Similarity Search of Flexible 3D Molecules combining Local and Global Shape Descriptors

Apostolos Axenopoulos , *Member*, IEEE, Dimitrios Rafailidis, Georgios Papadopoulos, Elias Houstis, and Petros Daras, *Senior Member*, IEEE

Abstract— In this paper, a framework for shape-based similarity search of 3D molecular structures is presented. The proposed framework exploits simultaneously the discriminative capabilities of a global, a local and a hybrid local-global shape feature to produce a geometric descriptor that achieves higher retrieval accuracy than each feature does separately. Global and hybrid features are extracted using pairwise computations of diffusion distances between the points of the molecular surface, while the local feature is based on accumulating pairwise relations among oriented surface points into local histograms. The local features are integrated into a global descriptor vector using the bag-of-features approach. Due to the intrinsic property of its constituting shape features to be invariant to articulations of the 3D objects, the framework is appropriate for similarity search of flexible 3D molecules, while at the same time it is also accurate in retrieving rigid 3D molecules. The proposed framework is evaluated in flexible and rigid shape matching of 3D protein structures as well as in shape-based virtual screening of large ligand databases with quite promising results.

Index Terms — Bioinformatics (genome or protein) databases; flexible 3D molecular shape comparison; virtual screening.

– – – – – – – – – – Φ – – – – – – – – – –

1 Introduction

THE THREE DIMENSIONAL STRUCTURE of a biological molecule is very important in order to understand its function and biological action. Comparison of the 3D molecular structures is useful in a variety of applications such as protein function prediction, computer aided molecular design, rational drug design and protein docking. Following the *similarity property principle* [1], according to which similar structures are likely to have similar properties, several approaches for molecular structure comparison have been proposed, using different representations of the molecules. As an example, in rational drug design, the process of *virtual screening* is usually applied, where given a target molecule, a search is performed in a large database for compounds that are most similar to the target. Since these compound databases range from thousands to millions of structures, an ideal method should provide accurate and at the same time rapid similarity matching. Among the various existing structural comparison methods [3] [52], those that are based on comparison of structures by their mainchain orientation [53] or the spatial arrangement of secondary structure [5] are quite slow, thus, similarity search in large molecular databases can be time-consuming. Therefore, in order to accelerate the search time, methods of 3D shape matching have been proposed in the literature.

A. Axenopoulos and E. Houstis are with the Department of Computer & Communication Engineering, University of Thessaly, Volos, Greece (e-mail: axenop@iti.gr; enh@inf.uth.gr)

P. Daras and D. Rafailidis are with the Information Technologies Institute, Centre for Research & Technology Hellas, Thermi-Thessaloniki, Greece (e-mail: daras@iti.gr; drafail@iti.gr)

G. Papadopoulos is with the Department of Biochemistry & Biotechnology, University of Thessaly, Larissa, Greece (e-mail: geopap@med.uth.gr)

1.1 Related Work

Techniques for similarity matching of molecular structures can be classified into different categories based on the molecular representation [2]. These representations may include backbone Ca positions [3], distance maps [4], secondary structure elements [5] and backbone torsion angles [6]. The technique/algorithm that is used for comparison highly depends on the chosen representation. As an example, for backbone representations, a common technique is dynamic programming [3]; spatial arrangements are used with secondary structure elements [5], while Monte Carlo algorithms are used with distance maps [4]. A broad category of techniques for Molecular Shape Comparison (MSC) rely on finding an optimal superposition of the molecules that are compared (superposition methods) [11]. Superposition is also applied to protein structural alignment to compare a pair of structures, where the alignment between equivalent residues is not given a priori [4][64][65][66][67][68]. Although superposition methods are particularly effective (in terms of identifying similarities between molecular structures), they lack efficiency; they are extremely computationally expensive, which makes search in large molecular databases a time consuming task. As the need for rapid and accurate comparison is becoming even more critical, due to the increasing size of the databases, *descriptor-based* methods have been introduced [11][28]. These extract low level features (descriptors) that capture the spatial profile of the molecule as a multidimensional feature vector. In this case, similarity matching is reduced to descriptor comparison using a common distance measure, which obviates the need for superposition. Since the work presented in this paper belongs to the category of *descriptor-based* techniques, a more detailed state-of-the-art analysis of these methods is provided in the sequel.

In shape-based approaches, the molecule is treated as a three-dimensional (3D) object, on which an appropriate algorithm is applied to extract low-level descriptors that uniquely characterize its shape. A common representation that is extensively used is the molecular surface [7]. Considering the molecular surface as input, several features can be generated, such as Spin Images [8] or Shape Histograms [9]. Spin Images are local 2D descriptions of the surface based on a reference frame that is defined by the associated surface points. Shape Histograms, on the other hand, exploit global geometric properties of the molecule captured in the form of a probability distribution sampled from a shape function (e.g. angles, distances, areas). In [10], the 3D molecular surface is given as input and 2D views of the surface are taken from 100 uniformly sampled viewpoints. Comparison is performed by multi-view matching using 2D Zernike moments and Fourier descriptors for each 2D view. Multi-view representation has been proven quite effective for shape matching of 3D objects [59]; however, for most multi-view-based methods, the optimal performance is achieved when the

database objects have symmetries, i.e. in retrieval of generic objects [54]. In the case of molecular shapes, these symmetries are not present. This obstacle could be overcome by using features oblivious to symmetries, such as integrating local features into bag-of-words to represent each view [60]; however, such approaches have not been reported so far to address the MSC problem.

Apart from the molecular surface, other representations are also possible. The method presented in [11][12] describes the shape of a molecule through its set of interatomic distances, which is encoded as a geometrical descriptor vector. The method achieves very fast comparison times and is appropriate for virtual screening problems. Another shape representation for molecular structure comparison is *alpha shapes* [71][72], which provide a coarser representation of the Connolly surface. Due to their high computational complexity, molecular structure comparison algorithms are usually parallelized [73][74] in order to distribute the processing task into several processors, thus, speed-up the matching process. Finally, there is also a category of more recent methods with the ability to identify subtle differences among very similar proteins, which assists in finding small structural variations that create differences in binding specificity [75][76][77]. The latter is particularly interesting, taking into account the fact that the variation of just a few residues can be enough to alter activity or binding specificity.

An interesting category of shape-based approaches comprises methods that extract moments from the 3D object. These have been successfully applied in pattern recognition problems [13]. The moment-based representations result in compact descriptor vectors with high discriminative power. Examples of moments are based on the theory of orthogonal polynomials, such as 2D/3D Zernike moments and Legendre moments [14]. These descriptors allow also reconstruction of the object from its moments [15]. The method in [16] takes as input the volume of the 3D molecular structure producing a new domain of concentric spheres. In this domain, 2D Polar-Fourier coefficients and 2D Krawtchouk moments are applied, resulting in a completely rotation-invariant descriptor vector. Spherical Harmonics have been widely used in molecular similarity comparison problems such as virtual screening [17], protein structure representation and comparison [18] and molecular docking [19][20]. Spherical Harmonics have the advantage of allowing the surface information to be encoded in a compact form as an orthonormal 1D vector of real numbers allowing fast comparison. Their main disadvantages are: a) they represent only star-shape surfaces; and b) the handling of alignment problems is associated with the fast comparison of objects [21]. Recently, 3D Zernike descriptors (3DZD) have been introduced as a representation of the protein surface shape [22]. These are based on a series expansion of a given 3D function. 3DZDs are rotation invariant, with the protein structures not necessarily being

aligned to perform the molecular shape comparison. Another advantage of 3DZDs is that they allow other characteristics of a protein surface, such as electrostatic potentials, to be incorporated into the descriptor vector [22]. 3DZDs have been used in problems of protein structure retrieval [2], protein-protein docking [23] and virtual screening [24] with quite satisfactory results.

In all methods for molecular shape comparison described above, the 3D molecules are treated as rigid objects. A drawback of these approaches is that they are not robust to shape deformations of flexible molecules. Since many molecules are flexible and this flexibility is part of their function, it should by no means be underestimated. To address such problems, methods for non-rigid shape matching should be utilized. Such methods have been introduced to address problems that include articulation of the 3D objects (e.g. different human or animal poses in generic 3D object retrieval), as rigid shape descriptors have been proven inappropriate [30][61]. The two main categories of non-rigid approaches are: a) global-shape-based and b) local-shape-based methods. The former [25][26][27] usually transform the Euclidean space or Euclidean metrics [40] to a metric space where the pairwise distances between points of the 3D object surface are invariant to deformations of the 3D object. These distances are usually accumulated into a histogram, which provides the final descriptor vector. Examples include canonical forms [31], geodesic distances (GD) [32], inner distances (ID) [27] or diffusion distances (DD) [28]. The difference of DD comparing to GD and ID is that DD is computed as the average length of paths connecting two points, while GD and ID represent the length of the shortest path. This makes DD more robust to topological changes and, thus, it has been proven more efficient to flexible molecular shape comparison problems. In [28], the Diffusion Distance Shape Descriptor (DDSD) is a histogram of the diffusion distances between all sample point pairs on the molecular surface. Experiments in a database of flexible molecules show that DDSD outperforms similar approaches.

Local-shape-based methods sample the surface and extract descriptors for each of the sampled local regions. Then, a codebook is created and a bag-of-features method is applied to generate a global shape descriptor [33][34][35]. A main challenge in these problems is the selection of the most appropriate local shape descriptor [61]. Apart from the discriminative ability, the descriptors should fulfil additional criteria such as fast descriptor extraction, compactness and rotation invariance. Several descriptors have been proposed that fulfil the above selection criteria. The Shape Impact Descriptor (SID) was introduced in [69] as a shape similarity measure for 3D objects and it is based on the idea that objects of similar shape will have similar surrounding fields created by the insertion of the 3D object in the space. The Local Spectral Descriptor has been proposed in [33] for retrieval of non-rigid 3D meshes and it is based on the extraction of geometric de-

scriptors, that is eigenvectors of the Laplace-Beltrami operator (LBO), from a surface patch centered around a sample point on the mesh. The Surflet-Pair-Relation Histograms method was introduced in [37] for global shape representation; furthermore this method was exploited in [63] as a local feature for non-rigid 3D object retrieval. It computes intrinsic geometric properties (azimuthal angle, cosine of polar angle, direction and distance) between pairs of oriented surface points in a 3D surface. Another approach, which was introduced in [29] for fast screening of proteins, is based on extraction of local patches from the protein surface and computation of a geometric fingerprint (distribution of curvatures) for each patch. It exploits local surface similarities and achieves rapid shape comparisons.

Although local-shape-based methods are appropriate for non-rigid shape matching problems, most of them have inferior performance in rigid shape retrieval over rigid methods [35]. In fact, only few methods achieve high performance in both rigid and non-rigid 3D objects [60][62]. It has been recently proven that combining multiple shape descriptors can significantly improve the performance of rigid 3D shape retrieval [36]. In [35], a combination of global and local features is proposed, where the Local Distance Feature (LDF) enhances the local descriptors extracted in 3D meshes by adding spatial context. LDF combines local characteristics – as it is computed on uniformly-sampled keypoints of the 3D surface – with global characteristics – as it takes into account the set of diffusion-like distances from each keypoint to the surface points of the entire mesh. These diffusion-like distances are computed by using a Manifold Ranking algorithm [41]. Following the same concept, a framework that combines multiple shape descriptors to address both rigid and flexible molecular shape matching problems is proposed in this paper.

1.2 Method Overview and Contributions

In Fig. 1, the block diagram of the proposed method is depicted. The crystal structure of the molecule is given as input (e.g. PDB file) and its Solvent Excluded Surface (SESs) is generated in the form of a triangulated mesh. Then, a mesh simplification step is performed on SES, resulting in two sets of points: a set of N_S oriented points and a set of N_K keypoints ($N_K < N_S$) that provide a coarser representation of the 3D molecule. In the descriptor extraction step, two different descriptor vectors are proposed in this work: the *Bag of Augmented Local Descriptor (BoALD)* and the *Modal Representation of the Diffusion-Distance Matrix* (*DDMR* descriptor). These descriptor vectors are combined into a common distance measure in order to calculate the dissimilarity between the query molecule and the molecules of a database.

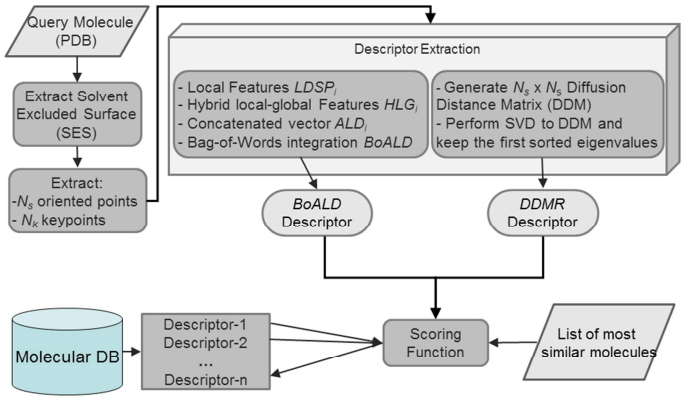


Fig. 1. Block diagram of the proposed method

The main contribution of the proposed work is that it successfully addresses the problem of shape comparison of flexible 3D molecules by combining *a global, a local and a hybrid local-global feature* into a unified descriptor. Such an approach has not been reported so far to the best of the authors' knowledge. Although numerous non-rigid shape matching approaches have been introduced [56][57], which deal effectively with deformations of articulated objects, it cannot be inferred that they are also applicable to flexible molecules [26]. The peculiarities of the molecular shape as well as the complexity of molecular shape deformations, as opposed to deformations of articulated objects (e.g. humans, animals), indicate the need for a method that captures the molecular conformations in a more efficient manner. The method should be highly discriminative and at the same time able to handle shape deformation of molecules with topological changes. Based on the fact that combination of different shape features produces more discriminative descriptors [36][58], in our case, we exploited the properties from a diversity of features, such as a global, a local and a hybrid local-global feature to produce an effective descriptor. Additionally, the global shape feature that has been integrated in our framework is based on the diffusion distance, which is able to capture topological changes in molecular shapes [28]. The proposed unified framework demonstrates superior performance to existing methods for shape comparison of flexible molecules. Experiments performed in a benchmark Database of Macromolecular Movements (MolMovDB) [38] show that our method clearly outperforms other state-of-the-art approaches. At the same time, our method achieves high accuracy in retrieval of rigid molecules as well. More specifically, it outperforms existing molecular shape matching approaches in three datasets. Thus, the proposed framework is applicable to both rigid-body and flexible molecular shape comparison problems.

Additional contributions of our work include:

Introduction of an accurate global shape descriptor, by improving existing work on Diffusion-Distance-based descriptor: the work of [28] based on Diffusion Distances is further extended in this paper. Instead of computing histograms of diffusion distances between all sample point pairs on the molecular surface, we provide a new representation by performing singular value decomposition (SVD) on the matrix that summarizes all point-to-point diffusion distances on the molecular mesh. The proposed Modal Representation of the Diffusion Distance Matrix achieves better results in similarity matching of flexible molecules, than the method in [28].

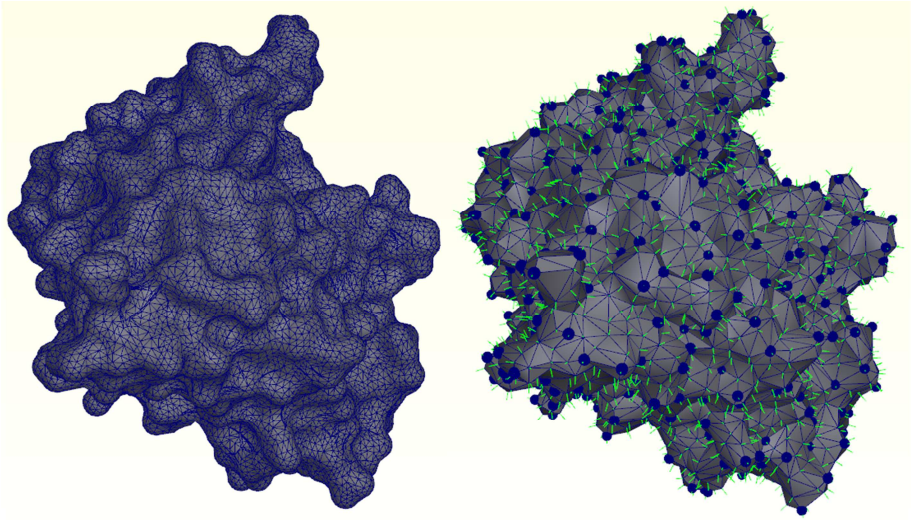
Evaluation of several state-of-the-art local shape descriptors in order to select the local feature that best fits to our framework: the selection criteria include computational efficiency in descriptor extraction, compactness of the descriptor, rotation-invariance and improved discrimination capacity in the flexible molecular shape comparison problem. Eventually, a shape descriptor, which is based on Surflet-Pair relations [37] and fulfils the above criteria, has been selected.

Finally, it is worth mentioning that the resulting shape descriptors constitute a compact representation of the molecular shape. Since it is a pure shape-based method (i.e. the descriptors do not rely on physicochemical information), it is applicable to both macromolecules (e.g. proteins) and small ligands. Thus, throughout the description of the approach, in Sections 2, 3 and 4, the term "molecule" will be used referring to both proteins and ligands. A distinction will be made in Experiments section, though, since the first two datasets refer to proteins and other macromolecules and the last two datasets refer to small ligands.

The rest of the paper is organized as follows: in Section 2, the pre-processing procedure is described, Section 3 analyses the computation of modal representation based on diffusion distances (DDMR) and Section 4 the computation of the Augmented Local Descriptor (ALD). The combined matching scheme that includes the global, the local and the hybrid feature is described in Section 5. Experiments performed in four benchmark datasets are reported in Section 6. Finally, conclusions are drawn in Section 7.

2 PREPROCESSING

The preprocessing procedure consists of two steps: the first step involves computation of the Solvent Excluded Surface (SESs) of the molecule, while, during the second step, the SES is remeshed so that each molecule is represented by the same number of oriented points. These preprocessing steps are required for descriptor extraction. Input to the system is the crystal structure of the molecule (e.g. in PDB file format), which represents its atoms in the 3-dimensional space (x, y, z coordinates). In order to generate a SES, the Maximal Speed Molecular Surface (MSMS) [51] software has been utilized, which is based on rolling a probe sphere (of size equal to the size of the solvent molecule) over the exposed contact surface of each atom.



a) b)

Fig. 2. a) the SES of a protein that consists of 25144 vertices and 50278 faces; b) the surface that is produced after the remeshing step, consisting of N_S = 3000 vertices and 5996 faces. The dark blue spheres depict the N_K =500 sub-sampled points, while the green lines depict the normals $\mathbf{n}_i$.

The output mesh is then used for the extraction of global and local shape descriptors. In order to apply the descriptor extraction algorithms, all molecules of the dataset should have the same number of mesh vertices. Since by using the MSMS software we cannot determine the exact number of the extracted vertices, a remeshing step follows to produce a mesh with the exact number of vertices N_S. For this remeshing, the Computational Geometry Algorithms Library (CGAL[1]) has been used. Let $\mathbf{p}_i$ be the i^{th} vertex, $i = 1,\ldots,N_S$. For each $\mathbf{p}_i$ its normal vector $\mathbf{n}_i$ is computed resulting in a set of N_S oriented points $(\mathbf{p}_i, \mathbf{n}_i)$. These oriented points are further sub-sampled to generate a new set of N_K keypoints $\mathbf{q}_i$, $i = 1,\ldots,N_K$, where $N_K < N_S$, that provide a coarser representation of the 3D molecule. Sub-sampling is performed using quasi-random sequence, which is a deterministic sequence that produces sample points more uniformly distributed than a pseudo-random sequence. In our case the Sobol sequence has been utilized [39]. In Fig. 2a, the SES of a protein is depicted. This mesh consists of 25144 vertices and 50278 faces. The new surface after the remeshing step consists of 3000 vertices and 5996 faces and it is shown in Fig. 2b. The normals $\mathbf{n}_i$ of the N_S oriented points are given in green lines, while the dark blue spheres depict the centers of the N_K sub-sampled points.

3 A Global Shape Descriptor based on Diffusion Distances

The computation of DD over the molecular surface is performed in three main steps: (a) calculation of the Markov probability matrix; (b) Singular Value Decomposition (SVD) of the matrix to generate the diffusion

[1] http://www.cgal.org/

map space; and (c) computation of the diffusion distances.

Let $\mathbf{p}_i$ be the set of N_S vertices. Let $K(\cdot)$ be a kernel function with bandwidth h. The Gaussian kernel $K(\mathbf{p}_i, \mathbf{p}_j) = \exp\left(-\left\|\mathbf{p}_i - \mathbf{p}_j\right\| / 2h^2\right)$ is one of the most commonly used, where the bandwidth h controls the local scale of each data point's neighborhood and $\left\|\mathbf{p}_i - \mathbf{p}_j\right\|^2$ is the Euclidean distance between surface points i and j. Then, the diffusion matrix $\mathbf{L}$ with elements $L_{ij} = K(\mathbf{p}_i, \mathbf{p}_j)$ is normalized as $\mathbf{M} = \mathbf{D}^{-1}\mathbf{L}\mathbf{D}^{-1}$ by the degree matrix $\mathbf{D}$ with $D_{ij} = \sum_i L_{ij}$. The normalized diffusion matrix $\mathbf{M}$ is a stochastic matrix with all row sums equal to one, and according to [42] it can be interpreted as a random walk on a graph, where the vertices of the graph are the surface points $i = 1, \ldots, N_S$ and the weights of the $\langle i, j \rangle$ edges correspond to M_{ij} values. Thus, M_{ij} denotes the $p(1, i \mid j)$ transition probability from the surface point j to point i in one time step ($t = 1$). For any finite time t the Markov probability matrix $\mathbf{M}^t$ with elements M_{ij}^t is computed as $M_{ij}^t = p(t, i \mid j)$, expressing the probability distribution of reaching surface point i, given a starting point j at time $t = 0$. Thus, the transition probability is given by $p(t, i \mid j) = \mathbf{e}_j \mathbf{M}^t$, where $\mathbf{e}_j$ is a row vector of zeros with a single entry equal to one at the *j*-th coordinate. Let the SVD of matrix $\mathbf{M}^t$ be $\mathbf{M}^t = \mathbf{A}\,\mathbf{\Sigma}\,\mathbf{B}^T$, where $\mathbf{\Sigma} = diag(\sigma_0, \sigma_1, \ldots, \sigma_k)$ and $\sigma_0 \geq \sigma_1 \geq \ldots \geq \sigma_k \geq 0$ are the $k+1$ singular values of $\mathbf{M}^t$, $\mathbf{A} = [\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_k]$ and $\mathbf{B} = [\mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_k]$ with $\mathbf{a}_i = \{a_i(1), a_i(2), \ldots, a_i(N_S)\}$ and $\mathbf{b}_i = \{b_i(1), b_i(2), \ldots, b_i(N_S)\}$ are the left and right singular vectors, respectively, and $\mathbf{a}_0$ and $\mathbf{b}_0$ are the first left and right eigenvectors, corresponding to the first $(\sigma_0 = 1)$ eigenvalue. Note that following [42], the first eigenvalue and the respective eigenvectors are excluded from the diffusion process and are used only for normalization purposes. The diffusion distance between surface points i, j at time t is calculated as:

$$D_t^2(i, j) = \left\|\mathbf{\Psi}_t(i) - \mathbf{\Psi}_t(j)\right\|^2 \qquad (1)$$

where $\mathbf{\Psi}_t(i) = \left(\sigma_1^t \cdot b_1(i), \sigma_2^t \cdot b_2(i), \ldots, \sigma_k^t \cdot b_k(i)\right)$ is the mapping of the *i*-th surface point from the original kernel space (formed by the kernel function $K(\cdot)$) to the diffusion map space at time t.

3.1 Modal Representation of Diffusion Distance

Given the computation of diffusion distances between the molecular surface points, the next step is to exploit this feature for the computation of a global shape descriptor. A common technique that has been already fol-

lowed in similar works [28] is to accumulate these pairwise distances into a histogram. In this paper, we propose an alternative approach based on a modal representation. The idea is to apply Singular Value Decomposition to the Diffusion Distance Matrix $\mathbf{DDM} = \left\{D_t^2(i,j)\right\}$, where $i, j = 1,\ldots,N_S$. In this way, **DDM** is separated into a matrix that contains intrinsic shape information and a matrix with information about the corresponding points. The SVD of $\mathbf{DDM}$ yields:

$$\mathbf{DDM} = \mathbf{U}\,\mathbf{L}\,\mathbf{V}^T \tag{2}$$

where the singular value matrix $\mathbf{L} = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$, contains the intrinsic information about geometry, and matrices $\mathbf{U}, \mathbf{V}$ contain the information about correspondences between points. The first n singular values $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ constitute the *Modal Representation of Diffusion Distance* (*DDMR*) descriptor $\mathbf{D}^{DDMR}$ of the 3D object. It has been proven in [32] that the eigenvalue matrix is invariant to sampling order of the surface points. Keeping only a relatively small percentage of the first singular values (Section 6.1) provides a highly compact shape descriptor with significantly discriminative power and robustness to molecular shape conformations.

4 An Augmented Local Descriptor

The proposed Augmented Local Descriptor (ALD) is computed on each of the N_K keyponts (Section 2) that provide a coarse approximation of the molecular surface. This results in a total of N_K ALD descriptor vectors $\mathbf{D}_i^{ALD}$ (i=1,…, N_K) that are extracted for each 3D molecule. Each descriptor vector $\mathbf{D}_i^{ALD}$ consists of two parts: the former is a purely local feature, the Local Descriptor based on Surflet-Pair Relations $\mathbf{D}_i^{LDSP}$, and the latter is a Hybrid Local-Global feature $\mathbf{D}_i^{HLG}$.

4.1 A Local Descriptor based on Surflet-Pair Relations (LDSP)

The first step for the extraction of local descriptors is to define a local region (patch) on the 3D surface, on which the descriptor is computed. In our case, the local descriptor is defined on a spherical region of radius R centered at each keypoint $\mathbf{q}_i$, $i = 1,\ldots,N_K$ (Fig. 3). Regarding the computation of geometric features on the local patch, the Surflet-Pair-Relation Histogram descriptor [37] has been selected comparing with other local descriptors (Shape Impact Descriptor (SID) [69] and Local Spectral Descriptor (LSD) [33]), since it achieved the highest performance in molecular similarity search, while being at the same time fast to compute, compact and rotation invariant. Given the set of oriented points $(\mathbf{p}_i, \mathbf{n}_i)$, $i = 1,\ldots,N_S$ of the 3D mole-

cule, the LDSP is computed on the subset $Q = \{(\mathbf{p}_1, \mathbf{n}_1), (\mathbf{p}_2, \mathbf{n}_2), \ldots, (\mathbf{p}_N, \mathbf{n}_N)\}$ of oriented points within a spherical region around keypoint $\mathbf{q}$ with $\|\mathbf{p}_i - \mathbf{q}\| \leq R$. For each pair of oriented points $(\mathbf{p}_1, \mathbf{n}_1), (\mathbf{p}_2, \mathbf{n}_2)$, four attributes α, β, γ and δ are computed, representing the azimuthal angle, the cosine of polar angle, the direction and the length of the translation from $\mathbf{p}_1$ to $\mathbf{p}_2$, respectively. Then, all 4-tuples $(\alpha, \beta, \gamma, \delta)$ of Q are collected into a 4-dimensional joint histogram.

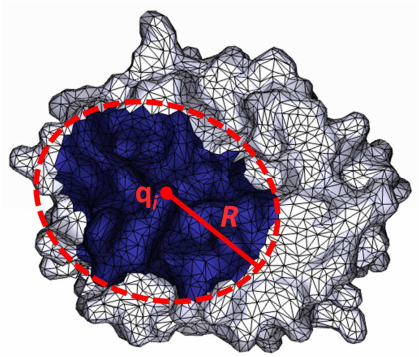


Fig. 3. The the local descriptor is defined on a spherical region (blue surface patch) of radius *R* centered at a keypoint $\mathbf{q}_i$.

A more detailed description in the computation of attributes α, β, γ and δ is available in [37]. An important parameter that needs to be analyzed, though, is the number of bins k_L for each dimension of the joint histogram. Taking into account that the LDSP descriptor $\mathbf{D}_i^{LDSP}$ for each keypoint $\mathbf{q}_i$ is a 1D vector of dimension k_L^4, the selection of parameter k_L should be such that the number of bins is adequate to produce a discriminative descriptor, while at the same time k_L is not very high so as to keep the descriptor dimensionality low. For $k_L < 5$, the discriminative power of the local feature was negatively affected, while for $k_L > 5$, the descriptor dimensionality was increasing dramatically without achieving significant improvement of accuracy, thus, $k_L = 5$ was selected, resulting in a descriptor vector $\mathbf{D}_i^{LDSP}$ of size 625. The optimal value for radius R has been estimated in a similar manner: very low values of R result in spherical regions with trivial shape information; for very high values of R, the local character of the descriptor, which gives its robustness to non-rigid problems, disappears. Eventually, an optimal choice for our experiments was $R = 0.4 \cdot R_A$, where R_A is the radius of the 3D molecule's smallest bounding sphere.

4.2 A Hybrid Local-Global feature (HLG)

Similar to LDSP, the *Hybrid Local-Global feature* (*HLG*) is computed for each keypoint $\mathbf{q}_i$, $i = 1, \ldots, N_K$. More specifically, the following set is computed for each $\mathbf{q}_i$:

$$DD_{\mathbf{q}_i} = \left\{ dd\left(\mathbf{q}_i, \mathbf{p}_1\right), dd\left(\mathbf{q}_i, \mathbf{p}_2\right), \ldots, dd\left(\mathbf{q}_i, \mathbf{p}_{N_S}\right), \right\}, \quad (3)$$

where $dd\left(\mathbf{q}_i, \mathbf{p}_j\right)$ is the diffusion distance from the keypoint $\mathbf{q}_i$ to sample point $\mathbf{p}_j$, $j = 1, \ldots, N_S$. The N_S diffusion distances of the set $DD_{\mathbf{q}_i}$ are accumulated into a 1D histogram of $k_H = 100$ bins. Again, the dimension k_h has been experimentally determined [35]. This histogram, which is normalized so that the sum of all values equals 1, constitutes the *HLG descriptor* $\mathbf{D}_i^{HLG}$ of keypoint $\mathbf{q}_i$.

According to the above definition, the HLG descriptor is neither a purely local feature nor a global descriptor. It combines local characteristics – as it is computed for each keypoint – with global characteristics – as it takes into account the set of diffusion distances of the entire molecule. HLG resembles to the Local Distance Feature (LDF) that was proposed in [35]. However, in [35], the distances to all points $\mathbf{p}_j$ are computed using a Manifold Ranking algorithm [41], according to which each keypoint $\mathbf{q}_i$ is used as the source of diffusion of ranking score for the MR. The resulting histogram is created by all ranking scores at sample points $\mathbf{p}_j$. In this paper, the distances $dd\left(\mathbf{q}_i, \mathbf{p}_j\right)$ are computed using the framework presented in Section 3. Thus, diffusion distances are computed only once for both the DDMR and the HLG descriptors.

4.3 Creating a Bag of Augmented Local Descriptors (BoALD)

During this step, the local LDSP descriptors and the hybrid HLG descriptors are integrated into a global histogram. This process is summarized in Fig. 4. Initially, for each keypoint $\mathbf{q}_i$ with LDSP descriptor $\mathbf{D}_i^{LDSP} = \left(d_i^{LDSP}(1), \ldots, d_i^{LDSP}\left(k_L^4\right)\right)$ and HLG descriptor $\mathbf{D}_i^{HLG} = \left(d_i^{HLG}(1), \ldots, d_i^{HLG}\left(k_H\right)\right)$, the ALD descriptor is given by:

$$\mathbf{D}_i^{ALD} = \left(d_i^{LDSP}(1), \ldots, d_i^{LDSP}\left(k_L^4\right), d_i^{HLG}(1), \ldots, d_i^{HLG}\left(k_H\right)\right) \quad (4)$$

$\mathbf{D}_i^{ALD}$ is a histogram of dimension $k_A = k_L^4 + k_H = 625 + 100 = 725$. To produce a global descriptor from the N_K local descriptors $\mathbf{D}_i^{ALD}$, the Bag-of-Features approach has been utilized. Let $V = \left\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{N_V}\right\}$ be a set of visual words. The dimension of each visual word is equal to k_A i.e. of the ALD histogram. The set V is created by applying *k*-means clustering to a subset (training set) of the ALD descriptors $\mathbf{D}_i^{ALD}$ of the molecular database. The descriptors that constitute the training set are selected randomly (10% of the local features of the database) in order to capture a representative view of the database. Each visual word $\mathbf{v}$ is the

center of a cluster. Then, each ALD descriptor $\mathbf{D}_i^{ALD}$ of the 3D molecule is vector quantized into a visual word and a histogram of N_V visual words is produced. This histogram $\mathbf{D}^{BoALD}$ is called *Bag-of-ALD descriptors* or *BoALD*.

The size of vocabulary N_V should be carefully chosen since it affects both retrieval accuracy and computational cost. For large datasets, which imply also a large number of samples to cluster, an increase of size N_V would require high computation times for the k-means clustering. On the other hand, retrieval accuracy is improved as vocabulary size increases, until a specific upper limit is reached, above which no further improvement is observed. Based on the aforementioned criteria, the optimal choice of vocabulary size is $N_V = 1000$, as it has been experimentally found.

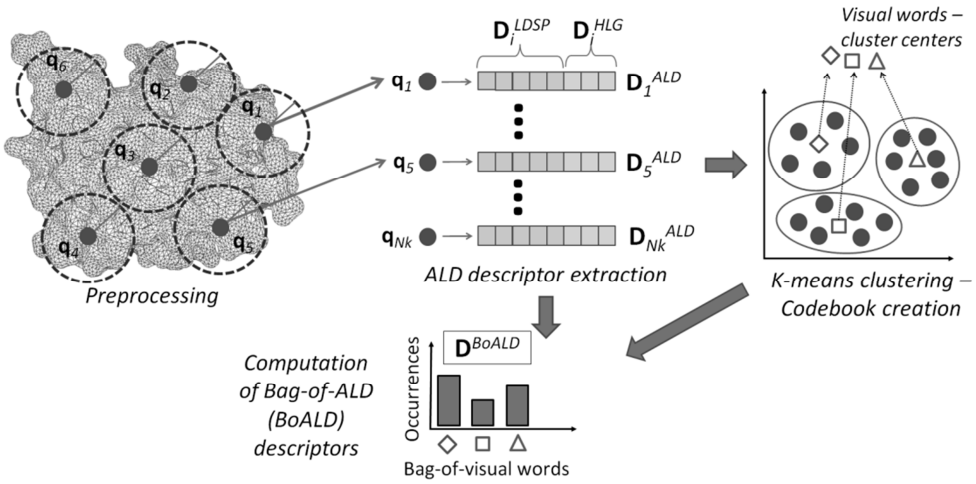


Fig. 4. The process for computing the BoALD descriptors.

5 SIMILARITY MATCHING

Let $\mathbf{D}^{DDMR}$ and $\mathbf{D}^{BoALD}$ be the DDMR and BoALD descriptor vectors that are extracted using the methods described in Sections 3 and 4, respectively. The overall shape dissimilarity between two 3D molecules A and B can be calculated as the weighted sum of the dissimilarities of each descriptor separately:

$$dis(A,B) = w^{DDMR} \cdot dis^{DDMR}(A,B) + w^{BoALD} \cdot dis^{BoALD}(A,B), \qquad (5)$$

where dis^{DDMR} and dis^{BoALD} are the dissimilarities of DDMR and BoALD descriptors, respectively, and w^{DDMR}, w^{BoALD} their corresponding weights. In general, the selection of the optimal distance metric for each descriptor is not trivial. An extensive study on the performance of the most well-known dissimilarity metrics is available in [36]. In the case of the DDMR descriptor, the *X-Distance* (or *normalized Manhattan Distance*) was

experimentally proven to be the optimal metric:

$$dis^{DDMR}(A,B) = 2 \cdot \sum_{i=1}^{N_D} \frac{\left| D_A^{DDMR}(i) - D_B^{DDMR}(i) \right|}{\left| D_A^{DDMR}(i) + D_B^{DDMR}(i) \right|}, \quad (6)$$

where $\mathbf{D}_A^{DDMR}$, $\mathbf{D}_B^{DDMR}$ are he descriptors of molecules A and B, respectively and N_D is the dimensionality of the descriptor vector. Similarly, the optimal distance metric for the BoALD descriptor is the *Kullback-Leibler Divergence*:

$$dis^{BoALD}(A,B) = \sum_{i=1}^{N_V} \left(D_A^{BoALD}(i) - D_B^{BoALD}(i) \right) \ln \frac{D_B^{BoALD}(i)}{D_A^{BoALD}(i)}, \quad (7)$$

where $\mathbf{D}_A^{BoALD}$, $\mathbf{D}_B^{BoALD}$ are the descriptors of molecules A and B, respectively and N_V is the dimensionality of the descriptor vector.

After selecting the optimal dissimilarity metrics, the weights w^{DDMR}, w^{BoALD} need to be determined. In our case, we followed the Particle Swarm Optimization (PSO) strategy [36] for the weight optimization. PSO is an algorithm for global optimization. It is motivated by the social behavior of organisms such as bird flocking and fish schooling. PSO optimizes a problem in which a best solution can be represented as a point or surface in an n-dimensional space. It iteratively tries to improve a candidate solution based on a given quality measure (fitness function). PSO establishes a population (swarm) of candidate solutions, known as particles that move around in the search space, and are guided by the best found positions, updated while better positions are found by the particles.

The population of candidate solutions, in our case, is the weights w^{DDMR}, w^{BoALD}, which can take arbitrary values between $[0,1]$. The fitness function to be optimized is the *average Tier-1 precision*, which is calculated on a train dataset. More specifically, each 3D molecule of the dataset is used as query to retrieve similar objects, using (5) as dissimilarity metric. The retrieved results are ranked in ascending order. The Tier-1 precision is given by the following equation:

$$P_{T1} = \frac{R^C(K)}{K}, K = |C| - 1 \quad (8)$$

where K is the number of first retrieved objects, $R^C(K)$ is the number of retrieved objects within the *K*-first, which are of the same class C with the query, and $|C|$ is the number of objects that belong to class C. PSO resulted in the following weights: $w^{DDMR} = 0.62$, $w^{BoALD} = 0.38$.

6 EXPERIMENTAL RESULTS

For the experimental evaluation of the proposed method, four different datasets have been selected. The first dataset is part of the Database of Macromolecular Movements (MolMovDB) [38], which comprises molecules with large conformational changes (http://www.molmovdb.org/), also including the intermediate morphs [70]. It consists of 2695 PDB files classified into 214 categories [55]. Each category consists of a collection of morphs representing different states of the same molecule. This dataset is used for parameter selection and for comparison with existing flexible molecular shape matching approaches [27][28]. The second dataset consists of 2631 3D protein structures. It is a subset of the FSSP database [49] and was created by us to demonstrate the performance of the Spherical Trace Transform (STT) in [16]. The 2631 proteins are classified into 27 classes according to the FSSP/DALI algorithm [50]. Each class consists of different protein structures, which have at least 25% similarity in their amino-acid sequence (according to the FSSP/DALI classification). The high classification accuracy achieved by STT in this dataset reveals that the proteins that belong to the same class, apart from their 25% sequence similarity, demonstrate also rigid shape similarity. The second dataset has been used to evaluate the performance of the proposed method in rigid shape matching of 3D protein structures and it is publicly available at vcl.iti.gr/protein_retrieval/PDB_FSSP.zip. It is worth mentioning that the first two datasets are different in nature and cannot be compared, since they measure different aspects of the molecular shape comparison problem (flexible vs rigid shape similarity), their classes have been created based on different criteria and none of the datasets is subset of the other. Finally, the third and fourth dataset are used to demonstrate the performance of our framework in large-scale virtual screening of ligands. Experiments have been performed on a PC with i5 2.8GHz processor, 4GB RAM.

6.1 Parameter Selection for the DDMR Descriptor

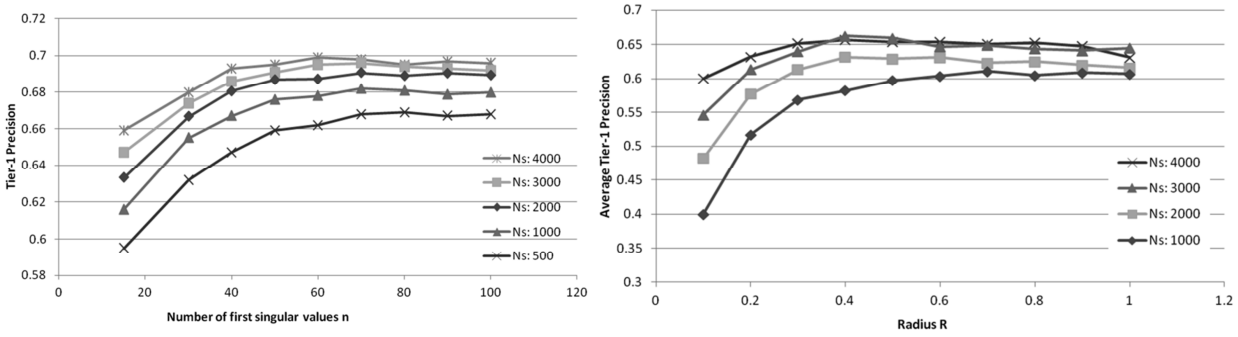
For the implementation of the DDMR descriptor (Section 3.1), the Matlab Toolbox for Dimensionality Reduction[2] (v0.8.1) has been selected, using the default parameters $h = 1$ and $t = 1$. The discriminative power of DDMR mainly depends on two parameters: a) the number N_S of sample points $\mathbf{p}_i$ on the molecular surface, and b) the dimensionality of DDMR descriptor vector, i.e. the number n of first singular values $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ of SVD (2). By increasing the number of sample points N_S, a higher-quality representation of the surface is achieved and accuracy is improved, however, this results in higher descriptor extraction times. Additionally, an increase of n may also improve the accuracy. We run several sets of experiments us-

[2] http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html

ing different values of N_S and n. As a performance metric, the average *Tier-1 Precision* has been selected (8).

In Fig. 5 a), the average Tier-1 Precision for different values of N_S and n, in MolMovDB is presented. It is obvious that as the number of sample points N_S increases a higher precision is achieved. Using a mesh resolution higher than 2000 points, though, the improvement in accuracy is negligible. Similar conclusions are drawn regarding the number n of first singular values. For values n higher than 50-60, there is no significant improvement in precision.

A critical factor for the parameter selection is the descriptor extraction time. Since the process of extracting the DDMR descriptor involves computations on $N_S \times N_S$ matrices, the processing time may increase prohibitively as the number of sample points N_S increases. This is highlighted in Table I, where it is obvious that for meshes consisting of 4000 points it takes approximately one minute for descriptor extraction, while for meshes of 1000 points the extraction time is less than 2 seconds. For the experiments that will be presented in the following subsections the values $N_S = 2000$ and $n = 50$ have been selected for DDMR.



a) b)

Fig. 5. a) Parameter selection for DDMR descriptor: the average Tier-1 Precision in MolMovDB for different values of n and N_S; **b)** Parameter selection for BoLDSP descriptor: the average Tier-1 Precision in MolMovDB for different values of radius R and N_S.

Table I: Average extraction times of the DDMR descriptor for different numbers of sample points.

Number of sample points N_S	DDMR Descriptor Extraction Time (s)
500	0.47
1000	1.34
2000	4.69
3000	14.08
4000	46.52

6.2 Parameter Selection for the BoALD Descriptor

The BoALD descriptor has been implemented by us in C++ based on the works presented in [35] and [37]. The performance of BoALD is affected by several parameters: a) the radius R of the local descriptor LDSP; b)

the number of sample surface points N_S ; c) the number of local points N_K , and d) the vocabulary size N_V of the codebook. The number of surface points N_S is related to radius R as follows: a small R provides sufficient locality to the descriptor but it requires a high N_S so that the local histograms are well populated. The number of local points N_K affects the selection of the vocabulary size N_V : for a given N_K , an increase of N_V improves the accuracy until a specific upper limit is reached. Beyond that limit a further increase of N_V has no effect in accuracy. If we increase N_K , then we can achieve a higher upper limit for N_V resulting in a more discriminative descriptor. It is worth mentioning that since *k*-means clustering (used in bag-of-features) involves random selection of cluster centers, the mean values of Tier-1 accuracy are reported, where each experiment was repeated ten times. However, in many cases, the differences between the Tier-1 accuracy values are minimal. To verify this, for all experiments, we applied statistical pairwise t-tests, where the calculated differences of means were insignificant at level 0.05. In Fig. 5 b), the average Tier-1 Precision of the BoLDSP (bag-of-features to LDSP) descriptor in MolMovDB for different values of radius R and number of sample points N_S is depicted. Starting from $R = 0$, precision increases as R increases, until a maximum is reached. As an example, for meshes with 3000 points, the maximum precision is achieved for $0.4 \cdot R_A$.

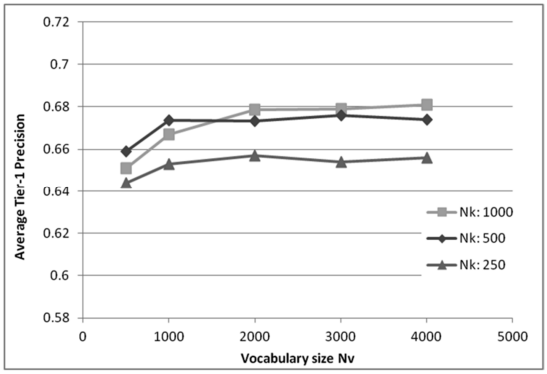


Fig. 6. Parameter selection for BoALD descriptor: the average Tier-1 Precision in MolMovDB for different values of vocabulary size N_V and N_K .

In Fig. 6, the average Tier-1 Precision of the BoALD descriptor in MolMovDB for different values of vocabulary size N_V and number of local points N_K is depicted. For $N_K = 250$, an increase of N_V does not affect the average precision. Similarly, for $N_K = 500$ the precision is not improved for $N_V \geq 1000$. Finally, for $N_K = 1000$, the improvement in accuracy comparing to $N_K = 500$ is negligible. It is also worth mentioning that the dimensionality N_V of BoALD should be kept relatively low to achieve faster matching times.

For the experiments that will be presented in the sequel the values $N_S = 3000$, $R = 0.4 \cdot R_A$, $N_K = 500$ and $N_V = 1000$ have been selected for BoALD.

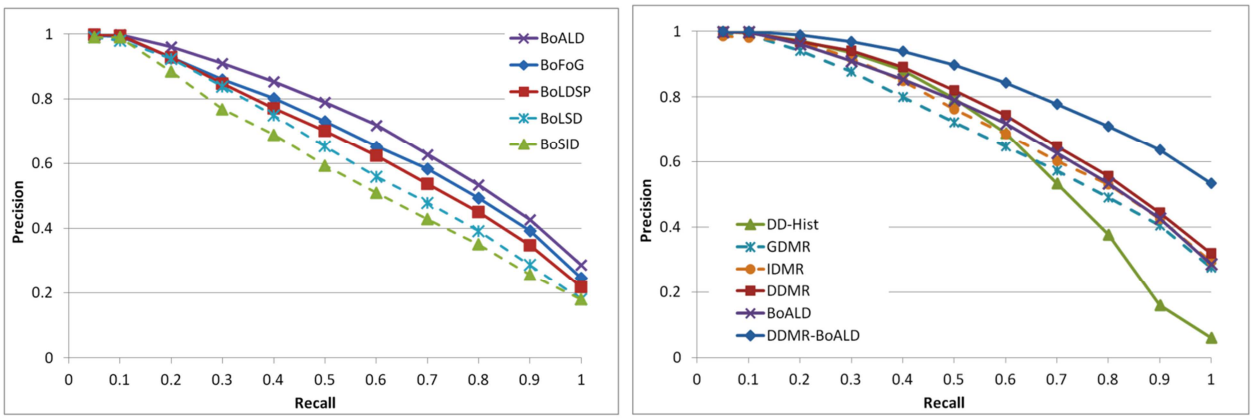
Table II: Average extraction times of the BoALD descriptor.

Number of sample points N_S	LDSP Descriptor Extraction Time (s)	HLG Descriptor Extraction Time (s)	BoALD bag-of-feature integration time (s) (N_K=500)
1000	0.28	0.34	2.34
2000	0.42	1.66	
3000	0.75	3.08	
4000	0.95	8.74	

The processing times for extraction of local features LDSP and HLG and for the BoALD bag-of-feature integration are given in Table II. The codebook learning via *k*-means clustering is a computationally expensive process. For the MolMovDB dataset with 2695 molecules and $N_K = 500$ local features per molecule, the total number of training samples (10% of the dataset) is 134750 local features. The *k*-means clustering of 134750 features with vocabulary size $N_V = 1000$ took about 1700s (28 minutes). Then, the bag-of-features integration time for each molecule is 2.34s, thus, 6300s (105 minutes) for the entire database. These computations need to be performed only once, during the pre-processing stage.

6.3 Performance Evaluation in MolMovDB – Flexible Similarity Matching

For performance evaluation in MolMovDB the *precision-recall* curve has been used, where precision is the proportion of the retrieved molecules that are relevant to the query and recall is the proportion of relevant molecules in the entire database that are retrieved. In a benchmark dataset that is classified, such as MolMovDB, relevant items are those belonging to the same category with the query. In Fig. 7 a), a comparison of different local surface descriptors in MolMovDB is presented. All local descriptors are extracted on the same set of keypoints $\mathbf{q}_i$, following a bag-of-features computation to produce a global descriptor vector. For the local descriptors reported in section 4.1, namely the Shape Impact Descriptor (SID), the Local Spectral Descriptor (LSD) and the Local Descriptor based on Surflet-Pair Relations (LDSP), the Bag-of-SID (BoSID), Bag-of-LSD (BoLSD) and Bag-of-LDSP (BoLDSP) are created, respectively. BoLDSP achieves better retrieval accuracy than the other two candidates, which justifies its selection as a local feature. Moreover, the contribution of spatial context as a complementary feature to the purely local descriptors is also demonstrated in Fig. 7 a). Combining LDSP and the hybrid HLG into the proposed BoALD descriptor achieves significantly higher performance than the purely local descriptors. It is worth mentioning that BoALD is more discriminative than the BoFoG descriptor presented in [35], which also combines a local with a hybrid descriptor.



a) b)

Fig. 7. a) Comparison of BoSID, BoLSD, BoLDSP , BoFoG and BoALD in MolMovDB; **b)** Comparison of DD-Hist, GDMR, IDMR, DDMR, BoALD and DDMR-BoALD in MolMovDB.

Another innovative feature of the proposed work is the modal representation of the diffusion distance matrix, which results in the DDMR descriptor. In Fig. 7 b), DDMR is compared against the method of [28], which accumulates the pairwise diffusion distances into a histogram (DD-Hist). The proposed DDMR descriptor outperforms DD-Hist especially for higher values of recall. The superiority of Diffusion Distance over Geodesic Distance and Inner Distance, in capturing molecular flexibility, is also demonstrated in Fig. 7 b). The modal representations of GD (GDMR) and ID (IDMR) are derived by substituting the Diffusion Distance Matrix in equation (2) with the Geodesic Distance Matrix and Inner Distance Matrix, respectively. Again, the proposed DDMR descriptor achieves higher retrieval accuracy. Finally, the combination of DDMR with BoALD, using the weighted sum of dissimilarities (5), is presented in Fig. 7 b). DDMR-BoALD clearly outperforms the rest of descriptors, which confirms our assumption that the combination of a global feature (DDMR) with a local feature (BoALD) achieves higher retrieval accuracy than each descriptor separately. Fig. 13 shows three morph deformations for each of the following macromolecules: a) Dehydroquinase, b) NHP6A and c) trp repressor. The molecule in the first column is given as query and the respective ones in second and third columns are retrieved within the first ranking positions. Despite the changes in their global shape due to molecular flexibility, the morphs still demonstrate high similarity to the query.

6.4 Evaluation of Rigid Similarity Matching

In Fig. 8, the precision-recall curves for the second dataset (subset of FSSP) are depicted. Our DDMR-BoALD descriptor is compared with STT [16], which is a rigid shape matching method. It is obvious that DDMR-BoALD outperforms STT in a rigid-shape dataset as well. This is mainly due to the fact that the combination of intrinsically different features (a global, a local and a hybrid local-global) increases the robustness of the resulting descriptor.

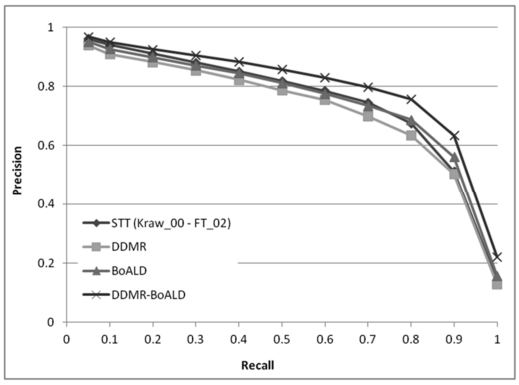


Fig. 8. Comparison of the proposed method with STT in the subset of FSSP database that was used in [16].

6.5 Comparison with structural alignment methods

In the experiments presented in the previous subsections, the proposed method is compared with descriptor-based approaches. A comparison of DDMR-BoALD descriptor with two structural alignment methods, FATCAT [65] and TM-Align [64], is presented in Fig. 9, where their superiority over our descriptor in terms of performance is obvious. In general, structural alignment methods achieve better performance than descriptor-based methods. However, for a thorough comparison between these intrinsically different approaches, additional parameters need to be taken into account. First of all, it is worth mentioning that DDMR-BoALD relies on geometric information only, which makes it appropriate for use in a wide range of molecules, both large macromolecules and small ligands. This is not possible in the case of structural alignment methods, which are looking for correspondences between atoms/residues. As an example, FATCAT and TM-Align cannot be applied to the ligands of section 6.7. Another important parameter is the efficiency of the method. In Table III, the times for comparing a pair of molecules using TM-Align, FATCAT and the proposed DDMR-BoALD descriptor are reported. DDMR-BoALD is 12500 times faster than TM-Align and 70000 times faster than FATCAT. Based on the above, descriptor-based and structural alignment methods should not be competitive but they should work collaboratively, i.e. a descriptor-based method can be used for fast filtering, at a first stage, and a structural alignment method can be used to refine a smaller subset of the results.

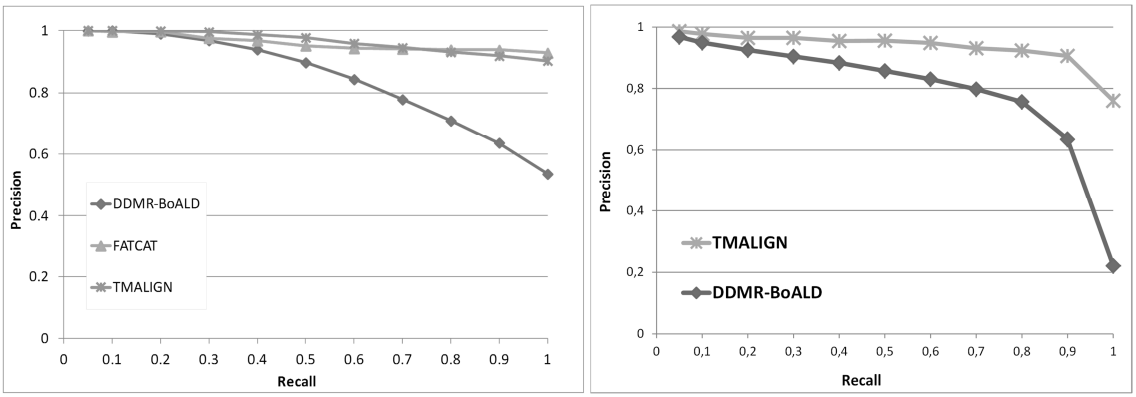
Table III: Average CPU times for comparing a pair of molecules using TM-Align, FATCAT and the proposed DDMR-BoALD descriptor.

Method	TM-Align	FATCAT	DDMR-BoALD
Average CPU Time for Pairwise Comparison	0.25s	1.4s	0.02ms

In Table IV and

Table V, the performance of combining the proposed DDMR-BoALD with TM-Align method is demonstrated in MolMovDB and the subset of FSSP datasets, respectively. More specifically, each item of the dataset is used as query and DDMR-BoALD is applied to match the query with all items of the dataset (fast filtering stage). Then, TM-Align is applied only to the first ranked results for re-ranking. Different percentages of the

first ranked results are shown (from 20% of first ranked to 80%). Performance is measured in Nearest Neigbour, Tier-1 precision and Tier-2 precision. These evaluation measures share the similar idea, that is, to check the ratio of models in the query's class that also appear within the top K matches, where K= 1 for Nearest Neighbor, $K = |C| - 1$ for Tier-1, and $K = 2*(|C| - 1)$ for Tier-2 and $|C|$ is the number of class members. The reported scores are averaged by all the objects in database. It should be stressed that it was not possible to compute precision-recall diagrams, since, for some items (queries) of the dataset, recall of all relevant (to the query) items (100%recall) may require retrieval of more than 80% of the first ranked results.



a) b)

Fig. 9. Comparison of the proposed method with a) FATCAT and TM-Align in MolMovDB and b) TM-Align in the subset of FSSP database

From Table IV and

Table V, it is clear that when we apply the TM-Align as a re-ranking step to different percentages of the initially ranked results using the DDMR-BoALD fast filtering, we preserve the accuracy high in terms of NN, Tier1 and Tier2, while we significantly speed up the matching time. In this set of experiments, the percentages of the initially ranked results vary from 20 to 100%, where in the case of 100% TM-Align is applied to all the ranked results without the DDMR-BoALD fast filtering. By increasing the percentage of the first ranked results to apply TM-Align from 20% to 100%, NN is almost not affected, the improvement in Tier-1 and Tier-2 is minor (less than 2%), while the matching time can be up to 5 times faster.

Table IV: Performance (average Nearest Neighbour, Tier-1 precision and Tier-2 precision) and matching times of the proposed DDMR-BoALD method, the TM-Align method and their combination, in MolMovDB dataset. Percentages show the amount of first items ranked by DDMR-BoALD that are kept for re-ranking with TM-Align. In the last column, only the TM-Align method has been used (no filtering step with the proposed method has been applied).

	DDMR-BoALD	TM-Align (20%)	TM-Align (40%)	TM-Align (60%)	TM-Align (80%)	TM-Align (100%)
NN	0.991	0.996	0.996	0.997	0.997	0.997
Tier-1	0.773	0.927	0.930	0.933	0.934	0.934
Tier-2	0.419	0.476	0.479	0.481	0.482	0.482
All-to-all Matching Time (s)	155	388523	776891	1165260	1553628	1941842

Table V: Performance (average Nearest Neighbour, Tier-1 precision and Tier-2 precision) and matching times of the proposed DDMR-BoALD method, the TM-Align method and their combination, in the subset of FSSP dataset. Percentages show the amount of first items ranked by DDMR-BoALD that are kept for re-ranking with TM-Align. In the last column, only the TM-Align method has been used (no filtering step with the proposed method has been applied). For 20%, average Tier-2 precision cannot be computed, since there are classes in the dataset, where the number of items needed for Tier-2 is greater than the 20% of the dataset.

	DDMR-BoALD	TM-Align (20%)	TM-Align (40%)	TM-Align (60%)	TM-Align (80%)	TM-Align (100%)
NN	0.958	0.958	0.959	0.959	0.959	0.959
Tier-1	0.759	0.898	0.906	0.915	0.921	0.922
Tier-2	0.432	-	0.466	0.476	0.481	0.481
All-to-all Matching Time (s)	138	346246	692354	1038462	1384570	1730540

6.6 Virtual Screening of Ligands

The proposed method has been also evaluated in large-scale virtual screening of ligand molecules, where the investigation of an accurate algorithm for rapid shape matching is a major scientific challenge. Two benchmark datasets have been used in our tests. The first is called the "*Directory of Useful Decoys*" (DUD) [43]. DUD is derived from the ZINC database of commercially available compounds for virtual screening [44]. A subset of DUD[3] was downloaded, which consists of 13 targets and has been already used in recent studies [24]. The dataset is presented in Table VI. More specifically, each of the 13 *targets* is used as query to retrieve similar molecules from its corresponding set of *actives*+*decoys* (e.g. *ace* is used as query in the set of 46 actives and 1796 decoys and so on). The more *actives* are included among the first retrieved results the better the accuracy of the search algorithm is. The data in Table VI are adapted from the work in [24], however, we provide it here as well, in order to have a better visualization of the dataset.

Table VI: The subset of DUD dataset [24] that was used in our experiments

Target	PDB	Actives	Decoys	Decoys per Active
angiotensin-converting enzyme (ace)	1o86	46	1796	39.04
acetylcholinesterase (ache)	1eve	100	3859	38.59
cyclin-dependent kinase 2(cdk2)	1ckp	47	2070	44.04
cyclooxygenase-2(cox2)	1cx2	212	12606	59.46
epidermal growth factor receptor(egfr)	1m17	365	15560	42.63
factor Xa(fxa)	1f0r	64	2092	32.69
HIV reverse transcriptase(hivrt)	1rt1	34	1494	43.94
enoyl ACP reductase(inha)	1p44	57	2707	47.49
P38 mitogen activated protein(p38)	1kv2	137	6779	49.48
phosphodiesterase(pde5)	1xp0	26	1698	65.31
platelet derived growth factor receptor kinase(pdgfrb)	1t46	124	5603	45.19
tyrosine kinase SRC(src)	2src	98	5679	57.95
vascular endothelial growth factor receptor(vegfr2)	1fgi	48	2712	56.5

The second benchmark is the anti-HIV dataset derived from the National Cancer Institute[4] (NCI) and is employed to simulate a typical virtual screening experiment. It consists of 42687 compounds [45], which are split into 423 confirmed actives, 1081 moderately actives and 41185 confirmed inactives. The structures are

[3] http://dud.docking.org/
[4] http://dtp.nci.nih.gov/docs/aids/aids_data.html

available for download[5] in SDF format. The objective of the virtual screening experiment in this dataset is to use the 1081 moderately actives as queries and search into the database of actives and inactives. The more confirmed actives are retrieved among the first ranked results the higher the accuracy of the algorithm is.

Three different metrics have been used to evaluate the performance of the proposed method in these datasets. The first is the *Enrichment Factor (EF)* [46], which describes the ratio of actives retrieved relative to the percentage of the database scanned:

$$EF^x = \frac{N_a / N_x}{T_A / T_D} \tag{9}$$

where T_A is the total number of actives in the database of size T_D and N_a is the number of actives in the top x percent N_x of the database. Another metric is the *Boltzmann Enhanced Discrimination of Receiver Operating Characteristic (BEDROC)* [47], calculated as:

$$BEDROC = \frac{\sum_{i=1}^{n} e^{-a\frac{r_i}{N}}}{\frac{n}{N}\left(\frac{1-e^{-a}}{e^{a/N}-1}\right)} \times \frac{R_a \sinh(a/2)}{\cosh(a/2)-\cosh(a/2-aR_a)} + \frac{1}{1-e^{a(1-R_a)}} \tag{10}$$

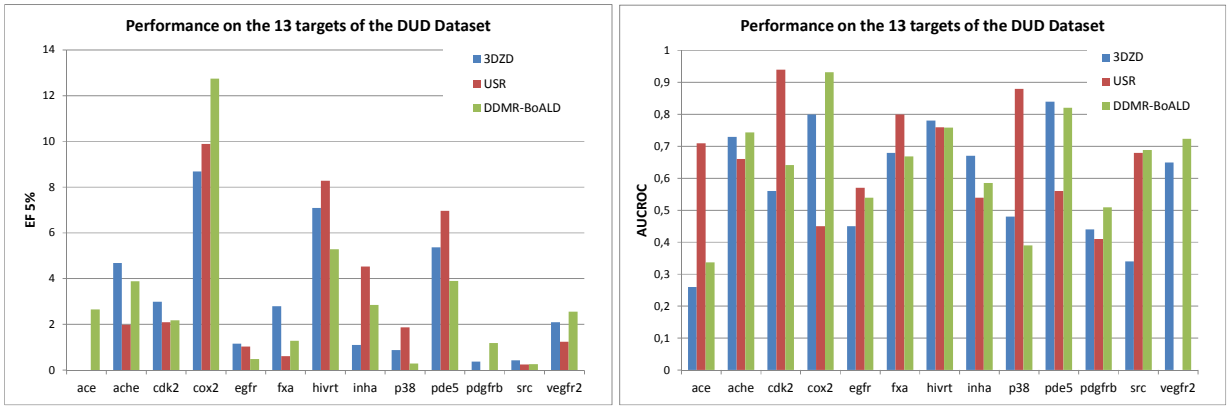
where n is the number of actives among N compounds, $R_a = n/N$, r_i is the rank of the i^{th} active and a is a weighting parameter. In our experiments, $a = 32.2$ is selected, which corresponds to $x = 5\%$ of the relative rank. Similarly, $x = 5\%$ is also selected for the EF metric (9).

Finally, the *Area Under Curve for Receiver Operator Characteristic (ROCAUC)* [24] is computed by:

$$AUCROC = 1 - \frac{1}{N_a}\sum_{i}^{N_a} \frac{N_{decoys}^i}{N_d} \tag{11}$$

where N_a and N_d is the number of actives and decoys, respectively, and N_{decoys}^i is the number of decoys ranked above the i^{th} active. The proposed DDMR-BoALD descriptor is compared with two approaches for fast virtual screening, which are also based on shape similarity matching. The first one is the 3D Zernike Descriptor (3DZD) [24], which is based on a series expansion of a given 3D function. The second one is the Ultrafast Shape Recognition (USR) scheme [11], which represents the molecular shape as a set of statistical moments generated from all-atom distance distributions that are calculated with respect to preselected reference locations. Both aforementioned methods are rotation-invariant, i.e. are able to capture the shape information independent of orientation.

[5] http://ligand.info



a) b)

Fig. 10. Performance of the 3DZD, USR and the proposed method on the 13 targets of the DUD dataset, using a) the Enrichment Factor metric; b) the AUCROC metric.

In Fig. 10 a), Fig. 10 b) and Fig. 11, the performance of 3DZD, USR and DDMR-BoALD on the 13 targets of the DUD dataset is given for the metrics EF ($x = 5\%$), AUCROC and BEDROC ($a = 32.2$), respectively. For 3DZD, the descriptor of order-12 using Correlation Coefficient as distance metric is reported, while for USR, the descriptor of order-16 using Correlation Coefficient as distance metric is reported [24].

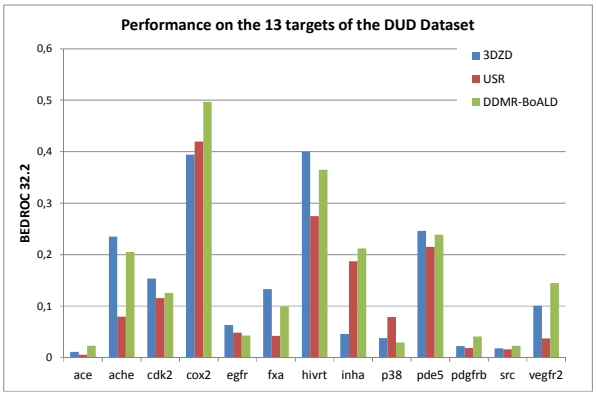


Fig. 11. Performance of the 3DZD, USR and the proposed method on the 13 targets of the DUD dataset, using the BEDROC metric.

Regarding the EF metric, the proposed method outperforms the other two in 4 out of 13 targets of the DUD Dataset, while 3DZD and USR are better in 5 and 4 targets, respectively. For the AUCROC metric, DDMR-BoALD is better in 5 targets, 3DZD in 3 and USR in 5. Finally, regarding the BEDROC metric, the proposed method outperforms others in 6 targets, 3DZD in 6 and USR in 1 target. The average scores are given in Table VII. The results derived using the 3 different metrics are not fully consistent, since e.g. USR is better than 3DZD in EF and AUCROC but it is worse in BEDROC. Overall, the proposed method is slightly better than the other two approaches in all metrics.

Table VII: EF, AUCROC and BEDROC (average) in DUD dataset for 3DZD, USR and DDMR-BoALD.

Descriptors	Metric	Order	EF 5%	AUCROC	BEDROC 32.2
3DZD	Correlation coefficient	12	2.90	0.59	0.14
USR	Correlation coefficient	16	2.99	0.62	0.12
DDMR-BoALD	-	-	**3.05**	**0.64**	**0.16**

The performance of 3DZD, USR and DDMR-BoALD is also compared in the anti-HIV dataset. In Table VIII, the average values of EF ($x = 5\%$), AUCROC and BEDROC ($a = 32.2$), for the three methods, are presented. Several results are available for both 3DZD and USR depending on the order of expansion of descriptor and the distance metric used. Again, the proposed method outperforms others in all three metrics.

A critical parameter that should be taken into account in virtual screening, especially in large databases, is the similarity matching time. In the anti-HIV dataset, which consists of more than 40000 molecules, the search times for USR are approximately 0.74-0.76s, while for 3DZD are 2.62-2.70s. These methods are significantly faster than non-shape-based approaches, which may take several hours for the same virtual screening task. The reason is that the shape-based descriptor vectors constitute a very compact representation of the molecular structure, thus, similarity matching using a common distance metric is rapid. The proposed DDMR-BoALD descriptor takes about 2.83s for a one-to-all matching in the anti-HIV dataset, thus, it is comparable to 3DZD. Consequently, since DDMR-BoALD outperforms 3DZD and USR in terms of retrieval accuracy, it can provide a better solution for rapid geometric virtual screening.

Table VIII: Average values of EF, AUCROC and BEDROC in the anti-HIV dataset for 3DZD, USR and the proposed method

Descriptors	Metric	Order	EF 5%	AUC ROC	BEDROC 32,2
3DZD	Correlation coefficient	4	1.298	0.421	0.0485
		6	1.334	0.423	0.0500
		8	1.297	0.430	0.0490
		10	1.208	0.435	0.0461
		12	1.297	0.430	0.0490
		14	1.146	0.444	0.0440
	Euclidean (DE)	4	1.307	0.411	0.0471
		6	1.292	0.416	0.0473
		8	1.301	0.427	0.0477
		10	1.255	0.435	0.0464
		12	1.301	0.427	0.0477
		14	1.263	0.455	0.0470
	Manhattan (DM)	4	1.281	0.412	0.0466
		6	1.267	0.418	0.0463
		8	1.250	0.431	0.0462
		10	1.201	0.442	0.0448
		12	1.251	0.431	0.0462
		14	1.222	0.463	0.0461
USR	Correlation coefficient	12	1.248	0.417	0.0461
		16	1.357	0.422	0.0480
	Euclidean (DE)	12	1.301	0.392	0.0486
		16	1.296	0.386	0.0485
	Manhattan (DM)	12	1.403	0.395	0.0515
		16	1.335	0.386	0.0497
DDMR-BoALD	-	-	**1.923**	**0.479**	**0.0521**

We have implemented an online tool for shape similarity search using the proposed method. Search is performed in the following datasets: a) the subset of the FSSP database; b) the subset of MolMovDB and c) the

anti-HIV dataset. A snapshot of the online tool[6] is given in Fig. 12. In this example, a search task is performed into MolMovDB using as query the molecule "*ff3*" that belongs to class "*062670-27261*". The first 14 retrieved results are presented. It is worth mentioning that the first 12 retrieved results belong to the same class with the query. These are actually representing the same molecule but with conformational changes. Despite the flexibility that is observed in the lower left part of the molecules, the algorithm is robust in capturing their global shape similarity. The online tool allows visualization of the 3D molecular structures using the Jmol[7] open-source Java viewer for chemical structures in 3D. By clicking on the thumbnail image of a retrieved molecule, a pop-up window of Jmol viewer appears.

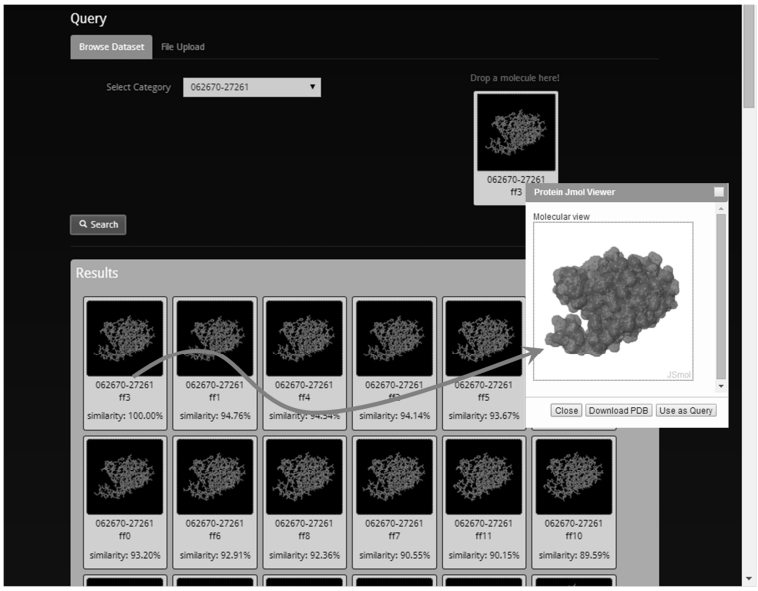


Fig. 12. Example of similarity search in MolMovDB using the proposed method (http://vcl.iti.gr/protein_retrieval/).

7 Conclusions and Discussion

We have presented a framework for similarity search of flexible molecules, which exploits both local and global geometric features. The global feature is based on pairwise computations of diffusion distances over the points of the surface and a singular value decomposition of the resulting diffusion distance matrix. The local feature is computed on each keypoint of the surface by accumulating pairwise relations among oriented surface points into a local histogram. Finally, the hybrid local-global feature is computed for each keypoint, taking into account the diffusion distances from the keypoint to all surface points, thus, enhancing the local keypoint with spatial context. The local and the hybrid features are concatenated into a joint histogram per keypoint and the multiple histograms are integrated into a global descriptor using the bag-of-features ap-

[6] http://vcl.iti.gr/protein_retrieval/
[7] http://www.jmol.org/

proach. The global and local features are combined to produce a geometric descriptor that achieves higher retrieval accuracy than each feature does separately.

The proposed method achieves high retrieval accuracy in similarity search of flexible molecules. In the MolMovDB dataset, which consists of molecules with large conformational changes, the proposed framework clearly outperforms other existing approaches in terms of precision-recall. At the same time, DDMR-BoALD descriptor achieves high retrieval performance in datasets of rigid molecules. Additionally, DDMR-BoALD provides a compact representation of the 3D molecular structure; therefore, it is appropriate for large-scale search tasks such as the virtual screening in large ligand databases. DDMR-BoALD is appropriate for retrieving small ligands as well, since it is comparable to slightly better than existing state-of-the-art approaches in two benchmarks for virtual screening.

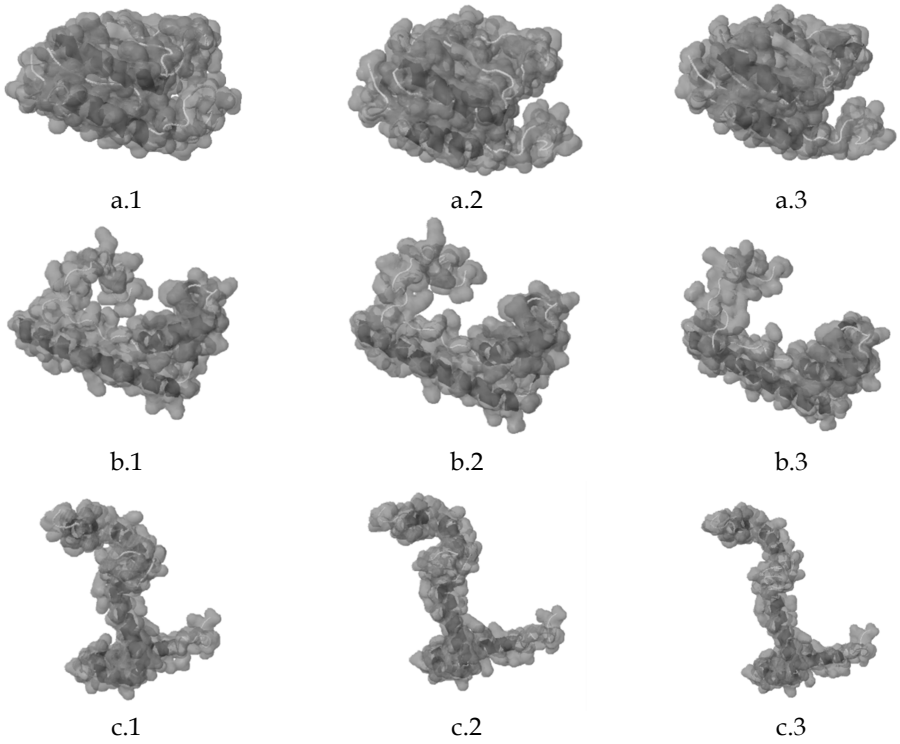


Fig. 13. Morph deformations for the following macromolecules a) Dehydroquinase, b) NHP6A and c) trp repressor. The molecule in the 1st column is given as query and the respective ones in 2nd and 3rd columns are retrieved within the first ranking positions, demonstrating high similarity to the query.

Nevertheless, the retrieval accuracy especially in virtual screening can be further improved, by enhancing the geometric features with non-geometric ones, such as physicochemical properties. At the moment, the latter are exploited by approaches that are extremely time-consuming, which, in combination with the rapid increase in size of the molecular databases, leads to prohibitively large search times. The effective integration of non-geometric information into a compact representation along with the shape-based features still remains a challenge for future research. Another important issue is that the number of mesh vertices sampled on the

molecular surface remains the same irrespective of the size of the molecule. The motivation behind this is the fact that a variable number of samples (proportional to the size of molecule) would result in descriptor vectors that are not comparable to each other. On the other hand, a fixed number of sample vertices produces scale invariant descriptor vectors, that is molecules of similar shape but with different size are regarded as similar. While the latter could be regarded as an advantage in the case of generic object retrieval, in molecular similarity comparison it introduces a limitation to the proposed descriptor. Thus, a challenge for future research is to investigate methods that are able to embed size information to the resulting descriptors.

REFERENCES

[1] A. Bender and R. C. Glen. "Molecular similarity: a key technique in molecular informatics." *Organic & biomolecular chemistry 2*, no. 22 (2004): 3204-3218.

[2] V. Venkatraman, L. Sael and D. Kihara. "Potential for protein surface shape analysis using spherical harmonics and 3D Zernike descriptors." Cell biochemistry and biophysics 54.1-3 (2009): 23-32.

[3] D. Kihara and J. Skolnick, "The PDB is a covering set of small protein structures", *Journal of Molecular Biology*, 334, 793-802, 2003.

[4] L. Holm and C. Sander. "Protein structure comparison by alignment of distance matrices." *Journal of molecular biology* 233.1 (1993): 123-138.

[5] K. Mizuguchi and N. Gö, "Comparison of spatial arrangements of secondary structural elements in proteins", *Protein Engineering*, 8.4 (1995): 353-362.

[6] M. R. Betancourt and J. Skolnick, "Local propensities and statistical potentials of backbone dihedral angles in proteins", *Journal of molecular biology* 342.2 (2004): 635-649.

[7] K. Kinoshita and H. Nakamura, "Identification of protein biochemical functions by similarity search using the molecular surface database eF-site", *Protein Science* 12.8 (2003): 1589-1595.

[8] M. E. Bock, G. M. Cortelazzo, C. Ferrari and C. Guerra, "Identifying similar surface patches on proteins using a spin-image surface representation", *In Combinatorial Pattern Matching*. Springer Berlin Heidelberg, 2005. p. 417-428.

[9] M. Ankerst, G. Kastenmüller, H. P. Kriegel and T. Seidl, "3D shape histograms for similarity search and classification in spatial databases." *Advances in Spatial Databases*. Springer Berlin Heidelberg, 1999.

[10] J. S. Yeh, D.Y. Chen, B.Y. Chen, M. Ouhyoung, "A web-based three-dimensional protein retrieval system by matching visual similarity", *Bioinformatics 2005*, 21(13):3056-3057.

[11] P. J. Ballester and W. G. Richards, "Ultrafast shape recognition to search compound databases for similar molecular shapes", *Journal of Computational Chemistry* 28.10 (2007): 1711-1723.

[12] P. J. Ballester and W. G. Richards, "Ultrafast shape recognition for similarity search in molecular databases", *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science* 463.2081 (2007): 1307-1321.

[13] M-K. Hu, "Visual pattern recognition by moment invariants", *IRE Transactions on Information Theory*, 8.2 (1962): 179-187.

[14] M. R. Teague, "Image analysis via the general theory of moments", *J. Opt. Soc*. Am 70.8 (1980): 920-930.

[15] C-H. Teh and R. T. Chin, "On image analysis by the methods of moments", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10.4 (1988): 496-513.

[16] P. Daras, D. Zarpalas, A. Axenopoulos, D. Tzovaras and M. G. Strintzis, "Three-dimensional shape-structure comparison method for protein classification", *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2006, 3(3), 193-207.

[17] W.Cai, J. Xu, X. Shao, V. Leroux, A. Beautrait and B. Maigret, "SHEF: a vHTS geometrical filter using coefficients of spherical harmonic molecular surfaces", *Journal of molecular modeling*, 2008, 14(5), 393-401.

[18] R. J. Morris, R. J. Najmanovich, A. Kahraman and J. M. Thornton, "Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons" *Bioinformatics*, 2005, 21(10), 2347-2355.

[19] D. W. Ritchie and G. J. Kemp. "Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces." *Journal of Computational Chemistry* 20.4: 383-395, 1999.

[20] D. W. Ritchie and G. J. Kemp, "Protein docking using spherical polar Fourier correlations." *Proteins: Structure, Function, and Bioinformatics*, 39.2 (2000): 178-194.

[21] L. Mak, S. Grandison and R. J. Morris, "An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison", *Journal of Molecular Graphics and Modelling* 26.7 (2008): 1035-1045.

[22] L. Sael, B. Li, D. La, Y. Fang, K. Ramani, R. Rustamov and D. Kihara, (2008). Fast protein tertiary structure retrieval based on global surface shape similarity. Proteins: Structure, Function, and Bioinformatics, 72(4), 1259-1273.

[23] V. Venkatraman, Y. Yang, L. Sael and D. Kihara, "Protein-protein docking using region-based 3D Zernike descriptors", *Bmc Bioinformatics*, 10(1), 407, 2009.

[24] V. Venkatraman, P. R. Chakravarthy and D. Kihara, "Application of 3D Zernike descriptors to shape-based ligand similarity searching", *J Cheminform* 17.1 (2009): 19.

[25] Y. Fang, Y-S Liu and K. Ramani, "Three dimensional shape comparison of flexible proteins using the local-diameter descriptor", *BMC Structural Biology 2009*, 9:29 doi:10.1186/1472-6807-9-29.

[26] Y-S Liu, Y. Fang and K. Ramani, "IDSS: deformation invariant signatures for molecular shape comparison", *BMC Bioinformatics 2009*, 10:157 doi: 10.1186/1471-2105-10-157.

[27] Y-S Liu, K. Ramani and M. Liu, "Computing the Inner Distances of Volumetric Models for Articulated Shape Description with a Visibility Graph", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, December 2011.

[28] Y-S Liu, Q. Li, G-Q. Zheng, K. Ramani, W. Benjamin, "Using diffusion distances for flexible molecular shape comparison", *BMC Bioinformatics 2010*, 11:480.

[29] S. Yin, E. A. Proctor, A. A. Lugovskoy, and N. V. Dokholyan, "Fast screening of protein surfaces using geometric invariant fingerprints", *PNAS 2009*, vol. 106, no. 39, pp. 16622–16626, September 29, 2009.

[30] P. Heider, A. Pierre-Pierre, R. Li and C. Grimm, "Local shape descriptors, a survey and evaluation", *In Proceedings of the 4th Eurographics conference on 3D Object Retrieval*, pp. 49-56. Eurographics Association, 2011.

[31] Z. Lian, A. Godil, X. Sun, H. Zhang, "Non-rigid 3D shape retrieval using multidimensional scaling and bag-of-features", *in Proceedings of the International Conference on Image Processing (ICIP2010)*, 2010, pp.3181–3184.

[32] D. Smeets, T. Fabry, J. Hermans, D. Vandermeulen, P. Suetens, "Isometric deformation modeling for object recognition", *in Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns (CAIP'09)*, 2009, pp.757–765.

[33] G. Lavoue, "Bag of words and local spectral descriptor for 3D partial shape retrieval", *in Proceedings of the Eurographics Workshop on 3D Object Retrieval (3DOR'11),* 2011, pp.41–48.

[34] R. Ohbuchi, K. Osada, T. Furuya, T. Banno, "Salient Local Visual Features for Shape-Based 3D Model Retrieval", *Proc. IEEE International Conference on Shape Modeling and Applications (SMI'08)*, Stony Brook University, June 4 - 6, 2008.

[35] S. Kawamura, K. Usui, T. Furuya, and R. Ohbuchi. "Local goemetrical feature with spatial context for shape-based 3D model retrieval." *In Proceedings of the 5th Eurographics conference on 3D Object Retrieval*, pp. 55-58. Eurographics Association, 2012.

[36] P. Daras, A. Axenopoulos, G. Litos, "Investigating the Effects of Multiple Factors towards more Accurate 3D Object Retrieval", *IEEE Transactions on Multimedia*, Vol. 14, No. 2, Page(s): 374 – 388, April 2012.

[37] E. Wahl, U. Hillenbrand and G. Hirzinger, "Surflet-pair-relation histograms: a statistical 3D-shape representation for rapid classification", *IEEE Fourth International Conference on 3-D Digital Imaging and Modeling*, 3DIM 2003.

[38] N. Echols, D. Milburn and M. Gerstein, "MolMovDB: Analysis and visualization of conformational change and structural flexibility", *Nucleic Acids Research 2003*, 31:478-482.

[39] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, "Numerical recipes in C+: the art of scientific computing", *Cambridge: Cambridge University Press*, Vol. 994, 2009.

[40] R. Osada, T. Funkhouser, B. Chazelle and D. Dobkin, "Shape distributions", *ACM Transactions on Graphics,* 2002, 21(4):807-832.

[41] D. Zhou, O. Bousquet, T. N. Lal, J. Weston and B. Schölkopf, "Learning with Local and Global Consistency", *NIPS 2003*.

[42] B. Nadler, S. Lafon, R. R. Coifman, I. G. Kevrekidis, "Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators", *in Advances in Neural Information Processing Systems* 18, 2005.

[43] N. Huang, B. K. Shoichet, J. J. Irwin, "Benchmarking sets for molecular docking", J. Med. Chem. 2006, 49:6789-6801.

[44] J.J. Irwin, B. K. Shoichet, "ZINC–a free database of commercially available compounds for virtual screening", J. Chem. Inf. Model 2005, 45:177-182.

[45] O.S. Weislow, R. Kiser, D. L. Fine, J. Bader, R. H. Shoemaker, M. R. Boyd, "New soluble-formazan assay for HIV-1 cytopathic effects: application to highflux screening of synthetic and natural products for AIDS-antiviral activity", J. Natl. Cancer. Inst. 1989, 81:577-586.

[46] A. Bender, R. C. Glen, "A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication", J. Chem. Inf. Model 2005, 45:1369-1375.

[47] J. F. Truchon, C. I. Bayly, "Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem", J. Chem. Inf. Model 2007, 47:488-508.

[48] M. Hattori, Y. Okuno, S. Goto, M. Kanehisa, "Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways", J. Am. Chem. Soc. 2003, 125:11853-11865.

[49] L. Holm and C. Sander, "The FSSP Database: Fold Classification Based on Structure-Structure Alignment of Proteins," Nucleic Acids Research, vol. 24, pp. 206-210, 1996.

[50] L. Holm and C. Sander, "Touring Protein Fold Space with Dali/FSSP," Nucleic Acids Research, vol. 26, pp. 316-319, 1998.

[51] M.F. Sanner, A.J. Olson, and J.-C. Spehner. "Fast and robust computation of molecular surfaces". *In 11th ACM Symposium on Computational Geometry*, 1995.

[52] R. Kolodny, D. Petrey, and B. Honig, "Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction", *Curr. Opin. Struct. Biol.* 2006;16:393–398.

[53] I. N. Shindyalov, P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path", *Protein Eng.* 1998;11:739–747.

[54] P. Daras, A. Axenopoulos, "A Compact Multi-View Descriptor for 3D Object Retrieval" *IEEE 7th International Workshop on Content-Based Multimedia Indexing (CBMI 2009)*, Chania, Greece, Jun 2009.

[55] S. C. Flores, L. J. Lu, J. Yang, N. Carriero, and M. B. Gerstein, "Hinge Atlas: relating protein sequence to sites of structural flexibility", *BMC Bioinformatics* 8: 167, 2007.

[56] R. Gal, A. Shamir and D. Cohen-Or. "Pose-oblivious shape signature." *IEEE Transactions on Visualization and Computer Graphics*, 13.2 (2007): 261-271.

[57] R. M. Rustamov, "Laplace-Beltrami eigenfunctions for deformation invariant shape representation." *Proceedings of the fifth Eurographics symposium on Geometry processing. Eurographics Association*, 2007.

[58] L. Nanni, J. Y. Shi, S. Brahnam and A. Lumini, "Protein classification using texture descriptors extracted from the protein backbone image" *Journal of theoretical biology*, 264(3), 1024-1032, 2010.

[59] D.-Y. Chen, X.-P. Tian, Y.-T. Shen and M. Ouhyoung, "On visual similarity based 3D model retrieval" *In Computer graphics forum*, vol. 22, no. 3, pp. 223-232. Blackwell Publishing, Inc, 2003.

[60] T. Furuya and R. Ohbuchi, "Dense sampling and fast encoding for 3D model retrieval using bag-of-visual features", *CIVR 2009*, Article 26, doi>10.1145/1646396.1646430.

[61] Z. Lian, A. Godil, B. Bustos, M. Daoudi, J. Hermans, S. Kawamura, Y. Kurita, G. Lavoué, H. Van Nguyen, R. Ohbuchi, Y. Ohkita, Y. Ohishi, F. Porikli, M. Reuter, I. Sipiran, D. Smeets, P. Suetens, H. Tabia, D. Vandermeulen, "A comparison of methods for non-rigid 3D shape retrieval", *Pattern Recognition*, Volume 46 Issue 1, January, 2013, Pages 449-461.

[62] B. Li, A. Godil, M. Aono, X. Bai, T. Furuya, L. Li, R. López-Sastre, H. Johan, R. Ohbuchi, C. Redondo-Cabrera, A. Tatsuma, T. Yanagimachi, and S. Zhang, "SHREC'12 Track: Generic 3D Shape Retrieval", *Eurographics Workshop on 3D Object Retrieval 2012*.

[63] Y. Ohkita, Y. Ohishi, T. Furuya, R. Ohbuchi, "Non-rigid 3D Model Retrieval Using Set of Local Statistical Features", *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp.593-598, 2012.

[64] Y. Zhang, J. Skolnick, "TM-align: A protein structure alignment algorithm based on TM-score ", *Nucleic Acids Research*, 2005 33: 2302-2309.

[65] Y. Ye and A. Godzik, "Flexible structure alignment by chaining aligned fragment pairs allowing twists", *Bioinformatics*, 2003, 19(Suppl 2):II246-II255.

[66] I.N. Shindyalov, P.E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path", *Protein Engineering* 11(9) 739-747, 1998.

[67] M. Shatsky, H.J. Wolfson, and R. Nussinov, "Flexible protein alignment and hinge detection", *Proteins: Structure, Function, and Genetics*, 48:242-256, 2002.

[68] Z. Li, P. Natarajan, Y. Ye, T. Hrabe, A. Godzik, "POSA: a user-driven, interactive multiple protein structure alignment server", *Nucl. Acids Res.* (2014) doi: 10.1093/nar/gku394.
[69] A.Mademlis, P.Daras, D.Tzovaras and M.G.Strintzis, "3D Object Retrieval based on Resulting Fields", *29th International conference on EUROGRAPHICS 2008*, workshop on 3D object retrieval, Crete, Greece, Apr 2008.
[70] S. C. Flores and M. B. Gerstein, "FlexOracle: predicting flexible hinges by identification of stable domains", *BMC bioinformatics* 8.1 (2007): 215.
[71] Y.Y. Tseng, J. Dundas and J. Liang. "Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns." *Journal of molecular biology* 387.2 (2009): 451-464.
[72] T. A. Binkowski, L. Adamian and J. Liang. "Inferring functional relationships of proteins from local sequence and spatial surface patterns." *Journal of molecular biology* 332.2 (2003): 505-526.
[73] J. Konc and D. Janežič. "ProBiS-2012: web server and web services for detection of structurally similar binding sites in proteins." *Nucleic acids research* 40.W1 (2012): W214-W221.
[74] A. Sharma, A. Papanikolaou and E. S. Manolakos. "Accelerating all-to-all protein structures comparison with TMalign using a NoC many-cores processor architecture." *Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW)*, 2013 IEEE 27th International. IEEE, 2013.
[75] B. Y. Chen and B. Honig. "VASP: a volumetric analysis of surface properties yields insights into protein-ligand binding specificity." *PLoS computational biology* 6.8 (2010): e1000881.
[76] B. Y. Chen, "VASP-E: Specificity Annotation with a Volumetric Analysis of Electrostatic Isopotentials." *PLoS computational biology* 10.8 (2014): e1003792.
[77] S. R. Amin et al. "Prediction and experimental validation of enzyme substrate specificity in protein structures." *Proceedings of the National Academy of Sciences* 110.45 (2013): E4195-E4202.

Apostolos Axenopoulos was born in Thessaloniki, Greece, in 1980. He received the Diploma degree in electrical and computer engineering and the M.S. degree in advanced computing systems from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2003 and 2006, respectively, and the PhD in Electrical and Computer Engineering from the University of Thessaly in 2014. He is an Associate Researcher at the Information Technologies Institute, Thessaloniki. His main research interestsinclude 3D object indexing, content-based search and retrieval and bioinformatics.

Dimitrios Rafailidis was born in Larissa, Greece, in 1982. He received the Diploma in Informatics from the Computer Science Department, the M.Sc. degree in Information Systems and the Ph.D. degree in Information Retrieval from the Aristotle University of Thessaloniki, Greece in 2005, 2007 and 2011, respectively. His main research interests include Machine Learning and Pattern Recognition, Multimedia Information Retrieval, Databases, Social Media and Artificial Intelligence Systems.

Petros Daras (M'07) was born in Athens, Greece, in 1974. He received the Diploma degree in electrical and computer engineering, the M.Sc. degree in medical informatics, and the Ph.D. degree in electrical and computer engineering, all from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1999, 2002, and 2005, respectively. He is a Researcher Grade C, at the Information Technologies Institute (ITI) of the Centre for Research and Technology Hellas (CERTH). His main research interests include search, retrieval and recognition of 3D objects, 3D object processing, medical informatics applications, medical image processing, 3D object watermarking and bioinformatics. He regularly serves as a reviewer/evaluator of European projects and he is a member of IEEE, a key member of the IEEE MMTC 3DRPC IG and chair of the IEEE Image, Video and Mesh Coding IG.

Georgios Papadopoulos received his diploma in Physics from the Aristotle University of Thessaloniki/Greece in 1977. He studied further theoretical Biophysics in the department of Physics of the Freie Universitat Berlin/Germany and received his Ph.D degree in Biophysics from the same department in 1989. He worked with short term contracts as guest scientist in the Hahn-Meitner Institut/Berlin/Germany and in the Fosrchungszentrum Julich/Germany. Since 1994 he has been teaching Physics, Biostatistics, Bioinformatics, Physical Chemistry and Biophysics in the University of Thessaly/Greece, Democritus University of Thrace/Greece and the Aristotle University of Thessaloniki/Greece. Since 2009 he is lecturer of Biophysics in the department of Biochemistry & Biotechnology/UTh. His Research interests are focused on the study of the structure of biological macromolecules and of their interactions using theoretical and computational methods.

Elias. N. Houstis is currently a full Professor of Computer Engineering and Communications department at University of Thessaly, Greece, Director of Research Center of Thessaly (CE.RE.TE.TH.), and Emeritus Professor of Purdue University. USA. Most of his academic career is associated with Purdue University. He has been a Professor of Computer Science and Director of the Computational Science & Engineering Program of Purdue University. He is a member of working groups WG2.5 IFIP on mathematical software and European ICT Directors. Houstis' current research interests are in the areas of problem solving environments, networking and parallel computing, enterprise systems, computational intelligence and finance, and e-services.