

STRUCTURE OF MATTER AND QUANTUM CHEMISTRY

Dedicated to my dear teacher Dr. Alexander I. Shtern on occasion of his 80th birthday

Quantum Metrics for Continuous Shape Measures of Molecules

A. L. Tchougréeff^{a,*}

^a*A.N. Frumkin Institute of Physical Chemistry and Electrochemistry of RAS, Moscow, Russia*

**e-mail: andrei.tchougreeff@gmail.com*

Received October 10, 2020; revised November 19, 2020; accepted December 10, 2020

Abstract—The field of continuous molecular shape and symmetry (dis)similarity quantifiers habitually called measures (specifically continuous shape measures (CShM) or continuous symmetry *measures* (CSM)) is obfuscated by the combinatorial numerical algorithms used in the field which restricts the applicability to the molecules containing up to twenty equivalent atoms. In the present paper we analyze this problem using various tools of classical probability theory as well as of one-particle and many-particle quantum mechanics. Applying these allows us to lift the combinatorial restriction and to identify the adequate renumbering of atoms (vertices) without considering all $N!$ permutations of an N -vertex set so that in the end purely geometric molecular shape (dis)similarity quantifier can be defined. Developed methods can be easily implemented in the relevant computer code.

DOI: 10.1134/S0036024421090272

1. INTRODUCTION

Since the introduction of continuous shape [1, 2] and symmetry [3] measures¹ (CShM and CSM, respectively) for molecules treated as sets of colored vertices in $\mathbb{R}^3$ in early 90-ies, this field of research has won a considerable attention as witnessed by the recent review [4] featuring 157 references of which a good fraction is devoted to various chemical applications. This interest is stipulated by the proposed new approach to symmetry and/or to the concept of molecular shape as not a binary *quality* which may either be present or absent rather as a continuous *quantity* which may be present or absent to a degree expressed by a real number. This approach wins a benevolence in the chemical community since strictly symmetric molecules are really rare whereas those which are in various vaguely defined senses “close” to symmetric ones dominate the class of the molecules considered as “symmetric” in some approximate or pragmatic sense. Thus, a possibility to somehow quantify the deviation from a symmetric pattern is in a high demand. Analogous considerations stipulate the interest to the shape measures which may help to decide how much a given trial set of vertices in $\mathbb{R}^3$ is similar (or not) to, say, (set of vertices of a) cuboctahedron. Although, extended recently to vectors, matrices,

operators and functions [5] the central question in this area of research is establishing CShM and/or CSM of a set of vertices in $\mathbb{R}^3$.

Despite that high interest and considerable effort invested in studying this problem in the course of the elapsed quarter of a century this research field is obfuscated by the atoms’ labeling problem which significantly restricts its practical applicability [4]. The source of the problem is as follows. Assume, someone wants to find out how much (dis)similar a (shape of a) figure made of three vertices in a plane differs from the perfect triangle (Fig. 1). According to [1–4, 6] the quantity to be calculated to find it out (in more practical situation of N -vertex sets in $\mathbb{R}^3$) reads:

$$\min \frac{1}{N} \sum_{k=1}^N |\mathbf{T}_k - \mathbf{P}_k|^2, \quad (1)$$

where $\mathbf{T}_k$ and $\mathbf{P}_k$ are the radii-vectors of the vertices in the studied (trial) and pattern sets, respectively, and by min it is understood that a set of allowed transformations $\Omega(\omega)$ —translations, rotations and uniform scaling—is applied to each $\mathbf{T}_k$. The pitfall of this definition is a tacit requirement of a *correspondence* between the labels used in the studied and the pattern sets as expressed by the *equal* subscripts of the radii-vectors in Eq. (1). This is, of course, an obstacle since nobody can guarantee that in a practical situation an arbitrary

¹ They are not *measures* in the strict mathematical sense of this word, like area, volume or probability, rather more precisely can be dubbed as *quantifiers*.

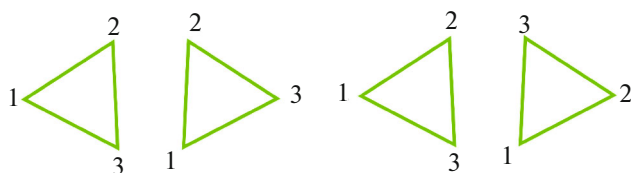


Fig. 1. Two different numberings of equilateral triangles with unity side length, one of which (right) representing problem for the comparison of their shapes or symmetries.

accepted numbering suits the purpose. Apparently, operating by only rotations on the triangles with the numbering shown on the right of Fig. 1 will never allow to put the vertices in a due correspondence so that the calculated quantifier—(squared) Euclidean distance between two *equal* triangles will be always positive and actually quite large—equal to two. This result: having a nonvanishing dissimilarity quantifier for coinciding figures, is counterintuitive and needs to be eliminated.

In general a *shape* is an entity invariant with respect to continuous—shift, rotation and scaling transformations, but as well to discrete ones—permutations of labels of its vertices. In the available algorithms the latter is purchased pretty dearly: as soon as an unfortunate labeling or numbering accidentally appears,² an alternative numbering, rectifying the situation, is recovered by a brute force search, ultimately leading to the $N!$ computational costs. The authors of [4, 6] point out the combinatorial nature of the problem of constructing and selecting the allowed permutations which requires the time proportional to $N!$ so that the exact computation of CShMs (CSMs) is not feasible for structures with more than 20 vertices.

Clearly, the pivot of the problem is that computed numerical value of quantifiers proposed so far heavily depends on either voluntary established or involuntary emerging correspondence between the (numerical or whatever) *labels* assigned to the vertices of the respective figures. If, by chance, the vertices in the *trial* and *pattern* figures bearing the *same* labels turn out to be close to each other, the problem solves easily, and a sensible value of the CShM/CSM recovers under the action of allowed transformations—(not very large) rotations of $\mathbb{R}^3$ and some scaling. If this is not the case (and nobody can guarantee the opposite) the value accessible under the optimization with respect to parameters of the allowed transformations of the shapes to be compared are bound from *below* as it happened to the triangles on the right of Fig. 1.

The way out proposed in the present paper is to put to work the objects possessing the full symmetry of the problem, that is, also with respect to all permutations

² Among $N!$ permutations it is pretty easy to pick a wrong one since only *one* is correct.

of the labels of vertices. When it comes to chemical applications the most obvious equivalence is that of the atoms of the same element (those with the same atomic number or, *equivalently*, nuclear charge). This is, however, not mandatory and one may equally take different isotopes of the same element as differently colored vertices or in opposite to consider, say, elements of the same group to be interchangeable or even consider as such simply the atoms with the same coordination numbers, thus opening doors to treating as equivalent structures of water ices, silica polymorphs and allotropes of the group IV elements [7]. This all is only to say that treating items as interchangeable, i.e., equivalent or not is the matter of our free discretion.

2. CONCEPT AND HEURISTIC CONSIDERATIONS

In general, N vertices in $\mathbb{R}^3$ $\{\mathbf{Q}_k \in \mathbb{R}^3 | k = 1-N\}$ form a *shape* Q . It remains the same if all N vertices are shifted by a vector $\mathbf{S}$: $\mathbf{Q}_k \Rightarrow \mathbf{S} + \mathbf{Q}_k$, $\mathbf{S} \in \mathbb{R}^3$, or radius-vector of each vertex is rotated: $\mathbf{Q}_k \Rightarrow \mathcal{R}\mathbf{Q}_k$ by the same rotation $\mathcal{R} \in SO(3)$, or radius-vector of each vertex is multiplied by a positive number κ : $\mathbf{Q}_k \Rightarrow \kappa\mathbf{Q}_k$; $\kappa \in \mathbb{R}$, $\kappa > 0$.³ In order to find out whether and to what extent the trial set $T = \{\mathbf{T}_k \in \mathbb{R}^3 | k = 1-N\}$ is similar to the pattern set $P = \{\mathbf{P}_k \in \mathbb{R}^3 | k = 1-N\}$ the vectors in T are allowed to be changed all at a time by the three above transformations (shift, rotation and scale) and the minimum of the sum of the (squared) Euclidean distances between the vertices of these shapes is used as the CShM Eq. (1).

The number of variables of the transformation to be applied to the trial set for continuous optimization can be reduced. E.g., it has been suggested [5] to superimpose two sets so that their (geometric) centers given by:⁴

$$\mathbf{GC}_P = \frac{1}{N} \sum_k \mathbf{P}_k, \quad \mathbf{GC}_T = \frac{1}{N} \sum_k \mathbf{T}_k \quad (2)$$

coincide, that is to assume the coordinates of the vertices in the pattern set to be given relative to the origin of the coordinate frame and to set the shift applied to the trial shape to be $\mathbf{S} = \mathbf{GC}_P - \mathbf{GC}_T$, namely, to put the center of the trial set in the origin of the coordinates as well. This transfers the consideration from the affine $\mathbb{R}^3$ space featuring two types of objects—points

³ Positiveness of κ is important as is of $\mathcal{R}$ being a *proper* rotation, otherwise a *chiral* shape could be superimposed with its *enantiomer*.

⁴ This slightly modifies the suggestion of [5] where it has been proposed to superimpose centers of mass of the molecules to be compared. As far as it goes about the shapes formed by equivalent vertices this definition gives the same result as one of [5].

and vectors to the vector $\mathbb{R}^3$ space [8]. After doing this, we can think that the shapes T and P are represented by whole rays protruding from the origin of the direct sum $\bigoplus_{i=1}^N \mathbb{R}^3$ of N copies of the $\mathbb{R}^3$ vector space, irrespective to nonvanishing scaling factors κ^5 and/or rotations $\mathcal{R}$ applied simultaneously to all 3-dimensional components of the direct sum. This improvement, however, does not solve the combinatorial problem: the shapes are invariant not only with respect to scaling and other elementary geometry transformations, but also with respect to permutations of the vertex labels. The latter symmetry does not hold for vectors (or rays) in $\bigoplus_{i=1}^N \mathbb{R}^3$ representing the shapes. Permuting the coordinates of a vector or the components of the direct sum makes a different vector, but not a different shape. Thus, some more involved instrumentation is needed.

2.1. Classical Statistics Based Quantifier

The pattern and trial shapes P and T can be put in one-to-one correspondence with vertex (or particle) densities in $\mathbb{R}^3$:

$$\rho_P(\mathbf{R}) = \sum_k \delta(\mathbf{R} - \mathbf{P}_k), \quad (3)$$

being sums of δ -distributions [9] and analogous expression for T . Hereinafter $\mathbf{R}$ is a point in $\mathbb{R}^3$ where the density is considered (measured or probed). Of paramount importance is that the labeling (numbering) of the vertices in either set is now immaterial: summation is commutative, thus the order in which the vertices $\mathbf{P}_k$ (or $\mathbf{T}_j$) appear in Eq. (3) does not matter.

A quantifier of a difference between densities has been proposed by Carbó and coworkers in a slightly different context as the integral over $\mathbb{R}^3$ of the squared difference of the two densities:

$$\varepsilon_{PT} = \int d^3\mathbf{R} |\rho_P(\mathbf{R}) - \rho_T(\mathbf{R})|^2, \quad (4)$$

that is, as the (squared) distance in the sense of the $L^2(\mathbb{R}^3)$ space⁶ [10]. It apparently has that advantage that as soon as the positions of all N vertices in the sets to be compared pairwise coincide *irrespective to their labels*, this quantity precisely vanishes provided one, first, takes the difference, and then squares it (modulo is unnecessary since the densities are real) and then takes the integral of the vanishing integrand. However, as everybody knows δ -distribution is not a usual function, so, mathematically, the story evolves not so smoothly, since an attempt to get the same result, but

first taking the square, then integrating the terms and then bookkeeping the contributions stumbles on the square of the δ -distribution, on which nobody knows what it is.

The ε_{PT} based approach can be set back on a track by replacing the δ -distributions by the normal ones centered at the vertices of P and T with an arbitrary, but small dispersion σ ,⁷ assumed to be equal for all vertices. By this we arrive to:

$$\begin{aligned} \rho_T^{(\sigma)}(\mathbf{R}) &= \sum_k \delta^{(\sigma)}(\mathbf{R} - \mathbf{P}_k) \\ &= \frac{1}{\sqrt{8\pi^3}\sigma^3} \sum_k \exp\left(-\frac{(\mathbf{R} - \mathbf{P}_k)^2}{3\sigma^2}\right) \end{aligned} \quad (5)$$

and analogously for $\rho_P^{(\sigma)}(\mathbf{R})$.

Inserting this in Eq. (4) immediately yields the (squared) $L^2(\mathbb{R}^3)$ distance between the sums of blurred δ -distributions as expressed through the finite quantities $W_{PT}^{(\sigma)}$, $W_{PP}^{(\sigma)}$, $W_{TT}^{(\sigma)}$ [10] and baptized the (molecular quantum) similarity and self similarity measures:

$$\begin{aligned} \varepsilon_{PT}^{(\sigma)} &= W_{PP}^{(\sigma)} + W_{TT}^{(\sigma)} - 2W_{PT}^{(\sigma)} \\ &= \int d^3\mathbf{R} [\rho_P^{(\sigma)}(\mathbf{R})]^2 + \int d^3\mathbf{R} [\rho_T^{(\sigma)}(\mathbf{R})]^2 \\ &\quad - 2 \int d^3\mathbf{R} \rho_P^{(\sigma)}(\mathbf{R}) \rho_T^{(\sigma)}(\mathbf{R}), \\ W_{PP}^{(\sigma)} &= \sum_{kk'} \int d^3\mathbf{R} \delta^{(\sigma)}(\mathbf{R} - \mathbf{P}_k) \delta^{(\sigma)}(\mathbf{R} - \mathbf{P}_{k'}), \\ W_{TT}^{(\sigma)} &= \sum_{jj'} \int d^3\mathbf{R} \delta^{(\sigma)}(\mathbf{R} - \mathbf{T}_j) \delta^{(\sigma)}(\mathbf{R} - \mathbf{T}_{j'}), \\ W_{PT}^{(\sigma)} &= \sum_{kj} \int d^3\mathbf{R} \delta^{(\sigma)}(\mathbf{R} - \mathbf{P}_k) \delta^{(\sigma)}(\mathbf{R} - \mathbf{T}_j), \end{aligned} \quad (6)$$

reducing to overlaps⁸ $\langle \delta_Q^{(\sigma)} | \delta_{Q'}^{(\sigma)} \rangle$ of the normal distributions with dispersion σ centered at the vertices $\mathbf{Q}$, $\mathbf{Q}'$. These latter evaluate explicitly (see e.g., [11]):

$$\langle \delta_Q^{(\sigma)} | \delta_{Q'}^{(\sigma)} \rangle = \frac{1}{8\pi^{3/2}\sigma^3} \exp\left(-\frac{(\mathbf{Q} - \mathbf{Q}')^2}{4\sigma^2}\right).$$

For $\mathbf{Q}' \neq \mathbf{Q}$ and small enough σ ($\ll |\mathbf{Q}' - \mathbf{Q}|$) the overlaps rapidly decay as σ decreases. By contrast, for the same estimate of the otherwise arbitrary dispersion, the terms with $\mathbf{Q}' = \mathbf{Q}$ contribute each $\frac{1}{8\pi^{3/2}\sigma^3}$, which

⁷ An inverse of the trick known yet to Kirchhoff and Helmholtz who used a limit of Gaussians to represent an identity impulse –

the δ -“function”: $\delta(x) = \lim_{\sigma \rightarrow 0} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right)$ to be used only under integral over x .

⁸ $L^2(\mathbb{R}^3)$ scalar products for which we use the standard quantum theoretical bra-ket notation (see e.g., [5, 11–13]).

⁵ Size doesn't matter. Irrespective to it, cube remains a cube.

⁶ That of the functions integrable with their squares.

diverges as $\sigma \rightarrow 0$. The strong dependence of $\varepsilon_{PT}^{(\sigma)}$ on σ in the sensitive range of small σ 's is an unpleasant feature which needs to be avoided. In the context of the Carbó's approach a way around is available. Specifically, the similarity measure W_{PT} is divided by the square root of the product of W_{PP} and W_{TT} .⁹ This defines so called *Carbó index* of a pair of distributions:

$$c_{PT}^{(\sigma)} = \frac{W_{PT}^{(\sigma)}}{\sqrt{W_{PP}^{(\sigma)} W_{TT}^{(\sigma)}}}. \quad (7)$$

For small enough σ , the self-similarities $W_{PP}^{(\sigma)}$ and $W_{TT}^{(\sigma)}$ are both close to $\frac{N}{\sqrt{8\pi^3}\sigma^3}$ and so is the square root of

their product. Dividing $W_{PT}^{(\sigma)}$ by that latter, apparently, removes the divergence while $\sigma \rightarrow 0$. Then at a fixed (small enough σ), the Carbó indici for pairs of individual normal distributions $\delta_Q^{(\sigma)}$ entering Eq. (5) centered at the vertices of either set rapidly decay for the “distant” vertices $|\mathbf{Q}_{k'} - \mathbf{Q}_k| \gg \sigma$ and quadratically flow to unity for the “close” vertices $|\mathbf{Q}_{k'} - \mathbf{Q}_k| \ll \sigma$ (see Fig. 3). For the comparison of a trial and pattern sets of vertices this means that the Carbó index of two sums of normal distributions flows to unity for coinciding trial and pattern densities (apparently referring to the same shape) and to zero for those which are $L^2(\mathbb{R}^3)$ -orthogonal.

Here, as well, the labeling of the vertices in either set is immaterial: it is all the same what label bears the vertex of the trial set which comes close to a given vertex in the pattern one—the only condition to be fulfilled is that the Euclidean distances between the vertices *within* either set are larger than σ . If so, then in a σ -vicinity of a vertex of one set can occur not more than *one* vertex of another set.¹⁰ Thus for the pair of trial and pattern sets which almost coincide, that is *all* vertices of the pattern set have *someone* vertex of the trial set in their σ -vicinity $\sigma > |\mathbf{P}_k - \mathbf{T}_j|$, the amount of the Carbó index approximately is:

$$c_{PT}^{(\sigma)} \sim \frac{1}{N} \sum_m \left[1 - \frac{1}{4\sigma^2} (\mathbf{T}_m - \mathbf{P}_m)^2 \right], \quad (8)$$

where the subscript m refers to properly renumbered vertices. For the pair of the sets, which contain pairs of vertices with $|\mathbf{P}_k - \mathbf{T}_j| \gg \sigma$ for *all* k, j , the summation in the above equation extends to the vertices in the σ -vicinity of each other only.

We see that Carbó indici of *all* pairs of the Gaussian densities Eq. (5) centered in the pattern and trial sets

⁹ Squared $L^2(\mathbb{R}^3)$ norms of $\rho_P^{(\sigma)}$ and $\rho_T^{(\sigma)}$ considered as its elements. It naturally applies only to distributions which are representable by the functions integrable with their squares.

¹⁰ The direct consequence of the triangle inequality.

act as a filter in the style of [14] allowing by an appropriate selection of σ to single out for each vertex in the pattern set one and only one vertex in the trial set at the closest separation between them. This holds, of course for the pattern and trial sets being in a *general position* relative to each other. For those accidentally having two vertices in the trial set at an equal distance from a given vertex in the pattern set,¹¹ a small stirring will destroy this degeneracy. Pragmatically, the problem reduces to a calculation of the $N \times N$ matrix of vertex-vertex (squared) *distances* with, say, rows corresponding to the vertices in the pattern set and columns to those in the trial set so that in each row labeled by k there is a unique column labeled by j , yielding $\min_j (\mathbf{P}_k - \mathbf{T}_j)^2$ and, respectively,

$\max_k \left[\exp \left(-\frac{(\mathbf{P}_k - \mathbf{T}_j)^2}{4\sigma^2} \right) \right]$. If we replace in each row the minimal (resp., maximal) element by unity and all others by zeroes we arrive to a *permutation* matrix of the *labels* assumed in Eq. (1), but purchased with the computational price of only $O(N^2)$ instead of factorial, by this lifting the combinatorial restriction mentioned in the Introduction.

The Carbó index Eq. (7) can be interpreted as the cosine of an angle between the rays in $L^2(\mathbb{R}^3)$ defined by $\rho_P^{(\sigma)}$ and $\rho_T^{(\sigma)}$. Taking its arccosine, we arrive to a true quantifier of the difference between the trial and pattern densities:

$$\gamma(\rho_P^{(\sigma)}, \rho_T^{(\sigma)}) = \arccos \rho_{PT}^{(\sigma)}, \quad (9)$$

being a special case of the Fubini-Study metric [15] which by definition satisfies the axioms describing distance. Inserting the estimate of the Carbó index Eq. (8) in the asymptotic expansion for the arccosine in the vicinity of its unity argument, we obtain:

$$\gamma(\rho_P^{(\sigma)}, \rho_T^{(\sigma)}) \sim \sqrt{\frac{\sum_m (\mathbf{T}_m - \mathbf{P}_m)^2}{2N\sigma^2}},$$

which following the tradition of taking the CShM/CSM's as squares of the corresponding distances yields:

$$\gamma^2(\rho_P^{(\sigma)}, \rho_T^{(\sigma)}) \sim \frac{\sum_m (\mathbf{T}_m - \mathbf{P}_m)^2}{2N\sigma^2},$$

by this regaining the description of the difference between the shapes in terms of the Euclidean distances between their vertices. Furthermore, redefining the shape dissimilarity quantifier as

$$2N\sigma^2\gamma^2(\rho_P^{(\sigma)}, \rho_T^{(\sigma)}), \quad (10)$$

¹¹ $(\mathbf{P}_k - \mathbf{T}_j)^2 = (\mathbf{P}_k - \mathbf{T}_{j'})^2; j \neq j'$.

we see that for small σ it is only weakly dependent on σ and flows to the sum of Euclidean distance between the properly renumbered vertices of the trial and pattern sets, basically being that what is needed in Eq. (1).

One more remark refers to the precision/sensitivity of the introduced quantifiers. Apparently, in the present setting either is controlled by the dispersion σ . For small values of σ the sensitivity of the quantifier, Carbó index, is very high—it readily responds to the small deviation of the trial vertex set from the pattern set. However, it is not precise: as soon as one of the vertices of the trial set leaves the σ -vicinity of its pattern counterpart it does not significantly increase further no matter how far the vertex moved, so that very close values of the quantifier may correspond to very different large Euclidean distances between the vertices in the coordinate space. By contrast, if the dispersion is large, for each given vertex in the pattern set the overlaps between the normal distribution in it and those in several vertices in the trial set may noticeably contribute to the quantifier. In plain words smaller σ makes the proposed quantifier closer to strict mathematical definition of a shape. Although, apparently the contribution of the closest vertices in two sets will be dominating for either small or large σ it is not clear a priori, how a numerical optimization procedure is going to behave in the latter case.

3. QUANTUM STATISTICS BASED (DIS)SIMILARITY QUANTIFIERS

The heuristic moves of Section 2.2.1 were based on the representation of the sets of vertices in $\mathbb{R}^3$ by density distributions. The name of Molecular Quantum Similarity Measure for the quantity Eq. (4) has been coined in [10] since, originally, the compared (electron) densities were assumed to come from a *quantum* chemical calculations of electronic structure. This wording is, however, not 100% precise, since fundamentally the used quantities are *classical* probability densities: the superposition principle of production of new probability densities from several given ones $\{\pi_k(\mathbf{R}) | k = 1-N\}$ is *classical*:

$$\frac{\sum_k p_k \pi_k(\mathbf{R})}{\sum_k p_k}; \quad p_k, \pi_k(\mathbf{R}) \geq 0, \quad \int d^3\mathbf{R} \pi_k(\mathbf{R}) = 1;$$

that is, the probabilities sum themselves [16]. This way is internally consistent (removing $\sum_k p_k$ from the denominator as adopted in Eqs. (3) or (5) changes merely the overall normalization); it applies even to singular distributions like δ -ones, but leads to problems with similarity quantifier defined through Eq. (4).

Transition to the Carbó index Eq. (7) scales the probability densities by their $L^2(\mathbb{R}^3)$ -norms analogously to [5]. The question is then what these rescaled subjects are? The rule of deriving probability densities from them is characteristic for *wave functions* (see e.g., [12]), alternatively dubbed probability density *amplitudes*: not the densities themselves rather the squares of rescaled densities integrate to unity and must be considered as probability densities. The superposition principle—the rule stipulating production of new probability amplitudes from given ones $\{\chi_k \in L^2(\mathbb{R}^3) | k = 1-N\}$ —is well known in quantum mechanics [12]:

$$\sum_k a_k \chi_k(\mathbf{R}), \quad (11)$$

$a_k \in \mathbb{C}$, so that the corresponding probability *density* is given by:

$$\frac{\sum_{kj} a_k^* a_j \chi_k^*(\mathbf{R}) \chi_j(\mathbf{R})}{\sum_{kj} a_k^* a_j \langle \chi_k | \chi_j \rangle}, \quad (12)$$

with $*$ standing for complex conjugate, which immediately shows that two amplitudes differing by a common numerical multiplier produce the same density, that is, describe the same quantum state, which makes these latter rays in the functional Hilbert space $L^2(\mathbb{R}^3)$ or, equivalently, points on the unit sphere in $L^2(\mathbb{R}^3)$ [17]. This allows us to go even further in obtaining elegant description of (dis)similarity of two shapes using the machinery pertinent to quantum mechanics.

3.1. Symmetric Quantum Quantifier of Shape (Dis)similarity

In order to obtain a shape (dis)similarity quantifier based on quantum mechanics, we introduce amplitudes for the blurred densities (normal distributions) Eq. (5) as the square roots of the latter:

$$\chi_{\mathbf{Q}}^{(\sigma)}(\mathbf{R}) = \frac{1}{(\sqrt{2\pi}\sigma)^{3/2}} \exp\left(-\frac{(\mathbf{R}-\mathbf{Q})^2}{4\sigma^2}\right). \quad (13)$$

Their overlaps are:

$$\langle \chi_{\mathbf{Q}}^{(\sigma)} | \chi_{\mathbf{Q}'}^{(\sigma)} \rangle = \exp\left(-\frac{(\mathbf{Q}-\mathbf{Q}')^2}{8\sigma^2}\right), \quad (14)$$

and the amplitude, describing the pattern set P , is:

$$\psi_P^{(\sigma)}(\mathbf{R}) = \sum_k \chi_{\mathbf{p}_k}^{(\sigma)}(\mathbf{R}), \quad (15)$$

and analogously for the trial set.

Applying to the amplitudes the Fubini–Study metric [15] which in the quantum mechanical bra-ket notation reads:

$$\gamma(\psi_P^{(\sigma)}, \psi_T^{(\sigma)}) = \arccos \sqrt{\frac{\langle \psi_P^{(\sigma)} | \psi_T^{(\sigma)} \rangle \langle \psi_T^{(\sigma)} | \psi_P^{(\sigma)} \rangle}{\langle \psi_P^{(\sigma)} | \psi_P^{(\sigma)} \rangle \langle \psi_T^{(\sigma)} | \psi_T^{(\sigma)} \rangle}}, \quad (16)$$

and is nothing, but the angle between the rays in the Hilbert space of amplitudes, defined, respectively, by $\psi_P^{(\sigma)}$ and $\psi_T^{(\sigma)}$ yields the natural distance between the corresponding points on the unity sphere in $L^2(\mathbb{R}^3)$. The conjugation—inverse order of the multipliers in the second scalar product in the numerator—is immaterial here since the amplitudes we are dealing with are always real.

Although, the quantum density Eq. (12) does not coincide with the sum of blurred densities Eq. (5) for finite values of σ ,¹² they have the same limit Eq. (3) as $\sigma \rightarrow 0$. Similarly, the Fubini–Study metric Eq. (16) when applied to the density amplitudes of the form Eq. (15) has the same analytic properties as the arccosine of the Carbó index Eq. (9). It vanishes for the coinciding shapes and flows to the maximal value of $\pi/2$ for orthogonal amplitudes, which, actually, never appear for the setting given by Eqs. (13), (15). More specifically, for the small enough σ the squared norms of the set amplitudes are:

$$\langle \psi_P^{(\sigma)} | \psi_P^{(\sigma)} \rangle \approx \langle \psi_T^{(\sigma)} | \psi_T^{(\sigma)} \rangle \approx N,$$

and the overlap of the set amplitudes for the sets “close” in the sense of the previous section reads:

$$\langle \psi_P^{(\sigma)} | \psi_T^{(\sigma)} \rangle \approx N - \frac{1}{8\sigma^2} \sum_{m=1}^N (\mathbf{T}_m - \mathbf{P}_m)^2$$

(as previously summation over m refers to the properly renumbered vertices) which yields in the lowest order

$$\gamma(\psi_P^{(\sigma)}, \psi_T^{(\sigma)}) \approx \arccos \left(1 - \frac{1}{8N\sigma^2} \sum_{m=1}^N (\mathbf{T}_m - \mathbf{P}_m)^2 \right),$$

and for the “close” sets recovers the Euclidean distance between them following the recipe Eq. (10) merely replacing coefficient 2 by 4.

So far we were dealing with a shape similarity quantifier between two fixed sets of discrete vertices in $\mathbb{R}^3$. The standard setting in the CShM/CSM context includes a mutual superimposition of the two sets by applying allowed transformations Ω to the vectors of the trial set and optimizing the value of so obtained quantifier with respect to the variational parameters ω

defining Ω . This formally corresponds to a replacement of the vectors $\mathbf{T}_j$ in the distributions $\rho_T, \rho_T^{(\sigma)}$ Eqs. (3), (5) or in amplitudes $\psi_T^{(\sigma)}$ Eq. (15) by $\Omega\mathbf{T}_j$ and maximizing either the Carbó index $c_{P,\Omega(\omega)T}$ or the overlap $\langle \psi_P^{(\sigma)} | \psi_{\Omega(\omega)T}^{(\sigma)} \rangle$ with respect to ω . As the shifts are eliminated by superimposing the centers of the sets Eq. (2) the optimization with respect to rotation and scaling implied by the definitions of CShM and CSM can be easily implemented with use of analytic gradients of the overlaps of the pattern and trial density amplitudes with respect to the variational parameters. More specifically, the derivative of the overlap of the density amplitudes of the trial and pattern shapes $\langle \psi_P^{(\sigma)} | \psi_{\Omega(\omega)T}^{(\sigma)} \rangle$ expresses through the derivatives of the overlaps of each pair of normal distribution amplitudes centered, respectively, on the vertices of the pattern and trial sets:

$$\frac{\partial \langle \psi_P^{(\sigma)} | \psi_{\Omega T}^{(\sigma)} \rangle}{\partial \omega} = \sum_{k,j} \frac{\partial \langle \chi_{P_k}^{(\sigma)} | \chi_{\Omega T_j}^{(\sigma)} \rangle}{\partial \omega}.$$

As for the latter, their explicit analytical expression is:

$$\begin{aligned} \frac{\partial \langle \chi_{P_k}^{(\sigma)} | \chi_{\Omega T_j}^{(\sigma)} \rangle}{\partial \omega} &= \frac{\partial}{\partial \omega} \exp \left(-\frac{(\mathbf{P}_k - \Omega\mathbf{T}_j)^2}{8\sigma^2} \right) \\ &= -\frac{2}{4\sigma^2} \frac{\partial \Omega}{\partial \omega} \mathbf{T}_j (\mathbf{P}_k - \Omega\mathbf{T}_j) \exp \left(-\frac{(\mathbf{P}_k - \Omega\mathbf{T}_j)^2}{8\sigma^2} \right). \end{aligned} \quad (17)$$

The allowed transformation matrix dependent on the above mentioned parameters can be explicitly written:

$$\Omega(\kappa, \boldsymbol{\varphi}) = \kappa \mathcal{R}(\boldsymbol{\varphi}) = \kappa \prod_{\alpha} \mathcal{J}_{\alpha}(\varphi_{\alpha}),$$

where by $\mathcal{R}$ we as previously denote a *rotation* of $\mathbb{R}^3$ by a triple of angles which is representable by a product of three Jacobi matrices $\mathcal{J}_{\alpha}(\varphi_{\alpha})$, $\alpha = x, y, z$ describing rotations around three coordinate axes of the frame fixed on the trial set (see the right side of Fig. 2). Their derivatives with respect to the angles can be written straightforwardly and that with respect to κ is obvious.

3.2. Antisymmetric Quantum Quantifier of Shape Similarity

Employing amplitudes (wave functions) opens yet another possibility of treating shape (dis)similarity, complementary to one described in Sections 2.2.1 and 3.3.1. There the vertex set was represented as a (weak) limit of a one-particle density as defined by Eqs. (6), (12)¹³ with decaying dispersion. Here we consider an

¹²The difference is the sum of interference terms (known in quantum chemistry as “differential overlaps”) $\chi_Q^{(\sigma)}(\mathbf{R})\chi_Q^{(\sigma)*}(\mathbf{R})$ —the price of the consistent quantum description, which any way rapidly decays with increase of the separation between $\mathbf{Q}$ and $\mathbf{Q}'$ and decrease of dispersion.

¹³Either of functions $\rho_Q, \rho_Q^{(\sigma)}, \psi_P^{(\sigma)}$ is that of one vector $\mathbf{R}$ as its argument which can eventually represent coordinates of only one particle, thus in the terminology of statistical physics and quantum mechanics those functions are one-particle densities or amplitudes.

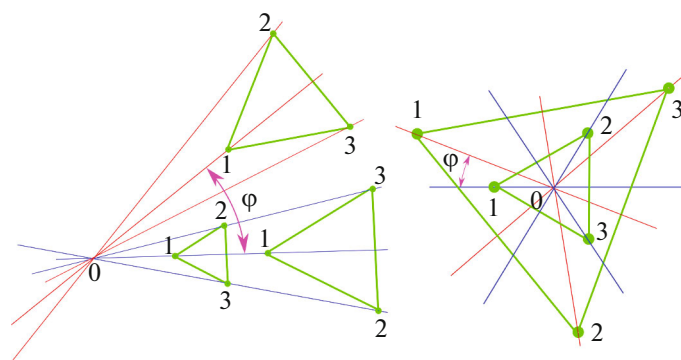


Fig. 2. Two cases of comparison of shapes of equilateral triangles. Left: the affine space picture where the pattern (smaller triangle) and trial (larger triangles) shapes are positioned not in the origin of the coordinate frame; the upper triangle is “correctly” numbered so that rotation by an angle φ and an appropriate scaling superimposes the trial and the pattern shapes so that the distance Eq. (1) ultimately vanishes. The distance Eq. (1) for the lower trial triangle cannot be made smaller than two unless an appropriate renumbering is found. Right: the vector space picture.

approach based not on a one-particle density/amplitude rather on *many-particle* amplitudes (wave functions).

Recollecting the chemical roots of the task of comparing *molecular* shapes we realize that even at the absolute zero of temperature the nuclei do not occupy fixed positions (as assumed in all CShM/CSM settings) rather each resides in a ground state of a harmonic oscillator, describing its *zero vibrations* [12]. Depending on the total spin of the nuclei of each sort (isotope) the wave function (amplitude) of their entire set is either symmetrized (for bosonic nuclei with integer spin) or antisymmetrized (for fermionic ones with half-integer spin) product [13] or, respectively, permanent or determinant¹⁴ of the ground states of the individual oscillators. The ground state of the harmonic oscillator is known to be a Gaussian function centered at the minimum of the molecular potential (and normalized according to Eq. (13)). In general the dispersions of these Gaussians depend on the characteristics of the potential and on the masses of nuclei, but here we can take them all with the same dispersion and ignore the existence of bosonic nuclei (one can think that for each element we evoke a fermionic isotope with half-integer spin). Then the many-particle amplitude for the pattern shape is a determinant:

$$\begin{aligned} \Pi^{(\sigma)}(P|\{\mathbf{R}_j\}) &= |\chi_{\mathbf{P}_1}^{(\sigma)}(\mathbf{R}_1), \dots, \chi_{\mathbf{P}_N}^{(\sigma)}(\mathbf{R}_N)| \\ &= \begin{vmatrix} \chi_{\mathbf{P}_1}^{(\sigma)}(\mathbf{R}_1) & \dots & \chi_{\mathbf{P}_1}^{(\sigma)}(\mathbf{R}_N) \\ \vdots & \ddots & \vdots \\ \chi_{\mathbf{P}_N}^{(\sigma)}(\mathbf{R}_1) & \dots & \chi_{\mathbf{P}_N}^{(\sigma)}(\mathbf{R}_N) \end{vmatrix}, \end{aligned} \quad (18)$$

¹⁴Like electrons where respective functions are known as Slater determinants.

and analogously for the trial shape. Eq. (18) is a *function* of N radii-vectors $\mathbf{R}_1, \dots, \mathbf{R}_N$ of involved nuclei dependent on N vector parameters $\mathbf{P}_1, \dots, \mathbf{P}_N$ —the positions where the component Gaussians are centered.

The scalar product in the Hilbert space spanned by such determinants¹⁵ (the overlap of the determinants) is the determinant of the matrix of the overlaps of the Gaussians forming the many-particle amplitudes Eq. (18) of the respective sets [13]:

$$\begin{aligned} \langle \Pi^{(\sigma)}(P) | \Pi^{(\sigma)}(T) \rangle &= \begin{vmatrix} \langle \chi_{\mathbf{P}_1}^{(\sigma)} | \chi_{\mathbf{T}_1}^{(\sigma)} \rangle & \dots & \langle \chi_{\mathbf{P}_1}^{(\sigma)} | \chi_{\mathbf{T}_N}^{(\sigma)} \rangle \\ \vdots & \ddots & \vdots \\ \langle \chi_{\mathbf{P}_N}^{(\sigma)} | \chi_{\mathbf{T}_1}^{(\sigma)} \rangle & \dots & \langle \chi_{\mathbf{P}_N}^{(\sigma)} | \chi_{\mathbf{T}_N}^{(\sigma)} \rangle \end{vmatrix} \\ &= \det S^{(\sigma)}(PT), \end{aligned} \quad (19)$$

with $S_{kj}^{(\sigma)}(PT) = \langle \chi_{\mathbf{P}_k}^{(\sigma)} | \chi_{\mathbf{T}_j}^{(\sigma)} \rangle$ given by Eq. (14). Despite their scaring appearance the determinants can be calculated relatively easily (in variance with permanent, which is the pragmatic reason of our choice in favor of fermionic nuclei). Indeed, for the square (but not necessarily symmetric) *matrix* $S^{(\sigma)}$ the singular value decomposition (SVD) of the form:

$$S^{(\sigma)} = U^{(\sigma)} \Lambda^{(\sigma)} V^{(\sigma)\dagger}, \quad (20)$$

where $U^{(\sigma)}$ and $V^{(\sigma)}$ are square unitary (for the real $S^{(\sigma)}$ orthonormal) matrices and $\Lambda^{(\sigma)}$ is a diagonal matrix with non-negative elements—singular values of $S^{(\sigma)}$, is always possible [18]; the statement known as well as the Löwdin parity theorem [11]. Then the sought determinant is given by [19]:

$$\det S^{(\sigma)} = \det U^{(\sigma)} \det \Lambda^{(\sigma)} \det V^{(\sigma)\dagger} = \prod_{i=1}^N \lambda_i^{(\sigma)}.$$

¹⁵The outer or Grassmann product $\bigwedge^N L^2(\mathbb{R}^3)$ of the $L^2(\mathbb{R}^3)$ Hilbert spaces of functions of each variable $\mathbf{R}_i$.

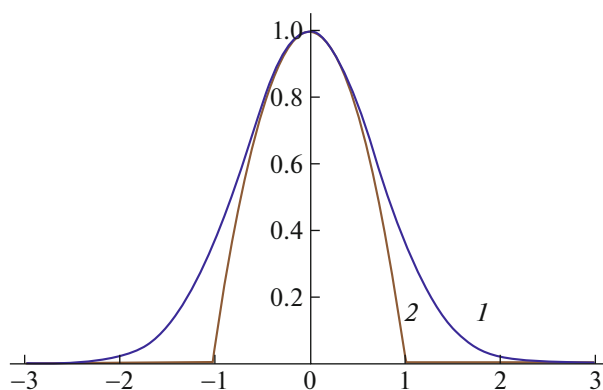


Fig. 3. Graphical representation of the course of the Gaussian (blue) and quadratic approximation to it (orange) in the vicinity of the vanishing argument.

This is an alternative solution of the combinatorial problem of the previous CShM/CSM approaches since the linear algebra algorithms [18, 19] scale as $O(N^3)$ with the number of vertices in the sets to be compared, although efficiently taking into account all permutations of the trial and pattern vertex labels due to the nature of determinant as of fully antisymmetric function of its arguments—by far better than the factorial one as reported in [3, 4, 6].

In the limiting case of the scalar squares of the individual many-particle amplitudes—those when $P = T$ —the norm comes out to surface as the *determinants* of the Gram matrices $S^{(\sigma)}(PP)$ and $S^{(\sigma)}(TT)$ of the Gaussian amplitudes pertaining to either set, which in their turn can be efficiently recovered as the product of their eigenvalues accessible at the $O(N^3)$ computational costs.

For $P = T$ the Gram matrices $S^{(\sigma)}(PP)$ and $S^{(\sigma)}(TT)$ have unities on their respective diagonals whereas the off-diagonal matrix elements can be made arbitrarily small making σ small enough. This eventually produces the eigenvalues, of which the k th, according to the Gerschgorin theorem [20], falls in the segment

$$1 \pm \sum_{j \neq k} \exp\left(-\frac{(\mathbf{Q}_k - \mathbf{Q}_j)^2}{8\sigma^2}\right), \quad Q = P, T \text{ which again can}$$

be made arbitrary narrow by selecting appropriate value of σ . By this the norm of the determinant amplitude Eq. (18) as well can be made arbitrarily close to unity.

Turning to comparison of the shapes P and T , we see that manipulating, as previously, with σ one can make all matrix elements in the row/column of the overlap matrix $S^{(\sigma)}(PT)$ much smaller than the maximal one. Then applying a similar reasoning (based on the Gerschgorin theorem, as above) to the product of overlap matrices $S^{(\sigma)}(PT) S^{(\sigma)}(TP)$ we conclude that k th eigenvalue of the latter can be made to flow to

$\exp\left(-\frac{(\mathbf{P}_k - \mathbf{T}_j)^2}{4\sigma^2}\right)$ for j giving $\min_j (\mathbf{P}_k - \mathbf{T}_j)^2$ so that the k th singular value of $S^{(\sigma)}(PT)$ flows to

$\exp\left(-\frac{(\mathbf{P}_k - \mathbf{T}_j)^2}{8\sigma^2}\right)$. Thus, a *unique* renumbering of vertices in the sets to be compared, resp., rows and columns in the overlap matrix, placing maximal elements of each row on the diagonal, is found. In the numeric SVD the required permutation is built in to the matrices $U^{(\sigma)}$ and $V^{(\sigma)}$. At finite (small enough) values of dispersion, the $U^{(\sigma)}$ and $V^{(\sigma)}$ are going to have a simple structure of the above permutation matrix multiplied by some orthogonal transformation not much different from the identity matrix. Since the singular values in $\Lambda^{(\sigma)}$ are almost equal to the above largest matrix elements, their product flows asymptotically to

$$\exp\left(-\frac{\sum_{m=1}^N (\mathbf{T}_m - \mathbf{P}_m)^2}{8\sigma^2}\right),$$

where the subscript m refers to properly renumbered vertices. Obviously, taking the logarithm of the above expression and multiplying it by $-8\sigma^2$ brings us back to the original definition of the (nonnormalized) shape measure Eq. (1) based on the Euclidean metric in $\mathbb{R}^3$. More formally, we introduce the family of shape (dis)similarity quantifiers parameterized by σ :

$$-\ln \det S^{(\sigma)}(PT)$$

and redefine the target quantifier by

$$-8\sigma^2 \ln \det S^{(\sigma)}(PT) \propto \sum_{m=1}^N (\mathbf{T}_m - \mathbf{P}_m)^2, \quad (21)$$

which, as we see, is again weakly dependent on σ and flows to the sum of (squared) Euclidean distances in the properly renumbered $\mathbb{R}^3$ spaces. This multiplicative quantifier works even better than the additive ones of Sections 2.2.1, 3.3.1: in variance with them which show the quadratic, respectively, Euclidean asymptotic only for “close” vertices $\sigma > |\mathbf{P}_k - \mathbf{T}_j|$, the σ^2 -rescaled logarithm of the multiplicative quantifier Eq. (21) flows to the (squared) Euclidean distance between the vertex sets for arbitrary vertex-vertex separations.

For the purpose of optimization with respect to rotation and scaling as implied by the definitions of CShM and CSM the vertices in the trial set are to be replaced by their images under the action of the allowed transformations Ω . Since the asymptotic for small enough σ is given by Eq. (21) the recipe:

$$-8\sigma^2 \min_{\omega} \ln \det S^{(\sigma)}(P, \Omega(\omega)T)$$

may be used for finding the traditional CShM stemming from Eq. (1).

The above considerations easily reformulate also in terms of the Fubini-Study metric in the Grassmann

product $\bigwedge^N L^2(\mathbb{R}^3)$ of the $L^2(\mathbb{R}^3)$ Hilbert spaces. Indeed the cosine of the angle between the rays in $\bigwedge^N L^2(\mathbb{R}^3)$ defined by the amplitudes $\Pi^{(\sigma)}(P)$, $\Pi^{(\sigma)}(T)$ is given by

$$\begin{aligned} & \cos \gamma(\Pi^{(\sigma)}(P), \Pi^{(\sigma)}(T)) \\ &= \sqrt{\frac{\langle \Pi^{(\sigma)}(P) | \Pi^{(\sigma)}(T) \rangle \langle \Pi^{(\sigma)}(T) | \Pi^{(\sigma)}(P) \rangle}{\langle \Pi^{(\sigma)}(P) | \Pi^{(\sigma)}(P) \rangle \langle \Pi^{(\sigma)}(T) | \Pi^{(\sigma)}(T) \rangle}} \\ &= \sqrt{\frac{\det[S^{(\sigma)}(PT)S^{(\sigma)}(TP)]}{\det[S^{(\sigma)}(PP)S^{(\sigma)}(TT)]}} \end{aligned}$$

Since $\det[S^{(\sigma)}(PP)S^{(\sigma)}(TT)]$ can be made arbitrary close to unity and $\det S^{(\sigma)}(PT)$ does not change under replacement of rows and columns, taking logarithm of the above expression immediately yields $\ln \det S^{(\sigma)}(PT)$ as previously.

3.3. Geometric Quantifier for the Shapes' (Dis)similarity

In the Sections 2.2.1, 3.3.1, 3.3.2 we identified some objects in functional spaces, which are in one-to-one correspondence with the shapes defined by finite vertex sets in $\mathbb{R}^3$, being fully (anti)symmetric with respect to permutations of the vertices' labels. The scalar products (overlaps) between these objects in the respective functional spaces, per se were not whatever quantifiers of the (dis)similarity of the shapes to be compared, but possessed that crucial property that for coinciding shapes the distances between their images in functional spaces were automatically vanishing—the feature which in the traditional CShM/CSM setting is accessible only with the factorial computational costs. In all cases the technical part of the solution consisted in ranking the set of vertex-vertex separations. By this we could obtain proper renumbering of the pattern and trial sets for which the Euclidean distances Eq. (1) yield correct result (zero) for the coinciding shapes. The said ranking allows us to go further and to reconstruct purely geometric quantifier for the shapes' (dis)similarity without explicit addressing the functional spaces.

Indeed, we recollect that the difference between rays quantifies by the angle between them, which according to the high-school geometry is:

$$\theta = \arccos \frac{(\mathbf{a}, \mathbf{b})}{ab}; \quad a = \sqrt{(\mathbf{a}, \mathbf{a})}; \quad b = \sqrt{(\mathbf{b}, \mathbf{b})} \quad (22)$$

(here $\mathbf{a}$, $\mathbf{b}$ are arbitrary vectors defining rays as sets of vectors $\kappa \mathbf{a}$, $\lambda \mathbf{b}$; $\kappa, \lambda > 0$ and $(\mathbf{a}, \mathbf{b})$ is their scalar product). The angle between the rays does not depend on their origins—in the setting accepted here the origins of all rays are reduced to the common center, but it is not mandatory as one can understand from Fig. 2.

The high-school definition Eq. (22) easily extends to the shapes. Indeed, recollecting the option of representing

shapes by rays in the direct sum $\bigoplus_{m=1}^N \mathbb{R}^3$ of $\mathbb{R}^3$ spaces, mentioned in the beginning of Section 2, and assuming the proper renumbering of the vertices in the trial and pattern sets to be established by whatever method described above we represent the vertex sets by vectors

$$\bigoplus_m \mathbf{P}_m, \bigoplus_m \mathbf{T}_m$$

in the direct sum $\bigoplus_{m=1}^N \mathbb{R}^3$ of properly renumbered $\mathbb{R}^3$ spaces.¹⁶ The respective scalar product is

$$\left(\bigoplus_m \mathbf{P}_m, \bigoplus_m \mathbf{T}_m \right) = \sum_{m=1}^N (\mathbf{P}_m, \mathbf{T}_m),$$

where $(\mathbf{P}_m, \mathbf{T}_m)$ are the scalar products in $\mathbb{R}^3$ as those in Eq. (22) and the norms of the respective sets as vectors in the direct sum are $|\mathcal{Q}| = \sqrt{\sum_{m=1}^N (\mathbf{Q}_m, \mathbf{Q}_m)}$; $\mathcal{Q} = P, T$ [19]—the usual Euclidean norms. With use of these concepts we can define purely geometric quantifier for the shapes' (dis)similarity, namely, the angle between the rays representing the shapes:

$$\cos \theta(P, T) = \left(\frac{\bigoplus_m \mathbf{P}_m}{|P|}, \frac{\bigoplus_m \mathbf{T}_m}{|T|} \right) = \frac{\sum_{m=1}^N (\mathbf{P}_m, \mathbf{T}_m)}{|P||T|} \quad (23)$$

(proper ordering of vertices defined above is crucial) which up to the definition of the meaning of those or other brackets coincides with the Carbó index for distributions or the (cosine of the) Fubini–Study metric in the spaces of (symmetric or antisymmetric) amplitudes. The optimization with respect to scaling factor is unnecessary within such a setting, since all scaling operations are absorbed in the vector norms in the denominator. The optimization with respect to *relative* rotations reduces to maximizing the following simple expression:

$$\sum_{m=1}^N (\mathbf{P}_m, \mathcal{R}(\boldsymbol{\varphi}) \mathbf{T}_m),$$

with respect to the components of $\boldsymbol{\varphi}$ ($|P|$ and $|T|$ are rotation invariant). Then, minimizing $\theta(P, T)$ with respect to three rotation angles $(\varphi_x, \varphi_y, \varphi_z) = \boldsymbol{\varphi}$ eventually produces an adequate (dis)similarity quantifier for a pair of shapes, vanishing for the shapes represented by coinciding rays.

4. CONCLUSION AND OUTLOOK

In the present paper we addressed the issue important for the area of “continuous shape and symmetry measures” [1–6], namely the contradiction between

¹⁶One can understand by $\bigoplus_m \mathbf{Q}_m$ the $3N$ components of N vectors $\mathbf{Q}_m$ written one after another or N column vectors $\mathbf{Q}_m$ arranged in a $3 \times N$ matrix like in Eq. (15) of [4].

the *idea* of the (molecular) shape as of an invariant characteristic of a *set* of vertices in the three-dimensional space and the only available *representation* of a set in a computer program as an array or a list either assuming certain *order* not pertinent to the concept of set. While the invariance of shapes with respect to continuous symmetries like translation, rotation and scaling is covered pretty straightforwardly, the invariance with respect to reordering the elements of arrays or lists requires for its correct reproduction in the traditional approach [1–6] the program tools scaling as a factorial of the number of vertices. In order to get around this complication we undertook an excursion to the area of functional spaces where we could construct representations of the shapes, symmetric (one-particle densities and density amplitudes) and anti-symmetric (many-particle density amplitudes) with respect to the permutations of the vertices' labels.

With use of this instrumentation we were able to develop a classification to be applied to all pairwise vertex-vertex distances between the sets to be compared so that an adequate renumbering of the vertices in the sets in question could be established. This ultimately reduces to a simplistic selection of the shortest distance in each row of the $N \times N$ vertex-vertex distance matrix. After this is done, the functional shape (dis)similarity quantifiers flow to the Euclidean distance between the properly reordered coordinate arrays provided the vivid recommendation by Littlewood: "Thus σ should be made as small as possible" [21] is fulfilled.¹⁷ Moreover, the proper reordering allows us to introduce purely geometrical shape dissimilarity quantifier: the angle between the rays defined by properly ordered vectors, representing shapes to be compared, which effectively rounds up our excursion in the realm of functional spaces.

Presently, we restrict ourselves by sets of vertices of one sort (molecules made of equivalent atoms). The generalization to the case of multiple sorts is straightforward. As the problems of the factorial growth of the computational costs are common for the shape measures considered here and the symmetry ones, these latter will be addressed elsewhere.

¹⁷"As small as possible" is not actually that small. Taking the separation between the nearest vertices in the pattern set for unity we estimate $\sigma = \frac{1}{2\sqrt{2}}$, which assures the factor $\frac{1}{e^2}$ —almost one tenth—for the Carbó index of two normal distributions centered in these vertices. Then in order to assure that for the "close" distance the quadratic approximation for the Gaussian filter is valid, it is enough to set its argument $< \frac{1}{9}$ that is the distance between the "close" vertices must be less than $\frac{1}{3\sqrt{2}}$ which assures the precision of the quadratic approximation at the level of 0.7%—more than enough for whatever practical applications.

ACKNOWLEDGMENTS

The author is thankful to Prof. Pere Alemany (Barcelona) who some years ago drew his attention to the problem of continuous shape and symmetry measures. Dr. Olga Chugreeva (Stuttgart) is acknowledged for valuable discussion.

REFERENCES

1. M. Pinsky and D. Avnir, *Inorg. Chem.* **37**, 5575 (1998).
2. D. Casanova, J. Cirera, M. Llunell, P. Alemany, D. Avnir, and S. Alvarez, *J. Am. Chem. Soc.* **126**, 1755 (2004).
3. H. Zabrodsky, S. Peleg, and D. Avnir, *J. Am. Chem. Soc.* **114**, 7843 (1992); H. Zabrodsky and D. Avnir, *J. Am. Chem. Soc.* **117**, 462 (1995).
4. P. Alemany, D. Casanova, S. Alvarez, C. Dryzun, and D. Avnir, *Rev. Comput. Chem.* **30**, 289 (2017).
5. C. Dryzun and D. Avnir, *Phys. Chem. Chem. Phys.* **11**, 9653 (2009).
6. M. Pinsky, C. Dryzun, D. Casanova, P. Alemany, and D. Avnir, *J. Comput. Chem.* **29**, 2712 (2008). <https://doi.org/10.1002/jcc.20990>
7. G. G. Malenkov, *J. Struct. Chem.* **57**, 793 (2016); A. L. Tchougreeff, *AIP Conf. Proc.* **1906**, 030004 (2017). <https://doi.org/10.1063/1.5012283>; A. L. Tchougreeff, *Theor. Chem. Acc.* **137**, 138 (2018). <https://doi.org/10.1007/s00214-018-2322-0>
8. A. I. Kostrikin and Yu. I. Manin, *Linear Algebra and Geometry* (Nauka, Moscow, 1986; CRC, Boca Raton, 1989).
9. L. Schwartz, *Theorie des distributions* (Hermann, Paris, 1951); A. Ya. Khelemskii, *Lectures and Exercises on Functional Analysis* (MTsNMO, Moscow, 2014; Am. Math. Soc., Philadelphia, 2006).
10. R. Carbo and B. Calabuig, *Int. J. Quantum Chem.*, **42**, 1695 (1992); E. Besalu, R. Carbo, J. Mestres, and M. Solà, *Top. Curr. Chem.* **173**, 31 (1995); P. E. Bowen-Jenkins, D. L. Cooper, and W. G. Richards, *J. Phys. Chem.* **89**, 2195 (1985); R. Carbó and E. Besalú, *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches* (Kluwer Academic, Amsterdam, 1995); R. Carbó and B. Calabuig, *Int. J. Quantum Chem.* **42**, 1681 (1992); R. Carbó, B. Calabuig, and E. Besalú, *Adv. Quantum Chem.* **25**, 255 (1994); R. Carbó, L. Leyda, and M. Arnau, *Int. J. Quantum Chem.* **17**, 1185 (1980).
11. I. Mayer, *Simple Theorems, Proofs, and Derivations in Quantum Chemistry* (Kluwer Academic/Plenum, New York, 2003).
12. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics, Vol. 3: Quantum Mechanics: Non-Relativistic Theory* (Nauka, Moscow, 1989; Pergamon, Oxford, 1977).
13. L. Zülicke, *Quantenchemie, Ein Lehrgang, Band 1: Grundlagen und allgemeine Methoden* (Deutscher Verlag der Wissenschaften, Berlin, 1973).

14. L. Piela, J. Kostrowicki, and H. A. Scheraga, *J. Phys. Chem.* **93** (8) (1989).
15. G. Fubini, *Atti Reale Ist. Veneto Sci., Lett. Art.* **63**, 502 (1904); *E. Study, Math. Ann.* **60**, 321 (1905).
16. A. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Springer, Berlin, 1933); B. N. Tutubalin, *The Theory of Probabilities and Stochastic Processes. Fundamentals of the Mathematical Apparatus and Applied Aspects* (Mosk. Gos. Univ., Moscow, 1992) [in Russian].
17. D. C. Brody and L. P. Hughston, *J. Geom. Phys.* **38**, 19 (2001); P. Facchi, R. Kulkarni, V. I. Man'ko, G. Mar-mo, E. C. G. Sudarshan, and F. Ventriglia, *Phys. Lett. A* **374**, 4801 (2010).
18. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed. (Cambridge Univ. Press, Cambridge, 1992), Chap. 2.6.
19. F. R. Gantmacher, *The Theory of Matrices* (Chelsea, New York, 1960).
20. S. Gerschgorin, *Bull. Acad. Sci. URSS, Math. Natur.*, No. 6, 749 (1931).
21. J. L. Littlewood, *A Mathematician's Miscellany* (Methuen, London, 1953), p. 38.