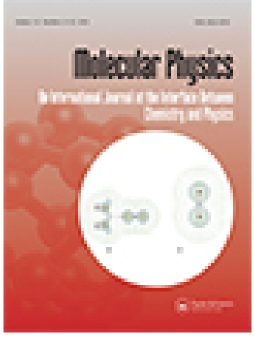


Taylor & Francis
Taylor & Francis Group

Molecular Physics

An International Journal at the Interface Between Chemistry and Physics

ISSN: 0026-8976 (Print) 1362-3028 (Online) Journal homepage: http://www.tandfonline.com/loi/tmph20

All molecular surfaces are equal: demanding invariance of predictions in linear single-variable models

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

To cite this article: Eirik F. Kjønstad, John F. Moxnes, Tomas L. Jensen & Erik Unneberg (2016): All molecular surfaces are equal: demanding invariance of predictions in linear single-variable models, Molecular Physics, DOI: 10.1080/00268976.2016.1140842

To link to this article: http://dx.doi.org/10.1080/00268976.2016.1140842

Published online: 03 Feb 2016.

MOLECULAR PHYSICS, 2016
http://dx.doi.org/10.1080/00268976.2016.1140842

Taylor & Francis
Taylor & Francis Group

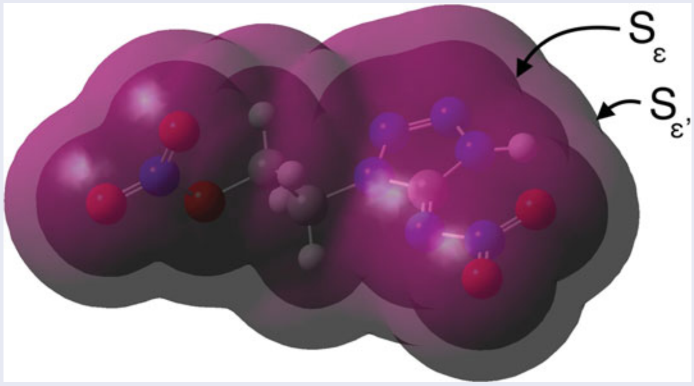
RESEARCH ARTICLE

All molecular surfaces are equal: demanding invariance of predictions in linear single-variable models

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

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

ABSTRACT
The molecular surface has been suggested to be a region of the molecule, where information of non-covalent intermolecular interactions is present. Many workers have pursued this idea by constructing models based on statistical parameters Φ extracted from the electrostatic potential on a particular molecular surface. We claim that a better approach is to define a family of equivalent molecular surfaces, each associated with a particular electron density ϵ. The demand that any model must give the same predictions on all such molecular surfaces yields a mathematical requirement restricting the space of permissible parameters. We prove that linear single-variable models of the form property $=$ $\alpha\,\Phi + \beta$ will only yield invariant predictions if the parameter values of Φ computed on equivalent surfaces are linearly related. This claim is not restricted to the use of the electrostatic potential, but holds for any parameter extracted from the surface of molecules. By using a set of 44 molecules, we also demonstrate that a frequently used aspect of the electrostatic potential, that of 'imbalance' of negative and positive values, fails to satisfy the linearity requirement. It is argued that multi-variable models should only include parameters that satisfy the single-variable requirement.



ARTICLE HISTORY
Received 9 September 2015
Accepted 4 January 2016

KEYWORDS
Molecular surface; electrostatic potential; electron density; linear modelling; density functional theory

1. Introduction

Several attempts to model chemical and physical properties have made use of the electrostatic potential $\mathscr{V}(\boldsymbol{r})$, the electric potential felt by a unit charge at position $\boldsymbol{r}$. Some of these efforts have been based on the assumption that non-covalent intermolecular interactions are, in principle, determined by the values of $\mathscr{V}$ on the molecular surface [1]. For an isolated molecule, the electrostatic potential is uniquely determined from (1) its geometry, given by the nuclear coordinates $\boldsymbol{R}_i$, and (2) its electron density $n(\boldsymbol{r})$, the probability density to find an electron at $\boldsymbol{r}$. Both $\boldsymbol{R}_i$ and $n(\boldsymbol{r})$ are easily determined by quantum mechanical methods, and hence so is $\mathscr{V}(\boldsymbol{r})$. In particular, we have the relation

$$\mathscr{V}(\boldsymbol{r}) = \sum_i \frac{Z_i}{|\boldsymbol{R}_i - \boldsymbol{r}|} - \int \frac{n(\widetilde{\boldsymbol{r}})}{|\widetilde{\boldsymbol{r}} - \boldsymbol{r}|}\,\mathrm{d}\widetilde{\boldsymbol{r}}, \qquad (1)$$

where Z_i is the charge of nucleus i. An intriguing possibility is that, when restricted to the surface of an isolated molecule, one might quantify non-covalent interactions between identical molecules by analysing $\mathscr{V}$.

Bader *et al.* [2] proposed that the surface of molecules $\mathscr{S}$ is well represented by surfaces of constant and small electron density $n(\boldsymbol{r})$. They observed that $n(\boldsymbol{r})$ values of 0.0010 a.u. and 0.0020 a.u. served the purpose well. Inspired by this work, others adopted one of these

CONTACT Erik Unneberg ✉ erik.unneberg@ffi.no

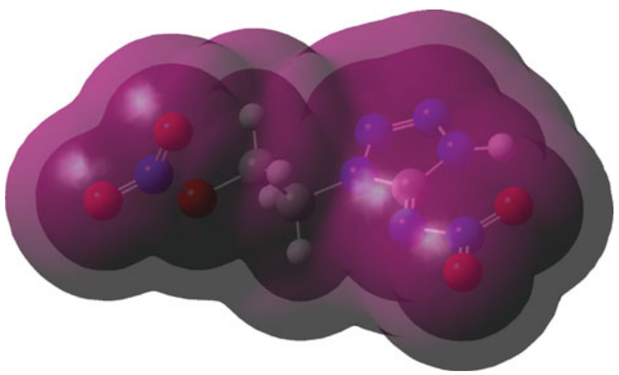


Figure 1. Surfaces of constant electron density for compound 1. The electron density on the outer surface is 0.0001 a.u., while it is 0.0040 a.u. on the inner. Both surfaces were computed with the B3LYP functional and a basis set of type 6-31G(d). The illustration was produced with the GaussView 4 software [13].

choices, usually the former [3–12]. It is vital to our work, however, that a particular surface is not selected. Rather, we define a family of molecular surfaces

$$\mathscr{S}_\epsilon = \{\boldsymbol{r} : n(\boldsymbol{r}) = \epsilon\}, \quad \epsilon_{\min} \leq \epsilon \leq \epsilon_{\max}, \tag{2}$$

and consider them all equivalent in terms of information content. By increasing ϵ, the surface $\mathscr{S}_\epsilon$ will at some point invade the inner parts of the molecule, making it a poor representation of its surface. Similarly, as ϵ is decreased, electrons become increasingly rare, and eventually $\mathscr{S}_\epsilon$ is far away from the nuclei. Consequently, we propose that there exists an interval $[\epsilon_{\min}, \epsilon_{\max}]$ for which $\mathscr{S}_\epsilon$ is an adequate representation of the molecular surface. The sudden boundaries are fictional, the true picture resembling one of gradual loss of information as the endpoints are reached and surpassed. We assume that $\mathscr{S}_{0.0001}$ and $\mathscr{S}_{0.0040}$ are molecular surfaces. They are illustrated in Figure 1.

Given the assumption that a property is to some extent determined by non-covalent intermolecular interactions, workers have restricted $\mathscr{V}$ to a particular surface $\mathscr{S}_\epsilon$, extracting statistical quantities deemed relevant for predicting said property. The properties modelled by this procedure include impact sensitivities [10,11], critical and boiling points [14], sublimation enthalpies and solvation free energies [1], solubilities [12,15], partition coefficients [16], crystal densities [5,8] and the potency of inhibitors as anti-HIV agents [17,18]. In general, it is assumed that there exists a function *f*, known as a general interaction properties function (GIPF) [4], such that

$$\text{property} = f(\Phi^{(1)}, \Phi^{(2)}, \ldots, \Phi^{(n)}), \tag{3}$$

where the $\Phi^{(i)}$ are parameters extracted from $\mathscr{V}$ restricted to the molecular surface. The objective of the developers of the GIPF procedure was to find means of making satisfactory predictions for properties of practical interest with a particular focus on molecular surfaces. In spite of the fact that it appears that a reasonable accuracy has been obtained in many cases, competing models *f* have often been left non-validated and judged solely by their ability to fit the data, regardless of complexity [4,9–11,14,16,17,19]. It is good practice to validate models, for instance by performing cross-validation or boot-strapping [20]. If validation is not possible, models of differing complexity are often compared by information criteria, which favour parsimonious models [21]. Given the fact that many of the GIPF models are very flexible (allowing several variables) and hence, vulnerable to overfitting, it is important that they are tested on unseen data-sets. Adopting the practice of validation will be beneficial to the further study of the effectiveness of these models.

Our work addresses a more general issue: if a model *f* is proposed whose parameters $\Phi^{(i)}$ are computed on the molecular surface $\mathscr{S}_\epsilon$, the choice of ϵ should be of minor or no importance, provided one stays within the as-yet unknown range $[\epsilon_{\min}, \epsilon_{\max}]$ yielding adequate representations of the molecular surface. That is, if the parameters were computed on a different molecular surface, say $\mathscr{S}_{\epsilon'}$, the predictions made by the model should not change. While others have remarked that the particular choice of electron density should be of minor importance [1,17,18], the mathematical consequences of this claim have not been investigated. Exploring these consequences is the subject of this paper.

In the special case of a linear single-variable model *f*, requiring that predictions remain the same on $\mathscr{S}_\epsilon$ and $\mathscr{S}_{\epsilon'}$ is mathematically equivalent to demanding the existence of constants α, β, α', β', such that

$$\text{property}_i = \alpha \; \Phi_i(\epsilon) + \beta, \tag{4a}$$

$$\text{property}_i = \alpha' \; \Phi_i(\epsilon') + \beta'. \tag{4b}$$

In these equations, $\Phi_i(\epsilon)$ is the parameter Φ computed on the surface $\mathscr{S}_\epsilon$ of molecule *i*. Substituting Equation (4b) into Equation (4a), we find that

$$\Phi_i(\epsilon') = \widetilde{\alpha} \; \Phi_i(\epsilon) + \widetilde{\beta}, \tag{5}$$

with $\widetilde{\alpha} = \alpha/\alpha'$ and $\widetilde{\beta} = (\beta - \beta')/\alpha'$. We have discovered that if predictions are surface-invariant, then the parameter values of Φ computed on different surfaces must be linearly related to each other. Of particular interest is the contrapositive variant of this statement: if Equation (5) is false, then the predictions will differ on the two surfaces, i.e. Equations (4a) and (4b) cannot be true simultaneously. This means that parameters Φ which violate the linearity of Equation (5) should never be used in linear single-variable models. We will refer to the requirement that parameters satisfy Equation (5) as the principle of molecular surface invariance.

In order to illustrate the usefulness of the principle, we will study two aspects of $\mathscr{V}$: the variation of $\mathscr{V}$ and the imbalance of positive and negative values of $\mathscr{V}$. It has been argued that variation is a measure of 'local polarity', thus being a prerequisite for intermolecular interactions [22], and that imbalance renders favourable interactions less probable [1,4,7]. We will study the following two parameters quantifying variation:[1]

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

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\mathscr{V}(\boldsymbol{r}_i) - \overline{\mathscr{V}})^2}. \tag{6b}$$

Both Π and σ are measures of the spread about the mean. Here, N is the number of points on the discretised molecular surface, and $\overline{\mathscr{V}}$ is the average potential value on the surface. Since its introduction by Brinck *et al.* [22], Π has been used to model octanol/water partition coefficients [16], enzyme inhibition concentrations [17,18], crystal densities [5,7,8,23], impact sensitivities [4] and solute-induced frequency shifts [9]. As we wish to study aspects of $\mathscr{V}$ in general, rather than investigating particular parameters, we consider σ as an analogue of the frequently used Π. Similarly, to quantify imbalance, we study the following parameters:

$$\nu = \frac{\sigma_-^2 \sigma_+^2}{(\sigma_-^2 + \sigma_+^2)^2}, \tag{7a}$$

$$\Theta = \left| \frac{\frac{1}{N}\sum_{i=1}^{N} \mathscr{V}(\boldsymbol{r}_i)^3}{\sigma^3} \right|. \tag{7b}$$

In Equation (7a), σ_-^2 and σ_+^2 are variances of $\mathscr{V}$ on the parts of the molecular surface for which $\mathscr{V}$ is negative and positive, respectively. It should be noted that $\sigma_-^2 = \sigma_+^2$ (signifying balance) implies $\nu = 0.25$, while $\sigma_-^2 \neq \sigma_+^2$ (signifying imbalance) implies $\nu < 0.25$. The imbalance parameter ν has been used to model crystal densities [7,8], enzyme inhibition concentrations [17,18], impact sensitivities [10], boiling and critical points [14] and solubilities [4]. We treat Θ as an analogue to ν. To understand the reasoning leading to Θ, consider the coefficient of skewness [24],

$$s = \frac{\frac{1}{N}\sum_{i=1}^{N} (\mathscr{V}(\boldsymbol{r}_i) - \overline{\mathscr{V}})^3}{\sigma^3}. \tag{8}$$

We claim that imbalance is adequately represented by the 'absolute skewness about zero'. To obtain a measure of this, we first replaced $\overline{\mathscr{V}}$ by zero (denoting the resulting quantity by s') and then took its absolute value ($\Theta = |s'|$). In a recent paper, we studied the role of variation (σ, Π) and imbalance (ν, Θ) in the prediction of crystal densities [23].

Both Π and ν have been applied to model the impact sensitivity of a material, as given by h_{50}, the height at which 50 per cent of samples detonate, and I_{50}, the impact energy at which 50 per cent of samples detonate [4,10],

$$h_{50} = \alpha\ \Pi + \beta\ \Pi^2 + \gamma, \tag{9}$$

$$I_{50} = \alpha\ \nu + \beta\ \sigma_+^2 + \gamma. \tag{10}$$

It should be kept in mind that these models have multiple variables, and consequently that the linearity of Equation (5) is not strictly required for the invariance of predictions.

2. Procedure

A set of 44 compounds were considered, most of them comprised of carbon, hydrogen, nitrogen and oxygen, but some also contain chlorine, fluorine, phosphorous or sulphur. An overview of the studied molecules is given in Table 1.

All computations were performed with the GAUSSIAN03 Revision E.01 software [25]. The molecular geometries were optimised with density functional theory (DFT), applying the B3LYP functional and basis sets of type 6-31G(d). Both $n(\boldsymbol{r})$ and $\mathscr{V}(\boldsymbol{r})$ were computed on a 100 by 100 by 100 grid. The statistical quantities Π, σ, ν and Θ were obtained manually on the five molecular surfaces $\mathscr{S}_{0.0001}$, $\mathscr{S}_{0.0005}$, $\mathscr{S}_{0.0010}$, $\mathscr{S}_{0.0025}$ and $\mathscr{S}_{0.0040}$. Points were identified as being on the surface with tolerances producing 4000–10,000 points on each surface, depending on the size of the molecule, see Table 2.

3. Results and discussion

Having computed $\Phi = \Pi$, σ, ν and Θ on five molecular surfaces, each parameter was compared with its value on the commonly used surface $\mathscr{S}_{0.0010}$. The rate at which linearity deteriorates may, thus, be taken to indicate how far away from the conventional molecular surface one must

Table 1. Overview of studied compounds, their ν and Π values on the surfaces $\mathscr{S}_{0.0001}$, $\mathscr{S}_{0.0010}$ and $\mathscr{S}_{0.0040}$. The quantity Π is listed in units of Hartree, ν is dimensionless.

Formula	Nr.	CAS-ID	$\nu_{0.0001}$	$\nu_{0.0010}$	$\nu_{0.0040}$	$\Pi_{0.0001}$	$\Pi_{0.0010}$	$\Pi_{0.0040}$
$C_3H_5N_7O_5$	1	1334230-65-7	0.25	0.24	0.20	0.018	0.026	0.036
$C_5H_7N_3O_5S$	2	1310420-43-9	0.23	0.25	0.23	0.027	0.036	0.044
$C_{10}H_{18}O_6$	3	1310054-86-4	0.25	0.25	0.24	0.014	0.021	0.029
$C_3H_6N_6O_3$	4	1308657-70-6	0.24	0.24	0.22	0.022	0.031	0.040
$C_{13}H_9ClN_4O$	5	1313411-75-4	0.25	0.24	0.25	0.011	0.015	0.021
$C_4H_5N_3O_4S$	6	1310420-42-8	0.25	0.23	0.21	0.027	0.036	0.044
$C_3HN_7O_8$	7	1286279-08-0	0.13	0.11	0.09	0.016	0.023	0.033
$C_3H_6N_4O_8$	8	1255215-40-7	0.14	0.13	0.10	0.019	0.026	0.035
$C_4H_4N_8$	9	1240554-70-4	0.22	0.24	0.25	0.016	0.023	0.032
$C_{13}H_{11}Cl_2N_3$	10	1224197-56-1	0.25	0.25	0.24	0.017	0.022	0.027
$C_5H_9N_5O_8$	11	1195678-79-5	0.22	0.19	0.15	0.016	0.025	0.034
$C_7H_{10}N_2O_3S$	12	1195693-61-8	0.25	0.24	0.22	0.025	0.032	0.041
$C_{12}H_9Cl_2NO_3$	13	50471-44-8	0.25	0.25	0.24	0.011	0.016	0.022
$C_8H_3Cl_2NO_4$	14	24564-72-5	0.16	0.22	0.25	0.010	0.015	0.022
$C_{14}H_{10}N_2O_2$	15	5585-14-8	0.14	0.19	0.24	0.011	0.015	0.020
$C_9H_8N_2$	16	1126-00-7	0.20	0.23	0.25	0.010	0.015	0.021
$C_{12}H_7N_7O_6$	17	1198599-44-8	0.23	0.22	0.20	0.017	0.024	0.034
$C_4H_6N_{12}O_4$	18	1135076-68-4	0.22	0.20	0.18	0.019	0.027	0.037
$C_2H_3N_5O_2$	19	1163729-75-6	0.24	0.21	0.18	0.020	0.027	0.037
$C_{10}H_8O$	20	135-19-3	0.22	0.19	0.15	0.011	0.017	0.022
$C_3H_3N_9O_2$	21	212318-72-4	0.20	0.18	0.15	0.023	0.032	0.042
$C_6H_2N_2O_8$	22	479-22-1	0.19	0.15	0.10	0.021	0.030	0.041
$C_6H_{10}N_6O_2$	23	1135076-65-1	0.22	0.25	0.25	0.014	0.020	0.027
$C_{12}H_{16}N_2$	24	1117789-09-9	0.14	0.16	0.20	0.019	0.025	0.031
$C_6H_7N_5O$	25	1071952-90-3	0.20	0.22	0.25	0.022	0.030	0.037
$C_{12}H_{11}NO_3$	26	86311-67-3	0.10	0.13	0.20	0.011	0.015	0.020
$C_{12}H_{18}N_4S$	27	83277-80-9	0.22	0.21	0.24	0.011	0.015	0.019
$C_4H_6N_6O_8$	28	81360-42-1	0.11	0.10	0.08	0.017	0.026	0.036
$C_6H_8NO_3P$	29	80241-43-6	0.25	0.25	0.24	0.015	0.023	0.031
C_7H_9NOS	30	77555-27-2	0.24	0.25	0.25	0.012	0.017	0.023
$C_{13}H_9N$	31	229-87-8	0.18	0.22	0.24	0.010	0.015	0.019
$C_8H_6O_4$	32	121-91-5	0.25	0.24	0.23	0.012	0.018	0.024
$C_6H_4N_2O_5$	33	51-28-5	0.24	0.25	0.23	0.015	0.022	0.030
$C_{10}H_{16}O_5$	34	1338212-46-6	0.17	0.20	0.24	0.014	0.020	0.028
$C_8H_6O_4$	35	100-21-0	0.25	0.24	0.23	0.012	0.017	0.025
$C_{15}H_{16}O_2$	36	80-05-7	0.21	0.19	0.18	0.010	0.015	0.021
$C_9H_5N_5O_6$	37	23309-22-0	0.23	0.21	0.19	0.017	0.026	0.036
$C_6H_4N_2O_2$	38	23296-57-3	0.22	0.25	0.25	0.013	0.018	0.025
$C_9H_9NO_3$	39	19182-97-9	0.17	0.22	0.25	0.015	0.021	0.028
$CH_2N_6O_2$	40	18588-16-4	0.23	0.19	0.16	0.021	0.031	0.041
$C_8H_4N_6O_6$	41	18373-83-6	0.18	0.18	0.15	0.017	0.025	0.035
$C_5H_5ClN_2O$	42	17551-52-9	0.22	0.22	0.25	0.019	0.025	0.035
$C_5H_6N_2O_2$	43	17325-26-7	0.24	0.25	0.25	0.013	0.020	0.029
$C_4H_2N_4O_2$	44	17098-88-3	0.24	0.25	0.24	0.017	0.026	0.034

stray before the models cease to yield consistent predictions. The results for the measures of spread and imbalance are given in Figures 2 and 3, respectively.

It is evident from these figures that imbalance (quantified by ν and Θ) does not satisfy the linearity in Equation (5). We conclude that imbalance should never be used in linear single-variable models, i.e. models of the form

$$\text{property} = \alpha\ \nu + \gamma. \tag{11}$$

Table 2. Tolerances applied to identify points on the studied molecular surfaces.

Surface	$\mathscr{S}_{0.0001}$	$\mathscr{S}_{0.0005}$	$\mathscr{S}_{0.0010}$	$\mathscr{S}_{0.0025}$	$\mathscr{S}_{0.0040}$
Tolerance [10^{-4}a.u.]	0.1	0.5	1	2	3

We have, in fact, shown that such a model cannot make the same predictions on all molecular surfaces. Inconsistent predictions in single-variable models does not preclude the possibility of consistent predictions in multi-variable models, a particular example of which is given in Equation (10). For this to be true, it seems that some highly non-trivial cancellation is required between the many parameters. At the very least, our results serve as a warning against such models, since the imbalance does not give consistent predictions for even the simplest of cases, that of a single variable. One may rightly object that ν and Θ could be improper measures of imbalance. We note, however, that they are in fair agreement, see Figure 4. A symmetric distribution should imply $\Theta = 0$ and $\nu = 0.25$, while asymmetry should increase Θ and reduce ν. This is consistent with our findings.

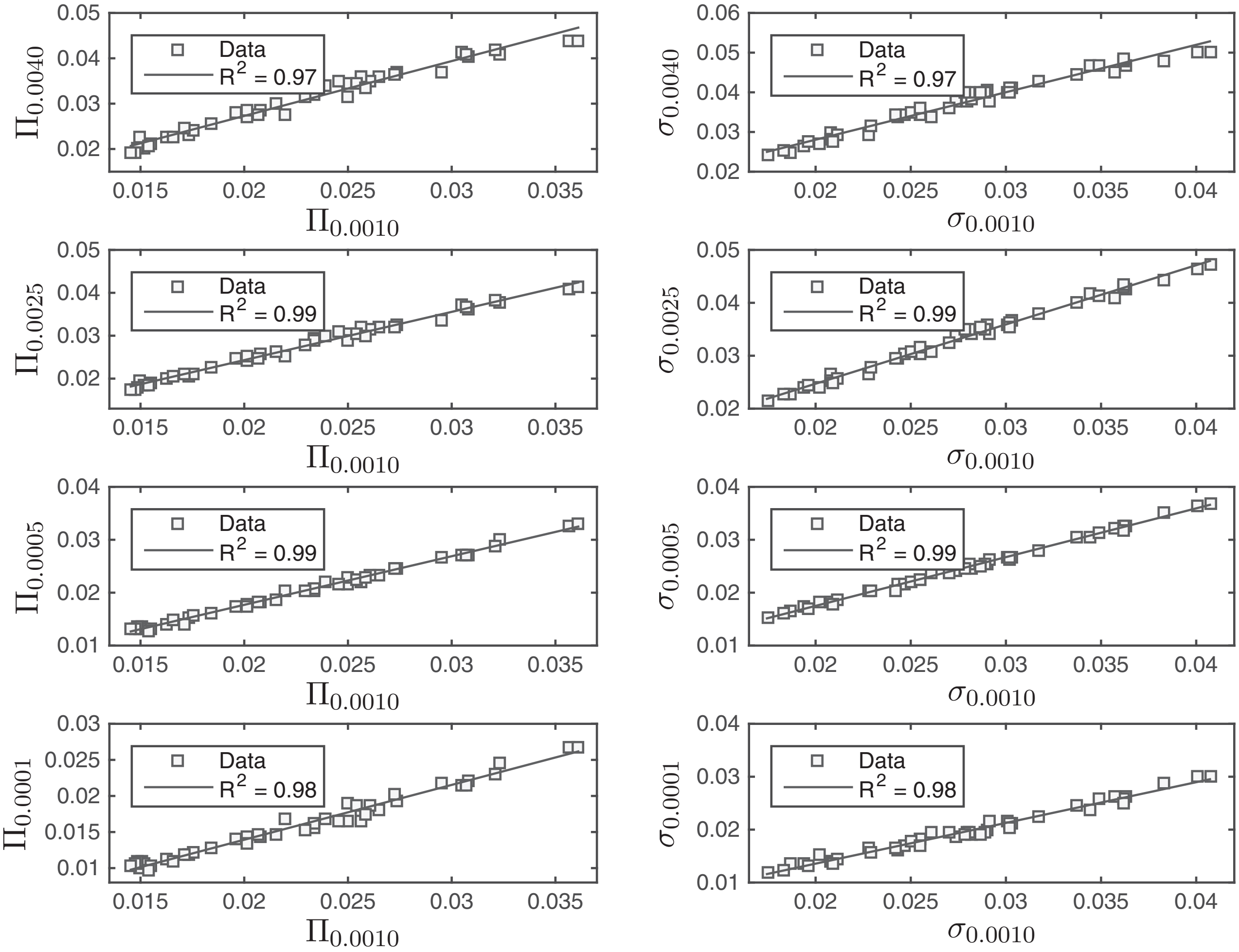


Figure 2. The parameter values of Π and σ on the isosurfaces $\mathscr{S}_{0.0001}$, $\mathscr{S}_{0.0005}$, $\mathscr{S}_{0.0010}$, $\mathscr{S}_{0.0025}$ and $\mathscr{S}_{0.0040}$. Both Π and σ are given in units of Hartree.

Having shown that measures of imbalance fail to satisfy surface-invariance, we proceed to ascertain the severity of the inconsistency, i.e. to analyse to what extent it is expected to cause different predictions on different surfaces. Observe that from the lower right plot of Figure 3 one readily identifies two compounds A and B for which, approximately,

$$\begin{aligned} \Theta_{0.0010}(A) &= 0.5, \quad && \Theta_{0.0010}(B) = 0.5, \\ \Theta_{0.0001}(A) &= 0, && \Theta_{0.0001}(B) = 1. \end{aligned} \tag{12}$$

Although these molecules have identical Θ values on $\mathscr{S}_{0.0010}$, their values differ markedly on $\mathscr{S}_{0.0001}$. The difference is significant, as is confirmed from the fact that $\Theta \in [0, 2]$ on $\mathscr{S}_{0.0001}$. Thus, $\Delta\Theta = 1$ covers half the range of Θ values in the study, clearly causing inconsistent predictions for A and B in models of the form property $= \alpha\,\Theta + \gamma$. In terms of impact sensitivity, A and B will be of equal and small sensitivity according to $\mathscr{S}_{0.0010}$, whereas A will be stable and B moderately sensitive on $\mathscr{S}_{0.0001}$. We conclude that the use of imbalance parameters in linear models will result in significantly inconsistent predictions.

Interestingly, while measures of imbalance do not satisfy the linearity of Equation (5), the skewness s' does, see Figure 4. The problem with imbalance is, thus, revealed to be that it treats left- and right-skewness on an equal footing. Both of them signify imbalance.

Whereas imbalance produces non-linearities, the measures of spread satisfy the linearity to a high degree ($R^2 \geq 0.97$). The implication is that this aspect of $\mathscr{V}$ is not inconsistent with its use in linear models. These results do not, however, imply that variation is relevant for predicting chemical or physical properties, only that applying it does not lead to any contradictions.

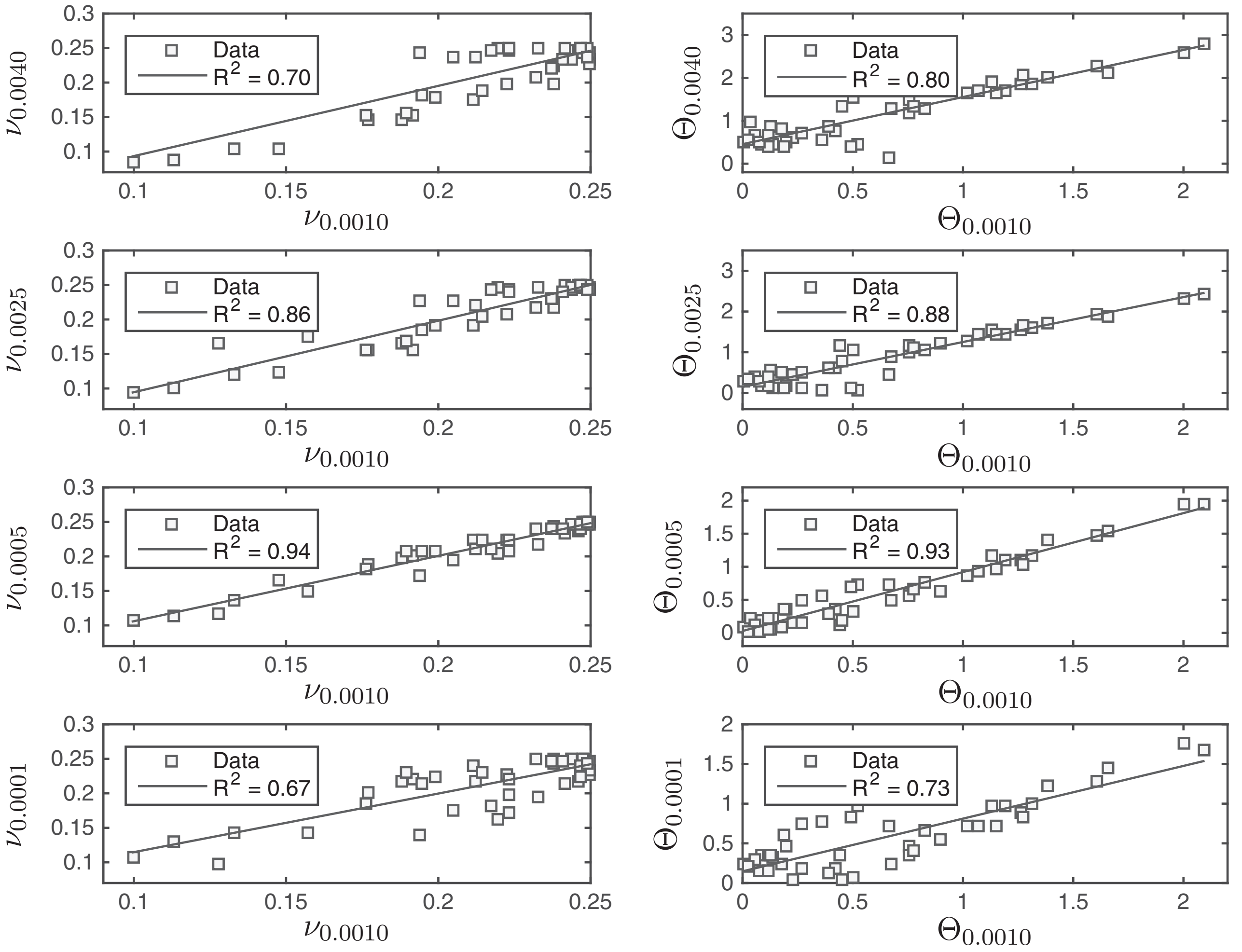


Figure 3. The parameter values of ν and Θ on the isosurfaces $\mathscr{S}_{0.0001}$, $\mathscr{S}_{0.0005}$, $\mathscr{S}_{0.0010}$, $\mathscr{S}_{0.0025}$ and $\mathscr{S}_{0.0040}$.

Considering the choice of electron density range, we note the similarity and proximity of $\mathscr{S}_{0.0001}$ and $\mathscr{S}_{0.0040}$, see Figure 1. Moreover, the trends shown by the electrostatic potential suggest that the information contained in the parameters of variation (σ, Π) and skewness (s') is preserved in the electron density range studied. Murray *et al.* have given a similar statement for the 0.0010 to 0.0050 a.u. range [26]. The appropriate values for $\epsilon_{\min}$ and $\epsilon_{\max}$ remain, however, an open question that will require more research to settle.

Earlier studies have considered the invariance of predictions on different molecular surfaces. Murray *et al.* [27] studied a model for the hydrogen-bond-acceptor parameter β. The molecular electrostatic spatial minima $V_{S,\min}$ and maxima $V_{S,\max}$ were used as parameters. By analysing their data (see tables 1 and 2 in [27]), we found that these parameters satisfy the principle of molecular surface invariance in Equation (5) on the three surfaces $\mathscr{S}_{0.0010}$, $\mathscr{S}_{0.0020}$ and $\mathscr{S}_{0.0050}$. Murray *et al.* [26] investigated the use of local ionisation energy as a tool for identifying reactive sites on defect-containing model graphene systems. The positions at which the local ionisation energy has a minimum were of particular interest since these sites are indicative of the locations of the least tightly bound, most reactive electrons. Considering their data (see table 4 in [26]), we found that the surface-minimum of the local ionisation energy $\bar{I}_{S,\min}$ fulfills the principle on $\mathscr{S}_{0.0010}$, $\mathscr{S}_{0.0030}$ and $\mathscr{S}_{0.0050}$.

Instead of focusing on particular models (with one or more parameters), our approach is to consider the parameters themselves and ask whether the predictions of any model using these parameters will be similar on different surfaces. The advantage of this approach is that it allows one to rigorously weed out parameters that should probably be avoided in all models. Constructing new models with parameters that satisfy molecular surface invariance is necessary for the predictions to remain the same on a range of surfaces.

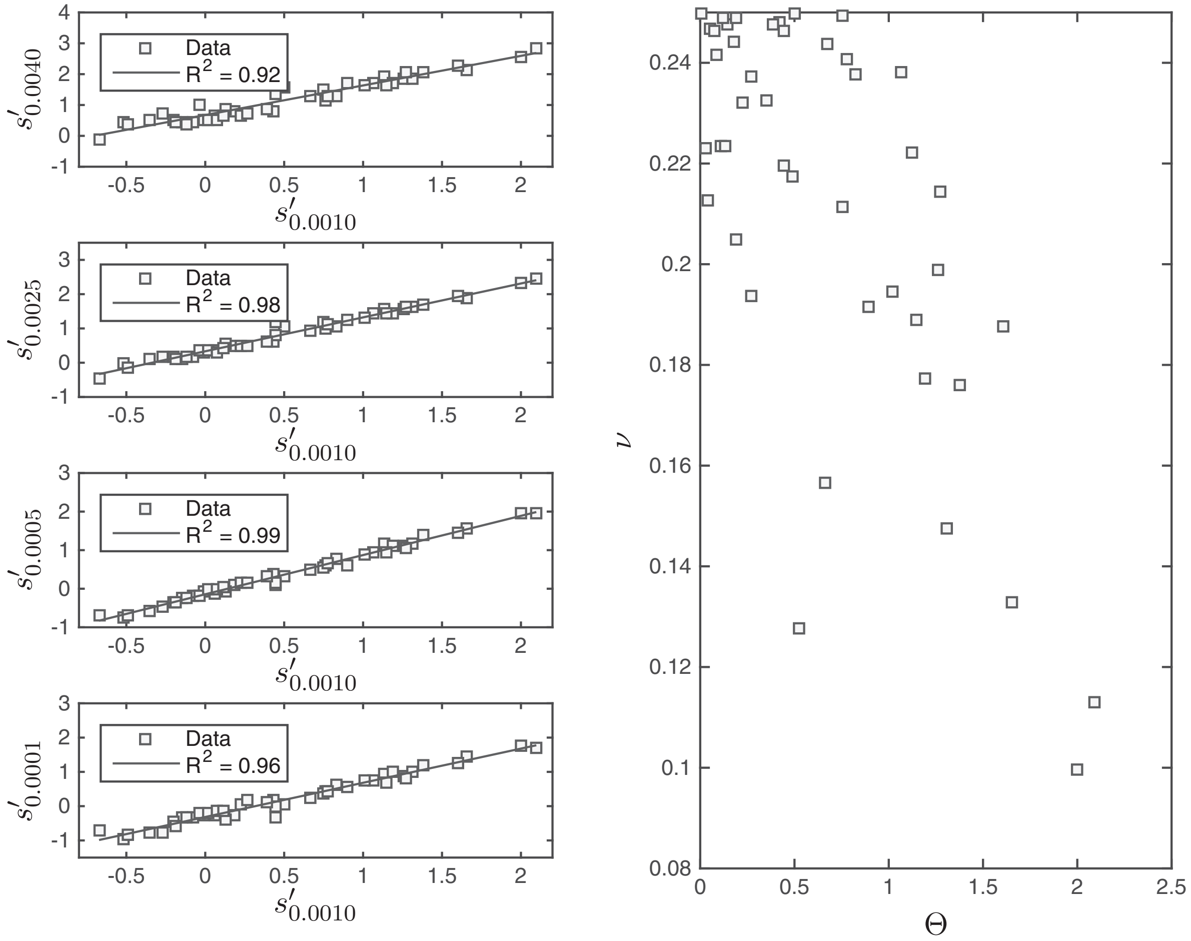


Figure 4. Left: the parameter values of s' on the isosurfaces $\mathscr{S}_{0.0001}$, $\mathscr{S}_{0.0005}$, $\mathscr{S}_{0.0010}$, $\mathscr{S}_{0.0025}$ and $\mathscr{S}_{0.0040}$. Right: the imbalance quantities Θ and ν.

4. Conclusions and summary

The arbitrary choice of the molecular surface has its pitfalls, especially in the context of modelling. While it is reasonable to define such a surface by regions of constant and small electron density, one is still left with having to choose a particular value. An essential question is how improbable the detection of an electron should be in order to be on the 'surface'. We take the view that there must be many equivalent choices, all equally well suited to represent the surface.

An appealing idea is to study the electrostatic potential on the molecular surface. Perhaps information of intermolecular interactions may be extracted, perhaps this information is relevant for a range of physical and chemical properties. Pursuing this idea, several workers have suggested models of the form

$$\text{property} = \alpha\ \Phi(\epsilon) + \beta, \tag{13}$$

where $\Phi(\epsilon)$ is the value of Φ on the surface of electron density ϵ. (These models often have several parameters Φ, but this is not important in the present discussion.) These efforts have in common that the importance of the particular choice of electron density ϵ has been overlooked. Our main claim is that there exists a range of electron densities whose corresponding surfaces are all equivalent representations of the molecular surface. Consequently, no model should make different predictions on different surfaces.

We proved that if linear single-variable models are to make the same predictions on every molecular surface, $\Phi(\epsilon)$ must be linearly related to $\Phi(\epsilon')$, where ϵ and ϵ' are electron densities corresponding to valid molecular surfaces. If this linearity is violated, no model of the form of Equation (13) is able to yield the same predictions on the two surfaces.

To illustrate the use of this requirement, we have demonstrated that imbalance, i.e. the predominance of negative or positive electrostatic potential on the surface, violates the linearity requirement in the electron density range [0.0001 a.u., 0.0040 a.u.]. Thus, any linear single-variable model applying this aspect will not produce consistent results on surfaces within this range. Moreover, we have demonstrated that the differences in predictions will in many cases be significant, and hence, that the violation of linearity is not merely a mathematical nuance. We also argue that multi-variable models applying imbalance should be avoided, reasoning that consistent predictions in this case will require highly non-trivial cancellations between the parameters making up the model. In the literature, linear models applying imbalance have been proposed to predict enzyme inhibition concentrations [18], impact energies [10], potencies of anti-HIV agents [17] and solubilities [4]. Our work indicates that none of these models should be used, since their predictions are expected to become inconsistent with slight changes in the molecular surface. This in turn suggests that the observed trends in the aforementioned studies might be coincidental. In another study, we have shown that this appears to be the case in the prediction of crystal densities [23].

In our consideration of imbalance and variation parameters, we have implicitly assumed that the range of electron densities [0.0001 a.u., 0.0040 a.u.] adequately represents the molecular surface. This appears to be a valid assumption, given that the information contained in the variation parameters (Π, σ) and the skewness parameter (s') is preserved throughout the range of electron densities.

We recommend that all molecular surface-based parameters should be checked for linearity, in the sense of Equation (5), before being considered as a relevant quantity in linear modelling of physical and chemical properties. In these modelling efforts, better predictive power may be obtained by using model parameters that are invariant to the change of molecular surface. As a final remark, we emphasise that these considerations are not restricted to the use of the electrostatic potential, but remain valid for any parameter obtained from the surface of molecules.

Note

1. We used the unbiased estimator for σ in our calculations, dividing by $N - 1$ instead of N. Since $N > 4000$ in all calculations, this makes no difference. We wrote N for readability.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] P. Politzer and J.S. Murray, Theor. Chem. Acc. **108** (3), 134–142 (2002).

[2] R.F.W. Bader, M.T. Carroll, J.R. Cheeseman, and C. Chang, J. Am. Chem. Soc. **109** (26), 7968–7979 (1987).

[3] A.F. Jalbout, Z.Y. Zhou, X. Li, M. Solimannejad, and Y. Ma, Comput. Theor. Chem. **664–665**, 15 (2003).

[4] J.S. Murray, T. Brinck, P. Lane, K. Paulsen, and P. Politzer, Comput. Theor. Chem. **307**, 55 (1994).

[5] P. Politzer, J. Martinez, J.S. Murray, M.C. Concha, and A. Toro-Labbé, Mol. Phys. **107** (19), 2095–2101 (2009).

[6] B.M. Rice, J.J. Hare, and E.F.C. Byrd, J. Phys. Chem. A **111** (42), 10874–10879 (2007).

[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] H. Hagelin, J.S. Murray, P. Politzer, T. Brinck, and M. Berthelot, Can. J. Chem. **73** (4), 483–488 (1995).

[10] J.S. Murray, M.C. Concha, and P. Politzer, Mol. Phys. **107** (1), 89–97 (2009).

[11] J.S. Murray, P. Lane, and P. Politzer, Mol. Phys. **93** (2), 187–194 (1998).

[12] P. Politzer, J.S. Murray, P. Lane, and T. Brinck, J. Phys. Chem. **97** (3), 729–732 (1993).

[13] R. Dennington II, T. Keith, and J. Millam, GaussView Version 4.1 (Semichem, Inc. Shawnee Mission, KS, 2007).

[14] J.S. Murray, P. Lane, T. Brinck, K. Paulsen, M.E. Grice, and P. Politzer, J. Phys. Chem. **97** (37), 9369–9373 (1993).

[15] P. Politzer, P. Lane, J.S. Murray, and T. Brinck, J. Phys. Chem. **96** (20), 7938–7943 (1992).

[16] T. Brinck, J.S. Murray, and P. Politzer, J. Org. Chem. **58** (25), 7070–7073 (1993a).

[17] P. Politzer, J.S. Murray, and Z. Peralta-Inga, Int. J. Quantum Chem. **85** (6), 676–684 (2001).

[18] T. Brinck, P. Jin, Y. Ma, J.S. Murray, and P. Politzer, J. Mol. Model. **9** (2), 77–83 (2003).

[19] T. Brinck, J.S. Murray, and P. Politzer, Int. J. Quantum Chem. **48** (2), 73–88 (1993).

[20] R. Wehrens, *Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences* (Springer, Heidelberg, 2011).

[21] K.P. Burnham and D.R. Anderson, Sociol. Methods Res. **33** (2), 261–304 (2004).

[22] T. Brinck, J.S. Murray, and P. Politzer, Mol. Phys. **76** (3), 609–617 (1992).

[23] E.F. Kjønstad, J.F. Moxnes, T.L. Jensen, and E. Unneberg, Mol. Phys., (2016) (submitted).

[24] 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).

[25] 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).

[26] J.S. Murray, Z.P. Shields, P. Lane, L. Macaveiu, and F.A. Bulat, J. Mol. Model. **19** (7), 2825–2833 (2013).

[27] J.S. Murray, T. Brinck, M.E. Grice, and P. Politzer, J. Mol. Struct. THEOCHEM **256**, 29 (1992).