.; Hogg, N. *Antioxid. Redox Signaling* **2012**, *17*, 969–980.

(75) Stamler, J. S.; Simon, D. I.; Osborne, J. A.; Mullins, M. E.; Jaraki, O.; Michel, T.; Singel, D. J.; Loscalzo, J. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 444–448.

(76) Liu, L.; Hausladen, A.; Zeng, M.; Que, L.; Heitman, J.; Stamler, J. S. *Nature* **2001**, *410*, 490–494.

(77) Seth, D.; Stamler, J. S. *Curr. Opin. Chem. Biol.* **2011**, *15*, Omics, 129–136.

(78) Doulias, P.-T.; Greene, J. L.; Greco, T. M.; Tenopoulou, M.; Seeholzer, S. H.; Dunbrack, R. L.; Ischiropoulos, H. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 16958–16963.

(79) Stamler, J. S.; Jia, L.; Eu, J. P.; McMahon, T. J.; Demchenko, I. T.; Bonaventura, J.; Gernert, K.; Piantadosi, C. A. *Science* **1997**, *276*, 2034–2037.

(80) Anand, P.; Stamler, J. S. *J. Mol. Med.* **2012**, *90*, 233–244.

(81) Mitchell, D. A.; Marletta, M. A. *Nat. Chem. Biol.* **2005**, *1*, 154–158.

(82) Bartberger, M. D.; Houk, K. N.; Powell, S. C.; Mannion, J. D.; Lo, K. Y.; Stamler, J. S.; Toone, E. J. *J. Am. Chem. Soc.* **2000**, *122*, 5889–5890.

(83) Arulsamy, N.; Bohle, D. S.; Butt, J. A.; Irvine, G. J.; Jordan, P. A.; Sagan, E. *J. Am. Chem. Soc.* **1999**, *121*, 7115–7123.

(84) Timerghazin, Q. K.; Peslherbe, G. H.; English, A. M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 1532–1539.

(85) Michael D. Bartberger, a. J. D. M.; Powell, S. C.; Stamler, J. S.; Houk, K. N.; Toone, E. J. *J. Am. Chem. Soc.* **2001**, *123*, 8868–8869.

(86) Williams, D. L. H. *Chem. Commun.* **1996**, 1085–1091.

(87) Timerghazin, Q. K.; Peslherbe, G. H.; English, A. M. *Org. Lett.* **2007**, *9*, PMID: 17616141, 3049–3052.

(88) Timerghazin, Q. K.; Talipov, M. R. *J. Phys. Chem. Lett.* **2013**, *4*, 1034–1038.

(89) Talipov, M. R.; Timerghazin, Q. K. *J. Phys. Chem. B* **2013**, *117*, 1827–1837.

(90) Flister, M.; Timerghazin, Q. K. *J. Phys. Chem. A* **2014**, *118*, 9914–9924.

(91) Baciu, C.; Cho, K.-B.; Gauld, J. W. *J. Phys. Chem. B* **2005**, *109*, 1334–1336.

(92) Toubin, C.; Yeung, D. Y. H.; English, A. M.; Peslherbe, G. H. *J. Am. Chem. Soc.* **2002**, *124*, 14816–14817.

(93) Laura L. Perissinotti,.; Estrin, D. A.; Leitus, G.; Doctorovich, F. *J. Am. Chem. Soc.* **2006**, *128*, 2512–2513.

(94) Abrams, A. J.; Farooq, A.; Wang, G. *Biochemistry* **2011**, *50*, 3405–3407.

(95) Oae, S.; Kim, Y. H.; Fukushima, D.; Shinhama, K. *J. Chem. Soc., Perkin Trans. 1* **1978**, 913–917.

(96) Park, J.-W. *Biochem. Biophys. Res. Commun.* **1988**, *152*, 916–920.

(97) Wong, P. S.-Y.; Hyun, J.; Fukuto, J. M.; Shirota, F. N.; DeMaster, E. G.; Shoeman, D. W.; Nagasawa, H. T. *Biochemistry* **1998**, *37*, 5362–5371.

(98) Dalle-Donne, I.; Rossi, R.; Colombo, G.; Giustarini, D.; Milzani, A. *Trends Biochem. Sci.* **2009**, *34*, 85–96.

(99) Flores-Santana, W.; Salmon, D. J.; Donzelli, S.; Switzer, C. H.; Basudhar, D.; Ridnour, L.; Cheng, R.; Glynn, S. A.; Paolocci, N.; Fukuto, J. M., et al. *Antioxid. Redox Signaling* **2011**, *14*, 1659–1674.

(100) Tocchetti, C. G.; Stanley, B. A.; Murray, C. I.; Sivakumaran, V.; Donzelli, S.; Mancardi, D.; Pagliaro, P.; Gao, W. D.; van Eyk, J.; Kass, D. A., et al. *Antioxid. Redox Signaling* **2011**, *14*, 1687–1698.

(101) Chen, Y.-Y.; Chu, H.-M.; Pan, K.-T.; Teng, C.-H.; Wang, D.-L.; Wang, A. H.-J.; Khoo, K.-H.; Meng, T.-C. *J. Biol. Chem.* **2008**, *283*, 35265–35272.

(102) Schreiter, E. R.; Rodrguez, M. M.; Weichsel, A.; Montfort, W. R.; Bonaventura, J. *J. Biol. Chem.* **2007**, *282*, 19773–19780.

(103) Weichsel, A.; Brailey, J. L.; Montfort, W. R. *Biochemistry* **2007**, *46*, 1219–1227.

(104) Han, S. *Biochem. Biophys. Res. Commun.* **2008**, *377*, 612–616.

(105) Lenari ivkovi, M.; Zarba-Kozio, M.; Zhukova, L.; Poznaski, J.; Zhukov, I.; Wysouch-Cieszyska, A. *J. Biol. Chem.* **2012**, *287*, 40457–40470.

(106) Stone, A. J. *J. Am. Chem. Soc.* **2013**, *135*, 7005–7009.

(107) Jackson, J. D., *Classical Electrodynamics*, 3rd ed.; John Wiley & Sons, Inc: 1999.

(108) Roman, S., *Advanced Linear Algebra*, 3rd ed.; Springer: New York, NY, USA, 2008.

(109) Tikhonov, A. N. *Dokl. Akad. Nauk SSSR* **1943**, *39*, 195–198.

(110) Phillips, D. L. *J. ACM* **1962**, *9*, 84–97.

(111) Hansen, P. C.; Pereyra, V.; Scherer, G., *Least Squares Data Fitting*; The Johns Hopkins University Press: Baltimore, MD, USA, 2013.

(112) Holland, J. H., *Adaptation in Natural and Artificial Systems*; MIT Press: Cambridge, MA, USA, 1992.

(113) Goldberg, D. E., *Genetic Algorithms in Search, Optimization and Machine Learning*, 1st; Addison-Wesley Longman Publishing Co., Inc.: Boston, MA, USA, 1989.

(114) Goldberg, D. E. *Complex Systems* **1990**, *5*, 139–167.

(115) Hansen, N.; Mller, S. D.; Koumoutsakos, P. *Evol. Comput.* **Mar. 2003**, *11*, 1–18.

(116) Hansen, N.; Ostermeier, A. *Evol. Comput.* **2001**, *9*, 159–195.

(117) Khoury, G. A.; Thompson, J. P.; Smadbeck, J.; Kieslich, C. A.; Floudas, C. A. *J. Chem. Theory Comput.* **2013**, *9*, 5653–5674.

(118) Wang, J.; Cieplak, P.; Li, J.; Hou, T.; Luo, R.; Duan, Y. *J. Phys. Chem. B* **2011**, *115*, 3091–3099.

(119) Wang, J.; Cieplak, P.; Li, J.; Wang, J.; Cai, Q.; Hsieh, M.; Lei, H.; Luo, R.; Duan, Y. *J. Phys. Chem. B* **2011**, *115*, 3100–3111.

(120) Wang, J.; Cieplak, P.; Cai, Q.; Hsieh, M.-J.; Wang, J.; Duan, Y.; Luo, R. *J. Phys. Chem. B* **2012**, *116*, 7999–8008.

(121) Wang, J.; Cieplak, P.; Li, J.; Cai, Q.; Hsieh, M.-J.; Luo, R.; Duan, Y. *J. Phys. Chem. B* **2012**, *116*, 7088–7101.

(122) Dickson, C. J.; Madej, B. D.; Skjevik, A. A.; Betz, R. M.; Teigen, K.; Gould, I. R.; Walker, R. C. *J. Chem. Theory Comput.* **2014**, *10*, 865–879.

(123) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(124) Wang, L.-P.; Martinez, T. J.; Pande, V. S. *J. Phys. Chem. Lett.* **2014**, *5*, 1885–1891.

(125) Wang, L.-P.; Chen, J.; Van Voorhis, T. *J. Chem. Theory Comput.* **Jan. 2013**, *9*, 452–460.

(126) Michalewicz, Z., *Genetic Algorithms + Data Structures = Evolution Programs (3rd Ed.)* Springer-Verlag: London, UK, UK, 1996.

(127) Wang, J.; Kollman, P. A. *J. Comput. Chem.* **2001**, *22*, 1219–1228.

(128) Betz, R. M.; Walker, R. C. *J. Comput. Chem.* **2015**, *36*, 79–87.

(129) Leonarski, F.; Trovato, F.; Tozzini, V.; Les, A.; Trylska, J. *J. Chem. Theory Comput.* **2013**, *9*, 4874–4889.

(130) Pahari, P.; Chaturvedi, S. *J. Mol. Model.* **2012**, *18*, 1049–1061.

(131) Larsson, H. R.; Duin, A. C. T.; Hartke, B. *J. Comput. Chem.* **2013**, *34*, 2178–2189.

(132) Strassner, T.; Busold, M.; Herrmann, W. A. *J. Comput. Chem.* **2002**, *23*, 282–290.

(133) Tafipolsky, M.; Schmid, R. *J. Phys. Chem. B* **2009**, *113*, 1341–1352.

(134) Courcot, B.; Bridgeman, A. J. *J. Comput. Chem.* **2011**, *32*, 240–247.

(135) Courcot, B.; Bridgeman, A. J. *J. Comput. Chem.* **2011**, *32*, 1703–1710.

(136) Chirlian, L. E.; Francl, M. M. *J. Comput. Chem.* **1987**, *8*, 894–905.

(137) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361–373.

(138) Momany, F. A. *J. Phys. Chem.* **1978**, *82*, 592–601.

(139) Tsiper, E. V.; Burke, K. *J. Chem. Phys.* **2004**, *120*, 1153–1156.

(140) Jakobsen, S.; Jensen, F. *J. Chem. Theory Comput.* **Dec. 2014**, *10*, 5493–5504.

(141) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollmann, P. A. *J. Am. Chem. Soc.* **1993**, *115*, 9620–9631.

(142) Hinsen, K.; Roux, B. *J. Comput. Chem.* **1997**, *18*, 368–380.

(143) Simmonett, A. C.; Gilbert, A. T. B.; Gill, P. M. W. *Mol. Phys.* **2005**, *103*, 2789–2793.

(144) Lawson, C. L.; Hanson, R. J., *Solving Least Squares Problems*; Series in Automatic Computation; Prentice-Hall: Englewood Cliffs, NJ 07632, USA, 1974,

(145) Herrera, F.; Lozano, M.; Verdegay, J. *Artificial Intelligence Review* **1998**, *12*, 265–319.

(146) Salomon, R. *Biosystems* **1996**, *39*, 263–278.

(147) Igel, C.; Hansen, N.; Roth, S. *Evol. Comput.* **Mar. 2007**, *15*, 1–28.

(148) Debiec, K. T.; Gronenborn, A. M.; Chong, L. T. *J. Phys. Chem. B* **2014**, *118*, 6561–6569.

(149) Becke, A. D. *J. Phys. Chem.* **Apr. 1993**, *98*, 5648–5652.

(150) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(151) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(152) Frisch, M. J. et al. Gaussian09 Revision D.01., Gaussian Inc. Wallingford CT 2009.

(153) Van der Walt, S.; Colbert, S.; Varoquaux, G. *Comput. Sci. Eng.* **Mar. 2011**, *13*, 22–30.

(154) Hunter, J. D. *Comput. Sci. Eng.* **2007**, *9*, 90–95.

(155) Arnautova, Y. A.; Abagyan, R.; Totrov, M. *J. Chem. Theory Comput.* **2015**, *11*, 2167–2186.

(156) Sprenger, K. G.; Jaeger, V.; Pfaendtner, J. *J. Phys. Chem. B* **2015**, *119*, 5882–5895.

(157) Mukhopadhyay, A.; Tolokh, I. S.; Onufriev, A. V. *J. Phys. Chem. B* **2015**, *119*, 6092–6100.

(158) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(159) Gabrieli, A.; Sant, M.; Demontis, P.; Suffritti, G. B. *J. Chem. Theory Comput.* **2015**, *11*, 3829–3843.

(160) Gao, Q.; Yokojima, S.; Fedorov, D. G.; Kitaura, K.; Sakurai, M.; Nakamura, S. *Chem. Phys. Lett.* **2014**, *593*, 165–173.

(161) Tschampel, S. M.; Kennerty, M. R.; Woods, R. J. *J. Chem. Theory Comput.* **Sept. 2007**, *3*, 1721–1733.

(162) Dixon, R. W.; Kollman, P. A. *J. Comput. Chem.* **1997**, *18.*

(163) Hansen, P. C., *Discrete Inverse Problems - Insight and Algorithms*; SIAM: Philadelphia, USA, 2010.

(164) Machta, B. B.; Chachra, R.; Transtrum, M. K.; Sethna, J. P. *Science* **2013**, *342*, 604–697.

(165) Transtrum, M. K.; Machta, B. B.; Brown, K. S.; Daniels, B. C.; Myers, C. R.; Sethna, J. P. *J. Chem. Phys.* **2015**, *143*, 010901.

(166) Ivanov, M. V.; Talipov, M. R.; Timerghazin, Q. K. *J. Phys. Chem. A* **2015**, *119*, 1422–1434.

(167) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Phys. Chem. B* **2002**, *106*, 7300–7307.

(168) Groetsch, C. W. *J. Phys.: Conf. Ser.* **2007**, *73*, 012001.

(169) Brown, K. S.; Sethna, J. P. *Phys. Rev. E* **2003**, *68*, 021904.

(170) Leang, S. S.; Zahariev, F.; Gordon, M. S. *J. Chem. Phys.* **2012**, *136*, 104101–104113.

(171) Steinmann, S. N.; Piemontesi, C.; Delachat, A.; Corminboeuf, C. *J. Chem. Theory Comput.* **2012**, *8*, 1629–1640.

(172) Parrish, R. M.; Hohenstein, E. G.; Martnez, T. J.; Sherrill, C. D. *J. Chem. Phys.* **2013**, *138*, 194107–194122.

(173) Lebedev, V. I.; Laikov, D. *Doklady Mathematics* **1999**, *59*, 477–481.

(174) Ahrens, C.; Beylkin, G. *Proc. R. Soc. A* **2009**, *465*, 3103–3125.

(175) Lebedev, V. I. *USSR Comp. Math. Math.* **1976**, *16*, 10–24.

(176) Sloan, I. *J. Approx. Theory* **1995**, *83*, 238–254.

(177) Rogers, D. M. *J. Chem. Phys.* **2015**, *142*, 074101–074111.

(178) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

(179) Shao, Y. et al. *Mol. Phys.* **2015**, *113*, 184–215.

(180) PyQuante: Python Quantum Chemistry., `https://github.com/gabrielelanaro/pyquante`.

(181) PyQuante2: Python Quantum Chemistry., `https://github.com/rpmuller/pyquante2`.

(182) Force Field Tool Box: Python library to optimize force fields for molecular mechanics., `https://github.com/maxivanoff/fftoolbox`, 2014–.

(183) Jones, E.; Oliphant, T.; Peterson, P., et al. SciPy: Open source scientific tools for Python., [Online; accessed 2016-03-23], 2001–.

(184) Baron, R.; McCammon, J. A. *Annu. Rev. Phys. Chem.* **2013**, *64*, 151–175.

(185) Jimnez-Moreno, E.; Gmez, A. M.; Bastida, A.; Corzana, F.; Jimnez-Oses, G.; Jimnez-Barbero, J.; Asensio, J. L. *Angewandte Chemie* **2015**, *127*, 4418–4422.

(186) Gresh, N.; Sponer, J. E.; Spakov, N.; Leszczynski, J.; Sponer, J. *J. Phys. Chem. B* **2003**, *107*, 8669–8681.

(187) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. *J. Chem. Theory Comput.* **2013**, *9*, 4046–4063.

(188) Zhang, C.; Lu, C.; Wang, Q.; Ponder, J. W.; Ren, P. *J. Chem. Theory Comput.* **2015**, *11*, 5326–5339.

(189) Marshall, G. R. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 107–14.

(190) Ivanov, M. V.; Talipov, M. R.; Timerghazin, Q. K. *J. Chem. Phys.* **2015**, *143*, 134102.

(191) Rafat, M.; Popelier, P. L. A. *J. Chem. Phys.* **2005**, *123*, 204103.

(192) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982–9985.

(193) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(194) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(195) Rappoport, D.; Furche, F. *J. Chem. Phys.* **2010**, *133*, 134105.

(196) Schuchardt, K. L.; Didier, B. T.; Elsethagen, T.; Sun, L.; Gurumoorthi, V.; Chase, J.; Li, J.; Windus, T. L. *J. Chem. Inf. Model.* **2007**, *47*, 1045–1052.

(197) Feller, D. *J. Comput. Chem.* **1996**, *17*, 1571–1586.

(198) Cances, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032–3041.

(199) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T., et al. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(200) Jensen, F. *J. Chem. Theory Comput.* **2008**, *4*, 719–727.

(201) Wolinski, K.; Hinton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251–8260.

(202) Gauss, J. *Chem. Phys. Lett.* **1992**, *191*, 614–620.

Appendices

Appendix A

Statistical Data on the Electrostatic Properties of the Model

To compare point charges models with the reference potential we used root mean square deviation (RMSD), root mean absolute error (RMAE), Pearson's correlation coefficient (R^2), maximum error, proportionality coefficient α in $\Phi_n^{PC} = \alpha\Phi^{ref} + \beta$, all of which were computed over the molecular electrostatic potential (MEP) grid and over the atomic electrostatic potential with the grid points defined as being the closest to the corresponding atom in the molecule:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_i^N \left[\Phi_n^{PC}(\mathbf{r}_i) - \Phi^{ref}(\mathbf{r}_i)\right]^2}{T}} \tag{A.1}$$

$$\mathrm{RMAE} = \frac{\sum_i^N \left|\Phi_n^{PC}(\mathbf{r}_i) - \Phi^{ref}(\mathbf{r}_i)\right|}{\sum_i^N \left|\Phi^{ref}(\mathbf{r}_i)\right|} \tag{A.2}$$

$$R^2 = \frac{\left[\sum_i^N \left(\Phi_n^{PC}(\mathbf{r}_i) - \overline{\Phi_n^{PC}}\right)\left(\Phi^{ref}(\mathbf{r}_i) - \overline{\Phi^{ref}}\right)\right]^2}{\sum_i^N \left(\Phi_n^{PC}(\mathbf{r}_i) - \overline{\Phi_n^{PC}}\right)^2 \sum_i^N \left(\Phi^{ref}(\mathbf{r}_i) - \overline{\Phi^{ref}}\right)^2} \tag{A.3}$$

A.1 Ammonia

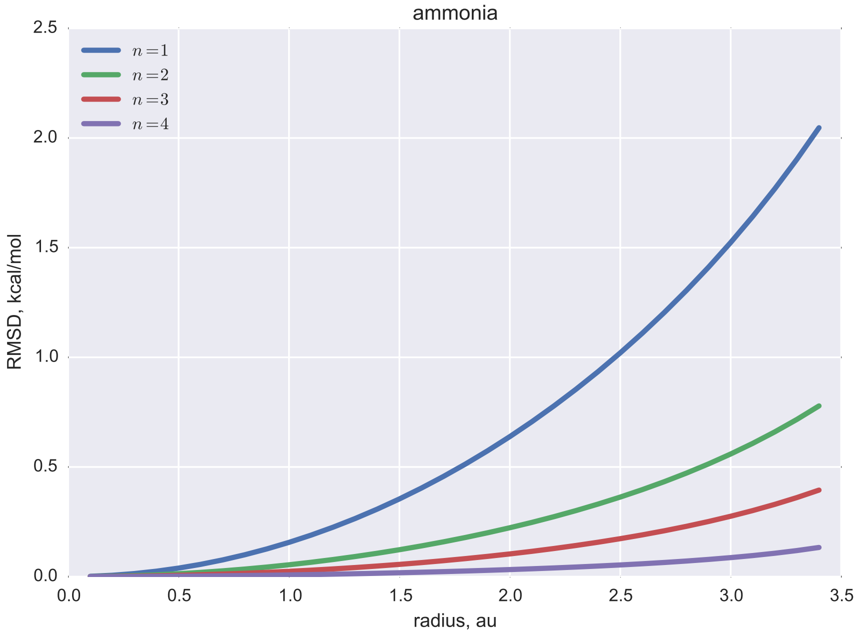


Figure A-A1: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

A.2 Bromomethane

Table A-B1: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	0.5	1.1	1.0	1.4

Table A-D1: Molecular multipole moments Q_{lm} of ammonia calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	-0.634	0.000	0.000	-2.258	0.000	0.000	0.000	0.000
AC PC	0.000	-0.660	0.000	0.000	-1.254	0.000	0.000	0.001	0.000
SS LM	0.000	-0.634	0.000	0.000	-2.258	0.000	0.000	0.000	0.000
MS LM	0.000	-0.634	0.000	0.000	-2.258	0.000	0.000	0.000	0.000

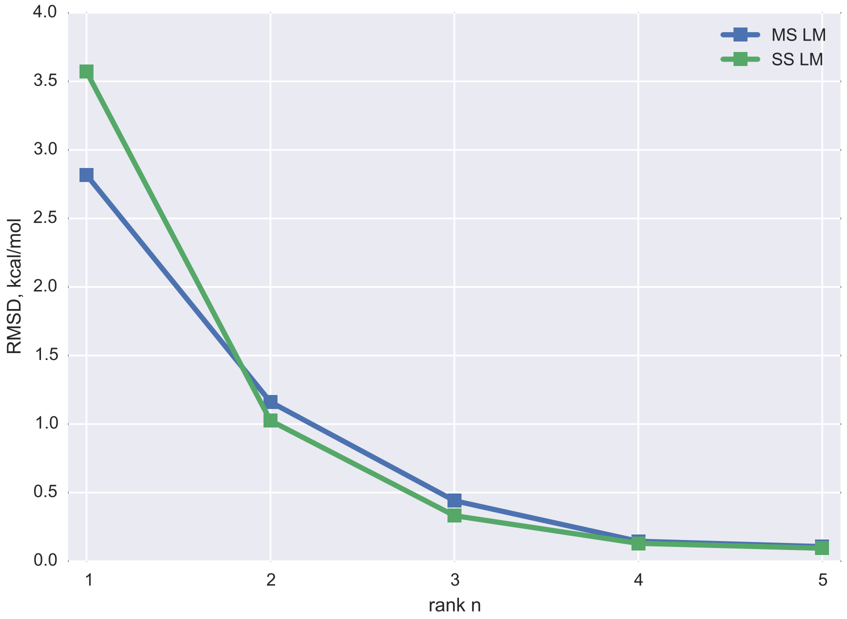


Figure A-E1: Convergence of the RMSD between ammonia QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F1: Comparison of ammonia QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	5.703	0.198	0.958	1.564	0.958
2	SS LM	3.326	0.125	0.983	1.025	1.001
3	SS LM	1.755	0.039	0.998	0.332	1.004
2	MS LM	3.357	0.147	0.978	1.161	1.001
3	MS LM	1.672	0.051	0.997	0.441	1.006

Table A-G1: Comparison of ammonia QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	N_1	AC PC	3.445	0.113	0.985	1.802	0.782
2	N_1	SS LM	1.772	0.042	0.995	0.737	1.140
3	N_1	SS LM	0.977	0.013	0.997	0.261	1.008
2	N_1	MS LM	2.863	0.067	0.991	1.171	1.212
3	N_1	MS LM	1.627	0.016	0.989	0.367	0.968

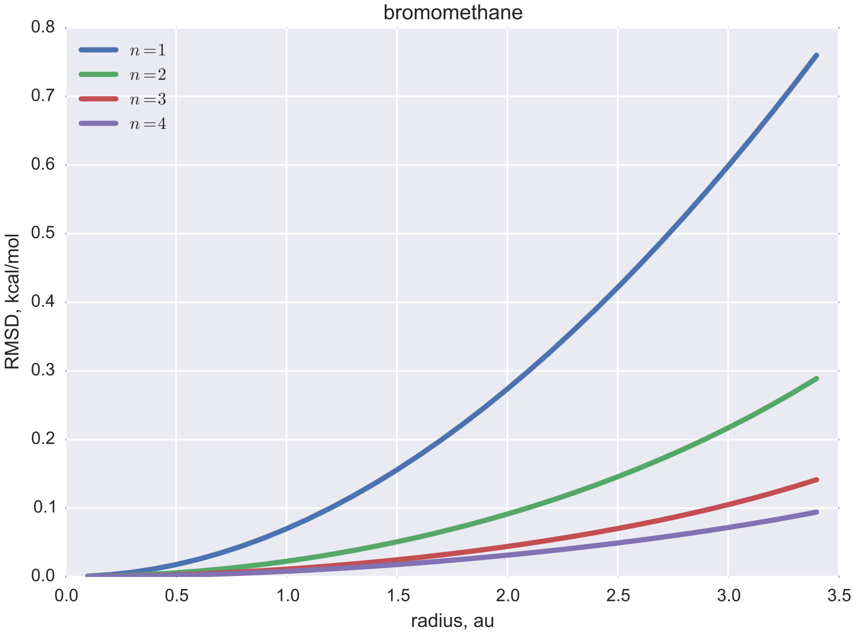


Figure A-A2: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B2: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	0.8	1.5	2.1	2.5

Table A-D2: Molecular multipole moments Q_{lm} of bromomethane calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.000	0.834	0.000	-1.370	0.000	0.000	2.371	0.000
AC PC	0.000	0.000	0.857	-0.001	-1.219	-0.006	0.003	2.110	-0.008
SS LM	0.000	0.000	0.834	0.000	-1.370	0.000	0.000	2.371	0.000
MS LM	0.000	0.000	0.834	0.000	-1.370	0.000	0.000	2.371	0.000

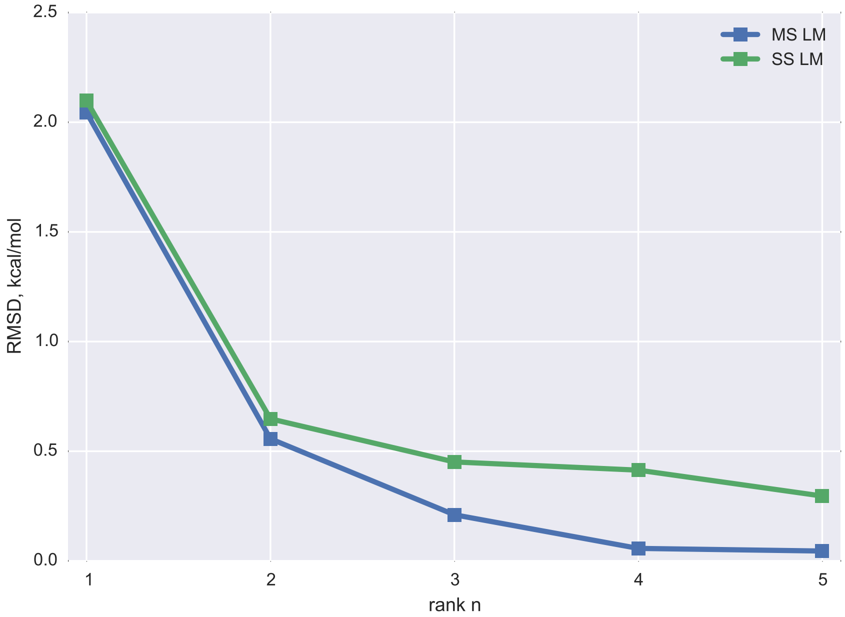


Figure A-E2: Convergence of the RMSD between bromomethane QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F2: Comparison of bromomethane QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	3.498	0.132	0.976	0.837	0.977
2	SS LM	2.827	0.102	0.986	0.648	0.972
3	SS LM	2.008	0.065	0.993	0.451	1.005
2	MS LM	2.726	0.082	0.990	0.557	0.980
3	MS LM	1.176	0.027	0.999	0.210	0.997

Table A-G2: Comparison of bromomethane QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	Br_5	AC PC	3.498	0.161	0.328	1.017	0.582
2	Br_5	SS LM	1.780	0.081	0.800	0.507	0.785
3	Br_5	SS LM	1.323	0.040	0.932	0.312	1.027
2	Br_5	MS LM	0.857	0.050	0.925	0.311	0.915
3	Br_5	MS LM	0.190	0.009	0.997	0.059	1.004

A.3 Chloromethane

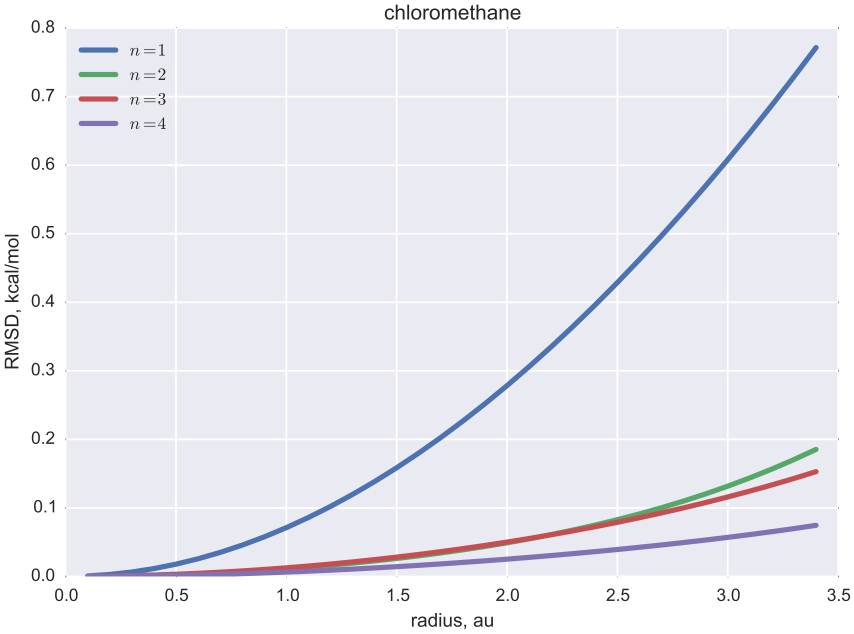


Figure A-A3: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B3: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	0.8	2.0	2.0	2.8

Table A-D3: Molecular multipole moments Q_{lm} of chloromethane calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.000	0.837	0.000	-0.640	-0.001	0.000	1.104	0.001
AC PC	0.000	0.000	0.851	0.001	-0.611	0.002	0.001	1.059	0.006
SS LM	0.000	0.000	0.837	0.000	-0.640	-0.001	0.000	1.104	0.001
MS LM	0.000	0.000	0.837	0.000	-0.640	-0.001	0.000	1.104	0.001

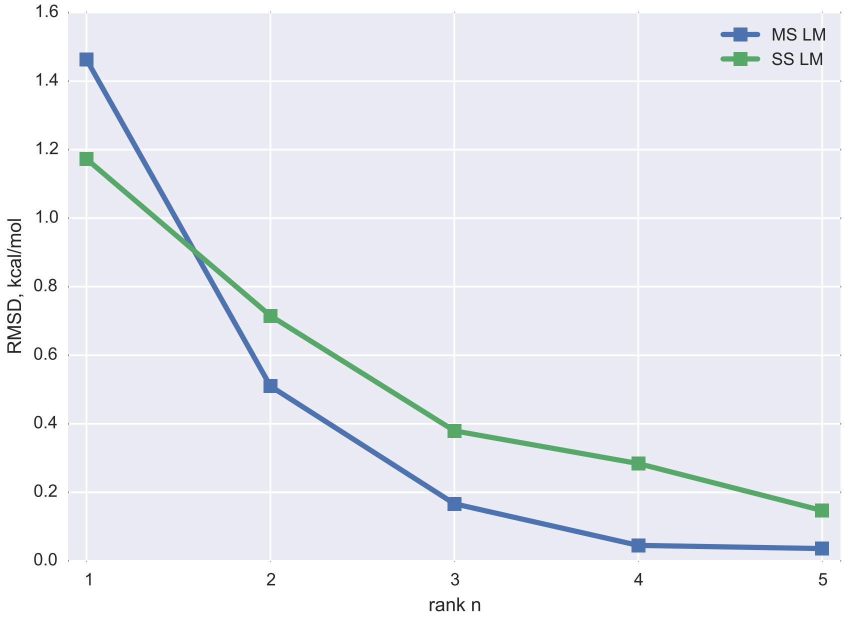


Figure A-E3: Convergence of the RMSD between chloromethane QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F3: Comparison of chloromethane QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	2.341	0.098	0.988	0.604	0.988
2	SS LM	2.946	0.112	0.983	0.715	0.971
3	SS LM	1.717	0.054	0.995	0.380	1.004
2	MS LM	2.442	0.071	0.991	0.510	0.988
3	MS LM	0.943	0.021	0.999	0.167	0.998

Table A-G3: Comparison of chloromethane QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	Cl_5	AC PC	2.341	0.106	0.709	0.684	0.956
2	Cl_5	SS LM	2.000	0.096	0.772	0.609	0.989
3	Cl_5	SS LM	1.250	0.033	0.943	0.261	0.927
2	Cl_5	MS LM	0.633	0.033	0.966	0.220	1.028
3	Cl_5	MS LM	0.240	0.007	0.998	0.046	0.997

A.4 cis-MeSNO

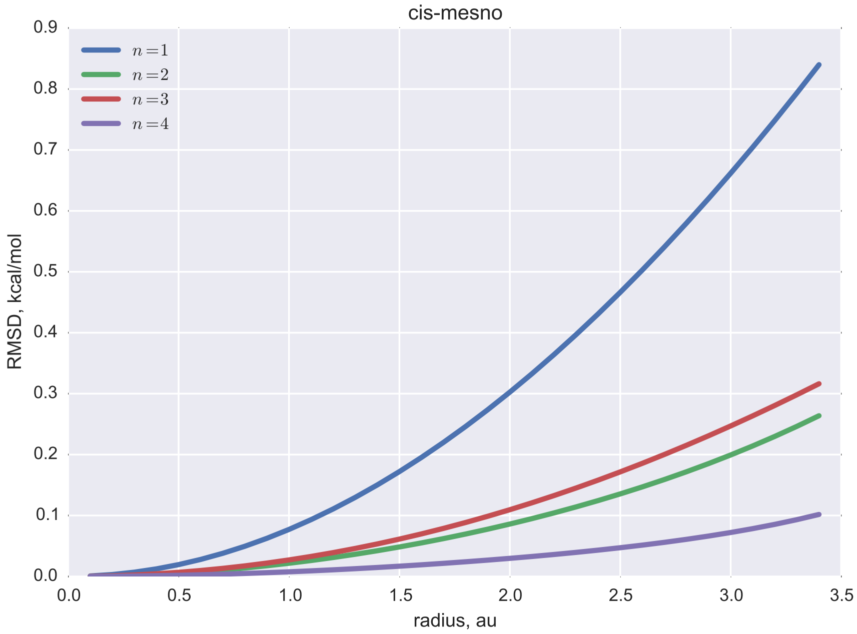


Figure A-A4: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B4: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	0.8	1.5	1.3	2.6

Table A-D4: Molecular multipole moments Q_{lm} of cis-mesno calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.000	0.937	0.305	0.582	0.000	0.000	2.668	-0.773
AC PC	0.003	0.000	0.919	0.291	-0.351	0.000	0.000	2.787	-1.191
SS LM	0.000	0.000	0.937	0.305	0.582	0.000	0.000	2.668	-0.773
MS LM	0.000	0.000	0.937	0.305	0.582	0.000	0.000	2.668	-0.773

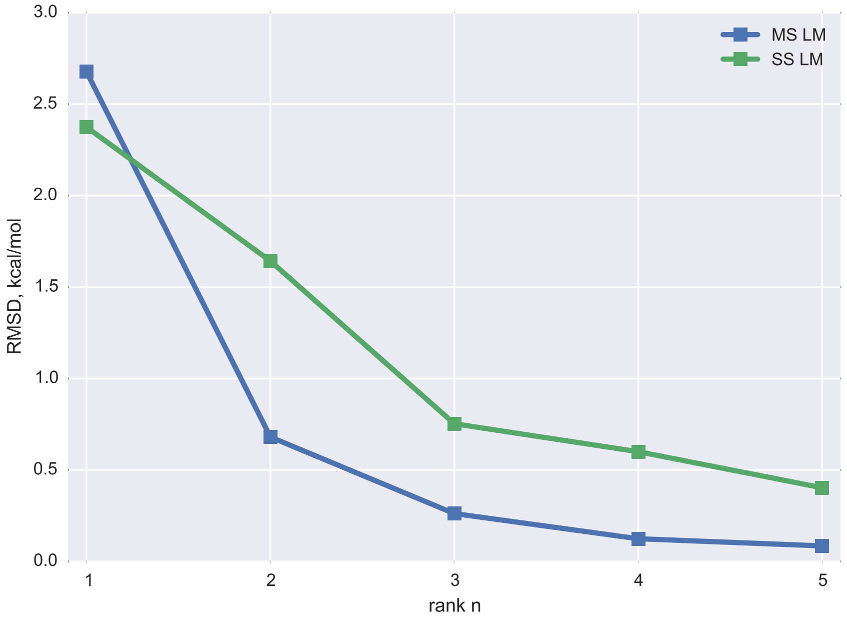


Figure A-E4: Convergence of the RMSD between cis-mesno QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F4: Comparison of cis-mesno QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	6.843	0.217	0.944	1.396	0.944
2	SS LM	6.916	0.254	0.926	1.642	0.978
3	SS LM	7.137	0.101	0.984	0.753	0.999
2	MS LM	3.047	0.105	0.987	0.680	0.989
3	MS LM	1.439	0.037	0.998	0.262	0.998

Table A-G4: Comparison of cis-mesno QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	O_1	AC PC	4.702	0.180	0.736	1.472	0.625
–	N_2	AC PC	5.707	0.243	0.622	1.670	0.494
–	S_3	AC PC	3.344	0.483	0.802	1.130	0.798
2	O_1	SS LM	5.957	0.198	0.637	1.723	0.519
2	N_2	SS LM	6.916	0.258	0.649	1.734	0.546
2	S_3	SS LM	4.662	0.590	0.627	1.468	0.955
3	O_1	SS LM	2.884	0.069	0.960	0.656	1.084
3	N_2	SS LM	3.096	0.120	0.902	0.867	1.002
3	S_3	SS LM	1.887	0.177	0.970	0.441	1.119
2	O_1	MS LM	1.747	0.072	0.962	0.609	0.861
2	N_2	MS LM	1.123	0.061	0.985	0.394	0.956
2	S_3	MS LM	1.401	0.223	0.930	0.526	0.918
3	O_1	MS LM	0.868	0.031	0.992	0.264	1.017
3	N_2	MS LM	1.051	0.038	0.993	0.265	0.949
3	S_3	MS LM	0.426	0.045	0.996	0.118	0.989

Table A-H4: Atom-centered point charge values of cis-mesno fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule.

O_1	N_2	S_3	C_4	H_5	H_6	H_7
-0.15	-0.01	-0.00	-0.25	0.15	0.13	0.13

Table A-I4: Atomic coordinates of cis-mesno optimized at mp2/aug-cc-pVTZ level of theory.

#	Element	x, Å	y, Å	z, Å
1	O	-1.435	-2.789	0.000
2	N	-2.272	-0.659	0.000
3	S	0.000	1.778	0.000
4	C	2.854	-0.053	0.000
5	H	4.418	1.279	0.000
6	H	2.923	-1.235	1.680
7	H	2.923	-1.235	-1.680

A.5 Fluoromethane

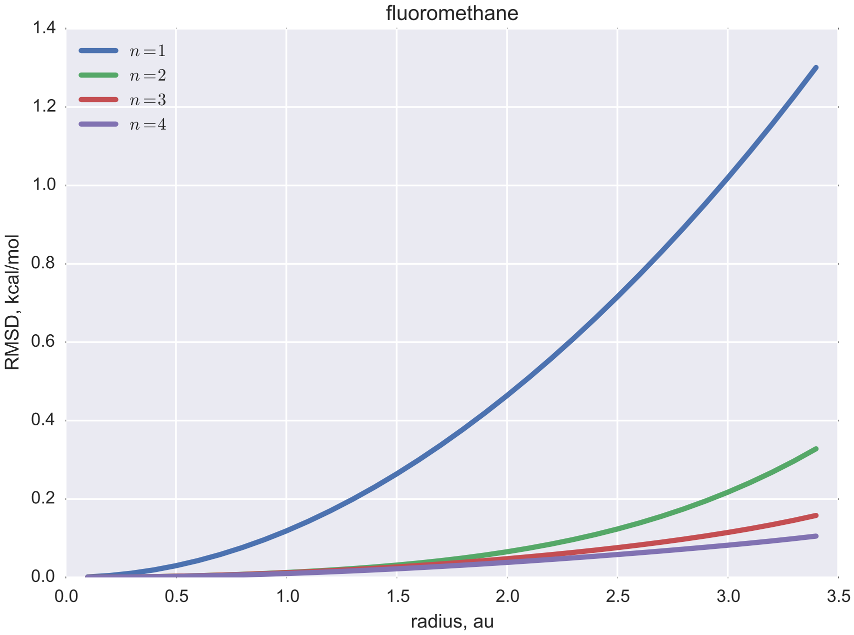


Figure A-A5: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B5: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	0.6	1.8	2.0	2.3

Table A-D5: Molecular multipole moments Q_{lm} of fluoromethane calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.000	0.816	0.000	0.352	0.000	0.000	-0.613	0.000
AC PC	0.000	0.000	0.821	0.000	0.061	0.000	0.000	-0.108	0.000
SS LM	0.000	0.000	0.816	0.000	0.352	0.000	0.000	-0.613	0.000
MS LM	0.000	0.000	0.816	0.000	0.352	0.000	0.000	-0.613	0.000

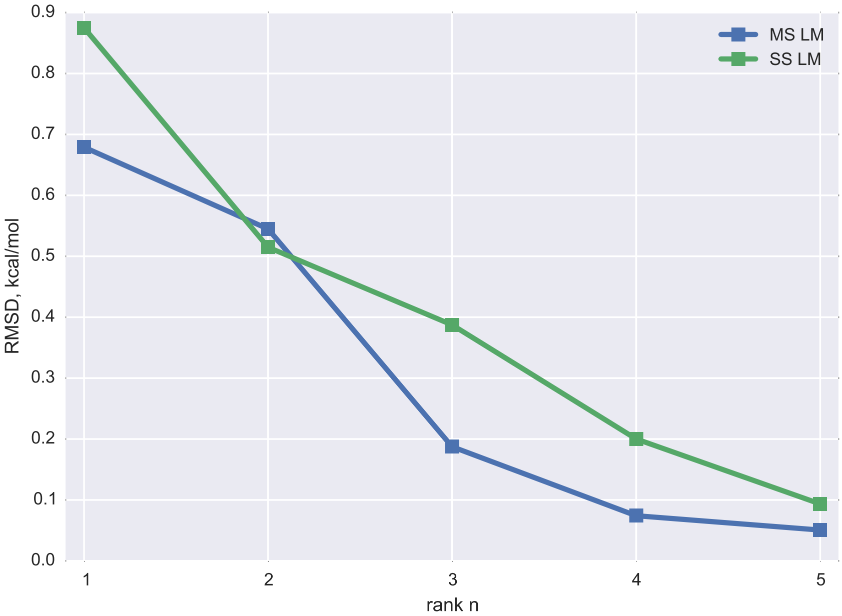


Figure A-E5: Convergence of the RMSD between fluoromethane QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F5: Comparison of fluoromethane QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	2.043	0.099	0.990	0.700	0.990
2	SS LM	1.979	0.063	0.994	0.515	0.996
3	SS LM	2.029	0.046	0.997	0.387	1.002
2	MS LM	2.067	0.069	0.994	0.545	1.003
3	MS LM	0.878	0.022	0.999	0.187	0.999

Table A-G5: Comparison of fluoromethane QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	F_5	AC PC	1.032	0.053	0.995	0.542	0.868
2	F_5	SS LM	0.708	0.017	0.997	0.212	1.052
3	F_5	SS LM	2.029	0.030	0.979	0.413	0.954
2	F_5	MS LM	0.844	0.025	0.991	0.275	0.963
3	F_5	MS LM	0.878	0.019	0.993	0.230	0.985

A.6 Formamide

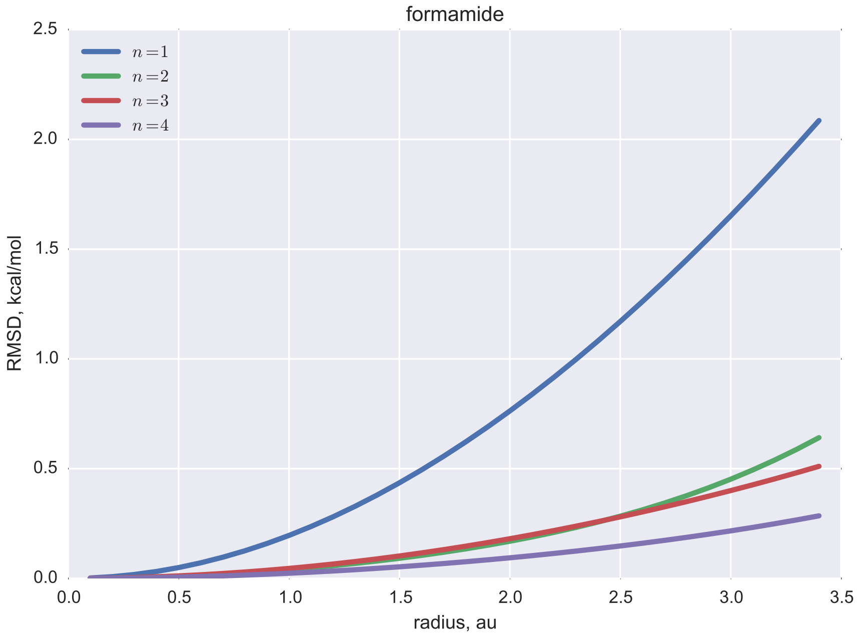


Figure A-A6: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B6: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	0.5	1.1	1.0	1.4

Table A-D6: Molecular multipole moments Q_{lm} of formamide calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.000	-1.684	-0.167	-1.671	0.000	0.000	-2.872	-0.425
AC PC	0.000	0.000	-1.682	-0.159	-1.636	0.000	0.000	-3.005	0.102
SS LM	0.000	0.000	-1.684	-0.167	-1.671	0.000	0.000	-2.872	-0.425
MS LM	0.000	0.000	-1.684	-0.167	-1.671	0.000	0.000	-2.872	-0.425

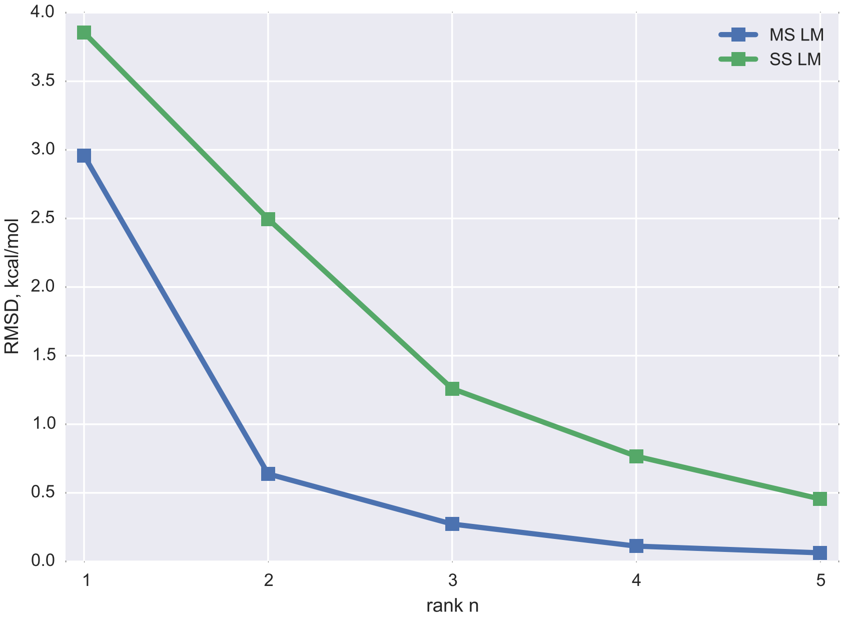


Figure A-E6: Convergence of the RMSD between formamide QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F6: Comparison of formamide QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	2.477	0.048	0.997	0.646	0.997
2	SS LM	8.695	0.178	0.963	2.497	1.016
3	SS LM	5.586	0.086	0.990	1.260	1.000
2	MS LM	2.771	0.045	0.997	0.638	1.003
3	MS LM	1.372	0.018	1.000	0.273	1.002

Table A-G6: Comparison of formamide QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	N_2	AC PC	0.886	0.194	0.985	0.401	1.192
–	O_3	AC PC	2.477	0.039	0.975	0.756	0.995
2	N_2	SS LM	5.984	1.405	0.848	2.854	1.902
2	O_3	SS LM	7.880	0.119	0.765	2.433	0.670
3	N_2	SS LM	5.548	1.632	0.749	2.957	1.323
3	O_3	SS LM	2.623	0.037	0.983	0.746	1.059
2	N_2	MS LM	1.216	0.319	0.981	0.591	1.092
2	O_3	MS LM	1.115	0.019	0.994	0.372	0.982
3	N_2	MS LM	0.685	0.243	0.991	0.426	0.945
3	O_3	MS LM	0.411	0.004	1.000	0.095	1.005

Table A-H6: Atom-centered point charge values of formamide fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule.

C_1	N_2	O_3	H_4	H_5	H_6
0.64	-0.93	-0.59	0.04	0.44	0.39

Table A-I6: Atomic coordinates of formamide optimized at mp2/aug-cc-pVTZ level of theory.

#	Element	x, Å	y, Å	z, Å
1	C	0.000	0.795	0.000
2	N	-1.782	-1.051	0.000
3	O	2.271	0.423	0.000
4	H	-0.827	2.702	0.000
5	H	-1.235	-2.871	0.000
6	H	-3.629	-0.626	0.000

A.7 Furan

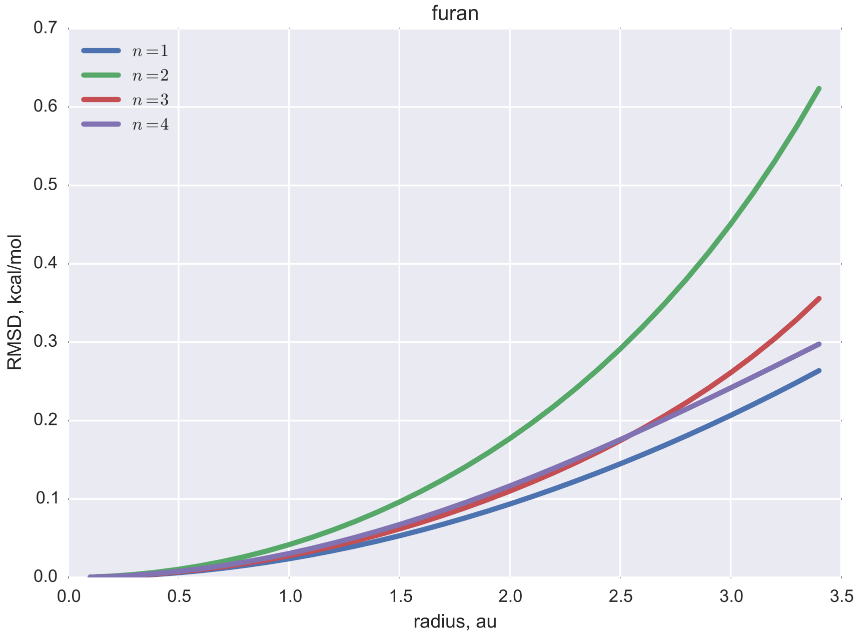


Figure A-A7: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B7: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	1.4	1.1	1.3	1.3

Table A-D7: Molecular multipole moments Q_{lm} of furan calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.000	0.000	0.302	-4.767	0.000	0.001	2.541	0.000
AC PC	0.000	0.000	0.000	0.294	-4.687	0.000	0.000	2.642	0.000
SS LM	0.000	0.000	0.000	0.302	-4.767	0.000	0.001	2.541	0.000
MS LM	0.000	0.000	0.000	0.302	-4.767	0.000	0.001	2.541	0.000

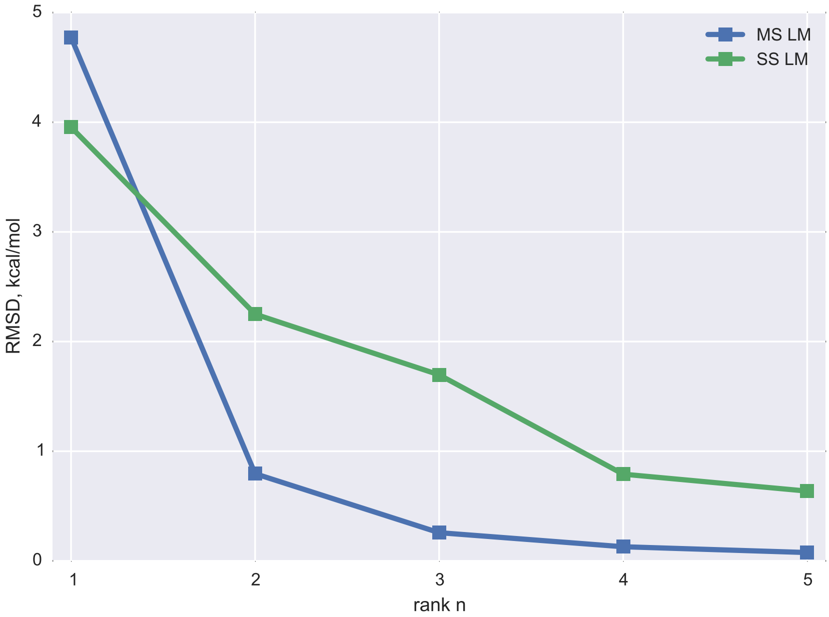


Figure A-E7: Convergence of the RMSD between furan QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F7: Comparison of furan QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	3.205	0.161	0.967	0.784	0.967
2	SS LM	11.242	0.461	0.771	2.250	0.956
3	SS LM	10.917	0.313	0.880	1.697	1.058
2	MS LM	2.825	0.167	0.968	0.795	1.015
3	MS LM	1.243	0.052	0.997	0.257	1.008

Table A-G7: Comparison of furan QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	O_5	AC PC	3.205	0.186	0.359	1.368	0.839
2	O_5	SS LM	11.242	0.470	0.061	3.540	0.750
3	O_5	SS LM	8.952	0.140	0.453	1.496	1.107
2	O_5	MS LM	2.680	0.121	0.728	0.975	1.123
3	O_5	MS LM	0.567	0.025	0.988	0.204	1.032

A.8 Imidazol

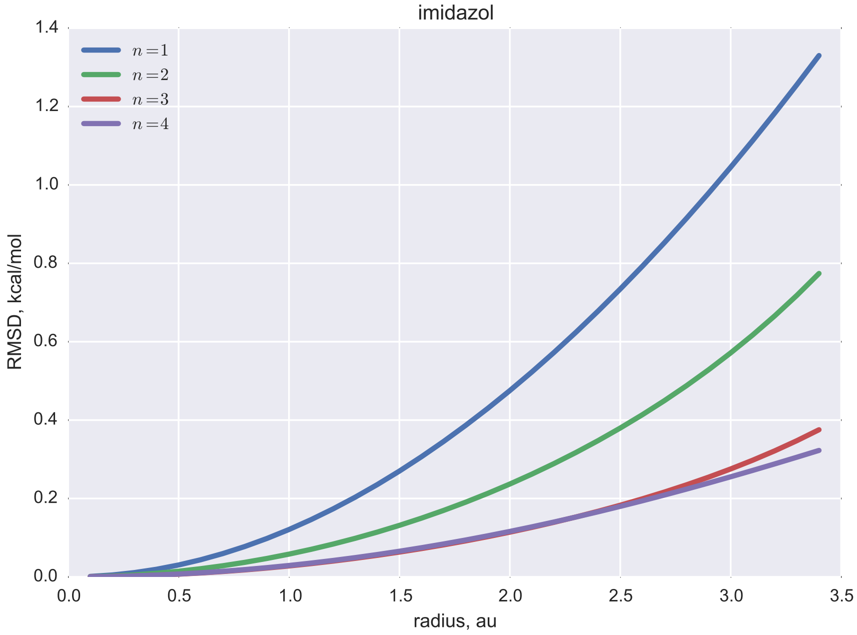


Figure A-A8: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B8: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	0.6	0.9	1.3	1.3

Table A-D8: Molecular multipole moments Q_{lm} of imidazol calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.000	0.411	1.466	-4.709	0.000	0.000	-0.474	-4.615
AC PC	0.000	0.000	0.411	1.454	-4.630	0.000	0.000	-0.370	-4.688
SS LM	0.000	0.000	0.411	1.466	-4.709	0.000	0.000	-0.474	-4.615
MS LM	0.000	0.000	0.411	1.466	-4.709	0.000	0.000	-0.474	-4.615

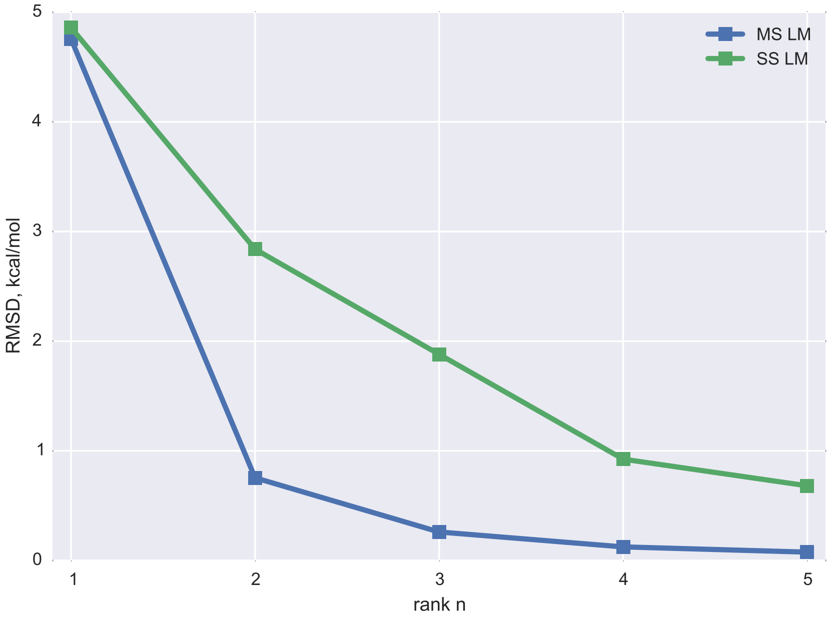


Figure A-E8: Convergence of the RMSD between imidazol QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F8: Comparison of imidazol QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	3.581	0.085	0.992	0.915	0.992
2	SS LM	12.419	0.267	0.930	2.839	1.006
3	SS LM	11.651	0.158	0.969	1.880	1.018
2	MS LM	2.603	0.071	0.995	0.754	1.008
3	MS LM	1.127	0.024	0.999	0.260	1.004

Table A-G8: Comparison of imidazol QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	N_5	AC PC	3.581	0.065	0.933	1.301	0.777
–	N_8	AC PC	2.932	0.300	0.993	0.925	1.311
2	N_5	SS LM	12.419	0.204	0.139	4.074	0.268
2	N_8	SS LM	10.713	0.912	0.979	2.936	2.237
3	N_5	SS LM	7.081	0.040	0.941	1.077	0.842
3	N_8	SS LM	11.323	1.914	0.886	5.078	1.964
2	N_5	MS LM	1.926	0.030	0.982	0.644	0.945
2	N_8	MS LM	2.119	0.493	0.979	1.215	1.141
3	N_5	MS LM	0.653	0.010	0.999	0.209	1.022
3	N_8	MS LM	0.580	0.111	0.996	0.296	0.915

Table A-H8: Atom-centered point charge values of imidazol fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule.

C_1	H_2	C_3	H_4	N_5	C_6	H_7	N_8	H_9
-0.43	0.24	0.14	0.12	-0.51	0.13	0.13	-0.12	0.29

Table A-I8: Atomic coordinates of imidazol optimized at mp2/aug-cc-pVTZ level of theory.

#	Element	x, Å	y, Å	z, Å
1	C	2.111	0.577	0.000
2	H	3.993	1.343	0.000
3	C	1.199	-1.860	0.000
4	H	2.270	-3.590	0.000
5	N	-1.396	-1.870	0.000
6	C	-2.061	0.538	0.000
7	H	-3.965	1.256	0.000
8	N	0.000	2.082	0.000
9	H	-0.024	3.983	0.000

A.9 Methanesulfonamide

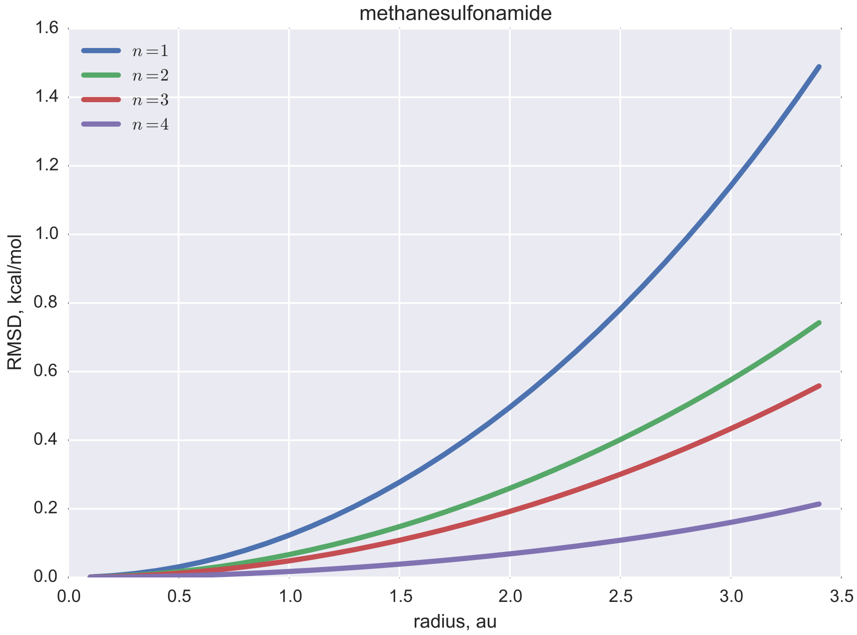


Figure A-A9: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B9: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	0.6	0.8	1.0	1.7

Table A-D9: Molecular multipole moments Q_{lm} of methanesulfonamide calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.011	-0.161	-1.525	-5.586	0.011	-0.013	7.591	-2.417
AC PC	0.000	0.011	-0.160	-1.523	-5.955	0.012	-0.016	7.726	-2.097
SS LM	0.000	0.011	-0.161	-1.525	-5.586	0.011	-0.013	7.591	-2.417
MS LM	0.000	0.011	-0.161	-1.525	-5.586	0.011	-0.013	7.591	-2.417

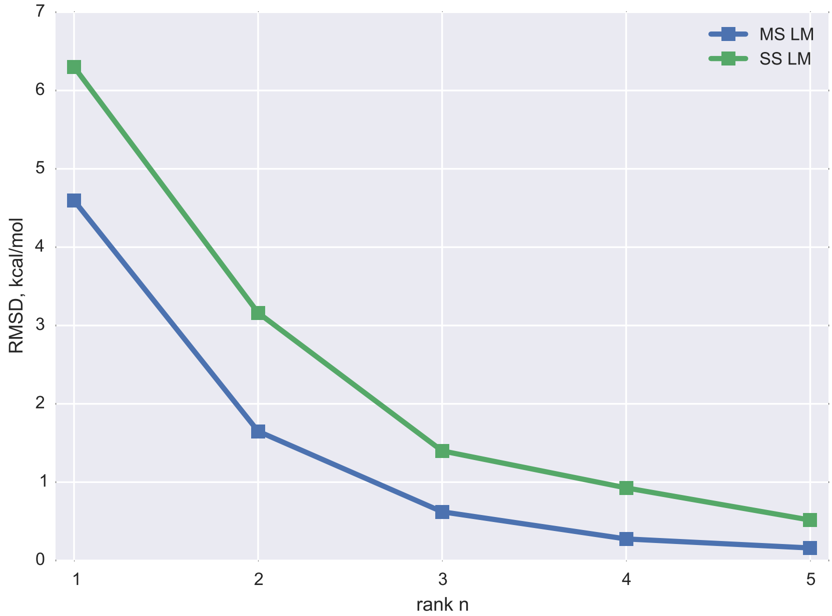


Figure A-E9: Convergence of the RMSD between methanesulfonamide QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F9: Comparison of methanesulfonamide QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	2.554	0.040	0.998	0.504	0.998
2	SS LM	15.135	0.253	0.915	3.162	0.966
3	SS LM	7.877	0.106	0.983	1.399	0.997
2	MS LM	6.331	0.140	0.976	1.650	0.988
3	MS LM	3.677	0.048	0.997	0.622	1.005

Table A-G9: Comparison of methanesulfonamide QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	O_2	AC PC	1.007	0.022	0.993	0.357	1.005
–	O_7	AC PC	1.026	0.021	0.994	0.355	1.003
–	N_8	AC PC	1.641	0.250	0.982	1.005	0.683
2	O_2	SS LM	13.269	0.183	0.772	2.958	1.193
2	O_7	SS LM	15.135	0.201	0.755	3.395	1.266
2	N_8	SS LM	11.427	1.997	0.399	7.874	-0.755
3	O_2	SS LM	5.549	0.057	0.945	1.105	1.047
3	O_7	SS LM	5.864	0.063	0.934	1.241	1.046
3	N_8	SS LM	3.424	0.394	0.433	1.801	0.973
2	O_2	MS LM	5.966	0.090	0.906	1.460	1.041
2	O_7	MS LM	5.988	0.120	0.849	1.895	0.987
2	N_8	MS LM	2.218	0.251	0.675	1.111	0.914
3	O_2	MS LM	1.996	0.026	0.990	0.431	0.992
3	O_7	MS LM	2.801	0.041	0.977	0.689	0.995
3	N_8	MS LM	1.642	0.191	0.951	0.803	1.223

Table A-H9: Atom-centered point charge values of methanesulfonamide fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule.

S_1	O_2	C_3	H_4	H_5	H_6	O_7	N_8	H_9	H_{10}
1.13	-0.57	-0.56	0.21	0.21	0.21	-0.56	-0.85	0.39	0.39

Table A-I9: Atomic coordinates of methanesulfonamide optimized at mp2/aug-cc-pVTZ level of theory.

#	Element	x, Å	y, Å	z, Å
1	S	0.077	0.311	-0.002
2	O	0.353	1.599	-2.410
3	C	-2.899	-1.223	0.010
4	H	-3.034	-2.358	1.714
5	H	-4.314	0.267	-0.001
6	H	-3.035	-2.380	-1.679
7	O	0.354	1.633	2.387
8	N	2.081	-2.126	0.015
9	H	3.157	-2.077	-1.569
10	H	3.159	-2.053	1.595

A.10 Methanesulfonic acid

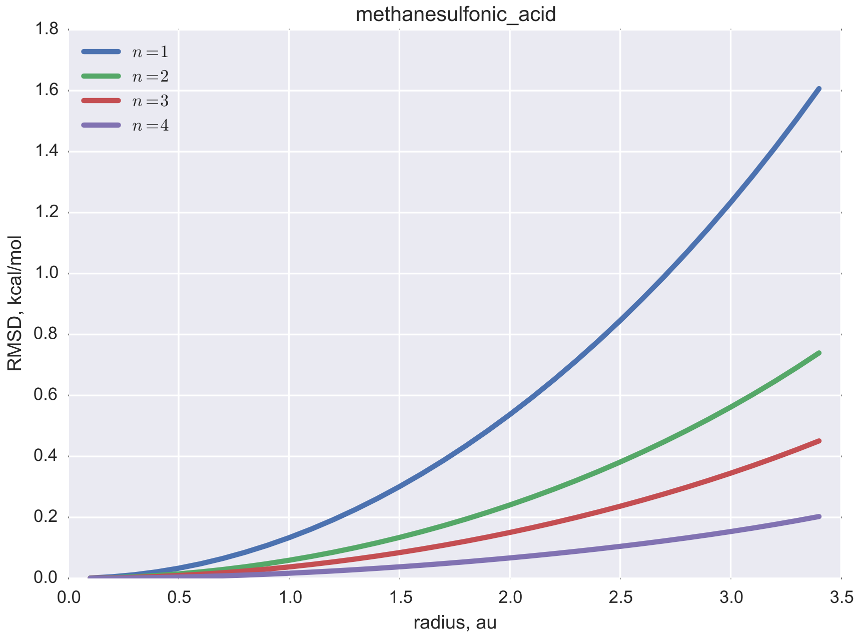


Figure A-A10: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B10: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	0.6	0.9	1.1	1.7

Table A-D10: Molecular multipole moments Q_{lm} of methanesulfonic acid calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.169	0.990	-1.349	-6.192	-1.244	-1.766	4.126	3.014
AC PC	0.000	0.172	0.992	-1.343	-6.325	-1.185	-1.819	4.133	2.992
SS LM	0.000	0.169	0.990	-1.349	-6.192	-1.244	-1.766	4.126	3.014
MS LM	0.000	0.169	0.990	-1.349	-6.192	-1.244	-1.766	4.126	3.014

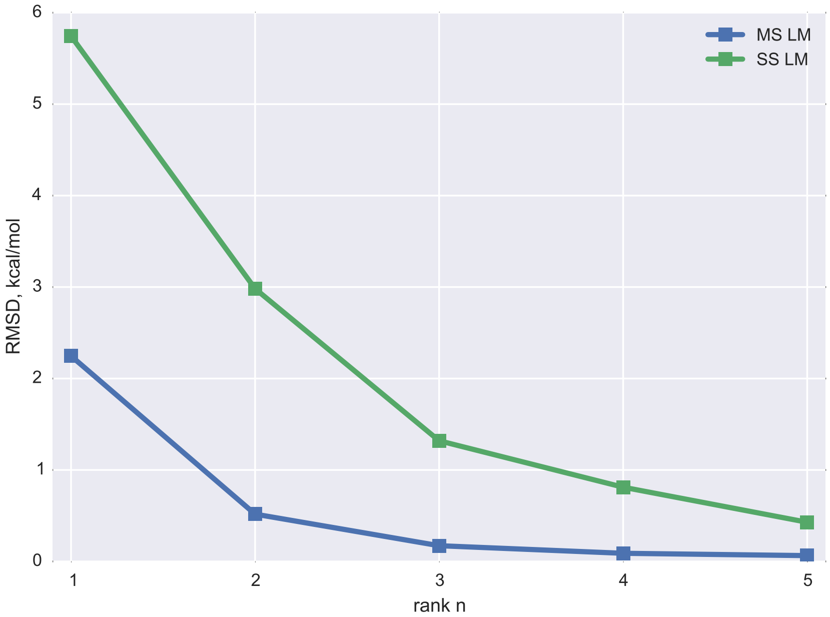


Figure A-E10: Convergence of the RMSD between methanesulfonic acid QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F10: Comparison of methanesulfonic acid QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	2.978	0.031	0.999	0.402	0.999
2	SS LM	17.860	0.222	0.931	2.982	0.975
3	SS LM	11.210	0.095	0.987	1.320	1.008
2	MS LM	2.131	0.039	0.998	0.517	0.998
3	MS LM	1.124	0.012	1.000	0.172	1.001

Table A-G10: Comparison of methanesulfonic acid QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	O_2	AC PC	1.278	0.019	0.993	0.305	0.998
–	O_7	AC PC	1.205	0.028	0.996	0.344	1.026
–	O_8	AC PC	1.910	0.075	0.994	0.475	1.033
2	O_2	SS LM	11.465	0.160	0.656	2.847	1.058
2	O_7	SS LM	10.170	0.243	0.808	2.915	1.170
2	O_8	SS LM	11.336	0.684	0.885	3.794	1.340
3	O_2	SS LM	4.161	0.049	0.945	0.898	0.985
3	O_7	SS LM	4.026	0.077	0.973	0.962	1.087
3	O_8	SS LM	4.730	0.209	0.970	1.219	1.119
2	O_2	MS LM	0.965	0.018	0.994	0.295	1.014
2	O_7	MS LM	0.998	0.021	0.997	0.267	0.997
2	O_8	MS LM	1.197	0.065	0.998	0.384	1.042
3	O_2	MS LM	0.792	0.005	0.999	0.104	0.991
3	O_7	MS LM	0.605	0.006	1.000	0.093	0.998
3	O_8	MS LM	1.051	0.030	0.999	0.191	0.982

Table A-H10: Atom-centered point charge values of methanesulfonic acid fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule.

S_1	O_2	C_3	H_4	H_5	H_6	O_7	O_8	H_9
1.12	-0.51	-0.69	0.22	0.26	0.26	-0.55	-0.53	0.42

Table A-I10: Atomic coordinates of methanesulfonic acid optimized at mp2/aug-cc-pVTZ level of theory.

#	Element	x, Å	y, Å	z, Å
1	S	-0.163	0.260	-0.134
2	O	-0.404	2.690	1.067
3	C	3.034	-0.699	0.025
4	H	3.188	-2.594	-0.743
5	H	4.098	0.644	-1.108
6	H	3.622	-0.623	1.989
7	O	-1.141	-0.210	-2.649
8	O	-1.458	-1.793	1.730
9	H	-2.471	-2.890	0.666

A.11 Methanethiol

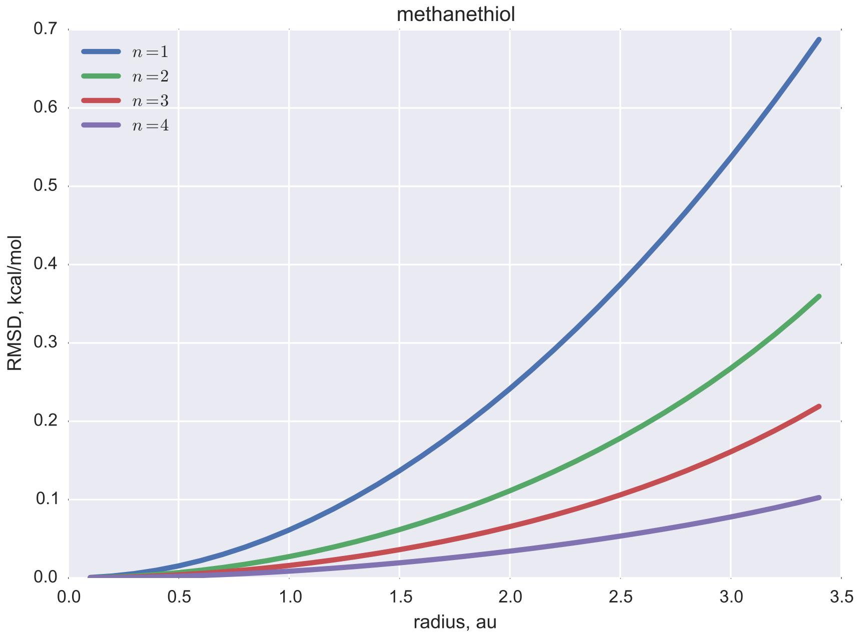


Figure A-A11: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B11: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	0.9	1.3	1.7	2.4

Table A-D11: Molecular multipole moments Q_{lm} of methanethiol calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.000	0.298	0.588	-2.107	0.000	0.000	0.830	-1.640
AC PC	0.000	0.000	0.300	0.602	-1.372	0.000	0.000	0.103	-2.150
SS LM	0.000	0.000	0.298	0.588	-2.107	0.000	0.000	0.830	-1.640
MS LM	0.000	0.000	0.298	0.588	-2.107	0.000	0.000	0.830	-1.640

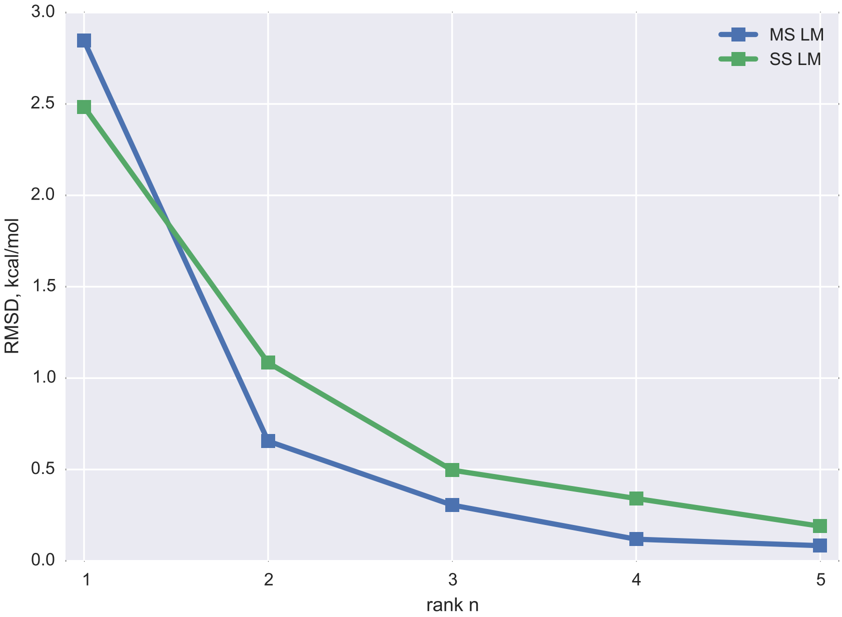


Figure A-E11: Convergence of the RMSD between methanethiol QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F11: Comparison of methanethiol QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	5.259	0.244	0.925	1.416	0.925
2	SS LM	4.070	0.183	0.956	1.084	0.964
3	SS LM	2.158	0.080	0.991	0.496	1.010
2	MS LM	2.855	0.106	0.984	0.655	0.985
3	MS LM	1.686	0.051	0.997	0.305	1.001

Table A-G11: Comparison of methanethiol QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	S_1	AC PC	4.526	0.189	0.237	1.665	0.426
2	S_1	SS LM	2.896	0.105	0.729	0.949	0.672
3	S_1	SS LM	2.032	0.034	0.965	0.404	1.064
2	S_1	MS LM	1.164	0.037	0.961	0.336	0.986
3	S_1	MS LM	0.902	0.027	0.979	0.247	0.999

A.12 Methanol

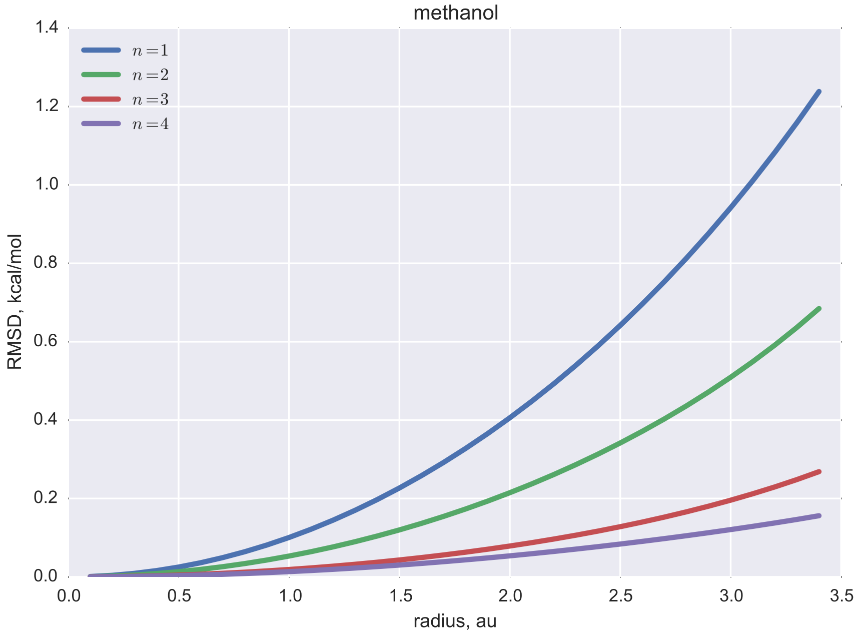


Figure A-A12: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B12: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	0.7	0.9	1.6	1.9

Table A-D12: Molecular multipole moments Q_{lm} of methanol calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.000	0.568	0.431	-0.849	0.000	0.000	1.170	-2.761
AC PC	0.000	0.000	0.561	0.436	-1.055	0.000	0.000	0.499	-2.855
SS LM	0.000	0.000	0.568	0.431	-0.849	0.000	0.000	1.170	-2.761
MS LM	0.000	0.000	0.568	0.431	-0.849	0.000	0.000	1.170	-2.761

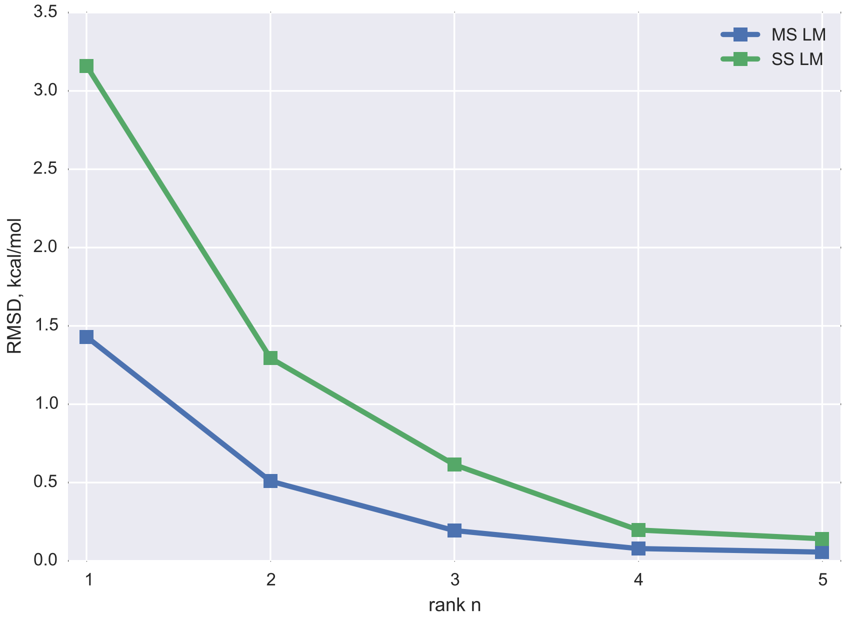


Figure A-E12: Convergence of the RMSD between methanol QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F12: Comparison of methanol QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	3.326	0.136	0.982	0.925	0.982
2	SS LM	5.108	0.182	0.966	1.295	1.003
3	SS LM	3.125	0.082	0.992	0.614	1.008
2	MS LM	1.950	0.071	0.995	0.509	1.004
3	MS LM	0.866	0.027	0.999	0.193	1.000

Table A-G12: Comparison of methanol QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	O_1	AC PC	2.232	0.062	0.943	0.853	1.019
2	O_1	SS LM	3.234	0.084	0.881	1.160	1.049
3	O_1	SS LM	2.827	0.041	0.961	0.676	1.003
2	O_1	MS LM	0.825	0.025	0.995	0.342	1.047
3	O_1	MS LM	0.866	0.015	0.994	0.231	0.963

A.13 Tetrazole

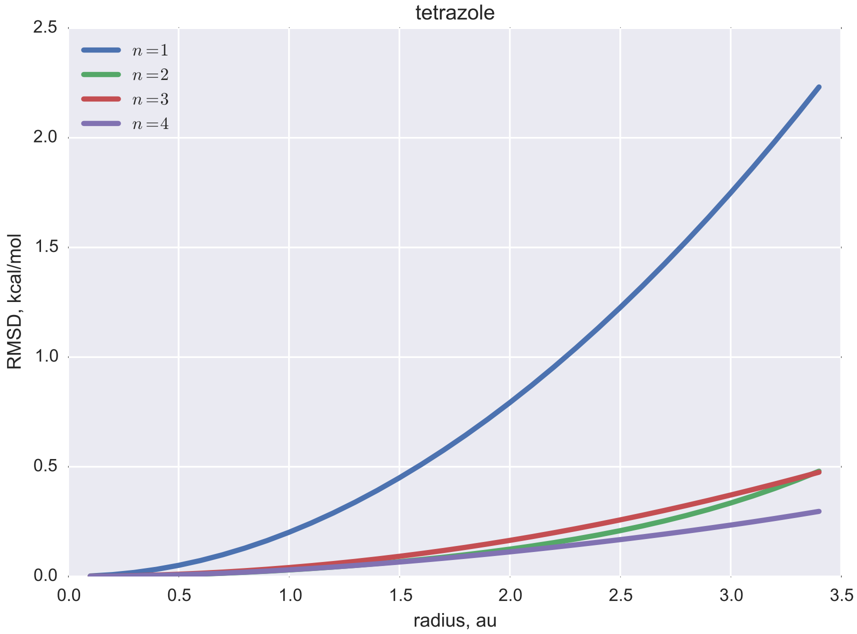


Figure A-A13: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B13: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	0.5	1.3	1.1	1.3

Table A-D13: Molecular multipole moments Q_{lm} of tetrazole calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.000	2.179	-0.334	-0.989	0.000	0.000	2.995	-0.745
AC PC	0.000	0.000	2.133	-0.329	-1.564	0.000	0.000	3.067	-0.678
SS LM	0.000	0.000	2.179	-0.334	-0.989	0.000	0.000	2.995	-0.745
MS LM	0.000	0.000	2.179	-0.334	-0.989	0.000	0.000	2.995	-0.745

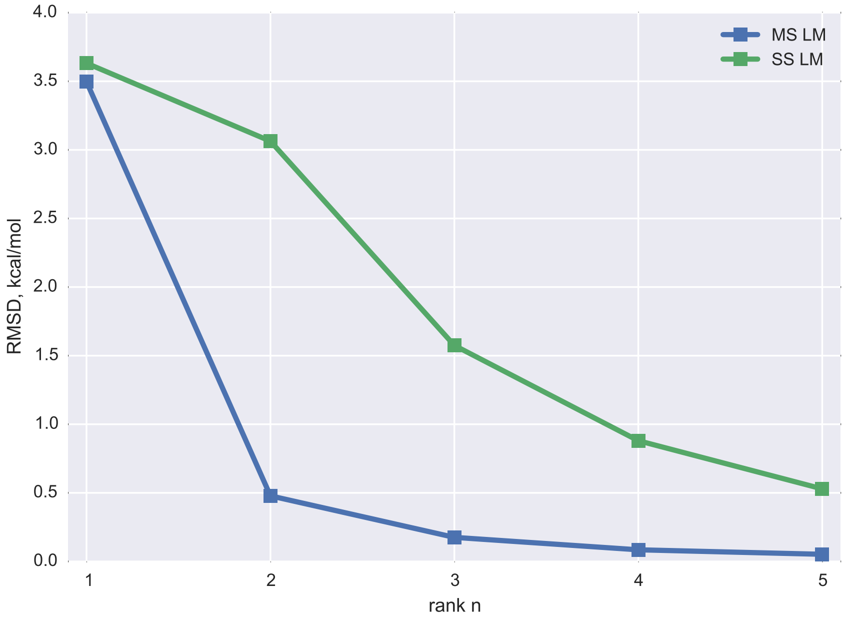


Figure A-E13: Convergence of the RMSD between tetrazole QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F13: Comparison of tetrazole QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	7.334	0.097	0.988	1.556	0.988
2	SS LM	12.000	0.188	0.958	3.063	1.015
3	SS LM	6.235	0.099	0.988	1.575	1.008
2	MS LM	2.272	0.028	0.999	0.478	1.005
3	MS LM	1.014	0.010	1.000	0.175	1.002

Table A-G13: Comparison of tetrazole QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	N_2	AC PC	3.691	0.201	0.990	1.243	1.421
–	N_3	AC PC	6.948	0.150	0.921	1.508	0.927
–	N_4	AC PC	7.334	0.103	0.818	2.047	0.613
–	N_5	AC PC	4.724	0.111	0.921	1.596	0.901
2	N_2	SS LM	7.772	0.651	0.986	3.862	1.919
2	N_3	SS LM	10.138	0.227	0.854	2.250	0.998
2	N_4	SS LM	11.333	0.142	0.607	3.141	0.496
2	N_5	SS LM	8.175	0.169	0.809	2.473	0.806
3	N_2	SS LM	6.235	0.560	0.832	3.083	1.091
3	N_3	SS LM	5.167	0.156	0.961	1.532	1.138
3	N_4	SS LM	4.081	0.082	0.912	1.552	1.062
3	N_5	SS LM	5.683	0.091	0.963	1.386	1.109
2	N_2	MS LM	0.637	0.057	0.996	0.327	1.025
2	N_3	MS LM	0.869	0.029	0.998	0.276	1.006
2	N_4	MS LM	0.844	0.013	0.997	0.261	0.968
2	N_5	MS LM	0.998	0.020	0.998	0.295	1.006
3	N_2	MS LM	0.512	0.042	0.999	0.245	0.947
3	N_3	MS LM	0.496	0.011	1.000	0.113	1.003
3	N_4	MS LM	0.646	0.004	1.000	0.099	1.015
3	N_5	MS LM	0.357	0.008	1.000	0.117	1.004

Table A-H13: Atom-centered point charge values of tetrazole fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule.

C_1	N_2	N_3	N_4	N_5	H_6	H_7
0.26	0.09	-0.26	-0.04	-0.37	0.23	0.09

Table A-I13: Atomic coordinates of tetrazole optimized at mp2/aug-cc-pVTZ level of theory.

#	Element	x, Å	y, Å	z, Å
1	C	1.584	1.330	-0.000
2	N	1.577	-1.209	0.000
3	N	-0.811	-2.046	-0.000
4	N	-2.231	-0.004	0.000
5	N	-0.789	2.113	-0.000
6	H	3.022	-2.450	-0.000
7	H	3.252	2.489	0.000

A.14 Thiazole

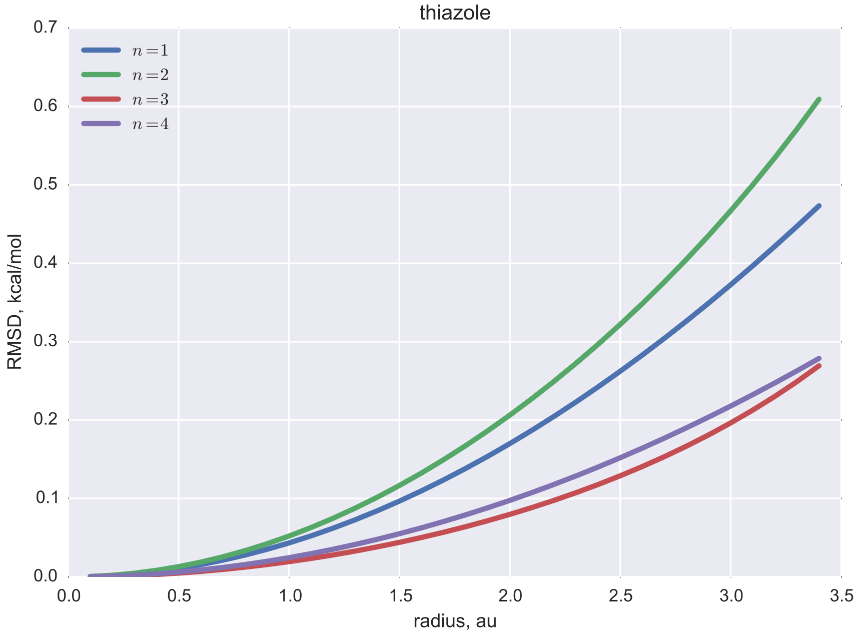


Figure A-A14: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B14: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	1.0	0.9	1.6	1.4

Table A-D14: Molecular multipole moments Q_{lm} of thiazole calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.000	0.372	-0.514	-3.310	0.000	0.000	4.639	4.602
AC PC	0.001	0.000	0.364	-0.511	-3.224	0.000	0.000	4.682	4.816
SS LM	0.000	0.000	0.372	-0.514	-3.310	0.000	0.000	4.639	4.602
MS LM	0.000	0.000	0.372	-0.514	-3.310	0.000	0.000	4.639	4.602

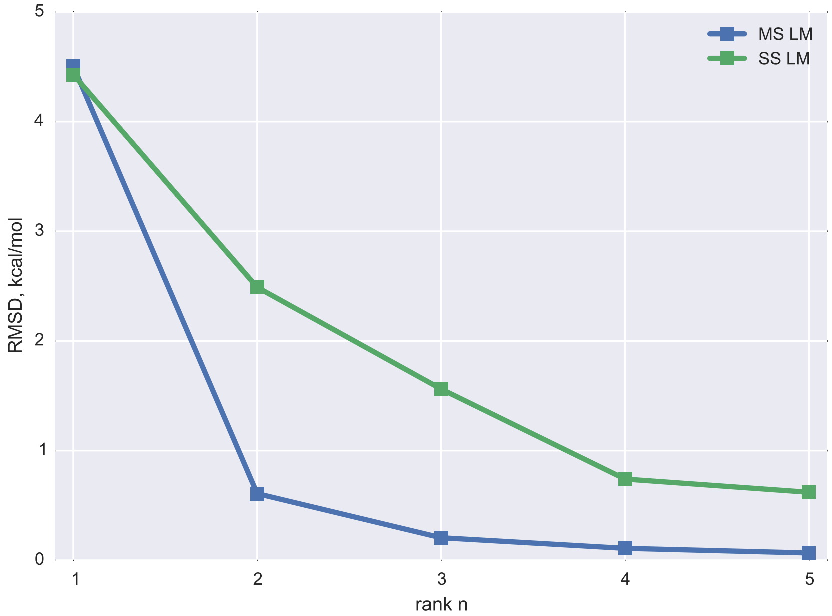


Figure A-E14: Convergence of the RMSD between thiazole QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F14: Comparison of thiazole QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	4.463	0.182	0.967	1.036	0.967
2	SS LM	9.586	0.440	0.820	2.489	0.909
3	SS LM	8.955	0.256	0.933	1.562	1.018
2	MS LM	2.725	0.097	0.989	0.608	1.002
3	MS LM	1.213	0.034	0.999	0.205	1.006

Table A-G14: Comparison of thiazole QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	N_5	AC PC	3.762	0.113	0.845	1.473	0.652
–	S_8	AC PC	3.436	0.482	0.812	0.936	0.805
2	N_5	SS LM	9.224	0.280	0.043	3.660	0.125
2	S_8	SS LM	8.599	0.743	0.690	1.492	1.145
3	N_5	SS LM	6.088	0.111	0.794	1.575	0.649
3	S_8	SS LM	8.549	0.479	0.871	1.158	1.287
2	N_5	MS LM	2.278	0.058	0.959	0.788	0.892
2	S_8	MS LM	0.867	0.088	0.991	0.193	0.985
3	N_5	MS LM	0.475	0.014	0.999	0.178	1.014
3	S_8	MS LM	0.437	0.040	0.998	0.089	0.981

Table A-H14: Atom-centered point charge values of thiazole fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule.

C_1	H_2	C_3	H_4	N_5	C_6	H_7	S_8
-0.34	0.23	0.14	0.12	-0.45	0.08	0.17	0.04

Table A-I14: Atomic coordinates of thiazole optimized at mp2/aug-cc-pVTZ level of theory.

#	Element	x, Å	y, Å	z, Å
1	C	2.299	0.041	0.000
2	H	4.277	-0.441	0.000
3	C	1.199	2.396	0.000
4	H	2.228	4.156	0.000
5	N	-1.380	2.430	0.000
6	C	-2.264	0.106	0.000
7	H	-4.249	-0.358	0.000
8	S	0.000	-2.226	0.000

A.15 trans-MeSNO

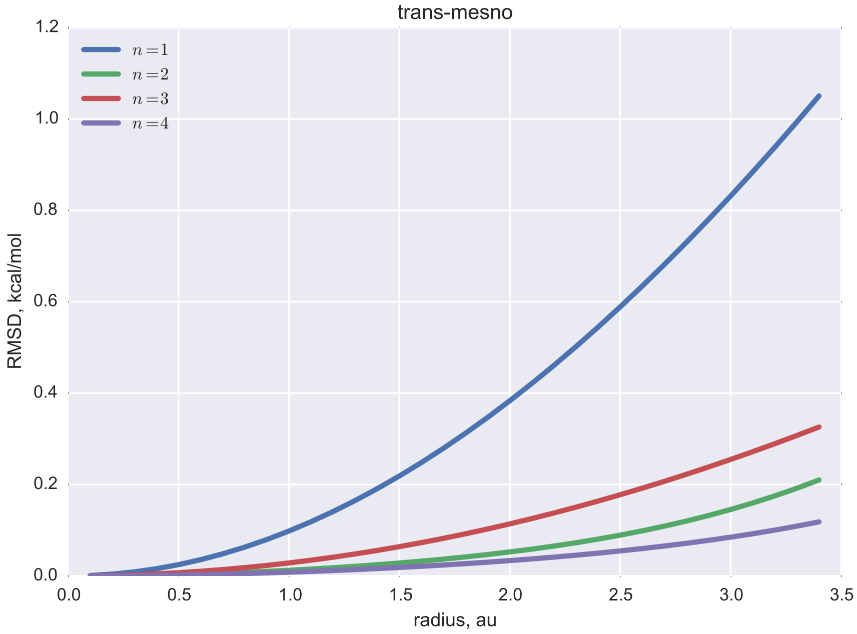


Figure A-A15: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B15: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	0.7	1.9	1.3	2.4

Table A-D15: Molecular multipole moments Q_{lm} of trans-mesno calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.000	1.132	0.436	0.370	0.000	0.000	1.542	-0.126
AC PC	0.002	0.000	1.133	0.426	-0.482	0.000	0.000	1.312	-0.353
SS LM	0.000	0.000	1.132	0.436	0.370	0.000	0.000	1.542	-0.126
MS LM	0.000	0.000	1.132	0.436	0.370	0.000	0.000	1.542	-0.126

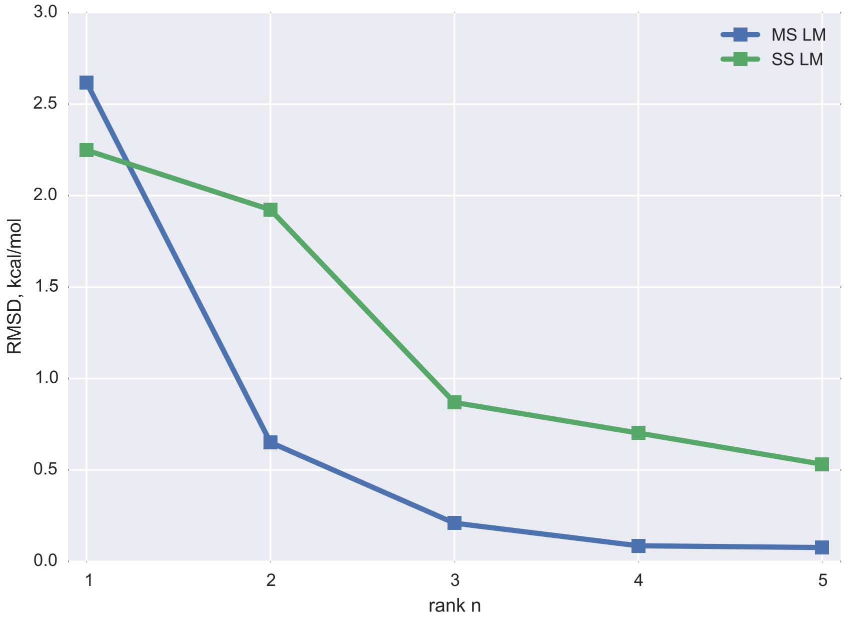


Figure A-E15: Convergence of the RMSD between trans-mesno QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F15: Comparison of trans-mesno QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	5.058	0.159	0.966	1.254	0.966
2	SS LM	8.471	0.244	0.928	1.923	1.012
3	SS LM	7.229	0.093	0.984	0.869	1.004
2	MS LM	2.528	0.084	0.991	0.651	0.989
3	MS LM	1.411	0.025	0.999	0.210	1.000

Table A-G15: Comparison of trans-mesno QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	O_1	AC PC	5.058	0.111	0.722	1.201	0.626
–	N_2	AC PC	4.492	0.539	0.183	2.190	0.328
–	S_3	AC PC	3.421	0.379	0.882	1.228	0.864
2	O_1	SS LM	8.471	0.214	0.161	2.145	0.235
2	N_2	SS LM	7.451	0.622	0.409	2.766	0.653
2	S_3	SS LM	5.914	0.502	0.782	1.803	1.306
3	O_1	SS LM	2.805	0.063	0.926	0.649	0.927
3	N_2	SS LM	4.495	0.354	0.962	1.589	1.506
3	S_3	SS LM	1.379	0.047	0.992	0.205	1.014
2	O_1	MS LM	1.111	0.038	0.975	0.374	0.945
2	N_2	MS LM	1.540	0.118	0.956	0.521	0.876
2	S_3	MS LM	1.566	0.171	0.951	0.590	0.955
3	O_1	MS LM	0.826	0.017	0.994	0.185	0.986
3	N_2	MS LM	0.481	0.038	0.994	0.185	0.964
3	S_3	MS LM	0.514	0.032	0.997	0.124	0.996

Table A-H15: Atom-centered point charge values of trans-mesno fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule.

O_1	N_2	S_3	C_4	H_5	H_6	H_7
-0.17	0.05	0.06	-0.60	0.25	0.20	0.20

Table A-I15: Atomic coordinates of trans-mesno optimized at mp2/aug-cc-pVTZ level of theory.

#	Element	x, Å	y, Å	z, Å
1	O	-3.186	-2.143	0.000
2	N	-0.937	-1.787	0.000
3	S	0.000	1.439	0.000
4	C	3.340	0.768	0.000
5	H	4.305	2.582	0.000
6	H	3.850	-0.279	1.689
7	H	3.850	-0.279	-1.689

A.16 Uracil

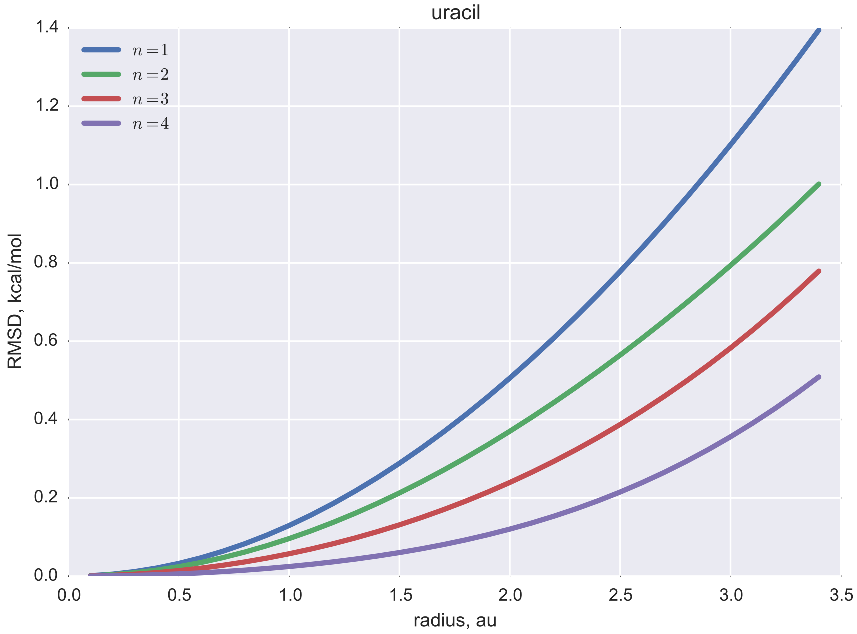


Figure A-A16: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B16: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	0.6	0.7	0.9	1.3

Table A-D16: Molecular multipole moments Q_{lm} of uracil calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.000	-0.469	1.926	1.847	0.000	0.000	-14.328	-2.227
AC PC	0.000	0.000	-0.468	1.933	2.082	0.000	0.000	-14.372	-2.201
SS LM	0.000	0.000	-0.469	1.926	1.847	0.000	0.000	-14.328	-2.227
MS LM	0.000	0.000	-0.469	1.926	1.847	0.000	0.000	-14.328	-2.227

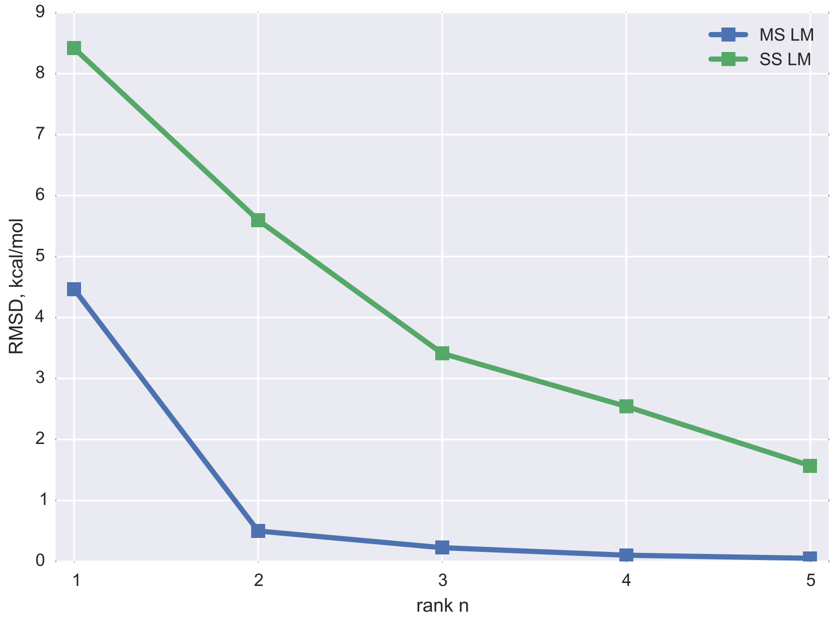


Figure A-E16: Convergence of the RMSD between uracil QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F16: Comparison of uracil QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	1.569	0.036	0.999	0.443	0.999
2	SS LM	18.215	0.486	0.793	5.599	0.937
3	SS LM	27.503	0.231	0.926	3.412	1.034
2	MS LM	2.335	0.039	0.998	0.499	1.001
3	MS LM	1.051	0.017	1.000	0.226	1.003

Table A-G16: Comparison of uracil QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	N_1	AC PC	1.531	0.073	0.956	0.672	1.017
–	N_5	AC PC	1.317	0.146	0.984	0.446	0.792
–	O_9	AC PC	1.356	0.024	0.995	0.412	1.048
–	O_11	AC PC	1.312	0.031	0.996	0.401	1.053
2	N_1	SS LM	14.847	0.740	0.833	6.576	2.683
2	N_5	SS LM	7.406	1.849	0.727	4.504	1.604
2	O_9	SS LM	9.592	0.293	0.154	4.525	0.178
2	O_11	SS LM	9.302	0.348	0.493	4.030	0.285
3	N_1	SS LM	8.418	0.414	0.564	3.622	1.318
3	N_5	SS LM	18.815	3.693	0.064	9.673	0.781
3	O_9	SS LM	8.349	0.135	0.784	2.418	0.721
3	O_11	SS LM	9.015	0.186	0.768	2.495	0.712
2	N_1	MS LM	1.198	0.094	0.983	0.733	0.950
2	N_5	MS LM	0.819	0.200	0.996	0.482	0.914
2	O_9	MS LM	0.745	0.016	0.997	0.259	0.976
2	O_11	MS LM	0.661	0.016	0.998	0.187	0.990
3	N_1	MS LM	0.882	0.028	0.980	0.279	0.928
3	N_5	MS LM	0.748	0.057	0.997	0.200	1.089
3	O_9	MS LM	0.372	0.007	1.000	0.119	0.993
3	O_11	MS LM	0.314	0.008	1.000	0.093	0.996

Table A-H16: Atom-centered point charge values of uracil fitted to the reference QM MEP. Subscript under the atom name corresponds to the order of the atom in the molecule.

N_1	C_2	C_3	C_4	N_5	C_6	H_7	H_8	O_9	H_{10}	O_{11}	H_{12}
-0.45	0.12	-0.60	0.88	-0.58	0.77	0.18	0.24	-0.63	0.35	-0.62	0.34

A.17 Water

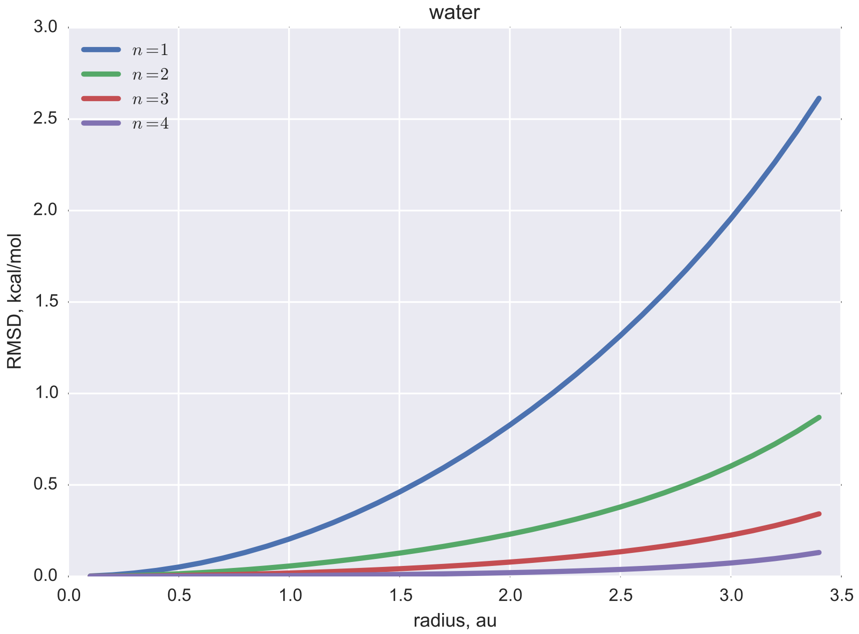


Figure A-A17: Effect of radius a on the RMSD between the MEP expansion Φ_n and electrostatic potential calculated using single-site Lebedev charge model.

Table A-B17: Values of the radius a required to reproduce MEP expansion Φ_n up to given degree n with less than 0.05 kcal/mol difference in RMSD.

n	1	2	3	4
a, au	0.5	0.9	1.6	2.7

Table A-D17: Molecular multipole moments Q_{lm} of water calculated using atom-centered point charges (AC PC), single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au, $n = 2$) as compared with the QM multipoles.

Model	Q_{00}	Q_{10}	Q_{11c}	Q_{11s}	Q_{20}	Q_{21c}	Q_{21s}	Q_{22c}	Q_{22s}
QM	0.000	0.000	0.000	-0.784	-1.727	0.000	0.000	1.276	0.000
AC PC	0.000	0.000	0.000	-0.802	-1.005	0.000	0.000	0.811	0.000
SS LM	0.000	0.000	0.000	-0.784	-1.727	0.000	0.000	1.276	0.000
MS LM	0.000	0.000	0.000	-0.784	-1.727	0.000	0.000	1.276	0.000

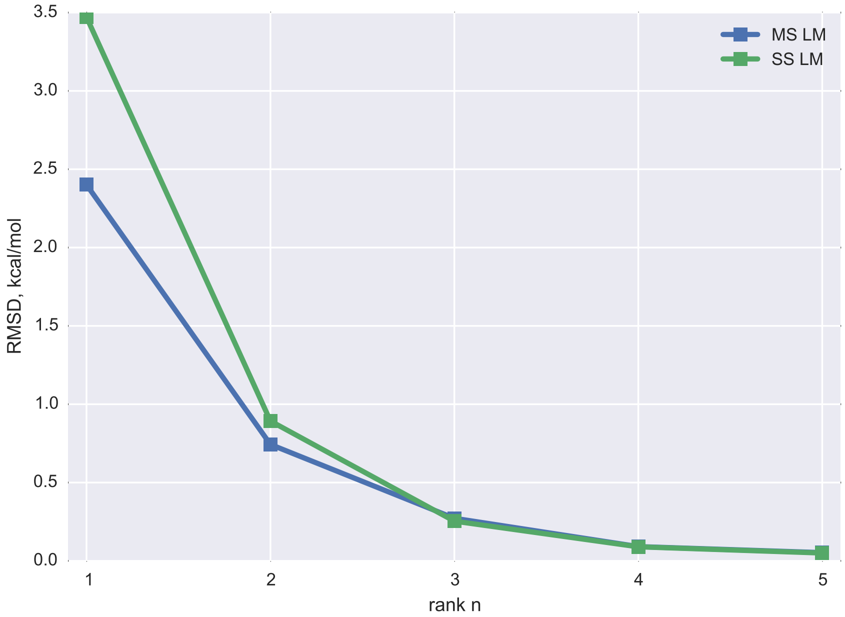


Figure A-E17: Convergence of the RMSD between water QM MEP and electrostatic potential calculated using charge models: single-site (SS LM) and multi-sites (MS LM) Lebedev models ($a = 0.5$ au).

Table A-F17: Comparison of water QM MEP (over molecular vdW grid) with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM) with $a = 0.5$. All dimensional quantities are in kcal/mol.

n	Model	Max. Error	RMAE	R^2	RMSD	α
–	AC PC	3.594	0.141	0.978	1.391	0.977
2	SS LM	3.080	0.087	0.991	0.893	1.007
3	SS LM	0.943	0.024	0.999	0.254	1.003
2	MS LM	2.077	0.073	0.994	0.743	1.006
3	MS LM	0.895	0.026	0.999	0.272	1.003

Table A-G17: Comparison of water QM AEPs with the potential produced by charge models: atom-centered point charges (AC PC), single-site (SS LM) and multi-sites Lebedev models (MS LM). All dimensional quantities are in kcal/mol. Subscript under the atom name corresponds to the order of the atom in the molecule.

n	Atom	Model	Max. Error	RMAE	R^2	RMSD	α
–	O_1	AC PC	2.994	0.075	0.921	1.069	1.129
2	O_1	SS LM	1.581	0.049	0.977	0.644	1.074
3	O_1	SS LM	0.664	0.015	0.996	0.214	0.964
2	O_1	MS LM	1.569	0.045	0.977	0.595	1.048
3	O_1	MS LM	0.862	0.018	0.994	0.260	0.956