Article:
Seddon, M.P., Cosgrove, D.A., Packer, M.J. et al. (2019) Alignment-free molecular shape comparison using spectral geometry: the framework. Journal of Chemical Information and Modeling, 59 (1). pp. 98-116. ISSN: 1549-9596

https://doi.org/10.1021/acs.jcim.8b00676

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 98−116

pubs.acs.org/jcim

Alignment-Free Molecular Shape Comparison Using Spectral Geometry: The Framework

Matthew P. Seddon,[†] David A. Cosgrove,[‡] Martin J. Packer,[§] and Valerie J. Gillet*[,†]

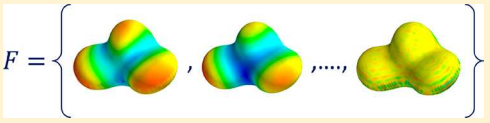
[†]Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, United Kingdom

[‡]Discovery Sciences, IMED Biotech Unit, AstraZeneca, Cambridge CB4 0WG, United Kingdom

[§]Chemistry, Oncology, IMED Biotech Unit, AstraZeneca, Cambridge CB4 0WG, United Kingdom

Supporting Information

ABSTRACT: A framework is presented for the calculation of novel alignment-free descriptors of molecular shape. The methods are based on the technique of spectral geometry which has been developed in the field of computer vision where it has shown impressive performance for the comparison of deformable objects such as people and animals. Spectral geometry techniques encode shape by capturing the curvature of the surface of an object into a compact, information-rich representation that is alignment-free while also being invariant to isometric deformations, that is, changes that do not distort distances over the surface. Here, we adapt the technique to the new domain of molecular shape representation. We describe a series of parametrization steps aimed at optimizing the method for this new domain. Our focus here is on demonstrating that the basic approach is able to capture a molecular shape into a compact and information-rich descriptor. We demonstrate improved performance in virtual screening over a more established alignment-free method and impressive performance compared to a more accurate, but much more computationally demanding, alignment-based approach.



INTRODUCTION

The development of in silico methods for shape-based searching of small molecules has been a topic of considerable interest for many years.[1−3] This is due to shape being fundamental to molecular recognition events such as a drug binding to a biological receptor. Given one or more known active compounds, shape searching can be used to identify molecules within databases that can adopt similar shapes to the query compound(s) and therefore may also bind to the receptor of interest. Shape-based searching forms one of a number of virtual screening methods that are applied in the absence of the 3D structure of the receptor. A key advantage of shape searching over 2D fragment-based methods is that it is more amenable to scaffold hopping, that is, finding hits that belong to different chemical series. This is important for drug discovery projects since it allows them to be moved into new areas of chemical space, thus increasing the chance of generating new intellectual property while also mitigating against potential side effects or synthetic intractability associated with existing compounds.

The different approaches to shape-based virtual screening can be divided into alignment-based and alignment-free approaches. Alignment-based methods require that a database molecule is superimposed on the query prior to calculating shape similarity. The aim of the alignment step is to maximize the similarity, for example, by finding the maximum overlap of the molecules which can be time consuming. ROCS, Rapid Overlay of Chemical Structures,[4] the industry-standard alignment method, uses Gaussian functions to represent atomic volumes which allow the rapid calculation of the overlap volume of aligned molecules. Other 3D alignment methods include the use of spherical harmonics[5] and field-based representations.[6] The computational complexity of alignment-based methods is such that they struggle to cope with the sizes of data sets, real and virtual, that are currently available for search. For example, the ExCAPE-DB database, compiled from ChEMBL and PubChem, consists of 70 million SAR (structure−activity relationship) data points.[7] The Enamine REAL database consists of 680 million compounds that are available for purchase through one-step synthesis, and the GDB-17 database of virtual compounds with up to 17 heavy atoms consists of 166 billion compounds.[8]

A number of alignment-free approaches have been developed in which shape is represented in vector form. They typically involve abstracting the 3D features of a molecule, such as interatomic distances and/or angles, into a vector representation. The resulting descriptors are intrinsic; that is, they are independent of the embedding space, which means that they are independent of the orientation of the molecule in 3D space and invariant to rotation and translation. Thus, the descriptors can be compared directly without the need to superimpose the molecules. For example, in UFSR (UltraFast Shape Recognition), a vector description of shape is calculated from interatomic distance distributions derived from a set of four reference locations.[9] Although, alignment-free methods allow rapid pairwise comparisons to be made, there is significant information loss in the representation which can affect performance accuracy. For comprehensive reviews of 3D similarity searching methods, see refs 1−3.

Received: October 1, 2018
Published: November 21, 2018

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

Whether alignment-based or alignment-free, the handling of molecular flexibility is a major challenge for 3D similarity searching. Most methods require explicit 3D models with conformational space being explored prior to searching through the enumeration of an ensemble of conformers for each molecule. The aim is to sample conformational space at a resolution that is sufficient to include all low energy conformations but not so exhaustive that excessive numbers of conformers are produced. Typical sampling strategies are based on a threshold on strain energy, or root-mean-square deviation of atom positions, or simply on the maximum number of conformers permitted.[10] Ideally, the sampling method will ensure that the bioactive conformations of the molecules are represented; however, it may be that the bound conformation is not a minimum energy conformation. The ensemble approach clearly increases the number of discrete comparisons required, typically by a further 1 or 2 orders of magnitude, depending on the resolution of the conformational search. Although 3D similarity methods are appealing conceptually, the issue of conformational flexibility makes them considerably more complex than 2D methods, and effective conformational sampling remains a challenging area.[11,12]

Shape matching is a topic which has received considerable attention in the field of computer vision where a very active area of research is the development of techniques to recognize shapes that can undergo deformations, for example, people or animals that can adopt different poses. Spectral geometry techniques have become the method of choice for this challenging problem. These techniques encode shape by capturing the curvature of the surface of an object into a compact, information-rich representation. As for the alignment-free descriptors mentioned above, spectral geometry descriptors are intrinsic and therefore independent of the embedding space. A key element of spectral geometry that differs from the above approaches is the encoding of geodesic distances, that is, distances over the surface of a shape rather than Euclidean or through-space distances. Taking the earth as an example, the geodesic distance between two cities is the distance measured over the surface of the earth, whereas the Euclidean distance is the shortest through-space distance and would pass through the earth. This property means that as well as being invariant to rotation and translation, the descriptors are also invariant to isometric deformations, that is, to changes to the "pose" of an object that preserve distances over its surface. The typical example from computer vision is recognizing a person in different poses, for example, standing and sitting. In this case, the property of isometry is illustrated by measuring the geodesic distance between the head and the feet over the surface of the body; this distance remains the same regardless of the pose, whereas the head and feet are much closer in the sitting pose when measured using Euclidean distance.

The original spectral shape method was developed by Reuter et al.[13] who showed that the spectrum (eigenvalues and eigenfunctions) of the Laplace−Beltrami operator over the surface of a shape can be used to develop a descriptor of the isometric geometry of the shape. The descriptors are generated by first representing the surface as a mesh. Geometric properties of the mesh then form the input to an eigendecomposition which transforms the original space into a set of eigenpairs (eigenvalues and corresponding eigenfunctions). The eigenpairs are orthonormal and are ordered on the extent to which they capture the original untransformed data. The process of transforming the geometric properties of a mesh over the surface of an object into eigenpairs can be considered analogous to transforming a high dimensional space to a low dimensional space using principal components analysis (PCA). Reuter et al.'s original descriptor consisted of the eigenvalues of the spectrum only, represented as a vector of non-negative numbers. They named the descriptor Shape-DNA to emphasize that it captures the intrinsic geometry of shape, in analogy to DNA, which characterizes an individual according to their gene sequence.

While the Shape-DNA descriptor was shown to be useful as a shape signature, it is limited in the extent to which it captures rich descriptions of shape. Following Reuter et al.'s work, local geometry descriptors were developed by mapping the geometric information contained in the eigenfunctions of the spectrum onto points on the surface of a shape. Local geometry descriptors are feature vectors that are assigned to each vertex. The mapping is achieved by applying functions to the spectrum to amplify different parts of it. The most commonly used functions are the Heat Kernel Signature, HKS,[14] and the Wave Kernel Signature, WKS.[15] Local geometry descriptors have been applied to tasks such as establishing correspondences between shapes and shape segmentation. Techniques have also been developed to aggregate local descriptors into whole object, or global, descriptors for domain-specific applications.[16] A survey of spectral geometry methods is provided by Li and Hamza[17] and the results of a recent benchmark study on a wide range of objects in different poses in which spectral geometry showed excellent performance is summarized by Lian et al.[18]

Diffusion distance (DD) is a related technique that has been applied to protein structure comparison.[19,20] DDs are based on topological (inner) distances which may be more appropriate for capturing the large articulations seen in protein conformations than spectral geometry, which is susceptible to distortions arising from surface "stretching" for large movements. DDs were also applied to small molecules;[20] however, we believe surface geometry will be significantly more discriminating than inner distances at representing the shapes of small molecules. We also believe that small molecules will be much less susceptible to distortion effects than proteins since the conformational changes of small molecules do not result in significant "stretching" of the surface.

In this paper, we present a framework for the application of spectral geometry methods to molecular shape comparison. To our knowledge, this is the first time that the technique has been applied to shape comparison of small molecules. The framework consists of a number of discrete steps each of which is parametrized for shape-based virtual screening. The aim of the work reported here is to demonstrate that the basic approach is effective in generating a compact and effective descriptor of molecular shape. To this end, we apply the methods to virtual screening based on a single conformer as input for each molecule and compare our results with established alignment-based and alignment-free methods. While we believe there is considerable potential for handling conformational flexibility, and indeed, this is one of the attractions of the approach, this aspect will be explored in future publications, following on from this work.

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

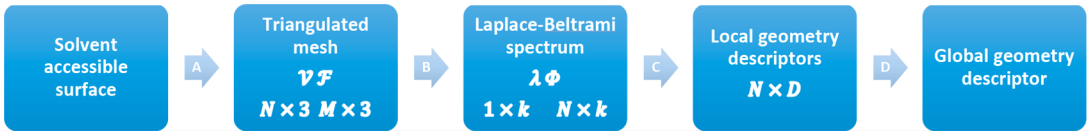


Figure 1. Framework for applying spectral geometry to generate molecular shape descriptors. The surface is represented by a triangulated mesh consisting of N vertices, $\mathcal{V}$, and M faces, $\mathcal{F}$. The eigendecomposition results in a k-dimensional vector of eigenvalues (λ) and an $N \times k$ matrix of eigenfunctions (ϕ). The local geometry descriptors of dimension D are derived for each vertex of the mesh. The local geometry descriptors are then aggregated to form an alignment-free global geometry descriptor.

METHODS

Overview. The framework for generating spectral geometry descriptors for molecular shapes is shown in Figure 1. (The discussion of methods has been simplified for ease of exposition, and the reader is referred to the Supporting Information for a detailed mathematical description of spectral geometry.) The first step is to generate a molecular surface in the form of a discrete triangulated mesh. Geometric properties of the surface are captured by solving the Laplace−Beltrami operator, Δ, over the mesh to give the spectrum:[7]

$$\Delta\phi = -\lambda\phi \tag{1}$$

The spectrum is a set of k eigenpairs represented as a k-dimensional vector of eigenvalues (λ) and an $N \times k$ matrix of eigenfunctions (ϕ). The spectrum then forms the basis of the computation of local geometry descriptors which are calculated for each point of the mesh. Finally, the local geometry descriptors are used to generate the global geometry descriptor which represents the shape descriptor of a molecule. Each of these steps is now described in further detail.

Representation of Molecular Shape. The mesh used to represent the surface of a molecule must meet a number of conditions in order that the Laplace−Beltrami operator can be solved.[21,22] First, the mesh should be triangulated. A mesh is a lattice graph in 3D space composed of vertices and edges: Each vertex has (x, y, z) coordinates, and a connection between two vertices is called an edge. The vertices are connected such that each edge forms the boundary of an enclosed region which is known as a face. Figure 2 shows a triangulated mesh with six

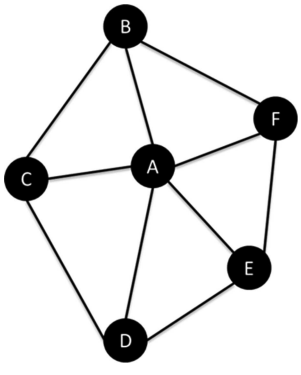


Figure 2. Triangular mesh consisting of six vertices $\{A, B, C, D, E, F\}$ and five faces $\{(A,B,C), (A,C,D), (A,D,E), (A,E,F), (A,F,B)\}$.

vertices, $\{A, B, C, D, E, F\}$ and five faces $\{(A,B,C), (A,C,D), (A,D,E), (A,E,F), (A,F,B)\}$. As all edges must form the boundary of an enclosed face, this means that each edge must be a member of at least one face and that a mesh may be described entirely by its vertices and faces. A 3D mesh, X, can therefore be defined by a set of N vertices, $\mathcal{V} = \{v_1, v_2, ..., v_i, ..., v_N\}$, and M faces, $f = \{f_1, f_2, ..., f_i, ..., f_M\}$.

Second, the mesh must be such that it is possible to trace a path over edges between any two vertices. This ensures that a notion of distance exists for all points on the mesh and that no parts of the mesh are disconnected from other parts. Third, there must be a strictly positive distance between all points. In practice, this means that duplicate vertices cannot exist. Finally, the mesh must not contain nonmanifold vertices and edges since these cannot be handled by most algorithms due to the geodesic behavior around them being poorly defined.[22] A nonmanifold vertex is one where two surfaces meet at a single point, as illustrated on the left of Figure 3. A nonmanifold edge

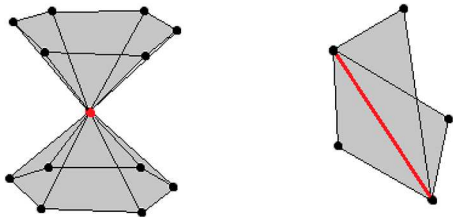


Figure 3. Illustration of nonmanifold vertices, see text for details.

is a member of more than two faces, creating a self-intersection, right of Figure 3. Note that a mesh may still have a boundary, that is, a collection of edges that only belong to one face. A mesh with no boundary edges is called a closed mesh.

Computing the Spectrum. The spectrum of the Laplace−Beltrami operator can be computed using techniques in linear algebra either directly or indirectly. The direct approach defines the Laplace−Beltrami operator as a matrix, $L_{N\times N}$, where N is the number of vertices in the mesh, and the elements of the matrix are assigned weights to represent the relationship between any two vertices. The cotangent weighting scheme, first described by Pinkall and Polthier,[23] defines the elements of the matrix as

$$L_{i,j} = \begin{cases} 1 & \text{if } i = j \\ w_{i,j} & \text{if } i \neq j \wedge j \in R(i) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where i and j are vertices and $R(i)$ is the set of vertices connected to vertex i. The diagonal of the matrix ($i = j$) is assigned values of 1. Geometric information is computed for adjacent vertices using the average cotangent of the opposite angles, that is, $w_{i,j} = \frac{1}{2}(\cot\theta_1 + \cot\theta_2)$ as illustrated in Figure 4 for the weight between vertices A and C. Then, in order to obtain the spectrum of the Laplace−Beltrami operator, eq 3 is solved numerically,

$$L\phi = \lambda\phi \tag{3}$$

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

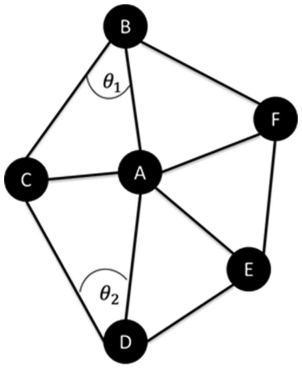


Figure 4. Cotangent weighting scheme. See text for details.

where ϕ are the eigenfunctions and λ the corresponding eigenvalues. Note this is a very sparse system as the vast majority of vertices are not connected, and therefore, it can be solved using sparse eigenvalue methods.

The spectrum can also by computed indirectly using the Finite Element Method (FEM), which makes it less dependent on the underlying mesh representation.[13] The FEM algorithm is mathematically complex and the details are not reported here as the cotangent method provides a more intuitive explanation of the method. In brief, the FEM algorithm constructs local stiffness and mass matrices for the vertices on the mesh and combines them over the whole mesh. The Laplace−Beltrami spectrum is then recovered from computing the generalized eigenvalue decomposition of these two matrices. Both the direct and indirect approaches were implemented and applied to the surface meshes used here (see later). Preliminary results, not reported here, demonstrated the superiority of FEM so that all spectra reported in this paper are computed using FEM.

The eigenpairs are orthonormal and are ordered on the extent to which they capture the original untransformed data (compare with principal components analysis where the first principal components are most significant in terms of encoding the variance of the original space). Eigenpairs that are near the beginning of the sequence encode global properties of the shape, whereas those lower down in the sequence encode more local features. For a mesh of N vertices, a full eigendecomposition would result in N eigenpairs; however, as the information content of successive eigenpairs is reduced, the number of eigenpairs is usually truncated. We use the variable k in Figure 1 to refer to the number of eigenpairs that is calculated so that the output from the eigendecomposition is a k-dimensional vector of eigenvalues (λ) and an $N \times k$ matrix of eigenfunctions (ϕ).

Local Geometry Descriptors. Local geometry descriptors consist of feature vectors that are assigned to each vertex (mesh point) on the surface and describe the shape around the vertex. They are generated using functions that act on the spectrum. The HKS consists of an exponential function of the eigenvalues that acts over the matrix of eigenfunctions and captures the notion of heat decay at each vertex, x, at time t, eq 4.[14]

$$\mathrm{HKS}_t(x) = \sum_{i=1}^{k} \exp(-\lambda_i t)\phi_i(x)^2 \tag{4}$$

The HKS function is evaluated as a $1 \times N$ vector where each element corresponds to a vertex on the mesh. For each vertex, it can be interpreted as a sample of the heat dissipation over time, t, and since heat dissipation is determined by the intrinsic geometry of the surface, it forms a descriptor of the local geometry of the vertex. The exponential function is shown graphically in Figure 5(a). For small values of t, greater weight

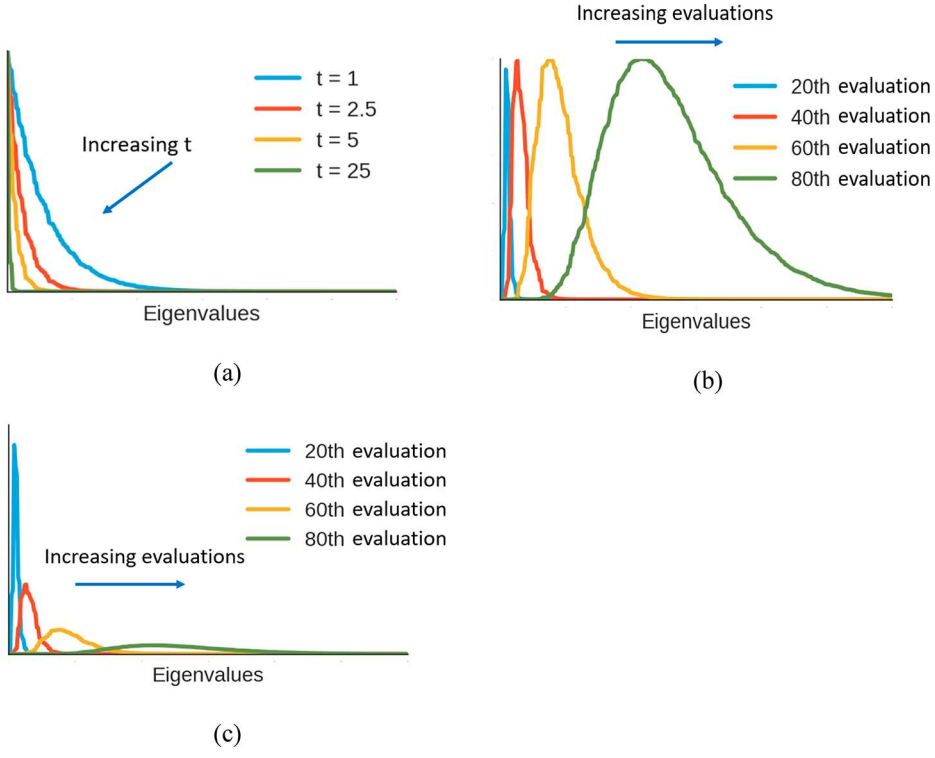


Figure 5. Illustration of the HKS and WKS functions used to derive local geometry descriptors.

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

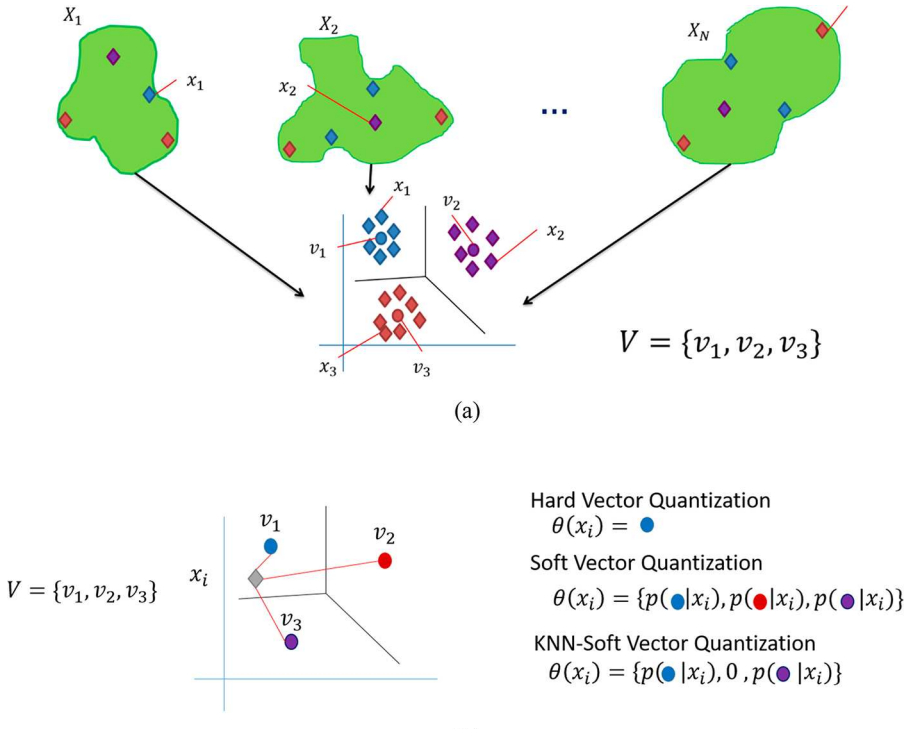


Figure 6. Bag-of-features method for generating global descriptors. (a) Vertices extracted from a set of molecules are clustered on their local geometry descriptors and a representative subset selected for use as a codebook. (b) Three different encoding schemes are used to map a given vertex onto the codebook.

is given to eigenfunctions that occur later in the spectrum and represent local geometry. As t increases, more emphasis is given to global features. Sampling the heat at D time points gives an $N \times D$ matrix where each row is a vertex in the mesh and each column is the HKS at a given time. Thus, each row of the matrix can be considered as a local geometry descriptor for that vertex represented by a D-dimensional vector with each element representing a different balance of global and local geometry features, eq 5.

$$\boldsymbol{p}(x) = (\mathrm{HKS}_1(x), ..., \mathrm{HKS}_D(x)) \tag{5}$$

The WKS is an alternative local geometry descriptor that has its roots in quantum mechanics and is parametrized by frequency rather than time.[15] The WKS samples the spectrum for a specified number of intervals, called evaluations in the original paper by Aubry et al. The spectrum is divided into equally sized intervals/evaluations and a Gaussian function is centered at the middle of each to amplify the signal around that point. For a given interval, E, the WKS at vertex x is

$$\mathrm{WKS}_E(x) = c_E \sum_{i=1}^{k} \exp\left(-\frac{(\log e - \log \lambda_i)^2}{2\sigma^2}\right)\phi_i(x)^2 \tag{6}$$

where e is the mean value in the Eth interval and the nominator is the squared distance of the log of the ith eigenvalue to the log of the middle of the interval. The σ^2 in the denominator is an arbitrary parameter that represents the variance of the log-normal distribution. Previous work has established that the value of $\sigma^2 = 7$ gives best performance and this is therefore adopted here.[15] As the WKS moves across the spectrum, the contribution shifts from global to local features (Figure 5(b)). In order to weight the contributions over different intervals equally, a normalization constant, c_E, is applied so that the area under each function is the same (Figure 5(c)). The number of intervals, or evaluations, used to calculate the WKS determines the number of elements in the local geometry descriptor assigned to each vertex and is referred to as D for consistency with the local geometry descriptors calculated using HKS. Thus, as for the HKS, each vertex is described by a D-dimensional vector with each element representing a different balance between local and global features:

$$\boldsymbol{p}(x) = (\mathrm{WKS}_1(x), ..., \mathrm{WKS}_D(x)) \tag{7}$$

Global Geometry Descriptors. The local geometry descriptors must be mapped to a global geometry descriptor in order to quantify the similarity of two objects in an alignment-free way. We have explored two options here: the covariance matrix method and the bag-of-features.

The covariance matrix method does not require any parametrization. It maps all shapes represented by $N \times D$ local geometry descriptors to the same space, that is, the space of $D \times D$ covariance matrices.[24] Therefore, the size of the covariance matrix is dependent only on D, the number of dimensions chosen for the local geometry descriptors, and not on the size of the mesh (or the underlying molecule). The covariance matrix is expressed as the covariance between the columns of the local geometry descriptors and is also independent of the ordering of the vertices in the mesh. The covariance matrices for two molecules in arbitrary orientation, C_1 and C_2, can be compared directly element-by-element using the Bray–Curtis metric[25] as follows:

$$d(C_1, C_2) = \frac{\sum_i |c_{1,i} - c_{2,i}|}{\sum_i |c_{1,i} + c_{2,i}|} \tag{8}$$

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

where i moves over all $D \times D$ elements of the two covariance matrices. This element-by-element metric was chosen over a vector−space method such as cosine, which relies on the inner product, since the inputs are flattened matrices.

Bag-of-features descriptors are the most common form of global geometry descriptor used in computer vision.[16] They have a longer history in the field of image processing and signal compression and originate as descriptors for text retrieval.[26] The method uses a *codebook* that represents key geometric features, or *codewords*, in feature space. The vertices of a given shape are mapped to the representative features, in a process known as *encoding*, and the frequency of occurrence of the codewords is aggregated. For example, suppose the codebook contains a codeword which is the local geometry descriptor of a vertex in a cupula-like region of a molecular shape. Then, each vertex of a given shape can be characterized on cupula-likeness by determining how close its local geometry descriptor is to the cupula codeword. In practice, codewords do not necessarily have nameable geometric properties. A key advantage of the bag-of-features approach compared to the covariance descriptor is that it is more compact: for a codebook with V codewords, an $N \times D$ matrix will be mapped to a $1 \times V$ vector rather than the D^2 matrix generated using the covariance method. However, there are a number of parameters that control the generation of bag-of-features descriptors, and these need to be optimized for the domain. The basic approach is presented here and parametrization experiments are reported below.

An overview of the workflow used to develop bag-of-feature descriptors is presented in Figure 6. First, a codebook, V, is computed from the set of local geometry descriptors, x_1, x_2...x_M, extracted from a sample of N molecules, X_1,X_2...X_N. The local geometry descriptors are clustered using k-means and a subset selected to form the representative features in the codebook. This process is shown schematically in Figure 6(a), where the local geometry descriptors are two dimensional. The blue vertices represent valley-like features. The red vertices represent cupula-like features, and the purple vertices represent flat regions. The codebook, V, is formed by selecting the centroid vertex in each cluster and consists of three codewords, v_1, v_2, and v_3, in the illustration. Given the large number of data points to be clustered, to avoid running into convergence issues with traditional k-means, we used the Mini-batch k-means algorithm which uses a subsampling strategy to provide fast training convergence.[27]

The process of calculating a global shape descriptor for an input molecule is illustrated in Figure 6(b) and involves comparing the local geometry descriptor at each vertex (mesh point) with each of the features in the codebook. The resulting encodings are aggregated to form a histogram. Three different encoding schemes were used here: Hard Vector Quantization; Soft Vector Quantization, and KNN-Soft Vector Quantization.

Let $\theta(x)$ denote the encoding of a vertex, $x \in X$. Hard Vector Quantization (HQ) is the simplest encoding method whereby each vertex, $x \in X$, is assigned to the closest codeword in the codebook based on its local descriptor $g(x)$,

$$\theta(x) = \underset{v_i \in V}{\operatorname{argmin}}\{d(g(x), v_i)\} \tag{9}$$

for codewords $v_i \in V$.

Soft Vector Quantization (SQ) attempts to reduce the amount of information lost by allocating a vertex to a single codeword by assigning a vector of probabilities to each vertex. Each vertex, $x \in X$, is assigned a vector of size $1 \times V$, where V is the number of codewords in the codebook. Then the ith element of the vector represents the probability that the local geometry of the vertex is close to the ith codeword in the codebook,

$$\theta(x) = \{p(v_1|x), ..., p(v_V|x)\} \tag{10}$$

where the probability scores are determined using the softmax formula,[16]

$$p(v_i|x) = c(x)\exp\left(\frac{\|g(x) - v_i\|_2^2}{2\sigma^2}\right) \tag{11}$$

where $c(x)$ is a normalization constant that ensures $\theta(x)_1 = 1$.

KNN-Soft Vector Quantization (KNN) is an attempt to balance the trade-off between the information loss of HQ with the increased noise of SQ due to the allocation of distant points to codewords. This encoding method assigns the softmax probability to the K-nearest codewords to each local descriptor. The encoding is therefore defined as

$$\theta(x) = \begin{cases} c(x)\exp\left(\frac{\|g(x) - v_i\|_2^2}{2\sigma^2}\right) & \forall\ v_i \in \text{KNN} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where KNN is the set of the K-nearest codewords to the local descriptor $g(x)$. In Figure 6(b), when $K = 2$ nearest neighbors, the descriptor will assign a stronger membership of the blue dot than the orange dot, which reflects its closer proximity to blue, and will assign a zero value to the purple feature as this feature is not a near neighbor.

For all encoding methods, the vertex encodings are aggregated over the shape to produce a frequency histogram. Finally, the histogram is l^2 normalized to give the global shape descriptor. This is done by dividing each element of the histogram by the l^2 norm, θ_2, which is the square root of the sum of the squared elements in the histogram

$$\|\theta\|_2 = \sqrt{\sum_v \theta_v} \tag{13}$$

for all v elements in the histogram. For all bag-of-features methods, the distance between the shape descriptors of two molecules is calculated using cosine distance. As these descriptors are vectors, the normalized inner-product distance was considered the natural metric in this descriptor space.

Data Sets. The performance of the spectral geometry descriptors has been evaluated on virtual screening experiments on the DUD-E data set.[28] DUD-E consists of sets of actives and decoys for 102 biological targets and was designed to provide a benchmark data set for docking programs by providing challenging decoys which were chosen to have similar physicochemical properties to the actives but dissimilar 2D topologies. DUD-E is not appropriate for evaluating ligand-based virtual screening using 2D chemical descriptors since the decoys are dissimilar by design thereby biasing the retrieval toward the known actives. Such a bias does not apply here, however, since the focus is on identifying ligands that have similar shape to a query active regardless of their 2D topology. In order to avoid dependence on conformation generation techniques, the 3D structures available in the data set download were used directly, and all of the experiments

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

reported here are based on this single rigid conformation per DUD-E entry.

First, a series of experiments was conducted on a subset of the targets extracted from DUD-E in order to parametrize the spectral geometry descriptors for virtual screening. The subset consisted of the following 10 targets: ada, comt, esr1, glcm, hxk4, kit, mapk2, pa2ga, ptn1, tysy. For each target, 20 active molecules were selected at random and mixed with 380 randomly selected decoy molecules. Each active was then used as a query, and the virtual screening performance was measured using AUC and BEDROC scores with the values averaged over each active and each target.

Following the parametrization experiments, virtual screening was carried out for all targets in DUD-E except for the 10 used in the parametrization experiments (and except target fgrf for which there were no 3D structures available in the download). First, the crystal ligand was used as query for each target. Next, a more extensive set of experiments was conducted in which 20 active molecules were selected randomly as queries for each target and searched over all actives and decoys in the set, and the results were averaged first for each target and then over all targets. To ensure the virtual screening results were not biased by multiple occurrences of the same molecule in the rankings, once the virtual screening had been performed against a query molecule, the results were filtered to find the best performing structure for each unique CHEMBL ID prior to the AUC and BEDROC scores being calculated.

RESULTS

Mesh Generation. The first step in generating spectral geometry descriptors is to compute meshes over the surfaces of the molecules. We generated meshes using the recent TMSmesh program, which was created for analytical use of meshes.[29,30] TMSmesh uses an atom-centered Gaussian that can be parametrized to approximate different molecular surfaces. The meshes output by TMSmesh are consistent with the requirements for computing the spectrum of the Laplace−Beltrami operator. Parameters were chosen that best approximate the solvent accessible surface.[30] These are a mesh density setting of H = 0.2, a Gaussian surface decay rate of D = 0.4, and the Gaussian surface isovalue, which controls the volume enclosed by the surface, of C = 1.2. TMSmesh is configured by specifying a density parameter rather than a specific number of mesh vertices so that the number of vertices in the mesh is not controlled. The overall distribution of mesh sizes is presented in Figure 7. The mean number of vertices in the mesh is 10,490 with the range being between 1367 and 20,130 (standard deviation 1,985). The mean number of vertices in the meshes generated for the actives and decoys for each DUD-E target ranges from 7900 to 12,900.

Visualizing the Spectrum. The eigenpairs of the spectrum are output in increasing order of eigenvalue and are structured such that the information content of the eigenfunctions moves from encoding global geometric features to more localized features. This is illustrated in Figure 8 for an example molecule which shows different eigenfunctions (columns of the $N \times k$ spectrum) plotted onto the surface of the molecule. Figure 8(c) shows the first eigenfunction with the colors aligned along the longest part of the molecule, which can be thought of as the x-axis. This eigenfunction can be considered analogous to the first component in PCA and shows the direction of largest variation in geometric terms. As the eigenvalues increase in size, the corresponding eigenfunctions show orthogonal *directions* (or axes) which capture smaller degrees of geometric variation. The fifth eigenfunction shows global shape variation in two directions, along the z-axis as well as starting in the middle and moving out along the y-axis. Figure 8(e) and (f) show the 10th and 250th eigenfunctions, respectively, which show more local variation over small sections of the surface of the molecule. Therefore, the smaller eigenvalues encode global intrinsic geometry, with the larger eigenvalues corresponding to local geometry.

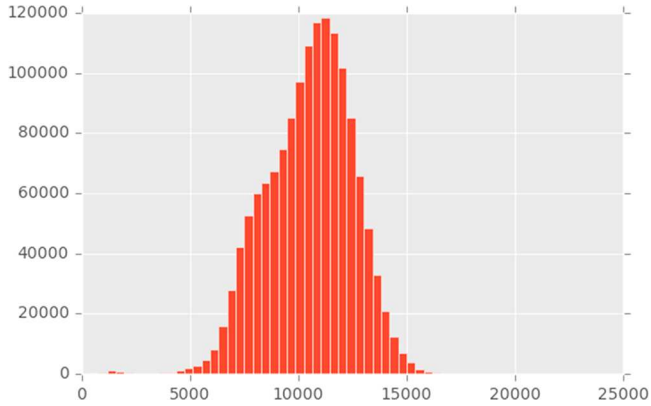


Figure 7. Distribution of the number of vertices in the meshes of the DUD-E data set.

Visualizing the Local Geometry Descriptors. Figure 9 illustrates local geometry descriptors calculated using the HKS at three different time points. For a given time t, the HKS function is evaluated as a $1 \times N$ vector where each element corresponds to a vertex on the mesh, and the distribution of values can be plotted over the surface. At time $t = 5$, the HKS emphasizes eigenfunctions that occur later in the spectrum and encode local geometry features. This is evident by the similar colors (red) assigned to the convex regions which differ from the concave regions (blue). Some noise from surface rendering is also evident, for example, at the bottom of the right-hand lobe, particularly at $t = 5$. At time $t = 15$, some smoothing has occurred as the earlier eigenfunctions which represent more global geometry are given greater emphasis, shown by the reduced noise effects, with the regions of different curvature still evident. At time $t = 300$, the convex features on the left and right are colored red/yellow; however, there is now a blue band across the middle of the molecule which also includes a convex region.

Figure 10 illustrates local geometry descriptors calculated using the WKS evaluated over 100 intervals. Figure 10(a) shows the second evaluation which encodes global curvature. At higher evaluations, Figure 10(b), more local features are encoded until the final evaluation in Figure 10(c) appears to encode noise and general artifacts of the mesh generation process.

Parametrizing the Number of Eigenpairs. The computational cost of performing the eigendecomposition increases with the number of eigenpairs computed. This effect is illustrated in Figure 11(a) for a single molecule of mesh size 5985 vertices at increasing numbers of eigenpairs and shows an increasing, nonlinear relationship between k and time. The computation time for 100 eigenpairs is around 1 s and increases to 5 s for 300 eigenpairs and 16 s for 500 eigenpairs. Figure 11(b) shows the time taken to compute 300 eigenpairs for a random sample of 250 molecules with mesh sizes varying from around 3000 to 12,000. In addition to depending on the

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

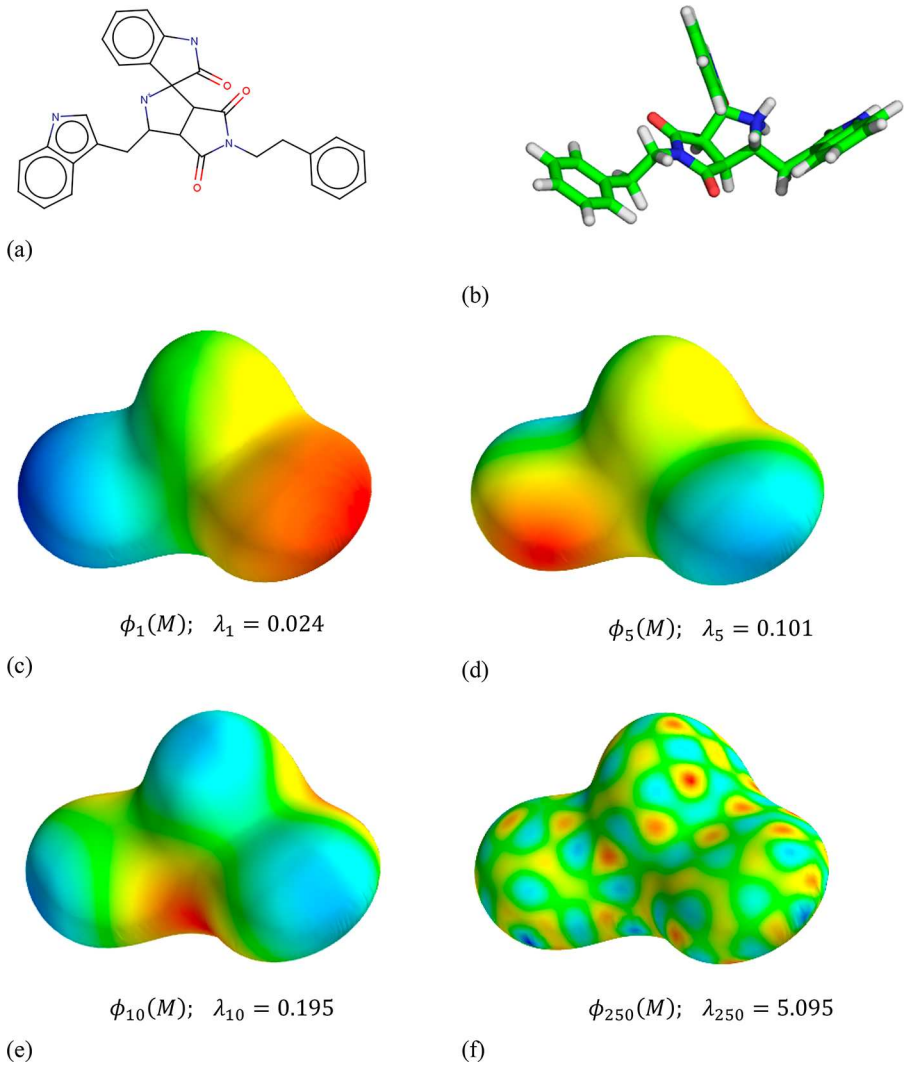


Figure 8. Molecule from the Zinc database (ZINC12885854) with eigenfunctions plotted over the surface along with their corresponding eigenvalues.

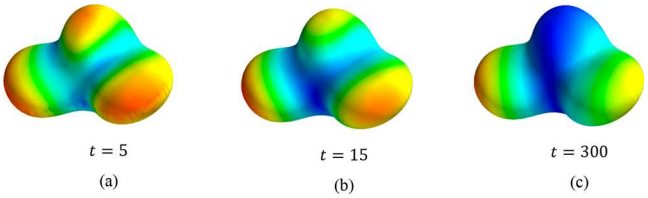


Figure 9. HKS plotted for all vertices at three different time points.

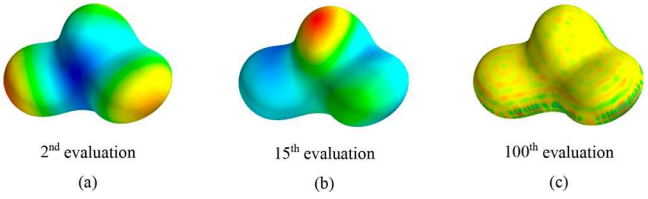


Figure 10. WKS plotted for all vertices for three different evaluations.

value of k, the time to compute the eigenpairs increases approximately linearly with the size of the mesh. The number of vertices for the sample molecule is at the lower end of the distribution of mesh sizes and therefore computation times; the longest computation time is around 15 s for a mesh with approximately 12,000 vertices. Although the eigendecomposition computation is a one-time cost (since it is required when computing the descriptors rather than during virtual screening run time), it is still beneficial to keep this to a minimum. Hence, the first parametrization experiment was to determine the optimum value of k.

Spectral geometry descriptors were calculated for the following numbers of eigenpairs, k: 5, 10, 20, 30, 40, 50, 100, 200, 300, 400, and used in virtual screening experiments on the 10 targets with each active compound used as query, in turn. In all cases, the local geometry descriptors were calculated using the WKS (with number of evaluations set at $D = 100$), and the local descriptors were aggregated to a global descriptor using the covariance matrix. The covariance matrix was chosen as the global descriptor as no parametrization is required, unlike for the bag-of-features global descriptors, the parametrization of which is described later. Virtual screening performance was evaluated using area under the curve (AUC), and BEDROC ($\alpha = 20$), and the results averaged over all targets are presented in Figure 12 with 95% confidence intervals computed using 10,000 bootstrap iterations. On average, the best performance measured using BEDROC was obtained for $k = 20$, and no performance gain was evident for

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

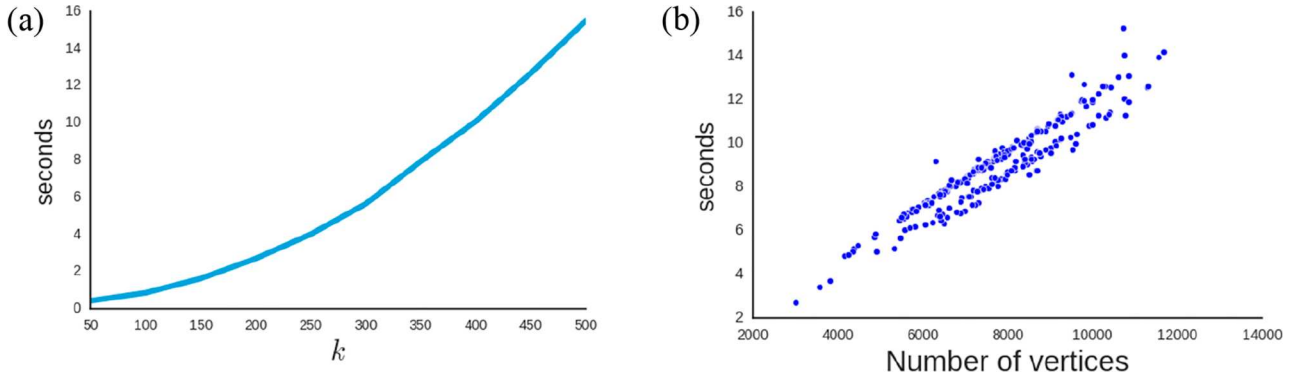


Figure 11. (a) Variation in computational time to compute increasing number of eigenpairs, k, for a sample molecule with mesh size 5985. (b) Variation in computational time to compute 300 eigenpairs for a sample of 250 molecules with varying numbers of mesh vertices.

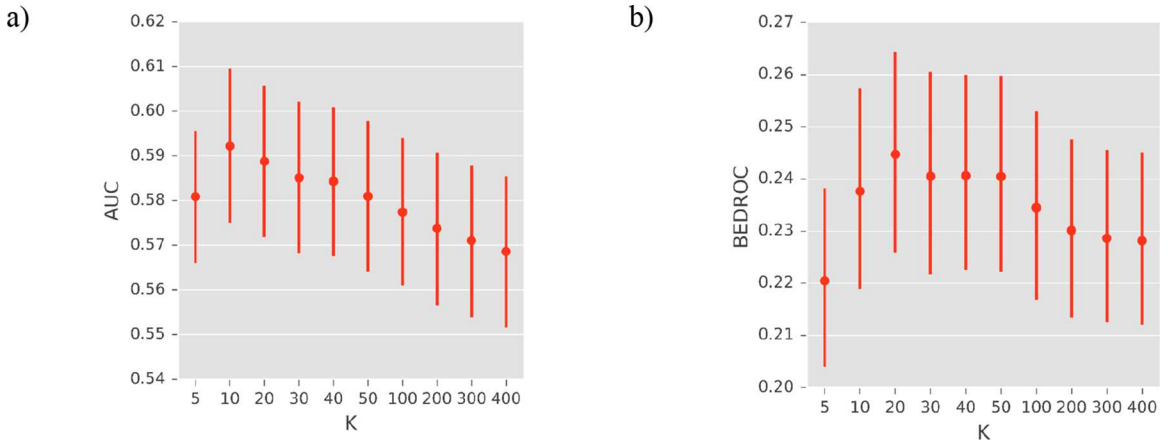


Figure 12. Virtual screening performance over 10 targets for different numbers of eigenpairs.

using higher numbers of eigenpairs. The AUC shows better average performance at $k = 10$. However, the early enrichment measures are better indicators of virtual screening performance; therefore, all subsequent experiments are based on $k = 20$ eigenpairs.

Parametrizing the Local Geometry Descriptors. The effect of varying the number of descriptors and the sample time points, t, for the HKS were investigated. Six sets of parameters were chosen as shown in Table 1. The local descriptors were aggregated to global descriptors using the covariance matrix method. The first set of parameters, T_0, consists of six time points that were found to be the optimum for deformable human shape data.[31] However, visual inspection of the HKS values suggested that there was little or no variation at higher time points, indicating that molecular shape has little local geometry variation in comparison to more complex deformable shapes such as human models. Therefore, smaller time ranges were also selected. Times T_1, T_2, and T_3 also have six elements but sample the time space up to 2500, 1500, and 700, respectively. To investigate whether performance would be substantially improved by sampling more data points, ranges up to 700 and 1000 were sampled at 1000 equally spaced points in time samples T_4 and T_5. In all cases, the time points were handled as real numbers.

Table 1. Number of Descriptors, D, and Time Points Used To Calculate Local Geometry Descriptors for HKS

Parameter	Values of t	Number of descriptors, D
T_0	[1024, 1351, 1783, 2353, 3104, 4096]	6
T_1	[50, 100, 500, 1000, 2500]	6
T_2	[20, 70, 300, 500, 900, 1500]	6
T_3	[20, 70, 150, 275, 400, 700]	6
T_4	In range 1..700	1000
T_5	In range 1..1000	1000

The results are shown in Figure 13 where it can be seen that the best parameters from the literature, T_0, perform the worst, showing that molecular shape has its own domain-specific features. Performance increases as the maximum value of t decreases. Overall, the higher dimension descriptors perform marginally better, and time samples T_3 (6 dimensions) and T_4 (1000 dimensions) were selected for subsequent experiments.

The effect of varying the number of evaluations for the WKS was investigated. Results are shown in Figure 14 for the following numbers of evaluations: 32, 64, 100, 500, 750, 1000. On average, $D = 32$ performs worse than all other parameters in terms of both the AUC and BEDROC statistics, and early enrichment performance improves with increasing D. However, as the covariance descriptors are of size D^2, the comparison time for each descriptor increases quadratically with D. Consequently, in order to balance the performance improvement of higher evaluations with comparison efficiency of lower evaluations, the parameters chosen for the virtual screening of the full DUD-E data set are $D = 64$ and $D = 100$.

Comparing the results obtained using the HKS (Figure 13) with those obtained using the WKS (Figure 14) shows that the

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

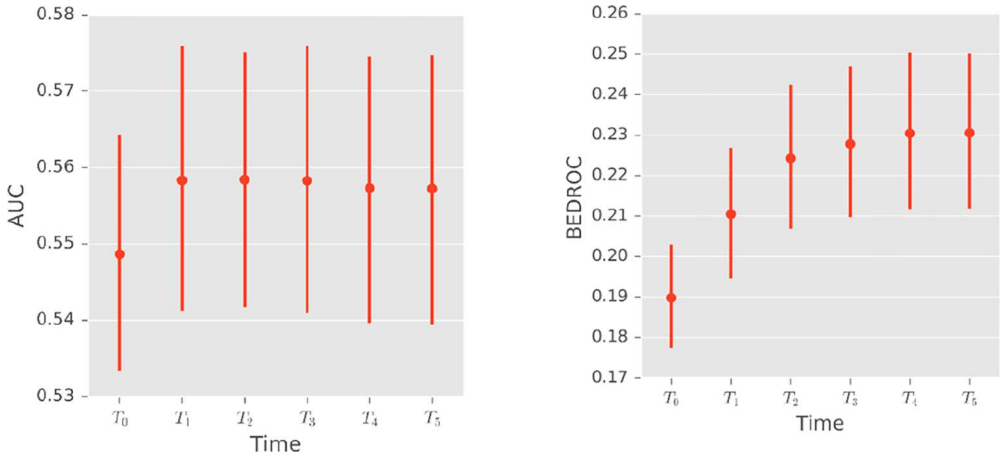


Figure 13. Virtual screening results evaluated on three targets using the HKS with the parameters shown in Table 1.

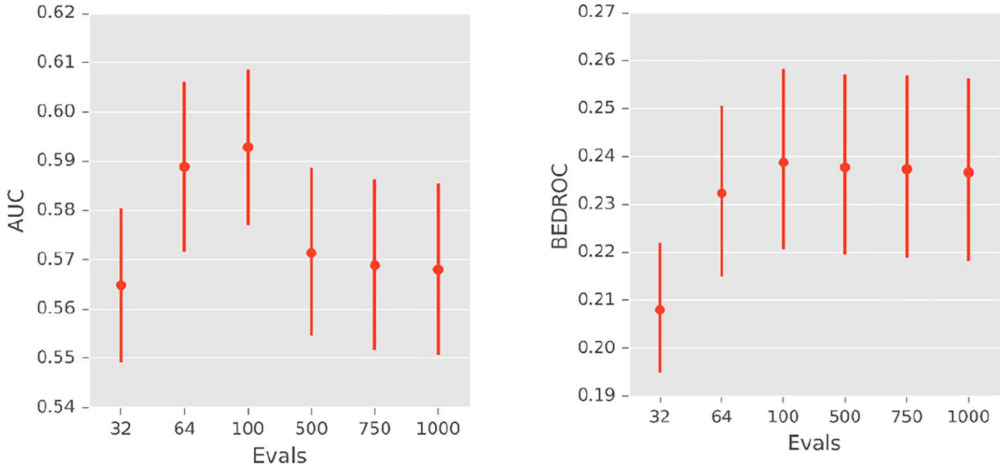


Figure 14. Virtual screening results evaluated on 10 targets using the WKS.

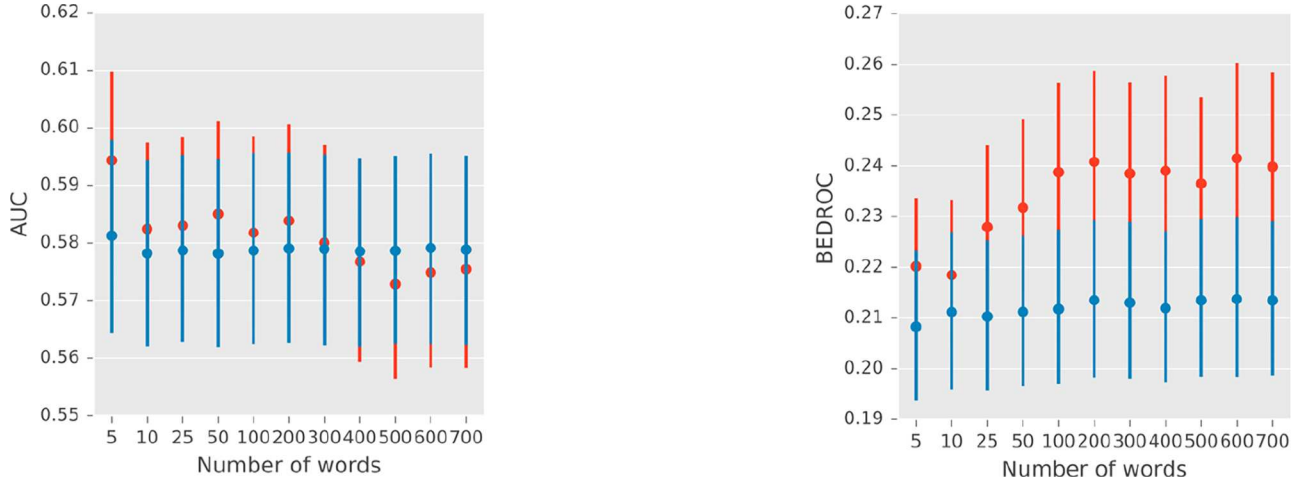


Figure 15. Effect of varying the number of codewords in the codebook on the HQ (red) and SQ (blue) encoding methods for bag-of-features.

WKS signature outperforms HKS both in terms of the AUC and early enrichment. Thus, only the WKS descriptors were considered for virtual screening on the full DUD-E data set.

Parametrizing the Global Geometry Descriptors. The two different methods for generating global geometry descriptors were investigated: the covariance matrix and the bag-of-features methods. As stated in the Methods, the covariance matrix does not require any parametrization with the size of the resulting descriptor being determined by the size of the local descriptors. Therefore, the two different covariance methods that are investigated below are Covariance 64 (corresponding to WKS with $D = 64$) and Covariance 100 (corresponding to WKS with $D = 100$).

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

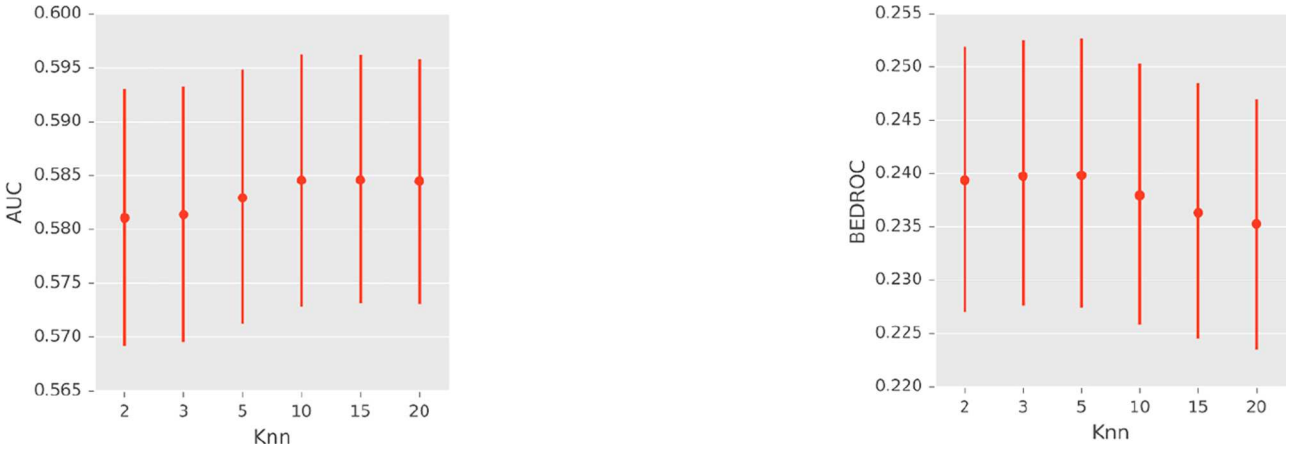


Figure 16. Effect of varying the number of nearest neighbors in the KNN encoding method for bag-of-features.

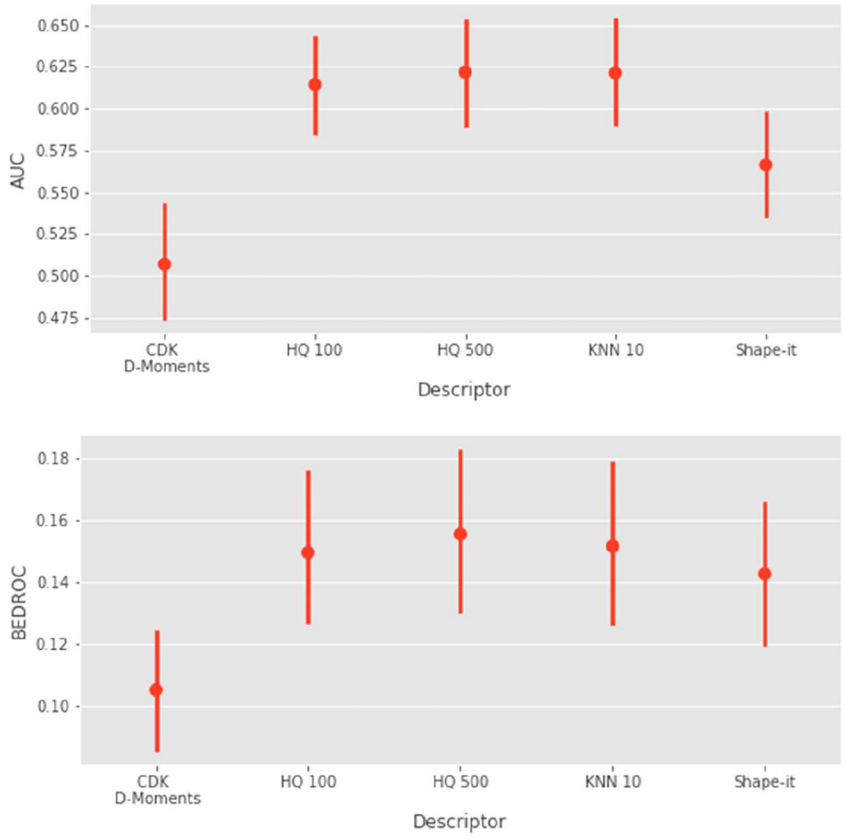


Figure 17. Average virtual screening results for the DUD-E data sets using the crystal ligand as query.

The bag-of-features approach has a number of steps which require parametrization and which were investigated. These are the number of molecules required to optimize the codebook, number of codewords in the codebook, and the encoding method. Local geometry descriptor space was sampled by randomly selecting 2000 molecules from the DUD-E data set. The WKS, D = 100, descriptors were computed for each molecule and collected in one large matrix which was used as the input for a mini-batch k-means algorithm. The centroids of the k-means clusters were then used as the codewords.

The effects of varying the number of codewords in the codebook are shown in Figure 15 for the HQ (Hard Vector Quantization) and SQ (Soft Vector Quantization) encoding methods. They show that the HQ encoding method performs best on average for the early recognition problem. The HQ histograms with 100 and 500 codewords, respectively, were taken forward to the final screen. Increasing the number of codewords in the descriptor increases the number of features in the descriptor space, which is demonstrated by the average improvement in early recognition. However, as these features are likely to be near each other in the descriptor space, there is likely to be a large amount of overlap in the SQ encoding, which would explain why the SQ encoding does not improve.

The effect of varying the number of nearest neighbors for the KNN encoding method is shown in Figure 16. The KNN descriptors were computed using the 100 word codebook that was used for the HQ 100 encoding. The results show that the virtual screening results are insensitive to the number of

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

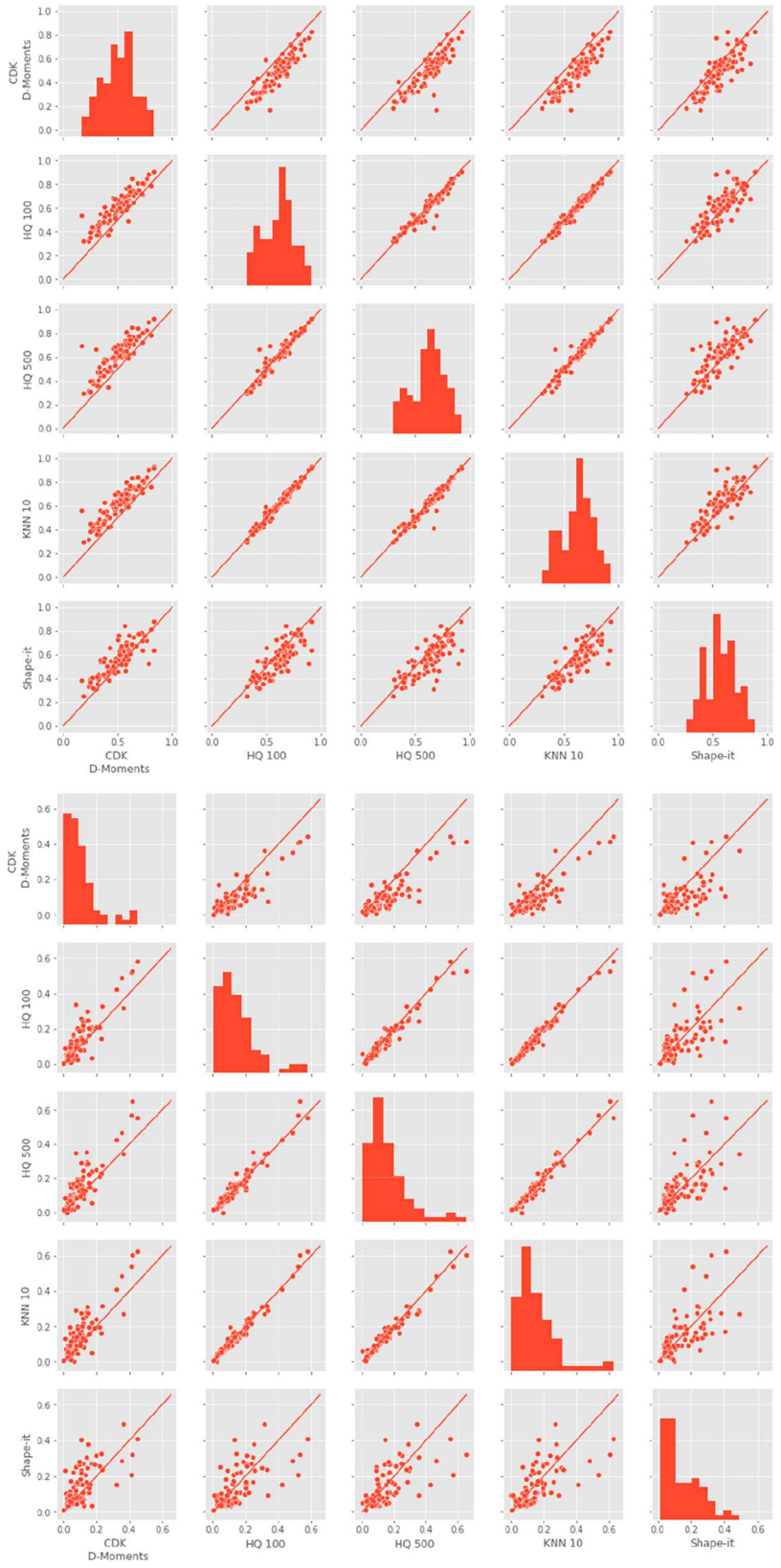


Figure 18. Pairwise virtual screening results for the DUD-E data sets using the crystal ligand as query: AUC, top; BEDROC, bottom. Each of the smaller plots compares two different methods with the individual data points showing the performance for a given target for each method. When the majority of plots are above the diagonal line, this indicates that the method on the *y*-axis outperforms that on the *x*-axis for the majority of the targets and vice versa. The histograms show the distribution of scores across all targets for a given method.

nearest neighbors and the performance values are worse than the HQ encoding; however, they still perform better than the SQ encoding. Therefore, a KNN encoding histogram $K = 10$ was taken forward to the full screen.

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

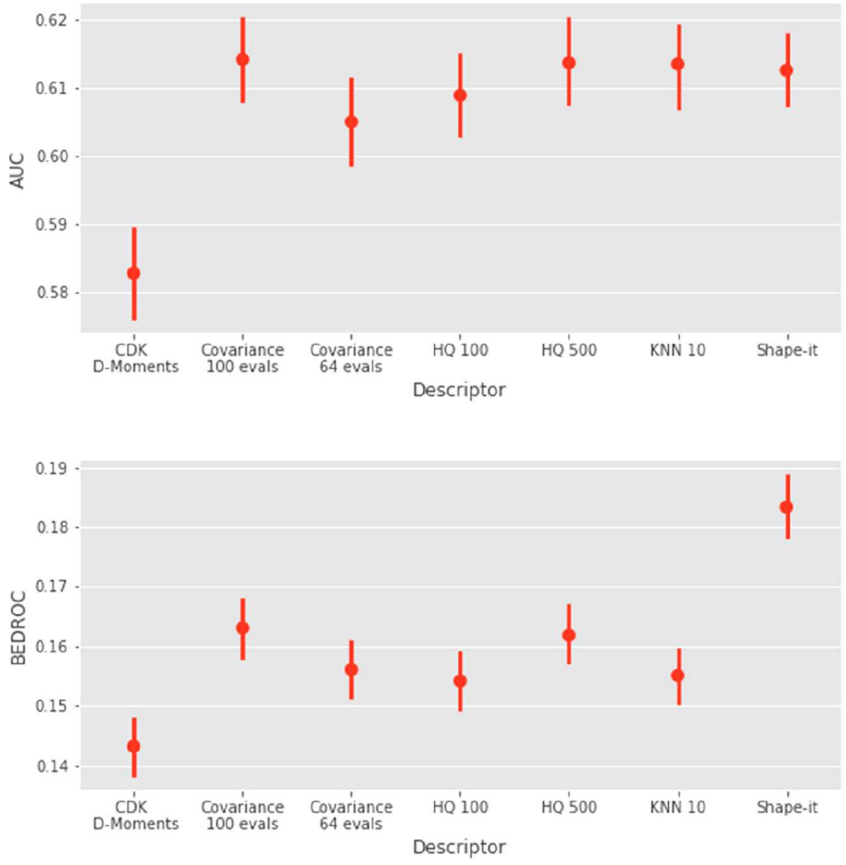


Figure 19. Average virtual screening results for the DUD-E data sets based on 20 actives per target.

Virtual Screening Experiments. Following the parametrization experiments, the spectral geometry methods were applied to large scale virtual screening experiments on the full DUD-E data set. The parameters for calculating the local geometry descriptors were fixed as $k = 20$ and WKS with $D = 100$, and the different methods for generating global geometry descriptors were compared. The latter including the covariance matrix method (Covariance 64 and Covariance 100) and three bag-of-features methods: HQ 100, HQ 500, and KNN 10. The spectral geometry methods were compared with established alignment-based and alignment-free methods, namely, Shape-it and CDK D-Moments, respectively. Shape-it is an open source alignment-based method that is similar to ROCS in that it is based on an atom-centered Gaussian representation of molecule shape (Shape-IT version v1.0.1 compiled against RDKit). The CDK D-Moments method is open source and similar to the UFSR method (CDK D-Moments in KNIME).

The first set of results shown in Figures 17 and 18 are based on using the crystal ligand as reference for each of the DUD-E targets. The virtual screening performance was evaluated statistically by comparing mean performance with a pairwise t test using posthoc Tukey HSD adjustments. Three of the bag-of-features spectral geometry methods, namely, KNN 10, HQ 100, and HQ 500, are compared with CDK D-Moments and Shape-IT. Figures 19 and 20 show the results of virtual screening experiments based on 20 actives chosen at random for each of the targets in DUD-E. Here, the covariance spectral geometry descriptors are also included for comparison with the bag-of-features spectral geometry methods. Full tabular results are presented in the Supporting Information.

DISCUSSION

Both the crystal ligand experiment (Figure 17) and the larger scale experiment based on multiple active compounds (Figure 19) indicate that there is little difference between the spectral geometry methods based on the performance measures (with the exception of Covariance 64 which shows reduced performance in the larger experiment). This finding is consistent for both the AUC, which considers the distribution of actives throughout the entire ranked lists, and the BEDROC statistic, which considers early enrichment and is arguably more relevant for virtual screening. As might be expected, the HQ 500 and HQ 100 descriptors are highly correlated, as shown by the pairwise plots (Figures 18 and 20); however, HQ 500 (consisting of a vector of 500 values) outperforms HQ 100 (a vector of 100 values) by a marginal amount indicating that the larger number of dimensions offers increased discriminatory power. The KNN 10 method is similar in performance to HQ 500, although it shows a decreased AUC for the larger virtual screening experiment. Although the performances of the Covariance 100 and HQ 500 descriptors are similar, the HQ 500 descriptor is significantly more compact than Covariance 100, which is a 100×100 matrix; thus, the similarity comparisons based on HQ 500 are significantly faster.

Comparing the spectral geometry descriptors with CDK D-Moments, which represents a more established alignment-free descriptor, the spectral geometry methods show significantly improved performance in both experiments and considering both early enrichment and AUC. For example, the pairwise comparison of the means shown in Figure 17 using Tukey HSD adjustments confirmed that all other methods are statistically significantly better than CDK D-Moments for

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

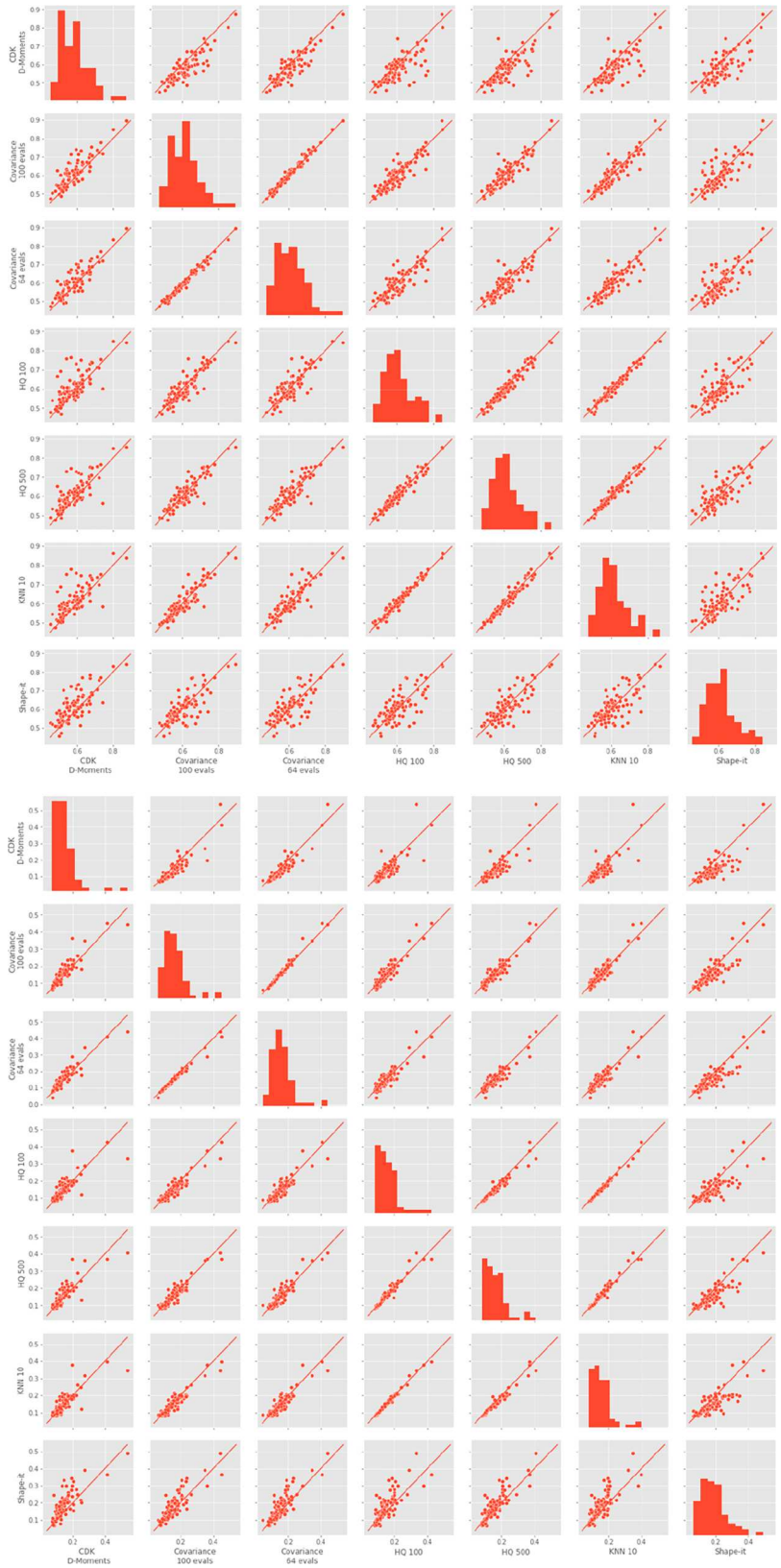


Figure 20. Pairwise virtual screening results for the DUD-E data sets based on 20 actives per target: AUC, top; BEDROC, bottom. See Figure 18 for an explanation.

both AUC (at an α = 0.001 significance level) and BEDROC (at an α = 0.005 significance level). Furthermore, Figures 18 and 20 show pairwise comparisons for each of the target classes in DUD-E where it can be seen that the spectral geometry descriptors generally outperform the CDK D-Moments in almost all cases. This finding supports our hypothesis in the Introduction that the spectral geometry descriptors, which capture properties of shape over the surface of a molecule,

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

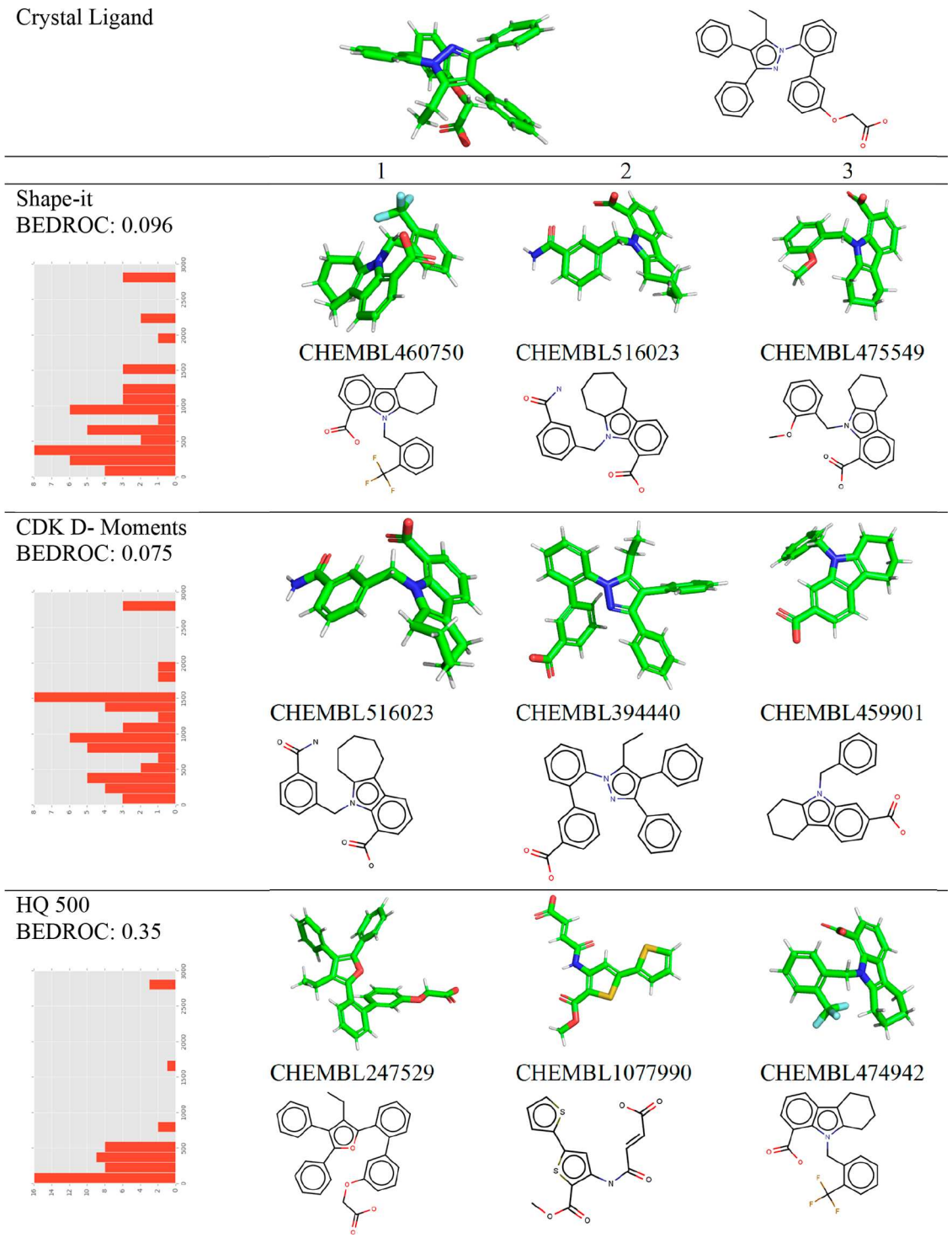


Figure 21. Top three actives retrieved for the crystal ligand in FABP4 for Shape-it, CDK D-Moments, and HQ 500, respectively. The histograms show the distribution of actives for each method.

provide a richer description of shape than descriptors that are based on interatomic distances. Although our best performing descriptors are larger than the CDK D-Moments which are based on 12 distance measures, this finding also holds true for the more compact spectral geometry descriptors, for example, HQ 100.

The spectral geometry descriptors also show good performance against the alignment-based method Shape-it which is based on the same principles as the industry-standard ROCS method. In the crystal ligand experiment, the HQ 500 descriptors perform better on average than Shape-it on both AUC and BEDROC and in the larger virtual screening experiment, HQ 500 is comparable to Shape-it in terms of AUC, although it is intermediate between CDK D-Moments and Shape-it in terms of early enrichment. Overall, however, this represents impressive performance given that the HQ 500 descriptors are invariant to alignment and are therefore considerably faster to compare than using a method such as

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

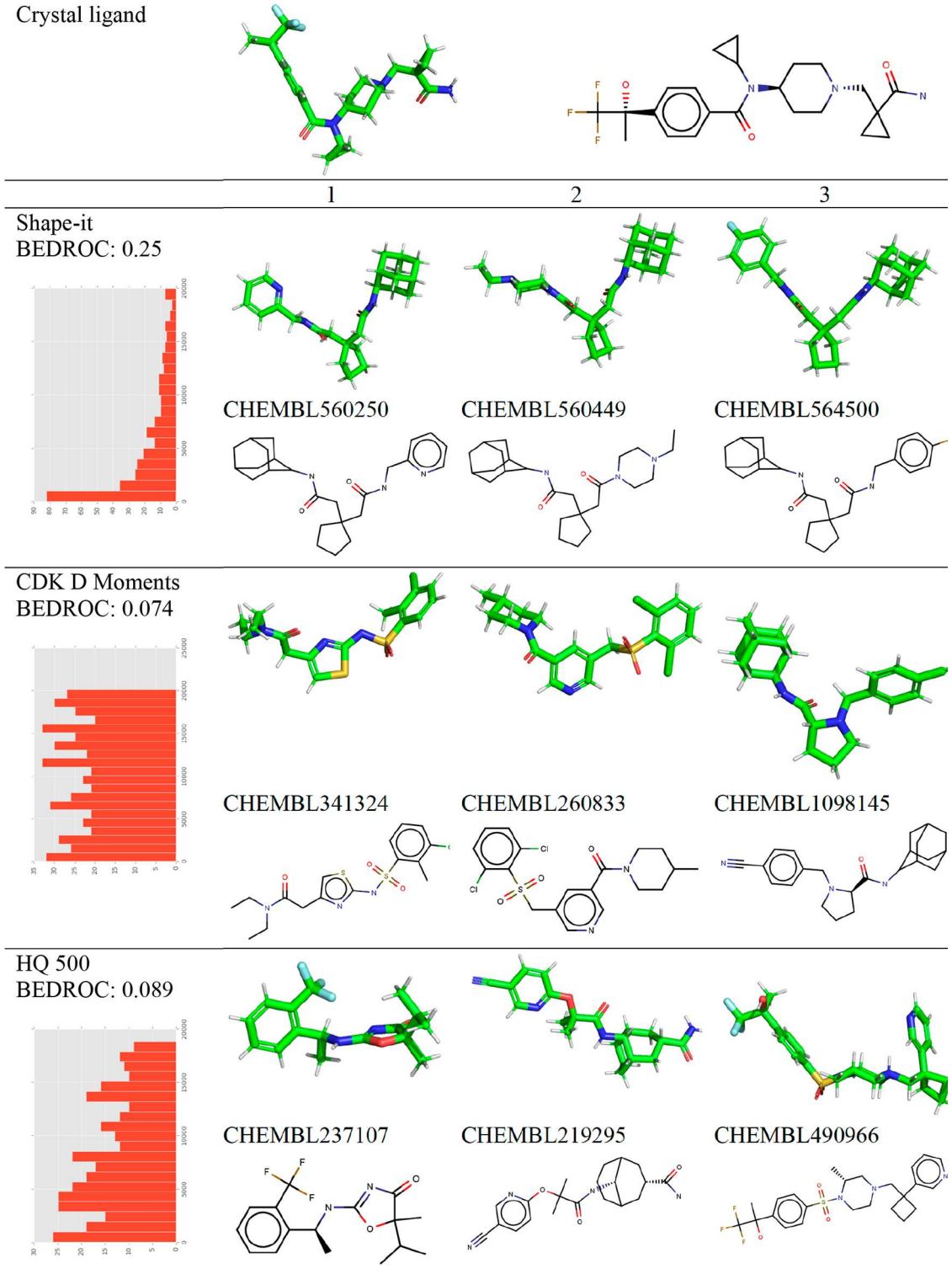


Figure 22. Top three retrieved actives against the crystal ligand in DHI1 for Shape-it, CDK D-Moments, Covariance 100, and HQ 500, respectively. The histograms show the distribution of actives for each method.

Shape-it that requires the optimum alignment to be found prior to calculating similarity.

Interestingly, optimal virtual screening performance for the spectral geometry global shape descriptors was achieved with relatively few eigenvalues (k = 20) with higher numbers of eigenvalues leading to reduced performance. As the lower eigenvalues correspond to the more global features of 3D molecular shape, this suggests that the more global features of a molecule's intrinsic geometry are the most important for discriminating between actives and inactives. In general, small molecules occupy a small subspace of all possible 3D shapes and minor variations in the 3D shapes of molecules are likely to come from predominantly global features resulting from atoms occupying volumes of space, rather than from minor local variations on the surface such as creases, which are more likely the result of the mesh generation with no chemical

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

meaning. Consequently, molecule shape can be described with a relative small number of eigenvalues. An alternative explanation is that the decrease in performance at higher numbers of eigenvalues suggests that the small variations in shape may be lost in noise from the surface generation process. The handling of noise and signal amplifications is at the heart of finding the optimal parameters for virtual screening. A similar case was observed in the parametrization of the WKS (Figure 14). The AUC performance declined markedly, and the BEDROC performance reached a plateau after $D = 100$ evaluations. As higher numbers of evaluations increase the granularity, and therefore the capacity to encode local geometry features, this might also suggest the dominance of surface noise in the descriptor. The balance of noise and signal for the covariance descriptors is managed at the level of the parameters of the local geometry descriptor. For the bag-of-features descriptors, this is handled through the number of codewords and the encoding method with HQ encoding and a codebook of 500 words performing best for virtual screening. A further interpretation of the codebook is that it captures the dominant intrinsic geometry features in molecular shape space. The relatively few number of necessary codewords, which is in the 100s rather than 1000s, also suggests that there is a relatively small number of geometric features in molecular shape.

While the virtual screening performance statistics give an overview of the retrieval of actives over a large number of targets, it does not show the types of actives that are being retrieved. Figures 21 and 22 show the shape properties of the top retrieved actives against the crystal ligands taken from two targets, FABP4 and DHI1, respectively. The FABP4 query in Figure 21 is an example where the spectral geometry descriptors performed better than the baseline methods. The top retrieved actives from Shape-it are structurally very similar both to themselves and the query: They all share the same fused ring system, and they differ only in the substituents on the peripheral phenyl ring. It would therefore be expected that a method that prioritizes volume overlap would identify these are being highly similar. On the other hand, although the top active retrieved by CDK D-Moments (CHEMBL516023) is also in the Shape-it top three, the next two hits are structurally more distinct. Nevertheless, all three exhibit similar global geometry properties. The top three actives retrieved by HQ 500 are more structural diversity compared to those retrieved by Shape-it and CDK D-Moments. The HQ 500 descriptor represents shape as a vector of local surface geometry features and, therefore, gives greater emphasis to local geometry compared to either Shape-it and CDK D-Moments, which can explain the greater diversity of the hits in terms of their 2D skeletons. HQ 500 is the top performing method with respect to the BEDROC statistic for this example, suggesting that local 3D shape features can be more important for 3D similarity than global volume overlap in some cases.

Figure 22 shows the top actives retrieved for the crystal ligand in the DHI1 data set. This is an example of a query molecule where the baseline methods performed significantly better than the spectral geometry methods. The best performing method, Shape-it, returns actives that are very similar in terms of the 3D conformation of the query. Again, there is very little structural diversity in the best performing actives. The top actives for CDK D-Moments retain the parabola-like global shape of the crystal ligand and the Shape-it hits, while providing more structural diversity. In contrast, the top performing actives for the HQ 500 method do not retain this global structural form. This suggests that the relative frequency of surface features allows a recognition of common local shape features that is independent of the rigid pose. Figure 22 suggests that the baseline methods perform best when all the actives have a common, dominant global shape structure, which is likely to be a dominant scaffold.

Overall, the examples suggest that the baseline methods (Shape-it and CDK D-Moments) perform best when the actives have the same pose or share a scaffold that dominates the global shape features. On the other hand, the spectral geometry descriptors retrieve a more diverse set of structures with common local shape features on the surface that are independent of global shape. The conformations in an ensemble are merely a (sometimes quite coarse) sample of a continuum that the molecule can adopt, so a descriptor that is tolerant of small changes of global shape is a better representation of the physical reality. Moreover, it suggests that spectral geometry descriptors encode different geometry features to those of the baseline methods and offer a complementary approach for virtual screening.

CONCLUSIONS

We have described a framework for applying spectral geometry to the problem of molecular shape comparison for 3D virtual screening. We have used the framework to develop a rich yet compact descriptor of molecular shape that is alignment independent. When compared to established 3D shape comparison methods on large scale virtual screening experiments on the DUD-E data set, the spectral geometry descriptors outperform the alignment-free CDK D-Moments, an open source implementation of UFSR. Furthermore, our descriptors show comparable performance to a Gaussian overlap method, namely, Shape-it, when measured using AUC. The alignment-based method gave better performance on early enrichment when averaged over a range of active molecules; however, the spectral geometry methods are considerably faster in operation since they involve vector comparisons only and do not require that an optimum alignment is found prior to calculating similarity. This decrease in computational time together with the compact nature of the spectral geometry descriptors can lead to a significant increase in the throughput of 3D virtual screening experiments. Given the reduced performance with respect to early enrichment, one possibility would be to use the spectral geometry method as a prescreen prior to using a more computationally demanding approach. Thus, spectral geometry could sit at the top of a shape-based virtual screening cascade providing a fast way of reducing a very large data set to a more manageable size for subsequent refinements using a more accurate alignment-based method.

As mentioned in the Introduction, spectral geometry has become the method of choice in computer vision for comparing deformable shapes such as people and animals, and there has been substantial development of the basic approach in this field, for example, to deal with issues such as color over the surface, partial matching, and mappings between different shapes. Our work presented here has focused on adapting the basic approach to the domain of small molecules. We describe this as a framework to emphasize the different steps required to produce the descriptors, each of which has required parametrization. We have optimized these here for the comparison of rigid shapes in large scale virtual screening

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116

experiments on the DUD-E data set. The rationale for this first publication has been to demonstrate the potential of the basic approach through comparing rigid molecules on shape only; thus, this first approach has been blind to chemistry. Having demonstrated significant potential on this more restricted problem, we are currently extending the basic approach to include properties encoded on the surface of a molecule and will describe this work in a future publication.

As well as being invariant to alignment, we see one of the key advantages of spectral geometry as the invariance of the descriptors to isometric deformations, which is the key property that has been exploited in the computer vision field. In the context of people, it is usually desirable to identify the same person whatever pose they have adopted. However, the relationship between conformation and shape in the world of small molecules is more complex given that a receptor recognizes a particular shape. In general, there is a conflict between the desire to identify a single molecule from its different conformations and to distinguish between the different conformations of a given molecule that may have different activity. Typically, a conformational ensemble is created by generating, either systematically or stochastically, a large number of conformations. Similar conformations are discarded, usually by RMSE of atom position, and the remaining ordered by ascending energy. The lowest energy N within an energy window are selected as the ensemble. The strength of the spectral geometry approach is that it encodes flexibility in a way that is formally defined and well understood which could provide a rational basis for handling conformational flexibility. The spectral geometry approach could allow the conformation selection to be based on a more relevant metric, the shape, allowing for a wider representation of shapes in the same number of conformers or a smaller number of conformers covering the same shape space. Either should improve virtual screening performance, the former by improving the quality of the ensembles, the latter by increasing the number of molecules that can be handled. A more rational approach to handling conformational flexibility compared to the current approaches of using energy or distance thresholds could have significance for a wide range of applications based on 3D structure, for example, pharmacophore mapping, 3D QSAR, and molecular dynamics simulations.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00676.

Formal description of spectral geometry methods. Full virtual screening results on the DUD-E data sets. The spectral geometry code is available via GitHub at https://github.com/SheffieldChemoinformatics/molsg. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: v.gillet@sheffield.ac.uk.

ORCID

Valerie J. Gillet: 0000-0002-8403-3111

Present Address

D.A. Cosgrove: CozChemIx Limited, 37 Coniston Way, Macclesfield, Cheshire SK11 7XR, United Kingdom.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the Biotechnology and Biological Sciences Research Council (BBSRC) and AstraZeneca for funding M.P.S. through a BBSRC studentship. We also thank Prof. Barry Pickup and Prof. Peter Willett for useful discussions during the course of this project.

■ REFERENCES

(1) Finn, P. W.; Morris, G. M. Shape-Based Similarity Searching in Chemical Databases. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3*, 226−241.

(2) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular Shape and Medicinal Chemistry: A Perspective. *J. Med. Chem.* **2010**, *53*, 3862−3886.

(3) Shin, W.-H.; Zhu, X.; Bures, M. G.; Kihara, D. Three-Dimensional Compound Comparison Methods and Their Application in Drug Discovery. *Molecules* **2015**, *20*, 12841−12862.

(4) Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein−Protein Interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.

(5) Pérez-Nueno, V. I.; Venkatraman, V.; Mavridis, L.; Clark, T.; Ritchie, D. W. Using Spherical Harmonic Surface Property Representations for Ligand-Based Virtual Screening. *Mol. Inf.* **2011**, *30*, 151−159.

(6) Cheeseright, T. J.; Mackey, M. D.; Melville, J. L.; Vinter, J. G. Fieldscreen: Virtual Screening Using Molecular Fields. Application to the Dud Data Set. *J. Chem. Inf. Model.* **2008**, *48*, 2108−2117.

(7) Sun, J.; Jeliazkova, N.; Chupakhin, V.; Golib-Dzib, J.-F.; Engkvist, O.; Carlsson, L.; Wegner, J.; Ceulemans, H.; Georgiev, I.; Jeliazkov, V.; Kochev, N.; Ashby, T. J.; Chen, H. Excape-Db: An Integrated Large Scale Dataset Facilitating Big Data Analysis in Chemogenomics. *J. Cheminf.* **2017**, *9*, 17.

(8) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database Gdb-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.

(9) Ballester, P. J. Ultrafast Shape Recognition: Method and Applications. *Future Med. Chem.* **2011**, *3*, 65−78.

(10) Schuffenhauer, A. Computational Methods for Scaffold Hopping. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 842−867.

(11) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martinez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.* **2012**, *52*, 867−881.

(12) Hawkins, P. C. D. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.* **2017**, *57*, 1747−1756.

(13) Reuter, M.; Wolter, F. E.; Peinecke, N. Laplace-Beltrami Spectra as 'Shape-DNA' of Surfaces and Solids. *Comput. Aided Des.* **2006**, *38*, 342−366.

(14) Sun, J.; Ovsjanikov, M.; Guibas, L. A Concise and Provably Informative Multi-Scale Signature Based on Heat Diffusion. *Computer Graphics Forum* **2009**, *28*, 1383−1392.

(15) Aubry, M.; Schlickewei, U.; Cremers, D. The Wave Kernel Signature: A Quantum Mechanical Approach to Shape Analysis. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Barcelona, Spain, 2011; 2011; pp 1626−1633.

(16) Bronstein, A. M.; Bronstein, M. M.; Guibas, L. J.; Ovsjanikov, M. Shape Google: Geometric Words and Expressions for Invariant Shape Retrieval. *ACM Transactions on Graphics* **2011**, *30* (1), 1−20.

(17) Li, C.; Ben Hamza, A. Spatially Aggregating Spectral Descriptors for Nonrigid 3D Shape Retrieval: A Comparative Survey. *Multimedia Syst* **2014**, *20*, 253−281.

(18) Lian, Z.; Godil, A.; Bustos, B.; Daoudi, M.; Hermans, J.; Kawamura, S.; Kurita, Y.; Lavoué, G.; Van Nguyen, H.; Ohbuchi, R.;

Ohkita, Y.; Ohishi, Y.; Porikli, F.; Reuter, M.; Sipiran, I.; Smeets, D.; Suetens, P.; Tabia, H.; Vandermeulen, D. A Comparison of Methods for Non-Rigid 3D Shape Retrieval. *Pattern Recognit* **2013**, *46*, 449−461.
(19) Liu, Y.-S.; Li, Q.; Zheng, G.-Q.; Ramani, K.; Benjamin, W. Using Diffusion Distances for Flexible Molecular Shape Comparison. *BMC Bioinf.* **2010**, *11*, 480.
(20) Axenopoulos, A.; Rafailidis, D.; Papadopoulos, G.; Houstis, E. N.; Daras, P. Similarity Search of Flexible 3D Molecules Combining Local and Global Shape Descriptors. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2016**, *13*, 954−970.
(21) Grinspun, E.; Desbrun, M.; Polthier, K.; Schröder, P.; Stern, A. *Discrete Differential Geometry: An Applied Introduction*; ACM SIGGRAPH Course Notes; ACM, 2006; p 7.
(22) Botsch, M.; Pauly, M.; Rossl, C.; Bischoff, S.; Kobbelt, L. *Geometric Modeling Based on Triangle Meshes*; ACM SIGGRAPH 2006 Courses; ACM, 2006.
(23) Pinkall, U.; Polthier, K. Computing Discrete Minimal Surfaces and Their Conjugates. *Experimental Mathematics* **1993**, *2*, 15−36.
(24) Tuzel, O.; Porikli, F.; Meer, P. Region Covariance: A Fast Descriptor for Detection and Classification. In *Computer Vision − Eccv 2006*; Leonardis, A.; Bischof, H.; Pinz, A., Eds.; Springer Berlin, Heidelberg, 2006; Vol. *3952*, pp 589−600.
(25) Masci, J.; Boscaini, D.; Bronstein, M. M.; Vandergheynst, P., Shapenet: Convolutional Neural Networks on Non-Euclidean Manifolds. *arXiv:1501.06297* **2015**.
(26) Salton, G.; Wong, A.; Yang, C. S. A Vector Space Model for Automatic Indexing. *Commun. ACM* **1975**, *18*, 613−620.
(27) Sculley, D. *Web-Scale K-Means Clustering*; ACM Press, 2010; p 1177.
(28) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (Dud-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.
(29) Chen, M. X.; Lu, B. Z. Tmsmesh: A Robust Method for Molecular Surface Mesh Generation Using a Trace Technique. *J. Chem. Theory Comput.* **2011**, *7*, 203−212.
(30) Chen, M. X.; Tu, B.; Lu, B. Z. Triangulated Manifold Meshing Method Preserving Molecular Surface Topology. *J. Mol. Graphics Modell.* **2012**, *38*, 411−418.
(31) Ovsjanikov, M.; Bronstein, A. M.; Bronstein, M. M.; Guibas, L. J. Shape Google: A Computer Vision Approach to Isometry Invariant Shape Retrieval. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, IEEE: Kyoto, Japan 2009; pp 320−327.

DOI: 10.1021/acs.jcim.8b00676
J. Chem. Inf. Model. 2019, 59, 98−116