Title

High-quality and universal empirical atomic charges for chemoinformatics applications

Journal

Journal of Cheminformatics, 7(1)

ISSN

1758-2946

Authors

Geidl, Stanislav

Bouchal, Tomáš

Raček, Tomáš

et al.

Publication Date

2015-12-01

DOI

10.1186/s13321-015-0107-1

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

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

High-quality and universal empirical atomic charges for chemoinformatics applications

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

Abstract

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

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

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

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

Background

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Methods

EEM parameterization (step 1)

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

EEM parameterization: selection of atom types to be parameterized

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

EEM parameterization: preparation of the training set

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

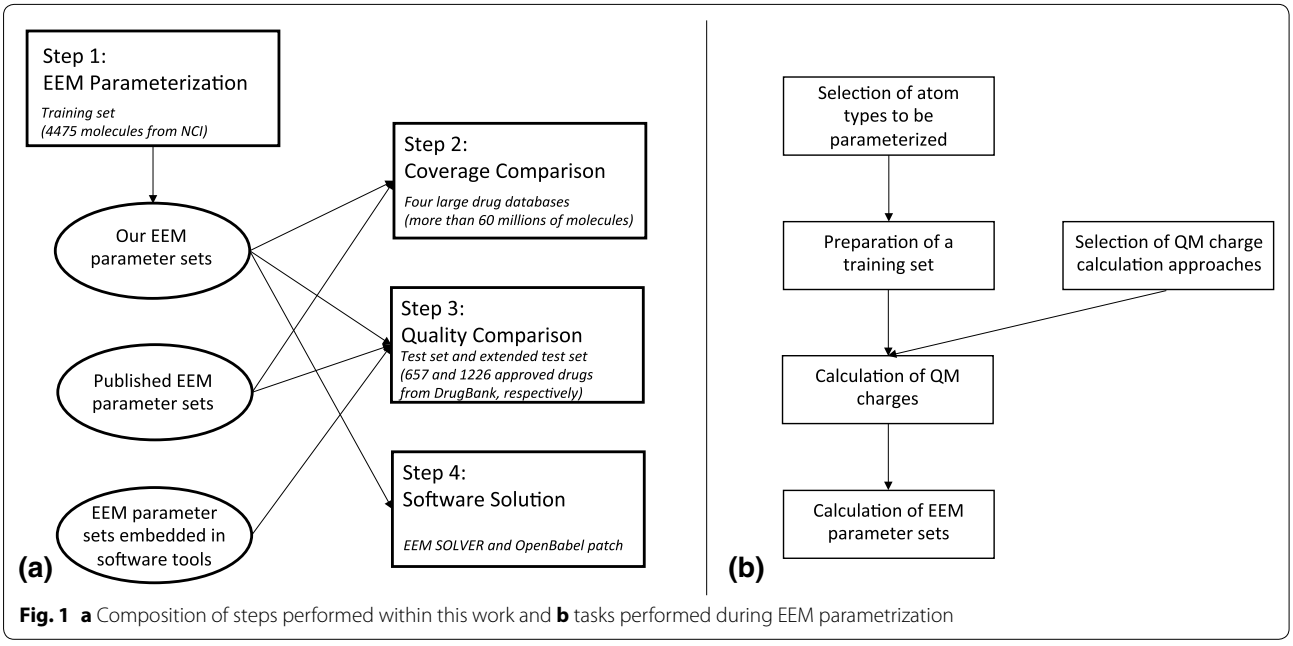

Fig. 1 **a** Composition of steps performed within this work and **b** tasks performed during EEM parametrization

Table 3 Occurrence of atom types in the training set

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

EEM parameterization: selection of QM charge calculation approach

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

Table 4 Quality criteria of our EEM parameter sets

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

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

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

Quality comparison: EEM parameter sets embedded in software tools

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

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

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

Coverage comparison and quality comparison combined

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

Software solution (step 4)

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

Conclusion

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

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

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

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

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

Additional files

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

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

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

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

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

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

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

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

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

Authors' contributions

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

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

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

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

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

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

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

References

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

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