A critical investigation of proposed electrostatic corrections to quantum mechanical volumes: the importance of variation and the irrelevance of imbalance

Eirik F. Kjønstad, John F. Moxnes, Tomas L. Jensen & Erik Unneberg

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

A critical investigation of proposed electrostatic corrections to quantum mechanical volumes: the importance of variation and the irrelevance of imbalance

Eirik F. Kjønstad, John F. Moxnes, Tomas L. Jensen and Erik Unneberg

Norwegian Defence Research Establishment (FFI), Land Systems Division, Kjeller, Norway

ABSTRACT
The crystal density of neutral and ionic molecular crystals is remarkably well approximated by the enclosed volume of molecular surfaces, where these surfaces are defined as regions of constant and small electron density. Several workers have proposed that estimates may be improved if one includes quantities extracted from the electrostatic potential on the surface of the molecule. The variation of the potential and the imbalance of positive and negative values have been considered to be of importance. In this study we demonstrate that whereas variation is important for improving crystal density predictions, imbalance is not. We present a density functional theory study on a set of 44 neutral molecular crystals. Ten-fold cross-validations were performed on models that incorporate variation, imbalance and combinations of both. Geometries were optimised using B3LYP and basis sets of type 6-31G(d). Electron densities and electrostatic potentials were computed with B3LYP and M05. Regardless of functional, models that correct for variation yield a relative decrease of 15%–18% in root-mean-square error of prediction. This correction appears to sharpen the error distribution about zero. Models based on imbalance yield no improvement, and we argue that it plays an insignificant role.
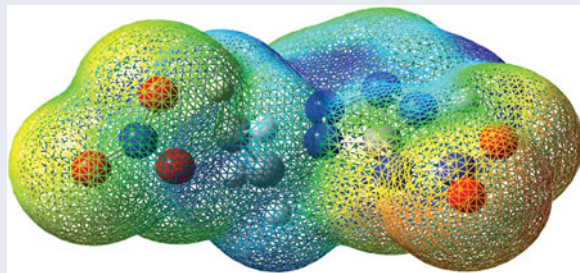
1. Introduction

The density is an important parameter for the detonation performance of a material. To a large extent, it determines the detonation velocity and detonation pressure [1], quantities of great interest in the design and development of energetic materials [2]. Theoretical prediction of the density is an attractive goal, not only because of its importance, but also because it avoids the high costs associated with synthesising and testing novel compounds.

Predicting the density of a crystal has turned out to be very difficult. One possible strategy is to make an attempt to determine the crystal structure, and subsequently deduce the density. Unfortunately, crystal structure prediction remains, in general, an unsolved problem [3,4]. A less ambitious but promising strategy consists of approximating the effective volume a molecule occupies in the crystal. This approach has been remarkably successful in the quantum mechanical setting, where estimates of the effective volume are given by enclosed volumes of surfaces of constant and small electron density [5–8]. Denoting the crystal density by ρ, we can define the molecular volume V by

$$\rho = \frac{M}{V}, \tag{1}$$

where M is the mass of a single molecule. Here V is the effective volume of a molecule in the crystal, implying that it may differ from the intuitive measure of volume as space. In fact, V will be smaller or larger depending on factors affecting the crystal structure, for instance the

molecular shape and intermolecular interactions. To a first approximation, V is simply the size of the molecule, and this can most definitely be estimated by quantum mechanics.

Bader *et al.* [9] proposed that the surface $\mathscr{S}$ of molecules may be represented by regions with a small and constant electron density $n(\mathbf{r})$, defined as the probability density to find an electron at the position $\mathbf{r}$. They suggested to apply surfaces on which $n(\mathbf{r}) = 0.0020$ a.u. or 0.0010 a.u. The latter,

$$\mathscr{S} = \{\mathbf{r} : n(\mathbf{r}) = 0.0010 \text{ a.u.}\}, \qquad (2)$$

has mostly been settled on in the literature [5–8,10–15], although it has been remarked that other surfaces would serve equally well as the molecular surface [16–18]. Rice and coworkers [5] found that V is exceptionally well approximated by the volume V_{encl} enclosed by $\mathscr{S}$. They achieved a root-mean-square error (RMSE) within 4% of experimentally obtained densities for a set of 180 neutral molecular crystals comprised of carbon, hydrogen, nitrogen and oxygen. The molecular surface $\mathscr{S}$ was computed by density functional theory (DFT), applying the B3LYP functional and basis sets of type 6-31G(d,p).

Although no systematic errors were found in the Rice *et al.* [5] study, it is clear that the arbitrary definition of $\mathscr{S}$ as well as the particular choice of functional and basis set might lead to such errors. Politzer and coworkers [6] observed that V_{encl} was systematically underestimated with the B3PW91 functional. Consequently, in order to correct for potential systematic errors, we will add a constant γ to V_{encl}:

$$\widehat{V} = V_{encl} + \gamma. \qquad (3)$$

In the following, we will denote approximations of V by $\widehat{V}$. It is implicitly assumed in Equation (3) that if a systematic error is present, it is independent of the volume. Another option is to scale V_{encl} by some constant α, i.e. $\widehat{V} = \alpha V_{encl}$, making the assumption that the error is proportional to the volume. One may also combine the two in $\widehat{V} = \alpha V_{encl} + \gamma$, but we have chosen to avoid this in order to keep the number of parameters to a minimum. The constant γ is absent in previous studies [6–8], making the comparison to more complicated models unfair. Moreover, these complicated models have, in addition to proposed electrostatic corrections, both a scaling constant α and an overall constant γ. These facts naturally raise the question of whether their effectiveness is due to α and γ, or as reported, electrostatics.

Although the molecular volume V is well approximated by V_{encl}, Politzer and coworkers [6] pointed out that the role of intermolecular electrostatic interactions

are completely ignored. Building on previous work, they proposed that information about such interactions are present in the electrostatic potential $\mathscr{V}(\mathbf{r})$, the electric potential felt by a unit charge positioned at $\mathbf{r}$:

$$\mathscr{V}(\mathbf{r}) = \sum_i \frac{Z_i}{|\mathbf{R}_i - \mathbf{r}|} - \int \frac{n(\widetilde{\mathbf{r}})}{|\widetilde{\mathbf{r}} - \mathbf{r}|} \, d\widetilde{\mathbf{r}}. \qquad (4)$$

Restricting the potential $\mathscr{V}(\mathbf{r})$ to $\mathbf{r}$ on $\mathscr{S}$, a number of studies have been conducted in which statistics from $\mathscr{V}$ assumed to be relevant for predicting some physical or chemical property have been extracted. Such properties include solubilities [15,19], impact sensitivities [13,14], partition coefficients [20], crystal densities [6–8], chemical inhibition [16,18], critical points and boiling points [21], as well as sublimation enthalpies and solvation free energies [17]. The procedure has been conceptually generalised as

$$\text{property} = f(\Phi^{(1)}, \Phi^{(2)}, \ldots, \Phi^{(n)}), \qquad (5)$$

where $\Phi^{(i)}$ are parameters computed from $\mathscr{V}$ restricted to $\mathscr{S}$. Applied parameters $\Phi^{(i)}$ in Equation (5) include the extrema of $\mathscr{V}$, average values, variances and other measures of spread, and areas of negative and positive regions [16]. The functional relationship f is unknown and has been termed a general interaction properties function (GIPF) [11]. While it is our view that modelling based on $\mathscr{V}$ is a sensible approach that should be further explored, we consider much of the research based on Equation (5) to be inconclusive, because the performance of a model has often been equated with the extent to which it fits the data-set, with no consideration of the model's complexity [11–14,16,20–22]. The complexity may be taken into account by considering information criteria, such the Akaike information criterion (AIC) or Bayesian information criterion (BIC) [23], providing more suitable quantities in the comparison of unvalidated models. Models should nevertheless be validated by testing them on unseen data-sets, for instance by test-set/training-set division, cross-validation or bootstrapping [24].

In this paper we will investigate proposed electrostatic corrections to V_{encl} and identify the cause of the observed improvement. As such, the work presented provides a firm foundation on which to formulate more accurate models and to interpret the reported results. We begin by describing the quantities that have been applied as corrections. Brinck and coworkers [25] introduced the 'charge

separation' quantity Π, defined by

$$\Pi = \frac{1}{N} \sum_{i=1}^{N} |\mathscr{V}(\boldsymbol{r}_i) - \overline{\mathscr{V}}|. \qquad (6)$$

A non-zero value of Π has been interpreted as a prerequisite for intermolecular interactions [6]. In this equation, N is the number of points on the discretised surface $\mathscr{S}$, and $\overline{\mathscr{V}}$ is the average value of $\mathscr{V}$ on $\mathscr{S}$. The quantity Π can also be identified as the mean absolute deviation from $\overline{\mathscr{V}}$, a measure of the spread about the mean. A related and more commonly used measure of variation is the standard deviation estimate [26],

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\mathscr{V}(\boldsymbol{r}_i) - \overline{\mathscr{V}})^2}, \qquad (7)$$

a quantity that we will consider as an alternative to Π.[1] Politzer et al. [19] proposed another measure of spread, $\sigma_{\text{tot}}^2 = \sigma_-^2 + \sigma_+^2$, in which negative and positive variances were treated separately:

$$\sigma_+^2 = \frac{1}{N_+} \sum_{\substack{\boldsymbol{r}_i \in \mathscr{S} \\ \mathscr{V}(\boldsymbol{r}_i) > 0}} (\mathscr{V}(\boldsymbol{r}_i) - \overline{\mathscr{V}}_+)^2,$$

$$\sigma_-^2 = \frac{1}{N_-} \sum_{\substack{\boldsymbol{r}_i \in \mathscr{S} \\ \mathscr{V}(\boldsymbol{r}_i) < 0}} (\mathscr{V}(\boldsymbol{r}_i) - \overline{\mathscr{V}}_-)^2. \qquad (8)$$

In this expression, the mean positive/negative $\mathscr{V}$ value on $\mathscr{S}$ is denoted $\overline{\mathscr{V}}_{+/-}$.

While the spread is a parameter that quantifies the variability of $\mathscr{V}$ on $\mathscr{S}$, other aspects may be important for intermolecular interactions. It has been suggested that one such aspect is the imbalance of negative and positive $\mathscr{V}$ values, the argument being that if significant imbalance is present, the probability for favourable intermolecular interactions diminishes [7,11,17]. Murray and coworkers [21] proposed to use the quantity

$$\nu = \frac{\sigma_-^2 \sigma_+^2}{\left(\sigma_-^2 + \sigma_+^2\right)^2} \qquad (9)$$

for this purpose, a reasonable choice if σ_-^2 and σ_+^2 correspond to 'amounts' of negative and positive values on $\mathscr{S}$. A perhaps simpler way to construct a parameter signifying imbalance is to make use of well-known measures of the asymmetry of distributions. To define such a quantity, we thus apply the statistical notion of the coefficient of skewness [27],

$$s = \frac{\frac{1}{N} \sum_{i=1}^{N} (\mathscr{V}(\boldsymbol{r}_i) - \overline{\mathscr{V}})^3}{\sigma^3}. \qquad (10)$$

It is our understanding of the term 'imbalance' that it should correspond to the 'absolute skewness of $\mathscr{V}$ about the origin'. Consequently, we define an imbalance parameter Θ by first replacing $\overline{\mathscr{V}}$ by zero in s, and then computing its absolute value. We will consider Θ as a possible correction to quantum mechanical volumes. To incorporate both variation and imbalance, Murray et al. [21] combined these two aspects of $\mathscr{V}$ in the parameter $\nu \sigma_{\text{tot}}^2$.

Politzer and coworkers [6] reasoned that the likelihood of attractive intermolecular interactions is maximised when the imbalance is close to zero (ν is large) and the variation is considerable (σ_{tot}^2 is large). With this interpretation, the molecules for which $\nu \sigma_{\text{tot}}^2$ is large are the ones that one would expect to be more closely packed in the solid than suggested by their volume. They therefore investigated two models for predicting the crystal density, one correcting for variation, and one correcting for both variation and imbalance:

$$\widehat{\rho} = \alpha \left(\frac{M}{V_{\text{encl}}} \right) + \beta \, \Pi + \gamma, \qquad (11)$$

$$\widehat{\rho} = \alpha \left(\frac{M}{V_{\text{encl}}} \right) + \beta \, \nu \sigma_{\text{tot}}^2 + \gamma. \qquad (12)$$

Fitting these equations to a set of 36 CHNO compounds, the training-set RMSE obtained was 0.053 g/cm^3 for Equation (11), and 0.047 g/cm^3 for Equation (12). The uncorrected V_{encl} yielded an RMSE of 0.068 g/cm^3. In an evaluation [8], Equation (12) was validated on a training-set and test-set consisting of 180 and 38 CHNO compounds, respectively. They found a mean absolute error (MAE) of 0.035 g/cm^3, compared to 0.050 g/cm^3 for the uncorrected V_{encl}. To our knowledge, Π has never been validated. But it does deserve such a test, given that the training-set RMSE obtained by Politzer and coworkers [6] was very similar for Π and $\nu \sigma_{\text{tot}}^2$.

While it has been noted [16–18] that the particular choice of surface $\mathscr{S}$ should be of minor importance, this claim has important mathematical consequences severely restricting the space of permissible parameters Φ used in linear models. This is the subject of a recent paper [28], in which we present a proof of the following. Consider a linear model of the form

$$\text{property} = \alpha \, \Phi + \beta. \qquad (13)$$

For such a model to make the same predictions on a range of surfaces $\mathscr{S}_\epsilon$ of electron density ϵ, the parameter values
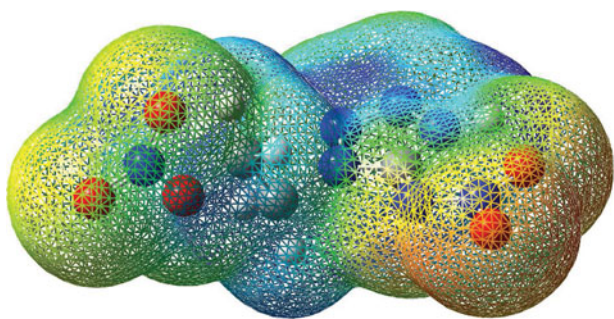
Figure 1. The electron density isosurface $\mathscr{S}_{0.0010}$ for compound 1 with the B3LYP functional. The illustration was produced with the GaussView 4.1 software [29].

of Φ computed on $\mathscr{S}_{\epsilon}$ and $\mathscr{S}_{\epsilon'}$ must be linearly related to each other. This requirement is independent of the modelled property, and hence of considerable importance for all attempts to apply molecular surface parameters in modelling. We tested the mathematical requirement in the particular case of imbalance and variation, showing that whereas variation ($\Phi = \sigma, \Pi$) satisfies it in the electron density range [0.0001 a.u., 0.0040 a.u.], imbalance ($\Phi = \nu, \Theta$) does not. This implies, in particular, that the model in Equation (16b) will fail to make the same predictions on these surfaces. The surface $\mathscr{S}_{0.0010}$ is illustrated in Figure 1.

2. Procedure

We study a set of 44 neutral molecular crystals. Among these, 23 have empirical formula of the form $C_xH_yN_zO_w$, six of the form $C_xH_yO_z$ and four can be written $C_xN_yO_z$. The remaining 11 compounds also contain carbon, hydrogen, nitrogen and oxygen, but some of them contain either chlorine, phosphor or sulphur. These latter compounds were included to shed some light on the composition-dependence of the studied models, and, in particular, to investigate the suggestion by Rice *et al.* that the density of halogen-containing compounds might be much less accurately estimated by their volume (10%–40%) [5]. An overview of the compounds is given in the Supporting Information Table 1. The experimental densities were obtained from the SciFinder database [30].

Table 1. Model prediction errors (RMSEP). All volumes are given in units of cm³/mol.

	$\widehat{V}_{(14)}$ Isosurface	$\widehat{V}_{(15a)}$ Π	$\widehat{V}_{(15b)}$ σ	$\widehat{V}_{(16a)}$ $\nu\sigma_{tot}^2$	$\widehat{V}_{(16b)}$ Θ
B3LYP	7.40	6.27	6.04	6.59	7.63
M05	7.49	6.39	6.16	6.75	7.72

Molecular geometry optimisations were performed with the GAUSSIAN03 Revision E.01 software [31], applying DFT with the B3LYP functional and basis sets of type 6-31G(d). Adopting the approach of Qiu *et al.* [32], we constructed the initial geometries using no information of the molecular geometry in the crystal. This procedure was different from that of Rice *et al.* [5,8] who selected the initial geometry as that of the crystal, thereby minimising the error caused by an inaccurate guess of the initial geometry, leading to a good volume estimate on which to add further corrections. It is of course also of interest to be able to predict the density for novel or notional compounds for which there exists no crystalline data. The viability of this approach rests on the assumption that geometries are, for the most part, of sufficiently small importance for the resulting volume estimate.

We performed single-point calculations with the B3LYP and M05 functionals in the search for possible functional dependencies of the volume estimates and the electrostatic parameters. The electron density $n(\mathbf{r})$ and the electrostatic potential $\mathscr{V}(\mathbf{r})$ were evaluated on a 100 by 100 by 100 grid. Points were identified as being on the surface $\mathscr{S}_{0.0010}$ with a tolerance of 10^{-4} a.u., yielding 6000–10, 000 points for each molecule. The statistical quantities Π, σ_-^2, σ_+^2, σ and Θ were subsequently evaluated on this surface, and UCSF Chimera [33] was applied to find the enclosed volume V_{encl}.

Five models for the molecular volume V are presented. While earlier studies have proposed models where corrections are linear in ρ [6–8], we will continue to assume linearity in our models of V. First, we study the uncorrected model

$$\widehat{V}_{(14)} = V_{encl} + \gamma. \tag{14}$$

Equation (14) will provide a good reference to ascertain the efficiency of the more complicated models. These have been selected to identify the factor responsible for the reported improvement from electrostatic considerations [6–8]. We consider two models that correct for the variation of $\mathscr{V}$:

$$\widehat{V}_{(15a)} = V_{encl} + \beta\,\Pi + \gamma, \tag{15a}$$

$$\widehat{V}_{(15b)} = V_{encl} + \beta\,\sigma + \gamma. \tag{15b}$$

We also investigate whether the imbalance of positive and negative $\mathscr{V}$ values is a relevant factor, both in combination with variation ($\nu\sigma_{tot}^2$) and as a potential correction in its own right (Θ):

$$\widehat{V}_{(16a)} = V_{encl} + \beta\,\nu\sigma_{tot}^2 + \gamma, \tag{16a}$$

$$\widehat{V}_{(16b)} = V_{encl} + \beta\,\Theta + \gamma. \qquad (16b)$$

The models which include the quantities Π and $\nu\sigma_{tot}^2$ suggested by Politzer *et al.* [6] are Equations (15a) and (16a), respectively. Although the predictions of the model based on Θ will be altered with changes in the surface [28], we have included it for the purpose of ascertaining its usefulness on the particular $\mathscr{S}_{0.0010}$ surface.

Due to the relatively small set of 44 compounds, we performed a 10-fold cross-validation for each model. Dividing the set into 10 separate folds, the models were trained using 9 out of 10 folds in 10 consecutive runs. In each run i, we computed the root-mean-square error of prediction (RMSEP), defined by [24]

$$\text{RMSEP}_i = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\widehat{V}_j - V_j^{exp})^2}. \qquad (17)$$

In this equation, N is the number of predicted compounds in the given run, and V^{exp} the experimental molecular volume, as determined by Equation (1) from the experimental density ρ^{exp}. The model RMSEP was then taken to be the average error of prediction of the 10 consecutive runs:

$$\text{RMSEP} = \frac{1}{10}\sum_{i=1}^{10}\text{RMSEP}_i. \qquad (18)$$

3. Results and discussion

The main results of our work can be divided into two parts. We begin by studying the predictive ability of the models that include electrostatic corrections based on variation, imbalance and the combination of the two. Subsequently we take a closer look at the relation of $\nu\sigma_{tot}^2$ to variation measures, such as σ_{tot}^2, Π and σ, arguing that the success of $\nu\sigma_{tot}^2$ is due to variation and not imbalance.

3.1. Predictive ability

In this section, we review the results of the 10-fold cross-validation. The predictive ability of the models, given in terms of their RMSEP values, is shown in Table 1.

The best performing models are those based on the variation parameters Π and σ, i.e. Equations (15a) and (15b). Compared to the uncorrected enclosed volume estimates, the Π and σ parameters show a 15% and 18% relative decrease in RMSEP, respectively, regardless of functional. Moreover, it is evident that the imbalance-variation parameter $\nu\sigma_{tot}^2$ leads to no improvement over

the simpler variation parameters Π and σ. Since applying imbalance as a linear term leads to inconsistent predictions on different surfaces [28], it is not surprising that Θ performs poorly, yielding no improvement over uncorrected volume estimates. Employing the surface invariant skewness s' about zero (see [28]) in $\widehat{V} = V_{encl} + \beta\,s' + \gamma$ leads to no improvement over $\Theta = |s'|$; however, we believe this is due to its irrelevance to the problem at hand. This is in contrast to Θ, where the poor performance is likely attributable to its non-linear variability with changes in the molecular surface. In conclusion, our study provides evidence in favour of the view that the variation of $\mathscr{V}$ on the molecular surface is related to the error of naive volume estimates. Furthermore, the findings clearly indicate that the imbalance of negative and positive $\mathscr{V}$ values is of no relevance, as this factor does not improve upon variation parameters and appears to be of no use on its own.

In order to investigate the observed changes in RMSEP further, we inspected the error distributions of the models, see Figure 2. Observe that the effect of Π and σ sharpens the peak of the error distribution, causing more compounds to be estimated within 5 cm^3/mol of experiment. The tails are left mostly unchanged. Considering B3LYP, the σ correction increases the number of compounds within 5 cm^3/mol from 22 to 27 for Π and to 29 for σ. The figures are similar for the M05 functional. The similar accuracy of the two functionals is due to the fact that they agree to a high extent on the value of the parameters ($R^2 > 0.99$) and that their volume estimates V_{encl} are very similar on the B3LYP geometries (standard deviation equal to 0.25 cm^3/mol). It should be noted, however, that the functionals might disagree to a significant extent if the optimisations are done independently. Thus, if one observes a considerable functional dependence, one may expect that this is mainly due to their influence on the optimised geometry.

A concern that must be addressed is that optimised geometries of isolated molecules will differ from the molecular geometries in the crystal. To address this issue, we compared our values with those obtained by Rice *et al.* [5,8], as they ran the optimisation using the geometry of the compound in the crystal as the initial guess. Their values for the four compounds in common with our study are in good agreement with our data, the deviations being less than 1 cm^3/mol, see Table 2. We also achieve comparable accuracy, with our estimates yielding for Equations (14) and (15b) a mean average error of 0.070 and 0.054 g/cm^3 from the experimental densities, compared to the respective values 0.050 g/cm^3 (V_{encl}) and 0.035 g/cm^3 ($\nu\sigma_{tot}^2$) obtained by Rice and coworkers [8]. Although a more extensive study is required to draw definite conclusions, we note that the considerable
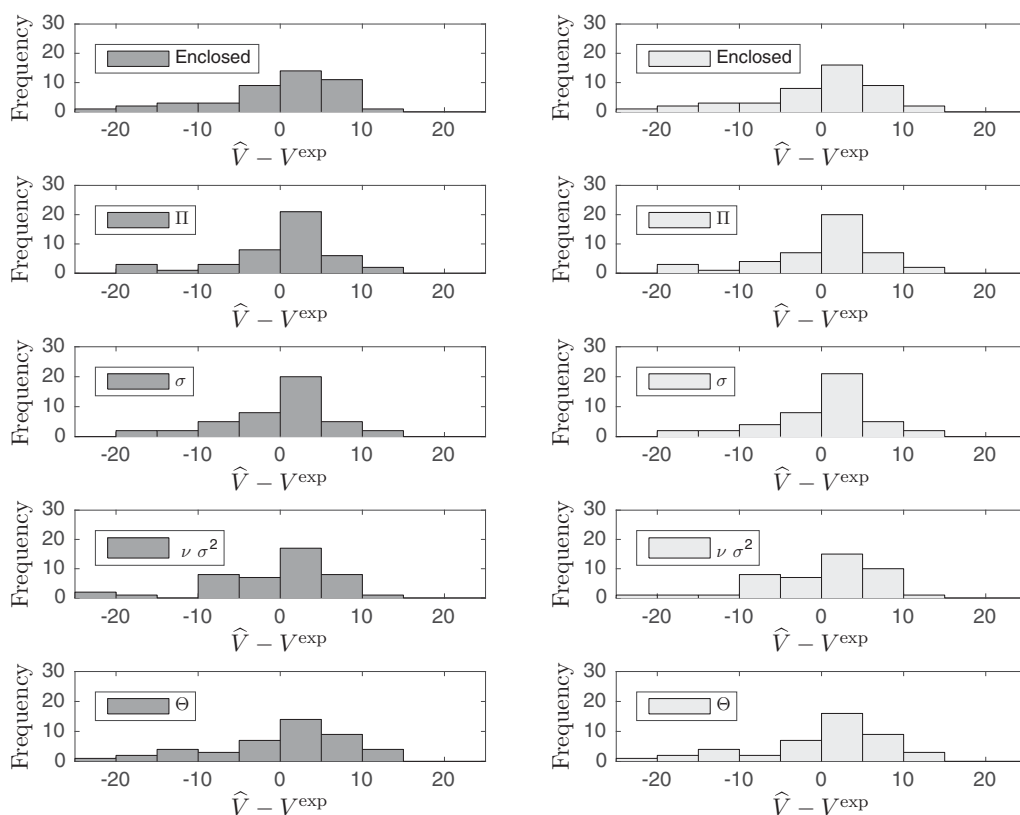
Figure 2. Error distributions of the models. The M05 distributions are illustrated in dark grey, while those of B3LYP are light grey. The errors are given in cm^3/mol.

agreement and comparable accuracy suggest that crystal densities are in many cases effectively predicted without any knowledge of the molecular geometry in the solid. It may be that our results underestimate the importance of the initial geometry in general, however. We note that many of the compounds studied have limited

Table 2. Listed are the B3LYP/6-31G(d) enclosed volumes V_{encl} in our study, and the B3LYP/6-31G(d,p) enclosed volumes $V_{encl}^{crystal}$ obtained by Rice *et al.* [5,8] using the geometry in the crystal as the initial guess. All values are given in cm^3/mol.

Structure	Formula	Compound	V_{encl}	$V_{encl}^{crystal}$
	$C_3H_5N_7O_5$	1	126.0	125.1
	$C_{12}H_{11}NO_3$	26	151.6	152.3
	$C_6H_4N_2O_5$	33	108.1	108.2
	$CH_2N_6O_2$	40	74.2	73.7
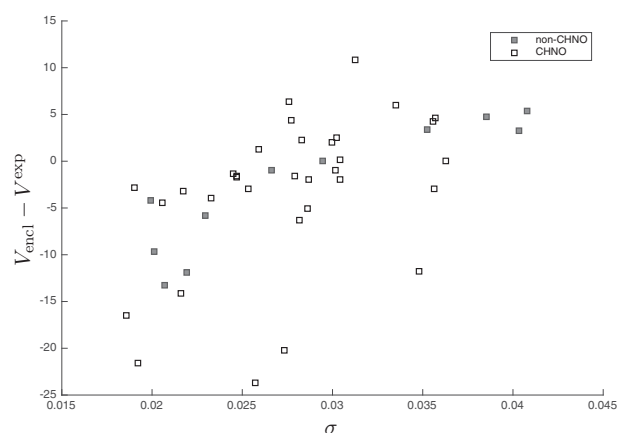
Figure 3. The error of the enclosed volume estimates V_{encl} plotted against σ. Non-CHNO compounds are shown in dark grey, and CHNO compounds in light grey. The values of σ and V_{encl} were computed with the M05 functional and are given in units of Hartree and cm^3/mol, respectively.

conformational freedom to change as a consequence of the altered environment in the solid.

Since the parameters considered in previous studies [6–8] have only been applied to CHNO compounds, we decided to look closer at the σ values of non-CHNO

Table 3. Fitted parameters for the enclosed volume estimates and the models including Π and σ corrections. The AICc values are listed with respect to the model the lowest value. The unit of β is cm^3/(mol Hartree) and the unit of γ is cm^3/mol.

		β	γ	AICc − AICc$_{\min}$
B3LYP/6-31G(d)	Enclosed (14)	–	2.332	15.8
	Π (15a)	−703.735	18.380	1.9
	σ (15b)	−720.314	22.296	0.0

compounds, see Figure 3. The effectiveness of variation (here given by σ) is evident from the figure, as it is seen that the higher the variation of the potential $\mathscr{V}$, the more likely V_{encl} is an overestimate. This is consistent with the interpretation given by Politzer *et al.* [6], who suggested that variation of $\mathscr{V}$ makes attractive intermolecular interactions more likely, causing the molecules to be closer in the solid than one would expect from their volume. While the non-CHNO compounds appear to be accurately described by both V_{encl} and σ, as compared to the CHNO systems, we note that the three compounds of highest σ all contain sulphur. A more extensive study is thus needed to ascertain the domain of validity of the variation parameter σ.

For future reference and use, we have fitted the uncorrected models as well as those corrected with Π and σ for B3LYP, see Table 3. It should be kept in mind that these models, being based on all compounds, have not been validated. For this reason, we have also listed their AIC values, corrected for sample size (AICc). A rule-of-thumb lends 'substantial support' (<2) to the Π and σ corrected models, but 'essentially no support' (>10) to the uncorrected model [23].

3.2. The possible irrelevance of imbalance

We have demonstrated that the imbalance-variation parameter $\nu\sigma_{\text{tot}}^2$ leads to no noticeable improvement over the simpler variation parameters Π and σ. Our work suggests that the findings of Rice and Byrd [8] is due to the close relation of $\nu\sigma_{\text{tot}}^2$ to σ_{tot}^2, a measure of variation. A comparison of the two reveals R^2 to be about 0.75, see Figure 4.

Furthermore, since σ_{tot}^2 is a measure of spread, it is also related to Π and σ. Considering the relation of $\nu\sigma_{\text{tot}}^2$ to σ^2, we found a correlation coefficient of $R^2 = 0.40$–0.45. Though not particularly high, it might be sufficiently high for both of them to correct for the same aspect of $\mathscr{V}$, namely variation. Politzer *et al.* [6] reported $R^2 = 0.68$ between Π and $\nu\sigma_{\text{tot}}^2$, but concluded that this value was sufficiently low for them to signify different aspects of $\mathscr{V}$. While it is true that ν might be a small correction, there is as yet no evidence for this, neither in our results nor in the work of Politzer and coworkers [6]. The slight decrease in the training set RMSE is not significant (0.053–0.047 g/cm^3). We therefore conclude that ν is adding noise, not correcting.
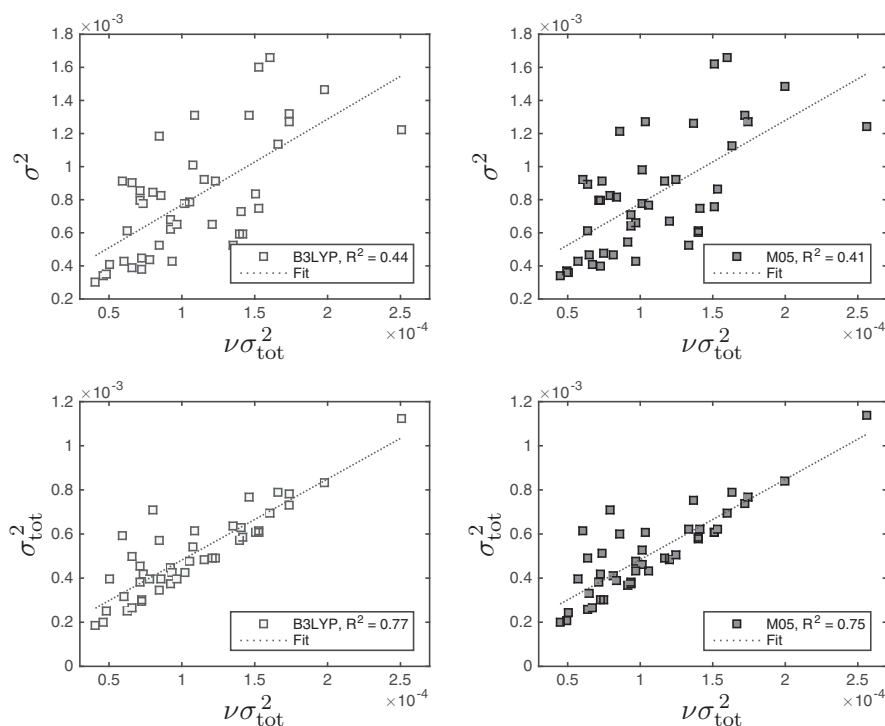
Figure 4. The values of $\nu\sigma_{\text{tot}}^2$ plotted against the variation quantities σ^2 (top) and σ_{tot}^2 (bottom). All values are given in Hartree^2.

This discussion has presupposed that Θ or ν are suitable measures of imbalance. We note that they are at the very least consistent with each other, as they have been shown to be in fair agreement [28]. By this we mean that, if the absolute skewness Θ is zero, it is likely that the imbalance parameter ν is equal to 0.25 (indicating no imbalance). Similarly, as Θ increases, the value of ν is seen to decrease.

4. Conclusions and summary

In agreement with previous studies, the enclosed volume of surfaces of small electron density is seen to efficiently predict the crystal density of neutral molecular crystals. Moreover, we have found that, even when the geometry is not taken to be the one experimentally found in the solid, the predictive ability is largely maintained. Our work shows that the presence of other atoms than carbon, hydrogen, nitrogen and oxygen does not result in a breakdown of the method, as the eleven such compounds studied are equally well approximated as the CHNO systems. Our study presents evidence in favour of the view that variation of the electrostatic potential (measured by the standard deviation and related quantities) on the surface of the molecule is responsible for small but possibly significant corrections to quantum mechanical volumes. Applying two measures of variation, we obtained a relative decrease in RMSEP of 15%–18%. The corrections sharpen the error distribution peak, yielding more predictions within 5 cm^3/mol of experiment.

The imbalance of the electrostatic potential on the surface of molecules is demonstrated to be irrelevant for the prediction of crystal densities. First, when used as a linear term, it leads to no improvement over the uncorrected estimates. In view of the fact that imbalance yields inconsistent predictions on different surfaces [28], this finding is not surprising. Second, when combined with the variation of the electrostatic potential, it does not improve predictions compared to the simpler corrections based on variation alone. The previously reported success of imbalance combined with variation is therefore likely due to variation alone, i.e. the role of imbalance is to add noise rather than to correct.

Note

1. We have used the unbiased estimator of the standard deviation, dividing by $N - 1$ instead of N. This choice does not affect the results since N is greater than 6000 in our calculations.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] M.J. Kamlet and S.J. Jacobs, J. Chem. Phys. **48** (3), 23–35 (1968).
[2] T.M. Klapötke, *Chemistry of High-Energy Materials* (Walter de Gruyter, Berlin, 2012).
[3] A. Gavezzotti, Acc. Chem. Res. **27** (10), 309–314 (1994).
[4] S.M. Woodley and R. Catlow, Nat. Mater. **7** (12), 937–946 (2008).
[5] B.M. Rice, J.J. Hare, and E.F.C. Byrd, J. Phys. Chem. A **111** (42), 10874–10879 (2007).
[6] P. Politzer, J. Martinez, J.S. Murray, M.C. Concha, and A. Toro-Labbé, Mol. Phys. **107** (19), 2095–2101 (2009).
[7] P. Politzer, J. Martinez, J.S. Murray, and M.C. Concha, Mol. Phys. **108** (10), 1391–1396 (2010).
[8] B.M. Rice and E.F.C. Byrd, J. Comput. Chem. **34** (25), 2146–2151 (2013).
[9] R.F.W. Bader, M.T. Carroll, J.R. Cheeseman, and C. Chang, J. Am. Chem. Soc. **109** (26), 7968–7979 (1987).
[10] A.F. Jalbout, Z.Y. Zhou, X. Li, M. Solimannejad, and Y. Ma, J. Mol. Struct. **664–665**, 15 (2003). ISSN 0166-1280.
[11] J.S. Murray, T. Brinck, P. Lane, K. Paulsen, and P. Politzer, J. Mol. Struct. **307**, 55 (1994). ISSN 0166-1280.
[12] H. Hagelin, J.S. Murray, P. Politzer, T. Brinck, and M. Berthelot, Can. J. Chem. **73** (4), 483–488 (1995).
[13] J.S. Murray, M.C. Concha, and P. Politzer, Mol. Phys. **107** (1), 89–97 (2009).
[14] J.S. Murray, P. Lane, and P. Politzer, Mol. Phys. **93** (2), 187–194 (1998).
[15] P. Politzer, J.S. Murray, P. Lane, and T. Brinck, J. Phys. Chem. **97** (3), 729–732 (1993).
[16] P. Politzer, J.S. Murray, and Z. Peralta-Inga, Int. J. Quant. Chem. **85** (6), 676–684 (2001).
[17] P. Politzer and J.S. Murray, Theor. Chem. Acc. **108** (3), 134–142 (2002).
[18] T. Brinck, P. Jin, Y. Ma, J.S. Murray, and P. Politzer, J. Mol. Model. **9** (2), 77–83 (2003). ISSN 1610-2940.
[19] P. Politzer, P. Lane, J.S. Murray, and T. Brinck, J. Phys. Chem. **96** (20), 7938–7943 (1992).
[20] T. Brinck, J.S. Murray, and P. Politzer, J. Org. Chem. **58** (25), 7070–7073 (1993).
[21] J.S. Murray, P. Lane, T. Brinck, K. Paulsen, M.E. Grice, and P. Politzer, J. Phys. Chem. **97** (37), 9369–9373 (1993).
[22] T. Brinck, J.S. Murray, and P. Politzer, Int. J. Quant. Chem. **48** (2), 73–88 (1993). ISSN 1097-461X.
[23] K.P. Burnham and D.R. Anderson, Socio. Meth. Res. **33** (2), 261–304 (2004).
[24] R. Wehrens, *Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences* (Springer, Heidelberg, 2011).
[25] T. Brinck, J.S. Murray, and P. Politzer, Mol. Phys. **76** (3), 609–617 (1992).
[26] W. Ledermann, E. Lloyd, *Handbook of Applicable Mathematics. Vol. VI: Statistics, Part A.* (Wiley, New York, NY, 1984).
[27] G.A. Korn and T.M. Korn, *Mathematical Handbook for Scientists and Engineers: Definitions, Theorems, and Formulas for Reference and Review* (Dover Publications, New York, NY, 2000).
[28] E.F. Kjønstad, J.F. Moxnes, T.L. Jensen, and E. Unneberg, Mol. Phys. doi:10.1080/00268976.2016.1140842 (2016).

[29] R. Dennington, T. Keith, and J. Millam, *Gaussview Version 4.1* (Semichem Inc. Shawnee Mission KS, 2007).

[30] Scifinder, *Chemical Abstracts Service* (Columbus, OH, 2014).

[31] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, J.A. Montgomery, Jr., T. Vreven, K.N. Kudin, J.C. Burant, J.M. Millam, S.S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G.A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J.E. Knox, H.P. Hratchian, 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, P.Y. Ayala, K. Morokuma, G.A. Voth, P. Salvador, J.J. Dannenberg, V.G. Zakrzewski, S. Dapprich, A.D. Daniels, M.C. Strain, O. Farkas, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J.V. Ortiz, Q. Cui, A.G. Baboul, S. Clifford, J. Cioslowski, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, C. Gonzalez, and J.A. Pople, *Gaussian 03, Revision E.01*. (Gaussian, Inc., Wallingford, CT, 2007).

[32] L. Qiu, H. Xiao, X. Gong, X. Ju, and W. Zhu, J. Hazard. Mater. **141** (1), 280–288 (2007). ISSN 0304–3894.

[33] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin, J. Comput. Chem. **25** (13), 1605–1612 (2004).