MASARYK UNIVERSITY
Faculty of Science

Dissertation

Brno 2021 Stanislav Geidl

MASARYK UNIVERSITY
Faculty of Science
National Centre for Biomolecular Research

Partial Atomic Charges and Their Chemoinformatics Application

Dissertation

Stanislav Geidl

Supervisor: prof. RNDr. Jaroslav Koča, DrSc.

Brno 2021

To my grandma.

Bibliografický záznam

Autor:	RNDr. Stanislav Geidl
	Přírodovědecká fafulta, Masarykova univerzita
	Národní centrum pro výzkum biomolekul

Název práce:	Parciální atomové náboje a jejich aplikace v chemoinformatice
Studijní program:	Biomolekulární chemie a bioinformatika
Vedoucí práce:	prof. RNDr. Jaroslav Koča, DrSc.
Konzultant práce:	doc. RNDr. Radka Svobodová Vařeková, Ph.D.
Akademický rok:	2020/2021
Počet stran:	
Klíčová slova:	

Bibliographic Entry

Author:	RNDr. Stanislav Geidl
	Faculty of Science, Masaryk University
	National Centre for Biomolecular Research

Title of Thesis:	Partial Atomic Charges and Their Chemoinformatics Application
Degree Programme:	Biomolecular chemistry and bioinformatics
Supervisor:	prof. RNDr. Jaroslav Koča, DrSc.
Supervisor specialist:	doc. RNDr. Radka Svobodová Vařeková, Ph.D.
Academic Year:	2020/2021
Number of Pages:	
Keywords:	

Abstrakt

Chemoinformatické přístupy pro výpočet fyzikálních a chemických vlastností organických molekul, speciálně pak molekul dosud nesyntetizovaných, jsou velmi užitečné v rámci procesu vývoje léčiv a v dalších oblastech moderních přírodních věd. Jednou z velmi důležitých vlastností organických molekul je disociační konstanta (pK_a). pK_a lze úspěšně predikovat pomocí chemoinformatických modelů založených na parciálních atomových nábojích. Tyto modely vyžadují molekulární 3D strukturu, kterou však lze připravit mnoha různými způsoby.

Prvním tématem, kterým jsem se v rámci své práce zabýval, je právě vliv zdroje 3D struktury molekul na přesnost predikce pK_a. Zjistil jsem, že výběr zdroje 3D struktury je klíčový pro úspěšnou predikci pK_a, přičemž 3D struktury z databází DTP NCI a PubChem je jevily jako nejvhodnější. Z této analýzy byla rovněž patrná nutnost kvalitních a rychle vypočítatelných parciálních atomových nábojů, sloužících jako vstupy pro predikci pK_a. Uvedená oblast se stala druhým tématem mé práce. Konkrétně jsem se zaměřil na parametrizaci metody Electronegativity Equalization Method (EEM), sloužící pro rychlý výpočet parciálních atomových nábojů. Výsledkem mé práce byly EEM parametry, poskytující kvalitní náboje pro léčiva a jim podobné organické molekuly. Výsledkem mé práce byly EEM parametry, poskytující vysoce kvalitní náboje pro léčiva a jim podobné organické molekuly. Při přípravě těchto parametrů jsem si rovněž uvědomil nutnost mít k dispozici softwarový nástroj pro výpočet nábojů a parametrizaci nábojových metod. Tato problematika se stala třetím tématem mé práce – spolupracoval jsem na vývoji software NEEMP, který slouží k parametrizaci EEM, validaci EEM parametrů a výpočtu nábojů pomocí metody EEM.

Celkově má práce poskytuje metodiky a nástroje pro stavbu kompletního workflow, sloužícího pro výpočet i u dosud nesyntetizovaných molekul. Toto workflow zahrnuje získání 3D struktury molekul, výpočet jejich parciálních atomových nábojů metodou EEM a aplikace nábojů pro predikci pK_a.

Abstract

Chemoinformatic approaches for predicting physicochemical properties, especially in the case of unsynthesized molecules, are beneficial in the drug design process and other modern life science fields. One of the most important properties of organic molecules is the dissociation constant (pK_a). pK_a is successfully predictable by chemoinformatics models based on partial atomic charges – these models require a molecules' 3D structure that can be obtained using different approaches. The first topic of my work is to analyze the influence of 3D structure sources on the quality of pKa prediction. I found out that the correct source of 3D structure is crucial for pK_a prediction while structures from databases DTP NCI and PubChem appear most suitable. This work shows a need for quality and quickly calculated charges used as input for pK_a prediction. This field became my second topic. Specifically, I focused on Electronegativity Equalization Method (EEM) for quick partial atomic charges calculation. The result of my work was the EEM parameters, which provide high-quality charges for drug-like molecules. During EEM parameters preparation, I realized a need to have a tool for partial charge calculation and parametrization. This problem became my third topic – I cooperated on the development of NEEMP software that can parametrize EEM, validate EEM parameters and calculate charges via EEM method.

Overall, my work provides a methodology and tools for building the whole workflow used for pK_a prediction, which can be used for unsynthesized molecules. This workflow contains obtaining 3D structures of molecules, partial atomic charges calculation, and pKa prediction.

Acknowledgements

...

Declaration

I hereby declare that this disertation thesis is my original authorial work, which I have worked on alone. All sources, references and literature used or excerpted during the elaboration of this work are properly cited and listed in a complete reference with regard to the source.

Brno, xxth June 2021 .....
Stanislav Geidl

Publication list with definition of autor's contribution

This dissertation is based on 3 articles of dissertation author (Stanislav Geidl, SG):

Geidl S, Svobodová Vařeková R, Bendová V, Petrusek L, Ionescu C-M, Jurka Z, Abagyan R, Koča J: **How Does the Methodology of 3D Structure Preparation Influence the Quality of pK_a Prediction?** *J Chem Inf Model* 2015, **55**:1088–1097.

SG prepared the input dataset (by extension of published datasets), performed all the calculations, participated in the analysis of the results, and wrote a part of the manuscript, including all tables and graphics.

Geidl S, Bouchal T, Raček T, Svobodová Vařeková R, Hejret V, Křenek A, Abagyan R, Koča J: **High-quality and universal empirical atomic charges for chemoinformatics applications.** *J Cheminform* 2015, **7**:59.

SG participated in the study's design, and cooperated in the preparation of the input data (molecules and published EEM parameters) and in QM charge calculation. SG performed the analyses and the interpretation of the data.

Raček T, Pazúriková J, Svobodová Vařekova R, Geidl S, Křenek A, Falginella FL, Horský V, Hejret V, Koča J: **NEEMP: Software for validation, accurate calculation and fast parameterization of EEM charges.** *J Cheminform* 2016, **8**:1.

SG prototyped the DE-Min approach and designed some new NEEMP features, such as the usage of RMSE instead of R^2. SG also implemented a preparation of validion reports.

Contents

Part I

Introduction

1

Introduction

In recent years, a vast amount of data about various types of molecules became available. For example, we can obtain the complete human genome of a selected individual in a few days, and about 150 thousand biomacromolecular structures have been determined and published (Protein Data Bank [1]). Furthermore, more than 100 million various small molecules are described in freely accessible databases (e.g., Pubchem [2], ZINC [3], ChEMBL [4]). This richness of data caused the formation of novel modern life-science research fields focused on the utilization of this data. The best-known modern life sciences are bioinformatics, structural bioinformatics, systems biology, genomics, proteomics, and also chemoinformatics. These current research specializations have provided many key results in basic and applied research (e.g. [5–11]).

One fascinating and beneficial field utilizing and processing newly available data about small molecules (i.e., drug-like compounds) is chemoinformatics. This discipline offers methodologies for comparing molecular similarity, molecular database search, virtual screening, and the prediction of molecules' properties and activities. This prediction is based on the idea that molecular structures' similarity has a consequence – a similarity in molecular properties. In chemoinformatics, the structure is first described using mathematical characteristics (so-called descriptors) – numbers containing 3D (or 2D or 1D) structure information and applicable as inputs of mathematical models. Then, these models are constructed based on a relation between descriptors and known values of the property or the activity. Such models are called Quantitative Structure-Property Relationship (QSPR) models or Quantitative Structure-Activity Relationship (QSAR) models.

A property, which is strongly required and is therefore often a target of chemoinformatics prediction models is the acid dissociation constant, K_a, and its negative logarithm pK_a. Those pK_a values are of interest in chemical, biological,

environmental, and pharmaceutical research [12–14]. pK_a values have found applications in many areas, such as evaluating and optimizing drug candidate molecules, pharmacokinetics, ADME profiling, understanding protein-ligand interactions, etc. Moreover, the critical physicochemical properties such as permeability, lipophilicity, solubility, etc., are pK_a dependent. Unfortunately, experimental pK_a values are available only for a limited set of molecules. In addition to that, obtaining experimental pK_a values for newly designed molecules is very time-consuming because they must be synthesized first. Chemoinformatics approaches for pK_a prediction are therefore currently intensively examined.

For this reason, I also focused on the chemoinformatics way of pK_a prediction in my work. Very promising descriptors for pK_a prediction are partial atomic charges [15–20] because they hold information about the distribution of electron density within the molecule. Specifically, electron densities on atoms close to the dissociating hydrogen provide a clue about its dissociation ability. The most common and accurate method for calculating partial atomic charges is an application of quantum mechanics (QM). QM calculation can be performed via various approaches, introducing different approximation levels (i.e., approximating a wave function by different sets of mathematical equations, which are called basis sets). QM outputs electron distribution in orbitals and this distribution can be divided into individual atoms using several charge calculation schemes (e.g., MPA, NPA, AIM, Hirshfeld, MK, etc.). Therefore, the correlation between pK_a and relevant atomic charges calculated by different QM approaches has been analyzed [19]. I also focused on this file in my bachelor thesis [21], developed a workflow for calculation of pK_a using QM partial atomic charges and examined, which types of QM are the most suitable.

QM charges are accurate, but their calculation is very time-consuming. A faster Alternative to QM charges is empirical charge calculation approaches. Furthermore, if we would like to apply chemoinformatics pK_a prediction models practically – for example, in pre-screening large sets of drug candidates – we need a fast approach. Therefore, in my master thesis [22], I developed a pK_a prediction workflow based on charges (including Electronegativity Equalization Method).

However, several pieces of the puzzle were still missing. For example, the developed pK_a prediction workflows [20] were strongly dependent on 3D structure source, and also, the quality of available EEM charges was low.

Therefore, my dissertation's goal was to develop a workflow that predicts pK_a for molecules not synthesized yet and without available experimental 3D structures.

Specifically, the thesis examined how to improve the process of pK_a prediction via providing suitable inputs. First, the influence of 3D structure source on pK_a prediction accuracy was analyzed. Afterward, the work focused on obtaining high-quality partial atomic charges, which served as descriptors for pK_a

calculation. In the end, the authors also support the development of methodology and software tools for obtaining these high-quality charges.

The thesis structure is the following: First, an overview of key fields is provided (Part II), i.e. – 3D structure and approaches for its prediction, charge calculation methods, and pK_a prediction approaches. Next, the achieved results, which we published in three research papers, are briefly described (Part III), and full-texts of the respective published papers are attached in Main papers. During the elaboration of this thesis, I was also involved in other projects. Most of them were not related to pK_a prediction but tightly connected to the field of chemoinformatics or structural bioinformatics. The outcome of these projects consists of several papers and a book I have co-authored. Their title pages are attached in Appendix: All My Publications.

Part II

Theory and Methods

2

Structure

2.1 Molecular Structure in Computer

The chemical structure is the essential information for chemoinformatics and computational chemistry calculations. We recognize different types of chemical structures according to the complexity of information [23].

The empirical or chemical formula provides information about molecule composition – elements and their count. The structural formula (2D structure) extends this information about topology – bonds between atoms. The three-dimensional structure also provides the conformation of a molecule – the relative placement of atoms in space. We try to provide conformation with the lowest energy representing the most probable conformation of molecule in reality. For some applications, there can even be an assembly of these 3D structures.

In chemoinformatics, two-dimensional structures are often used, but the three-dimensional structure can often bring new information into the *in silico* calculations or models. On the other hand, this 3D structure can be obtained experimentally for a limited number of small molecules. What with other molecules, including those which were not synthesized yet?

2.2 3D Structure Calculation

We apply more computationally efficient methods for 3D structure computation because we use them as input for high-throughput methods. For this reason, many resources were devoted to the development of fast and accurate 3D structure prediction methods [24]. These can be classified into the following groups [24]: rule-based and data-based, fragment-based, numerical methods,

and conformational analysis. These rule-based and data-based, fragment-based methods are partially overlying.

2.2.1 Rule-Based and Data-Based Methods

These methods use chemical knowledge of geometric and energetic rules known from experiments and theoretical calculations. In these methods, we use rules explicitly to describe, e.g., bond lengths and angles; we use data implicitly to describe, e.g., ring conformation.

2.2.2 Fragment-Based Method

The fragment-based method is the incremental method using rules in the first step to fragment a structure into parts. According to the following rules, the parts are assembled by linking fragment templates from a library (database). Predicted structures are created from the most similar and largest fragments in a database as possible.

2.2.3 Numerical Method

These numerical methods consist of three methods: molecular mechanics (MM), quantum mechanics, and distance geometry (DG). Distance geometry is a great tool to prepare a reasonable initial structure, which is very close to some low-energy conformation. For this structure, we can use the optimization process from MM or/and QM.

2.2.4 Conformational Analysis

This method generates a set of conformations for one molecule using different approaches - genetic algorithms, systematic methods, random techniques, Monte Carlo or MD simulation. The one or more different structures are selected based on criteria such as the number of conformers, minimum RMSD, only conformations with the lowest MM energy (low-energy conformers).

3

Partial Atomic Charges

3.1 The Concept of Atomic Charges

Atomic charges are a theoretical concept for the quantitative description of electron density around every atom in a molecule. The first basic concept came from early chemistry, where an integer expressed these charges (e.g. -1, +2). Later, they were a real number (partial charge) in organic chemistry and physical chemistry [25]. It is a great approach to explain the mechanism of a lot of chemical reactions. Recently, partial atomic charges also became popular in chemoinformatics, as they proved to be informative descriptors for QSAR and QSPR modeling [16, 26] and for other applications [17, 27, 28]; they can be utilized in virtual screening [29, 30] and similarity searches [31, 32]. In reality, we are not able to measure these numbers, only calculate or estimate them. For such reasons, many different approaches for the calculation of partial atomic charges were developed.

3.2 Overview of Charge Calculation Methods

3.2.1 QM Charge Methods

These methods use a wave function as a starting point and then apply subsequent population analysis, charge calculation scheme, or fit to some physical observation [33].

Mulliken population analysis (MPA) [34, 35] simply calculates a charge of an atom as a sum of an electron density from its molecular orbitals and a half

of an electron density from its bonding orbitals. Natural population analysis (NPA) [36, 37] sophisticatedly improves the MPA method by orthogonalization of specific atoms and after this, NPA performs charge assignment from electron density the same way as in MPA. NPA atomic charges are more stable and independent of the size of basis sets. Other possible population analyses are Löwdin population analysis [38], Hirshfeld population analysis [39].

AIM (atoms-in-molecules) charge calculation scheme is based on the idea that electron density measured by X-ray can help with the calculation of partial charges. Bader and his coworkers [40, 41] defined an atomic volume that is used for charge calculation. Other well-known approaches are electrostatic potential fitting methods (ESP) like CHELPG [42] or MK (Merz-Singh-Kollman) [43] and their extension – RESP methods [44].

Cramer and at [45] also developed semiempirical methods – charge model 5 (CM5), which extends Hirshfeld population analysis by empirical parameters to reproduce charge-dependent observables.

3.2.2 Empirical Methods

Empirical approaches use only empirical parameters, and some of these can calculate charges from the 3D structure or only from the topology (2D structure) of a molecule. Therefore, they are distinctly faster than QM approaches.

One of the first empirical methods developed, CHARGE [46], performs a breakdown of the charge transmission by polar atoms into single-bond, double-bond, and triple-bond additive contributions. Other empirical methods have been developed on the electronegativity equalization principle. One group of these empirical approaches are using the Laplacian matrix formalism and the product is a redistribution of electronegativity: Gasteiger-Marsili (PEOE, partial equalization of orbital electronegativity) [47], GDAC (geometry-dependent atomic charge) [48], KCM (Kirchhoff charge model) [49], DENR (dynamic electronegativity relaxation) [50] or TSEF (topologically symmetric energy function) [50].

The second group of approaches applies the full equalization of orbital electronegativity. For example, this group contains EEM (electronegativity equalization method) [51] and its extensions (ABEEM [52], SFKEEM [26]), QEq (charge equilibration) [53], EQEq (extended QEq) [54], or SQE (split charge equilibration) [55].

Group of conformationally independent methods (based on the 2D structure) contains CHARGE, Gasteiger-Marsili, KCM, DENR, and TSEF. Group of conformationally dependent – geometrical charges (based on the 3D structure) also consider an influence of conformation and includes the following methods: GDAC, EEM, ABEEM, SFKEEM, QEq, EQeq, and SQE.

A typical representative of the topological method is the Gasteiger-Marsili method, which first assigns charges based on atom types and then iteratively updates atomic charges based on the closest partners. The correction is smaller and smaller in every step until the sixth step when these corrections are too small and atomic charges are final. Empirical parameters for this method were calculated from QM.

On the other hand, the EEM method needs a complete 3D structure and more applicable charges for some of the applications.

3.3 EEM Calculation

EEM (electronegativity equalization method) [51] is one of the most popular empirical charge calculation methods and was developed more than twenty years ago. This method's new parameterizations [26, 51, 56–61] and extension [26, 52] are still under development. An advantage of EEM calculation is that it considers the influence of the molecule's conformation on the atomic charges. For this reason, EEM charges are often used in predictive models as chemoinformatics regressors (descriptors) [62].

EEM is based on three principles:

The first principle is Sanderson's electronegativity equalization principle. It assumes that the effective electronegativity of each atom in the molecule is equal to the molecular electronegativity:

$$\chi_1 = \chi_2 = \cdots = \chi_x = \bar{\chi} \tag{3.1}$$

where χ_x is the effective electronegativity of the atom x and $\bar{\chi}$ is the molecular electronegativity.

The second principle is the principle of the charge balance. The sum of all charges is equal to the total charge Q:

$$\sum_{x=1} q_x = Q \tag{3.2}$$

where q_x is the charge of the atom x.

And the last principle is the principle of charge-dependent electronegativity. This principle is the definition of atomic electronegativity, and states that the electronegativity of each atom can be expressed as a function of its charge:

$$\chi_i = A_i + B_i \cdot q_i + \kappa \sum_{j=1\, i\neq j}^{N} \frac{q_j}{R_{i,j}} \tag{3.3}$$

where $R_{i,j}$ is the distance between atoms i and j, and the coefficients A_i, B_i and κ are empirical parameters.

These principles can be summed up to a system of equations with N + 1 unknowns (where $q_1, q_2, \ldots, q_N$ and $\bar{\chi}$):

$$\begin{pmatrix} B_1 & \frac{\kappa}{R_{1,2}} & \cdots & \frac{\kappa}{R_{1,N}} & -1 \\ \frac{\kappa}{R_{2,1}} & B_2 & \cdots & \frac{\kappa}{R_{2,N}} & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\kappa}{R_{N,1}} & \frac{\kappa}{R_{N,2}} & \cdots & B_N & -1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \\ \bar{\chi} \end{pmatrix} = \begin{pmatrix} -A_1 \\ -A_2 \\ \vdots \\ -A_N \\ Q \end{pmatrix} \tag{3.4}$$

The first values of parameters A_i and B_i were modifications of experimental hardness and electronegativity [51]. κ was equal to 1. Nowadays, these parameters are calculated from the QM charges [26, 56–61]. Therefore, EEM charges were correlated with the QM charge calculated with the same method used for parametrization.

3.4 Quality and Usability of EEM parameters

The quality of EEM parameters describes how the empirical charges computed using these EEM parameters correspond with QM charges used for EEM parameterization. Three main characteristics (statistics [63, 64]) can describe the quality of EEM parameters – the coefficient of determination root mean square error (RMSE) and average absolute error ($\bar{\Delta}$).

The coefficient of determination R^2 is the squared value of the Pearson coefficient (equation 3.5). This value describes the linear correlation rate. Values close to 1 mean that values correlate very well, and values close to 0 mean no correlation.

$$R = \sqrt{\frac{\sum_{x=1}^{N} \left((q_x^{calc} - \overline{q}_x^{calc}) \cdot (q_x^{ref} - \overline{q}_x^{ref}) \right)}{\sum_{x=1}^{N} (q_x^{calc} - \overline{q}_x^{calc})^2 \cdot \sum_{x=1}^{N} (q_x^{ref} - \overline{q}_x^{ref})^2}} \tag{3.5}$$

where q^{ref} is the reference value of charge calculated by QM and q^{calc} is charge value calculated by EEM. $\overline{q}^{ref}$, $\overline{q}^{calc}$ are the average value of q^{ref}, respectively q^{calc}.

Root mean square error RMSE is the normalized sum of squared error describing the reliability of the model calculated by:

$$\mathrm{RMSE} = \frac{\sum_{x=1}^{N} (q_x^{calc} - q_x^{ref})^2}{N} \tag{3.6}$$

Average absolute error $\overline{\Delta}$ is an averaged difference between corresponding EEM and QM charges in a molecule and is calculated according to an equation:

$$\overline{\Delta} = \frac{\sum_{x=1}^{N} |q_x^{calc} - q_x^{ref}|}{N} \tag{3.7}$$

Their **coverage** describes the applicability of EEM parameters. Coverage is a percentage value describing EEM parameters' ability to calculate charges for individual molecules in a dedicated dataset. *De facto*, this coverage depends on the representation of atom types in EEM parameters.

$$\text{coverage} = \frac{N_{pos}}{N_{tot}} \tag{3.8}$$

where N_{pos} is the number of molecules able calculated by EEM parameters and N_{tot} is the total number of molecules in a dataset.

3.5 EEM Parametrization

For the parameterization of EEM charges, a lot of different methods have been introduced. We can summarize it into two groups: one group contains a method that analytically solves equation [**?**] – linear regression [59, 60] and the second group contains optimization methods [61] such as Accelerated Random Search, Particle Swarm Optimization, and Differential Evolution algorithms. Both of these groups need input – a set of molecules with 3D structures and QM atomic charges. In my work, linear regression and differential evolution were used, and therefore, they are described in more detail below:

The linear regression (LR) method is based on these two equations:

$$A_i + B_i \cdot x = y \tag{3.9}$$

$$\begin{aligned} x &= q_i \\ y &= \chi_i - \kappa \sum_{j=1\, i \neq j}^{N} \frac{q_j}{R_{i,j}} \end{aligned} \tag{3.10}$$

Equations are derived from equations 3.9 and 3.10, which define the EEM method. In the LR method, the dataset of molecules with QM charges can change in every iteration to improve the quality of resulting charges. Quality criterium can be the Pearson correlation coefficient or the coefficient of determination, and the root mean square error or different types of errors. An advantage of the

LR method is its straightforwardness and the possibility to optimize κ by another iteration. On the other hand, this method is not possible to make parametrization for some extensions of EEM like SFKEEM and ABEEM.

Differential Evaluation (DE) [65] is a heuristics method to focus on finding a global minimum of a function. This method works similar to other optimization methods – iteratively optimize parameters to improve the final solution. Parameters of function are set up randomly, mutated, and evaluated until there is no best solution.

4

Acid Dissociation Constant Prediction

4.1 Motivation

The acid dissociation constant (pK_a) is a physicochemical property that characterizes the strength of acids. It is one of the essential properties for pharmaceutical, chemical, biological and environmental research or industry. For example, it can be used in the chemoinformatics pipeline for evaluation and optimization of drug candidate [66–68], ADME profiling [69, 70], pharmacokinetics [12], understanding protein-ligand interactions [13, 71].

4.2 Overview of Approaches

Several different approaches for pKa prediction have been developed [71–74].

4.2.1 LFER (Linear Free Energy Relationships) Methods

This is one of the oldest methods [75, 76] for pK_a prediction. This method uses the linear relation of Gibbs energy and pK_a or the logarithm of a reaction rate constant – the Hammett and Taft equations. An advantage of this method is a simple, straightforward, and quick calculation, but on the other hand, we need substituent and reaction parameters. This method is still used in the programs ACD/pKa [77], EPIK [78], and SPARC [79].

4.2.2 Database Methods

These methods [80, 81] use a library (database) of molecules with known pKa values. The pK_a value is taken directly from this library, or it is interpolated or triangulated from most similar molecules in this library. Most accurate results are produced only for molecules that are similar to molecules in the database. For this reason, it is essential to have an extensive library.

4.2.3 Ab Inition Quantum Mechanical Calculations

These methods [82, 83] use the fact that the dissociation constant can be calculated from the Gibbs energy of the reaction and from the solvation based on equation 4.2. However, there is no general approach, and every specific calculation configuration needs to be calibrated based on experimental values. The significant disadvantage of these methods is that they are time-consuming. On the other hand, these methods can be very accurate if they use correct calibration parameters. It is only one of the few methods that can be used to extend the training dataset with experimental values or validate some of this experimental value. It means that other methods can be improved by this method. This method is implemented as a module of the Jaguar quantum chemical software package [84].

$$\mathrm{p}K_a = -\log_{10} K_a \tag{4.1}$$

$$K_a = e^{\frac{-\Delta G^\circ}{RT}} \tag{4.2}$$

4.2.4 QSPR Method

The quantitative structure-property relationship method [17,85,86] uses mainly a linear model to describe the relationship between molecular structure and a property of a molecule, in our case pK_a. In those models, structures are presented by descriptors [62] that are numerical expressions of molecular properties. For example, descriptors can be the number of hydrogen atoms, the ratio between carbon atoms and all atoms in the molecule, or solvent accessible surface area.

pK_a correlates well with the polarizability, HOMO energy [87], proton-transfer energy [87], partial atomic charges [15, 16, 18–20], the electrostatic potential of the molecule [88], etc. Partial atomic charges proved as very promising descriptors [15, 16, 18–20] for pK_a prediction.

Part III

Results

5

Synopsis of the Results

We published a series of articles about pKa prediction [19,20] where we showed that some specific atomic charges correlate with pK_a. Based on this, we were able to build QSPR models for the prediction of this property. We also compared QM, EEM charges, and their models. In this dissertation, I focused only on the last one:

Geidl S, Svobodová Vařeková R, Bendová V, Petrusek L, Ionescu C-M, Jurka Z, Abagyan R, Koča J: **How Does the Methodology of 3D Structure Preparation Influence the Quality of pK_a Prediction?** *J Chem Inf Model* 2015, **55**:1088–1097.

In this article, we utilized different approaches to generate the 3D structures of organic molecules. These structures were used for the building of pK_a prediction models based on charge descriptors. Then we analyzed various influences and relationships and found which methodologies for 3D structure preparation are applicable for pK_a prediction.

We examined not only pKa prediction models employing QM charges but also the models utilizing EEM charges. An application of EEM charges looked very promising. Moreover, EEM charge calculation is significantly faster than QM charge calculation.

However, in parallel, we found one significant limitation of EEM charges – the parameters. It was available a reasonable number of parameter sets, but they had only a low coverage. For example, Bultinc's parameter set [57] contains parameters only for these elements: C, F, H, N, O. This fact markedly reduced a dataset, which we were able to use for QM charges.

A lack of parameters disallows the usage of EEM charges in many chemoinformatics applications. For this reason, in our follow-up work, we focused on

the development of new and more robust EEM parameters. The first step was to develop a new parameterization tool that was easy to use and extendable. After the prototype, we carefully prepared a new dataset of molecules, and for this dataset, we computed EEM parameters with higher coverage. These new parameters were published in a scientific paper:

Geidl S, Bouchal T, Raček T, Svobodová Vařeková R, Hejret V, Křenek A, Abagyan R, Koča J: **High-quality and universal empirical atomic charges for chemoinformatics applications.** *J Cheminform* 2015, **7**:59.

After some tuning up and extensive research, we released and also published the tool for EEM parametrization – the NEEMP software:

Raček T, Pazúriková J, Svobodová Vařekova R, Geidl S, Křenek A, Falginella FL, Horský V, Hejret V, Koča J: **NEEMP: Software for validation, accurate calculation and fast parameterization of EEM charges.** *J Cheminform* 2016, **8**:1.

Sections 5.1, 5.2, and 5.3 contain a summary and Main papers full text of all these aforementioned articles. I was also involved in other projects, and the outcome of it is a list of additional articles and book chapters in Appendix: All My Publications.

5.1 How Does the Methodology of 3D Structure Preparation Influence the Quality of pKa Prediction?

From our previous articles [19, 20], we know that the prediction of pK_a is possible via QSPR models using partial atomic charges as descriptors. The article's goal was to discover how the methodology of 3D structure preparation influences the quality of pKa prediction. We prepared a dataset containing 60 phenols, 82 carboxylic acids, 48 anilines, and additional testing 53 phenols for these purposes. We took structures from 5 different sources for all these molecules: PubChem [2], DTP NCI database [89]; Balloon [90], Frog2 [91], OpenBabel [92], and RDKit [93] software. We used neutral forms of all the molecules and dissociated forms of phenols and carboxylic acids, and associated forms of anilines. We also optimized these structures with MM (Molecular mechanics) and QM. All combination led to 7220 structures for that we calculated four different QM, one semiempirical QM, four different EEM charges, and Gasteiger-

Marselli charges. We created 516 QSPR models for all these molecules and charges. The robustness of these models was tested by cross-validation and QM charges also by an independent test set of phenols.

We confirmed that QM and EEM charge descriptors could be used for pK_a prediction. About half of all models have excellent quality with $R^2 \geq 0.9$. We also showed that ab initio and semiempirical charges correlate with pK_a and their models are very accurate. In EEM, we had models with a little worse quality, but empirical charges are calculated much faster, and an application in chemoinformatics is much more appropriate. In our models, we were not able to use Gasteiger-Marsili charges to get an adequate quality.

We focused on different types of influence. For classes of the benzene derivates (phenols and anilines) was much easier to obtain high-quality models. Nevertheless, for aliphatic hydrocarbon derivates (carboxylic acid), it was more challenging.

The focus of this article was a comparison of the source of the 3D structure. The influence of input structure for models is essential because the result – the quality of QSPR models – depends on input structures and their quality. For example, structures taken from RDKIT generated only by distance geometry produced fragile models. On the other hand, the 3D structures from the DTP NCI and PubChem databases, formally structures generated by CORINA and Omega, exhibited the best performance for all the tested molecular classes and charge calculation approaches. Structures generated by Frog2 also performed very well. Other 3D structure sources can also be used, but with caution.

We also tried the influence of structure optimization on the quality of QSPR models. In most cases, differences between original structures and optimized structures were slight.

In this article, we summed up the best workflow for the fast and accurate prediction of pKa. This or similar workflow can also be easily applied to other important properties for *in silico* designed molecules. Flow is about preparing 3D structures by CORINA or Omega (with no further optimization), calculation of EEM charges for these structures, and then the EEM QSPR calculation of pK_a.

5.1.1 My contribution

I prepared the input dataset (by extension of published datasets), performed all the calculations, participated in the analysis of the results, and wrote a part of the manuscript, including all tables and graphics.

5.2 High-quality and universal empirical atomic charges for chemoinformatics applications

The EEM method for charge calculation was published several decades ago. Before our study, there were done some improvements of EEM [26, 52], parameterizations of empirical parameters [26, 51, 56–61], and developments of new parameterization methods [26, 57, 59]. However, there was a problem with the usability of EEM because the available parameter sets had a low coverage in chemical space.

We prepared a set of 4475 distinct small organic, drug-like molecules containing these elements: H, C, N, O, F, P, S, Cl, Br, and I, in different functional groups. This set was created so that each selected atom type is contained in at least 100 molecules. CORINA calculated the 3D structure for all molecules. The next step was the calculation of reference QM charges. We selected 6 different approaches: B3LYP/6-311G/MPA, B3LYP/6-311G/NPA, B3LYP/6-311G/AIM, HF/6-311G/MPA, HF/6-311G/NPA, and HF/6-311G/AIM. EEM parameterization was performed for six QM charge calculation approaches, and the whole set of prepared molecules was used.

Our new EEM parameters get very high quality – all coefficients of determination had a value greater or equal to 0.9. We also showed that the used QM approach did not prove any difference in the quality of parameters – B3LYP and HF produced comparable results. EEM parameters based on NPA and AIM population analysis are slightly better than EEM parameters based on MPA.

We also calculated coverage of our parameters previously published parameters on four big chemoinformatics databases of drug-like molecules — DrugBank, [94] ChEMBL, [4] PubChem, [2] and ZINC [3]. We found out that their coverage is less than 60% for most of the previously published parameters. Our newly produced parameters showed coverage of at least 90% for these databases. Consistency of coverage points out that this problem is not related to a database but concerns a chemical space of drug-like molecules and their atom types.

For evaluation of quality, we selected 657 approved drugs from the DrugBank database. We compared the coefficient of determination, root mean square error and found out that our new parameters outperform the previously published parameters. Coverage of the old parameters on this small evaluation dataset is like coverage on whole databases.

The quality of EEM parameters is also affected by a used QM charge scheme. EEM parameters derived from MPA, NPA, and AIM charges showed high quality. EEM parameters based on Hirshfeld charges were acceptable, and MK and

CHELPG charges cannot be used with EEM. On the other hand, none of the QM theory level and basis set combinations showed any problem in the quality of EEM parameters. This also confirmed that we used an appropriate selection of reference QM charges.

We also evaluated already existing tools for calculating partial atomic charges, and all the tools showed some issues. For example, OpenBabel tool [92] is using Bult2002_mpa parameter set [57], but developers extended this set about missing atom types with parameters for different atom types. This hack increases coverage paid by decreased EEM charges quality for molecules containing atom type missing in the original parameter set.

5.2.1 My contribution

I participated in the study's design, and I cooperated in the preparation of the input data (molecules and published EEM parameters) and in QM charge calculation. I performed the analyses and the interpretation of the data.

5.3 NEEMP: software for validation, accurate calculation, and fast parameterization of EEM charges

This article describes NEEMP – a software tool with three main functionalities – parametrization of EEM charges from reference QM charges, validation of EEM parameter sets (including quality and coverage calculation), and EEM charge calculation.

NEEMP provides two different parametrization approaches:

- linear regression (LE),
- differential evolution with the local minimization (DE-MIN) approach.

A combination of a global optimization method with a local optimization method improves EEM parameterization. This combined approach provides a more robust methodology than LR. Therefore it is applicable even for heterogeneous training sets. Specifically, we combined differential evolution (DE) [65] with the local minimization method NEWUOA [95]. Quality criteria for evaluation of each iteration of the parametrization process (both LR and DE-MIN) can be set up to coefficient of determination (R^2) or the root mean square error (RMSE).

The validation mode of NEEMP calculates quality metrics, coverage, and a graphical representation of EEM charge correlation with reference QM charges.

The calculation mode of NEEMP provides a calculation of EEM charges using an input parameter set.

The article also presents two case studies to show the functionality and performance of NEEMP – a parametrization and a validation case study.

The parametrization case study targets a comparison of the parameterization method (LR vs. DE-MIN) and metrics for model validation (R^2 vs. RMSE). The case study proved that LR (with both metrics) is suitable for smaller and homogeneous datasets. DE-MIN (with RMSD metric) is a more robust approach that can also handle the parametrization of larger and more heterogeneous datasets. The validation case study provided similar findings to the previous article – low coverage of the older parameter sets. Also, a quality validation agrees with the previous article for smaller datasets with molecules comprised of C, H, N, and O. On the other hand, the case study uncovered an interesting problem: in larger and more heterogeneous datasets – the parameters set from our previous article proved accuracy problems with some atom types. Using NEEMP, we computed parameter sets, which are also applicable for such problematic datasets.

5.3.1 My contribution

I prototyped the DE-Min approach and designed some new NEEMP features, such as the usage of RMSE instead of R^2. I also implemented a preparation of validion reports.

6

Follow-up work and future plans

From these results, we were able to create a tool for the calculation of charge descriptors – ChrgDescCalc.py [96], and we also successfully prepared the universal model for pK_a prediction [97].

The EEM parameters computed in our publications [98, 99] became a part of AtomicChargeCalculator [100] and also its successor AtomicChargeCalculator II [101].

A parameterization approach from an NEEMP article [99], was further extended by my coworker J. Pazúriková in an article [102]. A further extension of the parameterization approach (optGM) was done by my colleague T. Raček and it allowed a parameterization of the majority of empirical charge calculation methods (not only EEM). An article describing optGM is now in a review process.

In the future, the development of an empirical charge calculation approach for proteins and a parameter set fully covering Protein Data Bank is in development.

Part IV

Conclusion

7

Conclusion

The acid dissociation constant (pK_a) is an important property of organic molecules. Its prediction (especially for unsynthesized molecules) is beneficial in the drug design process and other modern life science fields.

Our previous articles [19, 20] (before my dissertation) proved that pK_a is successfully predictable by chemoinformatics models based on partial atomic charges.

In my thesis, I focused on the first topic – analysis of 3D structure sources on pK_a prediction. We proved that the source of the 3D structure had a significant impact on charges and, respectively, on the quality of pK_a prediction models. The models exhibited the best performance for two databases and two software used by these databases for 3D structure generation – a database DTP NCI (where CORINA generates 3D structures) and a database PubChem (3D structures generated by Omega). Other software tools for 3D structure generation required additional MM optimization to produce acceptable or good pK_a prediction models. In this work, we also showed that pK_a prediction models had the best performance when QM or EEM charges (with specific parameter sets) were used. Purely empirical and topological charges (e.g., Gasteiger-Marseli) proved as too approximated for pKa prediction. pK_a prediction models based on EEM charges seemed very promising because they were fast (no time-demanding QM charge calculation was required), and quality was high (comparable to models based on QM charges). However, we also found that the applicability of EEM parameters for drug-like molecules (e.g., if the parameters cover all atomic types present in the molecule) was significantly limited.

It motivated us to focus on the development of EEM parameters suitable for pK_a prediction. Specifically, we prepared a new molecule dataset and successfully computed EEM parameter sets with more extensive coverage for drug-like molecules and excellent quality ($R^2 > 0.9$). These newly published parameter

sets can be easily used in chemoinformatics applications such as virtual screening or QSAR/QSPR modeling. We also prepared an OpenBabel patch with these parameter sets.

During EEM parameters preparation, I realized a need to have a tool for partial charge calculation and parametrization. For this reason, I focused on the development of NEEMP software. NEEMP is the only available tool that provides EEM parametrization, validation of EEM parameters, and calculation of EEM charges. In NEEMP, we also included an improved parametrization process, including the DE-MIN method that can markedly increase the quality of final parameters for heterogeneous datasets. We published NEEMP, and in the article, we also provided two case studies demonstrating NEEMP capabilities. The publication also included new EEM parameters tailored for ligand molecules.

These articles together provide a solid base for preparing chemoinformatics workflows for pK_a prediction, including 3D structures and partial atomic charges. They sum up the current state of the art and distill the best of well-known approaches, tools, and parameters to increase the quality of the final result.

Part V

Appendix

Bibliography

[1] Berman, H. M., Kleywegt, G. J., Nakamura, H., and Markley, J. L. (2014) The protein data bank archive as an open data resource. *Journal of computer-aided molecular design*, **28**, 1009–1014.

[2] Bolton, E. E., Wang, Y., Thiessen, P. A., and Bryant, S. H. (2008) Pubchem: integrated platform of small molecules and biological activities. *Annual reports in computational chemistry*, **4**, 217–241.

[3] Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. (2012) Zinc: A free tool to discover chemistry for biology. *Journal of chemical information and modeling*, **52**, 1757–1768, pMID: 22587354.

[4] Gaulton, A., et al. (2012) Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, **40**, D1100–D1107.

[5] Paulsen, C. E., Armache, J.-P., Gao, Y., Cheng, Y., and Julius, D. (2015) Structure of the TRPA1 ion channel suggests regulatory mechanisms. *Nature*, **520**, 511–517.

[6] Cao, E., Liao, M., Cheng, Y., and Julius, D. (2013) Trpv1 structures in distinct conformations reveal activation mechanisms. *Nature*, **504**, 113–118.

[7] Prota, A. E., Bargsten, K., Zurwerra, D., Field, J. J., Díaz, J. F., Altmann, K.-H., and Steinmetz, M. O. (2013) Molecular mechanism of action of microtubule-stabilizing anticancer agents. *Science*, **339**, 587–590.

[8] Lu, J., et al. (2005) Microrna expression profiles classify human cancers. *Nature*, **435**, 834–838.

[9] Puente, X. S., et al. (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*, **475**, 101–105.

[10] Nayal, M. and Honig, B. (2006) On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*, **63**, 892–906.

[11] Xie, L., Xie, L., and Bourne, P. (2009) A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics*, **25**, i305–i312.

[12] Comer, J. and Tam, K. (2001) *Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies*. Verlag Helvetica Chimica Acta, Postfach, CH-8042 Zürich, Switzerland.

[13] Klebe, G. (2000) Recent developments in structure-based drug design. *Journal of molecular medicine*, **78**, 269–281.

[14] Kim, J. H., Gramatica, P., Kim, M. G., Kim, D., and Tratnyek, P. G. (2007) Qsar modelling of water quality indices of alkylphenol pollutants. *SAR and QSAR in environmental research*, **18**, 729–743.

[15] Citra, M. J. (1999) Estimating the pk_a of phenols, carboxylic acids and alcohols from semi-empirical quantum chemical methods. *Chemosphere*, **1**, 191–206.

[16] Gross, K. C., Seybold, P. G., and Hadad, C. M. (2002) Comparison of Different Atomic Charge Schemes for Predicting pKa Variations in Substituted Anilines and Phenols. *International journal of quantum chemistry*, **90**, 445–458.

[17] Zhang, J., Kleinöder, T., and Gasteiger, J. (2006) Prediction of pKa Values for Aliphatic Carboxylic Acids and Alcohols With Empirical Atomic Charge Descriptors. *Journal of chemical information and modeling*, **46**, 2256–2266.

[18] Kreye, W. C. and Seybold, P. G. (2009) Correlations between quantum chemical indices and the pk_as of a diverse set of organic phenols. *International journal of quantum chemistry*, **109**, 3679–3684.

[19] Svobodová Vařeková, R., Geidl, S., Ionescu, C.-M., Skrehota, O., Kudera, M., Sehnal, D., Bouchal, T., Abagyan, R., Huber, H., and Koca, J. (2011) Predicting pk(a) values of substituted phenols from atomic charges: Comparison of different quantum mechanical methods and charge distribution schemes. *Journal of chemical information and modeling*, **51**, 1795–806.

[20] Svobodová Vařeková, R., Geidl, S., Ionescu, C.-M., Ehota, O., Bouchal, T., Sehnal, D., Abagyan, R., and A, J. (2013) Predicting p ka values from eem atomic charges. *Journal of cheminformatics*, **5**, 18.

[21] Geidl, S. (2011), pka prediction based on atomic charges – *Bachelor's thesis*. URL: https://is.muni.cz/auth/th/ya74p/?lang=en;setlang=en.

[22] Geidl, S. (2013), Predicting pka values from eem atomic charges – *Master's thesis*. URL: https://is.muni.cz/auth/th/g67zh/?lang=en;setlang=en.

[23] Gasteiger, J. and Engel, T. (2006) *Chemoinformatics: a textbook*. John Wiley & Sons.

[24] Sadowski, J. (2008) *3D Structure Generation*, pp. 231 – 261.

[25] Atkins, P. and De Paula, J. (2011) *Physical chemistry for the life sciences*. Oxford University Press.

[26] Chaves, J., Barroso, J. M., Bultinck, P., and Carbo-Dorca, R. (2006) Toward an alternative hardness kernel matrix structure in the electronegativity equalization method (eem). *Journal of chemical information and modeling*, **46**, 1657–1665.

[27] Moller, H., Martinez-Yamout, M., Dyson, H., and Wright, P. (2005) Solution structure of the N-terminal zinc fingers of the Xenopus laevis double-stranded RNA-binding protein ZFa. *Journal of molecular biology*, **351**, 718–730.

[28] Ghafourian, T. and Dearden, J. (2000) The Use of Atomic Charges and Orbital Energies as Hydrogen-bonding-donor Parameters for QSAR Studies: Comparison of MNDO, AM1 and PM3 Methods. *Journal of pharmacy and pharmacology*, **52**, 603–610.

[29] Galvez, J., Garcia, R., Salabert, M. T., and Soler, R. (1994) Charge Indexes. New Topological Descriptors. *Journal of chemical information and modeling*, **34**, 520–525.

[30] Stalke, D. (2011) Meaningful structural descriptors from charge density. *Chemistry*, **17**, 9264–78.

[31] Lyne, P. D. (2002) Structure-based virtual screening: an overview. *Drug discovery today*, **7**, 1047–1055.

[32] Bissantz, C., Folkers, G., and Rognan, D. (2000) Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *Journal of medicinal chemistry*, **43**, 4759–4767.

[33] Cramer, C. (2003) *Essentials of computational chemistry*.

[34] Mulliken, R. S. (1955) Electronic Population Analysis on LCAO-MO Molecular Wave Functions. II. Overlap Populations, Bond Orders, and Covalent Bond Energies. *Journal of chemical physics*, **23**, 1841.

[35] Mulliken, R. S. (1955) Electronic Population Analysis on LCAO-MO Molecular Wave Functions. I. *Journal of chemical physics*, **23**, 1833.

[36] Reed, A. E. and Weinhold, F. (1983) Natural bond orbital analysis of near-Hartree-Fock water dimer. *Journal of chemical physics*, **78**, 4066–4073.

[37] Reed, A. E., Weinstock, R. B., and Weinhold, F. (1985) Natural population analysis. *Journal of chemical physics*, **83**, 735.

[38] Löwdin, P.-O. (1950) On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals. *Journal of chemical physics*, **18**, 365.

[39] Hirshfeld, F. L. (1977) Bonded-atom fragments for describing molecular charge densities. *Theoretica chimica acta*, **44**, 129–138.

[40] Bader, R. F. W. (1985) Atoms in molecules. *Accounts of chemical research*, **18**, 9–15.

[41] Bader, R. F. W. (1991) A quantum theory of molecular structure and its applications. *Chemical reviews*, **91**, 893–928.

[42] Breneman, C. M. and Wiberg, K. B. (1990) Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J. Comput. Chem.*, **11**, 361–373.

[43] Besler, B. H., Merz, K. M., and Kollman, P. A. (1990) Atomic charges derived from semiempirical methods. *Journal of computational chemistry*, **11**, 431–439.

[44] Bayly, C., Cieplak, P., Cornell, W., and Kollman, P. (1992) A well behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model. *J. Phys. Chem.*, **97**, 10269–10280.

[45] Marenich, A. V., Cramer, C. J., and Truhlar, D. G. (2009) Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *Journal of physical chemistry B*, **113**, 6378–96.

[46] Abraham, R., Griffiths, L., and Loftus, P. (2004) Approaches to charge calculations in molecular mechanics. *Journal of Computational Chemistry*, **3**, 407 – 416.

[47] Gasteiger, J. and Marsili, M. (1980) Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, **36**, 3219–3228.

[48] Cho, K.-H., Kang, Y. K., No, K. T., and Scheraga, H. A. (2001) A Fast Method for Calculating Geometry-Dependent Net Atomic Charges for Polypeptides. *Journal of physical chemistry B*, **105**, 3624–3634.

[49] Oliferenko, A. A., Pisarev, S. A., Palyulin, V. A., and Zefirov, N. S. (2006) Atomic Charges via Electronegativity Equalization: Generalizations and Perspectives. *Advances in quantum chemistry*, **51**, 139–156.

[50] Shulga, D., Oliferenko, A., Pisarev, S., Palyulin, V., and Zefirov, N. (2008) Parameterization of empirical schemes of partial atomic charge calculation for reproducing the molecular electrostatic potential. *Doklady Chemistry - DOKL CHEM*, **419**, 57–61.

[51] Mortier, W. J., Ghosh, S. K., and Shankar, S. (1986) Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules. *Journal of the American Chemical Society*, **108**, 4315–4320.

[52] Yang*, Z.-Z., , and Wang, C.-S. (1997) Atom-bond electronegativity equalization method. 1. calculation of the charge distribution in large molecules. *The journal of physical chemistry A*, **101**, 6315–6321.

[53] Rappe, A. K. and Goddard, W. A. (1991) Charge equilibration for molecular dynamics simulations. *Journal of physical chemistry*, **95**, 3358–3363.

[54] Wilmer, C., Kim, K. C., and Snurr, R. (2012) An extended charge equilibration method. *The Journal of Physical Chemistry Letters*, **3**, 2506–2511.

[55] Nistor, R. A., Polihronov, J. G., Müser, M. H., and Mosey, N. J. (2006) A generalization of the charge equilibration method for nonmetallic materials. *Journal of chemical physics*, **125**, 094108.

[56] Baekelandt, B. G., Mortier, W. J., Lievens, J. L., and Schoonheydt, R. A. (1991) Probing the reactivity of different sites within a molecule or solid by direct computation of molecular sensitivities via an extension of the electronegativity equalization method. *Journal of the American Chemical Society*, **113**, 6730–6734.

[57] Bultinck, P., Langenaeker, W., Lahorte, P., De Proft, F., Geerlings, P., Van Alsenoy, C., and Tollenaere, J. P. (2002) The Electronegativity Equalization Method II: Applicability of Different Atomic Charge Schemes. *Journal of physical chemistry A*, **106**, 7895–7901.

[58] Bultinck, P., Vanholme, R., Popelier, P. L. A., De Proft, F., and Geerlings, P. (2004) High-speed Calculation of AIM Charges Through the Electronegativity Equalization Method. *Journal of physical chemistry A*, **108**, 10359–10366.

[59] Svobodová Vařeková, R., Zuzanna, J., Jakub, V., Suchomel, S., and Koca, J. (2007) Electronegativity equalization method: Parameterization and validation for large sets of organic, organohalogene and organometal molecule. *International Journal of Molecular Sciences*, **8**.

[60] Jiroušková, Z., Vařeková, R. S., Vaněk, J., and Koča, J. (2009) Electronegativity equalization method: parameterization and validation for organic molecules using the Merz-Kollman-Singh charge distribution scheme. *Journal of computational chemistry*, **30**, 1174–8.

[61] Ouyang, Y., Ye, F., and Liang, Y. (2009) A modified electronegativity equalization method for fast and accurate calculation of atomic charges in large biological molecules. *Physical chemistry chemical physics*, **11**, 6082–9.

[62] Todeschini, R. and Consonni, V. (2009) *Molecular Descriptors for Chemoinformatics*.

[63] Urdan, T. (2011) *Statistics in Plain English*.

[64] Verzani, J. (2018) *Using R for Introductory Statistics*.

[65] Storn, R. and Price, K. (1997) Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, **11**, 341–359.

[66] Ishihama, Y., Nakamura, M., Miwa, T., Kajima, T., and Asakawa, N. (2002) A rapid method for pk_a determination of drugs using pressure-assisted capillary electrophoresis with photodiode array detection in drug discovery. *Journal of pharmaceutical sciences*, **91**, 933–942.

[67] Babić, S., Horvat, A. J., Pavlović, D. M., and Kaštelan-Macan, M. (2007) Determination of pk_a values of active pharmaceutical ingredients. *TrAC*, **26**, 1043–1061.

[68] Manallack, D. (2007) The pk_a distribution of drugs: Application to drug discovery. *Perspectives in medicinal chemistry*, **1**, 25–38.

[69] Wan, H. and Ulander, J. (2006) High-throughput pk_a screening and prediction amenable for adme profiling. *Expert opinion on drug metabolism & toxicology*, **2**, 139–155.

[70] Cruciani, G., Milletti, F., Storchi, L., Sforna, G., and Goracci, L. (2009) *In silico* pK_a prediction and adme profiling. *Chemistry & biodiversity*, **6**, 1812–1821.

[71] Lee, A. C. and Crippen, G. M. (2009) Predicting pk_a. *Journal of chemical information and modeling*, **49**, 2013–2033.

[72] Rupp, M., Körner, R., and Tetko, I. V. (2010) Predicting the pK_a of small molecules. *Combinatorial chemistry and high throughput screening*, **14**, 307–327.

[73] Fraczkiewicz, R. (2006) *In Silico Prediction of Ionization*, vol. 5. Elsevier.

[74] Ho, J. and Coote, M. (2010) A universal approach for continuum solvent pk_a calculations: Are we there yet? *Theoretica chimica acta*, **125**, 3–21.

[75] Clark, J. and Perrin, D. D. (1964) Prediction of the strengths of organic bases. *Quarterly reviews of the Chemical Society*, **18**, 295–320.

[76] Perrin, D. D., Dempsey, B., and Serjeant, E. P. (1981) *pK_a prediction for organic acids and bases*. Chapman and Hall: New York.

[77] *Software:* acd/pka. URL: https://www.acdlabs.com/products/percepta/predictors/pKa/.

[78] Shelley, J., Cholleti, A., Frye, L., Greenwood, J., Timlin, M., and Uchimaya, M. (2008) Epik: A software program for pka prediction and protonation state generation for drug-like molecules. *Journal of computer-aided molecular design*, **21**, 681–91.

[79] Hilal, S., Karickhoff, S., and Carreira, L. (1995) A rigorous test for sparc's chemical reactivity models: Estimation of more than 4300 ionization pkas. *Quantitative Structure-Activity Relationships*, **14**, 348 – 355.

[80] Sayle, R., Physiological ionization and pka prediction. URL: http://www.daylight.com/meetings/emug00/Sayle/pkapredict.html.

[81] Blower, P. E. and Cross, K. P. (2006) Decision tree methods in pharmaceutical research. *Current topics in medicinal chemistry*, **6**, 31–39.

[82] Liptak, M. D., Gross, K. C., Seybold, P. G., Feldgus, S., and Shields, G. (2002) Absolute pk_a determinations for substituted phenols. *Journal of the American Chemical Society*, **124**, 6421–6427.

[83] Toth, A. M., Liptak, M. D., Phillips, D. L., and Shields, G. C. (2001) Accurate relative pk_a calculations for carboxylic acids using complete basis set and gaussian-n models combined with continuum solvation methods. *Journal of chemical physics*, **114**, 4595–4606.

[84] *Software:* schroinger: Jaguar. URL: https://www.schrodinger.com/products/jaguar.

[85] Dixon, S. L. and Jurs, P. C. (1993) Estimation of pKa for Organic Oxyacids Using Calculated Atomic Charges. *Journal of computational chemistry*, **14**, 1460–1467.

[86] Jelfs, S., Ertl, P., and Selzer, P. (2007) Estimation of pk_a for druglike compounds using semiempirical and information-based descriptors. *Journal of chemical information and modeling*, **47**, 450–459.

[87] Gross, K. and Seybold, P. (2001) Substituent effects on the physical properties and pka of phenol. *International Journal of Quantum Chemistry*, **85**, 569 – 579.

[88] Liu, S. and Pedersen, L. (2009) Estimation of molecular acidity via electrostatic potential at the nucleus and valence natural atomic orbitals. *The journal of physical chemistry. A*, **113**, 3648–55.

[89] Nci open database compounds. Retrieved from http://cactus.nci.nih.gov/ on August 10, 2010.

[90] Vainio, M. J. and Johnson, M. S. (2007) Generating Conformer Ensembles Using a Multi-objective Genetic Algorithm. *Journal of chemical information and modeling*, **47**, 2462–2474.

[91] Miteva, M. A., Guyon, F., and Tufféry, P. (2010) Frog2: Efficient 3d conformation ensemble generator for small compounds. *Nucleic acids research*, **38**, W622–W627.

[92] O'Boyle, N., Banck, M., James, C., Morley, C., Vandermeersch, T., and Hutchison, G. (2011) Open Babel: An Open Chemical Toolbox. *Journal of cheminformatics*, **3**, 33–47.

[93] Landrum, G., Rdkit: Open-source cheminformatics. Retrieved from http://www.rdkit.org on January 10, 2014.

[94] Law, V., et al. (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, **42**, D1091–7.

[95] Zaslavski, A. and Powell, M. (2006) *The NEWUOA software for unconstrained optimization without derivatives*, pp. 255–297.

[96] Hejret, V. (2015), Charge descriptors application in chemoinformatics – *Bachelor's thesis*. URL: https://is.muni.cz/auth/th/vbxsz/?lang=en;setlang=en.

[97] Hejret, V. (2017), Prediction of physico-chemical properties via charge descriptors – *Master's thesis*. URL: https://is.muni.cz/auth/th/vbxsz/?lang=en;setlang=en.

[98] Geidl, S., Bouchal, T., Raček, T., Svobodová Vařeková, R., Hejret, V., Křenek, A., Abagyan, R., and Koča, J. (2015) High-quality and universal empirical atomic charges for chemoinformatics applications. *Journal of Cheminformatics*, **7**.

[99] Raček, T., Pazúriková, J., Svobodová Vařeková, R., Geidl, S., Křenek, A., Falginella, F., Horský, V., Hejret, V., and Koča, J. (2016) Neemp: Software for validation, accurate calculation and fast parameterization of eem charges. *Journal of Cheminformatics*, **8**.

[100] Ionescu, C.-M., Sehnal, D., Falginella, F., Pant, P., Pravda, L., Bouchal, T., Svobodová Vařeková, R., Geidl, S., and Koča, J. (2015) Atomicchargecalculator: Interactive web-based calculation of atomic charges in large biomolecular complexes and drug-like molecules. *Journal of Cheminformatics*, **7**, 50.

[101] Raček, T., Schindler, O., Toušek, D., Horský, V., Berka, K., Koča, J., and Svobodová, R. (2020) Atomic charge calculator ii: web-based tool for the calculation of partial atomic charges. *Nucleic acids research*, **48**.

[102] Pazúriková, J., Křenek, A., and Matyska, L. (2016) Guided optimization method for fast and accurate atomic charges computation. Évora Gómez, J. and Hernandéz-Cabrera, J. J. (eds.), *Proceedings of the 2016 European Simulation and Modelling Conference*, Ghent, Belgicko, pp. 267–274, EUROSIS - ETI.

Main papers

How Does the Methodology of 3D Structure Preparation Influence the Quality of pK_a Prediction?

Stanislav Geidl[1], Radka Svobodová Vařeková[1,*], Veronika Bendová[1], Lukáš Petrusek[1], Crina-Maria Ionescu[1], Zdeněk Jurka[1], Ruben Abagyan[2], Jaroslav Koča[1,*]

[1] National Centre for Biomolecular Research, Faculty of Science and CEITEC, Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic.
[2] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, 9500 Gilman Drive, San Diego, MC 0657, USA.

Journal of Chemical Information and Modeling 2015, **55**:1088–1097.

High-quality and universal empirical atomic charges for chemoinformatics applications

Stanislav Geidl[1], Tomáš Bouchal[1], Tomáš Raček[1,2], Radka Svobodová Vařeková[1,*], Václav Hejret[1], Aleš Křenek[3], Ruben Abagyan[4], Jaroslav Koča[1,*]

[1] National Centre for Biomolecular Research, Faculty of Science and CEITEC, Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic.
[2] Faculty of Informatics, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic.
[3] Institute of Computer Science, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic.
[4] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, 9500 Gilman Drive, San Diego, MC 0657, USA.

Journal of Cheminformatics 2015, **7**:59.

Geidl *et al. J Cheminform (2015) 7:59*
DOI 10.1186/s13321-015-0107-1

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

High-quality and universal empirical atomic charges for chemoinformatics applications

Stanislav Geidl[1†], Tomáš Bouchal[1†], Tomáš Raček[1,2†], Radka Svobodová Vařeková[1*], Václav Hejret[1], Aleš Křenek[3], Ruben Abagyan[4] and Jaroslav Koča[1*]

Abstract

Background: Partial atomic charges describe the distribution of electron density in a molecule and therefore provide clues to the chemical behaviour of molecules. Recently, these charges have become popular in chemoinformatics, as they are informative descriptors that can be utilised in pharmacophore design, virtual screening, similarity searches etc. Especially conformationally-dependent charges perform very successfully. In particular, their fast and accurate calculation via the Electronegativity Equalization Method (EEM) seems very promising for chemoinformatics applications. Unfortunately, published EEM parameter sets include only parameters for basic atom types and they often miss parameters for halogens, phosphorus, sulphur, triple bonded carbon etc. Therefore their applicability for drug-like molecules is limited.

Results: We have prepared six EEM parameter sets which enable the user to calculate EEM charges in a quality comparable to quantum mechanics (QM) charges based on the most common charge calculation schemes (i.e., MPA, NPA and AIM) and a robust QM approach (HF/6-311G, B3LYP/6-311G). The calculated EEM parameters exhibited very good quality on a training set ($R^2 > 0.9$) and also on a test set ($R^2 > 0.93$). They are applicable for at least 95 % of molecules in key drug databases (DrugBank, ChEMBL, Pubchem and ZINC) compared to less than 60 % of the molecules from these databases for which currently used EEM parameters are applicable.

Conclusions: We developed EEM parameters enabling the fast calculation of high-quality partial atomic charges for almost all drug-like molecules. In parallel, we provide a software solution for their easy computation (http://ncbr.muni.cz/eem_parameters). It enables the direct application of EEM in chemoinformatics.

Keywords: Partial atomic charges, Electronegativity Equalization Method, EEM, Quantum mechanics, QM, Drug-like molecules

Background

Partial atomic charges are real numbers describing the distribution of electron density in a molecule, thus providing clues as to the chemical behaviour of molecules. The concept of charges began to be used in physical chemistry and organic chemistry. Afterwards, partial atomic charges were adopted by computational chemistry and molecular modelling, where they serve for calculating electrostatic interactions, describe the reactivity of the molecule etc. Specifically, they are applied in molecular dynamics, docking, conformational searches, binding site predictions etc. Recently, partial atomic charges also became popular in chemoinformatics, as they proved to be informative descriptors for QSAR and QSPR modelling [1–9] and for other applications [10–12]; they can be utilised in pharmacophore design [13–15], virtual

*Correspondence: radka.svobodova@ceitec.muni.cz; jkoca@chemi.muni.cz
†Stanislav Geidl, Tomáš Bouchal and Tomáš Raček are joint first authors
[1] National Centre for Biomolecular Research, Faculty of Science and CEITEC, Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic
Full list of author information is available at the end of the article

screening [16–18], similarity searches [19–21], molecular structure comparison [22–24] etc.

The partial atomic charges cannot be determined experimentally or derived straightforwardly from the results of quantum mechanics (QM), and many different methods have been developed for their calculation. The most common method for charge calculation is an application of the QM approach and afterwards the utilisation of a charge calculation scheme. Charge calculation schemes can be based on orbital-based population analysis, on wave-function-dependent physical observables or on reproducing charge-dependent observables. Examples of orbital-based population analyses are Mulliken population analysis (MPA) [25, 26], Löwdin population analysis [27] and Natural population analysis (NPA) [28, 29]. Wave-function-dependent physical observables are used in the atoms-in-molecules (AIM) approach [30, 31], Hirshfeld population analysis [32–34], CHELPG [35] and Merz-Singh-Kollman (MK) [36, 37] method. The reproduction of charge-dependent observables is applied in the CM1, CM2, CM3, CM4, and CM5 approaches [38, 39].

Unfortunately, QM charge calculation approaches are very time-consuming. A markedly faster alternative is to employ empirical charge calculation approaches, which can also provide high-quality charges. These approaches can be divided into conformationally-independent, which are based on 2D structure (e.g., Gasteiger's and Marsili's PEOE [40, 41], GDAC [42], KCM [43], DENR [44]) and conformationally-dependent, calculated from 3D structure (e.g., EEM [45], QEq [46] or SQE [47, 48]). We would like to highlight that conformationally-dependent charges are considered to be more suitable for chemoinformatics applications [1–3, 7, 12, 20]. The reason is that these charges contain extensive information not only about chemical surrounding of atoms, i.e., its topology (2D structure based charges) but also geometry and "chemical quality" of the surrounding. Such information is missing, for example, in force field charges which use averaged atomic charges from large sets of structures. Therefore we only focus on conformationally-dependent atomic charges.

Electronegativity equalization method (EEM) is the most frequently used conformationally-dependent empirical charge calculation approach. It calculates charges using the following system of linear equations:

$$\begin{pmatrix} B_1 & \frac{\kappa}{R_{1,2}} & \cdots & \frac{\kappa}{R_{1,N}} & -1 \\ \frac{\kappa}{R_{2,1}} & B_2 & \cdots & \frac{\kappa}{R_{2,N}} & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\kappa}{R_{N,1}} & \frac{\kappa}{R_{N,2}} & \cdots & B_N & -1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \\ \bar{\chi} \end{pmatrix} = \begin{pmatrix} -A_1 \\ -A_2 \\ \vdots \\ -A_N \\ Q \end{pmatrix} \quad (1)$$

where q_i is the charge of an atom i; $R_{i,j}$ is the distance between atoms i and j; Q is the total charge of the molecule; N is the number of atoms in the molecule; κ is the molecular electronegativity, and A_i, B_i and κ are empirical parameters. The parameters A_i and B_i vary for individual atom types, where atom type is a combination of element type and maximal bond order of the atom i. For example, atom type C2 means that the atom is carbon and it creates at least one double bond with its neighbors. An atom X in the aromatic ring is therefore also included into X2 atom type. The parameters A_i, B_i and κ are molecule independent and they are calculated from QM atomic charges by a process of EEM parameterization [49]. EEM is not only a fast charge calculation approach, but it can also provide highly accurate charges, i.e., they can mimic the QM charges for which EEM has been parameterized. On the other hand, EEM charges can be outperformed in certain situations. Specifically, QEq showed better agreement with experimental dipole moments [46] and SQE is presented as an extension of the EEM to obtain the correct size-dependence of the molecular polarizability [47]. But this drawback is compensated by a fact that the quality of EEM charges was documented by many successful applications [2, 3, 50–55] and they are clearly the most cited empirical conformationally-dependent charges.

Therefore, many EEM parameter sets for various QM charge calculation approaches were published later or recently (see Table 1). In parallel, a few freely available software tools also include an EEM charge calculation method (see Table 2).

EEM recently began to be also used in chemoinformatics, giving very promising results [1–3, 64, 65]. Because of their rapid calculation, they can be easily computed for large sets of molecules (e.g., drug-like compounds). Unfortunately, a broader utilisation of EEM charges in chemoinformatics is now limited by the fact that available EEM parameter sets can only cover part of common organic molecules, as they only contain the parameters for some elements and certain bond orders (Table 1). For the above reasons, our aim with this work is to provide EEM parameter sets that cover most of the drug-like molecules and with accuracy comparable to QM charges. Specifically, we have parameterized EEM for frequently used charge calculation schemes, high enough QM theory levels and a large basis set. Afterwards, we compared the coverage and quality of our EEM parameter sets with previously published EEM parameter sets (see Table 1) and with EEM parameter sets embedded in software tools (see Table 2). Additionally, we have prepared a software solution, enabling the user to easily calculate EEM charges via our EEM parameters.

Table 1 Summary information about published EEM parameters evaluated in this study

QM theory Level + basis set	Charge calc. scheme	EEM parameter set name	Published by	Elements and bond orders included[†]
HF/STO-3G	MPA	Baek1991	Baekelandt et al. [56]	C, O, N, H, P, Al, Si
		Svob2007_cbeg2	Svobodova et al. [49]	C1, C2, O, N1, N2, H, S1
		Svob2007_cmet2	Svobodova et al. [49]	C1, C2, O, N1, N2, H, S1, Fe, Zn
		Svob2007_chal2	Svobodova et al. [49]	C1, C2, O, N1, N2, H, S1, Br, Cl, F, I
		Svob2007_hm2	Svobodova et al. [49]	C1, C2, O, N1, N2, H, S1, F, Cl, Br, I, Fe, Zn
HF/6-31G*	MK	Jir2008_hf	Jirouskova et al. [57]	C1, C2, O, N1, N2, H, S1, F, Cl, Br, Zn
B3LYP/6-31G*	MPA	Bult2002_mpa	Bultinck et al. [58]	C, O, N, H, F
	NPA	Bult2002_npa	Bultinck et al. [58]	C, O, N, H, F
		Ouy2009[‡]	Ouyang et al. [59]	C, O, N, H
		Ouy2009_elem	Ouyang et al. [59]	C, O, N, H
	Hir.	Bult2002_hir	Bultinck et al. [58]	C, O, N, H, F
	MK	Bult2002_mk	Bultinck et al. [58]	C, O, N, H, F
		Jir2008_mk	Jirouskova et al. [57]	C1, C2, O, N1, N2, H, S1, F, Cl, Br, Zn
	CHELPG	Bult2002_che	Bultinck et al. [58]	C, O, N, H, F
	AIM	Bult2004_aim	Bultinck et al. [60]	C, O, N, H, F

[†] An element symbol with no further information (e.g., C) means that the EEM parameters are available for this element bound by all possible bond orders. The element symbol followed by a number (e.g., C1) means that the EEM parameters are only available for this element bound by a bond with an order described using this number

[‡] For this parameter set, C1 represents sp^3 hybridization, C2 sp^2 hybridization, C3 sp hybridization, etc.

Table 2 Information about freely available software tools enabling EEM charge calculation

Software	EEM parameters used by a software
OpenBabel [61]	It contains the embedded EEM parameter set Bult2002_mpa, which was parameterized for B3LYP/6-31G*/MPA charges. It does not allow any other EEM parameter set to be used
Balloon [23]	It contains an embedded EEM parameter set published by Puranen et al. [62], which was calculated by fitting to the MEP field. Balloon's developers claim that the EEM charges calculated via Balloon should be comparable to B3LYP/cc-pVTZ/MPA. It does not allow any other EEM parameter set to be used
EEM SOLVER [63]	It allows the use of any input EEM parameter sets provided by the user. It does not contain any embedded EEM parameter sets

Methods

EEM parameterization (step 1)

All the steps performed during our work are depicted in Fig. 1a. The most challenging part of our work was the EEM parameterization. This step required several tasks (see Fig. 1b) and the quality of the calculated EEM parameters sets depends on the proper accomplishment of all these tasks.

EEM parameterization: selection of atom types to be parameterized

Our goal is to provide EEM parameter sets applicable for most common drug-like molecules. Therefore, we provide EEM parameters for the majority of atom types occurring in these molecules. These atom types are summarized in Table 3 (columns 1–3).

EEM parameterization: preparation of the training set

Our training set contains the 3D structures of 4475 distinct small organic molecules. The molecules were obtained from the DTP NCI database [66] and their 3D structures were generated with CORINA 3.60 [67], without any further geometry optimization. The DTP NCI database collects compounds tested as anticancer drugs (with positive or negative results), therefore it is a database of common drug-like molecules. The training set was created in such a way that each selected atom type is contained in at least 100 molecules. The occurrences of individual atom types in the training set are summarized in Table 3. The list of training set molecules, including their NSC numbers and summary formulas, can be found in (Additional file 1: Table S1).

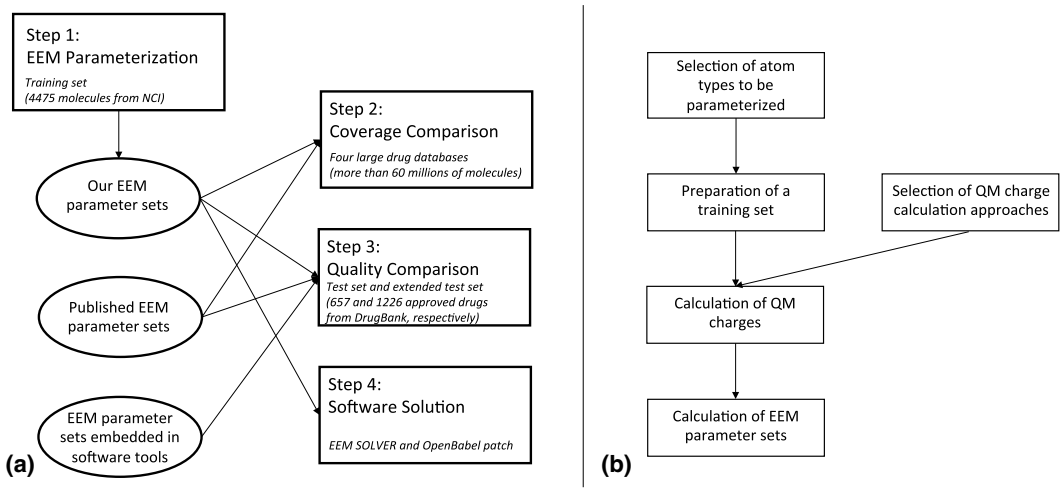

Fig. 1 **a** Composition of steps performed within this work and **b** tasks performed during EEM parametrization

Table 3 Occurrence of atom types in the training set

Denotation of atom type	Element symbol	Maximal bond order	Number of atoms with this atom type in the training set	Number of molecules containing this atom type in the training set
H1	H	1	57,119	4442
C1	C	1	15,220	3447
C2		2	38,097	4149
C3		3	345	266
N1	N	1	4151	2483
N2		2	3383	1879
N3		3	345	266
O1	O	1	5016	2525
O2		2	5793	3069
F1	F	1	938	395
P1	P	1	153	143
P2		2	251	213
S1	S	1	1034	770
S2		2	1391	1211
Cl1	Cl	1	1084	676
Br1	Br	1	336	261
I1	I	1	1734	1365
Total	–	–	136,390	4475

EEM parameterization: selection of QM charge calculation approach

We performed the EEM parameterization for two QM theory levels (B3LYP and HF), one basis set (6-311G) and three charge calculation schemes (MPA, NPA and AIM). We provide the EEM parameters for all combinations of these theory levels, the basis sets and the charge calculation schemes (see Table 4). Theory levels HF and B3LYP were selected, because they are very often used for QM charge calculation and were also successfully used for EEM parameterization several times [49, 56–60]. The basis set 6-311G was used, because it is robust, also covers iodine and moreover, Pople basis sets are very suitable for EEM parameterization. MPA and NPA

Table 4 Quality criteria of our EEM parameter sets

EEM parameter set name	Relevant QM charges	R^2	RMSD	$\bar{\Delta}$
Cheminf_b3lyp_mpa	B3LYP/6-311G/MPA	0.9007	0.1038	0.0727
Cheminf_b3lyp_npa	B3LYP/6-311G/NPA	0.9651	0.0746	0.0540
Cheminf_b3lyp_aim	B3LYP/6-311G/AIM	0.9499	0.0785	0.0558
Cheminf_hf_mpa	HF/6-311G/MPA	0.9178	0.1125	0.0776
Cheminf_hf_npa	HF/6-311G/NPA	0.9633	0.0805	0.0574
Cheminf_hf_aim	HF/6-311G/AIM	0.9441	0.0919	0.0651

Table 5 Size of database, used for comparison of EEM parameter set coverages

Database	Number of compounds
DrugBank	6874
ChEMBL	1,456,020
PubChem	63,676,639
ZINC	21,957,378

population analyses were employed, because they are the most known charge calculation schemes and additionally, EEM is able to mimic MPA and NPA charges very successfully [49, 58, 59]. AIM was selected, because it is based on a different principle from the other two, and EEM can also mimic AIM charges very efficiently [60]. Note that we do not provide EEM parameters for ESP and RESP charges, because it is known that EEM does not mimic these charges well [2, 58].

EEM parameterization: calculation of QM charges

For each molecule from the training set, six sets of QM charges were calculated via the above-mentioned six QM charge calculation approaches. The calculations of QM charges were carried out using Gaussian09 [68]. With the AIM population analysis, the output from Gaussian03 was further processed with the software package AIMAll [69].

EEM parameterization: calculation of EEM parameter sets

For each set of QM charges, the EEM parameterization was performed and the values of the parameters are provided in (Additional file 2: EEM parameters). The software NEEMP [70] was used for the parameterization. This software implements the parameterization methodology described by [49] and introduces several marked improvements into it. NEEMP provides EEM parameter sets together with their quality criteria, i.e., squared Pearson correlation coefficient (R^2), root mean square deviation (RMSD), and average absolute error ($\bar{\Delta}$), calculated via Eqs. (2), (3) and (4), respectively

$$R^2 = \frac{\left(\sum_{i=1}^{N}\left(q_i^{EEM}-\bar{q}^{EEM}\right)\left(q_i^{QM}-\bar{q}^{QM}\right)\right)^2}{\sum_{i=1}^{N}\left(q_i^{EEM}-\bar{q}^{EEM}\right)^2\sum_{i=1}^{N}\left(q_i^{QM}-\bar{q}^{QM}\right)^2} \quad (2)$$

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}\left(q_i^{EEM}-q_i^{QM}\right)^2}{N}} \quad (3)$$

$$\bar{\Delta} = \frac{\sum_{i=1}^{N}\left|q_i^{EEM}-q_i^{QM}\right|}{N} \quad (4)$$

where q_i^{EEM} is the EEM charge of an atom i; q_i^{QM} is the QM charge of an atom i; $\bar{q}^{EEM}$ is an average of all EEM charges; $\bar{q}^{QM}$ is an average of all QM charges, N is the number of atoms in the molecule.

Coverage comparison (step 2)

For comparison, we used our six EEM parameter sets and 15 published EEM parameter sets, described in Table 1 (all 21 of these EEM parameter sets will be below referred to as the tested EEM parameter sets). The coverage comparison was done on four very well-known databases of drug-like chemical compounds: DrugBank [71, 72], ChEMBL [73], PubChem [74], and ZINC [75]. The number of compounds in all these databases (from 10th February 2015) are summarized in Table 5. For each tested EEM parameter set, we analysed how many compounds from the four databases can be covered by them (i.e., contains only atom types present in the tested EEM parameter sets). This coverage analysis was done using NEEMP.

Quality comparison (step 3)

This evaluation was done for the 21 above-mentioned tested EEM parameter sets and was performed on two data sets—a test set (657 molecules) and an extended test set (1226 molecules). The extended test set contained all approved drugs (i.e., drugs which have received approval in at least one country) from the DrugBank database (downloaded 10th February 2015), for which it was possible to calculate all QM charges necessary for testing. The test set was a subset of the extended test set, which contained only molecules covered by all the tested EEM parameter sets. The 2D structures of all molecules were obtained from DrugBank. The lists of molecules from the test set and the extended test set, including their DrugBank IDs and summary formulas, can be found in (Additional file 3: Table S2a; Additional file 4: Table S2b, respectively). The 3D structures of all the molecules were

generated with CORINA 2.6 [67], without any further geometry optimization. For all the molecules, we calculated all the types of QM charges which corresponded to the tested EEM parameters. This means we used the 8 QM charge calculation approaches mentioned in Table 1 and the six QM charge calculation approaches employed for calculating our EEM parameter sets. The calculations of QM charges were done with Gaussian09 and the AIMAll software package was used for AIM charges. We compared the quality of the tested EEM parameter set on both the test set and the extended test set. The comparison was done using NEEMP, which provided quality criteria for all the tested EEM parameter sets. In the extended test set, some molecules were not covered by certain EEM parameter set(s). Therefore, we calculated quality criteria based purely on the covered molecules and in parallel, we also computed the coverage.

Quality comparison: EEM parameter sets embedded in software tools

The calculation of EEM charges can be done with a few software tools, e.g., EEM SOLVER, OpenBabel or Balloon. The software tools OpenBabel and Balloon contain embedded EEM parameter sets (see Table 2). Therefore, we also evaluated the quality of these embedded EEM parameter sets. This evaluation was done for the same data sets and via the same procedure as with the tested EEM parameter sets. The only difference was that the EEM charges were not calculated with NEEMP, but with OpenBabel and Balloon. Afterwards, these EEM charges were compared with the relevant QM charges using R statistical software [76], which provided their quality criteria.

Software solution (step 4)

We provide the user two such solutions, the first based on EEM SOLVER and the second on OpenBabel.

Results and discussion

EEM parameterization (step 1)

EEM parameterization was performed for six QM charge calculation approaches, and a training set containing 4475 drug-like molecules was used. Squared Pearson correlation coefficient (R^2), root mean square deviation (RMSD) and average absolute error ($\overline{\Delta}$) of the obtained EEM parameter sets, calculated for the training set, are summarized in Table 4. These quality criteria describe the correlation between QM charges and the corresponding EEM charges and they were calculated using NEEMP software.

These results show that the quality of our EEM parameter sets is very high, i.e., all the R^2 values are higher or equal to 0.9. Table 4 also illustrates that QM theory levels B3LYP and HF are both applicable for EEM parameterization, and EEM charges based on them have similar accuracy. From this table, we can also see that the quality of EEM parameters based on NPA and AIM population analysis is slightly better than for MPA.

Coverage comparison (step 2)

Information about the coverages of published EEM parameter sets and our EEM parameter sets are summarized in Table 6. The coverages were computed on four well-known databases of drug-like molecules—DrugBank, ChEMBL, PubChem and ZINC. Table 6 shows that the coverages of the published EEM parameter sets are low (<60 %). The only exception are the EEM parameter sets published by Svobodova et al. and Jirouskova et al., which have coverage between 70 and 80 %. In contrast, our EEM parameter sets have very high coverage—about 95 % or more for all the databases. The not covered molecules include atom types rare for drug-like molecules, e.g., metals or boron. An interesting fact is that the coverages are very similar for all four analyzed databases. Therefore, low EEM parameter set coverage is not merely an isolated issue related to one database, but a general problem.

Quality comparison (step 3)

Table 6 summarizes the main quality criteria (i.e., R^2 values) of all tested EEM parameter sets for the test set, which contained 657 approved drugs from DrugBank. Other quality criteria (RMSD and $\overline{\Delta}$) can be found in (Additional file 5: Table S3) and all values of partial atomic charges (represented as tables and as graphs) are in (Additional file 6). The table shows that our EEM parameter sets are among the best performing EEM parameter sets to have been published so far. The table also illustrates that the quality of EEM parameters is strongly influenced by the selection of QM charge calculation scheme. Specifically, EEM parameters based on MPA, NPA and AIM charges are very high quality, and EEM parameters based on Hirshfeld charges are still acceptable. EEM parameters based on MK and CHELPG charges are very low quality, which is in agreement with published data [2, 58]. Both theory levels (HF and B3LYP) and all three basis sets used (STO-3G, 6-31G* and 6-311G) are applicable for EEM parameterization. These results also confirm that our selection of QM theory level, basis set and charge calculation schemes is appropriate.

For the extended test set, the quality criteria exhibit similar trends (see Additional file 7: Table S4). In parallel, the coverages for this data set are slightly higher than for the complete DrugBank database. An interesting fact is that even for such common compounds as approved drugs, the

Table 6 Summary information about coverage and quality of all tested EEM parameters (see below for meaning of colours)

Relevant QM charges		EEM parameter set name	Coverage comparison				Quality comparison
QM theory level + basis set	Charge calc. scheme		Coverage [%]				R^2 Test set
			DrugBank	ChEMBL	PubChem	ZINC	
HF/STO-3G	MPA	Baek1991	58.1	42.3	40.5	40.1	0.8981
		Svob2007_cbeg2	55.0	49.5	47.3	51.9	0.9758
		Svob2007_chal2	71.7	75.2	77.2	80.2	0.9668
		Svob2007_chm2	72.2	75.2	77.3	80.2	0.9623
		Svob2007_cmet2	55.5	49.5	47.3	51.9	0.9676
HF/6-31G*	MK	Jir2008_hf	70.8	74.7	76.5	79.8	0.6872
B3LYP/6-31G*	MPA	Bult2002_mpa	55.4	49.4	48.2	49.6	0.9658
	NPA	Bult2002_npa	55.4	49.4	48.2	49.6	0.8131
		Ouy2009	49.0	41.1	39.1	40.0	0.9655
		Ouy2009_elem	50.0	41.2	39.1	40.0	0.9633
	Hirshfeld	Bult2002_hir	55.4	49.4	48.2	49.6	0.9061
	MK	Bult2002_mk	55.4	49.4	48.2	49.6	0.7844
		Jir2008_mk	70.8	74.7	76.5	79.8	0.7022
	CHELPG	Bult2002_che	55.4	49.4	48.2	49.6	0.7803
	AIM	Bult2004_aim	55.4	49.4	48.2	49.6	0.9739
HF/6-311G	MPA	Cheminf_hf_mpa	94.6	95.7	96.9	100.0	0.9606
	NPA	Cheminf_hf_npa					0.9713
	AIM	Cheminf_hf_aim					0.9791
B3LYP/6-311G	MPA	Cheminf_b3lyp_mpa					0.9552
	NPA	Cheminf_b3lyp_npa					0.9695
	AIM	Cheminf_b3lyp_aim					0.9800

Coverage	> 90%	> 80%	> 70%	> 60%	< 60%
R^2	> 0.95	> 0.9	> 0.85	> 0.8	< 0.8

coverages of published EEM parameter sets are low. Specifically, most published EEM parameter sets have coverages between 55 and 65 %. Further remarkable fact is that quality criteria of our EEM parameters are better for the test set than for the training set. The reason is that the training set is much larger and heterogeneous than the test set.

Quality comparison: EEM parameter sets embedded in software tools

EEM charges produced with OpenBabel were compared with QM charges calculated with B3LYP/6-31G*/MPA. The quality criteria for the test set were the same as for the EEM parameters Bult2002_mpa (i.e., R^2 about 0.97). This was expected, because OpenBabel uses Bult2002_mpa as its embedded EEM parameters. Very surprising was the behavior of OpenBabel on the extended set. The coverage was 100 %, but the quality criteria were markedly lower (e.g., R^2 about 0.82). The reason for this is that OpenBabel replaces the EEM parameters for atom types which are not provided in Bult2002_mpa with the EEM parameters for some other atom types. Unfortunately, this approach is not very reliable, i.e., the quality criteria for molecules which are in the extended test set but are not in the test set are very low ($R^2 = 0.66$). Additionally, this approach is relatively tricky. The user does not know whether the correct or the estimated EEM parameters are used and, therefore, whether the resulting EEM charges will be of a good quality.

The EEM charges produced by Balloon were compared with the QM charges calculated by the B3LYP/cc-pVTZ/MPA approach. The coverage was close to 100 %, but the correlation was also low ($R^2 < 0.8$). On the other hand, the Balloon developers mentioned that the EEM charges provided by Balloon do not correspond directly to some particular QM charges, and they should only be close to B3LYP/cc-pVTZ/MPA charges.

All the quality criteria and coverages for EEM parameter sets embedded in OpenBabel and Balloon are summarized in (Additional file 8: Table S5).

Coverage comparison and quality comparison combined

To date, there have been no EEM parameter sets available which would provide both high coverage and high-quality EEM charges (see Table 6). On the other hand, the EEM parameter sets calculated in this paper solve this problem, because they exhibit coverage close to 100 % and excellent quality criteria. Therefore, they can be used for chemoinformatics applications.

Software solution (step 4)

For the actual applicability of EEM in chemoinformatics, the user doesn't just need EEM parameter sets that are high quality and cover almost all molecules. They also need a software package that embeds these EEM parameter sets and calculates EEM charges based on them. We provide the user with two such solutions. First, we provide our EEM parameter sets in a format that can be directly used in EEM SOLVER (Additional file 2: EEM parameter sets). Second, we provide an OpenBabel patch which allows our EEM parameter sets to be used directly in OpenBabel (Additional file 9: OpenBabel patch). All the information including documentation is also accessible on the web: http://ncbr.muni.cz/eem_parameters. The parameters are also accessible via ACC web application [77].

Conclusion

We provide here six EEM parameter sets which enable the user to calculate EEM charges with quality comparable to frequently used QM charges computed by well-known charge calculation schemes (i.e., MPA, NPA and AIM) and based on a robust QM approach (HF/6-311G, B3LYP/6-311G). The training set for EEM parameterization contained more than 4000 molecules from the DTP NCI drug database, and all six calculated EEM parameter sets exhibited a very good quality on this training set ($R^2 > 0.9$).

The coverage of these computed EEM parameter sets was then compared with the coverages of 15 EEM parameter sets published in the past. This comparison was done on four key databases of drug-like molecules—DrugBank, ChEMBL, Pubchem and ZINC. The comparison showed that our EEM parameter sets enable us to calculate EEM charges for almost all molecules in these databases.

We then compared the quality of computed and published EEM parameter sets on two test data sets composed of approved drugs from DrugBank. This comparison also included EEM parameter sets embedded in the software tools OpenBabel and Balloon. The comparison showed that our EEM parameter sets are among the best performing EEM parameter sets published to date ($R^2 > 0.93$).

To summarize, charge calculation methodology suitable for chemoinformatics applications like virtual screening or QSAR should be fast, conformationally-dependent and accurate. EEM fulfils all these requirements. However, EEM parameter sets that would exhibit high coverage of drug-like molecule databases and provide high quality charges have not been available to date. The EEM parameters calculated in this paper solve this problem. They exhibit coverage close to 100 % and excellent quality criteria, therefore they are applicable in chemoinformatics.

Last but not least, we provide a software solution for the easy computing of EEM charges based on these EEM parameter sets—input files for EEM SOLVER and OpenBabel patch.

Additional files

Additional file 1: Table S1. List of training set molecules, including their NSC numbers and summary formulas.

Additional file 2: EEM parameters. Values of EEM parameter sets for these six charge calculation approaches (i.e. B3LYP/6-311G/MPA,B3LYP/6-311G/NPA, B3LYP/6-311G/AIM, HF/6-311G/MPA, HF/6-311G/NPA, and HF/6-311G/AIM). These EEM parameter sets are in a format which can be used as an input file for EEM SOLVER.

Additional file 3: Table S2a. A list of molecules from the test set including their DrugBank IDs and summary formulas.

Additional file 4: Table S2b. A list of molecules from the extended test set including their DrugBank IDs and summary formulas.

Additional file 5: Table S3. RMSD and $\overline{\Delta}$ values of all tested EEM parameter sets on the test set.

Additional file 6: Charge details. Values of partial atomic charges (represented as tables and as graphs) for all tested EEM parameter sets on the testset.

Additional file 7: Table S4. R^2, RMSD, $\overline{\Delta}$ and coverage values of all tested EEM parameter sets on the extended test set.

Additional file 8: Table S5. RMSD and $\overline{\Delta}$ values for OpenBabel and Balloon on the test set and extended test set.

Additional file 9: OpenBabel patch. A patch for OpenBabel, which enables it to use the EEM parameter sets calculated in this paper.

Authors' contributions

The concept of the study originated from JK and was reviewed and extended by RA, while the design was put together by RSV and SG and reviewed by JK and RA. TB and SG prepared the input data (molecules and published EEM parameters). TB, SG and VH performed QM charge calculation. TR updated and extended NEEMP software. TB and TR performed EEM parameterizations, EEM charges validation and calculation of statistical data. VH prepared an automatic workflow, which is able to reproduce all steps preformed in the article. AK reviewed, corrected and improved this workflow. TR wrote the OpenBabel patch. The data were analyzed and interpreted by RSV, SG and JK. The manuscript was written by RSV in cooperation with JK, and reviewed by all authors. All authors read and approved the final manuscript.

Author details
[1] National Centre for Biomolecular Research, Faculty of Science and CEITEC, Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic. [2] Faculty of Informatics, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic. [3] Institute of Computer Science, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic. [4] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, 9500 Gilman Drive, San Diego, MC 0657, USA.

Acknowledgements
This work was supported by the Grant Agency of the Czech Republic [13-25401S]; the European Community's Seventh Framework Programme (CZ.1.05/1.1.00/02.0068) from the European Regional Development Fund; and by the European Social Fund and the state budget of the Czech Republic (CZ.1.07/2.3.00/20.0042, CZ.1.07/2.3.00/30.0009).

This work was also supported in part by NIH Grants R01 GM071872, U01 GM094612, and U54 GM094618 to R.A.. The access to MetaCentrum supercomputing facilities provided under research intent MSM6383917201 is greatly appreciated.

Authors' information
Stanislav Geidl, Tomáš Bouchal and Tomáš Raček wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors. Radka Svobodová Vařeková and Jaroslav Koča wish it to be known that, in their opinion, they should be regarded as joint Corresponding Authors.

Competing interests
The authors declare that they have no competing interests.

Received: 7 July 2015 Accepted: 16 November 2015
Published online: 02 December 2015

References

1. Svobodová Vařeková R, Geidl S, Ionescu C-M, Skřehota O, Kudera M, Sehnal D, Bouchal T, Abagyan R, Huber HJ, Koča J (2011) Predicting pKa values of substituted phenols from atomic charges: comparison of different quantum mechanical methods and charge distribution schemes. J Chem Inf Model 51(8):1795–1806
2. Svobodová Vařeková R, Geidl S, Ionescu C-M, Skřehota O, Bouchal T, Sehnal D, Abagyan R, Koča J (2013) Predicting pKa values from EEM atomic charges. J Chem Inf 5(1):18
3. Geidl S, Svobodová Vařeková R, Bendová V, Petrusek L, Ionescu C-M, Jurka Z, Abagyan R, Koča J (2015) How does the methodology of 3D structure preparation influence the quality of pKa prediction? J Chem Inf Model 55(6):1088–1097
4. Dixon SL, Jurs PC (1993) Estimation of pKa for organic oxyacids using calculated atomic charges. J Comput Chem 14:1460–1467
5. Zhang J, Kleinöder T, Gasteiger J (2006) Prediction of pKa values for aliphatic carboxylic acids and alcohols with empirical atomic charge descriptors. J Chem Inf Model 46:2256–2256
6. Gross KC, Seybold PG, Hadad CM (2002) Comparison of different atomic charge schemes for predicting pKa variations in substituted anilines and phenols. Int J Quantum Chem 90:445–58
7. Ghafourian T, Dearden JC (2000) The use of atomic charges and orbital energies as hydrogen-bonding-donor parameters for QSAR studies: comparison of MNDO, AM1 and PM3 methods. J Pharm Pharmacol 52(6):603–610
8. Dudek AZ, Arodz T, Gálvez J (2006) Computational methods in developing quantitative structure-activity relationships (QSAR): a review. Comb Chem High Throughput Screen 9(3):213–228
9. Karelson M, Lobanov VS, Katritzky AR (1996) Quantum-chemical descriptors in QSAR/QSPR studies. Chem Rev 96(3):1027–1044
10. Todeschini R, Consonni V (2008) Handbook of molecular descriptors. Wiley-VCH Verlag GmbH, Weinheim
11. Galvez J, Garcia R, Salabert MT, Soler R (1994) Charge indexes. New topological descriptors. J Chem Inf Model 34(3):520–525
12. Stalke D (2011) Meaningful structural descriptors from charge density. Chemistry 17(34):9264–9278
13. Wermuth CG (2006) Pharmacophores: historical perspective and viewpoint from a medicinal chemist. In: Langer T, Hoffmann RD (eds) Pharmacophores and pharmacophore searches, vol 32. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
14. MacDougall PJ, Henze CE (2007) Fleshing-out pharmacophores with volume rendering of the Laplacian of the charge density and hyperwall visualization technology. In: Matta CF, Boyd RJ (eds) The quantum theory of atoms in molecules: from solid state to DNA and drug design. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, pp 499–514
15. Clement OO, Mehl AT (2000) HipHop: pharmacophores based on multiple common-feature alignments. In: Güner OF (ed) Pharmacophore perception, development, and use in drug design. International University Line, La Jolla, pp 69–84
16. Lyne PD (2002) Structure-based virtual screening: an overview. Drug Discov Today 7(20):1047–1055
17. Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 43(25):4759–4767
18. Park H, Lee J, Lee S (2006) Critical assessment of the automated AutoDock as a new docking tool for virtual screening. Proteins 65(3):549–554
19. Kearsley SK, Sallamack S, Fluder EM, Andose JD, Mosley RT, Sheridan RP (1996) Chemical similarity using physiochemical property descriptors. J Chem Inf Model 36(1):118–127
20. Nikolova N, Jaworska J (2003) Approaches to measure chemical similarity—a review. QSAR Comb Sci 22(910):1006–1006
21. Holliday JD, Jelfs SP, Willett P, Gedeck P (2003) Calculation of intersubstituent similarity using R-group descriptors. J Chem Inf Comput Sci 43(2):406–411
22. Tervo AJ, Rönkkö T, Nyrönen TH, Poso A (2005) BRUTUS: optimization of a grid-based similarity function for rigid-body molecular superposition. 1. Alignment and virtual screening applications. J Med Chem 48(12):4076–4086
23. Vainio MJ, Johnson MS (2007) Generating conformer ensembles using a multiobjective genetic algorithm. J Chem Inf Model 47(6):2462–2474
24. Lemmen C, Lengauer T, Klebe G (1998) FLEXS: a method for fast flexible ligand superposition. J Med Chem 41(23):4502–4520
25. Mulliken RS (1955) Electronic Population Analysis on LCAO-MO Molecular Wave Functions. I. J Chem Phys 23(10):1833
26. Mulliken RS (1955) Electronic population analysis on LCAO-MO molecular wave functions. II. Overlap populations, bond orders, and covalent bond energies. J Chem Phys 23(10):1841
27. Löwdin P-O (1950) On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals. J Chem Phys 18(3):365
28. Reed AE, Weinhold F (1983) Natural bond orbital analysis of near-Hartree-Fock water dimer. J Chem Phys 78(6):4066–4073
29. Reed AE, Weinstock RB, Weinhold F (1985) Natural population analysis. J Chem Phys 83(2):735
30. Bader RFW (1985) Atoms in molecules. Acc Chem Res 18(1):9–15
31. Bader RFW (1991) A quantum theory of molecular structure and its applications. Chem Rev 91(5):893–928
32. Hirshfeld FL (1977) Bonded-atom fragments for describing molecular charge densities. Theor Chim Acta 44(2):129–138
33. Ritchie JP (1985) Electron density distribution analysis for nitromethane, nitromethide, and nitramide. J Am Chem Soc 107(7):1829–1837
34. Ritchie JP, Bachrach SM (1987) Some methods and applications of electron density distribution analysis. J Comput Chem 8(4):499–509
35. Breneman CM, Wiberg KB (1990) Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. J Comput Chem 11(3):361–373
36. Singh UC, Kollman PA (1984) An approach to computing electrostatic charges for molecules. J Comput Chem 5(2):129–145
37. Besler BH, Merz KM, Kollman PA (1990) Atomic charges derived from semiempirical methods. J Comput Chem 11(4):431–439
38. Kelly CP, Cramer CJ, Truhlar DG (2005) Accurate partial atomic charges for high-energy molecules using class IV charge models with the MIDI! basis set. Theor Chem Acc 113(3):133–151
39. Marenich AV, Cramer CJ, Truhlar DG (2009) Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. J Phys Chem B 113(18):6378–6396

40. Gasteiger J, Marsili M (1978) A new model for calculating atomic charges in molecules. Tetrahedron Lett 19(34):3181–3184
41. Gasteiger J, Marsili M (1980) Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. Tetrahedron 36(22):3219–3228
42. Cho K-H, Kang YK, No KT, Scheraga HA (2001) A fast method for calculating geometry-dependent net atomic charges for polypeptides. J Phys Chem B 105(17):3624–3624
43. Oliferenko AA, Pisarev SA, Palyulin VA, Zefirov NS (2006) Atomic charges via electronegativity equalization: generalizations and perspectives. Adv Quantum Chem 51:139–156
44. Shulga DA, Oliferenko AA, Pisarev SA, Palyulin VA, Zefirov NS (2010) Fast tools for calculation of atomic charges well suited for drug design. SAR QSAR Environ Res 19(1–2):153–165
45. Mortier WJ, Ghosh SK, Shankar S (1986) Electronegativity equalization method for the calculation of atomic charges in molecules. J Am Chem Soc 108:4315–4320
46. Rappe AK, Goddard WA (1991) Charge equilibration for molecular dynamics simulations. J Phys Chem 95(8):3358–3363
47. Nistor RA, Polihronov JG, Müser MH, Mosey NJ (2006) A generalization of the charge equilibration method for nonmetallic materials. J Chem Phys 125(9):094108
48. Mathieu D (2007) Split charge equilibration method with correct dissociation limits. J Chem Phys 127(22):224103
49. Svobodová Vařeková R, Jiroušková Z, Vaněk J, Suchomel S, Koča J (2007) Electronegativity equalization method: parameterization and validation for large sets of organic, organohalogene and organometal molecule. Int J Mol Sci 8:572–572
50. Janssens GOA, Baekelandt BG, Toufar H, Mortier WJ, Schoonheydt RA (1995) Comparison of cluster and infinite crystal calculations on zeolites with the electronegativity equalization method (EEM). J Phys Chem 99(10):3251–3258
51. Heidler R, Janssens GOA, Mortier WJ, Schoonheydt RA (1996) Charge sensitivity analysis of intrinsic basicity of Faujasite-type zeolites using the electronegativity equalization method (EEM). J Phys Chem 100(50):19728–19734
52. Sorich MJ, McKinnon RA, Miners JO, Winkler DA, Smith PA (2004) Rapid prediction of chemical metabolism by human UDP-glucuronosyltransferase isoforms using quantum chemical descriptors derived with the electronegativity equalization method. J Med Chem 47(21):5311–5317
53. Bultinck P, Langenaeker W, Carbó-Dorca R, Tollenaere JP (2003) Fast calculation of quantum chemical molecular descriptors from the electronegativity equalization method. J Chem Inf Comput Sci 43(2):422–428
54. Smirnov KS, van de Graaf B (1996) Consistent implementation of the electronegativity equalization method in molecular mechanics and molecular dynamics. J Chem Soc Faraday Trans 92(13):2469
55. Ionescu C-M, Geidl S, Svobodová Vařeková R, Koča J (2013) Rapid calculation of accurate atomic charges for proteins via the electronegativity equalization method. J Chem Inf Model 53(10):2548–2548
56. Baekelandt BG, Mortier WJ, Lievens JL, Schoonheydt RA (1991) Probing the reactivity of different sites within a molecule or solid by direct computation of molecular sensitivities via an extension of the electronegativity equalization method. J Am Chem Soc 113(18):6730–6734
57. Jiroušková Z, Vařeková RS, Vaněk J, Koča J (2009) Electronegativity equalization method: parameterization and validation for organic molecules using the Merz-Kollman-Singh charge distribution scheme. J Comput Chem 30(7):1174–1178
58. Bultinck P, Langenaeker W, Lahorte P, De Proft F, Geerlings P, Van Alsenoy C, Tollenaere JP (2002) The electronegativity equalization method II: applicability of different atomic charge schemes. J Phys Chem A 106(34):7895–7901
59. Ouyang Y, Ye F, Liang Y (2009) A modified electronegativity equalization method for fast and accurate calculation of atomic charges in large biological molecules. Phys Chem Chem Phys 11(29):6082–6089
60. Bultinck P, Vanholme R, Popelier PLA, De Proft F, Geerlings P (2004) High-speed calculation of AIM charges through the electronegativity equalization method. J Phys Chem A 108(46):10359–10366
61. O'Boyle N, Banck M, James C, Morley C, Vandermeersch T, Hutchison G (2011) Open Babel: an open chemical toolbox. J Chem Inf 3(1):33–47
62. Puranen JS, Vainio MJ, Johnson MS (2010) Accurate conformation-dependent molecular electrostatic potentials for high-throughput in silico drug discovery. J Comput Chem 31(8):1722–1732
63. Svobodová Vařeková R, Koča J (2006) Optimized and parallelized implementation of the electronegativity equalization method and the atom-bond electronegativity equalization method. J Comput Chem 3:396–405
64. Bultinck P, Carbó-Dorca R, Langenaeker W (2003) Negative Fukui functions: new insights based on electronegativity equalization. J Chem Phys 118(10):4349
65. Burden FR, Polley MJ, Winkler DA (2009) Toward novel universal descriptors: charge fingerprints. J Chem Inf Model 49(3):710–715
66. Open NCI Database (2012) Release 4. http://cactus.nci.nih.gov/download/nci/
67. Sadowski J, Gasteiger J (1993) From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. Chem Rev 93:2567–2581
68. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA Jr, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA. Gaussian 09, Revision E.01. http://www.gaussian.com
69. Todd A Keith (2015) AIMAll 15.05.18. http://aim.tkgristmill.com
70. Raček T, Svobodová Vařeková R, Křenek A, Koča J NEEMP—tool for parameterization of empirical charge calculation method EEM. http://ncbr.muni.cz/neemp/
71. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M(2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36(Database issue):901–906
72. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS (2004) DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 42(Database issue):1091–1097
73. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger FA, Light Y, Mak L, McGlinchey S, Nowotka M, Papadatos G, Santos R, Overington JP (2014) The ChEMBL bioactivity database: an update. Nucleic Acids Res 42(Database issue):1083–1090
74. Bolton EE, Wang Y, Thiessen PA, Bryant SH (2008) PubChem: integrated platform of small molecules and biological activities. In: Wheeler R, Spellmeyer D (eds) Annual Reports in Computational Chemistry, vol. 4, Chap 12. Elsevier, Oxford
75. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) ZINC: a free tool to discover chemistry for biology. J Chem Inf Model 52(7):1757–1768
76. R Core Team R: A Language and Environment for Statistical Computing. http://www.r-project.org/
77. Ionescu CM, Sehnal D, Falginella FL, Pant P, Pravda L, Bouchal T, Svobodová Vařeková R, Geidl S, Koča J (2015) AtomicChargeCalculator: interactive web-based calculation of atomic charges in large biomolecular complexes and drug-like molecules. J Cheminf 7(1):50

NEEMP: Software for validation, accurate calculation and fast parameterization of EEM charges

Tomáš Raček[1,2,3], Jana Pazúriková[1,3], Radka Svobodová Vařekova[1,2,*], Stanislav Geidl[1,2], Aleš Křenek[1,2], Francesco Luca Falginella[1], Vladimír Horský[1,3], Václav Hejret[1], Jaroslav Koča[1,2]

[1] CEITEC – Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic.
[2] National Centre for Biomolecular Research, Faculty of Science, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic.
[3] Faculty of Informatics, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic.
[4] Institute of Computer Science, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic.

Journal of Cheminformatics 2016, **8**:1

Raček et al. J Cheminform (2016) 8:57
DOI 10.1186/s13321-016-0171-1

Journal of Cheminformatics

SOFTWARE

Open Access

NEEMP: software for validation, accurate calculation and fast parameterization of EEM charges

Tomáš Raček[1,2,3], Jana Pazúriková[1,3], Radka Svobodová Vařeková[1,2*], Stanislav Geidl[1,2], Aleš Křenek[1,4], Francesco Luca Falginella[1], Vladimír Horský[1,3], Václav Hejret[1] and Jaroslav Koča[1,2]

Abstract

Background: The concept of partial atomic charges was first applied in physical and organic chemistry and was later also adopted in computational chemistry, bioinformatics and chemoinformatics. The electronegativity equalization method (EEM) is the most frequently used approach for calculating partial atomic charges. EEM is fast and its accuracy is comparable to the quantum mechanical charge calculation method for which it was parameterized. Several EEM parameter sets for various types of molecules and QM charge calculation approaches have been published and new ones are still needed and produced. Methodologies for EEM parameterization have been described in a few articles, but a software tool for EEM parameterization and EEM parameter sets validation has not been available until now.

Results: We provide the software tool NEEMP (http://ncbr.muni.cz/NEEMP), which offers three main functionalities: EEM parameterization [via linear regression (LR) and differential evolution with local minimization (DE-MIN)]; EEM parameter set validation (i.e., validation of coverage and quality) and EEM charge calculation. NEEMP functionality is shown using a parameterization and a validation case study. The parameterization case study demonstrated that LR is an appropriate approach for smaller and homogeneous datasets and DE-MIN is a suitable solution for larger and heterogeneous datasets. The validation case study showed that EEM parameter set coverage and quality can still be problematic. Therefore, it makes sense to verify the coverage and quality of EEM parameter sets before their use, and NEEMP is an appropriate tool for such verification. Moreover, it seems from both case studies that new EEM parameterizations need to be performed and new EEM parameter sets obtained with high quality and coverage for key structural databases.

Conclusion: We provide the software tool NEEMP, which is to the best of our knowledge the only available software package that enables EEM parameterization and EEM parameter set validation. Additionally, its DE-MIN parameterization method is an innovative approach, developed by ourselves and first published in this work. In addition, we also prepared four high-quality EEM parameter sets tailored to ligand molecules.

Keywords: Partial atomic charges, Electronegativity equalization method, EEM, EEM parameterization, wwPDB CCD database

*Correspondence: radka.svobodova@ceitec.muni.cz
[1] CEITEC – Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic
Full list of author information is available at the end of the article

Background

Information about electron density distribution in a molecule is very useful, because it gives us an insight into the chemical behavior of the molecule and helps us to understand its reactivity. We can express this information via the electron populations of orbitals. But this approach is highly complex, resource-demanding and inconvenient for applications. A markedly more efficient solution is to summarize the electron density "belonging" to each atom into one overall number—partial atomic charge. The concept of partial atomic charges was first applied in physical and organic chemistry, and because of its usefulness and intuitiveness it was also adopted in computational chemistry (e.g., docking [1], conformers generation [2] or molecular dynamics [3, 4]), bioinformatics (e.g., similarity searches [5, 6], molecular structure comparison [7, 8]) and chemoinformatics (e.g., QSAR and QSPR modelling [9–14], pharmacophore design [15], virtual screening [16]).

The most common and also the most accurate charge calculation method is the application of quantum mechanics (QM). Specifically, QM is employed for calculating electron orbital populations, and the populations are divided among the individual atoms using a charge calculation scheme. Unfortunately, there is no one universal and best method for QM charge calculation. We can use various combinations of QM theory level and basis set to obtain information about electron distribution in the orbitals. In addition, we can also apply different charge calculation schemes to process this information and obtain a sum of electron density for each individual atom. Well-known charge calculation schemes are for example Mulliken population analysis (MPA) [17, 18], Natural population analysis (NPA) [19, 20], the atoms-in-molecules (AIM) approach [21, 22], CHELPG [23] and Merz-Singh-Kollman (MK) [24, 25] method. Therefore, many various combinations of QM theory level, basis set and charge calculation schemes can be used for QM charge calculation. Different combinations are suitable for different types of applications.

Although a wide spectrum of QM charge calculation methods are available, all the methods have a major limitation—they are very time-demanding. For this reason, empirical charge calculation approaches have been developed [3, 26–33]. One of the most frequently used empirical approaches is the electronegativity equalization method (EEM). It is based on DFT and it calculates the charges via the following equation set:

$$\begin{bmatrix} B_1 & \frac{\kappa}{R_{1,2}} & \cdots & \frac{\kappa}{R_{1,N}} & -1 \\ \frac{\kappa}{R_{2,1}} & B_2 & \cdots & \frac{\kappa}{R_{2,N}} & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\kappa}{R_{N,1}} & \frac{\kappa}{R_{N,2}} & \cdots & B_N & -1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \\ \bar{\chi} \end{bmatrix} = \begin{bmatrix} -A_1 \\ -A_2 \\ \vdots \\ -A_N \\ Q \end{bmatrix} \quad (1)$$

where q_i is the charge of an atom i; $R_{i,j}$ is the distance between atoms i and j; Q is the total charge of the molecule; N is the number of atoms in the molecule; $\overline{\chi}$ is the molecular electronegativity, and A_i, B_i and κ are empirical parameters. The parameters A_i and B_i vary for individual atom types, where atom type is a combination of element type and maximal bond order of the atom i. For example, the atom type N3 means that the atom is nitrogen and it creates at least one triple bond with its neighbors.

The main advantages of EEM are the following: It provides conformationally dependent charges (i.e., charges sensitive to conformational change), it has low time complexity (i.e., $\theta(N^3)$) and its accuracy is comparable to QM approaches. A limitation of EEM is that it requires a set of empirical parameters (i.e., A_i and B_i and κ). These empirical parameters are calculated from QM charges using a process of EEM parameterization. Consequently, EEM can mimic the QM charge calculation approach for which it was parameterized. In addition, because the EEM parameter set is calculated for a specific dataset of molecules, it provides the highest quality of charges on molecules similar to this dataset. Therefore, the EEM parameterizations are often performed for different QM charge calculation approaches and also for various types of molecules (small organic molecules, peptides, proteins, ligands, organometals etc.) to achieve the best accuracy of EEM charges. A lot of EEM parameter sets were published in the past [34–39] and new EEM parameter sets are still in development [40]. Unfortunately, the EEM parameter sets published in the past often only contain parameters for a few atom types and therefore cannot be used for molecules including other atoms.

Because of the strong demand for EEM parameterization, several EEM parameterization approaches were developed. The most widely known is an application of linear regression (LR), described by [31] and [35] and utilized for the preparation of many EEM parameter sets, e.g., in [34–37, 39, 40]. An alternative approach is differential evolution, described and used in [38]. Also other approaches (e.g., accelerated random search [38], particle swarm optimization algorithm [38]) were tested for EEM parameterization, but they were not applicable. Unfortunately, no software is currently available for EEM parameterization or for the validation of EEM parameters. All the software tools related to EEM (e.g., OpenBabel [41], Balloon [42], EEM Solver [43]) are focused purely on EEM charge calculation.

This motivated us to create such a tool and to provide it to the research community. Specifically, we developed NEEMP—a software for fast EEM parameterization, EEM parameters validation and also EEM charge calculation. NEEMP offers two approaches for EEM parameterization—the standard LR method and differential evolution with local minimization (DE-MIN) approach,

recently developed by ourselves. NEEMP also provides two validation modes—a validation of EEM charge quality and coverage. The quality validation compares EEM charges with relevant QM charges and reports common correlation coefficients. The coverage validation analyzes how large a proportion of the molecules from the input database can be processed using the validated EEM parameter set (therefore the validated EEM parameter set covers these molecules).

NEEMP is available here: http://ncbr.muni.cz/NEEMP, source codes are also in (Additional file 1). NEEMP is also documented in Bio.Tools [57]—a portal of bioinformatics resources world-wide.

NEEMP performance was demonstrated via two case studies—the first was focused on EEM parameterization and the second on EEM parameter validation. In both case studies, we worked with molecules from the databases which are very interesting and important for the life science community. Specifically, the wwPDB CCD database [44] of all ligands present in biomacromolecular structures, the DrugBank database [45] of drug compounds and the PubChem database [46], containing a huge amount of organic molecules.

Description of the tool

NEEMP offers the user three modes—calculation, parameterization and validation mode.

Calculation mode

In this mode, NEEMP calculates EEM charges for the input molecule(s) using a user-defined EEM parameter set. Therefore, this mode requires 3D structure(s) of the input molecule(s), information about their total charge (0 for neutral molecules, nonzero real number for charged molecules) and the input EEM parameter set. The charge calculation is performed using Eq. (1) and the values of EEM charges are returned.

Parameterization mode

This mode is for calculating EEM parameters. An input for this calculation is a training set of molecules (i.e., their 3D structures) and QM charges for each molecule. NEEMP can calculate EEM parameters for neutral molecules and also for ions. The parameterization can be performed via two approaches: LR and DE-MIN.

The LR approach is implemented according to its description in [35]. The only extension is that the previous implementation only enables the best performing EEM parameter set to be selected via searching for the highest squared Pearson coefficient (R^2). NEEMP also offers a selection based on the lowest average atom type root mean square difference ($avg(RMSD_a)$). The $avg(RMSD_a)$ is calculated as an average of root mean square difference values for individual atom types ($RMSD_a$). For simplification, the $avg(RMSD_a)$ metrics will be abbreviated below as "RMSD metrics". Optionally, the program can also attempt to discard some of the molecules in the training set, which may yield better results in some situations.

The DE-MIN algorithm is one that we recently developed ourselves. Advanced EEM parameterization approaches [38] usually combine global optimization methods (evolution algorithms, genetic algorithms, simulated annealing) with local optimization methods (simplex method, conjugated gradients or other). These advanced approaches search for the set of EEM parameters that fit QM charges from the training set in the best possible way. They offer a more robust approach than LR, therefore they are applicable even for the heterogeneous training set. We combined differential evolution (DE) with local minimization, which has not been done before. DE starts with generating a random population of vectors, each vector consisting of κ, A_i and B_i for all atom types. Afterwards, all vectors (i.e., EEM parameter sets) are evaluated: EEM charges are computed using the parameter set and compared to QM charges via the chosen metrics (R^2, $RMSD$). Vectors with at least slightly promising results (e.g., $R^2 > 0.2$ and $R > 0$) are minimized by the local minimization method NEWUOA [47]. This step significantly increases the quality of population vectors. Then evolution is mimicked over many iterations: a new vector is created as a combination of two vectors randomly selected from the population. Again, if the vector is promising, we apply local minimization. The best vector found during the evolution iterations is polished again via a few more iterations of NEWUOA and presented as the result. Because of the random generation of the vectors, the DE-MIN approach works stochastically, i.e., even for identical inputs, the results will slightly differ.

Validation mode

This mode enables us to perform two types of EEM parameter set validation—coverage validation and quality validation.

The coverage validation analyzes, how large a proportion of the molecules from the input database are composed only of the atom types included in the input EEM parameter set. This means they are "covered" by this EEM parameter set. Therefore, the coverage validation requires an input database (containing the 3D structures of molecules) and the validated EEM parameters. This validation returns a count and a percentage of molecules from the database which are covered by the parameters. Additionally, it identifies which particular molecules are covered and which are not.

The quality validation tests the accuracy of the EEM charges produced by the input EEM parameter set on the

validated dataset. Therefore, the inputs are the validated EEM parameters, the dataset (3D structures of molecules) and relevant QM charges for each molecule. This validation of quality provides three types of quality criteria—summary criteria (calculated for the whole dataset), atom type criteria (calculated for all the atom types available in the validated EEM parameter set) and criteria for individual molecules. The summary validation criteria are the Pearson coefficient (R), the squared Pearson coefficient (R^2), the Spearman coefficient, the squared Spearman coefficient, root mean square deviation ($RMSD$), absolute average difference (Δ) and maximal absolute difference (Δ_{max})). The atom type criteria and the criteria for individual molecules are the same, but they are calculated for all the relevant atom types or molecules, respectively. In addition, NEEMP also generates graphs depicting the correlation between reference charges and EEM charges. Specifically, it creates graphs showing the dependency for all the atoms (see Fig. 1a) and also graphs for individual atomic types (see Fig. 1b).

Implementation

NEEMP is implemented as a single C program which switches among its three modes (calculation, parameterization, and validation) according to a command line option. Therefore, its distribution is trivial—only a single binary and a few libraries for a particular platform are downloaded. In total, the program size is approximately 5000 lines of code.

The most compute-intensive part (common to all program modes) is the solution of the linear equation system (1). We use LAPACK DSPSV/DSPSVX calls [48] (Cholesky matrix factorization followed by backward substitution and optionally iterative refinement). The LR parameterization method solves another system of linear equations to do the least squares fitting; in this case we use a LAPACK DGELS call (QR factorization) which can handle nearly singular matrices more accurately.

Both open-source and Intel MKL LAPACK implementations are supported.

The DE-MIN parameterization method uses NEWUOA local minimization, the program links to Powell's implementation in its references [47].

The program utilizes simple, coarse grain parallelism. Using the Open MP programming paradigm, several loops—charge calculation for multiple molecules, evaluation over different κ values in the LR method, and minimization of multiple parameter sets in DE-MIN—run in parallel on available CPU cores. Of these, the first provides the best speedup.

The program only supports a single file format (SDF), using its internal file parsing routine, hence does not introduce dependencies on other libraries. Other file formats can be easily converted using 3rd-party tools (e.g., Open Babel).

Results and discussion

We prepared two case studies to show the functionality and performance of NEEMP. The first case study is focused on EEM parameterization and the second on the validation of EEM parameters.

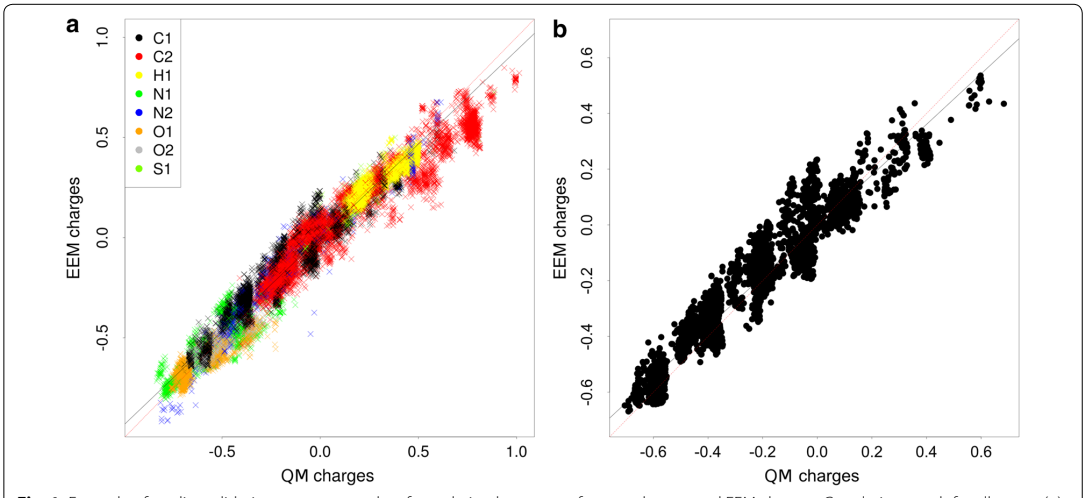

Fig. 1 Example of quality validation outputs–graphs of correlation between reference charges and EEM charges. Correlation graph for all atoms (**a**) and correlation graph for C1 atom type (**b**)

Parameterization case study

This case study first compares EEM parameterization approaches (Parameterization comparison case study), then shows the parameterization running times (Parameterization running time case study) and afterwards focuses on EEM parameterization for wwPDB CCD (Parameterization calculation case study).

Parameterization comparison case study

Goal Comparison of EEM parameterization approaches (LR vs. DE-MIN, R^2 metric versus *RMSD* metric) and evaluation of which are the most suitable for which types of data.

Datasets preparation In this case study, we used four datasets, which are described in Table 1.

We wanted to demonstrate that NEEMP is able to produce results comparable with previously published data. Therefore, we focused first on datasets for which EEM parameterization has been performed in the past. Specifically, our first two datasets—DTP_small and DTP_large (see Table 1) originate from the DTP NCI database [49] and were used in publications [35] and [40], respectively. In addition, we also wanted to provide new interesting and useful results for the research community. For this reason, we then focused on datasets of interest to bioinformatics and chemoinformatics, and which have never been subjected to EEM parameterization. Specifically, the next two datasets (CCD_gen and CCD_exp, see Table 1) were obtained from the wwPDB Chemical component dictionary (wwPDB CCD) database.

This database contains molecules which are parts of biomacromolecular structures deposited in Protein Data Bank [50]. Therefore, these molecules are highly biologically important and include drug molecules, metabolites, compounds from biochemical pathways, etc. For each molecule, the wwPDB CCD contains two types of coordinates, i.e, ideal coordinates generated by CORINA software [51] (included in our dataset CCD_gen) and model coordinates originating from experimental data (included in our dataset CCD_exp). wwPDB CCD is a database of "raw" structural data, therefore we had to perform several preprocessing steps to create our datasets. In this way we obtained the datasets CCD_gen_all and CCD_exp_all, which we used in the validation case study. But for our EEM parameterization goals, these datasets were too large (about four times larger than the dataset DTP_large). Therefore we reduced the size of datasets by a factor of four. Details about wwPDB CCD preprocessing and a summary of its results can be found in (Additional file 2) and (Additional file 3), respectively. Lists of the molecules in all datasets are in (Additional file 4).

The four datasets enable us to increase how demanding our EEM parameterization was in a stepwise manner and therefore show the strong and weak points of the LR and DE-MIN EEM parameterization approaches. The first dataset (DTP_small) is the easiest—small, with low variability of atomic types, molecules and structure sources. The second dataset (DTP_large) is more

Table 1 Description of datasets used in parameterization case study

		Dataset			
Denotation		**DTP_small**	**DTP_large**	**CCD_gen**	**CCD_exp**
Source database		**DTP NCI**		**wwPDB CCD**	
Number of molecules		1956	4475	4443	
Atomic types (elements and bond orders)		C1, C2, O1,O2, N1, N2,H, S1	H1, C1, C2,C3, N1, N2,N3, O1, O2,F1, P1, P2,S1, S2, Cl1,Br1, I1	H1, C1, C2, C3, N1,N2, N3, O1, O2, F1,P2, S1, S2, Cl1, Br1	
Size of molecules		6-176 atoms	5-124 atoms	3-305 atoms	
Type of molecules		Small organic molecules	Small organic molecules	Small organic and inorganic molecules, organometals, peptides	
Source of 3D structures		Generated by CORINA			Experimental structures
Characterization of a dataset	Variability of atomic types	Low	High		
	Variability of molecules	Low		High	
	Variability of structure sources	Low			High
Reference to publication		[35] (set beg2)	[40]	–	–

ambitious, because it is large and contains a large number of atomic types. The third dataset (CCD_gen) brings further complexity, since it contains heterogeneous types of molecules. The fourth dataset (CCD_exp) is the most challenging, because it has all the demands of CCD_gen and in addition, its structures originate from different experiments performed under highly varied conditions by various scientific teams.

Selection and calculation of QM charges The QM charge calculation approach B3LYP/6-311G/NPA was selected for calculating the QM charges used as inputs for the EEM parameterization. These charges were selected, because the B3LYP theory level, 6-311G basis set and NPA proved to be very suitable for EEM parameterization [37, 38, 40]. In addition, the same combination of B3LYP, 6-311G and NPA was used in publication [40], from which we took the dataset DTP_large. The QM charges were calculated by Gaussian [52] for all molecules from datasets DTP_small, DTP_large, CCD_gen and CCD_exp.

EEM parameterization The EEM parameterization was performed using NEEMP on all four prepared datasets and four different parameterization methodologies were used (LR with R^2 metrics, LR with *RMSD* metrics, DE-MIN with R^2 metrics and DE-MIN with *RMSD* metrics). Thus we obtained 16 EEM parameter sets, including their quality criteria. The molecules in all the datasets were not optimized before performing the EEM parameterization. This strategy was motivated by a fact, that the resulting EEM parameters should be utilized also for non optimized molecules, to keep the procedure of EEM charge calculation quick. The same strategy was successfully used in the past (e.g., in articles [9, 35, 40, 53]).

Comparison of EEM parameterization methods LR and DE-MIN using metrics R^2 and *RMSD*. The main quality criteria of the calculated EEM parameter sets are summarized in Table 2. Complete validation reports for all the EEM parameter sets are in (Additional file 5) and the particular EEM parameter sets are stored in (Additional file 6).

For the simple dataset DTP_small, both LR and DE-MIN provide excellent results and both R^2 and *RMSD* metrics are applicable. Only the combination of DE-MIN with R^2 metrics performs slightly weaker.

For the bigger dataset DTP_large, which contains more atom types, differences between the tested approaches started to appear. Summary quality criteria are still very good for all of the approaches, but only the combinations LR+*RMSD* and DE-MIN+*RMSD* also have acceptable atom types criteria. Interestingly, the performance of LR+*RMSD* and DE-MIN+*RMSD* is still almost the same.

For the dataset CCD_gen, which brings a heterogeneity of molecules, the differences between the approaches markedly increase. LR still has good summary quality criteria, but the atom types quality criteria significantly worsen, even with LR+*RMSD*. Therefore, only the combination DE-MIN+*RMSD* seems to be applicable for this dataset and provides very good quality criteria.

In the last and the most challenging dataset CCD_exp, the tested approaches demonstrate similar trends as for CCD_gen, but even more pronounced. LR also has weak summary quality criteria and the atom types quality criteria are highly problematic. Fortunately, the DE-MIN+*RMSD* approach is still applicable and provides quality criteria only slightly worse than for CCD_gen. A graph of the QM and EEM charges correlation for CCD_exp and the approaches LR+*RMSD* and DE-MIN+*RMSD* are shown in Fig. 2, and demonstrate that with such a large and heterogeneous dataset, the proper choice of EEM parameterization approach is crucial.

Summary of comparison results To conclude, we found that LR (with both metrics) is an appropriate approach for smaller and homogeneous datasets. On the other hand, DE-MIN (with *RMSD* metric) is a markedly more suitable solution for larger and more heterogeneous datasets.

Parameterization running time case study

The performance of NEEMP, measured on a standard personal computer is showed in Table 3.

All measurements were repeated 3 times, and we always considered the minimum running time of all the repetitions (in this way random interference of background activity of the operating system is masked out). Running time varies from a few minutes to several hours. As expected, there is no observable difference between CCD_gen and CCD_exp—the complexity depends on the number of molecules and atoms but not the specific values of atom coordinates or charges. In general, DE-MIN performs significantly better for all datasets. The difference becomes more apparent with a larger number of molecules, being caused by the discard algorithm, which has to examine more options (this step is not necessary for DE-MIN).

As described in the Implementation section, the code can run on multiple CPU cores in the heaviest computations, therefore the computation time can be markedly shortened. Figure 3 shows speedup of the parallel version on different number of CPU cores, i.e., how many times faster the parallel program runs compared to the single-core version. The experiments were run with the DE-MIN method and RMSD metric, on the CCD_exp dataset and using a machine with 4 Intel Xeon E7-4860 @ 2.27 GHz CPUs. Again, all measurements were repeated 3 times, and we always considered the minimum running time. The particular values of minimum running time are summarized in (Additional file 7: Table S2).

In the ideal case, if the workload was uniformly distributed among all the cores, the speedup would be the

Table 2 Quality criteria of EEM parameter sets calculated in parameterization comparison case study

Dataset	EEM parameterization		Quality criteria				
			Summary criteria			Atom types criteria	
	Method	Metric	R^2	RMSD	Δ	max($RMSD_a$)	max($Δ_a$)
DTP_small	LR	R^2	0.9718	0.0555	0.0416	0.1061	0.0882
		RMSD	0.9686	0.0592	0.0418	0.0917	0.0709
	DE-MIN	R^2	0.9728	0.1023	0.0809	0.2427	0.2392
		RMSD	0.9700	0.0584	0.0408	0.0926	0.0714
DTP_large	LR	R^2	0.9629	0.0780	0.0554	1.5287	1.3864
		RMSD	0.9531	0.0729	0.0545	0.1767	0.1402
	DE-MIN	R^2	0.9674	0.1875	0.1439	2.0937	2.0671
		RMSD	0.9599	0.0693	0.0515	0.1774	0.1436
CCD_gen	LR	R^2	0.9662	0.0732	0.0497	1.7118	0.6637
		RMSD	0.9484	0.0881	0.0609	0.7908	0.5174
	DE-MIN	R^2	0.9764	1.2229	0.9972	0.9267	6.9413
		RMSD	0.9696	0.0648	0.0449	0.1595	0.1128
CCD_exp	LR	R^2	0.9620	0.0803	0.0526	2.2310	0.8494
		RMSD	0.8848	0.1376	0.0830	2.1176	1.9218
	DE-MIN	R^2	0.9738	0.2764	0.2230	1.4040	1.3970
		RMSD	0.9687	0.0665	0.0466	0.1933	0.1383

Legend:

R^2	> 0.95	> 0.925	> 0.9	>0.85	> 0.8	< 0.8	
RMSD, Δ, max($RMSD_a$), max($Δ_a$)	< 0.05	< 0.1	< 0.15	< 0.2	< 0.3	< 0.4	>= 0.4

same as the number of cores. However, the measurement shows a decrease in efficiency of the parallel execution, which is a consequence of the non-uniform distribution of the workload (existence of non-parallel sections in fact). We can conclude that it is worth running NEEMP with **up to 20 CPU cores**, where we still get an approximately 6x speedup, but using more cores becomes a waste of resources. In general, with larger training sets, when there is more work to evaluate a single parameter vector, the efficiency will improve.

Parameterization calculation case study

Goal In this case study, we would like to obtain high-quality EEM parameters for the wwPDB CCD database and based on several frequently used QM charge calculation approaches. For this purpose, we will apply the knowledge obtained during our comparison of EEM parameterization approaches.

Dataset preparation During this comparison, we prepared two datasets for wwPDB CCD: CCD_gen and CCD_exp. CCD_gen provided EEM parameter sets with better quality criteria, therefore we will use this dataset.

Selection and calculation of QM charges The QM charge calculation approach B3LYP/6-311G/NPA was again selected—for the same reasons as in the parameterization comparison case study. Furthermore, B3LYP/6-311G/MPA was selected, because MPA is often used for EEM parameterization [31, 34–37] as well. Moreover, it was also used in combination with B3LYP/6-311G [40]. Then, the approaches B3LYP/6-311G*/NPA and B3LYP/6-311G*/MPA were selected. The reason for this was that the 6-311G* basis set had never been used for EEM parameterization, and EEM parameters for these approaches can be interesting and useful for the research community. The QM charges were calculated by Gaussian [52] for all molecules from the CCD_gen dataset, except for the B3LYP/6-311G/NPA charges, which were taken from the parameterization comparison case study.

EEM parameterization The EEM parameterization was performed by NEEMP on the CCD_gen dataset for the four above-mentioned QM charges. The DE-MIN+*RMSD* approach was used, because it provides the best results for CCD_gen in the parameterization comparison case study. Thus we obtained 4 EEM parameter sets, including their quality criteria.

Summary of EEM parameterization results The denotations and the quality criteria of the obtained EEM

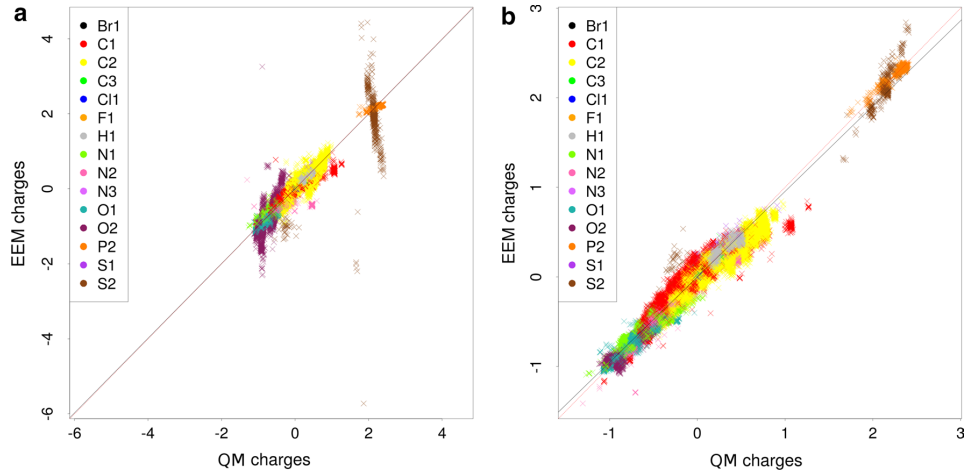

Fig. 2 Graph of QM and EEM charges correlation for dataset CCD_exp and LR+*RMSD* approach (**a**) and DE-MIN+*RMSD* approach (**b**)

Table 3 NEEMP performance on a standard personal computer (Intel i7-4790K CPU @ 4.00GHz)

Dataset	DTP_small		DTP_large		CCD_gen and CCD_exp	
EEM parameterization method	LR	DE-MIN	LR	DE-MIN	LR	DE-MIN
Running time	54 m	14 m	4 h 25 m	16 m	9 h 24m	25 m

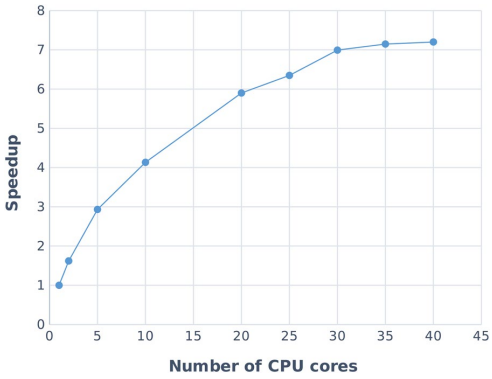

Fig. 3 Speedup achieved by the parallel version run at different number of CPU cores

parameter sets are summarized in Table 4. These results show that NEEMP provided us with four high-quality EEM parameter sets for the wwPDB CCD database. These EEM parameter sets are in (Additional file 6) and validation reports for them are in (Additional file 5).

Validation case study

This case study first analyses the coverage of selected EEM parameter sets (Coverage validation case study) and then also the quality of these sets (Quality validation case study).

Coverage validation case study

Goal In this case study, we would like to compare the coverage of selected EEM parameter sets on key databases of small molecules. In this way, we introduce NEEMP functionality focused on the validation of coverage.

EEM parameter sets Several sets of published EEM parameter sets [34, 35, 37, 38, 40], i.e., the sets which proved to be of good quality in the past [10, 11, 40], and also four EEM parameter sets calculated in the parameterization calculation case study were selected for the coverage comparison. A list of the compared EEM parameter sets, including basic information about them, can be seen in the first three columns of (Additional file 8: Table S3).

Databases The coverage comparison was done on three well-known databases of biologically important small molecules: wwPDB CCD, DrugBank, and PubChem. The number of compounds in all these databases (from March 2016) are summarized in Table 5. wwPDB CCD,

Table 4 Denotations and main quality criteria of EEM parameter sets calculated in parameterization calculation case study

EEM parameter set denotation	Relevant QM charges	Quality criteria				
		Summary criteria			Atom types criteria	
		R^2	RMSD	Δ	max($RMSD_a$)	max($Δ_a$)
Ccd2016_npa	B3LYP/6-311G/NPA	0.9696	0.0648	0.0449	0.1595	0.1128
Ccd2016_mpa	B3LYP/6-311G/MPA	0.9502	0.0711	0.0491	0.2138	0.1276
Ccd2016_npa2	B3LYP/6-311G*/NPA	0.9747	0.0595	0.0438	0.1904	0.1588
Ccd2016_mpa2	B3LYP /6-311G*/MPA	0.9676	0.0582	0.0417	0.1319	0.1019

Legend:

R^2	> 0.95	> 0.925	> 0.9	>0.85	> 0.8	< 0.8	
RMSD, Δ, max($RMSD_a$), max($Δ_a$)	< 0.05	< 0.1	< 0.15	< 0.2	< 0.3	< 0.4	>= 0.4

Table 5 Size of database, used for comparison of EEM parameter set coverages

Database	Number of compounds
DrugBank	7097
wwPDB CCD	21,741
PubChem	71,632,601

which was also used in the parameterization case study, is a medium-sized database including ligands incorporated in biomacromolecules. DrugBank is a relatively small database containing chemical compounds with medical applications. The PubChem database intends to include all common chemical substances, therefore it is very large and heterogeneous.

Coverage comparison procedure The coverage of all the tested EEM parameter sets was calculated via NEEMP for all three databases of interest. The results are summarized in (Additional file 8: Table S3).

Summary of results Interestingly, even though the databases are very different, the coverage values are very similar for all of them. Only the EEM parameter sets calculated recently (i.e., Cheminf2015 and Ccd2016 sets) exhibit sufficient coverage (> 93 % for all the databases). The other parameter sets have low coverage, specifically, they are only applicable for 40–80% of molecules from the tested databases. The coverage values for DrugBank and PubChem agree with information published in [40]. In general, this confirms that coverage is a weakness of the majority of currently published EEM parameter sets. We also showed that NEEMP enables us to easily obtain information about the EEM parameter set coverage for each database of interest.

Quality validation case study

Goal This case study compares the quality of selected EEM parameter sets on two datasets, which contain wwPDB CCD structures. It also shows NEEMP functionality focused on the validation of EEM parameter set quality.

Preparation of datasets Two datasets containing molecules from wwPDB CCD were used for quality comparison—a simple dataset for basic testing and a challenging dataset for deep analysis of EEM parameter set quality. The challenging dataset is specifically the dataset CCD_gen_all, which was prepared in the parameterization comparison case study and which includes structures generated by CORINA. This dataset contains all the wwPDB CCD molecules composed of atoms of C, H, N, O, S, P, F, Cl and Br and that have no structural errors. Therefore, it includes about 82 % of the whole of wwPDB CCD, it is highly chemically heterogeneous and demanding for calculating high-quality EEM charges. The simple dataset (denoted CCD_gen_CHNO) is a subset of CCD_gen_all. The list of molecules in this dataset can be found in (Additional file 4). This dataset was designed for a basic quality test of all the EEM parameter sets used in the coverage validation case study, and so it had to be completely covered by all these EEM parameter sets. For this reason, its molecules contain only the atoms C, H, N and O and do not include triple bonds. This fact implies its low chemical variability. Information about both datasets are summarized in Table 6.

EEM parameter sets The quality comparison was analyzed on the same EEM parameter sets as the coverage comparison. Specifically, when the quality comparison was performed on the dataset CCD_gen_CHNO, all these EEM parameter sets were used. The quality comparison on CCD_gen_all was only performed for the

Table 6 Description of datasets used in quality validation case study

	Datasets	
Designation	CCD_gen_CHNO*	CCD_gen_all*
Source database	wwPDB CCD	wwPDB CCD
Number of molecules	8144	17,769
Atomic types (elements and bond orders)	H1, C1, C2, N1, N2, O1, O2	H1, C1, C2, C3, N1, N2, N3, O1, O2, F1, P2, S1, S2, Cl1, Br1

*All other information about the dataset is the same as for the dataset CCD_gen, described in Table 1

Cheminf2015 and Ccd2016 EEM parameter sets, because only these sets can be applied on all the molecules from this dataset.

Calculation of QM charges For molecules from the dataset CCD_gen_CHNO, we calculated the same charges as in the coverage validation case study, because EEM charges calculated using the tested EEM parameter sets had to be compared with corresponding QM charges. Therefore, the following QM charges were calculated: HF/STO-3G/MPA and NPA, B3LYP/6-31G*/MPA and NPA, B3LYP/6-311G/MPA and NPA, and B3LYP/6-311G*/MPA and NPA. For molecules from the dataset CCD_gen_all, we only calculated QM charges corresponding to the EEM parameter sets Cheminf2015 and Ccd2016. Therefore we calculated the QM charges B3LYP/6-311G/MPA and NPA, and B3LYP/6-311G*/MPA. The QM charges were calculated by Gaussian [52] or (where possible) taken from the parameterization case study.

Quality comparison procedure For the dataset CCD_gen_CHNO, the EEM charges were calculated using the same EEM parameter sets as in the coverage validation case study. Afterwards, these EEM charges were compared with the corresponding QM charges via NEEMP and the validation reports were created. A summary of the most important quality criteria and the validation reports are in (Additional file 9: Table S4) and (Additional file 5), respectively. For the dataset CCD_gen_all, the EEM charges were calculated using only the EEM parameter sets Cheminf2015 and Ccd2016. We then employed NEEMP to compare EEM charges with relevant QM charges and produce validation reports. The most important quality criteria are summarized in Table 7, selected correlation graphs are shown in Fig. 4 and all the validation reports are in (Additional file 5).

Summary of results All the EEM parameter sets proved to be of very high quality on the dataset CCD_gen_CHNO. Both the summary quality criteria and the atom type quality criteria were excellent. Specifically, R^2 was mostly higher than 0.95, $RMSD < 0.08$ and $max(RMSD_a) < 0.12$. This documents the fact that all these EEM parameter sets are very well adjusted for EEM charge calculation on datasets with low atom type variability. The results for the dataset CCD_gen_all are more heterogeneous (see Table 7). The summary criteria are excellent ($R^2 > 0.95$) or at least acceptable ($R^2 \sim 0.9$) for all the EEM parameter sets. But the atom type quality criteria are sometimes problematic. The Cheminf2015_mpa parameter set in particular produced very high $max(RMSD_a)$ and $max(\Delta_a)$ values. Figure 4a and the validation report shows that there is a problem with the correlation of charges on carbon atoms with triple bonds (C3 atoms). Further EEM parameter sets have sufficiently low atom type quality criteria. Two of the EEM parameter sets (Ccd2016_mpa and Cheminf2015_npa) contain slide correlation issues for S2 or C3—see an example in Fig. 4b. The remaining EEM parameter sets exhibited no problems or issues. Furthermore, the quality criteria of all Ccd2016 parameter sets are comparable to the quality criteria obtained during the calculation process (see Table 4). This fact confirmed the robustness of the EEM parameterization performed via NEEMP. In general, these results show that the datasets with high atom type variability can still represent a challenge for the available EEM parameter sets. Therefore, the EEM parameter set quality validation implemented in NEEMP is a very important step in EEM usage and application.

Conclusion

We provide the software tool NEEMP, which offers three main functionalities: EEM parameterization (via the LR and DE-MIN method, with R^2 and *RMSD* metrics); EEM parameter set validation (i.e., validation of coverage and quality) and EEM charge calculation. NEEMP was implemented in C, contains parallelization and provides a fast and accurate solution for work with EEM. To the best of our knowledge, NEEMP is the only available software tool enabling EEM parameterization and EEM parameter set validation. In addition, the DE-MIN parameterization method is an innovative approach, developed by ourselves and first published in this work.

NEEMP functionality is demonstrated on two case studies—a parameterization and a validation case study.

The parameterization case study analyses the performance of both parameterization methods (LR and DE-MIN) and metrics (R^2 and *RMSD*) using four different datasets which increase the demands of EEM parameterization in a stepwise manner. We found that LR (with both metrics) is an appropriate approach for

Table 7 Quality criteria of the EEM parameter sets on the dataset CCD_gen_all

Relevant QM charges		EEM parameter set name	Quality criteria				
			Summary criteria			Atom types criteria	
QM theory level + basis set	Charge calc. scheme		R^2	RMSD	Δ	max($RMSD_a$)	max($Δ_a$)
B3LYP/6-311G	MPA	Cheminf2015_mpa	0.8986	0.1189	0.0597	2.2752	1.1929
		Ccd2016_mpa	0.9568	0.0732	0.0473	0.2884	0.0911
	NPA	Cheminf2015_npa	0.9763	0.0635	0.0446	0.281	0.1916
		Ccd2016_npa	0.9872	0.9254	0.0662	0.1682	0.1219
B3LYP/6-311G*	MPA	Ccd2016_mpa2	0.97	0.0595	0.0406	0.1365	0.102
	NPA	Ccd2016_npa2	0.9796	0.0603	0.0428	0.2002	0.1671

Legend:

R^2	> 0.95	> 0.925	> 0.9	> 0.85	> 0.8	< 0.8	
RMSD, Δ, max($RMSD_a$), max($Δ_a$)	< 0.05	< 0.1	< 0.15	< 0.2	< 0.3	< 0.4	>= 0.4

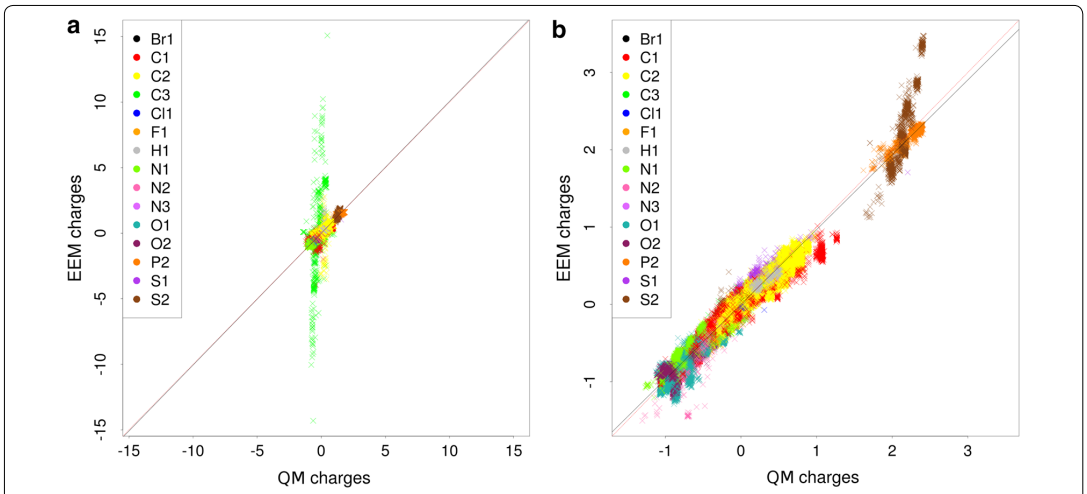

Fig. 4 Graph of QM and EEM charges correlation for Cheminf2015_mpa parameter set (**a**) and Cheminf2015_npa parameter set (**b**) on the dataset CCD_gen_all. The graph for Cheminf2015_mpa includes a marked correlation problem at C3 atoms (they are in *green*), the graph for Cheminf2015_npa shows a slight correlation issue at S2 atoms (they are in *brown*)

smaller and homogeneous datasets. On the other hand, DE-MIN (with *RMSD* metric) is a markedly more suitable solution for larger and more heterogeneous datasets. We also showed that NEEMP is able to perform EEM parameterizations in a reasonable time, and its execution on multiple processors produces a marked speedup. We then performed EEM parameterization via the DE-MIN method with *RMSD* metrics on wwPDB CCD—a database of ligands found in biomacromolecular structures. This database is frequently used by the life science community and it has never been subjected to EEM parameterization. Despite the high heterogeneity of the database, we produced 4 high-quality EEM parameter sets. This demonstrated, that NEEMP is highly applicable for the computation of new EEM parameter sets.

The validation case study focused first on coverage validation. Specifically, we validated the coverage of selected EEM parameter sets (i.e., several published EEM parameter sets and EEM parameter sets provided in this article) on three well-known databases of small molecules

(wwPDB CCD, PubChem and DrugBank). It was shown that the coverage of older EEM parameter sets is problematic. Specifically, they are only applicable for 40–80 % of molecules from the tested databases. Only the recently published Cheminf2015 EEM parameter sets and the EEM parameter sets provided in this article had sufficient coverage (>90–95 %). The case study then also focused on quality validation of the selected EEM parameter sets. All the sets performed very well on a small dataset with molecules comprised of C, H, N and O. On the other hand, the larger and more heterogeneous dataset (17,769 molecules; 15 atom types) was a challenge for most of the tested EEM parameter sets. The older parameter sets could not cover the dataset and the newer ones (i.e., Cheminf 2015) had accuracy problems with some atom types. The only applicable EEM parameter sets were the Ccd2016 sets provided in this article. From these results it can be seen that EEM parameter set coverage and quality can still be problematic. Therefore it makes sense to verify the coverage and quality of EEM parameter sets before their use, and NEEMP is an appropriate tool for such verification.

Moreover, from both case studies it seems that it is still necessary to perform new EEM parameterizations and obtain EEM parameter sets with high quality and coverage on key structural databases.

Last but not least, NEEMP can potentially help the community to perform EEM parameterizations which are challenging. For example, EEM parameterization based on HF/6-31G*/MK QM charges. Mimicking these QM charges via EEM is very important because they are used for AMBER partial-charge parameterization routine focused on biomolecular ligands [54–56]. On the other hand, EEM is documented as an approach which performs very weakly for MK charge calculation scheme [10, 37, 40]. Employing NEEMP can help us to override the problems with MK based EEM parameterizations, or it can confirm limitations of EEM in this domain. Further challenging EEM parameterizations, which can be potentially solved via NEEMP, are parameterizations focused on proteins or metalloproteins—large macromolecules containing long-range interactions.

Availability and requirements

Project name: NEEMP
Project home page: http://ncbr.muni.cz/neemp
Operating system(s): Linux (recommended), Windows, Mac OS X
Programming language: C, external library in Fortran
Other requirements: GNU Fortran, libxml2, LAPACK, zlib, OpenMP
License: GNU GPLv3
Any restrictions on use by non-academics: no restrictions

Additional files

Additional file 1. Source codes. Archive with source codes of NEEMP.

Additional file 2. Preprocessing information. Detailed description of wwPDB CCD preprocessing.

Additional file 3. Summarization of preprocessing results for individual molecules from wwPDB CCD.

Additional file 4. List of molecules. Lists of molecules in all datasets.

Additional file 5. Validation reports. Validation reports for EEM parameter sets, calculated in parameterization case study and for EEM parameter sets, tested in validation case study.

Additional file 6. EEM parameter sets. EEM parameter sets, calculated in the parameterization case study.

Additional file 7: Table S2. NEEMP running times on more CPUs.

Additional file 8: Table S3. Summary information about coverage of tested EEM parameter sets, performed on databases wwPDB CCD, DrugBank and PubChem.

Additional file 9: Table S4. Quality criteria of the EEM parameter sets on the dataset CCD_gen_CHNO.

Authors' contributions

TR: Design and development of NEEMP, work coordination; JP: DE-MIN module design and development, performing of analyses; RSV: Case study experiments design, results evaluation, writing publication; SG: Prototyping of DE-MIN and suggesting of RMSD instead R2, validation reports implementation; AK: NEEMP development, measuring NEEMP running times; FLF: Charge calculation, NEEMP testing, writing NEEMP documentation; VHo: wwPDB validation; VHe: Charge calculation; JK: Review of experimental design, writing publication. All authors read and approved the manuscript.

Author details

[1] CEITEC – Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic. [2] National Centre for Biomolecular Research, Faculty of Science, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic. [3] Faculty of Informatics, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic. [4] Institute of Computer Science, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic.

Acknowledgements

This research has been financially supported by the Ministry of Education, Youth and Sports of the Czech Republic under the project CEITEC 2020 (LQ1601).

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated. In addition, access to the CERIT-SC computing and storage facilities provided by the CERIT-SC Center under the programme "Projects of Projects of Large Research, Development, and Innovations Infrastructures" (CERIT Scientific Cloud LM2015085), is greatly appreciated

Competing interests

The authors declare that they have no competing interests.

Received: 8 July 2016 Accepted: 5 October 2016
Published online: 17 October 2016

References

1. Park H, Lee J, Lee S (2006) Critical assessment of the automated AutoDock as a new docking tool for virtual screening. Proteins 65(3):549–554

2. Vainio MJ, Johnson MS (2007) Generating conformer ensembles using a multiobjective genetic algorithm. J Chem Inf Model 47(6):2462–2474
3. Rappe AK, Goddard WA (1991) Charge equilibration for molecular dynamics simulations. J Phys Chem 95(8):3358–3363
4. Chenoweth K, Van Duin AC, Goddard WA (2008) Reaxff reactive force field for molecular dynamics simulations of hydrocarbon oxidation. J Phys Chem A 112(5):1040–1053
5. Kearsley SK, Sallamack S, Fluder EM, Andose JD, Mosley RT, Sheridan RP (1996) Chemical similarity using physiochemical property descriptors. J Chem Inf Model 36(1):118–127
6. Holliday JD, Jelfs SP, Willett P, Gedeck P (2003) Calculation of intersubstituent similarity using R-group descriptors. J Chem Inf Comput Sci 43(2):406–411
7. Tervo AJ, Rönkkö T, Nyrönen TH, Poso A (2005) BRUTUS: optimization of a grid-based similarity function for rigid-body molecular superposition. 1. Alignment and virtual screening applications. J Med Chem 48(12):4076–4086
8. Lemmen C, Lengauer T, Klebe G (1998) FLEXS: a method for fast flexible ligand superposition. J Med Chem 41(23):4502–4520
9. Svobodová Vařeková R, Geidl S, Ionescu C-M, Skřehota O, Kudera M, Sehnal D, Bouchal T, Abagyan R, Huber HJ, Koča J (2011) Predicting pKa values of substituted phenols from atomic charges: comparison of different quantum mechanical methods and charge distribution schemes. J Chem Inf Model 51(8):1795–1806
10. Svobodová Vařeková R, Geidl S, Ionescu C-M, Skřehota O, Bouchal T, Sehnal D, Abagyan R, Koča J (2013) Predicting pKa values from EEM atomic charges. J Cheminf 5(1):18
11. Geidl S, Svobodová Vařeková R, Bendová V, Petrusek L, Ionescu C-M, Jurka Z, Abagyan R, Koča J (2015) How does the methodology of 3D structure preparation influence the quality of pKa prediction? J Chem Inf Model 55(6):1088–1097
12. Gross KC, Seybold PG, Hadad CM (2002) Comparison of different atomic charge schemes for predicting pKa variations in substituted anilines and phenols. Int J Quantum Chem 90:445–458
13. Galvez J, Garcia R, Salabert MT, Soler R (1994) Charge indexes: new topological descriptors. J Chem Inf Model 34(3):520–525
14. Stalke D (2011) Meaningful structural descriptors from charge density. Chemistry 17(34):9264–9278
15. MacDougall PJ, Henze CE (2007) Fleshing-out pharmacophores with volume rendering of the laplacian of the charge density and hyperwall visualization technology. In: Matta CF, Boyd RJ (eds) The quantum theory of atoms in molecules: from solid state to DNA and drug design. Wiley, Weinheim, pp 499–514
16. Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 43(25):4759–4767
17. Mulliken RS (1955) Electronic population analysis on LCAO-MO molecular wave functions. II. Overlap populations, bond orders, and covalent bond energies. J Chem Phys 23(10):1841
18. Mulliken RS (1955) Electronic population analysis on LCAO-MO molecular wave functions. I J Chem Phys 23(10):1833
19. Reed AE, Weinhold F (1983) Natural bond orbital analysis of near-Hartree-Fock water dimer. J Chem Phys 78(6):4066–4073
20. Reed AE, Weinstock RB, Weinhold F (1985) Natural population analysis. J Chem Phys 83(2):735
21. Bader RFW (1985) Atoms in molecules. Acc Chem Res 18(1):9–15
22. Bader RFW (1991) A quantum theory of molecular structure and its applications. Chem Rev 91(5):893–928
23. Breneman CM, Wiberg KB (1990) Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. J Comput Chem 11(3):361–373
24. Singh UC, Kollman PA (1984) An approach to computing electrostatic charges for molecules. J Comput Chem 5(2):129–145
25. Besler BH, Merz KM, Kollman PA (1990) Atomic charges derived from semiempirical methods. J Comput Chem 11(4):431–439
26. Gasteiger J, Marsili M (1978) A new model for calculating atomic charges in molecules. Tetrahedron Lett 19(34):3181–3184
27. Gasteiger J, Marsili M (1980) Iterative partial equalization of orbital electronegativity–a rapid access to atomic charges. Tetrahedron 36(22):3219–3228
28. Cho K-H, Kang YK, No KT, Scheraga HA (2001) A fast method for calculating geometry-dependent net atomic charges for polypeptides. J Phys Chem B 105(17):3624–3634
29. Oliferenko AA, Pisarev SA, Palyulin VA, Zefirov NS (2006) Atomic charges via electronegativity equalization: generalizations and perspectives. Adv Quantum Chem 51:139–156
30. Shulga DA, Oliferenko AA, Pisarev SA, Palyulin VA, Zefirov NS (2010) Fast tools for calculation of atomic charges well suited for drug design. SAR QSAR Environ Res 19(1–2):153–165
31. Mortier WJ, Ghosh SK, Shankar S (1986) Electronegativity equalization method for the calculation of atomic charges in molecules. J Am Chem Soc 108:4315–4320
32. Nistor RA, Polihronov JG, Müser MH, Mosey NJ (2006) A generalization of the charge equilibration method for nonmetallic materials. J Chem Phys 125(9):094108
33. Mathieu D (2007) Split charge equilibration method with correct dissociation limits. J Chem Phys 127(22):224103
34. Baekelandt BG, Mortier WJ, Lievens JL, Schoonheydt RA (1991) Probing the reactivity of different sites within a molecule or solid by direct computation of molecular sensitivities via an extension of the electronegativity equalization method. J Am Chem Soc 113(18):6730–6734
35. Svobodová Vařeková R, Jiroušková Z, Vaněk J, Suchomel S, Koča J (2007) Electronegativity equalization method: parameterization and validation for large sets of organic, organohalogene and organometal molecule. Int J Mol Sci 8:572–582
36. Jiroušková Z, Vařeková RS, Vaněk J, Koča J (2009) Electronegativity equalization method: parameterization and validation for organic molecules using the Merz-Kollman-Singh charge distribution scheme. J Comput Chem 30(7):1174–1178
37. Bultinck P, Langenaeker W, Lahorte P, De Proft F, Geerlings P, Van Alsenoy C, Tollenaere JP (2002) The electronegativity equalization method II: applicability of different atomic charge schemes. J Phys Chem A 106(34):7895–7901
38. Ouyang Y, Ye F, Liang Y (2009) A modified electronegativity equalization method for fast and accurate calculation of atomic charges in large biological molecules. Phys Chem Chem Phys 11(29):6082–6089
39. Bultinck P, Vanholme R, Popelier PLA, De Proft F, Geerlings P (2004) High-speed calculation of AIM charges through the electronegativity equalization method. J Phys Chem A 108(46):10359–10366
40. Geidl S, Bouchal T, Raček T, Svobodová Vařeková R, Hejret V, Křenek A, Abagyan R, Koča J (2015) High-quality and universal empirical atomic charges for chemoinformatics applications. J Cheminf 7(1):59
41. O'Boyle N, Banck M, James C, Morley C, Vandermeersch T, Hutchison G (2011) Open babel: an open chemical toolbox. J Cheminf 3(1):33–47
42. Vainio MJ, Johnson MS (2007) Generating conformer ensembles using a multiobjective genetic algorithm. J Chem Inf Model 47(6):2462–2474
43. Svobodová Vařeková R, Koča J (2006) Optimized and parallelized implementation of the electronegativity equalization method and the atom-bond electronegativity equalization method. J Comput Chem 3:396–405
44. Feng Z, Chen L, Maddula H, Akcan O, Oughtred R, Berman HM, Westbrook J (2004) Ligand depot: a data warehouse for ligands bound to macromolecules. Bioinformatics 20(13):2153–2155
45. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M (2008) Drugbank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36(suppl 1):901–906
46. Bolton EE, Wang Y, Thiessen PA, Bryant SH (2008) Pubchem: integrated platform of small molecules and biological activities. Ann Rep Comput Chem 4:217–241
47. MJD Powell (2006) The NEWUOA software for unconstrained optimization without derivatives. In: Large-scale nonlinear optimization, pp. 255–297. Springer, Oxford
48. Anderson E, Bai Z, Bischof C, Blackford S, Demmel J, Dongarra J, Du Croz J, Greenbaum A, Hammarling S, McKenney A, Sorensen D (1999) LAPACK users' guide, 3rd edn. Society for Industrial and Applied Mathematics, Philadelphia
49. Open NCI Database, Release 4. http://cactus.nci.nih.gov/download/nci/
50. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide protein data bank. Nat Struct Mol Biol 10(12):980–980
51. Sadowski J, Gasteiger J (1993) From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. Chem Rev 93:2567–2581

52. MJ Frisch, GW Trucks, HB Schlegel, GE Scuseria, MA Robb, JRCheeseman, JA Montgomery Jr, T Vreven, KN Kudin, JC Burant, JMMillam, SS Iyengar, J Tomasi, V Barone, B Mennucci, M Cossi, GScalmani, N Rega, GA Petersson, H Nakatsuji, M Hada, M Ehara, KToyota, R Fukuda, J Hasegawa, M Ishida, T Nakajima, Y Honda, OKitao, H Nakai, M Klene, X Li, JE Knox, HP Hratchian, JB Cross, VBakken, C Adamo, J Jaramillo, R Gomperts, RE Stratmann, O Yazyev,AJ Austin, R Cammi, C Pomelli, JW Ochterski, PY Ayala, K Morokuma,GA Voth, P Salvador, JJ Dannenberg, VG Zakrzewski, S Dapprich, ADDaniels, MC Strain, O Farkas, DK Malick, AD Rabuck, K Raghavachari,JB Foresman, JV Ortiz, Q Cui, AG Baboul, S Clifford, J Cioslowski,BB Stefanov, G Liu, A Liashenko, P Piskorz, I Komaromi, RL Martin,DJ Fox, T Keith, MA Al-Laham, CY Peng, A Nanayakkara, M Challacombe,PMW Gill, B Johnson, W Chen, MW Wong, C Gonzalez, JA Pople, Gaussian09, Revision E.01. http://www.gaussian.com
53. Jelfs S, Ertl P, Selzer P (2007) Estimation of pka for druglike compounds using semiempirical and information-based descriptors. J Chem Inf Model 47(2):450–459
54. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. J Comput Chem 25(9):1157–1174
55. Bren U, Hodošček M, Koller J (2005) Development and validation of empirical force field parameters for netropsin. J Chem Inf Model 45(6):1546–1552
56. Udommaneethanakit T, Rungrotmongkol T, Bren U, Frecer V, Stanislav M (2009) Dynamic behavior of Avian Influenza A Virus Neuraminidase Subtype H5N1 in Complex with Oseltamivir, Zanamivir, Peramivir, and Their Phosphonate Analogues. J Chem Inf Model 49(10):2323–2332
57. Ison J, Rapacki K, Ménager H, Kalaš M, Rydza E, Chmura P, Anthon C, Beard N, Berka K, Bolser D, Booth T, Bretaudeau A, Brezovsky J, Casadio R, Cesareni G, Coppens F, Cornell M, Cuccuru G, Davidsen K, Vedova GD, Dogan T, Doppelt-Azeroual O, Emery L, Gasteiger E, Gatter T, Goldberg T, Grosjean M, Grüning B, Helmer-Citterich M, Ienasescu H, Ioannidis V, Jespersen MC, Jimenez R, Juty N, Juvan P, Koch M, Laibe C, Li J-W, Licata L, Mareuil F, Mičetić I, Friborg RM, Moretti S, Morris C, Möller S, Nenadic A, Peterson H, Profiti G, Rice P, Romano P, Roncaglia P, Saidi R, Schafferhans A, Schwämmle V, Smith C, Sperotto MM, Stockinger H, Vařeková RS, Tosatto SCE, de la Torre V, Uva P, Via A, Yachdav G, Zambelli F, Vriend G, Rost B, Parkinson H, Løngreen P, Brunak S (2016) Tools and data services registry: a community effort to document bioinformatics resources. Nucleic Acids Res 44(D1):D38–D47

Appendix: All My Publications

Curriculum Vitae