White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/247888/

Version: Published Version

Article:
Middleton, J., Ghiandoni, G.M., Packer, M.J. et al. (2026) Graph Laplacian spectral descriptors for alignment-free molecular shape and electrostatic comparison. Journal of Cheminformatics. ISSN: 1758-2946

https://doi.org/10.1186/s13321-026-01310-6

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:
https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

RESEARCH | Article in Press

Graph Laplacian spectral descriptors for alignment-free molecular shape and electrostatic comparison

James Middleton[1] · Gian Marco Ghiandoni[2] · Martin J. Packer[3] · Mengdie Zhuang[1] · Valerie J. Gillet[1]

Received: 6 June 2026 / Accepted: 10 September 2026
© The Author(s) 2026

Abstract

Shape-based virtual screening methods are widely used in drug discovery. Alignment-dependent approaches often provide high screening accuracy but are computationally expensive, whereas alignment-free methods are faster but generally less accurate. Here, we present GE-MolSG, a novel alignment-free molecular descriptor that combines shape and electrostatic information using spectral graph representations derived from the graph Laplacian. GE-MolSG outperforms the established alignment-free ElectroShape method and achieves performance comparable to ESP-Sim without requiring molecular alignment. To demonstrate the extensibility of the graph-based framework, we further explore a proof-of-concept patch-based extension termed GM-MolSG, which incorporates additional surface features. The graph-based formulation provides a flexible and extensible framework for incorporating surface-associated molecular properties into compact descriptors for rapid alignment-free similarity searching.

Scientific contribution

GE-MolSG introduces a novel alignment-free spectral graph approach for molecular comparison that combines shape and electrostatic information within compact spectral descriptors. The method outperforms the established alignment-free ElectroShape descriptor and achieves performance comparable to ESP-Sim without requiring molecular alignment, while providing a flexible basis for incorporating additional surface-associated molecular properties. As a proof-of-concept demonstration of this extensibility, we further explore a patch-based extension, GM-MolSG, incorporating local geometric and chemical surface descriptors.

Keywords Virtual screening · 3D · Shape · Graphs · Electrostatics

Introduction

Shape-based searching is widely used to screen molecular libraries and prioritise compounds for further investigation [1]. Compared with fingerprint-based methods, shape-based approaches can retrieve compounds with similar shapes despite differing chemical structures. This capability is particularly useful for scaffold hopping,

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

enabling the identification of alternative chemotypes to circumvent intellectual property constraints or replace candidate series with undesirable side effects [2].

Shape searching can be divided into alignment-dependent and alignment-free methods [3, 4]. Alignment-dependent methods involve first finding the optimum alignment of a database compound to the query and then evaluating its similarity based on, for example, volume overlap. In contrast, alignment-free methods aim to capture molecular shape in a vector-based representation that is insensitive to the orientation and position of molecules in a 3D space.

Alignment-dependent methods are generally regarded as providing state-of-the-art performance in shape-based virtual screening. For example, Rapid Overlay of Chemical Structures (ROCS) is considered the industry standard alignment-based shape method. In ROCS, atomic volumes are represented using atom-centred Gaussians [5, 6] and molecular similarity of a pair of molecules is determined as volume overlap which is calculated rapidly for a given alignment using Gaussian functions [7, 8]. The Gaussian approach forms the basis of many subsequent methods [3, 4]. Alignment-based methods are, however, more computationally demanding than alignment-free methods due to the need to find an optimum alignment.

Alignment-free methods typically involve capturing aspects of molecular shape into a vector representation. These approaches are typified by the Ultrafast Shape Recognition (USR) [9] method which is based on statistical moments (i.e., the mean, standard deviation and skewness) of distance distributions of all the atoms in a molecule to four geometrically defined reference points. The resulting descriptor is a vector of 12 continuous values which is independent of the orientation of a molecule and the similarity of two molecules can be calculated directly from their vector representations. This approach enables very rapid virtual screening; however, its performance has generally been reported as inferior to alignment-based methods [10].

While shape is fundamental to molecular recognition, it is well known that binding is driven by both shape and electrostatic complementarity between a ligand and its biological receptor. This has led to the development of a variety of approaches to capture both shape and chemical features. In ROCS, chemical ("color") similarity is accounted for by dummy atoms that represent pharmacophore features defined by the ROCS colour force field. The shape and colour similarity scores are calculated independently and combined into a single, weighted value. ESP-Sim [11] is an open-source alignment-dependent method that is conceptually similar to ROCS. It initially aligns molecules on shape and then evaluates similarity based on volume overlap and overlap of atom-centred Coulomb potentials. In FlexX [7], which is also Gaussian-based, the amplitudes of the Gaussians depend on the properties of the underlying atoms.

Alignment-free methods have also been extended to encode chemical properties alongside shape. For example, ElectroShape [12] extends the USR method to include up to two additional properties on the atoms with these being partial charge and lipophilicity. This is achieved by augmenting atomic positions with additional dimensions corresponding to the features of interest. When including partial charge information, atoms are represented as $(x, y, z, \mu q)$ where q represents the partial charge of an atom and μ is a scaling factor that modifies the units of the fourth dimension to represent charge per Angstrom. Other approaches based on spherical harmonics, Zernike polynomials and covariance matrices have also been applied to encapsulating 3D molecular shape and other meaningful molecular recognition properties in a concise vector-based representation [13–15].

Beyond cheminformatics, shape comparison is also a central task in computer vision, where spectral geometry methods have been widely applied [16]. In spectral geometry, the curvature of the surface of an object is encoded in an information rich vector-based representation. However, existing spectral geometry approaches are not naturally suited to incorporating heterogeneous physicochemical information alongside shape.

We previously reported the development of MolSG [17] which is an alignment-free shape searching method for molecule comparison based on spectral geometry. A related approach, RGMolSA, is also based on spectral methods [18]. In MolSG, the surface of a molecule is represented as a triangulated mesh to which the Finite Element Method (FEM) is applied to approximate the Laplace-Beltrami Operator (LBO), generating a spectrum (i.e., a set of eigenvalues and eigenvectors) which are then processed to produce an alignment-free descriptor.

A limitation of the FEM method for approximating the LBO is that it is based on geometric elements of the triangulated mesh and is not readily extendable to include additional dimensions such as properties on a surface.

Here we describe the extension of our shape-only method to encode both shape and electrostatics into a compact information-rich alignment-free descriptor called GE-MolSG. In this new method, the molecular surface is discretised as a graph rather than as a triangulated mesh, and the graph Laplacian is used to approximate the LBO and generate the spectrum. The edges of the graph can be weighted only on the distances between the connecting nodes so that the spectrum generated from the graph Laplacian yields a shape-only representation, comparable with that obtained using MolSG. However, because of the graph-based implementation, nodes can be further extended to include additional information. We have adopted this approach to combine shape with electrostatics to provide a novel alignment-free descriptor. To our knowledge, GE-MolSG is the first spectral geometry molecular descriptor that integrates electrostatics directly into the graph approximation of the LBO while retaining alignment-free comparison. To demonstrate the flexibility of the graph-based framework, we further explore a proof-of-concept extension termed GM-MolSG, which incorporates local surface patch representations containing additional geometric and chemical surface features.

The remainder of this manuscript is organised as follows. First, we evaluate the impact of replacing spectral geometry with spectral graph theory by comparing GE-MolSG in shape-only mode against MolSG. We then benchmark the full GE-MolSG framework across a range of experiments including comparing it against an established alignment-free method, and an alignment-based method. To demonstrate the flexibility of the graph-based formulation, we then present preliminary proof-of-concept results for GM-MolSG, which incorporates richer local surface representations through graph-based surface patches. Finally, we describe the theoretical foundations of GE-MolSG, including the graph Laplacian approximation, the graph construction procedure, and the surface patch concepts underlying GM-MolSG. The implementation of all methods presented in this work is available as open-source software.

Results and discussion

Graph discretisation of Laplacian

The impact of moving from the triangulated surface representation and associated FEM approximation of the LBO used in MolSG to the graph-based representation and graph Laplacian approximation used in GE-MolSG was investigated. To isolate the effect of the discretisation strategy, comparisons were performed using shape-only similarity searches such that only spatial coordinates were used to calculate distances and affinities in the graph-based representation. This graph-based shape-only implementation is referred to as G-MolSG, where the omission of the prefix "E" indicates that electrostatic information was not included. The number of nearest neighbours used to construct the graph representation was set to 100 based on the parameterisation experiments described in the Methods section.

An additional graph-based implementation was also investigated in which graph connectivity was restricted to the edges of the triangulated mesh used in MolSG. In this approach, affinities were calculated only between vertices connected within the triangulated mesh, resulting in a substantially sparser graph representation. This implementation is referred to as GT-MolSG.

Figure 1 shows information contained in the first three eigenfunctions for each method for an active molecule (ChEMBL ID: 1096504_0) taken from the CXCR4 dataset (see Methods for a description of the experimental data). In each case, the eigenfunction values have been plotted on the molecular surface. The spectrum of the LBO (that is, the set of eigenvalues and eigenfunctions) can be considered as a non-linear low-dimensional embedding of the molecular surface. The first eigenfunction captures the direction of most geometric variation (i.e., longest axis), with the second and third eigenfunctions representing further sets of orthogonal basis functions which capture the next largest directions of geometric variation.

https://doi.org/10.1186/s13321-026-01310-6

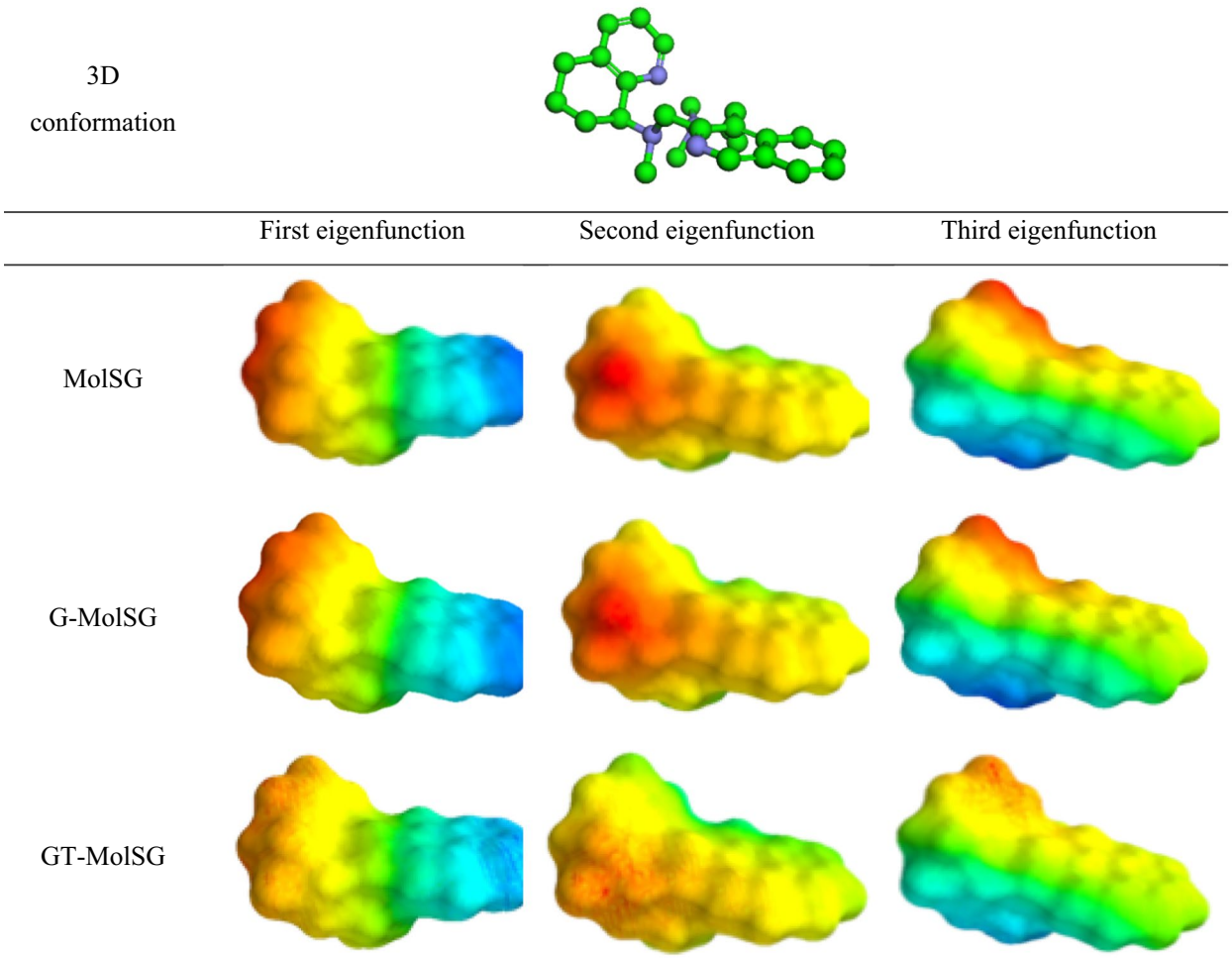


Fig. 1 Plots of the first three eigenfunctions for molecule ChEMBL ID: 1096504_0 generated using MolSG, G-MolSG and GT-MolSG. Surface colour captures the sign of each eigenfunction (red/blue for positive/negative values), the intensity of the colour describes the magnitude of the eigenfunction, with more intensely coloured regions corresponding to larger absolute values. The overall spatial distribution of the eigenfunctions is broadly similar for all three methods. However, the eigenfunctions produced by GT-MolSG appear to be less smooth than those generated by MolSG and G-MolSG. This behaviour is attributed to the sparse graph connectivity underlying GT-MolSG, which affects the information captured by the graph Laplacian

The distribution of colours is broadly similar for all three methods, however, the information captured by the GT-MolSG method appears less intense relative to MolSG and G-MolSG. This is attributed to the sparse nature of the graph underlying GT-MolSG which impacts on the information captured by the graph-based Laplacian. In contrast, the more densely connected graph used in G-MolSG more closely reproduces the patterns observed for MolSG. Since the primary difference between GT-MolSG and G-MolSG is the degree of graph connectivity, these results suggest that increased local connectivity, achieved by connecting each vertex to its 100 nearest neighbours, is important for preserving geometric information in the graph-based representation of the molecular surface.

Next, the performance of MolSG, GT-MolSG, and G-MolSG was evaluated using similarity searches across ten randomly selected targets. For each target, retrieval performance was averaged over 20 randomly selected active molecules used as queries, with identical targets and query sets employed for all methods. Results are shown in Fig. 2 where GT-MolSG yields lower BEDROC values than MolSG and G-MolSG for most targets.

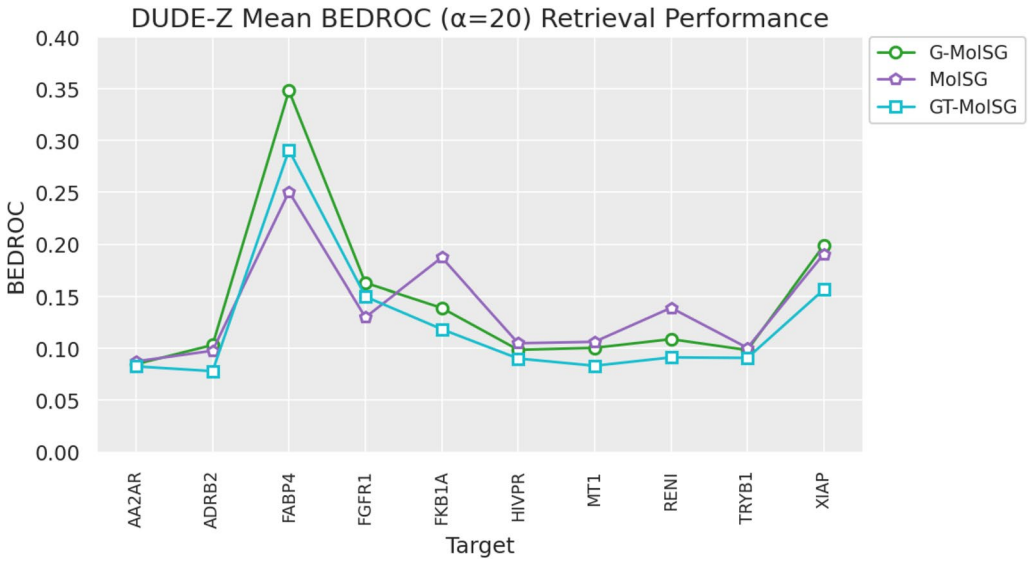


Fig. 2 Ten-target virtual screening benchmark. Mean BEDROC (α = 20) values for MolSG, G-MolSG, and GT-MolSG, averaged over 20 randomly selected active query molecules for each target

The reduced performance of GT-MolSG relative to G-MolSG suggests that restricting graph connectivity to the sparse triangulated mesh limits the ability of the resulting graph representation to capture geometric relationships effectively. In contrast, the denser *k*-NN graph employed by G-MolSG appears sufficient to preserve much of the geometric information encoded by the FEM-based representation used in MolSG. To further investigate this difference, we performed a mechanistic analysis of information propagation on the two graph representations using the graph heat kernel (Supporting Information, Section S10). This analysis illustrates that the restricted connectivity of GT-MolSG results in more localised propagation of information, whereas the denser G-MolSG graph permits information to propagate across a broader region of the molecular surface.

The similarity between the shape information represented by G-MolSG and MolSG extends beyond the low-frequency information shown in Fig. 1. Figure 3 shows the 50th eigenfunction plotted on the molecular surface of compound CHEMBL573550 in the XIAP dataset evaluated using the MolSG, G-MolSG and GT-MolSG methods, respectively. The distribution of colours is similar for MolSG and G-MolSG, whereas the surface for GT-MolSG deviates from these.

Combining shape and electrostatics

Figure 4 illustrates the impact of incorporating electrostatic information alongside shape in the spectral geometry descriptor. The electrostatic component was generated from partial charge information extracted from the ESP-DNN model developed by Rathi et al. [19] as described in the Methods. The first column (a) illustrates the ESP energy values on the surface of molecule CHEMBL573550 from the XIAP target dataset; column (b) shows the first eigenfunction for G-MolSG following the eigendecomposition using shape only; and column (c) shows the first eigenfunction for GE-MolSG when the eigendecomposition is based on shape and electrostatics.

The comparison between the representations in columns (b) and (c) reveals that incorporating electrostatic information substantially alters the resulting eigenfunction. Furthermore, several regions associated with strong positive and negative electrostatic potential in Fig. 4a are shown to have been captured in Fig. 4c. These observations suggest that electrostatic information is incorporated into the spectral representation and influences the resulting graph Laplacian eigendecomposition. The impact of this additional information on virtual screening performance is evaluated in the following section.

https://doi.org/10.1186/s13321-026-01310-6

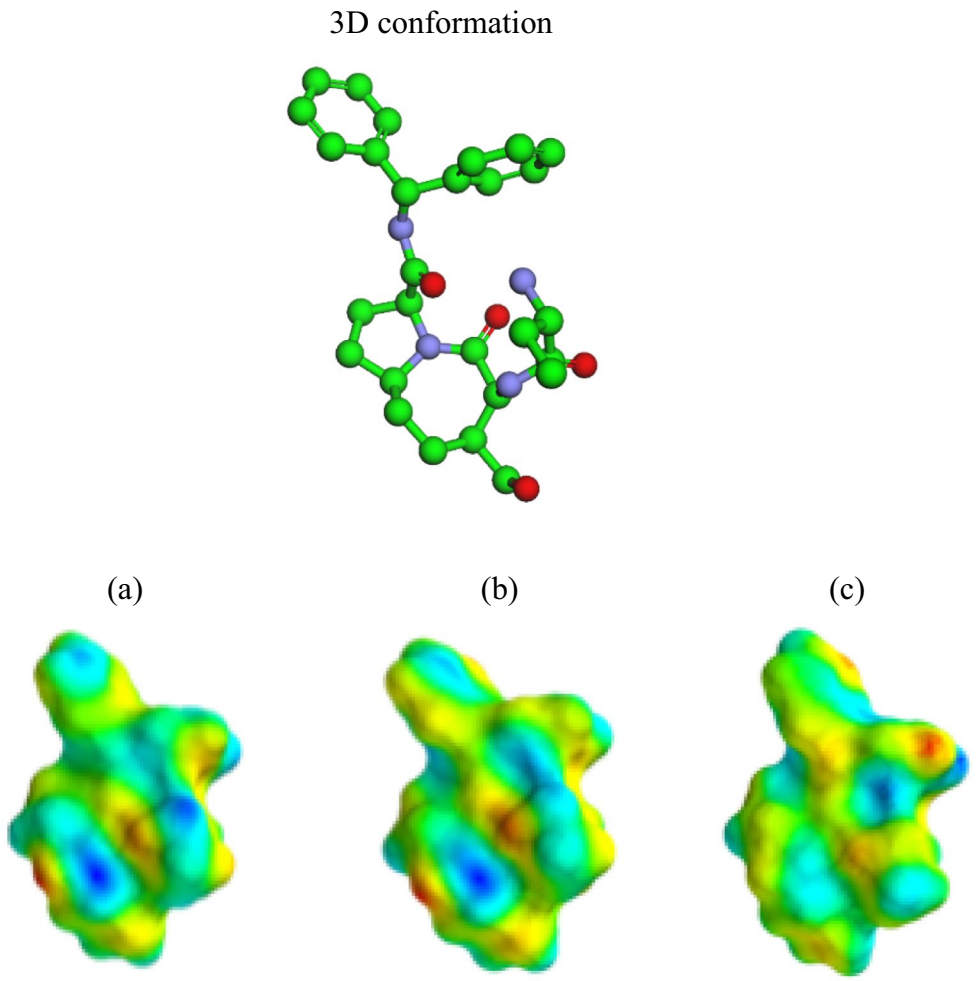


Fig. 3 Surface visualisation of the 50th eigenfunction for compound CHEMBL573550 (top) from the XIAP dataset generated using **a** MolSG, **b** G-MolSG, and **c** GT-MolSG

DUDE-Z performance

The performance of GE-MolSG, incorporating both shape and ESP-DNN-derived electrostatics, was compared with shape-only MolSG using similarity searches across all 43 targets in the DUDE-Z dataset. Figure 5 shows the mean BEDROC ($\alpha = 20$) and EF1% values for individual targets. For each target, 20 active compounds were randomly selected as queries, with the constraint that each query corresponded to a unique ChEMBL identifier. This restriction was introduced because DUDE-Z contains multiple tautomeric entries for compounds predicted to populate more than one protonation state at physiological pH.

Table 1 summarises the mean, median, and standard deviation of the performance metrics across all targets. GE-MolSG achieved substantially higher BEDROC and EF1% values than the shape-only MolSG implementation, indicating that the incorporation of electrostatic information provides information complementary to molecular shape alone.

A Wilcoxon signed-rank test, as described in Supporting Information S1, was applied to the BEDROC ($\alpha = 20$) and EF1% values obtained for each target. The results indicate that the inclusion of electrostatic information yields a statistically significant improvement in performance relative to the shape-only method ($p = 2.274 \times 10^{-13}$).

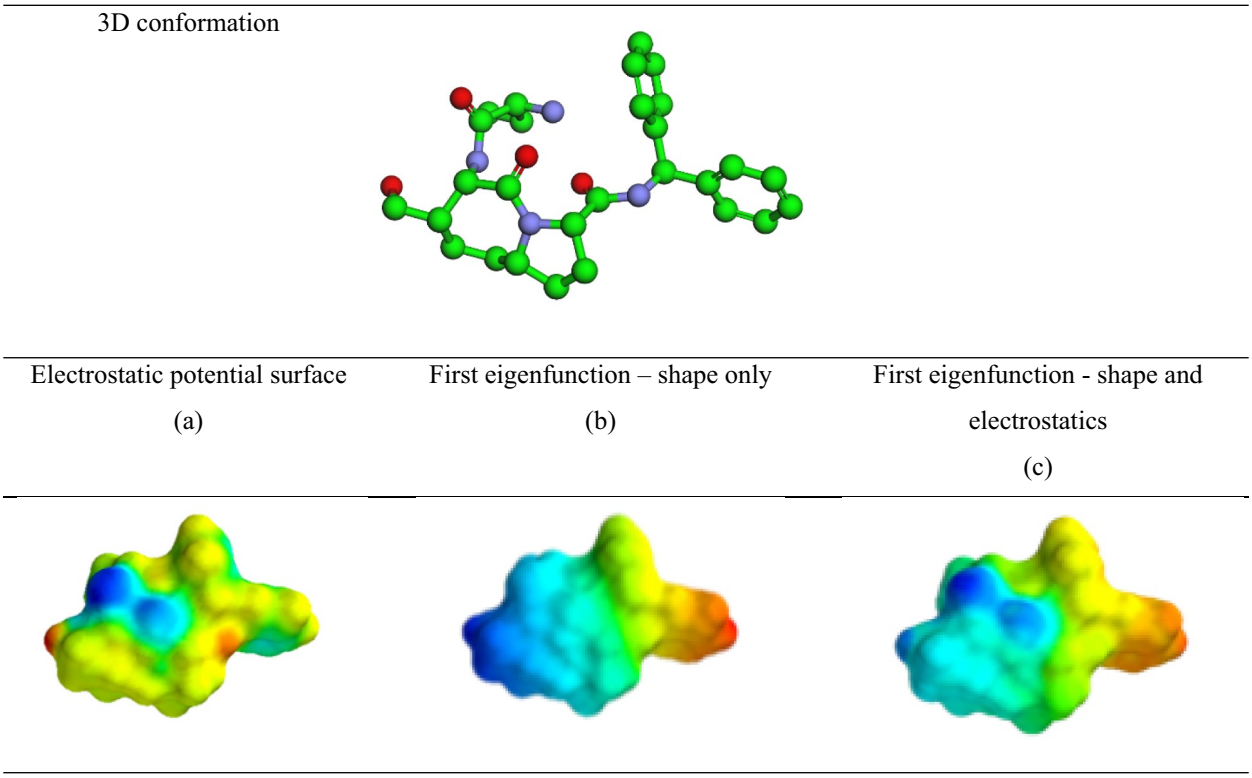


Fig. 4 Molecular surface represented as: **a** an electrostatic potential surface for CHEMBL573550 where the colour (red/blue positive/negative) represents ESP energy in kcal mol^(−1). In **b** and **c** the colours represent eigenfunction values captured during the eigendecomposition of the graph Laplacian for the shape-only and multi-property representations respectively. These results illustrate how incorporation of ESP energy into the graph Laplacian modifies the resulting eigenfunction representation across the molecular surface

To investigate whether the observed variation in GE-MolSG performance across DUDE-Z targets is associated with specific protein families, we further analysed BEDROC performance according to protein family categories. Figure 8S in Section S9 of the Supporting Information reports the BEDROC performance of GE-MolSG and MolSG across protein family categories derived from DUD-E [20] and UniProt [21].

Influence of electrostatics on the retrieval performance

To illustrate the contribution of electrostatic information, Fig. 6 shows the query compound ChEMBL113191_2 from the NRAM target together with representative retrieved actives from the NRAM dataset (5263 actives and decoys in total). The retrieved compounds contain a variety of Murcko scaffolds distinct from that of the relatively planar query compound and exhibit substantially different surface shapes. As MolSG relies solely on surface shape information, it struggles to retrieve these actives from the wider target set. In contrast, since GE-MolSG also captures electrostatic information, it enables the retrieval of compounds with different surface shapes but similar chemical features and charge distributions. In this experiment, the electrostatic information was derived from ESP-DNN partial charges [19].

Figure 7, which plots the rank position of the active compounds for GE-MolSG (y axis) and MolSG (x axis), further highlights the retrieval rank improvements obtained by GE-MolSG for the NRAM query compound ChEMBL113191_2. Most active compounds lie below the y=x diagonal, indicating improved retrieval ranks relative to MolSG across the wider target set. This suggests that incorporating electrostatic information enables

https://doi.org/10.1186/s13321-026-01310-6

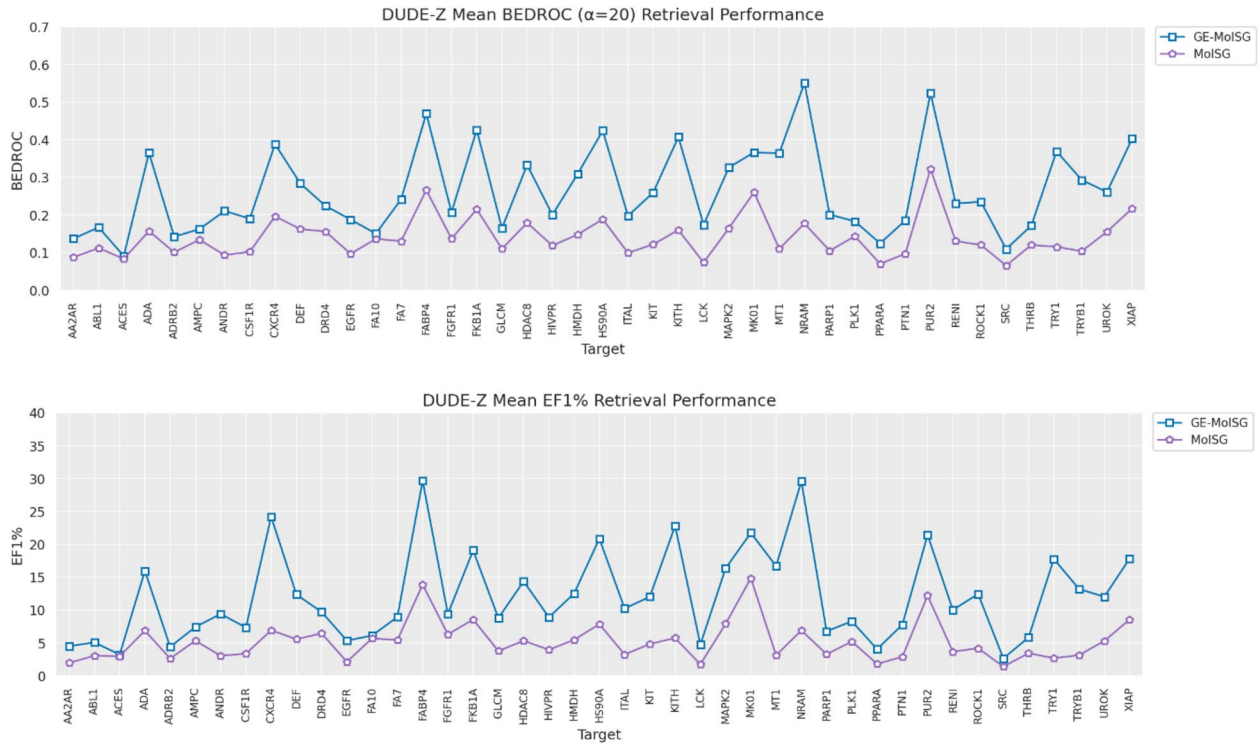


Fig. 5 Mean BEDROC ($\alpha = 20$) and EF1% values for MolSG and GE-MolSG across the 43 DUDE-Z targets. For each target, retrieval performance was averaged over 20 randomly selected active compounds used as queries

Table 1 Mean, median, and standard deviation of BEDROC ($\alpha = 20$) and EF1% values across the 43 DUDE-Z targets for MolSG and GE-MolSG

Method	BEDROC ($\alpha = 20$)		EF1%	
	Mean	Median	Mean	Median
GE-MolSG	0.27 (±0.11)	0.23 (±0.11)	12.13 (±6.90)	10.01 (±6.90)
MolSG	0.14 (±0.05)	0.13 (±0.05)	5.17 (± 3.00)	4.83 (±3.00)

GE-MolSG to retrieve chemically relevant compounds that are not effectively prioritised using surface shape information alone.

To further investigate the effect of incorporating electrostatic information, ESP energy distributions were analysed, and the results are included in Supporting Information S2. The investigated distributions include the ESP energy distributions of the query compound ChEMBL113191_2 and the six retrieved actives shown in Fig. 6, together with the distributions of the top 5% retrieved compounds for both GE-MolSG and MolSG using the same NRAM query compound.

Comparison of GE-MolSG with established shape and electrostatics methods

The performance of GE-MolSG was compared with the established alignment-free ElectroShape and the alignment-based ESP-Sim methods.

The ElectroShape descriptor [12] is an alignment-free descriptor that is similar in scope to GE-MolSG, however, it is considerably simpler and is essentially an extension of the shape-only USR [9] based on moments

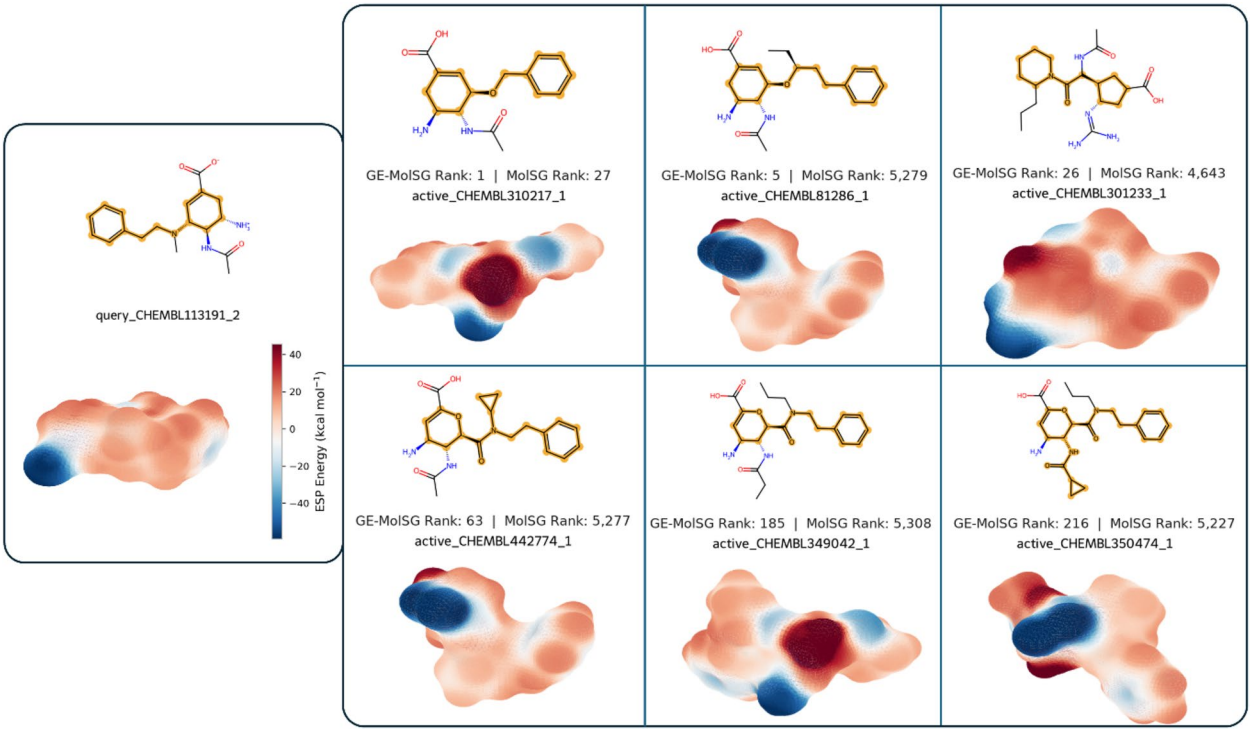


Fig. 6 Query compound ChEMBL113191_2 from the NRAM target (left) and representative active compounds retrieved from the NRAM dataset with distinct non-generic Murcko scaffolds (right). Electrostatic potential surface representations are shown beneath each compound. GE-MolSG assigns substantially improved retrieval ranks relative to MolSG for these examples

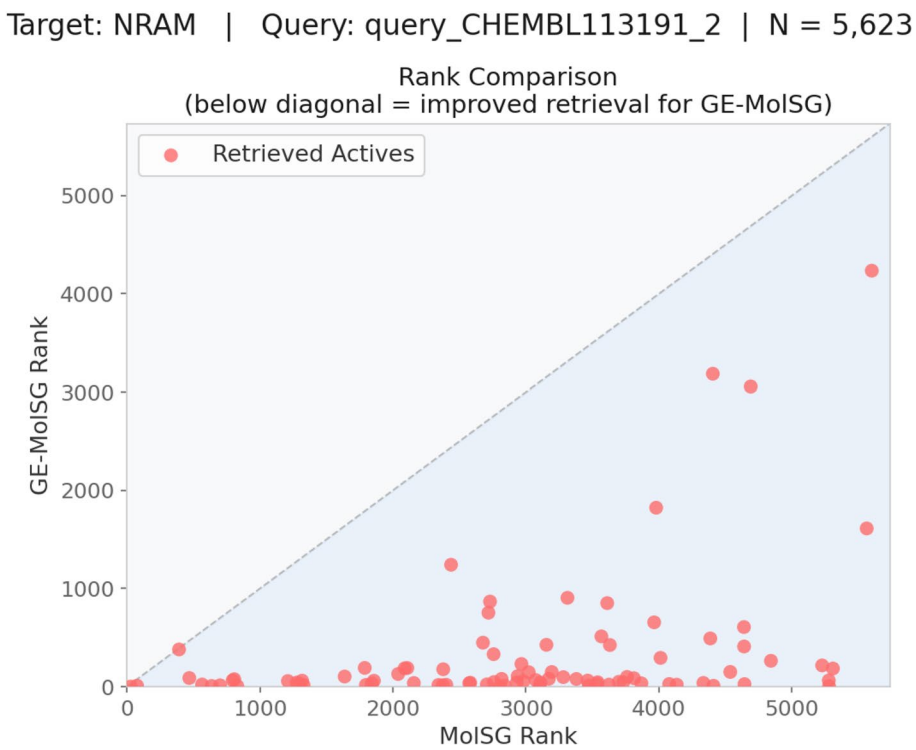


Fig. 7 Comparison of active compound retrieval ranks for GE-MolSG and MolSG using the NRAM query compound ChEMBL113191_2. Each point represents an active compound from the NRAM dataset, with points below the $y = x$ diagonal indicating improved retrieval rank for GE-MolSG relative to MolSG

https://doi.org/10.1186/s13321-026-01310-6

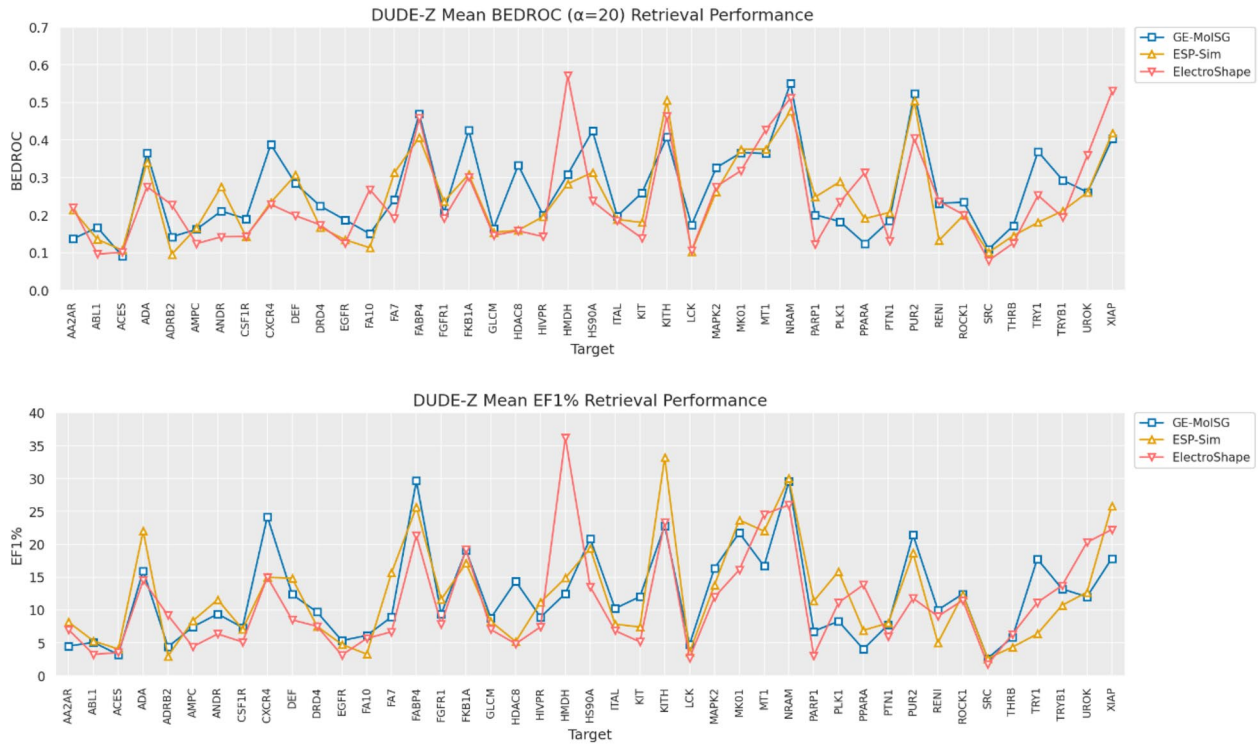


Fig. 8 Mean BEDROC (α = 20) (top) and EF1% (bottom) values for GE-MolSG, ElectroShape, and ESP-Sim across the 43 DUDE-Z targets. For each target, retrieval performance was averaged over 20 randomly selected active query compounds

with respect to a small number of reference points. It is possible to change the partial charge model used in ElectroShape and an initial series of experiments was conducted to determine which charge model gave the best performance. Results are shown in Supporting Information S3. The MMFF94 partial charge model [22] was found to give the best performance and was therefore used in the comparison with GE-MolSG. The Python code implementation of ElectroShape was obtained from the Open Drug Discovery Toolkit (ODDT) package (version 0.7, https://github.com/oddt/oddt). The same single-conformer for each molecule as used in GE-MolSG (and as described in the Methods section) was used to permit a direct comparison between methods.

ESP-Sim is an open-source alignment-dependent descriptor introduced by Bolcato et al. [11] available on GitHub (Version 1.0, 2023). Quantifying shape and electrostatic similarity is achieved by first aligning a molecule to the query based on shape, and then summing the ESP similarity and shape similarity. Shape is represented as a series of Gaussian functions with shape similarity based on volume overlap. Electrostatic information is similarly represented by Gaussian functions based on the Coulomb potential. ESP-Sim was shown to be competitive with a variety of published methods, including ROCS [6], in a benchmark on the 102 DUD-E targets. ESP-Sim was implemented here using a configuration analogous to that of the DUD-E benchmark provided in the GitHub repository (Version 1.0, 2023). The charge model used was the ML model as it exhibited best performance on the DUD-E benchmark. The same single-conformer for each molecule as used in GE-MolSG (and as described in the Methods section) was used for ESP-Sim to permit a direct comparison between methods.

Across the 43 DUDE-Z targets, GE-MolSG using ESP-DNN charges [19] achieved the highest mean BEDROC value of the three methods evaluated (Fig. 8, Table 2). Wilcoxon signed-rank tests applied to the target-wise BEDROC (α = 20) values are reported in Table 3S in Section S4 of the Supporting Information, and indicate that the improvement over both ElectroShape and ESP-Sim is statistically significant. In contrast, EF1% values were comparable between GE-MolSG and ESP-Sim, with both methods outperforming ElectroShape. Although

Table 2 Mean, median and standard deviation BEDROC ($\alpha = 20$) and EF1 performance for GE-MolSG, ElectroShape and ESP-Sim over 20 queries for all 43 targets in DUDE-Z

Method	BEDROC ($\alpha = 20$)		EF1%	
	Mean	Median	Mean	Median
GE-MolSG	0.27 (±0.11)	0.23 (±0.11)	12.13 (±6.90)	10.01 (±6.90)
ESP-Sim	0.24 (±0.11)	0.21 (±0.11)	12.23 (±7.63)	11.18 (±7.63)
ElectroShape	0.24 (±0.13)	0.20 (±0.13)	11.06 (±7.50)	8.49 (±7.50)

GE-MolSG did not demonstrate a statistically significant improvement over ESP-Sim in terms of EF1%, its superior BEDROC performance indicates more effective early enrichment across the benchmark.

These results demonstrate that GE-MolSG combines the practical advantages of an alignment-free representation with retrieval performance comparable to that of the alignment-dependent ESP-Sim method. While improvements in EF1% were limited, the statistically significant increase in BEDROC suggests that the richer graph-based representation provides more effective early recognition of active compounds. In comparison with ElectroShape, GE-MolSG achieves superior performance while retaining the benefits of an alignment-free approach.

The performance of GE-MolSG was further compared with ESP-Sim and ElectroShape across protein family categories. Figure 9S in Section S9 of the Supporting Information reports the mean BEDROC performance of the three methods for protein family categories derived from DUD-E [20] and UniProt [21].

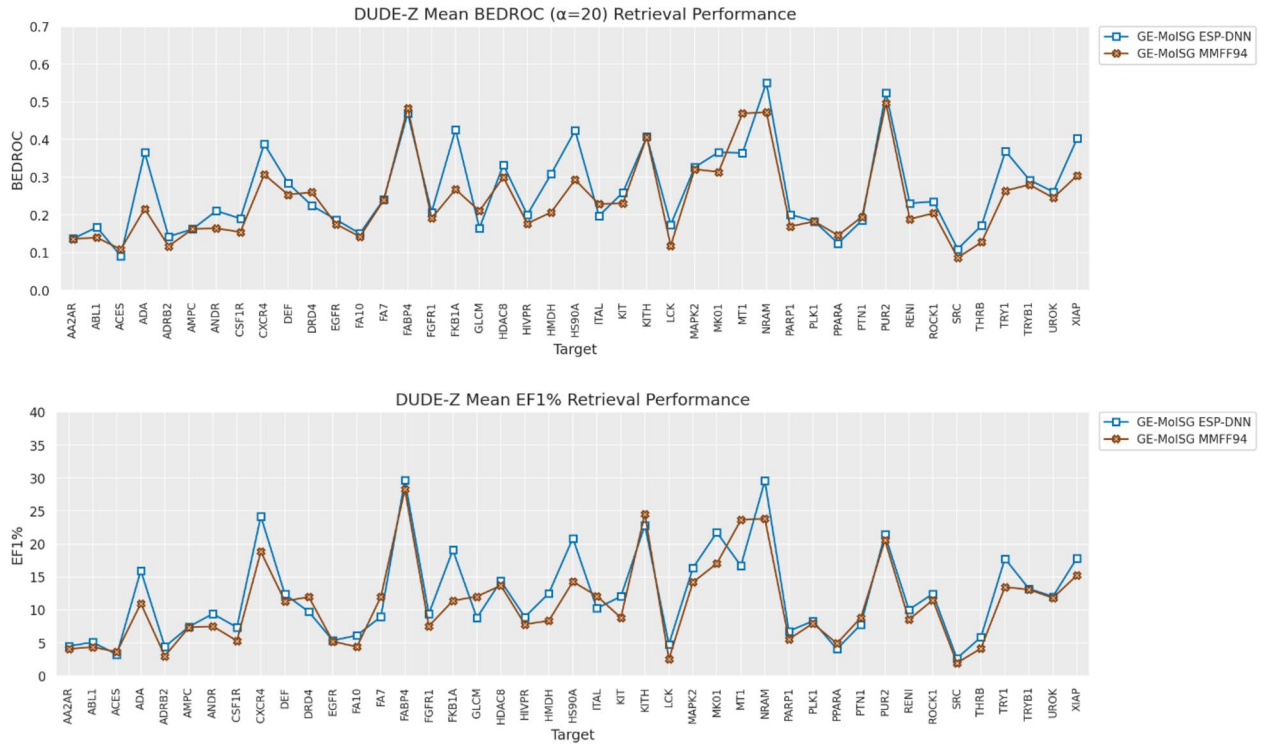


Fig. 9 Target-wise mean BEDROC (α = 20) (top) and EF1% (bottom) values for GE-MolSG using ESP-DNN and MMFF94 partial charge models across the 43 DUDE-Z targets. Results for each target were averaged over 20 randomly selected active query compounds

Limitations of the method

A key property of both MolSG and GE-MolSG is that both methods are isometric invariant. While this enables alignment-free comparison of molecular surfaces, it also renders the descriptors fundamentally incapable of distinguishing between enantiomers. Chirality corresponds to a reflection transformation that preserves intrinsic geometry, and therefore the geometric information encoded by the descriptors is identical for a pair of enantiomeric structures. Consequently, the same property that confers alignment invariance also eliminates sensitivity to molecular handedness.

Other molecular similarity methods address chirality through different mechanisms. For example, ElectroShape incorporates a symmetry-breaking component during descriptor construction that enables discrimination between enantiomers, whereas alignment-dependent methods such as ESP-Sim can capture chiral differences through the molecular alignment procedure itself.

The incorporation of chirality-aware information into the GE-MolSG framework therefore represents an important direction for future research. Potential approaches include the introduction of orientation-dependent surface features, chiral surface descriptors, or graph representations that are sensitive to reflection symmetry. Such extensions could enable discrimination between enantiomeric structures while preserving the alignment-free nature of the method.

Influence of charge model

We evaluated the effect of replacing ESP-DNN partial charges with those from MMFF94, which are computationally less expensive and were previously reported to be optimal for the ElectroShape descriptor using Open Babel version 3.1.1. Figure 9 compares the BEDROC ($\alpha = 20$) and EF1% performance of GE-MolSG using MMFF94 and ESP-DNN partial charges across all DUDE-Z targets. The performance profiles of the two charge models are broadly similar, with both producing comparable target-wise trends throughout the benchmark. Although differences are observed for individual targets, the relative ranking of targets is largely preserved. Pairwise BEDROC and EF1% values for all DUDE-Z targets are provided in Supporting Information S5.

As summarised in Table 3, GE-MolSG using ESP-DNN charges achieves higher mean BEDROC and EF1% values than the corresponding MMFF94 implementation. The improvements are modest but consistent across both metrics. Nevertheless, the relatively small magnitude of the differences suggests that GE-MolSG remains reasonably robust to the choice of partial charge model.

Proof-of-concept surface patch extension

To demonstrate the extensibility of the graph-based framework, we developed GM-MolSG, an extension of the GE-MolSG codebook representation that incorporates local surface patch descriptors (Fig. 11). The extended representation incorporates additional chemical and geometric surface features including ESP, Wildman-Crippen logP [23], Shape Index [24], and Jazzy hydrogen-bond donor and acceptor strengths [25].

Table 3 Mean, median, and standard deviation of BEDROC ($\alpha = 20$) and EF1% values across the 43 DUDE-Z targets for GE-MolSG using ESP-DNN and MMFF94 partial charge models

Method	BEDROC ($\alpha = 20$)		EF1%	
	Mean	Median	Mean	Median
GE-MolSG ESP-DNN	0.27 (±0.11)	0.23 (±0.11)	12.13 (±6.90)	10.01 (±6.90)
GE-MolSG MMFF94	0.24 (±0.10)	0.21 (±0.10)	10.86 (±6.33)	10.97 (±6.33)

GM-MolSG was benchmarked against GE-MolSG, MolSG, the alignment-free methods ElectroShape and ElectroShape5D [12], and the alignment-dependent methods Roshambo [26] and ESP-Sim [11]. Figure 10 shows the BEDROC ($\alpha = 20$) and EF1% performance obtained across the DUDE-Z benchmark. For each target, the same set of 20 active query compounds used in the previous experiments was employed to ensure a consistent comparison.

A Wilcoxon signed-rank test was used to compare GM-MolSG against the reference methods across the DUDE-Z benchmark using a 95% confidence level. GM-MolSG showed higher BEDROC and EF1% performance than the comparator methods across the benchmark, with statistically significant differences observed in the pairwise comparisons. Table 3S in Section S4 of the Supporting Information reports the exact p values for all pairwise comparisons, confirming the statistical significance of the observed differences. These preliminary results suggest that incorporating local surface patch information may provide a promising direction for extending graph spectral molecular descriptors.

To further assess whether performance differences between the evaluated methods are associated with specific protein families, we analysed their BEDROC performance across protein family categories. Figure 10S in Section S9 of the Supporting Information reports the mean BEDROC performance of GM-MolSG, GE-MolSG, MolSG, ESP-Sim, ElectroShape, Roshambo, and ElectroShape5D across protein family categories derived from DUD-E [20] and UniProt [21].

Timing analysis

Timings for descriptor generation and comparison for all methods detailed in this work are presented in Supporting Information S8 based on a randomly sampled set of 500 molecules from the target PUR2 in the DUDE-Z dataset. Since all methods use the same single conformation for each molecule the timings are for post-conformer generation processes. All timing analysis was performed on a system which features an AMD Ryzen 7 9800X3D CPU and 32 GB DDR5 RAM.

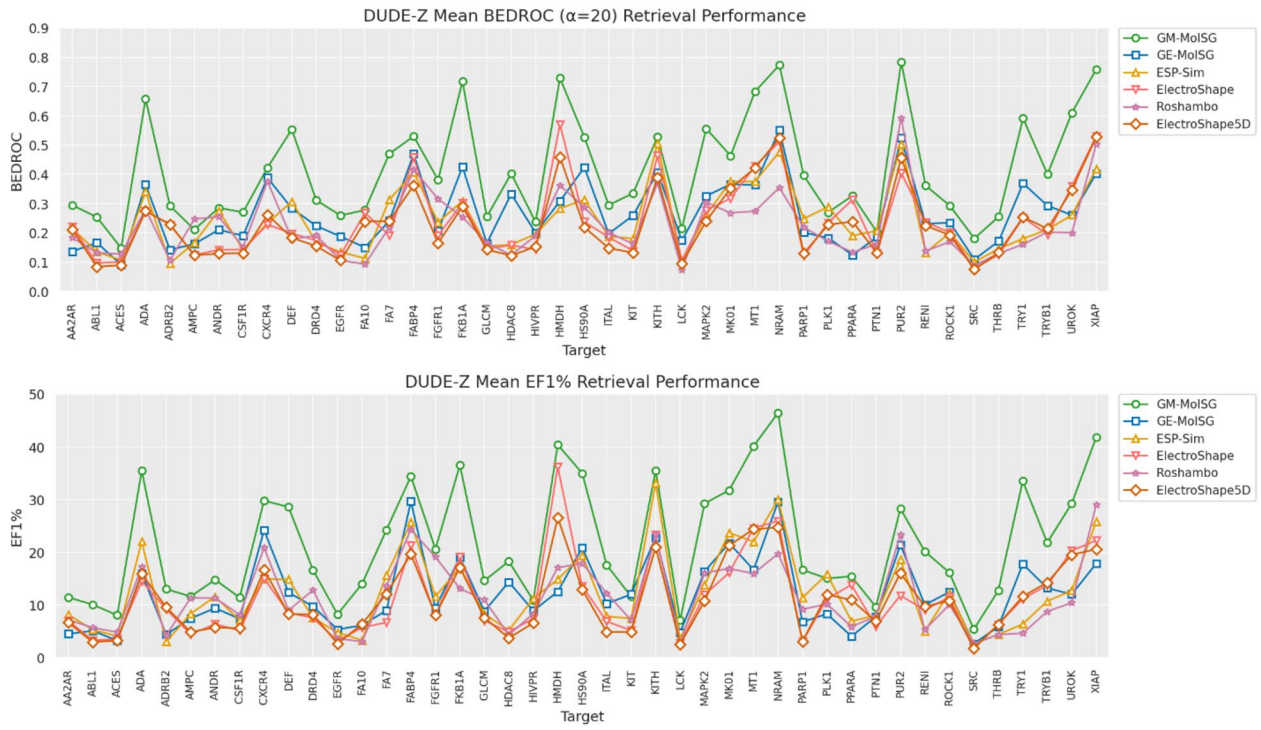


Fig. 10 Mean BEDROC ($\alpha = 20$) (top) and EF1% (bottom) values across the DUDE-Z targets for GM-MolSG compared with GE-MolSG, MolSG, ESP-Sim, ElectroShape, Roshambo, and ElectroShape5D

Figure 5S compares the time required by the ESP-DNN model [19] to generate predicted atom partial charges with that required to generate empirical MMFF94 partial charges on CPU. The average time per molecule was approximately 5.0×10^{-4} s for ESP-DNN and 2.5×10^{-5} s for MMFF94, making MMFF94 approximately 20-fold faster.

Figure 6S compares the end-to-end descriptor generation time, starting from individual 3D molecular conformations, for the FEM-based and graph Laplacian-based MolSG and GE-MolSG methods with those of ElectroShape and ESP-Sim. For the spectral methods, most of the computational cost is associated with molecular surface generation, resulting in longer descriptor generation times than the established methods. However, as alignment-free molecular descriptors, GE-MolSG descriptors can be generated once and stored as vector-based representations of the corresponding 3D molecular conformations for subsequent retrieval and similarity calculations. Furthermore, the current GE-MolSG implementation has not been specifically optimised for computational efficiency, and the timing comparison should therefore be considered preliminary.

Figure 7S compares the time required to calculate similarities between 50 query molecules, randomly subsampled from 500 randomly selected molecules in the PUR2 dataset, and each molecule in the dataset. For the alignment-free methods, similarity evaluation is rapid, with the differences in computational cost primarily reflecting the dimensionality of the descriptor representations. ElectroShape and ElectroShape5D require the least computation, whereas the spectral geometry methods incur a higher cost due to their richer, higher-dimensional representations. In contrast, ESP-Sim and Roshambo require substantially more computation because similarity evaluation involves rigid-body alignment and molecular overlay. These results highlight the computational advantage of the alignment-free representations for large-scale similarity searching.

Conclusions

We have presented GE-MolSG, a novel alignment-free molecular descriptor that combines molecular shape and electrostatic information through a spectral graph representation of molecular surfaces. By replacing the Finite Element Method approximation of the Laplace–Beltrami operator used in MolSG with a graph Laplacian formulation, GE-MolSG enables the direct integration of physicochemical information into compact alignment-independent descriptors suitable for rapid similarity searching, overcoming a key limitation of the original MolSG framework.

Incorporating electrostatic information led to improvements in virtual screening performance across the DUDE-Z benchmark. GE-MolSG consistently showed an overall improvement over the established alignment-free ElectroShape method and achieved retrieval performance comparable to, and in some cases exceeding, the ESP-Sim approach, despite avoiding computationally expensive molecular alignment. These results suggest that spectral graph representations offer the effectiveness advantages of alignment-based methods and the efficiency associated with alignment-free methods.

GE-MolSG was also able to retrieve active compounds with alternative chemotypes and distinct Murcko scaffolds while preserving similar electrostatic and surface characteristics to the query, highlighting its potential for scaffold hopping and exploration of chemically diverse bioactive space. The compact spectral embeddings generated by GE-MolSG may therefore provide a useful basis for searching molecular databases to rapidly retrieve compounds with similar three-dimensional recognition properties despite substantial structural differences.

A key advantage of the GE-MolSG framework is its flexibility and extensibility. Because the graph operates on vertex-associated feature vectors, arbitrary molecular surface properties can be incorporated beyond electrostatics, making our framework broadly applicable to multi-property molecular comparison. To explore this capability, we introduced GM-MolSG as a proof-of-concept extension incorporating local surface patch information together with electrostatics, lipophilicity, Shape Index and hydrogen-bonding properties. Preliminary benchmarking suggests that incorporating local surface information may provide a promising direction for further improving alignment-free molecular retrieval. However, GM-MolSG remains a prototype, and further

work is required to establish the optimal representation and contribution of its constituent features. In particular, the current implementation generates an individual patch around every surface vertex, resulting in substantial computational overhead. Restricting patch generation to geometrically and chemically salient surface points may provide a more efficient representation while retaining relevant local information.

The experiments reported here use a single low-energy conformer for each molecule, as is common practice when comparing three-dimensional ligand-based virtual screening methods [3, 17]. It is notable that spectral descriptors based on intrinsic surface geometry can exhibit some robustness to changes in pose and conformation, although they remain sensitive to non-isometric deformations that alter distances across a molecular surface [16, 27]. Preliminary work by Seddon et al. [28] suggests that spectral geometry descriptors calculated for different conformations of the same molecule can retain shared intrinsic geometric information, and a systematic assessment of conformational sensitivity using ensembles of molecular conformations would provide a useful extension of the present work.

The current implementation incurs a higher computational cost during descriptor generation than the established alignment-free methods considered here, primarily due to molecular surface generation and spectral decomposition. This cost is incurred during descriptor initialisation, while subsequent similarity evaluation using the precomputed vector representations is rapid and avoids repeated molecular alignment. Further optimisation of surface generation and eigendecomposition could therefore improve the scalability of GE-MolSG for large molecular databases. Furthermore, scalable indexing schemes could be explored to enable faster vector comparisons. Finally, and as mentioned above, the incorporation of chirality-aware information into the GE-MolSG framework is an area for future research.

Overall, GE-MolSG provides a flexible spectral framework for representing molecular surface geometry and electrostatic information in an alignment-free form. The results demonstrate its potential for three-dimensional virtual screening and scaffold exploration, while the graph-based formulation provides a foundation for extending the representation to additional molecular surface properties. The preliminary GM-MolSG results further indicate the potential of incorporating local surface information, motivating future work towards more efficient and systematically optimised multi-property surface representations.

Methods

The workflow for the GE-MolSG method is illustrated in Fig. 11. GE-MolSG operates on electrostatic potential (ESP) surfaces generated from 3D molecular conformations. The ESP surface is represented as a point cloud and converted into a graph using the k-nearest neighbour algorithm. The graph is described by an affinity matrix W, encoding spatial and electrostatic relationships between surface points, and a degree matrix D, whose diagonal entries contain the sum of edge weights connected to each node.

The affinity and degree matrices are used to construct the graph Laplacian, from which the eigenvectors and eigenvalues are obtained. These are processed in a manner similar to MolSG by first generating local geometry descriptors and subsequently converting them into alignment-independent global descriptors. Relative to MolSG, the main differences are the use of an ESP surface representation instead of a shape-only triangulated mesh, and the use of a graph Laplacian approximation of the LBO rather than the FEM discretisation, for computing the eigenpairs. The following sections describe each stage of the GE-MolSG workflow in detail.

Surface generation and molecular charge information

Molecular surfaces and charges were generated using the ESP-DNN toolkit developed by Rathi et al. [19], which was reported to produce near-DFT quality partial charges at substantially lower computational cost than traditional quantum chemical methods. ESP-DNN internally predicts partial charges, which are projected onto a triangulated Connolly surface mesh generated using a variant of the marching cubes algorithm. The option is also provided to

https://doi.org/10.1186/s13321-026-01310-6

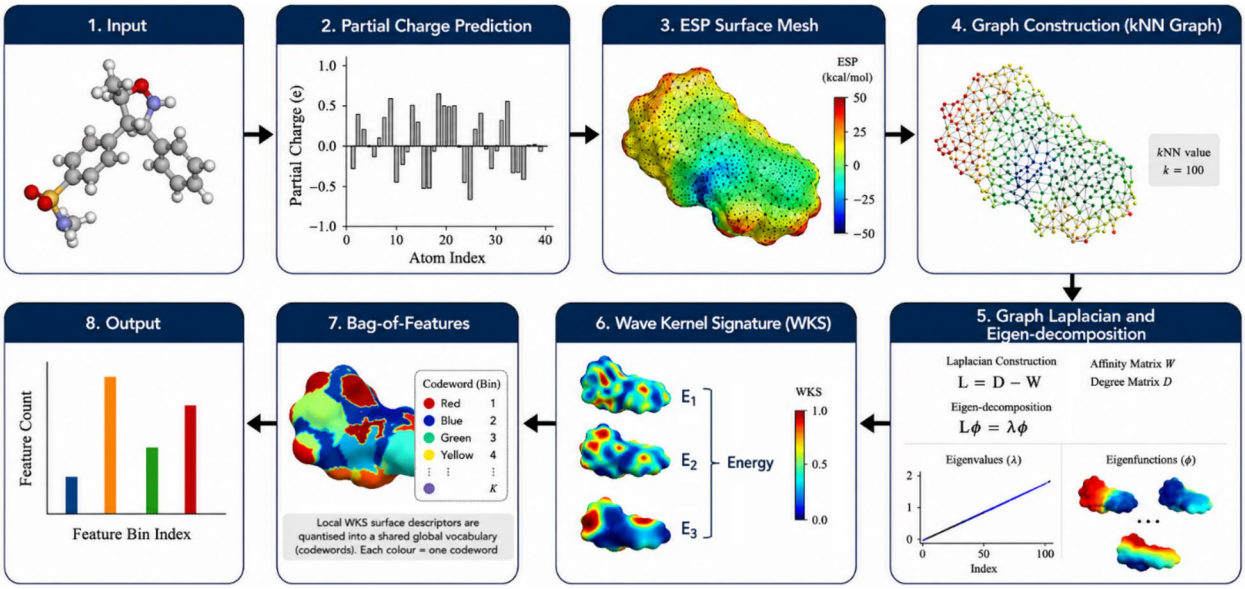


Fig. 11 Workflow of the GE-MolSG descriptor. A 3D molecular conformation is first processed to predict partial charges and generate an electrostatic potential (ESP) surface mesh. The ESP surface is converted into a k-nearest neighbours graph, from which the graph Laplacian is constructed and eigendecomposed to obtain eigenvalues and eigenfunctions. Wave Kernel Signatures (WKS) are then computed and encoded into a bag-of-features representation to produce the final alignment-independent molecular descriptor

generate ESP surfaces using MMFF94 partial charges [22] since these proved to exhibit comparable performance in the resulting descriptor.

Representing the surface as a graph

The original mesh connectivity is discarded, and the surface is represented as a four-dimensional point cloud $(x,y,z,\mu q)$, where x, y, and z are the 3D spatial coordinates of points on the ESP surface, q is the ESP energy value, and μ is a scaling factor. As in the ElectroShape method of Armstrong et al. [12], the electrostatic component is scaled to balance spatial and electrostatic contributions.

The point cloud is converted into a graph using the k-nearest neighbours algorithm. Pairwise distances between neighbouring vertices in the four-dimensional space are transformed into similarities using the variable-bandwidth Gaussian kernel described in Eq. (9).

The Laplace–Beltrami spectrum is approximated through the eigendecomposition of the graph Laplacian derived from the affinity and degree matrices. The resulting eigenpairs are orthonormal and ordered such that early eigenpairs encode global surface properties, whereas later eigenpairs capture increasingly local features. For a graph containing N vertices, a full eigendecomposition would produce N eigenpairs. However, as the contribution of successive eigenpairs decreases, the spectrum is truncated to the first 100 eigenpairs following our previous work on spectral shape descriptors. Figure 4S in Section S6 of the Supporting Information provides an assessment of this choice for GE-MolSG on DUDE-Z, showing that performance reaches a plateau at approximately $K = 100$. The eigendecomposition therefore produces a K-dimensional vector of eigenvalues (λ) and an $N \times K$ matrix of eigenfunctions (ϕ), where $K = 100$.

Local geometry descriptors

The spectrum of K eigenvalues and $N \times K$ eigenfunctions is then transformed using the Wave Kernel Signature (WKS) [29], as implemented in MolSG. The WKS is a spectral descriptor ((1) that captures local geometric properties at multiple spatial scales by passing a Gaussian filter over the spectrum to smooth out the local and global geometry effects. The spectrum is divided into D uniformly spaced levels/intervals, and a Gaussian function is centred at each level to amplify the contribution of eigenvalues around that point. The WKS generates a descriptor vector for each vertex where the length of the vector is determined by the number of defined intervals, i.e., the output is a matrix of vertex descriptors of dimensionality $N \times D$. In (1), e is the mean value in the Eth interval and the numerator is the squared distance of the log of the ith eigenvalue to the log of the middle of the interval; σ^2 is an arbitrary parameter that controls the width of the Gaussian weighting in log-eigenvalue space.

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

As the WKS moves across the spectrum, the contribution shifts from global to local features. In order to weight the contributions over different intervals equally, a normalisation constant, c_E, is applied so that the area under each function is the same. 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. Thus, each vertex is described by a D-dimensional vector with each element representing a different balance between local and global features as shown in (2):

$$\boldsymbol{p}(x) = (WKS_1(x), \ldots, WKS_D(x)) \tag{2}$$

Global geometry descriptor

The local geometry descriptors are defined at the vertices of the graph, G, and are transformed to global geometry descriptors to enable the comparison of molecules without the need for an alignment. This is achieved using the Bag-of-Features (BoF) method described by Bronstein et al. [30]. In an offline procedure, a codebook C consisting of geometric codewords, $\{c_1, c_2, \ldots, c_{q-1}, c_q\}$ is generated from a reference set of molecules. The local geometry descriptors from the reference molecules are clustered using the k-means algorithm with k-means++ initialisation [31], and the resulting cluster centroids are used as the codewords within the codebook. The size of the codebook is therefore determined by the parameter k.

The codebooks used to generate the global descriptors for GE-MolSG were derived from randomly sampled 400 molecules from the DUDE-Z dataset. This number was decided based on evaluating different rounds of sampling of molecules also of increasing sample size as described by Middleton [32]. The global descriptor for an unseen molecule is generated by mapping its local geometry descriptors to the codebook using a vector quantisation scheme. Two approaches were investigated: hard quantisation and k-nearest neighbour (kNN) quantisation, both illustrated in Fig. 12. Hard quantisation assigns each local geometry descriptor to its nearest codeword. In contrast, kNN quantisation assigns descriptors to multiple nearby codewords using a weighting scheme, reducing information loss by allowing stronger contributions from closer codewords. The resulting assignments are aggregated and normalised to produce a descriptor invariant to molecular size.

Validation data and virtual screening

The performance of GE-MolSG was evaluated in similarity searching experiments across all targets in the Directory of Useful Decoys Enhanced (DUDE-Z) dataset [33]. DUDE-Z is the third iteration of the DUD virtual screening benchmark series, following DUD [34] and DUD-E [20], and contains sets of active and decoy compounds

https://doi.org/10.1186/s13321-026-01310-6

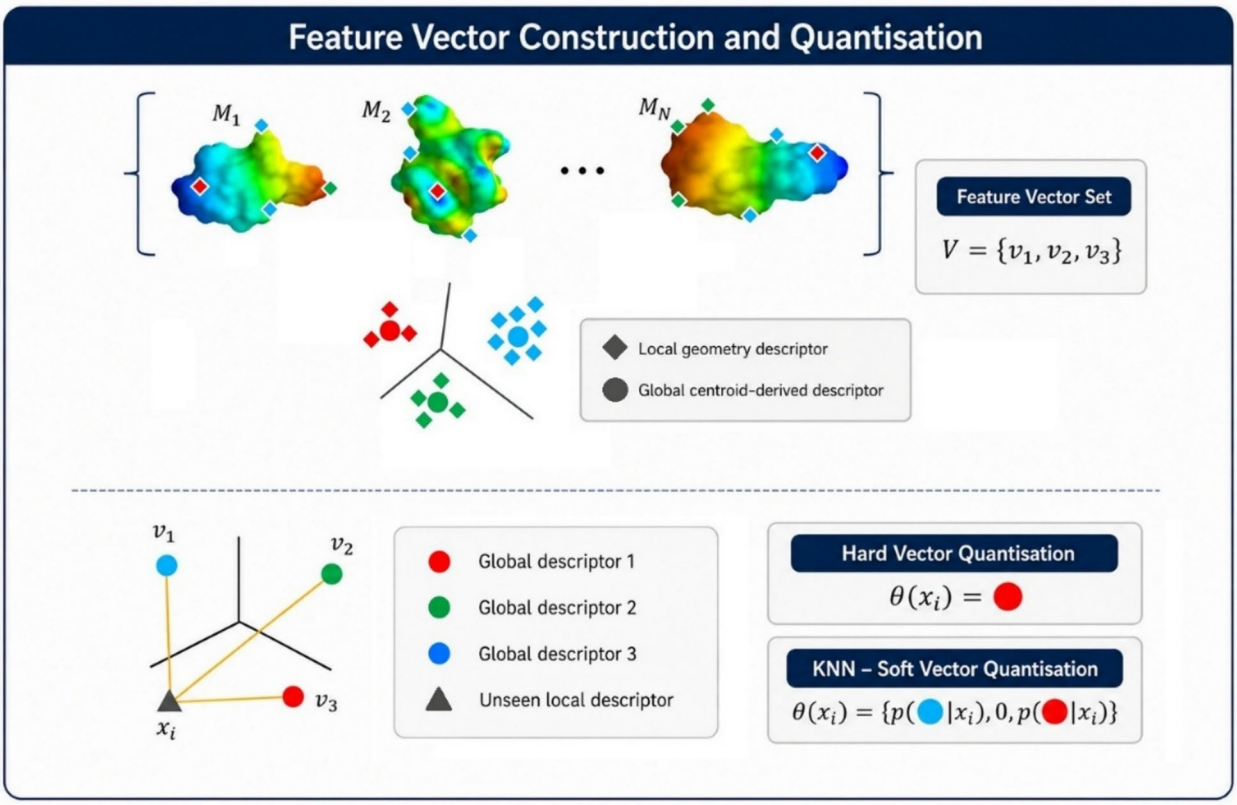


Fig. 12 Top: Local geometry descriptors extracted from reference molecules are clustered to generate centroid-derived global descriptors, which together form the Bag-of-Features codebook. Bottom: Local geometry descriptors from unseen molecules are mapped to the codebook using either hard vector quantisation or k-nearest neighbour (kNN) soft quantisation. In hard quantisation, each descriptor is assigned to its nearest codeword, whereas kNN quantisation distributes assignments across multiple nearby codewords using distance-based weights

for 43 protein targets. The decoys were primarily obtained from the ZINC15 database and were selected to match the physicochemical properties of the actives while remaining topologically dissimilar. As described by Réau et al. [35], molecular fingerprint similarity was used to ensure separation between active and putative inactive compounds, following the similar property principle whereby compounds with low topological similarity to known actives are assumed to lack biological activity. This strategy is commonly used in virtual screening benchmark construction, including DUDE-Z [33]. Unlike DUD-E, where actives were found to be more highly charged than decoys, DUDE-Z addresses this source of bias by ensuring matched charge distributions between active and decoy compounds.

The preparation workflow for the DUDE-Z dataset is illustrated in Fig. 13 and was implemented using RDKit. Although the dataset is available in both SDF and SMILES formats, SMILES representations were used following



Fig. 13 GE-MolSG data preparation workflow. SMILES representations were extracted and explicit hydrogens were added prior to 3D conformer generation using the RDKit ETKDGv3 algorithm. The generated conformers were optimised using the MMFF94 force field, and the lowest-energy conformer (LEC) was retained as the representative structure. The selected LEC was then processed using the GE-MolSG workflow described in Fig. 11 to generate an alignment-independent molecular descriptor

the methodology of Jiang et al. [3]. The protonation and tautomeric states provided in the original dataset were retained unchanged. Implicit hydrogens were added using the AddHs function in RDKit.

Three-dimensional conformations were generated using ETKDGv3 with 50 conformers produced per molecule. The conformers were subsequently optimised using the MMFF94 force field [22] via the MMFFOptimizeMolecule function in RDKit. Default parameters were used throughout, except that the maximum number of optimisation iterations was increased from 200 to 500. The lowest-energy conformer (LEC) for each molecule was retained as the representative structure.

As the ESP-DNN model [19] accepts input in PDB format, the selected LECs were saved as individual PDB files named according to their activity class and associated ChEMBL or ZINC identifier. DUDE-Z also contains multiple protomeric forms associated with the same ChEMBL or ZINC identifier. To avoid artificially inflating or reducing retrieval performance through repeated retrieval of the same compound, only the highest-ranking protomer for each identifier was retained in the ranked lists.

For each target, 20 query compounds were selected at random from the active set. Virtual screening performance was evaluated using the Enrichment Factor at 1% (EF1%) and BEDROC [36]. EF1% measures the proportion of active compounds retrieved within the top 1% of the ranked list but does not account for the relative ordering of actives within this subset. BEDROC addresses this limitation by weighting the early retrieval of active compounds. The BEDROC parameter α controls the importance assigned to the early portion of the ranked list, with $\alpha = 20$ corresponding to 80% of the score being determined by the top 8% of retrieved compounds.

Parameterisation of GE-MolSG

As with MolSG [17], GE-MolSG contains several adjustable parameters that influence the geometry information captured from molecular surface representations. A restricted grid search was performed to optimise these parameters. Table 4 summarises the parameter ranges explored together with the final optimised values. Parameter optimisation was conducted using 10 DUDE-Z targets, with 20 randomly selected actives per target used as query compounds (Table 5). Performance was evaluated using BEDROC ($\alpha = 20$) and EF1%, and the best-performing configuration was selected.

Three values for the number of nearest neighbours used during graph construction ($k = 10, 50,$ and 100) were evaluated, with $k = 100$ producing the best performance. Increasing k increases the local neighbourhood information available to each vertex, although at the cost of a denser graph Laplacian and increased eigendecomposition complexity. Two ESP scaling factors (0.2 and 0.3) and two distance metrics (Euclidean and cosine) were also investigated. The optimal configuration used an ESP scaling factor of 0.3 with Euclidean distance.

Table 4 Parameter settings explored during grid search optimisation and the final values selected based on validation across 10 DUDE-Z targets

Parameter	Tested values	Optimised value
Graph construction		
kNN nearest neighbours (kNN)	10, 50, 100	100
kNN distance metric (DM)	Euclidean, Cosine	Euclidean
Electrostatic weighting ($\boldsymbol{\mu}$)	0.20, 0.30	0.3
Local geometry descriptors		
WKS evaluations ($\boldsymbol{D}$)	100, 150	100
WKS standard deviation ($\boldsymbol{\sigma}$)	7, 15	15
Global geometry descriptors		
Number of codewords (CW)	250, 500, 1000	1000
Vector quantization (VQ)	Hard quantisation, k-nearest neighbours ($k = 3$)	k-nearest neighbours ($k = 3$)
Similarity metric		
Global descriptor similarity metric (SM)	Cosine similarity, Chi squared	Chi squared

Table 5 DUDE-Z targets used in the GE-MolSG parameterisation experiments

Target	Number of actives	Number of decoys	Total number of compounds
AA2AR	90	4489	4579
ADRB2	65	2932	2997
FABP4	45	2204	2249
FGFR1	71	3419	3490
FKB1A	46	2298	2344
HIVPR	64	3146	3210
MT1	28	2150	2178
RENI	115	3736	3851
TRYB1	38	1801	1839
XIAP	76	3734	3810

As GE-MolSG incorporates additional electrostatic information relative to MolSG, the local and global descriptor parameters were also re-optimised. Two values for the number of Wave Kernel Signature (WKS) evaluations ($D = 100$ and 150) were tested, with the optimal value remaining $D = 100$, consistent with MolSG. In contrast, the optimal WKS standard deviation increased from $\sigma = 7$ in MolSG to $\sigma = 15$ in GE-MolSG. Larger σ values place greater emphasis on global surface structure, suggesting that broad shape and electrostatic relationships are more informative for distinguishing active and decoy compounds than highly local surface detail.

Three codebook sizes were evaluated, with an optimal value of $CW = 1000$. This is larger than the value previously used for MolSG ($CW = 500$), suggesting that the additional electrostatic information benefits from a richer feature vocabulary. The optimal quantisation strategy also differed from MolSG. Whereas MolSG performed best using hard vector quantisation, GE-MolSG achieved improved performance using k-nearest neighbour quantisation, which assigns descriptors across multiple nearby codewords rather than a single closest codeword.

Finally, two similarity measures were evaluated for comparing the global geometry descriptors: cosine similarity, as previously used in MolSG, and the Chi-squared metric following the work of Yang et al. [37]. These are described in Supporting Information S7. The results from the grid search showing the sensitivity of the method to the number of k-nearest neighbours used to construct the surface graphs and choice of distance metric are reported in the Supporting Information S11.

Patch surface extension

The codeword vocabulary was further extended through construction of an additional codebook based on local surface patches. Patches were centred on each surface vertex using the Bag-of-Covariance matrices (BoC) methodology. Similar to the previously described Bag-of-Features (BoF) approach, the method generates a codebook representation of local surface environments. However, instead of representing each point using a Wave Kernel Signature (WKS) descriptor, each patch is represented by a covariance matrix capturing relationships between local geometric and chemical surface features within the neighbourhood of the vertex.

Our implementation follows the methodology proposed by Faraki et al. [54], which applies a Log-Euclidean transformation to covariance matrices. This transformation maps the curved Riemannian manifold of covariance matrices to a Euclidean tangent space, enabling the use of standard Euclidean distance metrics for covariance matrix comparison.

Additional chemical and geometric features were extracted from the molecular surface and underlying chemical structure to represent molecular recognition properties similarly employed in established patch-based frameworks such as MaSIF [39]. The extended feature representation included electrostatic potential (ESP) derived from MMFF94 partial charges projected onto the molecular surface, Wildman-Crippen logP [23] representing hydrophobicity potential, Shape Index [24] encoding local surface geometry, mean curvature and Gaussian

curvature, as complementary descriptors of local surface geometry derived from the principal curvatures, and Jazzy hydrogen-bond donor and acceptor strengths, which have recently demonstrated effectiveness for modelling ADME endpoints in a blind prediction challenge [25, 40]. All of the above features are projected onto the molecular surface to produce scalar field representations of the selected properties [41].

Theory

Graph Laplacian approximation

The Laplace–Beltrami operator (LBO) is a spectral differential operator that can be applied to capture the intrinsic geometry of a manifold. The LBO generalises the Laplacian operator to curved manifolds rather than Euclidean space. The LBO is widely applied in the analysis of geometric objects, including tasks such as heat diffusion and shape analysis. As illustrated in Eq. (3), the LBO captures the divergence of the gradient of a function on a manifold, where the manifold can be considered as a low-dimensional surface embedded in some higher-dimensional space.

$$\Delta_{\mathcal{M}}(f) = div_{\mathcal{M}}(grad_{\mathcal{M}}(f)) \tag{3}$$

Because the LBO is continuous, it must be discretised to obtain a computationally tractable approximation. The Finite Element Method (FEM), is a popular discretisation technique in computer vision where it has been applied to mesh-based representations of a surface. FEM relies on subdividing the surface into a set of finite elements (e.g., triangles or tetrahedra) and approximates the continuous LBO by combining piecewise basis functions defined at the nodes of the elements. The previously described MolSG molecular shape method is based on the FEM approximation of the LBO and has been shown to be effective in capturing the shapes of molecules into a compact alignment-free descriptor.

While the FEM approach to approximating the LBO is appropriate for capturing shape, it is not trivial to extend it to include properties on a surface, as would be required for a combined shape and electrostatics descriptor. However, it has been shown that a manifold structure can also be approximated by a graph, with the graph Laplacian acting as a discrete counterpart which approximates the continuous LBO [42–45].

The graph Laplacian is built from the adjacency and degree matrix representations of a graph. The formula for the graph Laplacian is defined in Eq. (4), where L represents the Laplacian matrix, D represents the degree matrix and A represents the adjacency matrix.

$$L = D - A \tag{4}$$

The adjacency matrix A, of a graph $G = (V, E)$, which consists of a set of n vertices, V, and a set of e edges, E, is described in Eq. (5), where e_{ij} represents an edge between vertices, v_i and v_j, and the values of the elements within the adjacency matrix are one when the vertices are connected, and zero otherwise.

$$A = \begin{cases} 1 if e_{ij} \in E \\ 0 if e_{ij} \notin E \end{cases} \tag{5}$$

The degree matrix, D as seen in Eq. (6), captures the connectivity of the vertices within the graph and is defined as the diagonal matrix of vertex degrees. Note that ϑ_{ij} is the number of edges incident on a node, v_{ij}, in the graph, G.

$$D_{ij} = \begin{cases} \sum_{j=0}^{j=n} \vartheta_{ij} if i = j \\ 0 if i \neq j \end{cases} \tag{6}$$

Another relevant formulation of the graph Laplacian is described in Eq. (7), where W represents the kernel or affinity matrix which captures similarities between adjacent vertices rather than binary connectivity relationships present in A.

$$L = D - W \tag{7}$$

The elements of the affinity matrix, W, are defined in Eq. (8), where w_{ij} represents the similarity or affinity between connected vertices v_i and v_j, with unconnected vertices represented as zero. While both the adjacency matrix, A, and affinity matrix, W, encode graph connectivity, the affinity matrix additionally captures the strength of the relationship between neighbouring vertices.

$$W = \begin{cases} w_{ij} if e_{ij} \in E \\ 0 if e_{ij} \notin E \end{cases} \tag{8}$$

In manifold learning, pairwise distances between vertices can be converted into affinity measurements using kernel functions [42]. In this work, the affinity matrix W is constructed using the variable bandwidth kernel shown in Eq. (9), where f_i and f_j represent feature vectors associated with vertices v_i and v_j, respectively. The feature vectors may consist solely of spatial coordinates or may additionally include physicochemical properties such as electrostatic information. The terms τ_i and τ_j represent local scaling parameters derived from the neighbourhood surrounding each vertex and account for variations in local sampling density.

$$W\left(v_i, v_j\right) = exp\left(-\frac{d\left(\boldsymbol{f}_i - \boldsymbol{f}_j\right)^2}{\tau_i \tau_j}\right) \tag{9}$$

While the unnormalised graph Laplacian L effectively captures graph topology [46], the normalised graph Laplacian $\mathcal{L}$ shown in Eq. (10) has been reported to provide improved convergence to the analytical LBO [47], where D represents the degree matrix.

$$\mathcal{L} = D^{-1} L D^{-1} \tag{10}$$

Non-uniform sampling of the manifold can bias approximation of the continuous LBO, particularly in regions with dense or sparse local connectivity [48]. The normalised graph Laplacian $\mathcal{L}$ is known to reduce the influence of non-uniform sampling of data on the manifold.

The spectrum of the graph Laplacian is obtained by solving the eigenvalue problem shown in Eq. (11). This produces a series of eigenvalues and eigenvectors that capture intrinsic structural properties of the underlying graph. The spectrum can therefore be viewed as a reduced-dimensional representation of the fundamental structural properties of the graph encapsulated in $\mathcal{L}$.

$$\mathcal{L}\phi = \lambda\phi \tag{11}$$

The eigenvalues and eigenvectors of the normalised graph Laplacian can be considered to capture the scales and modes of vibration of resonant frequencies across the graph structure respectively [44], analogous to vibrational modes in a physical system. The resulting spectral representation is influenced by both the local connectivity of vertices and the weighting scheme used to define relationships between neighbouring vertices [49].

In this work, the graph Laplacian is subjected to an eigendecomposition to capture how information or signals vary and propagate across a graph representing a molecular surface. The capacity of the graph Laplacian operator to capture signal propagation across an underlying graph structure is fundamentally analogous to the convolution operation in traditional grid-based convolutional neural networks. Consequently, the graph Laplacian serves as the theoretical foundation for spectral graph convolutional neural networks, such as those studied by Defferrard et al. [50], who used truncated Chebyshev polynomials to map spectral filters directly onto the graph, inherently

capturing how the structural and topological features of a graph influence how a signal behaves and propagates across the graph without requiring a computationally costly eigendecomposition.

Following the work by Defferrard et al. [50], Kipf and Welling [51] demonstrated that restricting this Chebyshev expansion to a localized, first-order linear approximation transforms global spectral operations into highly efficient spatial ones. In [51], it is shown that graph convolutional neural networks can effectively capture graph structure and topology using these efficient first-order approximations of spectral filter functions. Ultimately, these localised operations mimic the behaviour of the low-frequency eigenfunctions of the graph Laplacian, which inherently capture smooth, coherent patterns over the local topology.

Graph construction

Belkin and Niyogi [42] describe two approaches for constructing a graph from a surface which is represented as a set of unconnected points p defined in $\mathbb{R}^n$ space. The first is the ε-neighbourhood graph method, in which points separated by less than a distance threshold ε are connected. This approach produces a symmetric adjacency matrix but may be sensitive to sparsity and non-uniform sampling.

The alternative method for graph construction is k-nearest neighbours, which connects a point x_i to its k nearest points. Compared with ε-neighbourhood graphs, k-NN graphs are generally more robust to sparsity and non-uniform sampling because each vertex is guaranteed to maintain a fixed number of local connections [42]. In this work, the k-NN approach was adopted for graph construction. As noted by Belkin and Niyogi [42], k-NN adjacency matrices are not necessarily symmetric (i.e., point x_i exists in the k-nearest neighbours of x_j but not vice versa). Symmetry was therefore enforced by averaging the adjacency matrix with its transpose prior to construction of the affinity matrix: (i.e., $\frac{A+A^T}{2}$ where A^T represents the transpose of the adjacency matrix A). This ensures that the resulting graph Laplacian is symmetric and therefore diagonalisable with real eigenvalues [52].

Surface patches

Surface patches provide complementary local geometric and chemical information that extends beyond point-wise surface descriptors. By capturing relationships between neighbouring surface features, the patch representation enables the descriptor to encode richer local surface environments and more complex molecular recognition patterns. In contrast to approaches such as SHOT [53] and MaSIF [39], which rely on handcrafted geometric distributions or geodesic convolutional neural networks, our method is based on the established Bag-of-Covariance matrices (BoC) framework.

Faraki et al. [38] demonstrated the effectiveness of region covariance matrices as descriptors within a BoC model, termed Bag-of-Riemannian Words, for image classification tasks. Tabia et al. [55] subsequently extended the BoC paradigm to the representation of 3D objects for shape matching and retrieval applications.

Given a feature matrix F of dimensionality $N \times d$, where N is the number of surface points within a surface patch and d is the dimensionality of the surface pointwise descriptors, the covariance between the i^{th} and j^{th} feature columns of F is defined in (12) below.

$$Cov(f_i,f_j) = \frac{1}{N}\sum_{k=1}^{N}\left(f_{i,k} - \overline{f_i}\right)\left(f_{j,k} - \overline{f_j}\right) \tag{12}$$

In Eq. (12), $f_{i,k}$ denotes the kth element of the ith feature column f_i, and $\overline{f_i}$ describes the mean of that feature column. Applying Eq. (12) to all pairwise combinations within the surface patch feature matrix, F produces a $d \times d$ covariance matrix, C, where each element C_{ij} represents the covariance between the ith and jth feature columns. As established, the resulting covariance matrix is symmetric positive definite (SPD), with diagonal elements corresponding to individual feature variances and off-diagonal elements capturing pairwise feature covariances [38].

One challenge associated with the BoC framework is that covariance matrices do not lie in Euclidean space, meaning that comparisons between covariance matrices require an appropriate Riemannian metric [56]. Tuzel et al. [57] provided an early demonstration of covariance matrices as region descriptors for applications including object detection and texture classification in computer vision.

Faraki [54] further developed the BoC methodology through the application of the Log-Euclidean transformation, which maps covariance matrices from the curved Riemannian manifold to a flat tangent space, thereby enabling the use of Euclidean distance measures. In our work, the same Log-Euclidean transformation is applied to embed a covariance matrix, C, into Euclidean space using the matrix logarithm. Given an SPD covariance matrix, C, the log-Euclidean mapping is defined via the matrix logarithm in Eq. (13).

$$\dot{C} = \log(C) \tag{13}$$

For a covariance matrix, C, the matrix logarithm is computed through the eigendecomposition of the SPD matrix which is defined in Eq. (14).

$$C = U \Lambda U^T \tag{14}$$

where U is an orthogonal matrix of eigenvectors and $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_d)$ is a diagonal matrix of eigenvalues which given that C is an SPD matrix, must contain strictly positive eigenvalues. The matrix logarithm of C is shown in Eq. (15).

$$\log(C) = U diag(\ln \lambda_1, \ln \lambda_2, \ldots, \ln \lambda_d) U^T \tag{15}$$

This transformation ensures that the mapped matrices defined by Eq. (13) reside in a vector space where conventional Euclidean operations, including vector quantisation for codebook construction, are mathematically valid. Additionally, as detailed in Faraki [54], given that the covariance matrix is symmetric, this means that $\dot{C}$ contains only $\frac{d(d+1)}{2}$ unique elements which can be isolated by extracting the upper triangular portion of $\dot{C}$ as a 1D vector, $Vec(\dot{C})$.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1186/s13321-026-01310-6.

Acknowledgements The authors thank R&D and R&D IT at AstraZeneca for supporting James Middleton's research project. The authors also acknowledge Pierre Ducrot for his contributions to the early experimental work that preceded the development of the graph-based methods presented in GE-MolSG.

Author contributions Conceptualization: JM, GMG, VJG, MJP, MZ; Methodology: JM, GMG, VJG, MJP, MZ; Software: JM, GMG; Investigation: JM; Formal analysis: JM; Data curation: JM; Visualization: JM, GMG; Writing, original draft: JM, VJG; Writing, review & editing: JM, GMG, VJG, MJP, MZ; Supervision: VJG, GMG, MJP, MZ; Funding acquisition: VJG.

Funding JM received funding from the EPSRC and AstraZeneca via an Industrial CASE studentship (EP/T517835/1).

Availability of data and materials The code and data supporting the implementation and validation of GE-MolSG and GM-MolSG are publicly available at https://github.com/SheffieldChemoinformatics/ge-molsg.

Declarations

Competing interests GMG is an editor of the *Journal of Cheminformatics*. JM was a PhD student at the University of Sheffield during the course of this work. JM, MJP are employees of AstraZeneca and hold stock options in the company. GMG holds stock options in AstraZeneca and Vertex Pharmaceuticals.

Open Access #### #### ####

#

26. Atwi R, Wang Y, Sciabola S, Antoszewski A (2024) Roshambo: open-source molecular alignment and 3d similarity scoring. J Chem Inf Model 64:8098–8104
27. Reuter M, Biasotti S, Giorgi D et al (2009) Discrete Laplace-Beltrami operators for shape analysis and segmentation. Comput Graph 33:381–390
28. Seddon MP (2017) The application of spectral geometry to 3D molecular shape comparison
29. Aubry M, Schlickewei U, Cremers D (2011) The wave kernel signature: a quantum mechanical approach to shape analysis. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops). IEEE, pp 1626–1633
30. Bronstein AM, Bronstein MM, Guibas LJ, Ovsjanikov M (2011) Shape Google: geometric words and expressions for invariant shape retrieval. ACM Trans Graph 30:1–20
31. Arthur D, Vassilvitskii S (2007) k-means++: The advantages of careful seeding. In: Soda. pp 1027–1035
32. Middleton J (2025) Combining shape and electrostatics in a spectral geometry alignment-free three-dimensional molecular descriptor. Ph.D. thesis University of Sheffield
33. Stein RM, Yang Y, Balius TE et al (2021) Property-unmatched decoys in docking benchmarks. J Chem Inf Model 61:699–714
34. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49:6789–6801
35. Réau M, Langenfeld F, Zagury J-F et al (2018) Decoys selection in benchmarking datasets: overview and perspectives. Front Pharmacol 9:11
36. Truchon J-F, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model 47:488–508
37. Yang W, Xu L, Chen X et al (2015) Chi-squared distance metric learning for histogram data. Math Probl Eng 2015:352849
38. Faraki M, Harandi M (2014) Bag of riemannian words for virus classification. Case Stud Intell Comput Achiev Trends 271–284
39. Gainza P, Sverrisson F, Monti F et al (2020) Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods 17:184–192
40. Castellanos M, MacDermott-Opeskin H, Ainsley J, Walters WP (2026) Lessons learned from the OpenADMET—ExpansionRx blind challenge. https://openadmet.ghost.io/lessons-learned-from-the-openadmet-expansionrx-blind-challenge/. Accessed 21 May 2026
41. Besl PJ, Jain RC (1986) Invariant surface characteristics for 3D object recognition in range images. Comput Vis Graph Image Process 33:33–80
42. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15:1373–1396
43. Belkin M, Sun J, Wang Y (2009) Constructing Laplace operator from point clouds in $\mathbb{R}$ d. In: Proceedings of the twentieth annual ACM-SIAM symposium on discrete algorithms. SIAM, pp 1031–1040
44. Lévy B (2006) Laplace-beltrami eigenfunctions towards an algorithm that "understands" geometry. In: IEEE international conference on shape modeling and applications 2006 (SMI'06). IEEE, p 13
45. García Trillos N, Gerlach M, Hein M, Slepčev D (2020) Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator. Found Comput Math 20:827–887
46. Dong X, Thanou D, Frossard P, Vandergheynst P (2016) Learning Laplacian matrix in smooth graph signal representations. IEEE Trans Signal Process 64:6160–6173
47. Von Luxburg U, Belkin M, Bousquet O (2008) Consistency of spectral clustering. Ann Stat 555–586
48. Xu G (2006) Discrete Laplace-Beltrami operator on sphere and optimal spherical triangulations. Int J Comput Geom Appl 16:75–93
49. Zhu X, Change Loy C, Gong S (2014) Constructing robust affinity graphs for spectral clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1450–1457
50. Defferrard M, Bresson X, Vandergheynst P (2016) Convolutional neural networks on graphs with fast localized spectral filtering. Adv Neural Inf Process Syst 29
51. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv:1609.02907
52. Cullum JK, Willoughby RA (2002) Lanczos algorithms for large symmetric eigenvalue computations: vol. I: theory. SIAM
53. Salti S, Tombari F, Di Stefano L (2014) SHOT: unique signatures of histograms for surface and texture description. Comput Vis Image Underst 125:251–264
54. Faraki M, Palhang M, Sanderson C (2015) Log-Euclidean bag of words for human action recognition. IET Comput Vis 9:331–339
55. Tabia H, Laga H (2015) Covariance-based descriptors for efficient 3D shape matching, retrieval, and classification. IEEE Trans Multimed 17:1591–1603
56. Förstner W, Moonen B (2003) A metric for covariance matrices. In: Grafarend EW, Krumm FW, Schwarze VS (eds) Geodesy-the challenge of the 3rd millennium. Springer Berlin Heidelberg, Berlin, pp 299–309
57. Tuzel O, Porikli F, Meer P (2006) Region covariance: a fast descriptor for detection and classification. In: European conference on computer vision. Springer, pp 589–600

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

James Middleton[1] · Gian Marco Ghiandoni[2] · Martin J. Packer[3] · Mengdie Zhuang[1] · Valerie J. Gillet[1]

✉ Valerie J. Gillet
v.gillet@sheffield.ac.uk

[1] School of Information, Journalism and Communication, University of Sheffield, The Wave, Sheffield S10 4AH, UK

[2] Augmented DMTA Platform, Data Analytics and AI, R&D IT, AstraZeneca, The Discovery Centre (DISC), Francis Crick Avenue, Cambridge CB2 0AA, UK

[3] Chemistry, Oncology, IMED Biotech Unit, AstraZeneca, Cambridge CB4 0WG, UK