

A Morse Transform for Drug Discovery

Alexander M. Tanaka  , Aras T. Asaad  , Richard I. Cooper   , and Vidit Nanda   

ABSTRACT. We introduce a new ligand-based virtual screening (LBVS) framework that uses piecewise-linear (PL) Morse theory to predict ligand binding potential. We model ligands as simplicial complexes via a pruned Delaunay triangulation and catalogue the critical points across multiple directional height-functions. This produces a rich feature vector, consisting of crucial topological features – peaks, troughs and saddles – that characterise ligand surfaces relevant to binding interactions. Unlike contemporary LBVS methods that rely on computationally intensive deep neural networks, we require only a lightweight classifier. The Morse-theoretic approach achieves state-of-the-art performance on standard datasets while offering an interpretable feature vector and scalable method for ligand prioritization in early-stage drug discovery.

1. Introduction

Computational screening has transformed drug discovery from a predominantly physical endeavour to a digitally-augmented one, accelerating the quest for new medicines by several orders of magnitude (Brogi, 2019).

1.1. Virtual Screening. Drugs typically modulate biological processes by interacting with specific regions, called *binding pockets*, on certain target proteins. Candidate drug molecules, known as *ligands* (Di Cera, 2020), may be identified through two primary virtual screening approaches: *structure-based* and *ligand-based*. Structure based methods require detailed knowledge of the target and use this knowledge to identify those ligand conformations that are likely to be active, i.e. to have a strong affinity for the known binding pocket (Maia et al., 2020). Ligand-based virtual screening (LBVS) methods, on the other hand, only use prior knowledge of known active molecules – either existing drugs or natural compounds – to identify other molecules with similar properties within large databases (Ripphausen et al., 2011). Structure-based virtual screening is more computationally intensive, but has the advantage of providing direct insight into the properties that determine active ligands. Conversely, LBVS is typically faster and may be used even in the absence of structural information about the target protein, but may suffer from a lack of data for new targets and early-stage projects.

Recent advances in LBVS have been dominated by sophisticated machine learning architectures (Sabe et al., 2021). Deep neural networks, in particular, have demonstrated

 Mathematical Institute, University of Oxford, Woodstock Road, OX2 6GG, UK.

 Oxford Drug Design Ltd, Blue Boar Court, 9 Alfred St, Oxford, OX1 4EH, UK.

 Inorganic Chemistry Laboratory, University of Oxford, South Parks Road, Oxford, OX1 3QR.

 Corresponding author. E-mail: nanda@maths.ox.ac.uk.

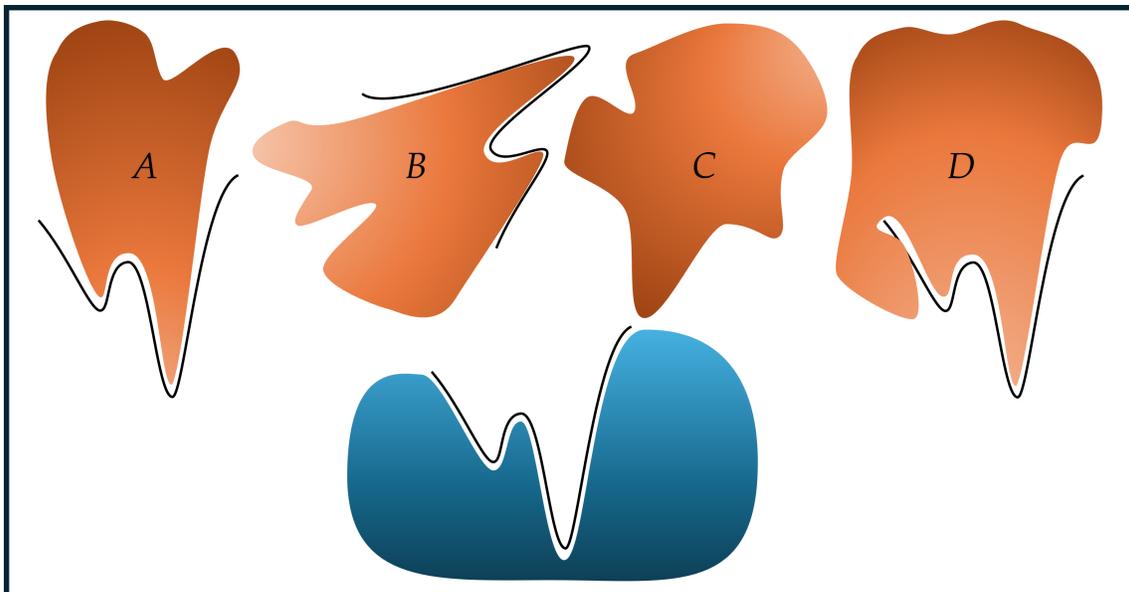


FIGURE 1. Two dimensional slices of a protein binding pocket along with four candidate ligands, A , B , C and D . Here A binds tightly with the target as drawn, while B binds after realignment. Both C and D are geometrically incompatible with the target. Note that the boundary of the binding pocket, drawn here as a W -shaped curve, must be (at least approximately) embodied in the surfaces of A and B for tight binding to be possible. Although this W -shaped region appears in the boundary of D , it is obstructed by the protrusion occurring on the left side.

impressive performance in the prediction of successfully binding ligands (Wu et al., 2024). There are two drawbacks ubiquitously encountered when using deep neural networks. First, they require large amounts of training data and computational resources. For most therapeutic targets and drug discovery projects, the number of experimentally validated active ligands is extremely small. This scarcity of data severely constrains the effectiveness of deep learning based approaches, particularly for novel or understudied protein targets. The second downside of deep neural networks is their inscrutability: the output they produce is the end-product of an enormous optimization procedure, and their reasoning remains difficult to decipher.

1.2. This Paper. Here we propose a new LBVS method which does not require prior knowledge of a large number of active ligands, and has the added benefit of employing an interpretable feature. The core idea is to map each ligand molecule to a vector of real numbers whose entries capture salient geometric and chemical properties. We describe the relevant geometry in Figure 1, which depicts a two-dimensional cartoon of a binding pocket along with four candidate ligands, labelled A through D . The key insight here is that the binding pocket must share an approximate (in the figure, W -shaped) boundary with the two well-fitting ligands. Ligand A fits tightly into the binding pocket as drawn, while ligand B also fits after a suitable rotation. On the other hand, ligands C and D do not fit at all for different reasons. In particular, C 's boundary does not have the required W -shape; and although D 's boundary does locally have the correct shape, this gets occluded

due to the protrusion towards its lower left side. In all cases, once the active ligand has been correctly aligned, its peaks and valleys will correspond to the peaks and valleys on the surface of the binding pocket.

An immediate advantage of thinking about binding potentials in terms of boundary compatibility is that it requires very few known active ligands to discover the correct binding shape for a given target. As long as the boundaries of A and B are sufficiently dissimilar from each other (away from the common W -shaped binding regions), there is hope that the knowledge of these two ligands alone may suffice when teaching a classifier to recognize the desired W -shape in new candidates. A second benefit, from a computational perspective, is that the signal we pursue in the ligand molecule is not global – only a small part of the boundary, probed to a (relatively) small depth, carries all of the desired information.

Conversely, there are three substantial difficulties that must be overcome before such geometric intuition can be translated to a practical and efficient LBVS pipeline. First, we must create a tractable *geometric model* of each ligand molecule, since ligands are typically presented as lists of atom types and locations or molecular graphs rather than smooth shapes. Having obtained such a model, the second challenge is to concoct a sufficiently discriminative *compressed representation* of the boundary region. Finally, we must solve the *alignment problem* – there is only a small part of the ligand boundary along which it may bind with the target, and we must discover the correct rotation that guarantees a tight fit. We address the first difficulty in a standard way, i.e. by modelling ligand molecules as Delaunay meshes built around the set of atom centres. There are only two modifications to keep in mind: we (a) weight each vertex with the van der Waal radius of the corresponding atom, and (b) discard all the simplices which contain a chemically irrelevant edge. An edge, for our purposes, is deemed irrelevant whenever its length exceeds the sum of van der Waal radii of the boundary vertices.

1.3. PL Morse Theory. If one pretends for a moment that the alignment problem has been solved, then piecewise-linear (PL) Morse theory (Kosiński, 1962; Brehm and Kühnel, 1987; Grunert et al., 2023) provides an excellent solution to our second difficulty. PL Morse theory can be used to extract a wealth of topological and geometric information from arbitrary simplicial complexes embedded in Euclidean space. Let $K \subset \mathbb{R}^n$ be such a complex; writing K_0 for its set of vertices, our starting point is a function $f : K_0 \rightarrow \mathbb{R}$. We call f a *PL Morse function* on K if it is injective on edges: in other words, we must have $f(v) \neq f(w)$ whenever the vertices v and w are connected by an edge of K . For each real number c , the *superlevel set* $K_{f \geq c}$ is the subcomplex of K consisting of all simplices whose vertices have f -value exceeding c . As we decrease the threshold c , the topology of $K_{f \geq c}$ evolves; the key observation is that this topology (namely, the homeomorphism type) only changes when we cross certain critical values of c . An analogous phenomenon occurs when dealing with a Morse function g defined on some smooth manifold X , as described in Milnor (1963): if we let $C(g)$ be the set of *critical points* $p \in X$ where the tangent space $T_x X$ lies in the orthogonal complement of the gradient vector $\nabla_x g$, then the superlevel set topology of X along g may change only at the set of critical values $\{g(x) \mid x \in C(g)\}$ (these are illustrated in Figure 2).

We have no recourse to gradient vectors and tangent spaces in PL Morse theory, since neither the underlying simplicial complex $K \subset \mathbb{R}^n$ nor the overlaid function $f : K_0 \rightarrow \mathbb{R}$ can be assumed to have any smooth structure. Nevertheless, there is a beautiful local description of critical vertices. Since f is required to be injective on edges, the neighbouring

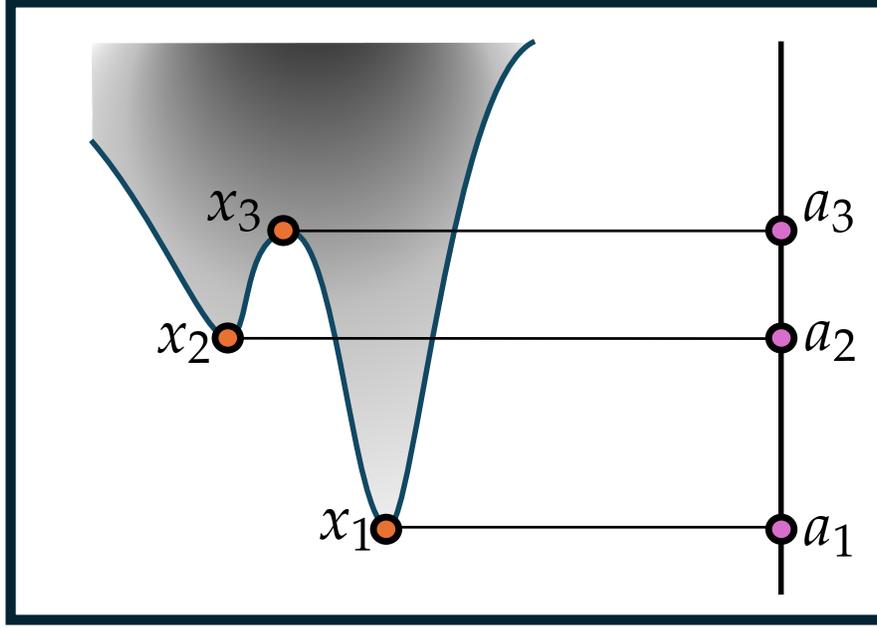


FIGURE 2. The three critical values a_1, a_2 and a_3 of the W -shaped boundary region from Figure 1 for the vertical height-function. These occur precisely at the heights of the critical points x_1, x_2 and x_3 where the tangent space is horizontal. Note that the boundaries of ligands A and B would Figure 1 would exhibit this critical value pattern along some direction. The ligand C has completely different critical values from all directions, and ligand D exhibits two additional critical values due to the obstructive protrusion in its boundary.

vertices of any given vertex $v \in K_0$ must have f -value either strictly greater or strictly smaller than $f(v)$. The *upper link* of v along f is the subcomplex of K spanned by all neighbouring vertices u for which $f(u) > f(v)$, as depicted in Figure 3. Now our vertex v is critical for f if and only if its upper link has non-trivial reduced Betti numbers¹. Thus, given real numbers $c > d$, the superlevel sets $K_{f \geq c}$ and $K_{f \geq d}$ can have different Betti numbers only when there is at least one critical vertex with f -value in the interval $[d, c]$.

Let $K \subset \mathbb{R}^n$ be an embedded simplicial complex, and select a vector ζ from the unit sphere $\mathbb{S}^{n-1} \subset \mathbb{R}^n$. We will denote the corresponding inner product height-function $x \mapsto \langle x, \zeta \rangle$ by $f_\zeta : K_0 \rightarrow \mathbb{R}$. A rich summary of the boundary of K in the direction ζ may be obtained by listing, for each critical vertex v of f_ζ , the vector of *Morse data*

$$\mu(v, \zeta) := [f_\zeta(v) \beta_{-1}(v, \zeta) \beta_0(v, \zeta) \cdots \beta_{n-2}(v, \zeta)]. \quad (1)$$

Here the first entry is the critical value $f_\zeta(v)$, while the remaining entries are the reduced Betti numbers of v 's upper link along f_ζ . This data would be an excellent combinatorial representation of K oriented along the direction ζ if all we cared to study were superlevelset homology. Since we seek a representation that is more aware of metric structure, the Morse

¹We defer the precise definition of Betti numbers (and more generally, of simplicial homology) to Appendix C; here we only remark that these numbers are efficiently computable (Kaczynski et al., 2004; Harker et al., 2014; Otter et al., 2017; The GUDHI Project, 2024).

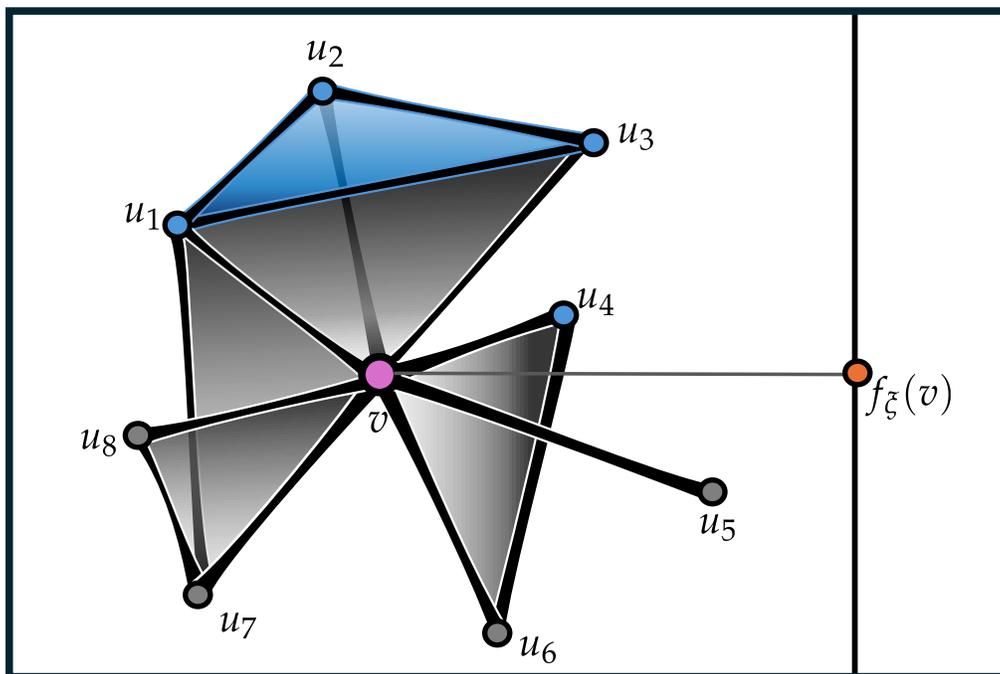


FIGURE 3. The neighbourhood of the vertex v in the illustrated simplicial complex has vertices $\{u_1, \dots, u_8\}$; its higher-dimensional simplices are $(u_1u_2u_3v)$, (u_1u_7v) , (u_4u_6v) , (u_5v) , (u_7u_8v) plus all their faces. The upper link of v with respect to the vertical height-function f_{ζ} is the subcomplex of this neighbourhood generated by those neighbours which are higher than v – explicitly, this is the blue region containing the 2-simplex $(u_1u_2u_3)$ plus all its faces along with the isolated vertex u_4 . The reduced Betti number of the upper link is non-trivial in dimension 0; therefore, v is a critical vertex for f_{ζ} . If we consider the same figure rotated clockwise by 90 degrees so that only $\{u_1, u_2, u_7, u_8\}$ were above v , then the upper link would have trivial Betti numbers and v would be non-critical for the corresponding height-function.

data leaves something to be desired. For instance, consider (a simplicial analogue of) the W -shaped region in Figure 2. One could deform the horizontal axis arbitrarily while preserving all critical points and their Morse data; and after such a deformation one could no longer expect the resulting shape to fit with the target from Figure 1.

1.4. The PL Morse Transform. Fortunately, the solution to this loss of metric information is the same as the solution to the alignment problem – consider more directions. Any horizontal deformation in Figure 2 would be immediately detected if we also were to examine Morse data along a slightly different direction $\zeta' \neq \zeta$. Moreover, by letting $\Xi \subset S^{n-1}$ be a sufficiently large collection of unit directions and collating the Morse data of height-functions along all $\zeta \in \Xi$, we are able to (at least partially) address the alignment problem. The ideal case $\Xi = S^{n-1}$, where we examine all possible directions, leads to deep classical results pertaining to the integral geometry of constructible sheaves and functions (Kashiwara and Schapira, 1990, Ch. IX). More recently, similar ideas have permeated topological data analysis in the guise of the *persistent homology transform* and related techniques for shape recognition (Turner et al., 2014; Munch, 2025).

For an arbitrary $\Xi \subset S^{n-1}$ and integer $d > 0$, we build a function from Ξ to the set of $d \times (n + 1)$ real matrices as follows. Given a $\xi \in \Xi$ with associated height-function $f_\xi : K_0 \rightarrow \mathbb{R}$, we sort the critical vertices $\{v_1, v_2, \dots, v_m\}$ of K along f_ξ in descending order of their critical values. Let us consider the matrix $\mathcal{M}_{K,d}(\xi)$ whose i -th row is the Morse data $\mu(v_i, \xi)$ from (1). We call this matrix-valued function $\xi \mapsto \mathcal{M}_{K,d}(\xi)$ defined on Ξ the **PL Morse transform** (of K , along Ξ , of depth d); it serves as the theoretical foundation upon which our feature vector is constructed.

1.5. The Feature Vector. Let us now return to the concrete setting where $K \subset \mathbb{R}^3$ is a pruned Delaunay complex corresponding to a ligand molecule. For computational reasons, we are forced to keep the set of directions Ξ finite; nevertheless, it is in our interest to distribute these directions uniformly across S^2 so that we may capture the geometry of K from several independent directions. To this end, in our experiments Ξ consists of the 32 directions corresponding to vertices of the pentakis dodecahedron, plus 68 directions chosen uniformly at random from S^2 . For each direction $\xi \in \Xi$, we let m_ξ denote the total number of critical vertices of the height-function f_ξ .

We begin by computing K 's Morse transform of depth $d = 20$ along Ξ – this amounts to a $m_\xi \times 4$ matrix assigned to each direction whose rows have the form of (1). Then, we extract nine column-wise percentiles of these matrix entries across all $\xi \in \Xi$, resulting in a thirty-six dimensional vector. Additionally, we record the lipophilicity, molar refractivity and partial charge of each atom in the ligand inside the multisets \mathcal{L} , \mathcal{R} and \mathcal{Q} , respectively. Thirty-six further numbers are obtained by computing the same earlier percentiles of the collection $\{m_\xi \mid \xi \in \Xi\}$, \mathcal{L} , \mathcal{R} and \mathcal{Q} , resulting at last in our 72-dimensional Morse feature vector of the ligand represented by K .

1.6. Classifier. We train a Light Gradient Boosting Machine (LGBM, [Ke et al., 2017](#)) for binary classification on the Morse features and features from the literature. We tune and evaluate the classifier using 5-fold cross-validation (for details, see Appendix I). The final evaluation metric scores are the mean scores across all the 5-folds.

2. Results

We evaluate the Morse feature vectorisation on two widely used benchmark datasets for virtual screening: the Directory of Useful Decoys, Enhanced (DUD-E, [Mysinger et al., 2012](#)) and the Maximum Unbiased Validation (MUV, [Rohrer and Baumann, 2009](#)) dataset. DUD-E consists of active ligands for 102 protein targets, accompanied by property-matched decoy molecules that resemble the ligands physically but are dissimilar in 2-D chemical fingerprint space to minimise the chance of binding. The DUD-E Diverse (D8) subset is a subset of DUD-E consisting of 8 targets that are representative of the diverse protein categories in DUD-E. Similarly, MUV consists of active ligands and decoys for 17 protein targets. MUV was designed to not be affected by artificial enrichment or analogue bias by ensuring that actives are close to decoys in simple chemical descriptor space.

We compare the performance of a LGBM classifier trained on different non-superpositional LBVS features to distinguish actives from decoys. We evaluate two versions of our features: 36-dimensional Morse features at depth 20 and 100 directions without the chemical percentiles (M); and the 72-dimensional chemistry-enhanced Morse features at depth 20 and 32 directions (M+C). We also test a 27-dimensional baseline feature consisting of nine percentiles of each 3D ordinate of the atoms of the randomly rotated molecule (R3P). We

compare against the following shape-based features: Ultrafast Shape Recognition (USR, [Ballester and Richards, 2007](#)); and the unweighted subset of Weighted Holistic Invariant Molecular descriptors (Wu, [Todeschini and Gramatica, 1997](#)). Additionally, we compare against the following hybrid shape- and chemistry-based features: Ultrafast Shape Recognition with CREDO Atom Types (UCT, [Schreyer and Blundell, 2012](#)); and the full set of Weighted Holistic Invariant Molecular descriptors (W).

2.1. The DUD-E database. Table 1 displays the mean and per target performance of tuned LGBM classifiers trained on a variety of different LBVS features generated from the D8 subset. Morse features have the highest mean AUROC of 0.84 ± 0.08 of all shape-based features, though for certain D8 targets (GCR and KIF11) Wu features have higher mean AUROC scores. The second-best shape-based feature is Wu scoring a mean AUROC of 0.81 ± 0.10 . Chemistry-enhanced Morse features achieve the highest mean AUROC score of 0.97 ± 0.03 of all features. UCT is the second-best overall feature with a mean AUROC of 0.92 ± 0.07 .

TABLE 1. The mean AUROC per D8 subset target and the overall mean AUROC of various LBVS methods. The method with the highest AUROC per target is displayed in bold and the highest shape-based method is underlined.

Target	Shape				Shape & Chemistry		
	M	R3P	USR	Wu	M+C	UCT	W
AKT1	<u>0.87</u>	0.65	0.77	0.82	0.99	0.98	0.92
AMPC	<u>0.80</u>	0.56	0.74	0.79	0.94	0.88	0.84
CP3A4	<u>0.66</u>	0.52	0.53	0.60	0.92	0.79	0.72
CXCR4	<u>0.85</u>	0.61	0.81	0.78	0.99	0.93	0.87
GCR	0.86	0.65	0.83	<u>0.87</u>	0.99	0.94	0.95
HIVPR	<u>0.93</u>	0.71	0.83	0.90	0.99	0.96	0.96
HIVRT	<u>0.82</u>	0.63	0.75	0.78	0.96	0.90	0.85
KIF11	0.90	0.67	0.87	<u>0.91</u>	0.98	0.98	0.96
mean	<u>0.84</u>	0.62	0.76	0.81	0.97	0.92	0.88
SD	0.08	0.08	0.10	0.10	0.03	0.07	0.08

2.2. The MUV database. Table 2 displays the performance of tuned LGBM classifiers trained on a variety of different LBVS features generated from the MUV dataset. The per target performance is also shown for the whole dataset. Morse features have the highest mean AUROC of 0.64 ± 0.11 of all shape-based features and the highest mean AUROC across 12/17 targets. Wu features are the second-best shape-based feature with a mean AUROC of 0.61 ± 0.13 and the highest mean AUROC across 7/17 targets (achieving the highest score of all types of features for the 692 target). Chemistry-enhanced Morse features achieve the highest mean AUROC score of 0.74 ± 0.12 of all features and the highest mean AUROC across 14/17 targets. Notably, W features and Morse features are the joint second-best method despite Morse features containing no explicit chemical information.

TABLE 2. The mean AUROC per MUV target and the overall mean AUROC of various LBVS methods. For each target the method with the highest AUROC is displayed in bold and the highest shape-based method is underlined.

Target	Shape				Shape & Chemistry		
	M	R3P	USR	Wu	M+C	UCT	W
466	<u>0.58</u>	0.54	0.56	0.56	0.60	0.64	0.61
548	<u>0.72</u>	0.57	0.62	0.65	0.85	0.74	0.69
600	0.58	0.59	0.58	<u>0.62</u>	0.59	0.59	0.64
644	<u>0.70</u>	0.57	0.59	<u>0.70</u>	0.83	0.83	0.75
652	<u>0.63</u>	0.57	0.57	0.43	0.71	0.57	0.60
689	<u>0.62</u>	0.62	0.46	0.45	0.75	0.57	0.44
692	0.64	0.49	0.62	<u>0.74</u>	0.64	0.59	0.72
712	<u>0.69</u>	0.51	0.61	0.62	0.74	0.63	0.71
713	<u>0.67</u>	0.66	0.65	0.57	0.74	0.70	0.72
733	<u>0.64</u>	0.45	0.44	0.53	0.66	0.45	0.59
737	<u>0.71</u>	0.50	0.63	0.65	0.83	0.65	0.67
810	<u>0.64</u>	0.43	0.46	0.62	0.77	0.48	0.61
832	0.58	0.43	0.52	<u>0.59</u>	0.79	0.81	0.53
846	<u>0.78</u>	0.47	0.45	0.73	0.89	0.70	0.81
852	0.64	0.55	0.67	<u>0.70</u>	0.81	0.73	0.67
858	0.52	0.49	0.57	<u>0.73</u>	0.76	0.60	0.69
859	<u>0.54</u>	0.52	0.53	<u>0.54</u>	0.62	0.41	0.50
mean	<u>0.64</u>	0.53	0.56	0.61	0.74	0.63	0.64
SD	0.11	0.10	0.12	0.13	0.12	0.14	0.14

3. Discussion

Performance. On both D8 and MUV, Morse features achieve the highest performance among the shape-based features and chemistry-enhanced Morse features are the best performing of all the evaluated features. Chemistry-enhanced Morse features score significantly better than vanilla Morse features, demonstrating the importance of chemical information for accurately identifying active ligands. Indeed, the average hybrid shape- and chemistry-based feature outperforms the average pure shape-based feature with the best shape-based feature being at best equal to the worst hybrid feature on MUV. Without exception, all features perform better on the D8 subset than the MUV dataset, confirming that MUV is a more challenging dataset. Tellingly, the baseline R3P has almost random performance on MUV (AUROC of 0.53 ± 0.10) whilst having some predictive power on D8 (AUROC of 0.62 ± 0.08), suggesting that MUV is also less biased in shape space.

Robustness. For Morse features, increasing both the depth and the number of directions improves the generalisation error with diminishing returns (Figure 4). On the D8 subset, the performance plateaus around a depth of 13 and 32 directions. The lack of any significant local peaks in the performance indicates that the features are well-behaved with respect to the depth and number of directions. Reassuringly, this means that these parameters do not have to be optimised or tuned to find a local maxima in performance –

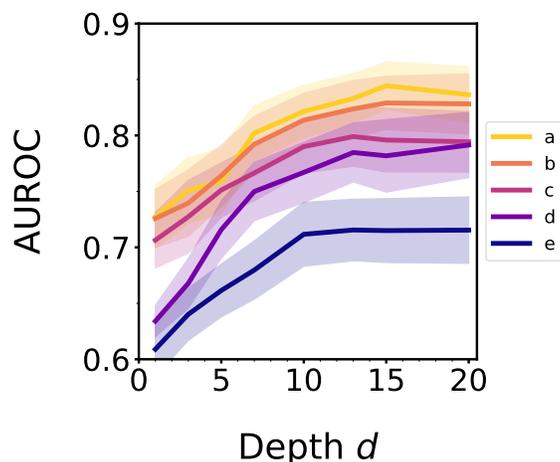


FIGURE 4. The mean AUROC score against depth of our LGBM classifier trained on Morse feature vectors computed using 100 directions (a), 32 pentakis dodecahedral directions (b), 12 icosahedral directions (c), 8 cubic directions (d) and 1 direction (e) for the D8 subset. Error bars are 95% confidence intervals.

just selecting a reasonably high value should be sufficient to extract most of the possible performance. This behaviour is expected as probing a molecule from more directions increases the chance of identifying salient geometric features, and after sufficiently many directions the sample percentiles of the components of the Morse feature should tend to the population percentiles of the Morse feature for an infinite number of directions.

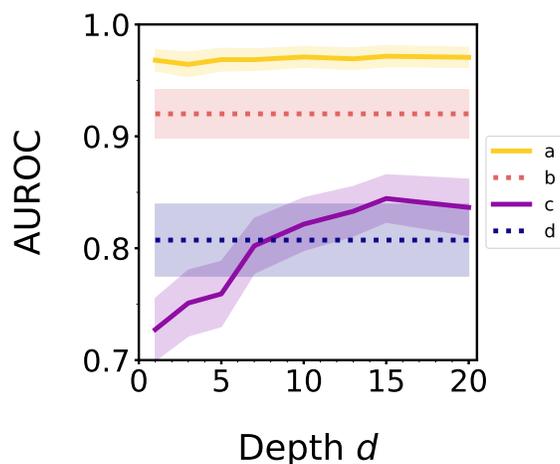


FIGURE 5. The mean AUROC score against depth of the LGBM classifier trained on chemically-enhanced Morse feature vectors computed using 32 directions (a) and Morse feature vectors computed using 100 directions (c) for the D8 subset. For comparison, the best-performing external shape- and chemistry-based feature UCT (b) and the best performing external shape-based feature Wu (d) are plotted with dotted lines. Error bars are 95% confidence intervals.

The performance of chemistry-enhanced Morse features is relatively unaffected by the depth unlike vanilla Morse features (Figure 5). For all depths, chemistry-enhanced Morse features dominate the performance of the second-best feature whilst Morse features only surpass the performance of the second-best shape-based method at a depth of around 10. This implies that the chemistry has a larger affect than the geometry on the performance in virtual screening datasets as the chemical components of the chemistry-enhanced Morse feature are the only components independent of depth.

Modifying the Morse transform to record the Morse data of the critical *and* non-critical vertices or randomly selecting vertices degrades the performance of the feature (see Figures N3 and N4), which empirically justifies the Morse-theoretic approach of focusing on critical vertices.

Acknowledgements

AMT thanks Paul Finn for helpful discussions. AMT thanks the EPSRC Centre for Doctoral Training in *Sustainable Approaches to Biomedical Science: Responsible and Reproducible Research* (grant number EP/S024093/1) for his studentship. VN's work is partially supported by the EPSRC under Grant EP/R018472/1 and in part by US AFOSR under Grant FA9550-22-1-0462. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility (<http://dx.doi.org/10.5281/zenodo.22558>).

References

- F. Aurenhammer and R. Klein. Voronoi diagrams. *Handbook of Computational Geometry*, pages 201–290, 2005. 14
- P. J. Ballester and W. G. Richards. Ultrafast shape recognition to search compound databases for similar molecular shapes. *Journal of Computational Chemistry*, 28(10):1711–1723, 2007. doi: <https://doi.org/10.1002/jcc.20681>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20681>. 7
- M. Bestvina and N. Brady. Morse theory and finiteness properties of groups. *Inventiones Mathematicae*, 129(3):445–470, Aug. 1997. doi: 10.1007/s002220050168. 15
- U. Brehm and W. Kühnel. Combinatorial manifolds with few vertices. *Topology*, 26(4):465–473, 1987. ISSN 0040-9383. doi: [https://doi.org/10.1016/0040-9383\(87\)90042-5](https://doi.org/10.1016/0040-9383(87)90042-5). URL <https://www.sciencedirect.com/science/article/pii/0040938387900425>. 3, 15
- S. Brogi. Computational approaches for drug discovery. *Molecules*, 24(17), 2019. ISSN 1420-3049. doi: 10.3390/molecules24173061. URL <https://www.mdpi.com/1420-3049/24/17/3061>. 1, 14
- Chemical Computing Group (CCG). Molecular operating environment. <https://www.chemcomp.com/>, 2025. [Accessed 03-03-2025]. 14
- A. E. Cleves, S. R. Johnson, and A. N. Jain. Electrostatic-field and surface-shape similarity for virtual screening and pose prediction. *J Comput Aided Mol Des*, 33(10):865–886, 10 2019. 20
- E. Di Cera. Mechanisms of ligand binding. *Biophysics Reviews*, 1(1):011303, 11 2020. ISSN 2688-4089. doi: 10.1063/5.0020997. URL <https://doi.org/10.1063/5.0020997>. 1, 14
- J.-P. Ebejer, G. M. Morris, and C. M. Deane. Freely available conformer generation methods: How good are they? *Journal of Chemical Information and Modeling*, 52(5):1146–1158, 2012. doi: 10.1021/ci2004658. URL <https://doi.org/10.1021/ci2004658>. PMID: 22482737. 16

- H. Edelsbrunner. Triangulations and meshes in computational geometry. *Acta Numerica*, 9: 133–213, 2000. doi: 10.1017/S0962492900001331. 15
- R. Grunert, W. Kühnel, and G. Rote. Pl Morse theory in low dimensions. *Advances in Geometry*, 23(1):135–150, Jan. 2023. ISSN 1615-715X. doi: 10.1515/advgeom-2022-0027. URL <http://dx.doi.org/10.1515/advgeom-2022-0027>. 3, 15
- S. Harker, K. Mischaikow, M. Mrozek, and V. Nanda. Discrete Morse theoretic algorithms for computing homology of complexes and maps. *Foundations of Computational Mathematics*, 14(1):151 – 184, 2014. 4
- C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020. doi: 10.1038/s41586-020-2649-2. URL <https://doi.org/10.1038/s41586-020-2649-2>. 25
- A. Hatcher. *Algebraic Topology*. Cambridge University Press, 2002. 15
- J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55. 25
- S. Jung, H. Vatheuer, and P. Czodrowski. Vsflow: an open-source ligand-based virtual screening tool. *Journal of Cheminformatics*, 15(1):40, Mar 2023. ISSN 1758-2946. doi: 10.1186/s13321-023-00703-1. URL <https://doi.org/10.1186/s13321-023-00703-1>. 20
- T. Kaczynski, K. Mischaikow, and M. Mrozek. *Computational Homology*. Springer, 2004. 4
- M. Kashiwara and P. Schapira. *Sheaves on Manifolds*. Springer Verlag, 1990. 5
- G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf. 6, 25
- D. R. Koes and C. J. Camacho. Shape-based virtual screening with volumetric aligned molecular shapes. *Journal of Computational Chemistry*, 35(25):1824–1834, 2014. doi: <https://doi.org/10.1002/jcc.23690>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.23690>. 20
- A. A. Kosiński. Singularities of piecewise linear mappings. i mappings into the real line. *Bulletin of the American Mathematical Society*, 68:110–114, 1962. URL <https://api.semanticscholar.org/CorpusID:45347127>. 3
- F. Li, K. Fujiwara, F. Okura, and Y. Matsushita. A closer look at rotation-invariant deep point cloud analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16218–16227, October 2021. 17
- R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018. 25
- E. H. B. Maia, L. C. Assis, T. A. De Oliveira, A. M. Da Silva, and A. G. Taranto. Structure-based virtual screening: from classical to artificial intelligence. *Frontiers in chemistry*, 8: 343, 2020. 1
- J. Milnor. *Morse theory*. Princeton University Press, Princeton, N.J, 1963. ISBN 0691080089.

- C. Mirabello and B. Wallner. Interlig: improved ligand-based virtual screening using topologically independent structural alignments. *Bioinformatics*, 36(10):3266–3267, 02 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa089. URL <https://doi.org/10.1093/bioinformatics/btaa089>. 20
- E. Munch. An invitation to the euler characteristic transform. *The American Mathematical Monthly*, 132(1):15–25, 2025. doi: 10.1080/00029890.2024.2409616. URL <https://doi.org/10.1080/00029890.2024.2409616>. 5
- M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem*, 55(14):6582–6594, Jul 2012. 6
- N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(17), 2017. 4
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 24
- Plotly Technologies Inc. Collaborative data science, 2015. URL <https://plot.ly>. 25
- S. Puertas-Martín, J. L. Redondo, P. M. Ortigosa, and H. Pérez-Sánchez. Optipharm: An evolutionary algorithm to compare shape similarity. *Scientific Reports*, 9(1):1398, Feb 2019. ISSN 2045-2322. doi: 10.1038/s41598-018-37908-6. URL <https://doi.org/10.1038/s41598-018-37908-6>. 20
- Rational Discovery LLC, G. Landrum, and J. Penzotti. Rdkit: Open-source cheminformatics, 2024. URL <https://www.rdkit.org>. 24
- P. Ripphausen, B. Nisius, and J. Bajorath. State-of-the-art in ligand-based virtual screening. *Drug Discovery Today*, 16(9):372–376, 2011. ISSN 1359-6446. doi: <https://doi.org/10.1016/j.drudis.2011.02.011>. URL <https://www.sciencedirect.com/science/article/pii/S1359644611000626>. 1, 14
- S. G. Rohrer and K. Baumann. Maximum unbiased validation (muv) data sets for virtual screening based on pubchem bioactivity data. *Journal of Chemical Information and Modeling*, 49(2):169–184, 2009. doi: 10.1021/ci8002649. URL <https://doi.org/10.1021/ci8002649>. PMID: 19161251. 6
- V. T. Sabe, T. Ntombela, L. A. Jhamba, G. E. Maguire, T. Govender, T. Naicker, and H. G. Kruger. Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *European Journal of Medicinal Chemistry*, 224:113705, 2021. 1, 14
- A. M. Schreyer and T. Blundell. Usrcat: real-time ultrafast shape recognition with pharmacophoric constraints. *Journal of Cheminformatics*, 4(1):27, Nov 2012. ISSN 1758-2946. doi: 10.1186/1758-2946-4-27. URL <https://doi.org/10.1186/1758-2946-4-27>. 7
- M. P. Seddon, D. A. Cosgrove, M. J. Packer, and V. J. Gillet. Alignment-free molecular shape comparison using spectral geometry: The framework. *Journal of Chemical Information and Modeling*, 59(1):98–116, 2019. doi: 10.1021/acs.jcim.8b00676. URL <https://doi.org/10.1021/acs.jcim.8b00676>. PMID: 30462505. 20
- E. Spanier. *Algebraic Topology*. Springer, 1966. 14
- The GUDHI Project. *GUDHI User and Reference Manual*. GUDHI Editorial Board, 3.10.1 edition, 2024. URL <https://gudhi.inria.fr/doc/3.10.1/>. 4, 24
- The Pandas Development Team. `pandas-dev/pandas`: Pandas, Feb. 2020. URL <https://doi.org/10.5281/zenodo.3509134>. 25

- P. Tiikkainen, P. Markt, G. Wolber, J. Kirchmair, S. Distinto, A. Poso, and O. Kallioniemi. Critical comparison of virtual screening methods against the muv data set. *Journal of Chemical Information and Modeling*, 49(10):2168–2178, 2009. doi: 10.1021/ci900249b. URL <https://doi.org/10.1021/ci900249b>. PMID: 19799417. 20
- R. Todeschini and P. Gramatica. Sd-modelling and prediction by whim descriptors. part 5. theory development and chemical meaning of whim descriptors. *Quantitative Structure-activity Relationships*, 16:113–119, 1997. URL <https://api.semanticscholar.org/CorpusID:95530317>. 7
- J.-F. Truchon and C. I. Bayly. Evaluating virtual screening methods: Good and bad metrics for the early recognition problem. *Journal of Chemical Information and Modeling*, 47(2):488–508, 2007. doi: 10.1021/ci600426e. URL <https://doi.org/10.1021/ci600426e>. PMID: 17288412. 18
- K. Turner, S. Mukherjee, and D. M. Boyer. Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA*, 3(4):310–344, 12 2014. ISSN 2049-8764. doi: 10.1093/imaiai/iau011. URL <https://doi.org/10.1093/imaiai/iau011>. 5
- D. Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005. URL <https://doi.org/10.1021/ci00057a005>. 16
- S. A. Wildman and G. M. Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, 1999. doi: 10.1021/ci9903071. URL <https://doi.org/10.1021/ci9903071>. 14
- H. Wu, J. Liu, R. Zhang, Y. Lu, G. Cui, Z. Cui, and Y. Ding. A review of deep learning methods for ligand based drug virtual screening. *Fundamental Research*, 2024. 2

Appendix A. Virtual screening

To a large extent, the biological activity of a ligand depends upon how well it binds to a target protein [Di Cera \(2020\)](#); this binding affinity is a complicated function of the geometric and chemical properties of both molecules. *Ligand-based virtual screening* (LBVS) is a high throughput, in-silico method for selecting molecules with high binding affinity from a dataset, using prior information about drugs which are known to bind successfully with the target ([Ripphausen et al., 2011](#)). The effectiveness of virtual screening can have a substantial contribution to the success of drug discovery projects ([Brogi, 2019](#); [Sabe et al., 2021](#)).

A crucial aspect of LBVS is the process used to measure the similarity of a candidate ligand to known actives. Here we describe a new method for comparing molecular geometry based on Morse theory. The chemical aspects considered here, on the other hand, are more standard and well-known. Explicitly, besides knowledge of the atom locations in 3D and their van der Waals radii, we only make use of the following chemical properties:

- (1) the MMFF94 modified *partial charge* as implemented in MOE ([Chemical Computing Group, CCG](#)), which accounts for the asymmetric distribution of electrons in the chemical bonds of a molecule;
- (2) the atomic contribution to Wildman-Crippen *molar refractivity* ([Wildman and Crippen, 1999](#)), which is a measure of the polarizability of the molecule; and
- (3) the atomic contribution to the Wildman-Crippen *lipophilicity* ([Wildman and Crippen, 1999](#)), which measures the equilibrium distribution of the molecule between a non-polar and a polar solvent.

In particular, we do not assume any prior knowledge of the target protein binding site, unlike in structure-based virtual screening.

Appendix B. Molecules as simplicial complexes

There are several reasonable discrete models for representing molecular data; among the most viscerally geometric is the *union of balls*, where one constructs balls (of van der Waals radius) around atom centres. This representation is visually appealing, but rather awkward from a computational perspective – checking whether or not a point in this union lies on the boundary is already a cumbersome task. It is therefore customary to represent a given molecule as a simplicial complex ([Spanier, 1966](#), Ch. 3.1) whose vertices correspond to atom centres.

For the purposes of building such a simplicial model, we assume access to the finite subset P of Euclidean space \mathbb{R}^3 consisting of atom-centres of a given molecule, and the function $w : P \rightarrow \mathbb{R}$ that sends each atom p to the corresponding van der Waals radius w_p . The weighted distance of a point $x \in \mathbb{R}^3 \setminus P$ to p is the (possibly negative) real number $d_w(x, p) := \|p - x\|^2 - w_p^2$, where $\|\bullet\|$ denotes the standard Euclidean norm. The *weighted Voronoi cell* of $p \in P$ is the subset $V(p) \subset \mathbb{R}^3$ given by

$$V_w(p) := \{x \notin P \mid d_w(x, p) \leq d_w(x, q) \text{ for all } p \neq q \in P\}.$$

Explicitly, this consists of all points in $\mathbb{R}^3 \setminus P$ which admit p as a nearest neighbour (in P , with respect to d_w). Each $V_w(p)$ is a closed convex set, and the collection of $\{V_w(p) \mid p \in P\}$ forms a regular cell decomposition of \mathbb{R}^3 – see [Aurenhammer and Klein \(2005\)](#) for details.

When P is in general position, the dual Voronoi cellulation forms a simplicial complex, which is called the *weighted Delaunay triangulation* (Edelsbrunner, 2000) of P and denoted $D_w(P) \subset \mathbb{R}^3$. More precisely: the vertex set of $D_w(P)$ is P , and there is a k -simplex for $k \in \{1, 2, 3\}$ spanning $\{p_0, p_1, \dots, p_k\}$ whenever the corresponding weighted Voronoi cells have nonempty intersection, i.e., when $\bigcap_{j=0}^k V_w(p_j) \neq \emptyset$. It is possible for simplices in $D_w(P)$ to contain edges which are far too long to accurately reflect the underlying molecular geometry. We therefore remove every simplex from $D_w(P)$ which contains an edge $\{p, q\}$ whose length $\|p - q\|$ exceeds the sum $w_p + w_q$ of van der Waals radii. The resulting simplicial subcomplex $K_w(P) \subset D_w(P)$, which we call the **pruned Delaunay triangulation** of P , serves as a convenient geometric representation of each given molecule.

Appendix C. Simplicial homology

To each finite n -dimensional simplicial complex K , one can associate a sequence of vector spaces called (*reduced*) *homology groups* as described below. We impose an arbitrary ordering on the vertices, so that each k -simplex σ is uniquely expressible as a tuple of vertices (v_0, v_1, \dots, v_k) written in ascending order. The faces of σ inherit this ordering: for each i in $\{0, \dots, \dim \sigma\}$, we let σ_{-i} denote the face of σ obtained by removing v_i .

We write C_k for the real vector space obtained by treating the k -simplices of K as a basis. The k -th *boundary operator* is the linear map $\partial_k : C_k \rightarrow C_{k-1}$ whose action on a basis k -simplex $\sigma \in K$ is given by $\partial_k(\sigma) = \sum_{i=0}^k (-1)^i \sigma_{-i}$. Let us also define $\partial_0 : C_0 \rightarrow \mathbb{R}$ as the map which sends each basis vertex to 1. Thus, we have a descending sequence of vector spaces and linear maps:

$$\dots \xrightarrow{\partial_{k+2}} C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} \dots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} \mathbb{R} \xrightarrow{\partial_{-1}} 0.$$

A routine calculation confirms that the composite $\partial_k \circ \partial_{k+1}$ is the zero map for all $k \geq -1$; or equivalently, the kernel of ∂_k contains the image of ∂_{k+1} . The quotient vector space $\tilde{H}_k(K) := \ker \partial_k / \text{img } \partial_{k+1}$ is called the **k -th reduced homology group** of K ; and its dimension, denoted $\tilde{\beta}_k(K)$, is called the k -th *reduced Betti number* of K . Homology groups satisfy several remarkable properties, including homotopy-invariance, functoriality, and efficient computability (Hatcher, 2002, Ch. 2). For our purposes here, it satisfies to note that the reduced Betti numbers of the one-point space are all zero, and that $\tilde{\beta}_{-1}(X)$ vanishes if and only if X is nonempty. We say that K is *acyclic* whenever $\tilde{\beta}_k(K) = 0$ holds for all k .

Appendix D. Piecewise-linear Morse theory

$K \subset \mathbb{R}^n$ be a finite simplicial complex and denote the set of its vertices by K_0 . An assignment $f : K_0 \rightarrow \mathbb{R}$ is called a *piecewise-linear (or, PL) Morse function* on K if it is injective on every simplex (Brehm and Kühnel, 1987; Bestvina and Brady, 1997; Grunert et al., 2023) – equivalently, if $f(v) \neq f(w)$ holds whenever $\{v, w\}$ forms an edge of K . Let us fix a PL Morse function $f : K_0 \rightarrow \mathbb{R}$; for each real number c we define the *superlevelset* $K^{\geq c}$ as the subcomplex of K spanned by all simplices σ whose vertices $v \in \sigma$ satisfy $f(v) \leq c$. PL Morse theory aims to explicitly describe how the homology groups of $K^{\geq c}$ evolve as a function of c .

The *upper link* of a vertex v with respect to f , denoted $L_f^+(v)$, is defined as the (possibly empty) simplicial subcomplex of K generated by all vertices $w \in K_0$ such that $\{v, w\}$ is an edge of K and $f(w) > f(v)$. It is straightforward to check that $L_f^+(v)$ has dimension at most

$n - 1$. Given a pair of real numbers $c > d$, note that we automatically have a containment $K^{\geq c} \subset K^{\geq d}$. The **link criterion** of PL Morse theory is as follows: if every vertex $v \in K_0$ with $c \leq f(v) \leq d$ has an acyclic $L_f^+(v)$, then the homology groups of $K^{\geq c}$ coincide with those of $K^{\geq d}$. Thus, the superlevelset homology can only change across those $c \in \mathbb{R}$ satisfying $f(v) = c$ for some vertex v whose upper link is *not* acyclic. These special vertices v are called the *critical points* of f , and the corresponding real numbers $c = f(v)$ are the *critical values* of f .

Appendix E. The PL Morse transform

Consider a simplicial complex $K \subset \mathbb{R}^n$. For each direction vector ξ lying on the unit sphere $S^{n-1} \subset \mathbb{R}^n$, let $f_\xi : K_0 \rightarrow \mathbb{R}$ be the inner product map $v \mapsto \langle v, \xi \rangle$. Under generic conditions, the map f_ξ is a PL Morse function for which no two critical points have the same critical value. Let $\Xi \subset S^{n-1}$ be any subset of generic unit vectors for K , and fix a positive integer d which is smaller than the number of critical points of f_ξ for each $\xi \in \Xi$. We may therefore order, the top d critical values of f_ξ as $c_1^\xi > \dots > c_d^\xi$, and write v_i^ξ for the unique critical point with value c_i^ξ . For brevity, the j -th reduced Betti number of the upper link $L_{f_\xi}^+(v_i^\xi)$ will be denoted $\tilde{\beta}(\xi)_j^i$. A routine calculation confirms that $\tilde{\beta}(\xi)_j$ is zero whenever $j \geq n - 1$.

The **Morse transform** of K (along Ξ , of depth d) is a function $\mathcal{M}_{K,d} : \Xi \rightarrow \text{Mat}_{\mathbb{R}}(d, n + 1)$ that sends each unit vector of Ξ to a certain $d \times (n + 1)$ matrix of real numbers. The i -th row of this matrix catalogues relevant Morse data of f_ξ at the i -th critical point p_i^ξ ; explicitly, it is given by

$$\mathcal{M}_{K,d}(\xi)_i := \left[c_i^\xi \quad \tilde{\beta}(\xi)_{-1}^i \quad \tilde{\beta}(\xi)_0^i \quad \dots \quad \tilde{\beta}(\xi)_{n-2}^i \right].$$

Thus, the first column of $\mathcal{M}_{K,d}(\xi)$ contains the top d critical values of f_ξ in descending order, while the remaining columns, which are all integer-valued, record reduced Betti numbers of the corresponding upper links.

Appendix F. Dataset preparation

We use a pipeline based on RDKit to curate and prepare DUD-E and MUV molecules for descriptor generation. Firstly, duplicate molecules are removed using comparison of their canonical Simplified Molecular Input Line Entry System (SMILES, [Weininger, 1988](#)) representation. The molecules are then standardised to ensure a consistent representation. The datasets contain molecules with varying protonation states; therefore, a ‘wash’ step is included so that each molecule is in a standard protonation state. We generate multiple conformations using RDKit ([Ebejer et al., 2012](#)) and after energy-minimisation the conformer with the lowest calculated force-field energy is retained and used for subsequent analysis.

Appendix G. The feature vector

Let $P' \subset \mathbb{R}^3$ be the collection of atom-centres of a candidate ligand molecule. We first construct a set Ξ of 100 directions in the unit sphere as follows – 32 directions correspond to vertices of the (origin-centred) pentakis dodecahedron; and the remaining 68 directions

are chosen at random by uniformly sampling S^2 . Our feature vector is constructed in five steps (MORSE):

- (1) **Modify the point cloud:** we centre P' about the origin and then apply a rotation so that the coordinate axes coincide with the principal components; let us call the resulting point cloud $P \subset \mathbb{R}^3$.
- (2) **Obtain the triangulation:** we construct the pruned Delaunay triangulation $K := K_w(P)$, where $w : P \rightarrow \mathbb{R}$ associates to each vertex the van der Waals radius of the corresponding atom.
- (3) **Realise the Morse transform:** we compute the Morse transform $\mathcal{M}_{K,d}$ of depth $d = 20$. Let $m_{\xi} \leq d$ be the total number of critical vertices found along the direction ξ up to depth d . Since $n = 3$ in our case, the Morse transform associates a $m_{\xi} \times 4$ matrix $M(\xi)$ to each direction $\xi \in \Xi$. We denote the i -th column of $M(\xi)$ by $M^i(\xi)$.
- (4) **Supplement with chemical data:** for each atom in the ligand we insert the atom's contribution to the partial charge, molar refractivity and lipophilicity into three multi-sets² \mathcal{Q} , \mathcal{R} and \mathcal{L} , respectively.
- (5) **Encapsulate in one vector:** for each integer $1 \leq i \leq 4$, let C_i be the union $\bigcup_{\xi \in \Xi} M^i(\xi)$ of i -th columns, recorded as a multi-set. We compute the p -th percentile C_i^p of C_i for all p in $\{0, 10, 25, 40, 50, 60, 75, 90, 100\}$. We also compute the same p -th percentiles of the multi-sets $\bigcup_{\xi \in \Xi} m_{\xi}$, \mathcal{Q} , \mathcal{R} and \mathcal{L} . Finally, we concatenate all the percentiles of the multi-sets together to obtain a single vector living in \mathbb{R}^{72} .

This is precisely the feature vector that we associate to the ligand represented by $P' \subset \mathbb{R}^3$.

Appendix H. Rotational invariance

The Morse transform $\mathcal{M}_{K,d}(\xi)$ (of depth d) of a simplicial n -complex $K_w(P')$ along a direction $\xi \in S^{n-1}$ is not intrinsically invariant under the action of the special orthogonal group $\text{SO}(n)$ on the underlying point cloud $P' \subset \mathbb{R}^n$ unless P' only consists of one point at the origin. However, the Morse feature vector of P' is invariant under the action of $\text{SO}(n)$ on P' owing to the intermediate step of orienting P' , as described below.

Let $C = \text{cov}(P', P')$ be the covariance matrix of P' and its eigendecomposition be $C = E\Lambda E^T$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ is the matrix of eigenvalues (*principal components*) of C and $E = [e^{(1)}, e^{(2)}, \dots, e^{(n)}]$ is the matrix of the associated eigenvectors (*principal component axes*). Then the point cloud P' is oriented by first aligning P' with the principal component axes and thus transforming it into a *canonical pose* $P_c = P'E$, which is invariant under the action of $\text{SO}(n)$ on P' . Due to the lack of uniqueness of the eigendecomposition, there is more than one canonical pose as there are $n!$ ways of ordering the eigenvectors in E and reflections of the eigenvectors $\pm e^{(i)}$ also satisfy the eigendecomposition leading to $n! \cdot 2^n$ canonical poses (Li et al., 2021). A *primary pose* P is chosen from the set of canonical poses $\{P_c\}$ by

- (1) ordering the eigenvalues such that $\lambda_1 > \lambda_2 > \dots > \lambda_n$, leaving 2^n canonical poses (assuming no repeated eigenvalues);
- (2) selecting the signs of the eigenvectors such that $|E| = 1$ (making E a proper rotation), leaving 2^{n-1} canonical poses; and

²Namely, we treat repeat occurrences of the same number as distinct elements.

- (3) keeping the eigenvectors with signs such that $\text{skew}[P'e^{(i)}] > 0$ for all $i < n$, leaving one canonical pose (assuming that P' does not have zero skew along any principal component axis).

In this way we can orient a point cloud P' into the primary pose P in a manner that is invariant under rotations of P' . In turn, the Morse feature vector that is derived from the primary pose is also rotationally invariant.

Appendix I. Hyperparameter tuning

We tune the hyperparameters of the LGBM classifier using a random search of 200 samples of the hyperparameter search space given in Table I1. For each 5-fold cross-validation split of the data, the best performing hyperparameters are selected by choosing the model with the lowest mean log loss across an *inner* 5-fold cross validation of the training subset of the data. Then the best hyperparameters are used to train the model using the whole training subset of the data, which is then assessed on the held-out test set. This whole process is repeated across each split of the 5-fold cross validation.

TABLE I1. LightGBM hyperparameters and their tuning search space. Given a set of hyperparameters \mathcal{S} , then $\text{uniform}(\mathcal{S})$ denotes uniform random sampling of \mathcal{S} and similarly $\text{loguniform}(\mathcal{S})$ denotes uniform logarithmic random sampling of \mathcal{S} .

Hyperparameter	Search Space
bagging_fraction	$\text{uniform}([0.3, 1.0])$
feature_fraction	$\text{uniform}([0.3, 1.0])$
max_depth	$\text{uniform}([2, 100] \cap \mathbb{N})$
min_data_in_leaf	$\text{loguniform}([20, 2000] \cap \mathbb{N})$
min_sum_hessian_in_leaf	$\text{loguniform}([10^{-5}, 20])$
num_leaves	$\text{loguniform}([2, 4095] \cap \mathbb{N})$

Appendix J. Evaluation metrics

The receiver operating characteristic (ROC) curve of a binary classifier is the plot of the true positive rate (TPR) against the false positive rate (FPR) as the threshold of the binary classifier varies. A common evaluation metric is the area under the ROC curve (AUROC) given by

$$\text{AUROC} = \int_{\infty}^{-\infty} \text{TPR} \frac{d\text{FPR}}{dT} dT \in [0, 1]. \quad (2)$$

It measures how likely it is for a randomly selected member of the positive class to be ranked above a randomly selected member of the negative class. A score of 0.5 indicates that the classifier has equivalent performance to a random classifier.

The Boltzmann-enhanced discrimination of the receiver operating characteristic at α (BEDROC_{α} , [Truchon and Bayly, 2007](#)) is given by

$$\text{BEDROC}_{\alpha} = \frac{w\text{AUAC} - w\text{AUAC}_{\min}}{w\text{AUAC}_{\max} - w\text{AUAC}_{\min}}, \quad (3)$$

where the weighted area under the accumulation curve (wAUAC) is given by

$$\text{wAUAC} = \frac{\int_0^1 F(x)w(x)dx}{\int_0^1 w(x)dx} \quad \text{with} \quad w(x) = e^{-\alpha x} \quad (4)$$

and $F(x)$ is the empirical cumulative distribution function (CDF) of the positive class. BEDROC_α is similar to the AUROC, except the contribution of the earlier part of the ROC curve is exponentially weighted to have a higher contribution. BEDROC_α is equal to 0.5 if the observed empirical CDF has the shape of the CDF produced by a probability density function proportional to the earlier exponential weight function with parameter α .

The enrichment factor at a fraction $\chi \in [0, 1]$ (EF_χ) is given by

$$\text{EF}_\chi = \frac{\int_0^1 F(x)w(x)dx}{\int_0^1 w(x)dx} \quad \text{with} \quad w(x) = \begin{cases} 1 & \text{if } x \leq \chi \\ 0 & \text{if } x > \chi \end{cases}. \quad (5)$$

It is a popular metric in virtual screening and measures how much more likely it is to find a member of the positive class in the first portion of a ranked sample than in the whole sample.

The relative enrichment factor at a fraction $\chi \in [0, 1]$ (REF_χ) is given by

$$\text{REF}_\chi = \frac{\text{EF}_\chi}{\text{EF}_{\chi, \max}}, \quad (6)$$

where $\text{EF}_{\chi, \max}$ is the maximum possible enrichment factor in the first χ of a ranked sample, which depends on the ratio of the positive to negative class.

Appendix K. Feature comparison criteria

There is a smorgasbord of virtual screening methods in the literature and an almost commensurate number of testing methodologies. Therefore, to ensure a fair comparison in the main paper we only evaluate our features against those that satisfy the following criteria:

- **LBVS feature.** We exclude features that use any information about the protein binding site.
- **Non-superpositional feature.** We exclude superpositional features that can only rank molecules using a similarity metric as these generally perform worse than classifiers trained on non-superpositional features (see Tables M1 and M2).
- **Classified with the same machine learning model.** We classify all the features in the main paper with a tuned LGBM classifier trained and evaluated in the same manner (see Section I).

Appendix L. Results for additional metrics

In Table L1 we record the mean scores of various metrics and their standard deviations of our tuned LGBM classifier trained on Morse features at depth 20 and 100 directions (M); chemistry-enhanced Morse features at depth 20 and 32 directions (M+C); R3P; USR; USRCAT; WHIM; and WHIMu descriptors. Morse features outperform all other shape-based features across all four metrics for the DUD-E Diverse dataset and are either best or joint-best across all metrics for the MUV dataset. Morse features augmented with chemical

data outperform all other shape and chemistry-based features across all metrics for both datasets.

TABLE L1. The mean metric scores and their standard deviations of various ligand-based virtual screening methods for the D8 and MUV datasets. For each target the method with the best metric score is displayed in bold and the best shape-based method is underlined.

Dataset	Metric	Shape				Shape & Chemistry		
		M	R3P	USR	WHIMu	M+C	USRCAT	WHIM
D8	AUROC	<u>0.84</u> ± 0.08	0.62 ± 0.08	0.76 ± 0.10	0.81 ± 0.10	0.97 ± 0.03	0.92 ± 0.07	0.88 ± 0.08
	BEDROC ₂₀	<u>0.43</u> ± 0.14	0.12 ± 0.05	0.30 ± 0.13	0.38 ± 0.15	0.86 ± 0.12	0.70 ± 0.17	0.56 ± 0.17
	EF _{1%}	<u>18</u> ± 9	3 ± 3	11 ± 8	16 ± 9	56 ± 14	44 ± 16	30 ± 12
	REF _{1%}	<u>0.29</u> ± 0.15	0.04 ± 0.05	0.18 ± 0.13	0.26 ± 0.14	0.87 ± 0.16	0.68 ± 0.22	0.48 ± 0.20
MUV	AUROC	<u>0.64</u> ± 0.11	0.53 ± 0.10	0.56 ± 0.12	0.61 ± 0.13	0.74 ± 0.12	0.63 ± 0.14	0.64 ± 0.14
	BEDROC ₂₀	<u>0.10</u> ± 0.08	0.06 ± 0.07	0.07 ± 0.08	<u>0.10</u> ± 0.09	0.24 ± 0.15	0.13 ± 0.13	0.12 ± 0.11
	EF _{1%}	<u>3</u> ± 6	1 ± 4	2 ± 6	<u>3</u> ± 6	10 ± 13	4 ± 8	3 ± 7
	REF _{1%}	<u>0.005</u> ± 0.012	0.002 ± 0.008	0.004 ± 0.012	<u>0.005</u> ± 0.013	0.019 ± 0.026	0.007 ± 0.015	0.006 ± 0.014

Appendix M. Comparison with similarity methods

Similarity (or superpositional) methods rank molecules by their similarity to a reference or template molecule using a custom similarity scoring function. In Table M1 we record the mean AUROC of similarity ligand-based virtual screening methods for DUD-E. We reproduce the results for eSim from Cleves et al. (2019); USR, USRCAT, ROCS (shape), ROCS (colour) and VAMS from Koes and Camacho (2014); CDK-D Moments, Shape-IT and Spectral Geometry Covariance 100 evaluations from Seddon et al. (2019); Interlig from Mirabello and Wallner (2020); and Optipharm-Robust and WEGA from Puertas-Martín et al. (2019).

In Table M2 we record the mean AUROC of similarity ligand-based virtual screening methods for MUV. We reproduce the results for VSFlow (shape COMBO) from Jung et al. (2023); Interlig from Mirabello and Wallner (2020); and BRUTUS and ROCS (colour) from Tiikkainen et al. (2009).

Generally, similarity methods perform worse than machine learning models trained on non-superpositional features, which can be seen by comparing Tables M1 and M2 with Tables 1 and 2 and in the main text.

TABLE M1. The mean AUROC per DUD-E Diverse (D8) subset target and the overall mean AUROC of the full set of 102 DUD-E (D102) targets for of various similarity methods. For each target the method with the highest AUROC is displayed in bold and the highest shape-based method is underlined.

Target	Shape								Shape & Chemistry			
	USR	CDK-D	Optipharm-Robust	ROCS (shape)	SGC	Shape-IT	VAMS	WEGA	eSim	Interlig	ROCS (colour)	USRCAT
AKT1	0.36	0.58	0.26	0.28	<u>0.62</u>	0.64	0.41	0.26	0.58	0.67	0.37	0.40
AMPC	0.53	0.63	0.63	0.58	0.58	0.59	<u>0.64</u>	<u>0.64</u>	0.62	0.76	0.76	0.73
CP3A4	0.53	0.52	0.53	<u>0.55</u>	0.51	0.54	0.54	0.53	0.58	0.70	0.51	0.51
CXCR4	0.62	0.67	0.71	<u>0.78</u>	0.66	0.65	0.72	0.73	0.79	0.92	0.78	0.65
GCR	0.50	0.63	0.52	0.49	0.66	0.77	0.49	0.50	0.64	0.76	0.59	0.63
HIVPR	0.74	0.62	0.70	0.72	0.63	0.62	<u>0.78</u>	0.71	0.84	0.84	0.69	0.73
HIVRT	0.62	0.50	0.52	0.63	0.50	0.46	<u>0.69</u>	0.52	0.71	0.67	0.61	0.67
KIF11	0.61	0.67	0.83	0.76	0.65	0.77	0.68	0.83	0.73	0.80	0.76	0.69
mean (D8)	0.56	0.60	0.59	0.60	0.60	<u>0.63</u>	0.62	0.59	0.69	0.77	0.63	0.62
mean (D102)	0.52	0.58	0.56	0.60	–	<u>0.61</u>	0.56	0.56	0.76	0.78	0.66	0.55

TABLE M2. The mean AUROC per MUV target of various similarity methods. An asterisk * indicates that the AUROC values were estimated from a figure. For each target the method with the highest AUROC is displayed in bold.

Target	Shape	Shape & Chemistry		
	VSFlow (shape COMBO)	Interlig	BRUTUS	ROCS (colour)
466	*0.62	0.57	0.51	0.56
548	*0.70	0.65	0.61	0.69
600	*0.59	0.59	0.50	0.58
644	*0.62	0.58	0.57	0.59
652	*0.57	0.61	0.42	0.57
689	*0.53	0.58	0.46	0.49
692	*0.63	0.56	0.63	0.65
712	*0.56	0.64	0.45	0.58
713	*0.51	0.51	0.41	0.48
733	*0.54	0.50	0.48	0.53
737	*0.65	0.59	0.46	0.53
810	*0.54	0.65	0.42	0.46
832	*0.70	0.65	0.55	0.62
846	*0.70	0.69	0.65	0.71
852	*0.78	0.67	0.63	0.73
858	*0.54	0.60	0.46	0.52
859	*0.49	0.59	0.51	0.51
mean	*0.60	0.64	0.51	0.58

Appendix N. Further ablation studies

We ablate the components of the Morse feature vector, then re-train and test our classifier on the ablated features to determine the relative effects of the components on the generalisation error. In Figures N1 and N2 we compare the per target performance of the chemistry-enhanced Morse features, Morse features and chemical percentile features. For nearly every target chemical information has stronger predictive power than shape information.

In Figures N3 and N4 we see that for both D8 and MUV the Morse-theoretic approach of focusing on critical vertices yields better performance than considering critical and regular vertices together, or randomly selecting vertices without replacement.

In Figure N5 we show the performance of various Morse features using different numbers of directions for MUV. Similar to the D8 results, the performance saturates around 32 directions and depth 15 (slightly higher than for D8) though the curves are less smooth with more uncertainty. In Figure N6 we compare chemistry-enhanced and vanilla Morse features against the best performing external methods in their category for MUV. Similar to D8, chemistry-enhanced Morse features are much less affected by depth with only a minor peak around depth 3.

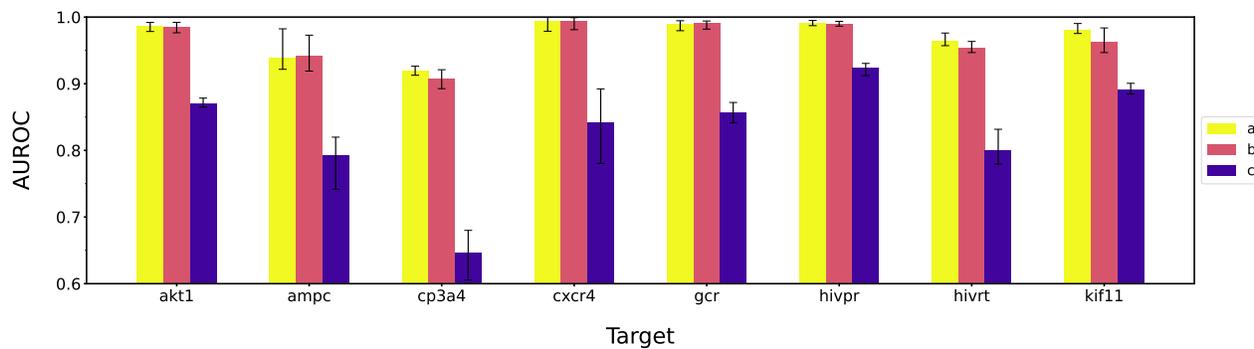


FIGURE N1. The mean AUROC score per DUD-E Diverse target of our classifier trained on Morse features at depth 3 and 32 directions enhanced with the chemical property percentiles (a), chemical property percentiles alone (b), and Morse features at depth 20 and 32 directions (c). Error bars are 95% confidence intervals.

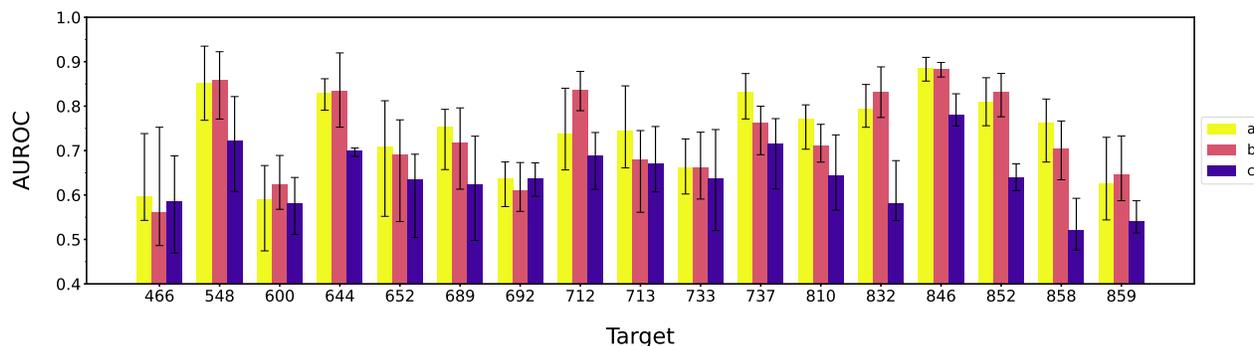


FIGURE N2. The mean AUROC score per MUV target of our classifier trained on Morse features at depth 3 and 32 directions enhanced with the chemical property percentiles (a), chemical property percentiles alone (b), and Morse features at depth 20 and 32 directions (c). Error bars are 95% confidence intervals.

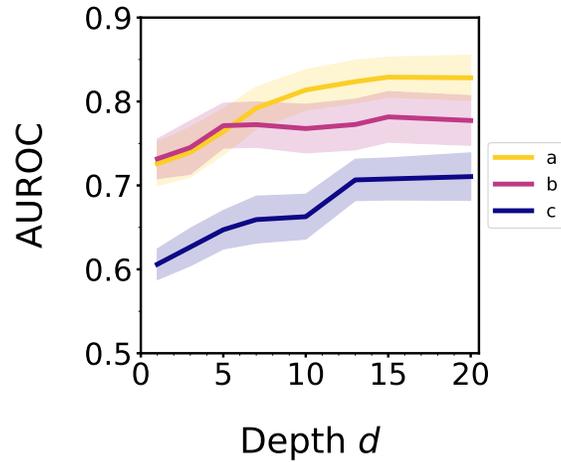


FIGURE N3. The mean AUROC score against depth of variants of the Morse feature vector (32 directions) for the DUD-E Diverse dataset. Here depth refers to the top number (by descending height) of critical vertices in the standard Morse transform (a), the top number (by descending height) of critical *and* regular vertices retained in a variant of the Morse transform (b), the number of randomly selected (without replacement) vertices in another variant of the Morse transform (c). Error bars are 95% confidence intervals.

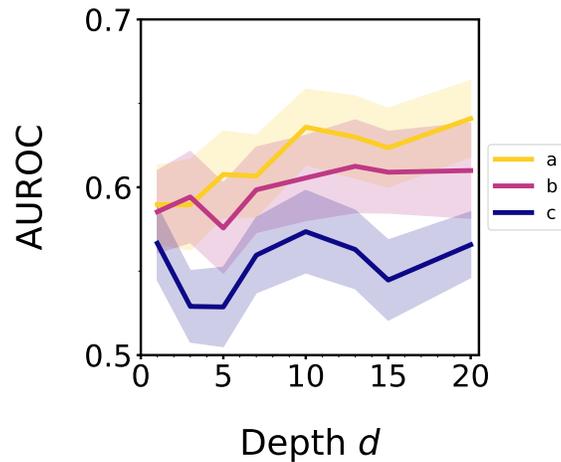


FIGURE N4. The mean AUROC score against depth of variants of the Morse feature vector (32 directions) for the MUV dataset. Here depth refers to the top number (by descending height) of critical vertices in the standard Morse transform (a), the top number (by descending height) of critical *and* regular vertices retained in a variant of the Morse transform (b), the number of randomly selected (without replacement) vertices in another variant of the Morse transform (c). Error bars are 95% confidence intervals.

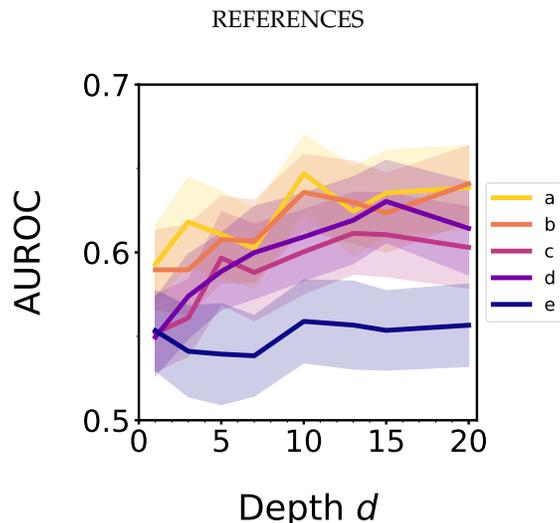


FIGURE N5. The mean AUROC score against depth of our LGBM classifier trained on Morse feature vectors computed using 100 directions (a), 32 pentakis dodecahedral directions (b), 12 icosahedral directions (c), 8 cubic directions (d) and 1 direction (e) for the MUV dataset. Error bars are 95% confidence intervals.

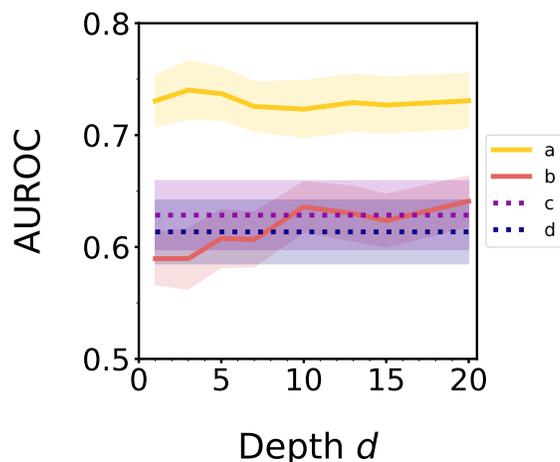


FIGURE N6. The mean AUROC score against depth of the LGBM classifier trained on the Morse feature vector computed using 32 directions (b) and Morse chemically enhanced feature vector computed using 32 directions (a) for the MUV dataset. For comparison, the best performing shape-based method unweighted WHIM (d) and best-performing external shape and chemistry-based method WHIM (c) are plotted with dotted lines. Error bars are 95% confidence intervals.

Appendix O. Implementation

Our pipeline is coded in PYTHON 3. We compute the WDTs and Betti numbers using the GUDHI library ([The GUDHI Project, 2024](#)); chemical properties with RDKit ([Rational Discovery LLC et al., 2024](#)); machine learning with SCIKIT-LEARN ([Pedregosa et al., 2011](#))

and LightGBM (Ke et al., 2017); hyperparameter tuning with RAY TUNE (Liaw et al., 2018); and throughout we use MATPLOTLIB (Hunter, 2007), NumPy (Harris et al., 2020), PANDAS (The Pandas Development Team, 2020), and PLOTLY (Plotly Technologies Inc., 2015).

Appendix P. Computational resources

Morse features and all their variants were computed on a cluster node consisting of two Intel Platinum 8628 CPUs (a 24 core 2.90 GHz Cascade Lake CPU) and 384 GB of memory.