Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PharmCast: rapid generation of three-dimensional pharmacophore fingerprints from two-dimensional structure without conformer generation

Steven M. Muskal

smuskal@eidogen-sertanty.com

Eidogen-Sertanty, Inc

Malcolm J. McGregor

SRI International

Research Article

Keywords: pharmacophore, fingerprint, scaffold hopping, surrogate model, virtual screening, applicability domain

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11662052/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

PharmCast: rapid generation of three-dimensional pharmacophore fingerprints from two-dimensional structure without conformer generation

Steven M. Muskal[1,*] and Malcolm J. McGregor[2]

[1] Eidogen-Sertanty, Inc. Oceanside, CA 92056;

[2] SRI International, Menlo Park, CA 94025

[*] Correspondence: smuskal@eidogen-sertanty.com

Abstract

A three-dimensional pharmacophore fingerprint records the binding features a molecule can present. It can help identify shared features among structurally distinct compounds, supporting scaffold hopping and similarity searching. We have leveraged it for decades, but its cost is dominated by conformer generation. We therefore removed the conformational stage, leveraging a feedforward neural network that predicts all 10,549 PharmPrint ensemble bits from a SMILES string. On the same machine it fingerprinted and compared two molecules in 0.584 ms vs. 5.71 s by the conformer based route, our reference calculation. Successive versions each added one kind of chemistry until accuracy stopped improving. Version 10 marks that convergence: 5,887,229 molecules from a screening collection, 1,202,183 activity backed ChEMBL compounds, and peptide loops from crystal structures. We evaluated three populations withheld from training: 155,648 catalog compounds, 1,453,042 ChEMBL compounds of 243 to 812 Da, and 13,500 peptide loops. On screening collection chemistry median fingerprint error was 0.008, Pearson r 0.980, and ranking accuracy 0.936. Loop peptides gave 0.016, 0.984 and 0.952, and the

ChEMBL arm 0.034, 0.910 and 0.870. Re-running the reference with a new conformer seed shifts it by 0.006 at r 0.995, the best any surrogate could do. Accuracy tracks how closely a molecule resembles the training set, and decreases on pharmacophores rare in training. Ensemble pharmacophore fingerprints can therefore be predicted from two-dimensional structure alone, at a cost suitable for collection-scale screening and virtual library exploration. PharmCast is downloadable at https://pharmcast.ai/.

Keywords: pharmacophore, fingerprint, scaffold hopping, surrogate model, virtual screening, applicability domain

1. Introduction

1.1 A century of the pharmacophore

The notion that binding depends on a small set of chemical features held in a particular spatial arrangement, and not on the skeleton carrying them, is among the oldest working ideas in medicinal chemistry. It is generally credited to Paul Ehrlich, whose 1909 account of chemotherapy illustrates the principle that a compound must possess both a group conferring affinity and a group conferring action.[1] Güner and Bowen have traced the concept to Ehrlich's late nineteenth century work, its movement toward a modern spatial definition through Schueler in 1960,[2] and its development into pharmacophore modeling by Kier from 1967,[3] who described the common features of ligands for central nervous system receptors and named the muscarinic pharmacophore.[4]

What made the idea computable was the ability to search three dimensional structures instead of reasoning about them by hand. Gund framed the problem of three dimensional pharmacophoric pattern searching,[5] and by the late 1980s systems such as 3DSEARCH[6] could take a query of typed features and interfeature distances and retrieve every molecule in a database able to present it. A two-dimensional substructure query asks what a molecule is made of; a pharmacophore query asks what it can present; and the two

questions have different answers. Asking the second question at scale has never been practical.

1.2 From query to descriptor

A query returns a list, whereas a descriptor returns a number for every molecule, and it is the descriptor that diversity analysis, similarity searching and quantitative modeling all require. The step from one to the other was to stop asking whether a molecule matches a single pharmacophore and to begin recording which of all possible pharmacophores it can present.

Pickett, Mason and McLay took that step with pharmacophore derived queries,[7] profiling libraries by the three-point pharmacophores their members express. We then defined the fingerprint used here[8, 9]: a fixed-length bit string in which each bit represents one three-point pharmacophore: a triple of feature types and the three binned distances between them. A molecule sets a bit when any accessible conformation presents that triangle. Mason and colleagues subsequently extended the construction to four-point pharmacophores,[10] which resolve the chirality that three-points cannot.

The descriptor has been applied and validated since. Deanda and Stewart applied the PharmPrint methodology independently to two protein kinases[11], and we applied it prospectively in a kinase-directed design campaign where the quantitative models driving compound selection were built on these fingerprints and pharmacophoric similarity established the novelty of the molecules chosen[12]. Here, we revisit how these fingerprints are generated.

The defining property of the descriptor is that it is nearly blind to scaffold. The 1999 paper makes the point with estradiol and diethylstilbestrol[8], both estrogen receptor ligands. They share a two-dimensional Morgan similarity of only 0.16, and they nevertheless present the same three features, an acceptor, a donor and an aromatic ring. Regenerating that geometry independently from a 100 conformer ETKDGv3 ensemble gives pairwise

feature separations of 2.8, 8.2 and 10.9 Å for estradiol against 2.8, 9.3 and 12.1 Å for diethylstilbestrol. Corresponding distances differ by as much as 1.2 Å, and yet each one falls in the same bin as its counterpart, 2.0 to 4.5, 7.0 to 10.0 and 10.0 to 14.0 Å respectively, so the two molecules set one and the same bit. Their pharmacophore similarity is 0.72 by the original conformation dependent method and 0.78 by PharmCast version 10. This example illustrates why topological pharmacophore searching became an established approach to scaffold hopping.[13-15]

Pharmacophore methods have since become standard in structure based form deriving models from protein bound ligands[16] and in ligand based form throughout virtual screening and library design. Leach, Gillet, Lewis and Taylor give a thorough account of the field as it stood.[17]

1.3 Computational cost

A molecule does not have a single three-dimensional shape, so the fingerprint is defined over a conformer ensemble, conventionally the bitwise OR across many conformers. Generating that ensemble dominates the computational cost.

Timed on screening collection compounds with the pipeline used here, the full route requires 2.86 s per molecule, of which 2.82 s is conformer generation with RDKit ETKDGv3 followed by UFF minimization and 0.039 s is the fingerprint calculation. At that rate a five million compound collection is roughly five and a half core months. Conformer generation uses established methods such as knowledge augmented distance geometry[18] and systematic torsion driving.[19]

The consequence has been that pharmacophore similarity is confined to small, curated sets. It is not usable as a screening objective over the full screening collection, nor inside an optimization loop, which is precisely where a three dimensional similarity measure would be most valuable. Two-dimensional fingerprints such as extended-connectivity fingerprints[20] are used instead because they are computationally affordable, although

they answer a different question. The fingerprint calculation accounts for only 1.4% of the total cost, so we focused on removing the conformational stage.

1.4 This work

We predict the ensemble fingerprint directly from two-dimensional structure and omit conformer generation. An ensemble fingerprint is an OR over conformations, so a bit records that a molecule can present a given triangle in some accessible conformation, without asserting that it presents it in any particular pose. That is closer to a property of molecular constitution than of any single geometry, and constitution is what a two-dimensional structure encodes.

PharmCast uses a feedforward network to map a circular fingerprint and eleven whole molecule descriptors to all 10,549 pharmacophore bits. More elaborate learned representations, including graph convolutional[21] and message-passing[22] networks, may improve performance.

Structurally distinct ligands for the same target are common enough to motivate scaffold hopping[13-15]. A pharmacophore fingerprint computed at screening collection scale can support it, and can prioritize compounds that are structurally novel. This work provides (i) a surrogate model and a measurement of its agreement with the reference calculation relative to that calculation's reproducibility; (ii) an explicit applicability domain that quantifies failure and relates it to training-set composition; and (iii) a version series showing that additional data from the same chemical space did not repair out-of-domain error, whereas adding the missing chemistry did.

2. Methods

2.1 The fingerprint

Each fingerprint contains 10,549 bits, one per three-point pharmacophore. A bit contains three feature types drawn from acceptor, donor, negative, positive, hydrophobic, aromatic

and other, together with the three pairwise distance bins between them. Distance bins are 2.0 to 4.5, 4.5 to 7.0, 7.0 to 10.0, 10.0 to 14.0, 14.0 to 19.0 and 19.0 to 24.0 Å. The enumeration is deterministic: distances are sorted ascending with ties broken by feature type so that each triangle has one canonical form, combinations violating the triangle inequality on the bin bounds are discarded, and the survivors are numbered sequentially. This encoding is the three-point PharmPrint, and a similarity comparison between two of them is a PharmSim score.

Reference fingerprints were generated identically for each source set, i.e. 100 conformers from isomeric SMILES using RDKit ETKDGv3 with a fixed seed followed by UFF minimization, largest fragment, hydrogens added, and per conformer fingerprints combined by bitwise OR into one ensemble record. This consistent procedure allows comparison across source sets.

The 100 conformer convention was checked against shallower and deeper ensembles on the flexible tail of the collection, 24 molecules carrying 8 or more rotatable bonds and spanning 154 to 977 Da. At 100 conformers the reference fingerprint is within 0.98 Tanimoto of one built from 200, median bit growth from 100 to 200 conformers is 2.5%, and every fingerprint at every depth is contained in the 200 conformer fingerprint (Online Resource 1, Table S6).

2.2 Training data

PharmCast version 10 uses three data sources: the Enamine screening collection, activity-backed ChEMBL compounds, and protein loop peptides. The June 2026 Enamine screening collection held 4,774,670 source compounds, of which 4,617,292 survived property filtering. Median molecular weight is 344 Da (range 116 to 599; 5th and 95th percentiles 255 and 465). Compounds have median values of 24 heavy atoms, 3 rings (2 aromatic), cLogP 2.7, and polar surface area 71 Å^2; 8% have defined stereochemistry. Compounds in the file are labeled by Enamine as HTS, Advanced, Legacy or Functional,

of which only HTS is the actively stocked screening library, a distinction that matters whenever a screening result is quoted as purchasable.

We selected the three sources because (i) the screening collection contains synthesized molecules intended for biological screening; (ii) activity-backed ChEMBL contains compounds that have been synthesized and tested against biological endpoints, demonstrating synthetic accessibility and potential biological relevance; and (iii) protein loop peptides come from regions of X-ray structures outside regular secondary structure.

The collection is diverse by scaffold and redundant by similarity. In a 200,000 molecule sample there are 135,768 distinct scaffolds, 88% of which appear exactly once, and the ten commonest scaffolds together hold 4.0% of the collection. The median compound nevertheless has a nearest neighbor at 0.714 Tanimoto on Morgan fingerprints, 54.9% have a neighbor at 0.70 or closer, and 64,473 compounds have a two-dimensional identical twin, so a raw compound count overstates how much distinct chemistry is present.

The full peptide set contains 136,494 loops of two to six residues excised from Protein Data Bank crystal structures[23], all from structures at 2.5 Å resolution or better and R-free 0.23 or better, and capped with the flanking backbone atoms present in the structure: the preceding residue's C and O as an acetyl and the following residue's N and CA as an N-methylamide. Each peptide fingerprint is the bitwise OR of 100 ETKDGv3 conformers, each optimized with UFF and converted to a reference fingerprint, together with every rigid crystal conformation observed for that peptide.

Protein loop peptides extend the conformational range represented in the training set. A surrogate for an ensemble fingerprint is hardest where a molecule has the most conformational freedom, and screening collection chemistry holds very little of it; the median loop has 16 rotatable bonds, compared with 5 for a screening-collection compound. We used loops excised from deposited structures to cover this range. Online Resource 1,

Fig. S1 shows one sequence in every deposited conformation. Median bit count is 412 pharmacophores per fragment against 534 for a collection compound.

The 13,500 held-out loops and the 122,994 training loops together account for all 136,494; section 2.4 gives how the holdout was stratified. Their reference fingerprints were generated by one procedure, so the only difference between them is whether a loop was available during training.

The ChEMBL set includes every ChEMBL molecule[24] from 142 to 1000 Da with at least one recorded activity, establishing that the compound was synthesized, purified, and assayed. Fingerprinting proceeds in ascending molecular weight so that coverage extends incrementally from chemistry already represented in the training set. Of these molecules, 1.6% fail to embed.

Each source set is sampled to the same size, 6,000 molecules, because nearest neighbor similarity rises with sample size and unequal samples cannot be compared. These values are therefore comparable to one another but not to a figure computed over a whole source set: the full 4.6 million compound collection has a median nearest neighbor of 0.714 against 0.389 in a 6,000 molecule sample of it. That the peptides are the least diverse of the three is expected, since the same loop sequence recurs across many structures. The three distributions are drawn in Online Resource 1, Fig. S2.

2.3 Model architecture and training

The network takes a 2,048 bit binary Morgan fingerprint of radius 2 alongside 11 whole molecule descriptors, 2,059 inputs in all, and predicts all 10,549 pharmacophore bits through two hidden layers of 1,024 and 512 units, for 8.05 million parameters. The descriptors are molecular weight, cLogP, topological polar surface area, rotatable bonds, rings, aromatic rings, hydrogen bond donors, hydrogen bond acceptors, heavy atoms, fraction of sp3 carbon and Labute accessible surface area. Training minimizes binary cross entropy.

Training and all timing measurements were performed on a single Apple Mac Studio with an M3 Ultra chip, 28 CPU cores, a 60 core GPU and 256 GB of unified memory, running PyTorch 2.9.1. From PharmCast version 8 onward, training used the Metal Performance Shaders backend on the GPU; the timing measurements reported here were made on the CPU, and inference uses no GPU. Each version retained the best set of weights obtained during training; version 10 ran for 100 epochs and restored the weights of epoch 68.

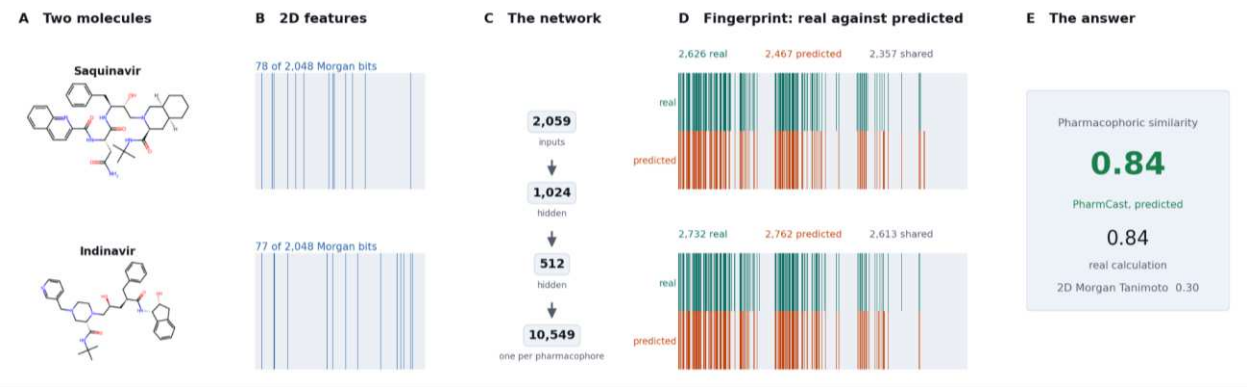


Fig. 1 The two routes to a fingerprint, shown for saquinavir and indinavir, HIV-1 protease inhibitors of unrelated scaffold. The conventional route generates a conformer ensemble and applies the reference calculation; PharmCast predicts the ensemble record from two-dimensional structure. PharmCast predicts a pharmacophore Tanimoto of 0.843, against 0.841 from the reference calculation and a two-dimensional Morgan Tanimoto of 0.303.

Saquinavir and indinavir illustrate this in Figure 1. Their Morgan similarity is 0.303, consistent with different chemical series, whereas their pharmacophore similarity is 0.841. Both compounds lie in the sparsely represented 600 to 700 Da range. Neither compound occurs in any training set. The two structures are CC(C)(C)NC(=O)[C@@H]1C[C@@H]2CCCC[C@@H]2CN1C[C@@H](O)[C@H](Cc1ccccc1)NC(=O)[C@H](CC(N)=O)NC(=O)c1ccc2ccccc2n1 for saquinavir and CC(C)(C)NC(=O)[C@@H]1CN(Cc2cccnc2)CCN1C[C@@H](O)C[C@@H](Cc1ccccc1)C(=O)N[C@H]1c2ccccc2C[C@H]1O for indinavir.

PharmCast version 10 was trained on 5,887,229 molecules drawn from screening collection compounds, activity-backed ChEMBL compounds from 142 to 1000 Da, and protein loop peptides, with a molecular-weight-stratified stopping set reserved for early stopping; feature normalization was fitted only on the training rows.

The stopping set is distinct from the three unseen test populations, which section 2.4 defines. None was used for training or early stopping.

2.4 Test sets and evaluation

Version 10 is evaluated on three chemically distinct populations because pooling would conceal their different failure modes. All are unseen during training.

- Loop peptides: 13,500 molecules, 9.9% of the full peptide set, reserved for testing and never trained on. Sampling was stratified by sequence length from two to six residues, by molecular-weight quintile, and by the number of crystal conformations deposited for that peptide, grouped as one, two or three, four to ten, and more than ten. Each stratum contributed 9.9%; the holdout median molecular weight is 575.7 Da and the training loops 576.6 Da.
- ChEMBL: 1,453,042 activity-backed molecules fingerprinted after the version 10 corpus was sealed, so none of them was available to train on. The arm spans 243.3 to 811.9 Da at a median of 458.5 Da, and it is the whole post-seal population rather than a sample of it.
- Screening collection: 157,335 real, purchasable Enamine screening-collection compounds from **the June 2026 Enamine** catalog that the training-set ingest filter excluded. Canonical SMILES were compared with all 4,617,292 screening collection structures, and two matches were removed. Molecular weight spans 42 to 1000 Da. Of these, 155,648 were fingerprinted and scored, and 1,687 could not be embedded.

Evaluation uses all 13,500 peptide holdouts, all 1,453,042 held-out ChEMBL molecules, and the 155,648 screening collection test cases with reference fingerprints. Pair statistics are estimated from five million pairs drawn from each population, and ranking accuracy from one million molecule triples.

Fingerprint error is the absolute difference between the predicted and the reference Tanimoto for a pair of molecules. Ranking accuracy is measured on triples of one query and two candidates. It is the fraction of triples in which the surrogate ranks the two candidates in the same order as the reference calculation, with exact ties set aside.

Agreement describes a single molecule rather than a pair. It is the Matthews correlation coefficient[25] between the predicted and the reference fingerprint bits, and it is labeled MCC in every table and figure. Accuracy would be dominated by bits that are zero in both fingerprints because 10,549-bit fingerprints can be sparse. The pair statistics follow from this per molecule distribution.

A ceiling is required in order to interpret either quantity. Repeating the reference calculation on the same molecules with a different embedding seed gives a fingerprint error of 0.006 at r 0.995, and no surrogate can be expected to do better than the target's own reproducibility.

Figure 5 scores every released model on the same 155,648 screening collection test cases. Online Resource 1, Table S9 reports each model's ranking accuracy on a fixed set of 5,000 held-out screening collection molecules and on 573 real ChEMBL compounds above 600 Da that were excluded from every model's training set. The fixed test populations allow direct comparison across all model versions.

2.5 Reference fingerprint coverage

Online Resource 1, Table S1 gives the reference fingerprints available when version 10 was trained against the eligible molecules in each source. The ChEMBL remainder is

chemistry that later versions can draw on, both as training data and as held-out molecules for testing. The peptide source contributes 136,494 reference fingerprints, of which 122,994 were used for training and 13,500 reserved for testing.

3. Results

All results in this section were obtained with PharmCast version 10. Figure 5 places it in the version series, and Online Resource 1 gives the ranking accuracy of every version and the training set each was built on. Agreement leveled off across versions 8, 9 and 10.

3.1 Computational performance

Table 1 Cost of generating one pharmacophore fingerprint, and of fingerprinting the complete collection of 4,653,831 compounds, by route.

Route	One fingerprint	Full screening collection
Conformer ensemble plus reference calculation	2.86 s	3,697 core hours
PharmCast from SMILES	0.288 ms	29.3 minutes on six workers

All timings were measured on one machine, an Apple M3 Ultra with 28 CPU cores, 20 performance and 8 efficiency, a 60 core GPU and 256 GB of unified memory, running macOS 26.5.1 with PyTorch 2.9.1 on CPU. The GPU was not used, and both routes were timed on the same hardware. These are inference timings, and a GPU would not materially change them: the forward pass is 0.009 ms of the 0.288 ms per molecule, so accelerating it cannot move the total. For training, the same GPU is 12.5 times faster per step.

Table 1 reports the end-to-end cost of predicting every fingerprint from SMILES and screening the complete collection: 29.3 min at 2,644 compounds per second on six worker processes, including screening collection input and featurization. Screening a collection whose fingerprints already exist is a bitwise comparison, which costs 0.012 ms per pair. The 3,697 core-hour conventional value was extrapolated from the per-molecule rate

because the collection has been fingerprinted by that route only in multiday production runs.

The per molecule figures split as follows. 2.82 s is conformer generation with RDKit ETKDGv3 followed by UFF minimization, and 0.039 s is the fingerprint calculation over the ensemble. PharmCast replaces both with featurization and one forward pass. In a library-scale batch, this takes 0.288 ms per molecule, including 0.009 ms for the forward pass. Processed alone, a molecule takes 0.357 ms, including 0.294 ms for featurization.

Comparing two fingerprints that already exist is a Tanimoto over 10,549 bits and costs 0.012 ms by either route, so the comparison is not where the difference lies. Starting from two SMILES strings, PharmCast returns a similarity in 0.584 ms at library scale against 5.71 s for the conventional route. Everything separating those two figures is conformer generation.

3.2 Fidelity by test population

PharmCast version 10 was trained on 5,887,229 molecules. It is evaluated on 155,648 Enamine screening collection test cases, 1,453,042 activity-backed ChEMBL compounds fingerprinted after the version 10 seal, and the 13,500-peptide reserved holdout. Median per molecule MCC is 0.881, 0.828 and 0.914 respectively, and Figure 3 draws that distribution for the screening collection arm.

Table 2 Fidelity by population for PharmCast version 10, against the reference calculation's own reproducibility. Screening collection is the Enamine test cases, ChEMBL is every activity-backed molecule fingerprinted after the version 10 seal, and loop peptides are the reserved holdout. Median MCC uses every molecule; the pair statistics come from five million pairs sampled per population and the ranking accuracy from one million triples, because the possible pairs grow with the square of the population. The final row compares the reference calculation with itself and gives the reproducibility ceiling.

Population	Molecules tested	Median MW	Median MCC	Median error	Pairs within 0.05	Pearson r	Ranking accuracy
Screening collection	155,648	211.3	0.881	0.008	89.3%	0.980	0.936
Loop peptides	13,500	575.7	0.914	0.016	88.3%	0.984	0.952
ChEMBL, activity backed, 243 to 812 Da	1,453,042	458.5	0.828	0.034	65.3%	0.910	0.870
Reference self-comparison	N/A	N/A	N/A	0.006	N/A	0.995	Ceiling

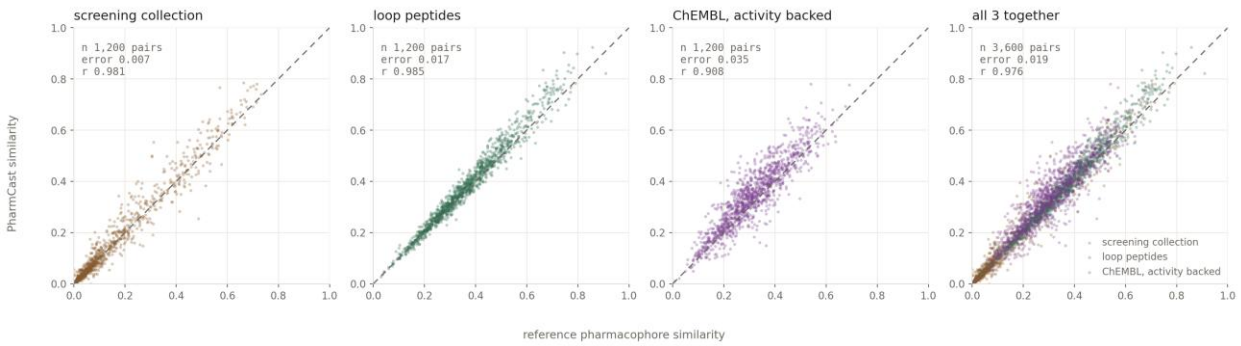


Fig. 2 Predicted against real pairwise similarity for PharmCast version 10. Peptides are the reserved holdout, screening collection compounds are Enamine test cases, and ChEMBL compounds were fingerprinted after the version 10 seal. Each population panel draws 1,200 pairs at random from 900 uniformly sampled molecules; the combined panel pools 3,600 pairs. Panel error and r use those pairs. The pair count keeps the scatter legible; Table 2 reports all tested molecules. The dashed line is exact agreement. The weakest population is ChEMBL, for which median error is 0.034 and about one pair in three falls outside 0.05.

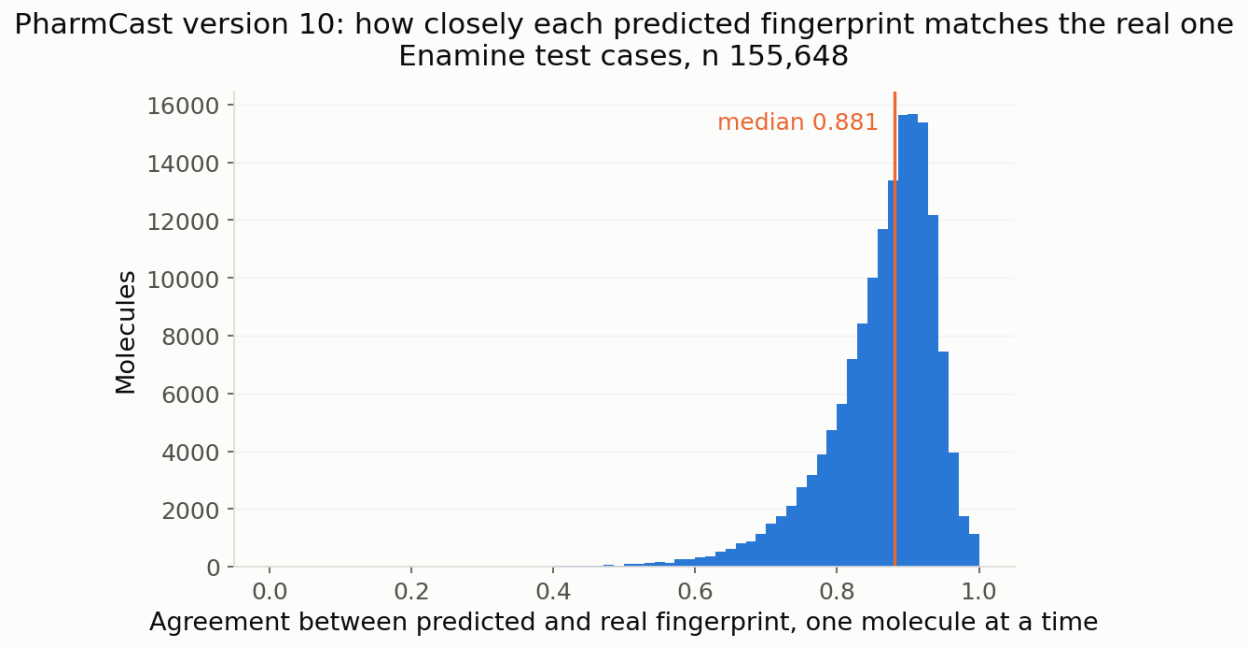


Fig. 3 How closely each predicted fingerprint matches the reference, as MCC per molecule, for PharmCast version 10 on 155,648 Enamine test cases. Median 0.881. The pairwise similarity results are a consequence of this distribution.

Aggregate agreement conceals which bits are recovered. Every one of the 10,549 pharmacophores was scored for how often the reference calculation sets it across the screening collection arm, and recall rises with that prevalence: median recall is 0.00 for pharmacophores the reference sets in fewer than one molecule in ten thousand, then 0.38, 0.59, 0.72 and 0.94 as prevalence rises through 0.1%, 1%, 10% and above. 827 pharmacophores are never recovered across the 155,648 screening collection molecules and 273 across the 1,453,042 ChEMBL molecules, so screening for a common feature combination and screening for a rare one are not the same proposition. False positives and false negatives are close to balanced on the screening collection and the peptides, 13 against 12 and 134 against 134 bits per molecule, and run 150 against 129 on the heavier ChEMBL arm (Online Resource 1, Table S5).

3.3 Independence from two-dimensional similarity

Figure 4 plots PharmCast's pairwise similarity against the reference calculation's, as Figure 2 does, with each pair colored by the two-dimensional Morgan similarity of the two molecules. The points follow the diagonal at every color, so PharmCast tracks the reference calculation whether or not the two molecules are two-dimensionally alike.

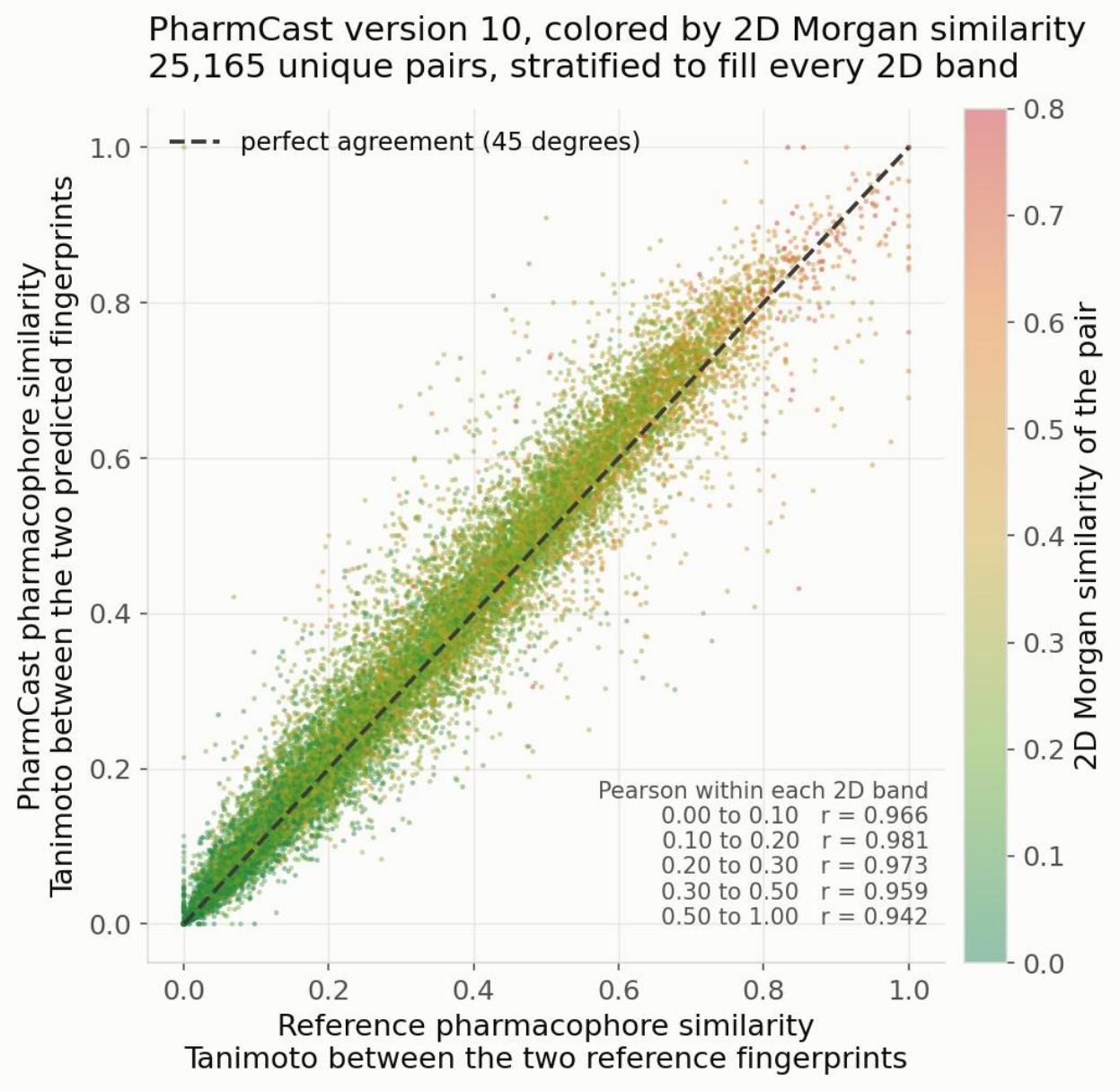


Fig. 4 Predicted against real pairwise similarity, colored by two-dimensional Morgan similarity, green for the most dissimilar pairs through red for the most similar. PharmCast version 10 on 5,000 held-out Enamine test cases. Agreement is flat across the color range, consistent with prediction of three-dimensional feature geometry independently of two-dimensional similarity.

The 5,000 molecules form 12,497,500 pairs. Each was scored for two-dimensional similarity, and the figure draws 6,000 from each of the bands 0 to 0.1, 0.1 to 0.2, 0.2 to 0.3 and 0.3 to 0.5, together with all 1,165 pairs above 0.5, for 25,165 in total. Balancing the bands is necessary because the library is not balanced: in a uniform draw, 97.5% of pairs fall below 0.20, and the two-dimensionally similar end of the panel would be nearly empty. Pearson agreement is 0.966, 0.981, 0.973, 0.959 and 0.942 across those bands, from the least two-dimensionally similar to the most. Recovering pharmacophore similarity across structurally dissimilar molecules is necessary for scaffold hopping.

3.4 Applicability domain

PharmCast version 10 is trained across 74 to 1000 Da. Performance within that span depends on training-set composition, which is given by source and by mass in Online Resource 1, Table S2. Coverage across that range is uneven, and the count by mass band says how uneven (Online Resource 1, Table S3).

Error did not increase uniformly with molecular weight. We therefore stratified the held-out molecules from Table 2 by weight, and Table 3 reports the result.

Table 3 Per molecule agreement against molecular weight for PharmCast version 10. Screening collection is the Enamine test cases, ChEMBL is the held-out activity-backed set, and loop peptides are the reserved holdout. Cells holding fewer than 20 molecules read not measured.

MW band	Screening collection	ChEMBL	Loop peptides
up to 300 Da	0.872	not measured	0.902
300 to 400 Da	0.887	0.858	0.938
400 to 500 Da	0.887	0.829	0.937
500 to 600 Da	0.899	0.812	0.922

600 to 700 Da	0.874	0.819	0.899
700 to 1000 Da	0.844	0.784	0.867

Screening collection and peptide agreement both vary by mass. Peptide agreement runs from 0.938 at 300 to 400 Da down to 0.867 above 700 Da, and screening collection agreement from 0.899 at 500 to 600 Da down to 0.844 above 700 Da. The held-out ChEMBL molecules span 243.3 to 811.9 Da and populate five bands, running from 0.858 at 300 to 400 Da to 0.784 above 700 Da, and they sit below the screening collection at every mass, 0.858 against 0.887 in the 300 to 400 Da band and 0.784 against 0.844 above 700 Da. Molecular weight alone therefore does not define the boundary. Training-set coverage and chemical class both matter, and ChEMBL is the most heterogeneous source and the least complete, at 44.4% fingerprinted.

PharmCast version 10 is calibrated across 74 to 1000 Da, most tightly on screening-collection chemistry and loop peptides, and least tightly on the diversity of activity-backed ChEMBL chemistry. Agreement declines at the upper end of the mass range, where the training set is thinnest. The donor, acceptor and rotatable-bond ranges are given in section 3.7 because a molecule can sit inside the mass range and outside the training set on any of them.

Mass is not the only axis on which a held-out molecule can sit far from the training set. Because the screening collection is redundant by similarity, a molecule drawn from it may have a close two-dimensional neighbor among the molecules the model trained on, and agreement should be read against that distance. For every sampled test molecule we computed the maximum Morgan Tanimoto to any of the 5,946,696 molecules in the version 10 sealed corpus, and Table 4 reports agreement inside similarity bands. Agreement rises with proximity to the training set in all three populations. The ChEMBL arm carries that result most cleanly: agreement runs from 0.752 for the most novel molecules to 0.916 for those with an exact two-dimensional twin in training, while the band medians span only

424 to 488 Da, a range over which Table 3 puts the mass effect near 0.01 against the 0.164 the bands move. In the screening collection the two move together, since the most novel band has a median mass of 139 Da against 488 Da for the near band, so that arm does not separate novelty from size. Only 6.7% of the screening collection arm has an exact twin in training, and that band does not score highest, which is not what a lookup would give. The peptide arm has no molecule below 0.50 and 39.3% of it has an exact twin, so it measures agreement on conformationally rich chemistry and says nothing about novel chemistry.

Table 4 Per molecule agreement against distance from the version 10 training set. Each cell gives the median MCC with the share of the arm in that band. Distance is the maximum Morgan Tanimoto from the test molecule to any molecule in the sealed training corpus, scanned over all 5,946,696 of them. Bands follow Mattsson and Walters[26], so the cut points were fixed independently of these results; the top band is separated at 0.999 so a molecule with an exact two-dimensional twin in training is counted on its own. Each arm is a random sample of 5,000 molecules from the populations of Table 2. The sealed corpus is about one percent larger than the 5,887,229 molecules used for training, the remainder being the stopping split, which can only understate novelty.

Nearest training neighbor	**Screening collection**	**ChEMBL, 243 to 812 Da**	**Loop peptides**
0.00 to 0.35	0.758 (0.6%)	0.752 (0.8%)	no molecules
0.35 to 0.50	0.829 (15.6%)	0.791 (14.6%)	no molecules
0.50 to 0.70	0.876 (53.5%)	0.822 (48.4%)	0.894 (0.2%)
0.70 to 1.00	0.909 (23.7%)	0.849 (32.6%)	0.916 (60.5%)

exactly 1.00	0.894 (6.7%)	0.916 (3.6%)	0.914 (39.3%)
Median nearest neighbor	0.62	0.65	0.96

3.5 Performance by training-set size and composition

All released models can be compared on one curve. The plotted agreement is the median MCC, and every version is measured on the same 155,648 screening collection test cases excluded from every training set. The curve therefore compares nine models on one population.

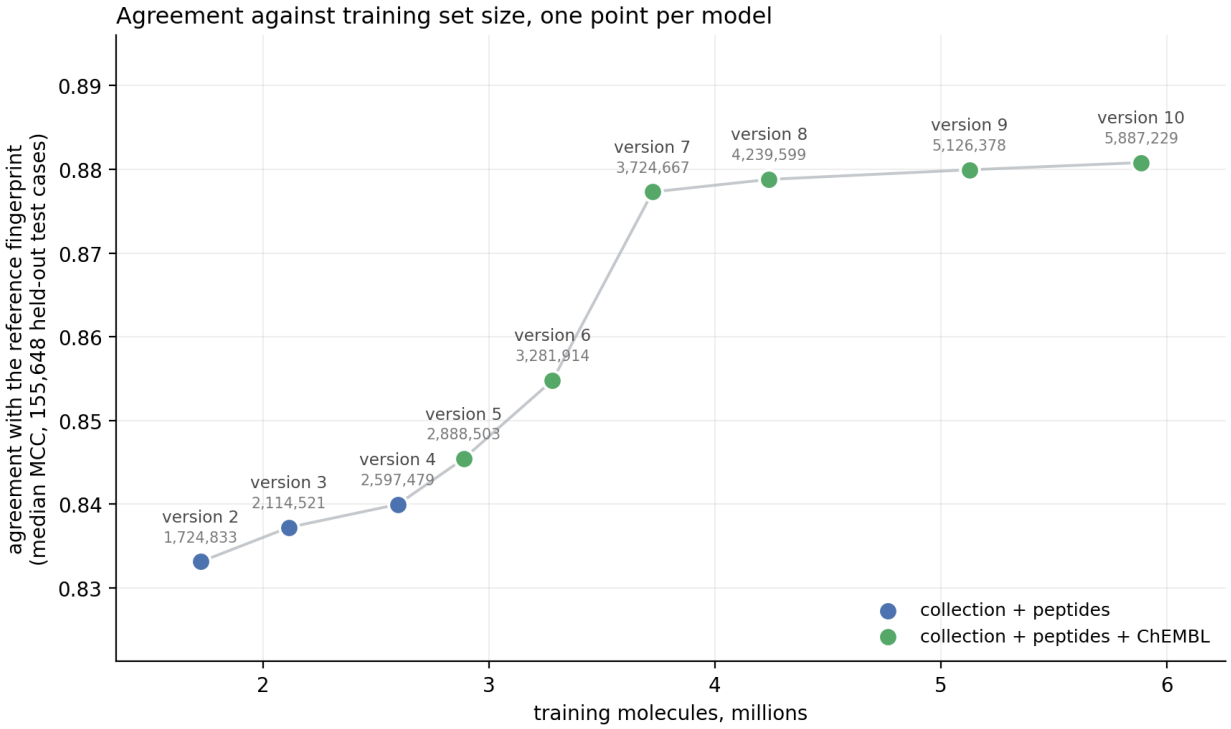


Fig. 5 Median MCC against training-set size on all 155,648 held-out screening collection test cases. Every model is scored on the same held-out test cases, so the points differ only by the model. The test cases are held out from every version by construction. The series ends at PharmCast version 10. Color marks the training sets; each point is labeled with its training set size.

Agreement rises from 0.833 at 1.72 million molecules to 0.881 at 5.89 million, and Figure 5 plots every point. Online Resource 1, Table S9 separates screening collection chemistry from real ChEMBL compounds above 600 Da. Screening collection ranking accuracy rises from 0.908 and then levels off in the high 0.93s. Large-ChEMBL ranking accuracy is 0.701 at PharmCast version 2, 0.710 at version 3 and 0.711 at version 4. Over that range the training set grew by 51%, but the added molecules came from a library filtered at 600 Da, so the large-compound result did not move.

Adding chemistry absent from the collection improved the out-of-domain result. Large-compound ranking accuracy rises from 0.711 at version 4 to 0.826 at version 10, and the step occurs at version 5, where ChEMBL chemistry first entered the training set.

Successive versions change training-set size and source composition together, so the series records what the released models do rather than isolating a cause, and the large-compound population of 573 molecules is small enough that adjacent versions from version 7 onward sit within its sampling noise. The main observation is that the training set grew by 51% to version 4 without improving the large-compound result, which improved when chemistry above the 600 Da filter entered.

3.6 Training-set composition

Online Resource 1, Fig. S4 and Table S1 report the composition of the version 10 training set. The screening collection contributes 4,609,488 compounds and is 99.8% complete. The loop-peptide set is complete at 136,494. ChEMBL contributes 1,214,214 compounds and is 44.4% complete. The source sets continue to be fingerprinted as they grow, and the ChEMBL molecules fingerprinted since the seal are the test arm of section 2.4.

3.7 Descriptor coverage

Section 3.4 reports error by molecular weight. Applicability also depends on other molecular properties. Donors and acceptors are two of the six feature types the fingerprint is built from, and the rotatable bond count sets how much conformational space the ensemble has to cover, so a molecule can sit inside the mass range and outside the training set on either. Online Resource 1, Table S4 gives each range as the 5th to the 95th percentile of a random sample of 20,000 molecules per source.

Full observed ranges over the same samples are as follows: the collection spans 173 to 597 Da, 0 to 5 donors, 0 to 11 acceptors and 0 to 10 rotatable bonds; ChEMBL spans 142 to 646 Da, 0 to 12 donors, 0 to 19 acceptors and 0 to 27 rotatable bonds; and the peptides span 159 to 961 Da, 0 to 19 donors, 2 to 13 acceptors and 2 to 32 rotatable bonds. The collection and peptide rows describe their full source sets. The ChEMBL row describes the reference fingerprints used for version 10; additional eligible ChEMBL compounds may extend these ranges in later versions. The rotatable-bond distribution separates the three source sets more strongly than molecular weight: median values are 5 for the screening collection, 4 for the ChEMBL set in the version 10 training set, and 16 for the peptides.

3.8 Agreement in a retrieval setting

We also compared candidate rankings from PharmCast and the reference calculation. The queries are 484 ChEMBL compounds with a measured pChEMBL of 7 or better against one of five targets from five protein families, none present in the version 10 training set. Each is scored against the other actives for its target together with compounds matched to it on mass, cLogP, donors, acceptors and rotatable bonds and drawn from the Enamine test cases, one to two thousand candidates per target.

Agreement over the whole ranking is high and agreement at the top of it is partial. Spearman between the reference ordering and the PharmCast ordering is 0.90, from 0.86 on Thrombin to 0.94 on mu opioid receptor. PharmCast places 0.54 of the reference

calculation's top one percent in its own top one percent and finds 0.62 of the actives the reference ranks there. Property matching is what makes this hard: every candidate is the size and shape of the query, so the similarities bunch and the top of the list turns on small differences.

These results support using PharmCast to rank a collection, followed by the reference calculation to evaluate the leading candidates. Online Resource 1, Table S10 reports every target.

3.9 Application: screening collection search against a pharmacophore reference

A fast pharmacophore comparison can be used as a search objective. Orforglipron, an 883 Da non-peptidic reference, was screened against 4,653,831 collection compounds, with a fingerprint generated from SMILES for each, in 29.3 minutes. That cost is paid once. With the collection fingerprints already stored, the same screen is 4,653,831 bitwise comparisons at 0.012 ms each, 55.85 s for the whole collection, which is what makes a pharmacophore screen a routine query rather than a campaign. The leading compounds were then recomputed with the reference calculation. The highest-scoring purchasable compound has an ensemble Tanimoto of 0.786 at 591 Da and a two-dimensional Morgan similarity of 0.113 to the reference, supporting scaffold hopping instead of analog retrieval.

The reference value of 0.786 establishes a baseline that subsequent designed molecules must exceed. This application lies in a sparsely populated region of the training set, motivating the expansion reported here.

4. Discussion

PharmCast predicts an ensemble pharmacophore fingerprint from two-dimensional structure with median error of 0.008 on screening collection chemistry, 0.016 on conformationally rich peptides and 0.034 across the whole post-seal ChEMBL population,

against 0.006 for the reference calculation's own reproducibility. Because the fingerprint is an OR over conformations, it records what a molecule can present without committing to a particular pose. This is primarily a property of molecular constitution, making the task more tractable than conformer prediction.

Error is governed by training-set coverage, and we expect this relationship to generalize beyond PharmCast. The difference between versions was training-set composition, which Online Resource 1, Fig. S3 traces across the version series. Additional screening collection data filtered at 600 Da did not improve the large-compound population, whereas adding ChEMBL chemistry did.

Coverage is not a proxy for size. Section 3.4 measures agreement against molecular weight: ChEMBL compounds of 300 to 400 Da read 0.858 against 0.887 for screening collection compounds of the same size, so at equal mass the gap reflects training-set coverage and chemical class. A single accuracy figure over a pooled evaluation set would have concealed all of this.

These results support reporting performance stratified by the property associated with failure and distributing the applicability domain with each model. The model card records the mass range, training-set composition, and measured degradation so that users can evaluate the boundary without consulting this paper.

The peptide set may also prove useful for more than training. The loops were extracted from deposited protein structures and are more conformationally flexible than screening collection compounds. They provide potential references for peptide mimicry, in which a small molecule is compared with the pharmacophoric features of a protein loop.

When predicted similarity is the objective of a search, the compounds that rise to the top include those the model overestimates, so the scores of the winners are optimistic. The leading candidates should therefore be rescored with the reference calculation. In a de novo design campaign against orforglipron run with ChIP, the synthesis-aware design protocol

introduced in [12] and since refined, PharmCast ranked every candidate and the survivors were then recomputed with the reference calculation, so each similarity reported for that campaign comes from the reference calculation: the delivered design matches the reference at 0.810 pharmacophore similarity while sharing 0.145 two-dimensional similarity with it, and in the refinement stage the best 250 candidates by PharmCast were rescored, the best of them measuring 0.797 against the parent's 0.754. We therefore recommend using the surrogate to rank candidates and rescoring a broad set of leading candidates with the reference calculation before selection.

5. Limitations

The model predicts the ORed ensemble fingerprint. Where an application requires the matching geometry of a particular conformer, that must come from the reference calculation, which retains per conformer records; the surrogate stands in for the ensemble, not for the conformers behind it. Bitwise aggregation also limits the descriptor itself: because bits are ORed across conformers, two bits set in one fingerprint need not be presentable in a single geometry. An ensemble fingerprint records what a molecule can present across its accessible conformations, not what it presents at once, and a similarity between two of them should be read that way.

Performance remains lowest outside screening-collection chemistry, particularly for the diversity of activity-backed ChEMBL compounds. Section 3.4 attributes this degradation to limited training-set coverage. Above 700 Da the training set is almost entirely peptidic, and ChEMBL agreement there falls to 0.784, its lowest in any mass band.

Predicted similarities have a small positive bias caused by systematic overprediction of set bits, with a median signed error of +0.002 on screening collection chemistry and +0.025 on the post-seal ChEMBL population. The bias grows with distance from the training set, and while it does not affect ranking, it should be considered when applying absolute thresholds.

A binary Morgan fingerprint of radius 2 saturates on a molecule built from repeated identical units: across five homologous series, 5 of 15 adjacent pairs have a Morgan block that is bit for bit identical, among them Gly3 against Gly4 and hexyl against octyl 4-alkylbenzoic acid. The input is not degenerate on those pairs, because six to ten of the eleven descriptors differ and nine of the eleven grow with the molecule, and the prediction separates them as the reference does: the reference pairwise Tanimoto reads 0.37, 0.39, 0.61, 0.81, 0.71 against 0.42, 0.39, 0.66, 0.79, 0.81 predicted (Online Resource 1, Tables S7 and S8). The saturation also measures what a two-dimensional fingerprint cannot see: where the Morgan block cannot separate two molecules at all, the reference pharmacophore fingerprint still reads them 0.37 to 0.81 apart, so two-dimensional similarity is not a substitute for the three-dimensional answer.

The screening collection has a median nearest-neighbor similarity of 0.714, so held-out molecules drawn from it are not fully independent of the version 10 training set.[26] For the version 10 screening collection test set the median nearest training neighbor is 0.62, and agreement increases from 0.829 for molecules in the 0.35 to 0.50 band to 0.909 in the 0.70 to 1.00 band. Table 4 reports this for all three populations. The in-domain values therefore describe performance on chemistry of the same kind, not on unseen chemistry. The peptide set and the collection differ in size and chemical composition, so performance on one should not be assumed to generalize to the other.

All results reported here are computed on real screening collection compounds, on peptides extracted from experimentally determined structures, and on real ChEMBL entries. No molecules were generated for evaluation.

Statements and Declarations

Funding

This work received no external funding. It was supported by Eidogen-Sertanty, Inc.

Author contributions

S.M.M. conceived the study, constructed the source sets and surrogate, performed the evaluation, and wrote the manuscript. M.J.M. co-developed the pharmacophore fingerprinting method on which this work builds and contributed to the analysis and manuscript.

Competing interests

S.M.M. is founder and CEO of Eidogen-Sertanty, Inc., which develops the PharmCast and ChIP software described here. M.J.M. declares no competing interests.

Data availability

The PharmCast package, model utilities, and documentation are available at https://github.com/smuskal/pharmcast under Apache-2.0, which covers both code and model weights. Trained weights are distributed from https://pharmcast.ai. The ChIP campaign reported in the Discussion is described at https://www.eidogen.com/chip2026_metoo.php.

The repository does not include the reference pharmacophore fingerprint generator or its toolchain and does not redistribute the training sets. Screening collection and building-block data are supplied under Enamine's terms. ChEMBL-derived data are licensed under CC BY-SA 3.0.

Each model checkpoint records its training-set composition, mass range, and applicability domain. These values are displayed by the card subcommand of the PharmCast command-line tool applied to the version 10 checkpoint, pharmcast_scp_v10.pt.

The released artifacts carry what is needed to reproduce the predictions. The package is released under a version tag, the version 10 checkpoint is published with its SHA-256 alongside an append-only checksum manifest, and a verify subcommand checks a local checkpoint against a shipped set of reference fingerprints and refuses to run under an RDKit release outside the supported range. That guard is load bearing rather than defensive: RDKit changed aromaticity perception for rings carrying an exocyclic carbonyl, which changes the Morgan block, which is model input, so the predicted fingerprint depends on the RDKit release. The identifier lists defining the three test populations of Table 2 are available from the corresponding author.

Ethics approval

Not applicable. This study involved no human participants, no animals and no clinical data. Consent to participate and consent for publication are therefore not applicable. A preprint of this work is deposited at bioRxiv.

References

1. Ehrlich P (1909) Über den jetzigen Stand der Chemotherapie. Ber Dtsch Chem Ges 42:17-47

2. Schueler FW (1960) Chemobiodynamics and drug design. McGraw-Hill, New York

3. Kier LB (1967) Molecular orbital calculation of preferred conformations of acetylcholine, muscarine, and muscarone. Mol Pharmacol 3:487-494

4. Güner OF, Bowen JP (2014) Setting the record straight: the origin of the pharmacophore concept. J Chem Inf Model 54:1269-1283

5. Gund P (1977) Three-dimensional pharmacophoric pattern searching. Prog Mol Subcell Biol 5:117-143

6. Sheridan RP, Nilakantan R, Rusinko A, Bauman N, Haraki KS, Venkataraghavan R (1989) 3DSEARCH: a system for three-dimensional substructure searching. J Chem Inf Comput Sci 29:255-260

7. Pickett SD, Mason JS, McLay IM (1996) Diversity profiling and design using 3D pharmacophores: pharmacophore-derived queries (PDQ). J Chem Inf Comput Sci 36:1214-1223

8. McGregor MJ, Muskal SM (1999) Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. J Chem Inf Comput Sci 39:569-574

9. McGregor MJ, Muskal SM (2000) Pharmacophore fingerprinting. 2. Application to primary library design. J Chem Inf Comput Sci 40:117-125

10. Mason JS, Morize I, Menard PR, Cheney DL, Hulme C, Labaudiniere RF (1999) New 4-point pharmacophore method for molecular similarity and diversity applications. J Med Chem 42:3251-3264

11. Deanda F, Stewart EL (2004) Application of the PharmPrint methodology to two protein kinases. J Chem Inf Comput Sci 44:1803-1809

12. Schürer SC, Tyagi P, Muskal SM (2005) Prospective exploration of synthetically feasible, medicinally relevant chemical space. J Chem Inf Model 45:239-248

13. Schneider G, Neidhart W, Giller T, Schmid G (1999) Scaffold-hopping by topological pharmacophore search: a contribution to virtual screening. Angew Chem Int Ed 38:2894-2896

14. Böhm H-J, Flohr A, Stahl M (2004) Scaffold hopping. Drug Discov Today Technol 1:217-224

15. Hu Y, Stumpfe D, Bajorath J (2017) Recent advances in scaffold hopping. J Med Chem 60:1238-1246

16. Wolber G, Langer T (2005) LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J Chem Inf Model 45:160-169

17. Leach AR, Gillet VJ, Lewis RA, Taylor R (2010) Three-dimensional pharmacophore methods in drug discovery. J Med Chem 53:539-558

18. Riniker S, Landrum GA (2015) Better informed distance geometry: using what we know to improve conformation generation. J Chem Inf Model 55:2562-2574

19. Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT (2010) Conformer generation with OMEGA. J Chem Inf Model 50:572-584

20. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742-754

21. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R, Hirzel T, Aspuru-Guzik A, Adams RP (2015) Convolutional networks on graphs for learning molecular fingerprints. Adv Neural Inf Process Syst 28:2215-2223

22. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE (2017) Neural message passing for quantum chemistry. In: Proceedings of the 34th international conference on machine learning, pp 1263-1272

23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235-242

24. Zdrazil B, Felix E, Hunter F et al (2024) The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res 52:D1180-D1192

25. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442-451

26. Mattsson B, Walters WP (2026) Identifying and addressing systematic data leakage in protein-ligand affinity benchmarks. bioRxiv 2026.06.29.735309

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf