Abstract

The electronegativity equalization method is classically used as a method allowing the fast generation of atomic charges using a set of calibrated parameters and provided knowledge of the molecular structure. Recently, it has started being used for the calculation of other reactivity descriptors and for the development of polarizable and reactive force fields. For such applications, it is of interest to know whether the method, through the inclusion of the molecular geometry in the Taylor expansion of the energy, would also allow sufficiently accurate predictions of spectroscopic data. In this work, relevant quantities for IR spectroscopy are considered, namely the dipole derivatives and the Cartesian Hessian. Despite careful calibration of parameters for this specific task, it is shown that the current models yield insufficiently accurate results.
1. Introduction

Electronegativity equalization is a pervasive concept in chemistry that is fundamental to our understanding of molecular charge distributions [1, 2]. One of its modern incarnations is the Electronegativity Equalization Method (EEM), which allows a rapid computation of \textit{ab initio quality} atomic charges from just the molecular geometry and a set of atomic electronegativity and hardness parameters [3, 4, 5]. Over the past 27 years, EEM strongly influenced related areas, such as in-silico screening [6], chemical reactivity descriptors [7, 8, 9] and empirical models for molecular potential energy surfaces, hereafter referred to as (polarizable and possibly reactive) force fields [10, 11, 12, 13, 14, 15, 16]. Extensive calibration studies have demonstrated the accuracy of atomic charges and the parameter transferability of EEM [17, 18, 19, 20, 21, 22, 14, 23] and its recent generalizations, the Split-Charge Equilibration (SQE) [15, 24, 25, 26] and Atom-Condensed Kohn-Sham approximated to Second order (ACKS2) [16].

EEM is essentially a semi-empirical form of density functional theory (DFT) [27, 28], in which the molecular electron density is modeled with a minimal set of variables (the atomic charges) and the energy of a system with a specific molecular geometry is expanded in a second order Taylor series of the charges. From that perspective, one may be tempted to derive other quantities from EEM beyond just the atomic charges of the ground state [7, 8, 9]. In this work, we explore the ability of EEM to reproduce key
quantities in the field of infrared spectroscopy, namely the Cartesian Hessian (i.e. force constants or second order derivative of the energy toward nuclear displacements) and the molecular dipole derivatives (first order derivative of the molecular dipole moment toward nuclear displacements). Such insights are essential for the development of force-field models and their applications in which IR spectra are derived from molecular dynamics simulations. It should be noted that such attempts have been made before by Mortier and co-workers [29, 30] although limited to mainly vibrations involving bonded atoms. Here the ambition is to go beyond this and consider the entire vibrational structure. For comparison, similar assessments for traditional fixed-charge models are also considered.

The paper is structured as follows. Analytic expressions for the key spectroscopic quantities are derived in section 2. Section 3 describes all computational aspects related to the model assessment, i.e. the test set of molecular reference computations, the cost functions used to calibrate EEM parameters and the calibration algorithm. All results are presented in section 4 and discussed in section 5. Finally, section 6 summarizes the main conclusions.

2. Key Spectroscopic Quantities in EEM

In this section, we start from the EEM energy expression. Justifications and a detailed description of the underlying physics can be found elsewhere [4, 12, 31, 16]. The molecular energy for a system with \( N \) atoms is approximated to second order in the atomic charges,

\[
E(\{q\}) = \chi^\dagger q + \frac{1}{2} q^\dagger \eta q, \tag{1}
\]
where $\mathbf{q}$ is a column vector with $N$ atomic charges, $\mathbf{\chi}$ is a column vector with $N$ atomic electronegativity parameters and $\mathbf{\eta}$ is a symmetric $N$ by $N$ matrix containing all second order coefficients, known as the hardness matrix. The diagonal elements of $\mathbf{\eta}$ are atomic hardness parameters, while the off-diagonal elements correspond to electrostatic interactions between the atoms.

In line with previous work, the electrostatics are based on Gaussian charge distributions, whose widths are derived from covalent radii [24, 25]. Note that our atomic hardness parameters do not include the factor $1/2$ of the Taylor series as in Parr’s operational definition of the atomic hardness [32].

The EEM *ground state* charges are found by minimizing this energy with a constraint on the total charge $q_{\text{mol}}$,

$$E_{\text{gs}} = E(\{q\}_{\text{gs}}) = \min_{\{q\}, d^\top \mathbf{q} = q_{\text{mol}}} E(\{q\}),$$

where $d$ is a column vector filled with $N$ times the number 1. This ground state is defined as the stationary point of the following Lagrangian:

$$L = E - \chi_{\text{mol}}(d^\top \mathbf{q} - q_{\text{mol}}).$$

To facilitate the remainder of the derivations, we introduce the vector with *uncounstrained* ground state charges:

$$\tilde{\mathbf{q}}_{\text{gs}} = -\mathbf{\eta}^{-1} \mathbf{\chi}$$

Using this notation, the stationary point of this Lagrangian takes the following form:

$$\chi_{\text{mol}} = \frac{q_{\text{mol}} - d^\top \tilde{\mathbf{q}}_{\text{gs}}}{d^\top \mathbf{\eta}^{-1} d} \quad \mathbf{q}_{\text{gs}} = \tilde{\mathbf{q}}_{\text{gs}} + \chi_{\text{mol}} \mathbf{\eta}^{-1} d$$

4
We will now treat the derivatives of the energy and the charges towards nuclear displacements. For the sake of compactness, nuclear Cartesian coordinates are labeled with Greek letters \( \alpha \) and \( \beta \) instead of explicitly writing \( R_{ij} \). (At most two are needed as we consider at most second order derivatives. Note that \( R_{ij} \) is component \( j \) of the Cartesian coordinates of atom \( i \) and not an interatomic distance between two atoms \( i \) and \( j \).) In EEM, the electronegativity parameters are geometry-independent; such derivatives are not considered below. The only geometry-dependence in the model is found in the off-diagonal elements of \( \eta \).

The forces acting on the atoms consist of first-order derivatives of the energy of the following form:

\[
\frac{\partial E_{gs}}{\partial \alpha} = \left( \chi^\top + q_{gs}^\top \eta \right) \frac{\partial q_{gs}}{\partial \alpha} + \frac{1}{2} q_{gs}^\top \frac{\partial \eta}{\partial \alpha} q_{gs} \tag{6}
\]

In analogy with the Helmann-Feynman theorem in quantum mechanics, the first term is zero because the charges are obtained from a variational principle. However, in the derivation of the force constants, the response of the charges does appear. Therefore, let us first derive the charge response and subsequently consider the force constants. The response of the charges with respect to a change in geometry can be found after careful application of the chain rule to Eq. (5):

\[
\frac{\partial q_{gs}}{\partial \alpha} = \frac{\partial q_{gs}}{\partial \alpha} + \frac{\partial \chi_{\text{mol}}^{-1}}{\partial \alpha} \eta d + \chi_{\text{mol}}^{-1} \frac{\partial \eta^{-1}}{\partial \alpha} d \tag{7}
\]
with
\[
\frac{\partial \eta^{-1}}{\partial \alpha} = -\eta^{-1} \frac{\partial \eta}{\partial \alpha}, \tag{8}
\]
\[
\frac{\partial \mathbf{q}_{gs}}{\partial \alpha} = -\eta^{-1} \frac{\partial \eta}{\partial \alpha} \mathbf{q}_{gs}, \tag{9}
\]
\[
\frac{\partial \chi_{mol}}{\partial \alpha} = \frac{q_{mol} - \mathbf{d}^T \mathbf{q}_{gs}(\mathbf{d}^T \frac{\partial \eta^{-1}}{\partial \alpha} \mathbf{d})}{(\mathbf{d}^T \eta^{-1} \mathbf{d})^2} - \frac{1}{\mathbf{d}^T \eta^{-1} \mathbf{d}} \frac{\partial \mathbf{d}^T \mathbf{q}_{gs}}{\partial \alpha}. \tag{10}
\]

The Cartesian Hessian can be found through the application of the chain rule to Eq. (6):
\[
\frac{\partial^2 E_{gs}}{\partial \alpha \partial \beta} = \frac{\partial \mathbf{q}_{gs}^T}{\partial \beta} \frac{\partial \eta}{\partial \alpha} \mathbf{q}_{gs} + \frac{1}{2} \mathbf{q}_{gs}^T \frac{\partial^2 \eta}{\partial \alpha \partial \beta} \mathbf{q}_{gs}. \tag{11}
\]

In EEM, the molecule is modeled as a superposition of spherical atomic charge distributions. Hence, the Cartesian components of the molecular dipole moment simply become
\[
D_j = \sum_{i=1}^{N} R_{ij} q_{gs,i}, \tag{12}
\]
where \(q_{gs,i}\) is the ground-state charge of atom \(i\). The dipole derivatives are obtained by differentiating \(D_j\) once with respect to a nuclear displacement:
\[
\frac{\partial D_j}{\partial \alpha} = \sum_{i=1}^{N} \delta_{\alpha,R_{ij}} q_i + R_{ij} \frac{\partial q_{gs,i}}{\partial \alpha}. \tag{13}
\]

In this paper, the EEM model will also be compared to a model with fixed atomic charges. Such a comparison is relevant as many force-field models treat the atomic charges as geometry-independent parameters. The above results for the Cartesian Hessian and the dipole derivatives are still valid for fixed-charge models, except that the terms containing derivatives of the charges can be dropped.
3. Assessment protocol

To test the accuracy of EEM for the key spectroscopic quantities of interest, DFT reference computations were carried out for a set of 166 organic drug-like molecules. [19, 20, 22] The molecular geometries were optimized at the PBE0/6-311+G(d,p) level of theory with Gaussian09 [33], followed by computation of the dipole moment, the Cartesian Hessian and the dipole derivatives, all at the same level of theory. Finally, the Hirshfeld-I atomic charges [34] are computed with Horton [35]. Previous studies have shown that Hirshfeld-I and Mulliken charges are suitable reference data for the calibration of EEM parameters. [22, 23, 24] However, in order to properly describe negatively charged moieties, one needs to include diffuse functions in the basis set, which leads to ill-defined Mulliken charges. In contrast, Hirshfeld-I charges are known to be robust with respect to the choice of basis set and are therefore used as reference data in this work. [36] A text file with all relevant results of the reference computations is provided as Supplementary data.

Five different sets of EEM parameters were calibrated, each to reproduce one class of the DFT reference data:

- **AC**: Hirshfeld-I atomic charges.
- **MD**: Molecular dipole moment.
- **MDD**: Molecular dipole derivatives.
- **FH**: Full Hessian.
• **LRH**: Long-range Hessian, i.e. only considering those matrix elements of Cartesian Hessian that correspond to pairs of atoms that are at least 5Å apart. At such larger distances, exchange-correlation effects and higher order terms should become negligible in DFT. One may expect that the EEM approximation, which is based classical electrostatics, becomes more reliable in this limit.

Each calibration was carried out by minimizing a corresponding least-squares cost function of the following form:

$$C_Z(p) = \sum_{m=1}^{M} \sum_{n=1}^{N_{z,m}} (x_{EEM,n}(p) - x_{DFT,n})^2$$

where $Z$ refers to the class of reference data (AC, MD, MDD, FH, LRH), $p$ is the vector of independent parameters. The first sum runs over all molecules ($M$ in total) and the second sum runs over all corresponding pairs of EEM predictions and reference data in class $Z$ for molecule $m$. Each term is a squared error between a reference result and the corresponding EEM prediction. The vector of parameters consists of atomic electronegativity and hardness parameters for each element found in the set of molecules (H, C, N, O, F and Cl). Since the electronegativity parameters can only be calibrated up to an unknown constant, the electronegativity of Hydrogen is used as a reference [25]. Hence, there are 11 independent parameters. The optimal parameters are denoted as $p_Z$, where $Z$ can be any of AC, MD, MDD, FH, LRH. Note that only the calibration AC relies on Hirshfeld-I charges while all other calibrations are independent method used to compute the charges. The implementation of all spectroscopic quantities was validated numerically with finite-difference approximations.
The covariance matrix adaption evolutionary strategy (CMA-ES) [37] was used to numerically minimize the five cost functions. Our choice for CMA-ES is motivated by several of its advantages. CMA-ES only requires that cost function values can be computed, i.e. no analytic gradients of the cost function must be implemented. Furthermore, CMA-ES has the robustness and (to some extent) the efficiency of a quasi-Newton method, i.e. it constructs a second order model of the function during minimization, without resorting to numerically instable finite difference approximations. In line with previous work, lower-bounds were imposed on the hardness parameters to guarantee that the hardness matrix remains positive definite [24, 25, 26]. All EEM predictions by the five sets of calibrated parameters are provided as Supplementary data.

To assess EEM’s ability to reproduce the five different quantities listed above, relative root-mean-square errors (RRMSEs) are computed for all combinations of optimal parameters and reference data:

$$RRMSE_{Z}(p) = \sqrt{\frac{C_{Z}(p)}{\sum_{m=1}^{M} \sum_{n=1}^{N_{Z,m}} x_{DFT,n}^2}}$$  \hfill (15)

An RRMSE of 0% corresponds to an exact reproduction of the reference data, while a value of 100% (or larger) indicates that the errors are equal to (or larger than) the reference data. A more detailed statistical analysis (including absolute RMSEs and parity plots) are provided as Supplementary data.

For the cases MD, MDD, FH and LH, the RRMSEs are also computed for a fixed-charge model. In analogy with the EEM approach, a fixed-charge model still describes the molecule as a superposition of spherical Gaussian
functions. However, the charges are now fixed at the Hirshfeld-I values and no geometry dependence of the charges is considered.

4. Results

All calibrations lead to a converged set of parameters shown in Table 1. Only for the calibrations AC and MD, the parameters have a reasonable order of magnitude, e.g. compared to previous calibrations [22, 24] or experimental properties of isolated atoms [32]. The condition number reported in Table 1 is based on the covariance estimate of the CMA-ES algorithm. A high condition number indicates the presence of a manifold of parameter vectors that approximately minimizes the cost function. In such cases, the parameters are not uniquely defined and the optimization landscape reflects many (near) degenerate solutions [25].

Table 2 lists the RRMSE computed for each combination of a cost function and a set of optimal parameters. This reveals how dependent the cost function is on the target function used in the optimization. Note that most results do not depend on the choice of the method for the population analysis. More specifically, all parameter sets in Table 1, except for the column \( p_{AC} \), are independent of the method used to compute the atomic charges.

Only the AC and MD calibrations lead to parameters that are sufficiently accurate for the description of atomic charges or molecular dipoles. This is in line with the previously observed good performance of EEM algorithms for Hirshfeld-I atomic charges [24, 25, 26] and it confirms that the CMA algorithm is an effective method for the calibration of EEM parameters. The AC and MD parameter sets do not give a useful approximation of the dipole
<table>
<thead>
<tr>
<th></th>
<th>$\mathbf{p}_{\text{AC}}$</th>
<th>$\mathbf{p}_{\text{MD}}$</th>
<th>$\mathbf{p}_{\text{MDD}}$</th>
<th>$\mathbf{p}_{\text{FH}}$</th>
<th>$\mathbf{p}_{\text{LRH}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi_C - \chi_H$</td>
<td>0.87</td>
<td>14.46</td>
<td>0.97</td>
<td>143.73</td>
<td>11.64</td>
</tr>
<tr>
<td>$\chi_N - \chi_H$</td>
<td>1.60</td>
<td>12.88</td>
<td>14.71</td>
<td>143.77</td>
<td>31.49</td>
</tr>
<tr>
<td>$\chi_O - \chi_H$</td>
<td>3.31</td>
<td>12.58</td>
<td>431.61</td>
<td>392.51</td>
<td>631.43</td>
</tr>
<tr>
<td>$\chi_F - \chi_H$</td>
<td>4.89</td>
<td>10.65</td>
<td>367.47</td>
<td>320.13</td>
<td>394.14</td>
</tr>
<tr>
<td>$\chi_{Cl} - \chi_H$</td>
<td>1.20</td>
<td>10.47</td>
<td>206.52</td>
<td>233.09</td>
<td>303.12</td>
</tr>
<tr>
<td>$\eta_H$</td>
<td>15.47</td>
<td>37.87</td>
<td>128.79</td>
<td>191.54</td>
<td>182.62</td>
</tr>
<tr>
<td>$\eta_C$</td>
<td>11.40</td>
<td>13.29</td>
<td>16.19</td>
<td>11.71</td>
<td>18.84</td>
</tr>
<tr>
<td>$\eta_N$</td>
<td>11.57</td>
<td>16.56</td>
<td>27.38</td>
<td>11.70</td>
<td>22.12</td>
</tr>
<tr>
<td>$\eta_O$</td>
<td>14.14</td>
<td>15.32</td>
<td>523.84</td>
<td>140.37</td>
<td>685.59</td>
</tr>
<tr>
<td>$\eta_F$</td>
<td>21.61</td>
<td>21.52</td>
<td>515.27</td>
<td>118.61</td>
<td>284.50</td>
</tr>
<tr>
<td>$\eta_{Cl}$</td>
<td>8.93</td>
<td>21.25</td>
<td>444.23</td>
<td>40.09</td>
<td>252.03</td>
</tr>
</tbody>
</table>

Table 1: Calibrated EEM Parameters for the five classes of reference data. All values of the different parameters are given in eV. The condition number is derived from the covariance matrix in the CMA-ES algorithm.
derivatives nor the Hessian, including when one only considers the long-range part of the Hessian. Parameters calibrated with the MDD, FH or LRH cost function fail to predict any property, i.e all RRMSEs are close to or above 100%. Note that, as was to be expected, the diagonal elements in Table 2 are lower than the off-diagonal elements on the same row or column. Again note that all RRMSEs in Table 2, except for the first row and the first column, are independent of the choice of population analysis.

The accuracy of the fixed-charge model for the classes MD, MDD, FH and LRH is shown in Table 3. From these results, it is clear that such a fixed-charge model is only capable of reproducing the molecular dipole moment to some accuracy. The errors on the dipole derivatives and the Cartesian Hessian are close to 100%, also when one only considers the long-range part of the Hessian.

5. Discussion

The diagonal elements of Table 2 show how well the EEM model can reproduce a class of reference data. As extensively demonstrated in the
literature, we observe that it is relatively easy to obtain EEM parameters that reproduce the Hirshfeld-I atomic charges [24, 25, 26]. A similar result is found for the calibration of parameters aimed at reproducing molecular dipole moments, albeit with a larger RRMSE. When considering the two EEM calibrations with reasonable parameters (AC and MD), the error in the reproduction of the dipole moment ($RRMSE_{MD}$ in Table 2) is worse than that of the fixed charge model (see Table 3). This was to be expected: the fixed-charge model contains a huge number of non-transferable parameters, i.e. one charge for every atom in the set, as compared to the 11 transferable EEM parameters. The Hirshfeld-I charges are known to reproduce the dipole moment of organic molecules [38, 39]. However, this attractive feature of Hirshfeld-I can not be fully reproduced by an EEM model based on Hirshfeld-I charges (or any other class of reference data). This corresponds to a known pathology of EEM. Namely, due to the metallic polarizability scaling in EEM, [40] molecular dipole moments are underestimated on average, especially in the limit of larger molecules. [26] This trend is also confirmed in Figure S9 (and to lesser extent in Figure S20) in the Supplementary data, showing that EEM tends to underestimate the molecular dipole moment.

The poor description of the dipole derivatives is somewhat disappointing.

<table>
<thead>
<tr>
<th>$RRMSE_{MD}$</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>$RRMSE_{MDD}$</td>
<td>77</td>
</tr>
<tr>
<td>$RRMSE_{FH}$</td>
<td>99</td>
</tr>
<tr>
<td>$RRMSE_{LRH}$</td>
<td>95</td>
</tr>
</tbody>
</table>

Table 3: RRMSE values [%] for the fixed-charge model.
Some authors previously suggested a good mapping between EEM and the charge-sensitivity analysis (CSA, i.e. a complete set of second order derivatives of the electronic energy towards AIM charges and nuclear coordinates) [41, 29, 30, 42]. This would imply a reasonable reproduction of dipole derivatives with EEM, which we do not observe. Compared to a fixed-charge model (see Table 3), EEM is only a marginal improvement. The parity plots of the dipole derivatives for each model (Figures S10, S21, S32, S43 and S54 in the Supplementary data) show that the large RRMSEs in row 3 of Table 2 are representative, i.e. not caused by a few outliers. These results have two important consequences. First, when EEM or fixed charges are used in the simulation of IR spectra, we expect similarly large errors on peak heights, i.e. IR intensities. Second, the good geometry dependence of atomic charges in EEM [39] is not valid for small displacements of the nuclei. This is reminiscent of the limitations of EEM for the description of the dipole polarizability [40, 15, 43, 24, 16] and suggests that EEM fails to describe any type of electronic linear response. An interesting topic of future research would therefore be the reproduction of dipole derivatives with recent generalizations of EEM that fix the dipole polarizability weaknesses, e.g. SQE [15] or ACKS2 [16].

The poor reproduction of the Hessian is in line with expectations. Several force-field models use the EEM energy as one of the contributions to the total energy [13, 14]. Obviously, the EEM energy alone is not sufficient to describe the curvature of the PES. Surprisingly, this observation still holds when only the long-range parts of the Hessian are considered. Again, the parity plots for the Hessian matrix elements (Figures S11, S12, S22, S23, S33, S34, S44, S45, S55, S56 in the Supplementary data) confirm that the
large RRMSEs in row 4 and 5 of Table 2 are not due to outliers. Although electrostatic interactions are the main long-range interactions in a force-field model, the EEM long-range contributions to the Hessian do not match the DFT reference data. Again, compared to a fixed-charge model (see Table 3), EEM is only a marginal improvement.

The long range Hessian model performs somewhat better than the Full Hessian model which is not unexpected. As EEM entails an energy expansion in terms of atomic charges only and considers only a rather simplified model for the electrostatic interaction, many aspects of the pair-wise interaction, such as exchange effects, are not taken properly into account. Such exchange effects grow smaller with increasing internuclear distance, as do higher order contributions to the electrostatic interaction. This explains why LRH is better than FH. On the other hand, in most CSA work mentioned above, the bonded, and thus shorter interactions where emphasized. Note, however, that Baekelandt et al. [30] used a force field model that corrects for covalent character of the bond.

6. Summary

We have shown that the bare EEM model can not describe (let alone predict) the key molecular properties related to IR spectroscopy: the dipole derivatives and the Cartesian Hessian. The potential improvements of EEM over a simple fixed-charge model are negligible. Although EEM (or a fixed charge model) is often used as one of many contributions in empirical models for the potential energy surface, the poor performance of the method implies that computer simulations of IR spectra require a more advanced model
of the response of atomic charges to a displacement of the nuclei. This complements the well-known limitations of EEM for the description of the dipole polarizability tensor and suggests that EEM fails to reproduce (even qualitatively) any linear response property. As an example of such a more advanced method, the ACKS2 model, which also includes a term for the electronic kinetic energy, may improve the accuracy the dipole derivatives.

Acknowledgments

T.V. is a post-doctoral researcher funded by the Foundation of Scientific Research - Flanders (FWO). Funding was also received from the Research Board of Ghent University (BOF), and BELSPO in the frame of IAP/7/05. P.B. acknowledges the Scientific Research Foundation - Flanders (FWO-Vlaanderen) and the Research Board of Ghent University (BOF) for continuous support.

Appendix A. Supplementary data

Supplementary data reflecting the molecular data set and raw EEM results as well as more in-depth statistical analysis associated with this article can be found in the online version, at http://dx.doi.org/10.1016/XXX.

References


[42] R. Heidler, G. O. A. Janssens, W. J. Mortier, R. A. Schoonheydt, Charge sensitivity analysis of the interaction of pyrrole with basic FAU-type