



HAL
open science

Sensaas: Shape-based Alignment by Registration of Colored Point-based Surfaces

Dominique Douguet, Frédéric Payan

► **To cite this version:**

Dominique Douguet, Frédéric Payan. Sensaas: Shape-based Alignment by Registration of Colored Point-based Surfaces. *Molecular Informatics*, 2020, 10.1002/minf.202000081 . hal-02879260

HAL Id: hal-02879260

<https://hal.science/hal-02879260v1>

Submitted on 23 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DOI: 10.1002/minf.202000081

SENSAAS: Shape-based Alignment by Registration of Colored Point-based Surfaces

Dominique Douguet*^[a] and Frédéric Payan*^[b]

Abstract: SENSAAS is a tool developed for aligning and comparing molecular shapes and sub-shapes. Alignment is obtained by registration of 3D point-based representations of the van der Waals surface. The method uses local properties of the shape to identify the correspondence relationships between two point clouds containing up to several thousand colored (labeled) points. Our rigid-body superimposition method follows a two-stage approach. An initial alignment is obtained by matching pose-invariant local 3D descriptors, called FPFH, of the input point clouds. This stage provides a global superimposition of the molecular surfaces, without any knowledge of their initial pose in 3D space. This alignment is then refined by optimizing the matching of colored points. In our study,

each point is colored according to its closest atom, which itself belongs to a user defined physico-chemical class. Finally, SENSAAS provides an alignment and evaluates the molecular similarity by using Tversky coefficients. To assess the efficiency of this approach, we tested its ability to reproduce the superimposition of X-ray structures of the benchmarking AstraZeneca (AZ) data set and, compared its results with those generated by the two shape-alignment approaches shaEP and SHAFTS. We also illustrated sub-matching properties of our method with respect to few substructures and bioisosteric fragments. The code is available upon request from the authors (demo version at <https://chemoinfo.ipmc.cnrs.fr/SENSAAS>).

Keywords: Shape-based alignment · molecular surfaces · point clouds · registration · molecular similarity

1 Introduction

3D shape is an important molecular feature in biological activity.^[1] Over years, a wide range of 3D shape similarity methods have been developed to compare and quantify the similarity between molecules. Molecular shape comparison methods can be classified according to the representation used. *Atomic-distance* based methods are alignment-free algorithms that describe the shape by using inter-atomic distance distributions. An example is the Ultrafast Shape Recognition (USR) tool developed by Ballester and Richards, which encodes the information into a string vector.^[2] *Atom-centered Gaussians* based methods, including the program ROCS, are probably the most widely used to evaluate the maximum volume overlap between two molecules.^[3,4] A third popular class uses *molecular fields* to describe, for example, the electrostatic or the hydrophobic potential, calculated at points of a grid around the molecule. In their review of 2018, Kumar and Zhang summarized several variations and hybridizations of these main classes.^[5] For example, shaEP is a tool that uses both *molecular fields* and *Gaussian functions* representations to estimate the molecular similarity^[6] while SHAFTS is an algorithm that combines a pharmacophore matching method and *Gaussian functions*.^[7] Alternative classes of shape similarity methods are *surface-based* approaches that use various representations of the surface such as spherical harmonics (PARAFIT,^[8] sHEms^[9] and HPCC^[10]) and their extensions (3D Zernike descriptors^[11]), probability distribu-

tion histograms as shape signature,^[12] alpha shapes as coarse representations of the surface^[13,14] and point-based surfaces. Point-based surface methods describe the molecular surfaces with sets of 3D points, clustered to form similar surface regions also called patches. Local surface properties of the surface patches are then extracted to represent the shape. Goldman and Wipke use quadratic shape descriptors of a subset of points called critical points to evaluate pairings between small sections of the surfaces.^[15] Cosgrove et al.,^[16] Exner et al.,^[17] Hofbauer et al.,^[18] Baum et al.^[19–21] and Krotzky et al.^[22] also developed approaches that divide the molecular surface into a set of patches enclosing an area of similar properties. Then, clique detection algorithms are usually used to identify maximal

[a] Dr. D. Douguet
Université Côte d'Azur, Inserm, CNRS, IPMC, 660 route des Lucioles
06560 Valbonne, France
E-mail: douguet@ipmc.cnrs.fr

[b] Dr. F. Payan
Université Côte d'Azur, CNRS, I3S, Les Algorithmes - Euclide B, 2000
route des Lucioles 06900 Sophia Antipolis, France
E-mail: fpayan@i3s.unice.fr

Supporting information for this article is available on the WWW under <https://doi.org/10.1002/minf.202000081>

© 2020 The Authors. Published by Wiley-VCH Verlag GmbH & Co. KGaA. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

common subgraphs and to perform a rigid-body superimposition between two sets of patches. Clique detection is a combinatorial problem known to be NP-hard and, thus, cliques are impracticable when one wants to work with a large number of points. In the current work, we describe a point-based surface method which aligns thousands of colored points by using registration methods, initially developed for matching dense 3D point clouds acquired by 3D sensors.

Representing a molecular surface by using a 3D point cloud is not a new concept.^[1,5,23-25] A point cloud has this advantage of being an intuitive description, suited for visualization and for building a mental model. For example, this representation highlights which atoms contribute to the surface and, thus, may interact with its surroundings. By contrast, working with 3D graphs do not allow to distinguish atoms that are masked by others. Another feature of point-based surfaces is the capacity to associate some extra attributes to the points such as color labels according to the nature of the underlying atoms. In their review, Kumar and Zhang pointed out that surface-based shape similarity methods are a particular class of shape similarity methods that are still in infancy.^[5,24] Indeed, although they underlined some particular interesting features of surface-based methods such as comparing shape of ligands with that of binding pockets, they indicated that their current concerns are the slowness, an efficiency lower than Gaussian overlay-based shape similarity methods and the inability to identify local similarities. In our study, we show that a surface-based shape similarity method can now run reasonably fast and is able to identify global and local similarities.

3D point set registration, point set matching or geometric registration refer to the same class of methods that aim to align several sets (or clouds) by identifying nearest-neighbor correspondences and minimizing the point-pair distances.^[26,27] The most popular approach is the ICP (Iterative Closest Point) method which gave rise to numerous variants. Point set registration is a fundamental problem in many domains: pattern recognition, computer vision or data reconstruction. Today, 3D registration methods that use geometrical features such as curvature, surface normals or higher order local descriptors are extensively used.^[28-30]

Registration is often considered as a hard optimization problem because perfect point-to-point correspondences rarely exist: the point clouds might partially overlap only, or the underlying objects may have particular local geometrical features. Nevertheless, we show in this paper that 3D point set registration can be particularly attractive to align molecular shapes and sub-shapes. The resulting approach is called SENSEAAS, for SENSITIVE SURFACE AS A SHAPE. The following sections detail the SENSEAAS workflow and show its ability to reproduce experimentally verified superimpositions of dissimilar molecules of the benchmarking AstraZeneca (AZ) data set.^[31] Originally, this data set was created for validating pharmacophore programs but it

also provides an appropriate benchmarking data set for evaluating the general applicability of a molecular alignment method.^[32] To assess the accuracy of alignments provided by SENSEAAS, we compared them with those generated by the state of the art methods, SHAEP and SHAFTS. Finally, the submatching property of SENSEAAS is emphasized by aligning some molecules of different sizes such as substructures and fragments.

2 Material and Methods

2.1 Sensaas Workflow

An overview of SENSEAAS is illustrated in Figure 1. It can be divided into four stages as follows. Let us consider two molecules that we want to align:

Step 1. Generate a 3D point cloud of the molecular surface for each molecule. Each point is described by its coordinates in 3D space (geometry) and a color according to the nature of the underlying atom (Figure 1b).

Step 2. Initialize the alignment. Point clouds are globally superimposed by finding the best matching in terms of geometry only (Figure 1c).

Step 3. Refine the initial alignment. The initial superimposition is locally refined by taking into account the geometry but also the physico-chemical properties mapped on the surface (pharmacophore features for example). This is done by i) defining classes of properties; ii) coloring each point according to its class (Figure 1d); iii) refining the initial superimposition by finding the best matching between these two colored point clouds (Figure 1e). This step results into a transformation matrix to translate and rotate the moving molecule.

Step 4. Calculate scores and display the superimposition. Fitness scores are calculated using point clouds and the final alignment of the 3D graphs is achieved by applying the transformation matrix to the 3D graph of the moving molecule (Figure 1f).

Each stage is described in detail in the next subsections.

a) Generation of 3D point clouds. Let us consider two molecules M_1 and M_2 to align. The goal of this step is to generate point-based 3D representations S_1 and S_2 of the molecular surface of M_1 and M_2 , respectively. S_1 and S_2 result from the computation of the van der Waals (vdW) surface of each molecule by using the program developed by Eisenhaber et al.,^[33] with van der Waals radii taken from A. Bondi.^[34] Finally, S_1 and S_2 consist in two sets of 3D points uniformly distributed on the vdW surface. Each point is depicted by its position in 3D space (x , y and z) and a label indicating the type of its closest atom.

Point-based surfaces → Global alignment → Coloration → Refinement → 3D graphs visualization

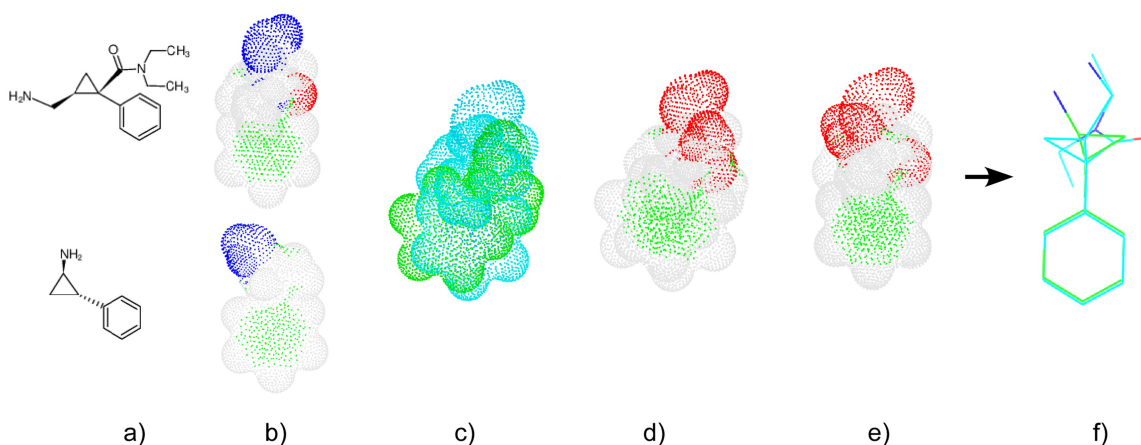


Figure 1. Schematic representation of SENSEAAS workflow for matching two molecules. a) Input molecules; b) Point-based representation of the molecular surfaces; points are colored according to the nature of the closest atom; c) Global superimposition (geometry-aware only); Point clouds are colored in blue and green for visualization purpose; d) Same global superimposition but colored according to the physico-chemical classes; e) Local refinement (geometry- and color-aware); f) Final alignment of the corresponding 3D graphs.

b) Geometry-based alignment. Considering two point sets, *Source* and *Target*, a rigid registration (or matching) consists in finding a transformation matrix T describing how to rotate and translate the *Source* in order to be superimposed to the *Target*. In our context, we have no *a priori* information about the initial pose of the point sets in 3D space, and about the existence of overlapping areas between the molecular surfaces. Therefore, we use a two-stage procedure: i) perform a global registration to obtain an initial “coarse” superimposition, and ii) refine the initial alignment by a local registration. This local registration, that would be inefficient without applying the global one, is explained in the next section.

To match 3D point sets globally, it is generally accepted that it is more efficient to match local 3D descriptors computed on a limited number of points rather than to match the points themselves. Local 3D descriptors provide pose invariant local geometrical properties of the surface inherent to a point cloud. In SENSEAAS, we use Fast Point Features Histograms (FPFH) proposed by Rusu et al.^[35] Such a histogram is defined by a 33-dimensional vector. In a nutshell, according to the shape of a histogram, one can know if the surface around a point looks like a plane, a cavity, a hump... As reduced sets of points on which the FPFH are calculated, we use simplified versions of S_1 and S_2 , called S_1^* and S_2^* . S_1^* and S_2^* are simply obtained by subdividing the bounding box of S_1 and S_2 into a set of small cubes, named voxels hereinafter. All points in a voxel are removed except the closest point to the centroid. This process is also referred to as down-sampling the point cloud. Figure 2 displays the reduced point clouds obtained for different voxel sizes (VS) for the example molecule me-indoxam.

A FPFH is computed for each point of the reduced point clouds. It results in two finite sets of FPFH. Then, the RANSAC algorithm is used to match them.^[36] RANSAC is an iterative model fitting framework widely used. At each iteration, it consists in picking N points from the reduced cloud S_1^* , and in identifying N corresponding points in the cloud S_2^* , by querying their nearest neighbor in the FPFH space. At this step, it is possible to compute the transformation matrix T , to align S_1^* and S_2^* . Before that, two checks are carried out to make sure the suggested transformation is meaningful in the 3D space.

- i) The first test estimates the proximity of points by applying T on the N points of S_1^* , and checking that the distance between these points and their corresponding points in S_2^* is lower than a given threshold ($VS \cdot 1.5$).
- ii) The second test checks that inter-distances (edge lengths) between the N points of S_1^* are close to the inter-distances between the N corresponding points in S_2^* . This is considered as satisfactory if each of them does not differ by more than 10%.

If these two tests are successful, the matching at this iteration is validated, the transformation matrix is applied to S_1^* and, a RMSE value (Root Mean Square Error) is computed between it and S_2^* . The maximal number of querying is 400000 but if tests are successful 1000 times, the procedure stops, the alignment with the smallest RMSE value is selected, and the associated transformation matrix T is applied to the entire point cloud S_1 . As a result, we obtain the point cloud S_1^g globally aligned on the other one (Figure 1c). From now on, this initial superimposition step is called GLOBAL.

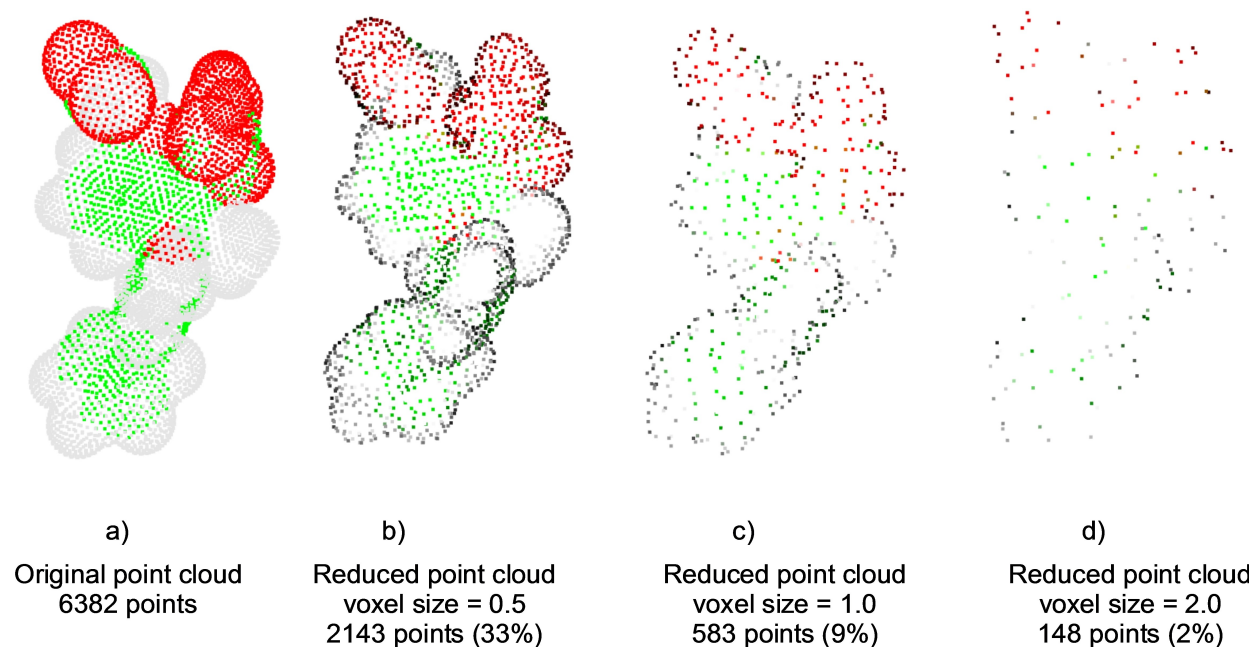


Figure 2. Effect of the voxel size (VS) on the number of points in the reduced point cloud for the molecule me-indoxam.

c) *Local refinement based on geometry and physico-chemical properties.* In many cases, an alignment obtained with a “global” approach is perfectible and can be improved by a “local” alignment.

3D shape is an important feature in molecular recognition but the distribution of physico-chemical properties in space as it is represented by pharmacophore features is determinant for the binding affinity of ligands. Such characteristics can be conveniently used to improve the alignment of molecules, by optimizing the matching of properties along with the geometry. To reach this objective, we chose to use an ICP variant for colored point clouds developed by Park et al.^[37] The interest of this approach is to obtain a tighter alignment by matching both geometry and color of points. This method is particularly appropriate for our molecular surfaces as each point has a label that can be seen as a coloration. We indicated in (a) that each input point is described by its position in 3D space (x, y and z) and its label indicating the type of its closest atom. We categorize atom types into four classes that will be seen as colors during this stage. Examples of colored clouds resulting from this classification are visible in Figure 1d. In our implementation, labels aim to recapitulate typical pharmacophore features such as aromatic, lipophilic and polar groups.^[38,39]

- The first class includes non polar hydrogen (H) and halogen atoms excepting fluorines (Cl, Br and I). Hydrogen and halogen atoms are molecule endings. They are the most frequent atoms that contribute to the surface geometry and coloration, and thus, highlight the apolar

surface area. Points belonging to this class are colored in white in our study (Figure 1d).

- The second class includes polar atoms able to be involved in hydrogen bonds such as N, O, S, H (if linked to N or O) and F.^[40] Points belonging to this class are colored in red in our study (Figure 1d).
- The third class includes “skeleton elements” such as C, P and B. Points belonging to this class are colored in green in our study (Figure 1d). This class sometimes contributes to the surface. For example, aromatic carbons may contribute to the surface by exposing a typical pattern formed by two parallel green patches as shown in Figure 1d. On the contrary, a sp³ carbon does not contribute significantly to the surface because hydrogen atoms that are linked to it masked it.
- The fourth class includes all elements not listed in the first three classes. This class is empty for most small organic molecules in medicinal chemistry. Points belonging to this class are colored in blue in our study.

Globally, ICP consists in i) selecting a set of points in the clouds *Source* and *Target*; ii) matching the selected points in *Source* with the “optimal” corresponding points in *Target* (historically the “closest points”); iii) estimating the transformation matrix *T* that gives the best alignment of *Source* and *Target*, by minimizing iteratively a cost function *E* depending on the corresponding points; iv) transforming *Source* with *T*, and going back to ii) until reaching convergence, or a maximum number of iterations.

The ICP variant for colored point clouds^[37] employed in SENSAAAS uses a cost function *E* that takes both geometric and color information for alignment: $E(T) = \delta E_g(T) + (1 - \delta)$

E_c , where δ , set to 0.8, balances the influence of the two terms. E_g is the objective function proposed in the *point-to-plane* ICP variant.^[29] The last matches the points by a procedure called “normal shooting” from *Source* to *Target* (i.e. projection in the normal direction), and then determines the matrix T that minimizes the RMSE between the corresponding pairs. E_c is the term that weights the alignment of points of same color. It measures the difference between the color of the *Source* points and the color of their projection on the tangent plane to their corresponding *Target* points. To summarize, the local registration stage of SENSAAAS consists in:

- i) taking as input the two point clouds aligned by GLOBAL (S_1^g and S_2);
- ii) matching the reduced point clouds S_1^{*g} and S_2^* ;
- iii) applying this ICP variant and stopping when the number of iterations reaches 100, or when the relative RMSE between two iterations is lower than $1e-6$.

As output, we obtain the final transformation matrix to superimpose S_1 to S_2 . From now on, this local registration stage is called COLOR.

2.2 Alignment Evaluation

To evaluate alignments, it is common to compute RMSE values and fitness scores. A RMSE indicates how close the points of two aligned clouds are paired whereas a fitness indicates how many points are paired. Points are considered paired if their distance is lower than a given threshold. In our implementation, we set the threshold value to 0.3 because it is the average distance between two adjacent points in our original point clouds. The fitness equals 1 if the alignment is perfectly achieved for two identical 3D molecular structures. Increasing this value would artificially increase the fitness by pairing more distant points. Fitness is also essential to calculate how many points with similar colors finally match. We defined three different fitness scores:

– $gfit$ is the ratio between the number of points of the transformed *Source* that match points of the *Target*, and its total number of points:

$$gfit = \frac{\text{number} - \text{of} - \text{matching} - \text{points} - \text{in} - \text{Source}}{\text{Total} - \text{number} - \text{of} - \text{points} - \text{in} - \text{Source}} \quad (1)$$

– $cfit$ is the sum of the fitness for each class, to specifically evaluate the matching of the colored points ($k=4$ classes):

$$cfit = \frac{\sum_{i=1}^k nb - \text{matching} - \text{points} - \text{in} - \text{Source} - \text{for} - \text{class} - i(2)}{\sum_{i=1}^k \text{number} - \text{of} - \text{points} - \text{in} - \text{class} - i}$$

– $hfit$ is the sum of the fitness for each class except the first class, to specifically evaluate the matching of polar and aromatic points (classes 2, 3 and 4):

$$hfit = \frac{\sum_{i=2}^k nb - \text{matching} - \text{points} - \text{in} - \text{Source} - \text{for} - \text{class} - i(3)}{\sum_{i=2}^k \text{number} - \text{of} - \text{points} - \text{in} - \text{class} - i}$$

Those scores are similar to a Tversky coefficient^[41] tuned to evaluate the embedding of one object into another one, here the molecules *Source* and *Target*. This implies that scores calculated with respect to *Target* may differ from those calculated with *Source*. Moreover, the smallest point cloud of the two will always obtain the highest fitness score as more points are paired, proportionally. In the following, “GLOBAL $gfit$ ” relates to the (intermediary) fitness value calculated after GLOBAL and terms $gfit$, $cfit$ or $hfit$ means that the whole registration algorithm (GLOBAL+COLOR) is applied before calculating fitness values.

2.3 Choice of the Voxel Size (VS)

To be efficient in a context of molecular surfaces, we had to tune the voxel size (VS) for GLOBAL and COLOR. All the functions used in GLOBAL and COLOR are implemented with functions of the open-source library Open3D0.7.0.^[42] As described in the workflow description, the VS influences the number of points in the reduced point cloud. It is important to select a value consistent with the density of the point cloud. In our study, points are uniformly distributed on the vdW surface and the distance between two adjacent points is ~ 0.3 Å (NSC program (version 2.0) from Eisenhaber et al.^[33]). Table 1 shows the average percentage of remain-

Table 1. Percentage of remaining points after reducing a point cloud.

Voxel size	Percentage of remaining points
0.1	99–100
0.2	95–96
0.3	67–70
0.4	43–46
0.5	29–31
0.6	21–22
0.7	16–17
0.8	12–13
0.9	10
1.0	8
1.1	7–8
1.2	5–6
3	0.9–1.1
7	0.3–0.6

ing points for various VSs. As expected, a $VS \leq 0.3$ does not significantly lower the number of points since more than

70% of initial points are kept. On the contrary, a $VS \geq 1$ leads to the removal of more than 90% of initial points. If too many points are removed, the matching fails. To select an appropriate VS value, we analyzed pairwise alignments for VS values ranging from 0.2 to 1.2 (*i.e.* from 95% to 5% of remaining points) with an increment of 0.1.

In the following examples, we studied self-matchings of molecules sorbate and Imatinib with rotated and translated versions of themselves (named “sorbate-moved” and “Imatinib-moved”, respectively). In such experiments, the perfect superimposition is expected, and *gfit* values must equal 1. We also assessed the substructure matching of Imatinib with a part of itself, Imatinib-part2 (Figure S1a and S1b in the Supporting Information). Figure 3 displays resulting alignments, and *gfit* scores plotted in function of VS . First, initial alignments obtained with GLOBAL are always improved by COLOR (for all runs and any VS): *gfit* is always higher than GLOBAL *gfit* (Figure 3a, b and c), and the final RMSE is always lower than the intermediary RMSE obtained by GLOBAL. Second, results show that there is no single value of VS for which the best solution could be found for all cases. For example, when VS equals 0.2 or 0.3, perfect superimpositions are observed for the sorbate case, but not for Imatinib. In the substructure matching (Figure 3c), GLOBAL *gfit* and *gfit* scores vary from one test to another, but good alignments and scores are obtained when VS is 0.4, 0.6, 0.7 or 0.8. In that case, VS values larger than 0.8 lead to more erratic results. In conclusion, VS strongly influences the accuracy of the alignments, and a unique and optimal value cannot be known *a priori*. Thus, as the down-sampling is reproducible for a given point cloud and for a given VS , we finally chose to successively execute eleven alignments with VS ranging from 0.2 to 1.2 with an increment of 0.1 and, to keep the best alignment. We consider that an alignment is good if it has a good shape overlap (high *gfit*) and a good superimposition of common pharmacophoric features (high *hfit*). Thus, the alignment with the largest sum of *gfit* and *hfit* values is considered as the best one.

2.4 Computing Time

All experiments have been carried out on a Dell PowerEdge R940 (Intel Xeon Gold 5220). SENSEAAS is running using the Debian version 5.3.15-1 of Linux. The processing time for molecules used in present tests takes between 3 to 25 seconds per run (Figure 4). Computation times of 25 seconds occur when a molecule is aligned on a shifted version of itself, for example in self-matching tests. However, aligning a molecule on itself is a rare and specific experiment, per se. In our study, we only used it as a control. No optimization of the code was carried out so far but an optimization of the point cloud preprocessing and the use of Python threads over voxel size could bring a substantial improvement in computation times.

2.5 Data Visualization

In this study, the Open3D function *visualization.draw_geometries* and the software PyMOL (version 1.3) were used to visualize alignments and generate illustrations.

2.6 SENSEAAS Validation with Self- and Substructure Matchings

Because of the stochastic nature of the registration methods used in SENSEAAS, we carried out an analysis of its reproducibility. In studied examples, experiments were repeated one hundred times, and resulting superimpositions and fitness scores were examined. Figure S1 in the Supporting Information displays the structures of molecules: the previously analyzed sorbate/sorbate-moved and Imatinib/Imatinib-moved pairs, and four substructures pairs sorbate/sorbateC, Imatinib/Imatinib-part1, Imatinib/Imatinib-part2 and Imatinib/Imatinib-part3, whose resulting alignments are also easy to validate.

2.7 Benchmarking AZ Data Set

Tests involving self- and substructure matchings are prerequisites but are not broad enough to evaluate the general applicability of a molecular alignment method. The only experimental evidence of alignments between different molecules is their co-crystallization with the same protein. Therefore, the performance of SENSEAAS was assessed by comparing its alignments using the 121 experimental overlays of the AZ test set available at the CCDC Center.^[31] This data set contains 1465 co-crystallized ligands extracted from the Protein Data Bank (PDB). The authors annotated overlays according to shape, feature and 2D fingerprint similarities between ligands and, finally classified overlays as either easy (22 cases), moderate (73 cases), hard (18 cases) or unfeasible (8 cases) to predict.^[43] For each of the 121 proteins, SENSEAAS was run using each experimental ligand as the *Target* in turn. Then, each resulting pose of other ligands (*Source*) was compared to their experimental pose and scored with the heavy atom root-mean-square deviation (RMSD). It is usually accepted that a pose is correctly predicted if the RMSD is below 2 Å. Results were then compared to those obtained with shaEP and SHAFTS.

2.8 Fragment Alignments

As we aim at developing a method that performs well in scaffold hopping and bioisostere replacement (*i.e.* replacing one chemical group by another one with a different chemotype),^[44,45] we evaluated the SENSEAAS ability to align the bioisosteric fragments tetrazole and carboxylate.^[46,47]

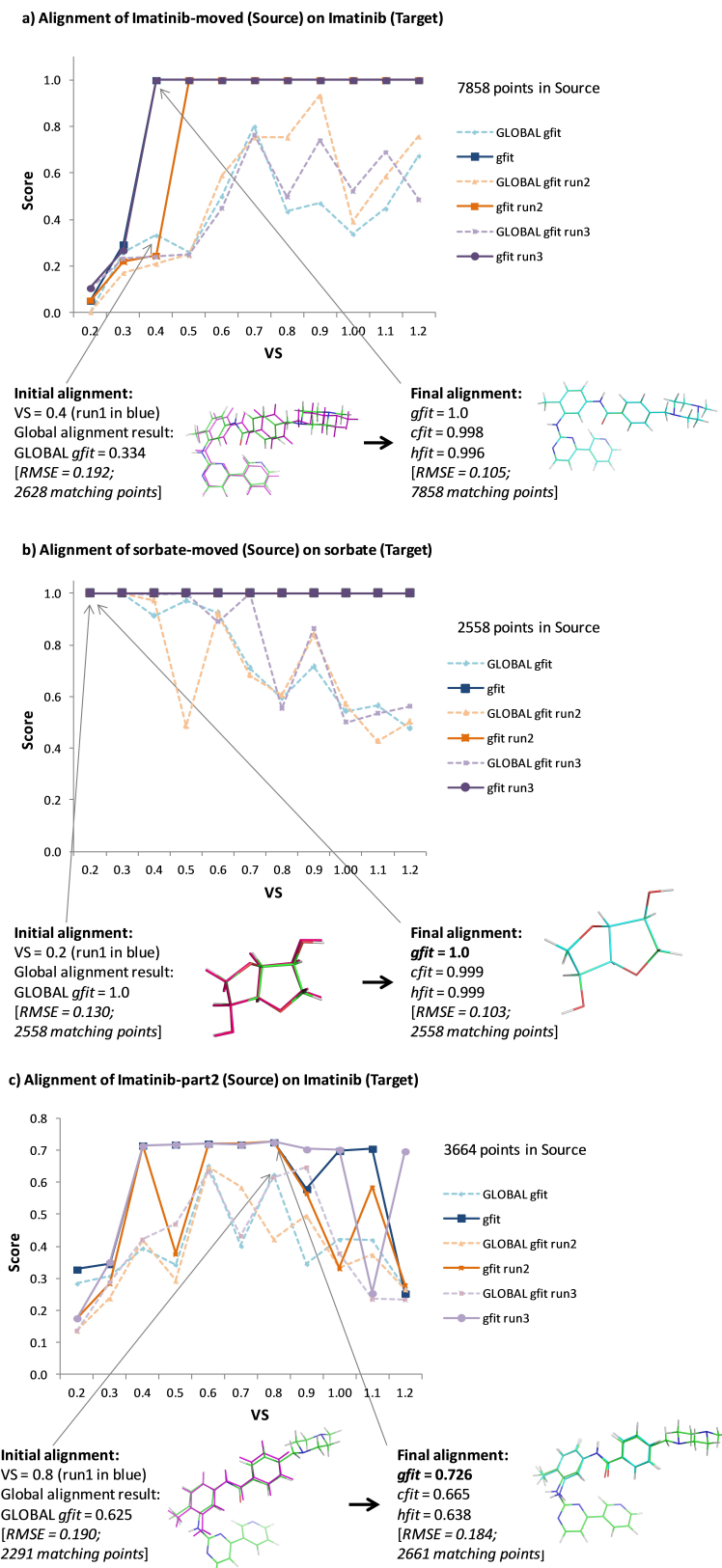
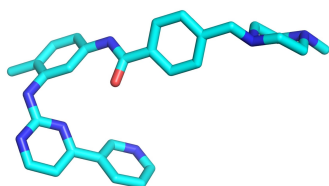


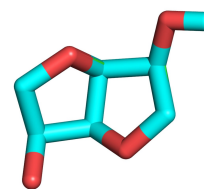
Figure 3. *gfit* versus GLOBAL *gfit* in function of VS. Three runs are plotted (blue, orange and mauve lines). The initial and final alignments are displayed with their fitness scores, number of matching points and RMSE values. The *Source* 3D graph aligned with GLOBAL only is colored in magenta (at left). *Target* and aligned *Source* 3D structures are colored in green and cyan, respectively (at right). (a) Imatinib/Imatinib-moved pair. (b) Sorbate/sorbate-moved pair. (c) Imatinib/Imatinib-part2 pair.

a) Imatinib self-matching



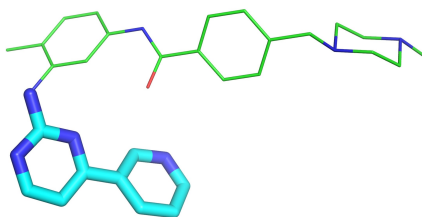
Imatinib fitness: *gfit+hfit* *gfit* *hfit*
 cpu ~ 25s / run **2.0** 1.0 1.0
 100% of times

b) sorbateC (Source) on sorbate (Target)



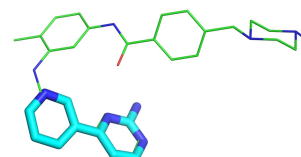
sorbateC fitness: *gfit+hfit* *gfit* *hfit*
 cpu ~ 6s / run **1.870** 0.871 0.999
 100% of times

c) Imatinib-part1 (Source) on Imatinib (Target)

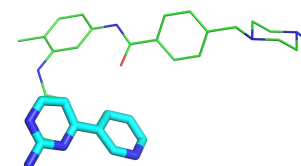


Imatinib-part1 fitness: *gfit+hfit* *gfit* *hfit*
 cpu ~ 4s / run **1.732** 0.916 0.816
 60% of times

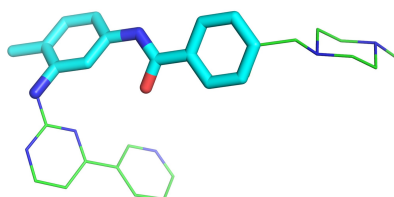
Second ranked alignment
gfit+hfit *gfit* *hfit*
 1.047 0.722 0.325



Third ranked alignment
gfit+hfit *gfit* *hfit*
 1.046 0.702 0.344

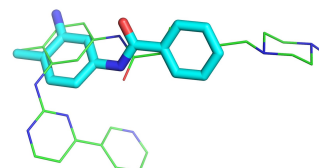


d) Imatinib-part2 (Source) on Imatinib (Target)

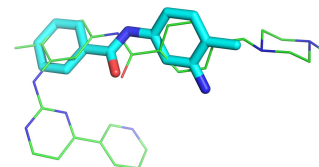


Imatinib-part2 fitness: *gfit+hfit* *gfit* *hfit*
 cpu ~ 4s / run **1.415** 0.721 0.694
 61% of times

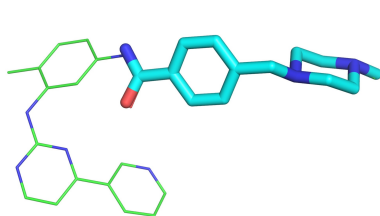
Second ranked alignment
gfit+hfit *gfit* *hfit*
 1.017 0.625 0.392



Third ranked alignment
gfit+hfit *gfit* *hfit*
 1.007 0.575 0.432

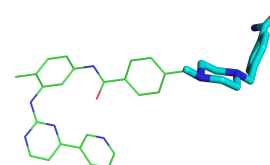


e) Imatinib-part3 (Source) on Imatinib (Target)



Imatinib-part3 fitness: *gfit+hfit* *gfit* *hfit*
 cpu ~ 5s / run **1.661** 0.913 0.748
 90% of times

Second ranked alignment
gfit+hfit *gfit* *hfit*
 0.633 0.544 0.089



Third ranked alignment
gfit+hfit *gfit* *hfit*
 0.499 0.487 0.012

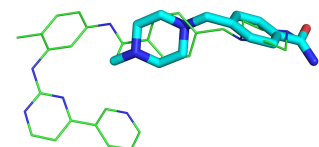


Figure 4. Results of self- and substructure-matchings. Each run was performed 100 times with a random orientation for the *Source* each time. a) Imatinib self-matching. b) sorbateC (*Source*) on sorbate (*Target*). c) Imatinib-part1 (*Source*) on Imatinib (*Target*). d) Imatinib-part2 (*Source*) on Imatinib (*Target*). e) Imatinib-part3 (*Source*) on Imatinib (*Target*). Corresponding superimposition(s) of 3D graphs, the percentage of times, the mean CPU time and the scores *gfit+hfit*, *gfit* and *hfit* are indicated. *Target* 3D structures are colored in green and *Source* 3D structures are colored in cyan. Hydrogen atoms are hidden in 3D graphs.

Table 2. Number of reproduced ligands and mean RMSD values for each category for SENSEAAS, shaEP and SHAFTS. Average computation time per pairwise alignment is indicated. It takes into account the entire procedure including the calculation of partial charges by cxcacalculator^[49] for programs shaEP and SHAFTS.

Category	Number of ligands	Reproduced ligands (RMSD \leq 2.0 Å)			Mean RMSD		
		SENSEAAS	ShaEP	SHAFTS	SENSEAAS	ShaEP	SHAFTS
easy	185	183 (98.9%)	178 (96.2%)	181 (97.8%)	0.37	0.5	0.46
moderate	991	927 (93.5%)	923 (93.1%)	922 (93%)	0.8	0.88	0.86
hard	190	148 (77.9%)	151 (79.4%)	144 (75.8%)	1.65	1.8	1.71
unfeasible	99	46 (46.4%)	53 (53.5%)	31 (31.3%)	2.69	2.58	3.47
Total =	1465 (100%)	1304 (89%)	1305 (89%)	1278 (87%)	0.97	1.06	1.09
average run time/pairwise alignment		5.7 s	0.51 s	0.46 s			

Such submatching properties requires that SENSEAAS must not only align molecules of the same size successfully, but also align substructures and bioisosteric fragments correctly, even if they are small. To this end, we selected the drug Valsartan which contains both a carboxylic acid and a tetrazole ring (Figure S1d–f). Results were compared to those obtained with shaEP and SHAFTS.

3 Results and Discussion

3.1 SENSEAAS Validation with Self- and Substructure Matchings

Figure 4 shows results of pairwise alignments for self- and substructure-matchings that are easy to validate. Because of the stochastic nature of the registration methods, those tests were repeated 100 times to identify possible alternate alignments. For the self-matching of Imatinib, the perfect alignment is always found (Figure 4a). Similarly, the sorbate/sorbateC alignment is a trivial test that SENSEAAS successfully achieves 100% of times (Figure 4b). Regarding substructure matchings, the best, expected, alignment is found most of the times and ranked first (Figure 4c, d and e, alignments at left). Over the repeats, alternate alignments are generated but their fitness scores are always lower (Figure 4c, d and e, alignments at right). We then assessed the ability of SENSEAAS to generate the expected alignment when a different conformation of the *Source* is used. Figure S2a in the Supporting Information shows the best alignment of such a conformer of the *Source* Imatinib-part3 that has a RMSD of 1.99 Å with the original substructure conformation. This alignment shows the expected matching of the aromatic and piperazine rings, but an inversion of the amide function. This is caused by the conformer that differs from the original Imatinib structure, precluding a perfect matching. Therefore, moderate variations in conformation do not prevent the identification of the correct alignment by SENSEAAS. We further investigated the capacity of SENSEAAS to identify the best Imatinib-part3 conformer from a library of 74 conformers generated with RDKit Open-Source Cheminformatics Software (<http://rdkit.org>).^[48] Figure S2b and S2c in the Supporting Information displays

score values for the 74 conformers and the first ranked superimposition, respectively. Conformer number 36 is the first ranked solution and is perfectly aligned with the Imatinib structure. This result shows that aligning a library of conformers with SENSEAAS succeeds in identifying the conformer allowing the best alignment.

3.2 Benchmarking AZ Data Set

As the only experimental evidence of alignments between different molecules is their co-crystallization with the same protein, we tested the ability of SENSEAAS to reproduce the superimposition of X-ray structures of the benchmarking AZ data set. Then, to assess its performance, we compared results with those generated with shaEP^[6] and SHAFTS.^[7] In this experiment, SENSEAAS was run ten times and the best ranked solution, *i.e.* best *gfit* + *hfit* value, was retained. shaEP and SHAFTS, being deterministic methods, were run once with the default settings and with partial charges calculated with cxcacalculator.^[49] Contrary to shaEP and SHAFTS, SENSEAAS does not require the calculation of partial charges. Figure 5 shows the distribution of the lowest RMSD obtained for the 1465 molecules. The three methods produce comparable distributions with small differences within each interval in favor of one method or the other. The mean RMSD value, computed over the entire data set, is also similar for the three methods, with a slight advantage for SENSEAAS with a value of 0.97 Å against 1.06 Å for shaEP and 1.09 Å for SHAFTS (Table 2). When considering a RMSD value \leq 2.0 Å, the number of ligands that are reproduced is still comparable: 1304 with SENSEAAS (89%), 1305 with shaEP (89%) and 1278 with SHAFTS (87%). This indicates that it exists at least one ligand structure that allows the correct alignment for nearly all of the ligands. However, when considering only the first, most accurate, RMSD interval [0–0.5], SENSEAAS outperforms the other methods with 941 molecules against 831 and 837 with shaEP and SHAFTS, respectively. Table 2 details the number of reproduced ligands and the mean RMSD value for each category. As expected, the number of reproduced ligands decreases according to the difficulty regardless of the method. More than 93% of ligands are reproduced in the

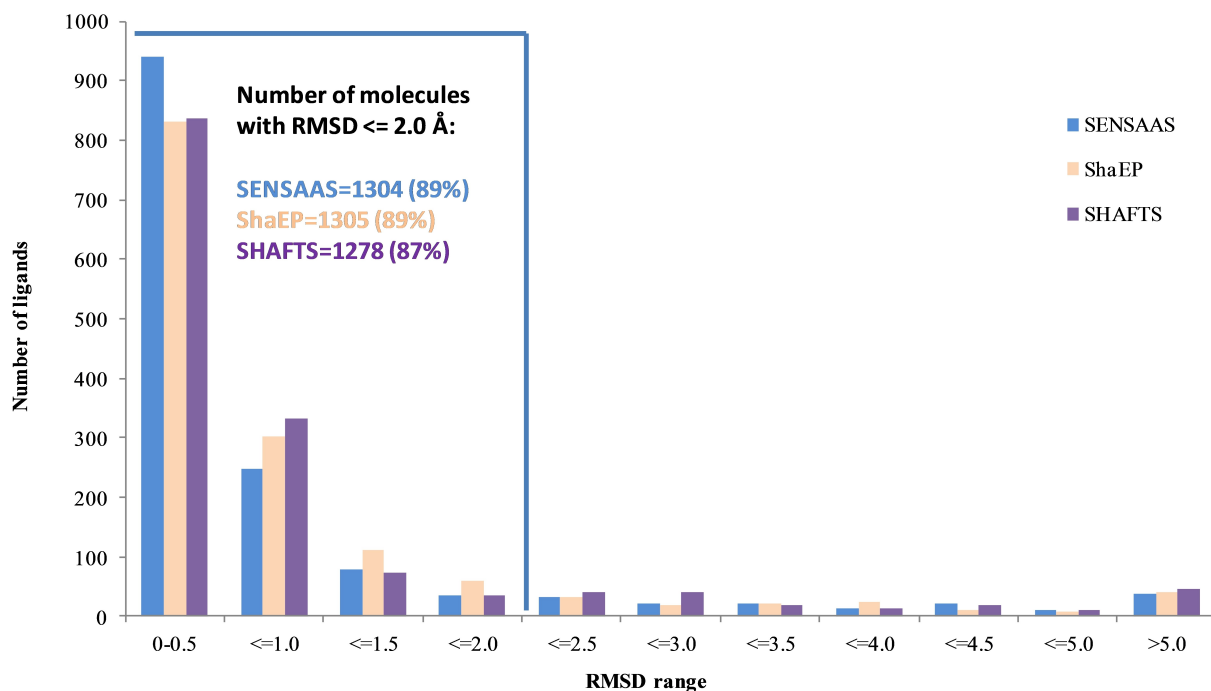


Figure 5. Distribution of the lowest RMSD obtained for each molecule of the AZ data set (1465 ligands). SENSEAAS results are colored in blue, shaEP results are colored in beige and SHAFTS results are colored in mauve. Inset: a large majority of molecules are successfully reproduced with a $\text{RMSD} \leq 2.0 \text{ \AA}$.

easy and moderate categories, between 75 and 79% in the hard category and between 31% and 53% in the unfeasible category. Consistently, the mean RMSD value for each category increases according to the difficulty. Of note, SENSEAAS generates more accurate results than shaEP and SHAFTS in the easy, moderate and hard categories and intermediate results in the unfeasible category. Figure S3 in the Supporting Information details results for the 121 overlays. The average computation time of SENSEAAS is in the order of a few seconds (Table 2). It is longer than that of shaEP and SHAFTS (by a factor of 10). SENSEAAS is not fully competitive in its current implementation with established Gaussian-shape methods in speed. Regarding the surface-based methods, the computation time of SENSEAAS is comparable to those published by Baum et al.,^[20] which were faster than other surface-based algorithms.

3.3 Fragment Alignments

In this work, we were particularly interested in the biososterism between the tetrazole and the carboxylate function (Figure S1d and e in the Supporting Information). Indeed, although these two functional groups are structurally different, several studies have demonstrated that they may be interchangeable in a bioactive molecule.^[46,47] Figure 6 shows pairwise alignments between a tetrazole or a carboxylate fragment and the drug Valsartan (Figure S1f).

As Valsartan contains both tetrazole and carboxylate functions in its structure, alignments with the two fragments are particularly interesting to evaluate possible alternate solutions. Our results show that the expected alignment between the tetrazole fragment and the Valsartan's tetrazole group is ranked first ($gfit + hfit = 1.520$) and that two other superimpositions with lower scores are generated: the alignment of the tetrazole fragment with the carboxylate function ($gfit + hfit = 0.826$) and with the alkane chain of Valsartan ($gfit + hfit = 0.756$) (fragment in cyan, magenta and orange in Figure 6a, respectively). Therefore, three solutions occur: the self-matching (ranked first), the biosisoteric matching (ranked second), and a geometric matching only with a $hfit$ score close to 0 (ranked third). Consistent results are obtained when the carboxylate fragment is aligned on Valsartan. Indeed, the self-matching is ranked first, the biosisoteric matching is ranked second, and the geometric matching is ranked third with a $hfit$ score of 0 (fragment in cyan, magenta and orange in Figure 6b, respectively).

The stochastic nature of SENSEAAS makes it possible to suggest some alternate superimpositions, contrary to deterministic methods that provide only one alignment. As a proof, Figure 6 also displays the alignments generated with shaEP and SHAFTS. shaEP correctly self-aligns the carboxylate fragment on Valsartan but not the tetrazole fragment whereas SHAFTS correctly aligns the ring of the tetrazole fragment (apart from the methyl carbon) but not

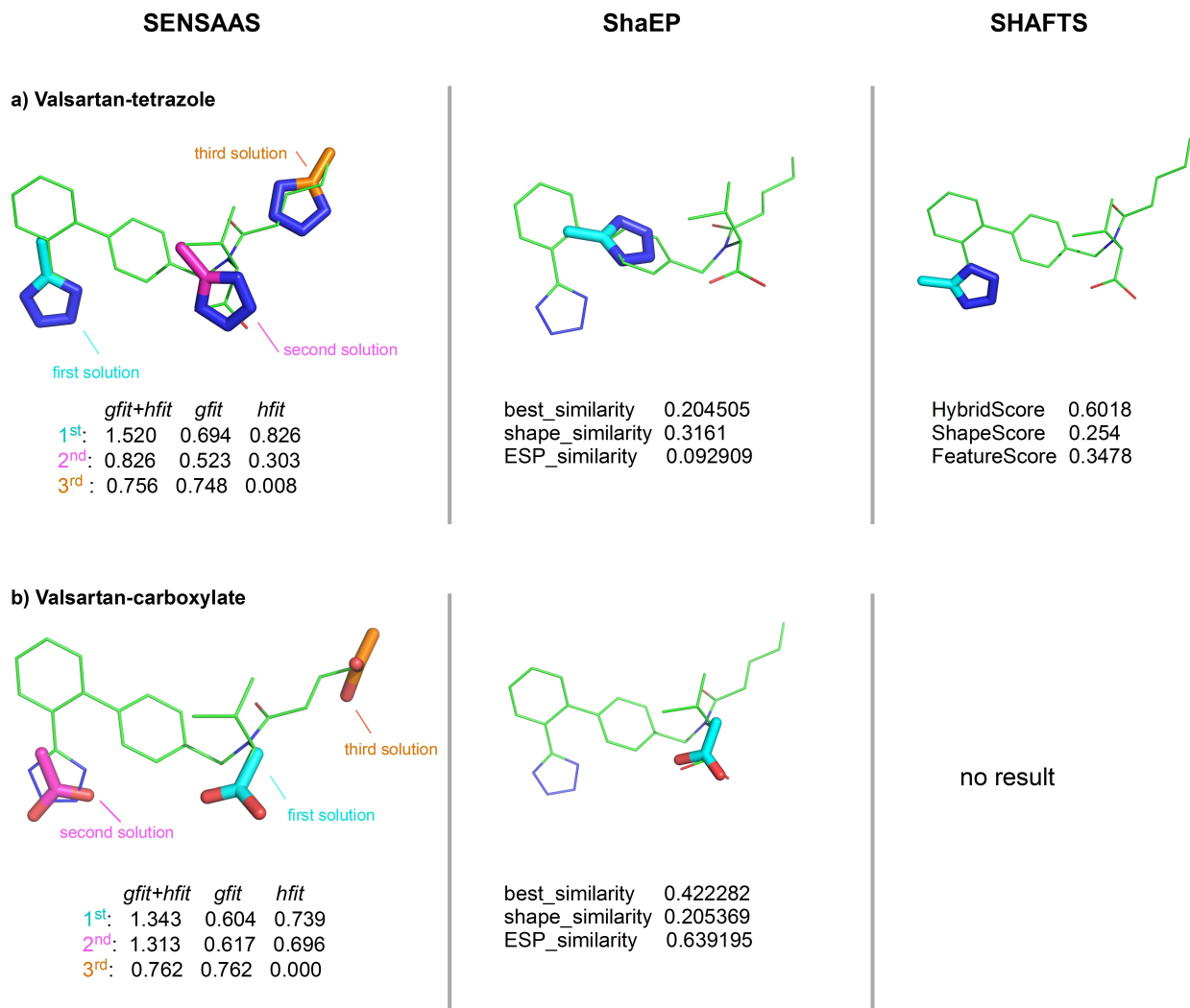


Figure 6. Results of fragment alignments. a) 3D graph superimpositions of Valsartan and tetrazole. b) 3D graph superimpositions of Valsartan and carboxylate. Fitness scores *gfit*, *cfi* and *hfit* are indicated for SENSEAAS. Fitness scores *Best_similarity*, *shape_similarity* and *ESP_similarity* are indicated for shaEP. Fitness scores *HybridScore*, *ShapeScore* and *FeatureScore* are indicated for SHAFTS. Target 3D structures are colored in green and Source 3D structures are colored in cyan (best ranked alignment), in magenta (second ranked alignment) or in orange (third ranked alignment).

the carboxylate fragment. Actually, these methods do not generate bioisosteric matchings. Similar results were obtained by SENSEAAS when fragments were aligned on the drug Adapalene (Figure S4 in the Supporting Information). In this example, neither shaEP nor SHAFTS perfectly aligns fragments on their counterparts. SHAFTS displays a correct alignment for the two oxygens of the carboxylate fragment on Adapalene but not for the methyl carbon perpendicularly oriented.

In summary, these experiments show that SENSEAAS is able to find the correct alignments but also to provide realistic superimpositions by identifying local similarities. Results also show that the algorithm appropriately ranks alternate solutions. This submatching property allows us to

envisage more complete studies in the search for bioisosteric replacements.

3.4 Discussion

In our approach, a point cloud contains hundreds to thousands colored points, depending on the down-sampling stage. For a reduced point cloud, between 5 and 95% of the points are kept when VS ranges from 0.2 to 1.2. This ensures to retain a detailed description of the shape and of the distribution of colors during the alignment. Up to now, the number of input points was a limiting factor in point-based surface alignment methods. Strategies were used to reduce the number of points by dividing the surfaces into

patches, and/or by excluding hydrogen atoms, or by using a united atom model to decrease the number of atoms contributing to the surface. While some of these methods use shape descriptors only,^[15,16] others associate shape and physico-chemical properties to describe their patches,^[17,18] or use distinct point clouds for molecular shape and for each physico-chemical property.^[19] Our algorithm, in contrast, uses all atoms and dense point-based surfaces calculated by the program NSC (600 dots per atomic sphere).^[33] The advantages of using dense point clouds are that a property is not associated with a single point in the space (the center of a patch for example), and that finer variations of physico-chemical properties can be mapped onto the surface. In the current setting, points are colored according to the pharmacophore feature of their closest atoms: polar, apolar, aromatic and other, but this could easily evolve towards more complexity: additional classes and colors, computed molecular properties such as the partial charge, potential energy... Evaluating the impact of such properties on alignments need to be investigated and assessed.

The present results show that SENSEAAS is able to identify the correct alignment of similar structures (self- and substructure matchings) and is able to rank them first. In our approach, the final superimposition strongly relies on a reasonable initial, geometric, alignment obtained with GLOBAL. This property is consistent with the importance of the shape in molecular interactions and is supported by the performance of methods that only use shape similarities.^[1,4,15,16] However, intermolecular interactions such as electrostatic, hydrophobic or hydrogen bondings must be also taken into account for stabilizing the recognition. Thus, aligning shapes in concert with physico-chemical features is expected to provide more accurate comparisons between molecules. A strength of our approach is that the initial alignment can be improved by matching both geometric and color information, to get alignments that GLOBAL could not provide on its own. We also showed that the combination of the scores *gfit* and *hfit* as final selection criterion allows us to discriminate alternate alignments and, specifically, to favor alignments with matching colored features. The general applicability of our alignment method was assessed by using the benchmarking AZ data set. Results showed that SENSEAAS provides accuracy performance equivalent to that of the reference methods shaEP and SHAFTS, but also that it generates more accurate alignments in the first precision interval ($\text{RMSD} \leq 0.5 \text{ \AA}$).

The stochastic nature of point set registration is known. In some usage cases, we would prefer a deterministic approach to always produce the same result, but, we can also take advantage of this stochastic property to suggest several realistic solutions when searching for bioisosteres or scaffold-hopping, for instance. The example of the alignment of the tetrazole fragment on the drug Valsartan shows that the fragment can match itself but also the carboxylate function as a bioisostere, which was intended.

As a result of this study, the applicability of SENSEAAS is in the field of lead optimization where scaffold hopping and bioisosteric replacement properties of a method are out of importance to identify promising compounds. Indeed, the execution speed of SENSEAAS in its current implementation is prohibitive for virtual screening of large chemical databases. However, even if SENSEAAS is not yet completely optimized and is not fully competitive with established Gaussian-shape methods in speed, we think that this approach provides a relevant contribution to shape-based alignment methods.

4 Conclusions

Molecular similarity is a central concept in drug discovery. A wide range of methods have been developed to describe molecules and to assess similarity by using representations such as the chemical graph in 2D or 3D, or the molecular shape. In this study, we investigated the use of 3D point set registration methods to identify similarities between colored point-based molecular surfaces. The resulting workflow, SENSEAAS, was evaluated and validated against several examples, including the benchmarking AZ data set. It shows potential for molecular similarity evaluation, scaffold hopping and bioisosteric replacement. In particular, as it uses open-source libraries and programs, it can be easily deployed for other pairwise comparison of shapes such as peptides, proteins, cavities (the negative image of the protein surface), or to search for shape complementarity.

Supplementary Materials

Structure of tested molecules, alignment results of the drug Imatinib with conformers of the substructure Imatinib-part3, results of the benchmarking AZ data set and results of fragment matching tests.

Abbreviations

SENSEAAS	SENSitive Surface As A Shape
GLOBAL	Global registration
COLOR	Colored point cloud registration
VS	voxel size
<i>gfit</i>	geometric fitness

Conflict of Interest

None declared.

Acknowledgements

The authors would like to thank Dr Antonini, M. (Université Côte d'Azur, CNRS, I3S) for fruitful discussions on the project. Authors D. Douguet and F. Payan received funding from the French National Research Agency (ANR) under the project 'Investissements d'Avenir' UCA^{JEDI} with the reference number ANR-15-IDEX-01.

References

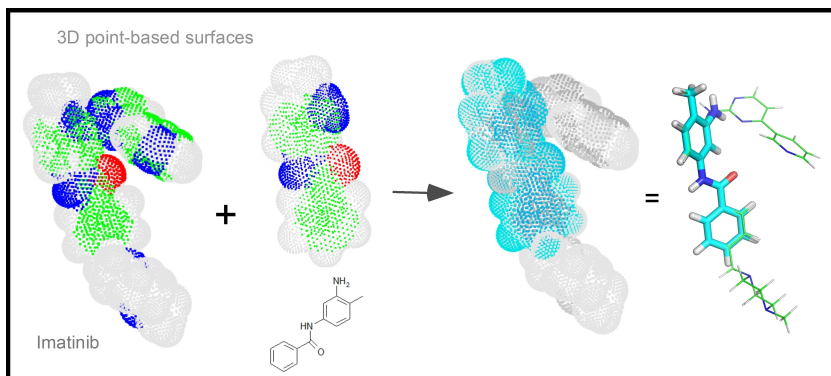
- [1] S. Kortagere, M. D. Krasowski, S. Ekins, *Trends Pharmacol. Sci.* **2009**, *30*, 138–147.
- [2] P. J. Ballester, W. G. Richards, *J. Comput. Chem.* **2007**, *28*, 1711–1723.
- [3] OpenEyeScientificSoftware ROCS 3.1.0.
- [4] J. A. Grant, M. A. Gallardo, B. T. Pickup, *J. Comput. Chem.* **1996**, *17*, 1653–1666.
- [5] A. Kumar, K. Y. J. Zhang, *Front. Chem.* **2018**, *6*, 315.
- [6] M. J. Vainio, J. S. Puranen, M. S. Johnson, *J. Chem. Inf. Model.* **2009**, *49*, 492–502.
- [7] X. Liu, H. Jiang, H. Li, *J. Chem. Inf. Model.* **2011**, *51*, 2372–2385.
- [8] L. Mavridis, B. D. Hudson, D. W. Ritchie, *J. Chem. Inf. Model.* **2007**, *47*, 1787–1796.
- [9] C. Cai, J. Gong, X. Liu, H. Jiang, D. Gao, H. Li, *J. Mol. Model.* **2012**, *18*, 1597–1610.
- [10] A. S. Karaboga, F. Petronin, G. Marchetti, M. Souchet, B. Maigret, *J. Mol. Graphics Modell.* **2013**, *41*, 20–30.
- [11] M. Novotni, R. Klein, *Eighth ACM Symposium on Solid Modeling and Applications*, Seattle, WA **2003**, 216–225.
- [12] R. J. Zauhar, G. Moyna, L. Tian, Z. Li, W. J. Welsh, *J. Med. Chem.* **2003**, *46*, 5674–5690.
- [13] H. Edelsbrunner, E. P. Mücke, *ACM Trans. Graph.* **1994**, *13*, 43–72.
- [14] J. A. Wilson, A. Bender, T. Kaya, P. A. Clemons, *J. Chem. Inf. Model.* **2009**, *49*, 2231–2241.
- [15] B. B. Goldman, W. T. Wipke, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 644–658.
- [16] D. A. Cosgrove, D. M. Bayada, A. P. Johnson, *J. Comput.-Aided Mol. Des.* **2000**, *14*, 573–591.
- [17] T. E. Exner, M. Keil, J. Brickmann, *J. Comput. Chem.* **2002**, *23*, 1176–1187.
- [18] C. Hofbauer, H. Lohninger, A. Aszodi, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 837–847.
- [19] D. Baum, PhD thesis **2007**.
- [20] D. Baum, H. C. Hege, *Computational Life Sciences: Second International Symposium*, ComplLife 2006 **2006**, Vol. 4216, 183–193.
- [21] D. Baum, *Computational Life Sciences: First International Symposium*, ComplLife 2005 **2005**, 198–207.
- [22] T. Krotzky, T. Fober, E. Hullermeier, G. Klebe, *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2014**, *11*, 878–890.
- [23] A. Nicholls, G. B. McGaughey, R. P. Sheridan, A. C. Good, G. Warren, M. Mathieu, S. W. Muchmore, S. P. Brown, J. A. Grant, J. A. Haigh, N. Nevins, A. N. Jain, B. Kelley, *J. Med. Chem.* **2010**, *53*, 3862–3886.
- [24] G. Maggiora, M. Vogt, D. Stumpfe, J. Bajorath, *J. Med. Chem.* **2014**, *57*, 3186–3204.
- [25] I. Akritopoulou-Zanze, J. T. Metz, S. W. Djuric, *Drug Discovery Today* **2007**, *12*, 948–952.
- [26] P. J. Besl, N. D. McKay, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1992**, *14*, 239–256.
- [27] F. Pomerleau, F. Colas, R. Siegwart, *Now Foundations and Trends* **2015**, *4*, 1–104.
- [28] R. Campbell, P. Flynn, *Computer Vision and Image Understanding* **2001**, *81*, 166–210.
- [29] S. Rusinkiewicz, M. Levoy, *Third International Conference on 3-D Digital Imaging and Modeling* **2001**, 145–152.
- [30] B. Bellekens, V. Spruyt, R. Berkvens, M. Weyn, *Fourth International Conference on Ambient Computing, Applications, Services and Technologies* **2014**, 8–13.
- [31] I. Giangreco, D. A. Cosgrove, M. J. Packer, *J. Chem. Inf. Model.* **2013**, *53*, 852–866.
- [32] T. C. Schmidt, D. A. Cosgrove, J. Bostrom, *J. Chem. Inf. Model.* **2018**, *58*, 747–760.
- [33] F. Eisenhaber, P. Lijnzaad, P. Argos, C. Sander, M. Scharf, *J. Comput. Chem.* **1995**, *16*, 273–284.
- [34] A. Bondi, *J. Phys. Chem.* **1964**, *68*, 441–451.
- [35] R. B. Rusu, N. Blodow, M. Beetz, *IEEE International Conference on Robotics and Automation (ICRA)* **2009**, 3212–3217.
- [36] M. A. Fischler, R. C. Bolles, *Commun. ACM* **1981**, *24*, 381–395.
- [37] J. Park, Q.-Y. Zhou, V. Koltun, *IEEE International Conference on Computer Vision (ICCV)* **2017**, 143–152.
- [38] S. Y. Yang, *Drug Discovery Today* **2010**, *15*, 444–450.
- [39] M. P. Sanders, A. J. Barbosa, B. Zarzycka, G. A. Nicolaes, J. P. Klomp, J. de Vlieg, A. Del Rio, *J. Chem. Inf. Model.* **2012**, *52*, 1607–1620.
- [40] N. K. Shinada, A. G. de Brevern, P. Schmidtke, *J. Med. Chem.* **2019**.
- [41] D. Stumpfe, J. Bajorath, *WIREs Comput. Mol. Sci.* **2011**, *1*, 260–282.
- [42] Q.-Y. Zhou, J. Park, V. Koltun, *arXiv* **2018**, 1801.09847.
- [43] I. Giangreco, T. S. Olsson, J. C. Cole, M. J. Packer, *J. Chem. Inf. Model.* **2014**, *54*, 3091–3098.
- [44] N. Brown, *Mol. Inf.* **2014**, *33*, 458–462.
- [45] M. Wirth, V. Zoete, O. Michielin, W. H. Sauer, *Nucleic Acids Res.* **2013**, *41*, D1137–1143.
- [46] J. R. Falck, S. R. Koduru, S. Mohapatra, R. Manne, K. R. Atcha, R. Atcha, V. L. Manthathi, J. H. Capdevila, S. Christian, J. D. Imig, W. B. Campbell, *J. Med. Chem.* **2014**, *57*, 6965–6972.
- [47] C. F. Matta, A. A. Arabi, D. F. Weaver, *Eur. J. Med. Chem.* **2010**, *45*, 1868–1872.
- [48] S. Riniker, G. A. Landrum, *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.
- [49] ChemAxon JChem Base version 19.19.0.

Received: April 9, 2020

Accepted: June 4, 2020

Published online on ■■■, ■■■■

FULL PAPER



Dr. D. Douquet, Dr. F. Payan**

1 – 14

**SENSAAAS: Shape-based Alignment
by Registration of Colored Point-
based Surfaces**

