MASARYK UNIVERSITY

Faculty of Informatics

Algorithms for calculating atomic charges

PhD Thesis Proposal

Brno 2016 Tomáš Raček

Declaration

I declare that I have worked on this thesis independently using only the sources listed in the bibliography. All resources, sources, and literature, which I used in preparing or I drew on them, I quote in the thesis properly with stating the full reference to the source.

Tomáš Raček

I certify that this text is adequate as a thesis proposal and I agree with it.

prof. RNDr. Luděk Matyska, CSc., advisor
Brno, January 2016

Acknowledgement

I would like to express my gratitude to my advisor prof. Luděk Matyska for his insight and useful comments to my work. I would also like to thank dr. Radka Svobodová Vařeková and dr. Aleš Křenek for their guidance and a large amount of time they devoted to me. Last but not least, my deepest thanks belongs to my family for their invaluable support.

Computational resources were provided by the MetaCentrum under the program LM2010005 and the CERIT-SC under the program Centre CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, Reg. no. CZ.1.05/3.2.00/08.0144.

Contents

1 Introduction **1**

2 State of the art **3**
2.1 Partial atomic charges and their applications 3
2.2 Methods of computing partial atomic charges 4
2.3 Overview of empirical methods 5
2.4 Implementations of empirical methods 8
2.5 Parameterization of empirical methods 9

3 Proposed framework **14**
3.1 Partial atomic charges framework 14
3.2 Common concepts in empirical methods 15
3.3 Domain specific language for empirical methods description . . . 15
3.4 Parameterization of empirical methods in DSL 16
3.5 Partial atomic charges database 18

4 Usage of the framework **19**

5 Plan of work **21**

6 List of publications **22**

Bibliography **23**

A Journal of Cheminformatics article **29**

Chapter 1

Introduction

Partial atomic charges are important quantities often used in classical sciences (physical chemistry and organic chemistry) and now even in modern life sciences (cheminformatics, biomolecular chemistry, computational chemistry or bioinformatics). Partial atomic charges are real numbers describing the distribution of electron density in the molecule. Specifically, these numbers indicate the amount of electrons the specific atom is able to draw towards itself, or on the contrary, to what degree it is able to hand over its own electrons to the neighbor atoms.

Despite the usefulness of the atomic charges, their obtaining poses a great challenge since their actual values can not be directly measured. However, they can be determined by the computation. In last few decades, there have been many methods developed for computing atomic charges. These methods are based on various physicochemical grounds and in general each provides slightly different atomic charge values. On the other hand, there are at least some similar trends in charges computed using different methods. Since we are unable to measure atomic charges, it's hard to say which method is actually the best. In practice, for some applications the most suitable charges are derived from one method, for another by using a different one.

Another problem related to the computation of atomic charges stems from the fact that the most accurate methods (i.e. quantum mechanical, QM) are computationally demanding. This inconvenience limits their use only to small systems, their application on large molecules is out of the question. Markedly faster alternatives to QM approaches are represented by the empirical methods. Again, there is a diversity of these available, although some share common ideas. However, improved time complexity comes at a price – empirical methods utilize empirical parameters which often have to be obtained from QM charges.

The variety of methods, different parameterization schemes and challenges related to the computation of atomic charges on large molecules cause that the area is difficult to grasp for life scientists. That's why it might be beneficial to look at those problems from another perspective – from a computer scientist's point of view.

For these reasons I focused on empirical charge calculation methods in my PhD thesis. My goal is to present a framework for notation, development, evaluation and parameterization of empirical methods. The first component of the framework will be the domain specific language (DSL) for empirical methods notation. The advantage of the language is the ability to modify existing methods easily or even to develop new ones. The core of the framework will provide all necessary functions for the methods, along with the DSL compiler which will enable the direct use of the methods in DSL, i.e. computation of atomic charges. For the parameterization of the methods, the framework will include several optimization approaches to be able to supply the user with high-quality parameters. Last but not least, the framework will be connected to a newly created database of atomic charges which will serve as a provider of QM charges required for the parameterizations. The database will be used also independently of the framework as a publicly accessible source of atomic charges.

Thesis proposal structure

This thesis proposal is structured in the following way: Chapter 2 presents the state of the art in the field – from introduction of the concept of the partial atomic charges and their applications to methods used for their computation and finally the parameterization of such methods. In Chapter 3, the partial atomic charges framework is proposed as a unified package for empirical methods – their design, parameterization and direct use. Chapter 4 describes several use cases in which the framework can be beneficial for the users. Chapter 5 briefly summarizes the overall study plan. List of publications is in Chapter 6. Bibliography follows. Finally, Appendix A contains a copy of an article published in the Journal of Cheminformatics.

Chapter 2

State of the art

In this chapter we describe what partial charges are and why they are important. Then we focus on description of the most common approaches for their calculation and present the state of the art results achieved.

2.1 Partial atomic charges and their applications

Partial atomic charges (or atomic charges) are the real numbers describing the distribution of electron density inside the molecule [1]. They are result of different electronegativities (atomic property which describes the tendency of an atom to attract electrons) of the constituent atoms.

For example, simple molecule such as *hydrogen fluoride* has two atoms – H and F. Fluorine is the most electronegative element in the periodic table with electronegativity $\chi_F = 3.98$, whereas for hydrogen $\chi_H = 2.20$. The difference in electronegativities causes that the fluorine atom attracts the electron of a hydrogen closer to itself thus creating a partial negative charge δ^- on the fluorine and partial positive charge δ^+ (of same magnitude but opposite sign) on the hydrogen (Figure 2.1).

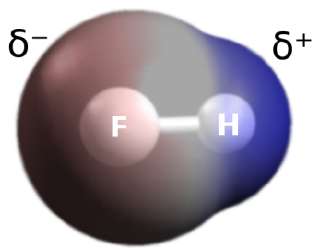

Figure 2.1: Partial atomic charges in hydrogen fluoride molecule

Because of their nature, partial atomic charges can't be physically observed or measured – they are just a theoretical concept. Nevertheless, there are methods which enable us to compute them. As follows from the previous sentences,

whether the results of such computations are physically and chemically meaningful can't be verified directly. However, many molecular properties which can be measured can be derived from atomic charges and therefore used indirectly for the verification. Example of such property is pK_a – dissociation constant which is the measure of the strength of acid in the solution [2, 3], $\log P$ – the ratio of concentrations of a compound in a mixture of two immiscible phases at equilibrium [4] or various cheminformatics charge descriptors [5].

The concept of atomic charges began to be used in physical chemistry and organic chemistry. Afterwards, partial atomic charges were adopted by computational chemistry and molecular modeling, 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. Recently, partial atomic charges also became popular in cheminformatics, as they proved to be informative descriptors for QSAR and QSPR modeling [3, 6, 7] and for other applications [8]; they can be utilized in pharmacophore design [9], virtual screening [10], similarity searches [11], molecular structure comparison [12] etc.

2.2 Methods of computing partial atomic charges

As mentioned before, partial atomic charges are a theoretical concept and can't be directly measured. However, there are many methods for their computation. The standard approaches are based on the principles of quantum mechanics (QM). However, in comparison to those methods, there exist faster yet accurate alternatives represented by empirical methods.

2.2.1 Quantum mechanical approaches

Quantum mechanical methods are the most common and accurate way of computing partial atomic charges. They are based on solving the Schröedinger equation [13]:

$$H\Psi = E\Psi \tag{2.1}$$

where H is the Hamiltonian of the system, Ψ is the wave function and E denotes the total energy. Equation 2.1 is analytically solvable only for single hydrogen atom, for more complex systems numerical methods have to be used. There are several approaches which are used, from simple Hartree-Fock (HF) [14] method to more sophisticated methods like B3LYP [15], Møller–Plesset perturbation theory methods (MPn) [16] and others. Molecular orbitals are approximated by linear combination of functions denoted as a basis set. There are hundreds of different basis sets, e.g. STO-3G which approximates the orbital by com-

bination of 3 primitive Gaussian functions. Last step is to perform the population analysis which partitions the electron density into the actual atomic charges. The examples are the Mulliken population analysis (MPA) [17], Natural population analysis (NPA) [18], Atoms In Molecules (AIM) [19], Merz-Singh-Kollman (MK) [20] and others.

The type of the QM charges is therefore denoted as a triplet – method/basis set/population analysis (e.g. HF/STO-3G/MPA or B3LYP/6-311G/NPA).

The biggest disadvantage of QM methods is computational complexity – $\mathcal{O}(B^4)$ or worse, where B is comparable to or higher than number of electrons in the system. This limits their use only for rather small molecules. Application on large molecules with more than 1000 atoms (e.g. protein complexes) is therefore out of the question.

2.2.2 Empirical approaches

A markedly faster alternative to the QM methods is to employ empirical charge calculation approaches which can also provide high-quality charges comparable to the QM ones. These approaches are based on equations derived from physicochemical laws but they simplify these laws by utilization of empirical constants (parameters). These parameters are calculated from known physicochemical quantities, computed from QM charges etc.

2.3 Overview of empirical methods

This sections presents an overview of several most common empirical methods for charge computation. These methods are widely used and they form a basis for many other methods built upon them.

The empirical methods are often divided into two groups: conformationally-independent and conformationally-dependent methods. The first category takes into the account only the topology of the molecule (information about constituent atoms and bonds), whereas the latter one is concerned also with the positions of the individual atoms.

2.3.1 Gasteiger Marsili

In 1978 Johann Gasteiger and Mario Marsili presented a paper describing a simple schema for calculating atomic charges [21, 22]. This conformationally-independent method is based on an iterative scheme.

In each iteration of the computation the charge between two bonded atoms with different electronegativities is shifted from the more electropositive atom to the more electronegative one. Then electronegativities, as functions of charges, are recalculated according to Equation 2.2 in which q_i is the charge on atom i, χ_i is the electronegativity of atom i and A_i, B_i and C_i are empirical parameters.

$$\chi_i = A_i + B_i q_i + C_i q_i^2 \tag{2.2}$$

Since the charge transfer forms an electrostatic field, the subsequent charge shift is more difficult. To cover this phenomenon, the method introduces a damping factor (set to 2 in original paper) which in each iteration effectively lowers the charge shifted. The convergence is therefore fast – according to authors, six iterations are sufficient. The method is described by the Equation 2.3.

$$q_i = \sum_{\alpha} \left\{ \left[\sum_{j} \frac{1}{D_i}(\chi_j - \chi_i) + \sum_{k} \frac{1}{D_k}(\chi_k - \chi_i) \right] \cdot \left(\frac{1}{f}\right)^{\alpha-1} \right\} \tag{2.3}$$

The meaning of the additional symbols is the following: α is number of iteration, f is a damping factor, D_i and D_k are simple functions of empirical constants (ionization potential and electron affinity); χ_i, χ_j and χ_k are electronegativities of atoms i, j and k where atom j is more electronegative than i and k is less electronegative than i. Note that all atoms denoted j and k are bonded to atom i.

The time complexity can be expressed as $\mathcal{O}(\alpha(N+M))$ where α is the number of iterations, N is number of atoms and M is number of bonds. However, since for practical considerations α is only a small integer and $M \sim N$ as molecules are sparse graphs, we can view the complexity as linear with respect to the number of atoms, i.e. $\mathcal{O}(N)$.

2.3.2 Electronegativity Equalization Method

According to the Sanderson's equalization principle [23] upon formation of the molecule, the individual atomic electronegativities χ_i gets equalized to $\overline{\chi}$ (Equation 2.4).

$$\overline{\chi} = \chi_1 = \ldots = \chi_N \tag{2.4}$$

This phenomenon is the base for the Electronegativity Equalization Method (EEM) invented by Mortier et al. in [24]. The defining formula is the following:

$$\chi_i = \chi_i^* + 2\eta_i^* q_i + \sum_{j \neq i} \frac{q_j}{R_{i,j}} \tag{2.5}$$

where χ_i is the electronegativity of atom i; q_i, q_j are the charges on atom i and j, respectively; χ_i^*, η_i^* are the effective electronegativity and hardness and are

the method's empirical parameters; $R_{i,j}$ is a internuclear distance between atoms i and j. Since these distances are required, the method is conformationally-dependent. The charge conservation principle (stated in Equation 2.6) also applies (Q is the total molecular charge).

$$Q = \sum_{i}^{N} q_i \tag{2.6}$$

Equations 2.4, 2.5 and 2.6 can be rewritten into the form of $N+1$ linear equations (see 2.7) and solved accordingly. The time complexity is therefore $\mathcal{O}(N^3)$ where N is number of atoms in the molecule.

$$\begin{bmatrix} 2\eta_1^* & R_{1,2}^{-1} & \cdots & R_{1,N}^{-1} & 1 \\ R_{2,1}^{-1} & 2\eta_2^* & \cdots & R_{2,N}^{-1} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ R_{N,1}^{-1} & R_{N,2}^{-1} & \cdots & 2\eta_N^* & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \\ -\overline{\chi} \end{bmatrix} = \begin{bmatrix} -\chi_1^* \\ -\chi_2^* \\ \vdots \\ -\chi_N^* \\ Q \end{bmatrix} \tag{2.7}$$

Since EEM is meant to reproduce chosen QM charges, the obtaining of the parameters A_i, B_i is essential and poses a major challenge in using the method.

Note that there are other methods directly based on EEM like Selfconstistent Functional Kernel Equalized Electronegativity Method (SFKEEM) [25] which replaces the ${R_{i,j}}^{-1}$ term in Equation 2.5 by more sophisticated function; or Atom-Bond Electronegativity Equalization Method (ABEEM) [26] which distributes the charge also on the bonds or its variant General atom-bond electronegativity equalization method (general ABEEM) [27] or ABEEM $\sigma\pi$ model [28].

2.3.3 Charge Equilibration

Another method following the Sanderson's principle is the one of Rappé and Goddard – Charge Equilibration (QEq) [29]. The structure of the defining formula is similar to EEM (Equation 2.8).

$$\overline{\chi} = \chi_i = \chi_i^0 + J_{i,i}^0 q_i + \sum_{i \neq j} q_j J_{i,j} \tag{2.8}$$

Apart from the symbols used in the previous methods, which have the same meaning, the new terms are $J_{i,j}$ – Coulombic interaction between unit charges on atoms i and j (which depends on their interatomic distance) and $J_{i,i}^0$ – Coulomb repulsion. These two terms represents the empirical parameters needed to be found.

The final system of linear equations resembles the one from EEM, however due to a unique nature of hydrogen the common formula for $J^0_{i,i}$ and χ^0_i gives chemically incorrect results. Thus those terms for hydrogen atoms are introduced as charge-dependent functions rendering QEq an iterative method. Usually six to ten iterations are sufficient to achieve consistency of hydrogen atom charges.

In terms of computational complexity, QEq can be viewed as the same as EEM, i.e. $\mathcal{O}(N^3)$, however in practice, QEq is more time demanding as it requires multiple iterations to be used.

Again, there are other methods built upon the original QEq, e.g. Consistent Charge Equilibration (CQEq) [30], An Extended Charge Equilibration Method (EQeq) [31], An Empirical Bond-Order-Corrected Extended Charge Equilibration Method (EQeq+C) [32] or Dynamic Charge Equilibration (DQEq) [33].

2.3.4 Other methods

Of course, the presented overview of methods is far from being a complete list of existing empirical charge calculation approaches. For example, there are other methods (although not that widely used) which adopt ideas from the theory of electrical circuits like Kirchhoff Charge Model (KCM) [34], Molecular Graph Charge model (MGC) or Orbital Graph Charge model (OGC) [35,36]; Split Charge Equilibration (SQE) [37] as an improvement on EEM and others.

2.4 Implementations of empirical methods

This section briefly summarizes available implementations of the methods mentioned in the previous section. The overview is presented in Figure 2.2 with additional comments below.

Method	Available implementations
EEM	EEM Solver [38], AtomicChargeCalculator [39], Open Babel [40], Balloon [37]
Gasteiger Marsili	RDKit [41], Open Babel [40], CDK [42]
QEq	Open Babel [40]

Figure 2.2: Overview of common empirical methods implementations

EEM has various implementations – the most sophisticated one is the recently published web service AtomicChargeCalculator. Balloon can provide charges comparable to B3LYP/cc-pVTZ/MPA QM charges, however no other parameter set can be used. Open Babel has integrated parameters taken from [43] (reproducing B3LYP/6-31G*/MPA charges), but substitutes parameters for

elements not covered with some default values without notifying the user. EEM Solver is a simple application which allows user to specify his own parameter set.

Gasteiger Marsili charges can be computed using Open Babel and RDKit or CDK, toolkit libraries. All the implementations have parameters for the most common elements hard-coded in the source code.

QEq method can be found in Open Babel, however, its implementation differs from the original description in Rappé's paper [29]. The $J_{i,j}$ terms (see Equation 2.8 for reference) for hydrogen atoms are not charge-dependent (as the charge influence term is always set to zero), the method is no longer iterative.

The implementation of QEq is also used internally in some molecular dynamics simulations, e.g. LAMMPS Molecular Dynamics Simulator [44] or Gaussian [45].

2.5 Parameterization of empirical methods

Empirical methods focused on imitating QM charges require set of parameters to be able to perform actual charge computation. These parameters are obtained by process called parameterization in which the objective is to find some combination of parameters using which we can achieve the best correspondence between empirical and reference QM charges.

Each of these methods uses several parameters in their defining equation (e.g. EEM has 2, ABEEM has 4) – however, they are actually different for each atom (or bond) type. Atom types can be based solely on element, or element and atom's highest bond order, element and atom's valence state, etc.

The quality of parameters is evaluated by statistical descriptors, the most commonly used in this area are: Pearson's correlation coefficient (R) or its square (R^2), root-mean-square deviation (RMSD) and average absolute difference (D_{avg}). These descriptors are defined as follows:

$$R(q^{qm}, q^{emp}) = \frac{\sum_{i=1}^{N}(q_i^{qm} - \overline{q^{qm}})(q_i^{emp} - \overline{q^{emp}})}{\sqrt{\sum_{i=1}^{N}(q_i^{qm} - \overline{q^{qm}})^2}\sqrt{\sum_{i=1}^{N}(q_i^{emp} - \overline{q^{emp}})^2}} \tag{2.9}$$

$$D_{avg}(q^{qm}, q^{emp}) = \frac{1}{N}\sum_{i=1}^{N}|q_i^{qm} - q_i^{emp}| \tag{2.10}$$

$$RMSD(q^{qm}, q^{emp}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(q_i^{qm} - q_i^{emp})^2} \tag{2.11}$$

where q_i^{qm} and q_i^{emp} are charges on atom i, the reference QM and the empirical one, respectively; N is a number of atoms in the molecule and $\overline{a}$ denotes an arithmetic mean of quantity a.

The parameterizations utilize two sets of molecules – the training set and the test (or validation) set. The first one is used for the calculation of parameters, so it has to be representative enough in the terms of the atom types selected. The latter one usually contains molecules not included in the training set and it is used for the validation of the calculated parameters. Selection of these sets is of crucial importance for the whole parameterization. However, we will not discuss it here further since it is a problem for chemists rather than computer scientists.

In general, the parameterization process consists of several phases following this outline:

1. Selection of atom types to be parameterized
2. Preparation of the training set
3. Calculation of QM charges for the training set
4. Calculation of empirical parameters
5. Validation of empirical parameters

Note that the actual computation of the parameters is an optimization problem. Therefore, the goal is to find a set of parameters p_i which minimizes the value of the chosen objective function. This function is often expressed in the terms of the statistical descriptors defined above.

The basic procedure for the calculation of the parameters can be summarized by the pseudocode in Figure 2.3.

```
divide the atoms into groups based on defined atom types;
while ¬ψ(...) do
    calculate parameters p_i according to the parameterization method;
    calculate empirical charges using p_i;
    calculate statistical descriptors between QM and empirical charges;
end
```

Figure 2.3: The pseudocode for the calculation of parameters

The ψ function determines when to end the computation. Commonly used criteria are: maximum number of iterations reached, time limit reached or a case when no significant increase in quality can be found.

2.5.1 Methods used for parameterization

There are several optimization approaches used for the parameterization of empirical methods. In this section, we describe the ones that are most widely-used in this area – linear regression, genetic algorithms and Nelder-Mead simplex method. It should be mentioned that there exist more advanced methods than for example the one of Nelder and Mead, however, the papers on parameterizations use mainly these three mentioned.

Genetic algorithms

The genetic algorithm (GA) is an optimization and search technique based on the principles of genetics and natural selection [50].

The objective is to maintain a set of candidate solutions (chromosomes) and in each generation to produce new offspring by combining the most promising candidates (crossover). However, if this was the only admissible change, the new generation would be limited by the properties of the parents. To mitigate the issue, the perturbations on offspring (mutations) are performed in a random fashion. The general outline of the genetic algorithm is depicted in Figure 2.4.

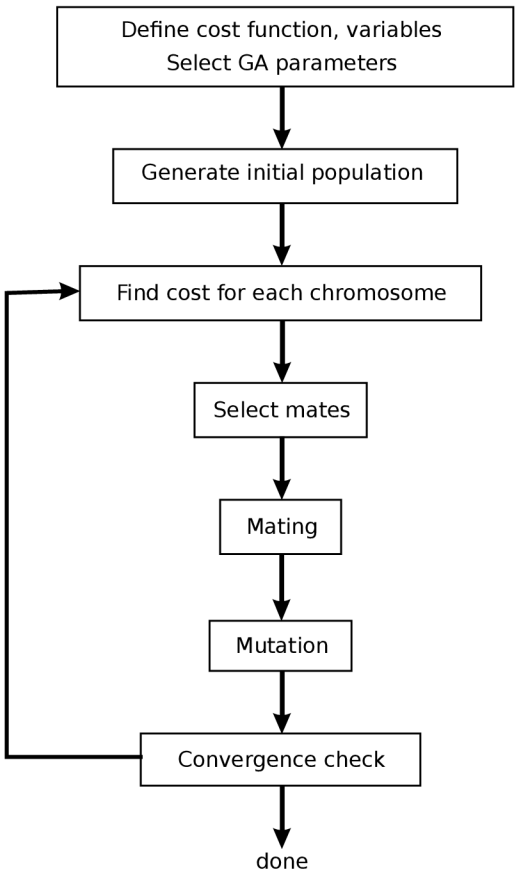

Figure 2.4: Outline of genetic algorithm adapted from [50].

The big advantage of the genetic algorithms is the ability to cope with the local minima, however, there is an increased cost stemming from the need of maintaining (possibly) large population.

Genetic algorithms were successfully used for parameterization of empirical methods, for example in [25, 51–53].

Linear regression

Linear regression is an approach for modeling the relationship between a response variable and a set of exploratory variables [46]. For the empirical charge computation methods like EEM, only one independent variable is used. Therefore, the goal is to find the parameters α and β for which the sum of squared differences is the smallest (Equation 2.12). Note that these two parameters are for single atom type, so that the complete parameter set contains a pair for each of them.

$$\min_{\alpha,\beta}\left(\sum_{i=1}^{N}(q_i^{qm}-(\alpha q_i^{emp}+\beta))^2\right) \tag{2.12}$$

Examples of the use of the linear regression for parameterization can be found in [24, 27, 47–49].

Simplex method

Nelder-Mead downhill simplex method is an optimization approach introducing a *simplex*, a geometrical figure with $N+1$ vertices in N dimensions [54].

The idea is to move the simplex in steps in the multidimensional space *downhill* until it encounters a minimum. The individual moves are represented as simplex transformations (Figure 2.5). Detailed description along with reference implementation can be found in [55].

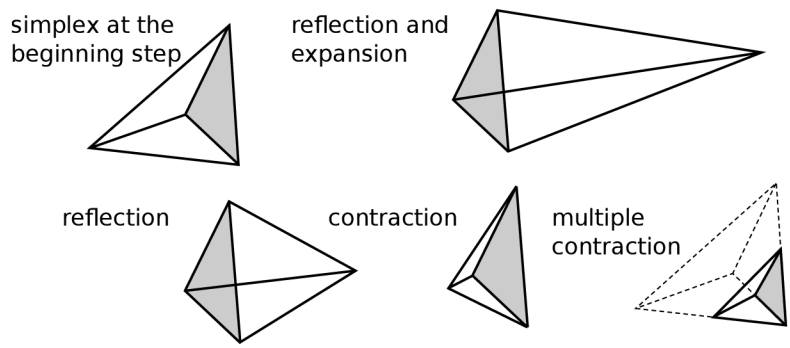

Figure 2.5: Transformations of the simplex [55].

The disadvantage of the simplex method is the need of the restart in case when the simplex reaches a local minimum. The method is therefore often used in conjunction with the genetic algorithm for refining the results, e.g. [25, 51–53].

2.5.2 Parameterization of EEM

Nowadays, as the EEM is one of the most widely-used empirical charge calculation methods, there has been a significant effort to provide parameters which are both universal (covering the most important elements in organic molecules) and accurate when compared to the reference QM charges.

The first EEM parameterization was performed by Mortier in his original paper where the EEM was introduced [24]. The parameters were calculated to reproduce HF/STO-3G/MPA charges using the least squares method. The training set included propane, ethene, propene, water, methanol, dimethyl ether, methylamine, dimethylamine, formaldehyde, formic acid and formamide, therefore only molecules containing H, C, N, O were covered by the parameters.

Since that time, many other authors published their own parameter sets. There are several goals to achieve – including more elements for better parameter coverage, using more sophisticated QM methods and basis sets and of course, maintaining the high quality of resulting EEM charges in comparison to QM ones.

These objectives pose non-trivial requirements on the parameterization process – from the selection of the representative training set, high QM calculation demands to choice of the suitable parameterization method.

Although there are available parameter sets published [48,53] which can produce high-quality empirical charges, they have limited coverage of elements [56] which reduces their practical applicability. Appendix A includes Table 1 with summarizing view of the parameter sets published.

One specific domain of EEM parameterization is the parameterization for biomacromolecules. The biomacromolecules are large molecules built from specific substructures and comprising of thousands of atoms. For example, proteins are systems made up of chains formed of amino acids; nucleic acids are created by repeating different nucleotides.

The great challenge in EEM parameterization for biomacromolecules is the fact that we are unable to obtain reference QM charges for them (because large system size is a limiting factor for QM calculations). Therefore, their parameterization is performed using fragments of the biomacromolecules along with the associated QM charges.

Another issue stems from the homogeneity of the structures, since they are composed of the same building blocks, they include chemically similar atoms. That's why QM charges on these atoms are similar as well and the training charge sets lack the diversity needed to perform high-quality parameterization.

Even though, there has been some acceptable EEM parameters published in [53,57].

Chapter 3

Proposed framework

This chapter forms an outline of the thesis goals – first, to devise a reasonable way for empirical methods notation, then their translation into the executive code used for the actual charge calculation, and finally to include a parameterization interface for methods which are meant to mimic quantum mechanical charges. These charges will be taken from the database of charges which will be developed as a part of the thesis. The database, however, will find use on its own since it will serve as publicly accessible source of partial charges for chemical structures. All the functionality will be provided by one framework, presumably a web service.

3.1 Partial atomic charges framework

As was stated before, there is no universal method of computing atomic charges that would be appropriate for all the applications. And the situation is similar also for the parameterizations of empirical charge calculation methods. Therefore, it's often required to choose different method and/or parameterization scheme for a certain class of molecules. To mitigate these issues, we propose the idea of the atomic charges framework.

The framework will use a domain specific language (DSL) for the empirical methods notation and it will provide a possibility to translate methods written in DSL into the executable code which can be used directly to compute atomic charges. Thus, the DSL will serve as a executable specification for the empirical charge calculation methods. Next sections discuss the reasoning behind the concept of DSL and its use within the framework.

The main objective is to translate the methods into the charge calculation modules for direct use within the framework. However, there are other possibilities which can be implemented as extensions – creation of standalone charge computation applications or libraries with defined API which could be included in user applications, e.g. molecular dynamics simulations. Chapter 4 describes several more use cases of the whole framework.

3.2 Common concepts in empirical methods

As could be seen from the previous chapter, the empirical methods can be expressed by using rather simple formulas, often in a form similar to $q_i = f(\ldots)$.

Besides, many methods share ideas, or are even built atop of each other like the EEM or QEq derivatives. For example, Oliferenko showed [58] that methods based on electronegativity equalization can be written in a concise matrix form (Equation 3.1).

$$\begin{bmatrix} \mathbf{H} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{q} \\ -\chi \end{bmatrix} = \begin{bmatrix} -\boldsymbol{\chi}^* \\ Q \end{bmatrix} \tag{3.1}$$

The $\mathbf{H}$ represents the hardness matrix, $\mathbf{1}$ is a vector of ones, $\mathbf{1}^T$ is its transpose, $\mathbf{q}$ is a vector of unknown charges, χ is the overall equalized electronegativity, $\boldsymbol{\chi}^*$ is the vector of the effective electronegativities and finally Q stands for the total molecular charge.

Matrix $\mathbf{H}$ then differentiates between the individual methods, for example, for QEq it has a following form:

$$\mathbf{H}_{QEq} = \begin{bmatrix} J^0_{1,1} & J_{1,2} & \cdots & J_{1,N} \\ J_{2,1} & J^0_{2,2} & \cdots & J_{2,N} \\ \vdots & \ddots & \ddots & \vdots \\ J_{N,1} & J_{N,2} & \cdots & J^0_{N,N} \end{bmatrix} \tag{3.2}$$

Importantly, the empirical methods do not include any complex mathematical operations (in contrast to the QM approaches) and they all use a common "language" – as they are expressed in the terms of electronegativity, hardness, interatomic distances, empirical parameters etc.

3.3 Domain specific language for empirical methods description

The above-mentioned similarities motivate the concept of the domain specific language for describing empirical methods. The goal is to provide a concise notation that would be easy to understand and to use. The language will be similar to LaTeX math in syntax since that's the familiar form for the potential users – computational chemists, physical chemists and others.

The general idea is to express the method's defining equation in DSL (in some cases along with other formulas describing terms used in the main one) in a form that resembles the original notation. In addition to that, the equation(s) will be accompanied by the formal annotations of the symbols used. To illustrate

the concept on example different from the aforementioned QEq, the sketch of the idea for Gasteiger Marsili method follows:

1. $\forall a \in M : \chi_a \leftarrow A_a + B_a q_a + C_a q_a^2$
2. $\forall a \in M : q_a \leftarrow q_a + \left(\sum_j \left(\frac{1}{D_a} (\chi_j - \chi_a) \right) + \sum_k \left(\frac{1}{D_k} (\chi_k - \chi_a) \right) \right) \cdot \left(\frac{1}{2} \right)^{\alpha - 1}$

- α – number of iteration
- j – atoms bonded to atom a with $\chi_a < \chi_j$
- k – atoms bonded to atom a with $\chi_a > \chi_k$
- D_a, D_k – empirical per atom type constants
- A_a, B_a, C_a – empirical per atom type constants

This annotation-based scheme is meant to be similar to the descriptions used in the scientific papers but here is presented in a formal way.

Apart from changing the formulas for the method itself, an interesting feature is the possibility to change the annotations as well to explore the modifications of the original method. As an example, consider changing the restriction on bonded atoms to the close atoms (R has its usual meaning – distance between atoms, 1 Å $= 10^{-10}$ m):

- atoms bonded to atom $a \rightarrow$ atoms with $R < 3$Å

Through the annotations the language should also include the option of specifying exceptions to the formulas. Hydrogen atoms, for example, sometimes require a special treatment – when a general rule does not apply, authors may come up with a suitable constant serving as a replacement.

Pragmatically, the goal of the language is not to express every possible notation, that would not be a manageable task to accomplish. Rather than that, it will first focus on the generalized description of the most commonly used methods mentioned in the previous chapter. Then, new classes of methods might be added.

3.4 Parameterization of empirical methods in DSL

With methods formulated in a common formalism, the framework can include a general interface allowing their parameterization against reference QM charges. The modular approach proposed makes the independent development of the parameterization part of the framework possible.

3.4.1 Parameterization of EEM

The framework should provide a generic way for parameterization of empirical methods, however, in this area we are not starting from scratch. In the last several years we were extensively engaged in parameterization of EEM, one of the most commonly used empirical methods. Specifically, we developed a software tool NEEMP which main function is the EEM parameterization. Apart from that it can be used for EEM charge calculation or validation of the EEM parameters.

The software extends the parameterization methodology presented in [48] using EEM in the form depicted at Equation 3.3.

$$\overline{\chi} = \chi_i = A_i + B_i q_i + \kappa \sum_{i \neq j} \frac{q_j}{R_{i,j}} \tag{3.3}$$

The κ is an independent variable in the parameterization process – NEEMP utilizes the Brent's optimization method [55] to find the best value in a given range.

The above-mentioned paper also suggests that choosing a suitable subset of the original training set can substantially increase quality of the resulting parameters. However, since there are 2^M possible subsets (as M stands for the number of molecules in the training set), the exhaustive search is definitely out of the question. For this purpose, NEEMP uses stochastic local search methods [59] to alleviate the problem. The software is also fully parallelized and uses high-performance mathematical libraries, the details of the implementation can be found in the Master's thesis [60].

NEEMP is not yet published (paper is in preparation, Jan 2016), however, its functionality was already evaluated. In December 2015, the most comprehensive parameterization of EEM on drug-like molecules was published in *Journal of Cheminformatics*. Utilizing NEEMP we were able to provide high-quality and universal parameters for EEM, the best available up to date.

In particular, we provided six sets of parameters for the most common elements in the drug-like molecules (H, C, N, O, S, F, Cl, Br, I, P with different multiplicities), all achieving great correlations with reference QM charges ($R^2 > 0.9$) on both training and validation set. The details can be found in the paper referenced in this work as Appendix A.

The part of the article was a prototype patch for Open Babel [40] which integrated our results into the popular open-source cheminformatics software. In January 2016, updated version of the patch was accepted to a development version of this project and it will be a part of the next release.

3.5 Partial atomic charges database

The parameterizations are meant to produce parameters which plugged in the empirical method reproduce the QM charges the best. However, obtaining these charges is a time-demanding process and as they can be used multiple times, the creation of a database connected to the framework seems to be a reasonable task.

The database can be utilized on its own though – to provide atomic charges for structures via a publicly-accessible interface, presumably a web service. Initially, the structures may come from the well-known databases of chemical structures, e.g. DrugBank [61] or Ligand Expo [62]. Later on, the charges and the structures in those databases can be linked against each other for user convenience.

Chapter 4

Usage of the framework

While the previous chapter presented the idea of the partial atomic charges framework and its proposed internals, the purpose from the user perspective has to be stated as well. This chapter describes several use cases which might be interesting for potential users.

1. Obtaining the atomic charges for a specific compound

In this most simple scenario the user just wants to get the atomic charges for a particular molecule (or set of molecules). If the database contains the molecule along with the required type of charges, he can download them instantly. If not, he can request their calculation by specifying a structure and a method to be used.

2. Obtaining a binary for the method

To be able to compute atomic charges locally, the framework can provide a binary (library or standalone application) compiled from a certain method in DSL. User can use the binary directly to compute charges on his computer or embed it in some molecular dynamics simulations.

3. Parameterization of a particular method

Third use case describes the situation in which the user is concerned with a specific charge calculation method and wants to find the optimal set of parameters for it. He selects a method, type of QM charges and two groups of molecules – one used for training the method, the other for validation of the parameters.

Again, QM charges can be provided by the charge database (if they are not present yet, then the parameterization request can cause their computation) or they can be uploaded by the user locally.

As the output, the framework provides user the best parameters found along with statistical descriptors assessing their quality.

4. Exploring the modifications of the method

Since the recent empirical methods are often based on their predecessors, the ability to modify the methods easily may come in useful. The user won't have to address the issues related to programming the method on his own, rather he may explore the modifications in the DSL.

If the method is meant to reproduce QM charges then the parameterization step as covered in the previous section will follow.

5. Selecting a suitable method for a given set of molecules

Finally, one is sometimes interested not in method itself, but his aim is to obtain high-quality charges for a certain class of molecules regardless of the method used for their computation. Apart from the *parameter sweep* used for a parameterization of a specific method, the framework could perform *algorithm sweep* to select the most appropriate method for the chosen QM charges.

Chapter 5

Plan of work

This chapter briefly summarizes the supposed schedule of the goals toward the PhD thesis completion.

Spring 2016: Paper on NEEMP (*Journal of Cheminformatics*). Design of DSL. State doctorate exam.

Fall 2016: Charge database, framework core implementation.

Spring 2017: Paper on the charge database (potential journals: *Nucleic Acid Research (database issue), PLOS Computational biology, Journal of Computational Chemistry, Journal of Cheminformatics*). DSL compiler.

Fall 2017: Parameterization interface of the framework.

Spring 2018: Publication on the framework (potential journals: *Journal of Chemical Information and Modeling, Journal of Computational Chemistry, Journal of Cheminformatics*). Writing the thesis.

Fall 2018: PhD thesis submission.

Chapter 6

List of publications

- Stanislav Geidl, Tomáš Bouchal, **Tomáš Raček**, Radka Svobodová Vařeková, Václav Hejret, Aleš Křenek, Ruben Abagyan and Jaroslav Koča. *High-quality and universal empirical atomic charges for chemoinformatics applications*. Journal of Cheminformatics, London: BioMed Central Ltd, 2015. ISSN 1758-2946. doi:10.1186/s13321-015-0107-1.
  - Shared first authorship.
  - I updated and extended NEEMP software to suit the needs of the article, participated in the EEM parameterization and EEM charges validation. I also wrote the Open Babel patch.
- **Tomáš Raček**, Radka Svobodová Vařeková, Aleš Křenek, Tomáš Bouchal and Jaroslav Koča. *NEEMP – Tool for parameterization of empirical charge calculation method EEM*. In Chemoinformatics Strasbourg Summer School 2014. 2014. [poster]
  - I developed a prototype version of NEEMP software.
- Tomáš Bouchal, Radka Svobodová Vařeková, **Tomáš Raček**, Crina-Maria Ionescu, Stanislav Geidl, Aleš Křenek and Jaroslav Koča. *Empirical charges for chemoinformatics applications*. In GCC 2013 - 9th German Conference on Chemoinformatics. 2013. doi:10.1186/1758-2946-6-S1-P60. [poster]
  - Same as above. I also participated in the parameterization process.
- **Tomáš Raček**, Radka Svobodová Vařeková and Aleš Křenek. *Advanced implementation of atomic charge calculation with electronegativity equalization*. In SVOČ 2013 v Opavě, 2013. [conference talk]
  - Same as above.

Bibliography

[1] V. Gupta, *Principles and Applications of Quantum Chemistry*. Academic Press, 2015.

[2] J. Zhang, T. Kleinöder, and J. Gasteiger, "Prediction of pKa values for aliphatic carboxylic acids and alcohols with empirical atomic charge descriptors," *Journal of chemical information and modeling*, vol. 46, no. 6, pp. 2256–2266, 2006.

[3] R. S. Vařeková, S. Geidl, C.-M. Ionescu, O. Skřehota, T. Bouchal, D. Sehnal, R. Abagyan, and J. Koča, "Predicting pKa values from EEM atomic charges.," *J. Cheminformatics*, vol. 5, p. 18, 2013.

[4] D. Eros, I. Kovesdi, L. Orfi, K. Takacs-Novak, G. Acsády, and G. Kéri, "Reliability of logP predictions based on calculated molecular descriptors: a critical review," *Current medicinal chemistry*, vol. 9, no. 20, pp. 1819–1829, 2002.

[5] R. Todeschini and V. Consonni, *Handbook of molecular descriptors*, vol. 11. John Wiley & Sons, 2008.

[6] R. S. Vařeková, S. Geidl, C.-M. Ionescu, O. Skřehota, M. Kudera, D. Sehnal, T. Bouchal, R. Abagyan, H. J. Huber, and J. Koča, "Predicting pKa values of substituted phenols from atomic charges: Comparison of different quantum mechanical methods and charge distribution schemes.," *Journal of chemical information and modeling*, vol. 51, no. 8, pp. 1795–1806, 2011.

[7] S. Geidl, R. S. Vařeková, V. Bendová, L. Petrusek, C.-M. Ionescu, Z. Jurka, R. Abagyan, and J. Koča, "How Does the Methodology of 3D Structure Preparation Influence the Quality of pKa Prediction?," *Journal of Chemical Information and Modeling*, vol. 55, no. 6, pp. 1088–1097, 2015.

[8] J. Galvez, R. Garcia, M. Salabert, and R. Soler, "Charge indexes. New topological descriptors," *Journal of chemical information and computer sciences*, vol. 34, no. 3, pp. 520–525, 1994.

[9] C. G. Wermuth, "Pharmacophores: historical perspective and viewpoint from a medicinal chemist," *Methods and Principles in Medicinal Chemistry*, vol. 32, p. 3, 2006.

[10] P. D. Lyne, "Structure-based virtual screening: an overview," *Drug discovery today*, vol. 7, no. 20, pp. 1047–1055, 2002.

[11] S. K. Kearsley, S. Sallamack, E. M. Fluder, J. D. Andose, R. T. Mosley, and R. P. Sheridan, "Chemical similarity using physiochemical property descriptors," *Journal of Chemical Information and Computer Sciences*, vol. 36, no. 1, pp. 118–127, 1996.

[12] C. Lemmen, T. Lengauer, and G. Klebe, "FLEXS: a method for fast flexible ligand superposition," *Journal of medicinal chemistry*, vol. 41, no. 23, pp. 4502–4520, 1998.

[13] A. R. Leach, *Molecular modelling: principles and applications*. Pearson education, 2001.

[14] J. C. Slater, "A simplification of the hartree-fock method," *Physical Review*, vol. 81, no. 3, p. 385, 1951.

[15] A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange," *The Journal of chemical physics*, vol. 98, no. 7, pp. 5648–5652, 1993.

[16] C. Møller and M. S. Plesset, "Note on an approximation treatment for many-electron systems," *Physical Review*, vol. 46, no. 7, p. 618, 1934.

[17] R. S. Mulliken, "Electronic population analysis on LCAO–MO molecular wave functions. I," *The Journal of Chemical Physics*, vol. 23, no. 10, pp. 1833–1840, 1955.

[18] A. E. Reed, R. B. Weinstock, and F. Weinhold, "Natural population analysis," *The Journal of Chemical Physics*, vol. 83, no. 2, pp. 735–746, 1985.

[19] R. F. Bader, "Atoms in molecules," *Accounts of Chemical Research*, vol. 18, no. 1, pp. 9–15, 1985.

[20] U. C. Singh and P. A. Kollman, "An approach to computing electrostatic charges for molecules," *Journal of Computational Chemistry*, vol. 5, no. 2, pp. 129–145, 1984.

[21] J. Gasteiger and M. Marsili, "A new model for calculating atomic charges in molecules," *Tetrahedron Letters*, vol. 19, no. 34, pp. 3181–3184, 1978.

[22] J. Gasteiger and M. Marsili, "Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges," *Tetrahedron*, vol. 36, no. 22, pp. 3219–3228, 1980.

[23] R. Sanderson, *Chemical Bonds and Bonds Energy*, vol. 21. Elsevier, 2012.

[24] W. J. Mortier, S. K. Ghosh, and S. Shankar, "Electronegativity-equalization method for the calculation of atomic charges in molecules," *Journal of the American Chemical Society*, vol. 108, no. 15, pp. 4315–4320, 1986.

[25] J. Chaves, J. M. Barroso, P. Bultinck, and R. Carbó-Dorca, "Toward an alternative hardness kernel matrix structure in the electronegativity equalization method (EEM)," *Journal of chemical information and modeling*, vol. 46, no. 4, pp. 1657–1665, 2006.

[26] Z.-Z. Yang and C.-S. Wang, "Atom-bond electronegativity equalization method. 1. Calculation of the charge distribution in large molecules," *The Journal of Physical Chemistry A*, vol. 101, no. 35, pp. 6315–6321, 1997.

[27] Y. Cong and Z.-Z. Yang, "General atom-bond electronegativity equalization method and its application in prediction of charge distributions in polypeptide," *Chemical Physics Letters*, vol. 316, no. 3, pp. 324–329, 2000.

[28] Y. Cong, Z.-Z. Yang, C.-S. Wang, X.-C. Liu, and X.-H. Bao, "Investigation of the regio-and stereoselectivity of diels–alder reactions by newly developed ABEEM $\sigma\pi$ model on the basis of local HSAB principle and maximum hardness principle," *Chemical physics letters*, vol. 357, no. 1, pp. 59–64, 2002.

[29] A. K. Rappe and W. A. Goddard III, "Charge equilibration for molecular dynamics simulations," *The Journal of Physical Chemistry*, vol. 95, no. 8, pp. 3358–3363, 1991.

[30] T. Ogawa, N. Kurita, H. Sekino, O. Kitao, and S. Tanaka, "Consistent charge equilibration (CQEq) method: application to amino acids and crambin protein," *Chemical physics letters*, vol. 397, no. 4, pp. 382–387, 2004.

[31] C. E. Wilmer, K. C. Kim, and R. Q. Snurr, "An extended charge equilibration method," *The Journal of Physical Chemistry Letters*, vol. 3, no. 17, pp. 2506–2511, 2012.

[32] G. C. Martin-Noble, D. Reilley, L. M. Rivas, M. D. Smith, and J. Schrier, "EQeq+ C: An empirical bond-order-corrected extended charge equilibration method," *Journal of chemical theory and computation*, vol. 11, no. 7, pp. 3364–3374, 2015.

[33] J. Sefcik, E. Demiralp, T. Cagin, and W. A. Goddard, "Dynamic charge equilibration-morse stretch force field: Application to energetics of pure silica zeolites," *Journal of computational chemistry*, vol. 23, no. 16, pp. 1507–1514, 2002.

[34] N. Zefirov, M. Kirpichenok, F. Izmailov, and M. Trofimov, "A new calculational scheme for electronegativity equalization according to Sanderson's principle," in *Dokl. Akad. Nauk*, vol. 304, pp. 887–891, 1989.

[35] A. Oliferenko, V. Palyulin, A. Neiman, and N. Zefirov, "A new topological model for the calculation of partial atomic charges," in *Doklady Chemistry*, vol. 375, pp. 281–284, Springer, 2000.

[36] A. A. Oliferenko, V. A. Palyulin, S. A. Pisarev, A. V. Neiman, and N. S. Zefirov, "Novel point charge models: reliable instruments formolecular electrostatic," *Journal of Physical Organic Chemistry*, vol. 14, no. 6, pp. 355–369, 2001.

[37] M. J. Vainio and M. S. Johnson, "Generating conformer ensembles using a multiobjective genetic algorithm," *Journal of chemical information and modeling*, vol. 47, no. 6, pp. 2462–2474, 2007.

[38] R. S. Vařeková and J. Koča, "Optimized and parallelized implementation of the electronegativity equalization method and the atom-bond electronegativity equalization method," *Journal of computational chemistry*, vol. 27, no. 3, pp. 396–405, 2006.

[39] C.-M. Ionescu, D. Sehnal, F. L. Falginella, P. Pant, L. Pravda, T. Bouchal, R. Svobodová Vařeková, S. Geidl, and J. Koča, "AtomicChargeCalculator: interactive web-based calculation of atomic charges in large biomolecular complexes and drug-like molecules.," *Journal of cheminformatics*, vol. 7, p. 50, jan 2015.

[40] N. M. OLBoyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," *J Cheminf*, vol. 3, p. 33, 2011.

[41] G. Landrum, "RDKit: Open-source cheminformatics," *Online). http://www. rdkit. org. Accessed*, vol. 3, no. 04, p. 2012, 2006.

[42] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen, "The chemistry development kit (cdk): An open-source java library for chemo-and bioinformatics," *Journal of chemical information and computer sciences*, vol. 43, no. 2, pp. 493–500, 2003.

[43] P. Bultinck, W. Langenaeker, P. Lahorte, F. De Proft, P. Geerlings, C. Van Alsenoy, and J. Tollenaere, "The electronegativity equalization method II: Applicability of different atomic charge schemes," *The Journal of Physical Chemistry A*, vol. 106, no. 34, pp. 7895–7901, 2002.

[44] S. Plimpton, "Fast parallel algorithms for short-range molecular dynamics," *Journal of computational physics*, vol. 117, no. 1, pp. 1–19, 1995.

[45] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, . Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, "Gaussian 09 Revision E.01." Gaussian Inc. Wallingford CT 2009.

[46] L. Ott, M. Longnecker, and R. L. Ott, *An introduction to statistical methods and data analysis*, vol. 511. Duxbury Pacific Grove, CA, 2001.

[47] C.-S. Wang, L. Shen-Min, and Z.-Z. Yang, "Calculation of molecular energies by atom-bond electronegativity equalization method," *Journal of Molecular Structure: THEOCHEM*, vol. 430, pp. 191–199, 1998.

[48] R. S. Vařeková, Z. Jiroušková, J. Vaněk, . Suchomel, and J. Koča, "Electronegativity equalization method: Parameterization and validation for large sets of organic, organohalogene and organometal molecule," *International Journal of Molecular Sciences*, vol. 8, no. 7, pp. 572–582, 2007.

[49] Z. Jiroušková, R. S. Vařeková, J. Vaněk, and J. Koča, "Electronegativity equalization method: parameterization and validation for organic molecules using the Merz-Kollman-Singh charge distribution scheme.," *Journal of computational chemistry*, vol. 30, no. 7, pp. 1174–1178, 2009.

[50] R. L. Haupt and S. E. Haupt, *Practical genetic algorithms*. John Wiley & Sons, 2004.

[51] P. Bultinck, W. Langenaeker, P. Lahorte, F. De Proft, P. Geerlings, M. Waroquier, and J. Tollenaere, "The electronegativity equalization method i: Parametrization and validation for atomic charge calculations," *The Journal of Physical Chemistry A*, vol. 106, no. 34, pp. 7887–7894, 2002.

[52] G. Menegon, K. Shimizu, J. Farah, L. Dias, and H. Chaimovich, "Parameterization of the electronegativity equalization method based on the charge model 1," *Physical Chemistry Chemical Physics*, vol. 4, no. 24, pp. 5933–5936, 2002.

[53] Y. Ouyang, F. Ye, and Y. Liang, "A modified electronegativity equalization method for fast and accurate calculation of atomic charges in large biological molecules," *Physical Chemistry Chemical Physics*, vol. 11, no. 29, pp. 6082–6089, 2009.

[54] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The computer journal*, vol. 7, no. 4, pp. 308–313, 1965.

[55] W. H. Press, *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[56] S. Geidl, T. Bouchal, T. Raček, R. S. Vařeková, V. Hejret, A. Křenek, R. Abagyan, and J. Koča, "High-quality and universal empirical atomic charges for chemoinformatics applications," *Journal of cheminformatics*, vol. 7, no. 1, pp. 1–10, 2015.

[57] C.-M. Ionescu, S. Geidl, R. Svobodová Vařeková, and J. Koča, "Rapid calculation of accurate atomic charges for proteins via the electronegativity equalization method," *Journal of chemical information and modeling*, vol. 53, no. 10, pp. 2548–2558, 2013.

[58] A. A. Oliferenko, S. A. Pisarev, V. A. Palyulin, and N. S. Zefirov, "Atomic charges via electronegativity equalization: Generalizations and perspectives," *Advances in Quantum Chemistry*, vol. 51, pp. 139–156, 2006.

[59] H. H. Hoos and T. Stützle, *Stochastic local search: Foundations & applications*. Elsevier, 2004.

[60] T. Raček, "Advanced implementation of atomic charge calculation with electronegativity equalization [online]," Master's thesis, Masaryk University, Faculty of Informatics, Brno, 2013 [cit. 2016-01-13].

[61] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "Drugbank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D668–D672, 2006.

[62] Z. Feng, L. Chen, H. Maddula, O. Akcan, R. Oughtred, H. M. Berman, and J. Westbrook, "Ligand depot: a data warehouse for ligands bound to macromolecules," *Bioinformatics*, vol. 20, no. 13, pp. 2153–2155, 2004.

Appendix A

Journal of Cheminformatics article

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

Chemistry Central

© 2015 Geidl et al. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

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).

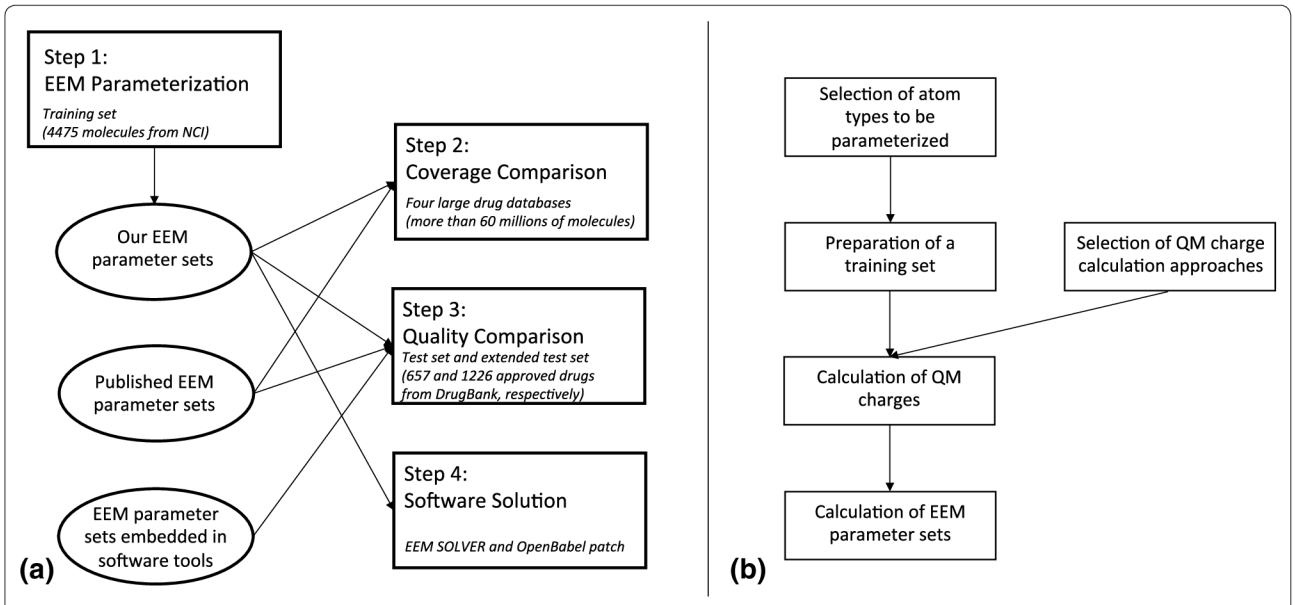

Fig. 1 **a** Composition of steps performed within this work and **b** tasks performed during EEM parametrization

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 ($\overline{\Delta}$), calculated via Eqs. (2), (3) and (4), respectively

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

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

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

where q_i^{EEM} is the EEM charge of an atom i; q_i^{QM} is the QM charge of an atom i; $\overline{q}^{EEM}$ is an average of all EEM charges; $\overline{q}^{QM}$ 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, Jirouškova 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. Jirouškova 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