# Predicting three-dimensional pharmacophore fingerprints from constitution alone

S. M. Muskal

*Eidogen-Sertanty, Inc.*

Technical report · PharmCast

**ABSTRACT**

Pharmacophore fingerprints describe which arrangements of chemical features a molecule can present in three dimensions, and they are among the most useful descriptors available for scaffold hopping, focused library design and similarity searching. Their cost is dominated not by the fingerprint but by the conformer ensemble it requires: of the 2.48 s needed to fingerprint one catalogue molecule over 100 conformers, 2.44 s is conformer generation and only 0.043 s is the fingerprint itself. We describe *PharmCast*, a feedforward model that predicts the complete 10,560-bit ensemble fingerprint directly from the molecule's two-dimensional structure, with no conformer generation at any stage. Trained on 1,192,657 catalogue molecules whose ensemble fingerprints were computed conventionally, PharmCast reproduces the held-out fingerprint at a median per-molecule Matthews correlation of 0.890. On the quantity that is actually consumed downstream, the pharmacophoric similarity between two molecules, it agrees with the reference calculation at Pearson 0.964 with a median absolute error of 0.019, and it orders candidate pairs correctly 92.1% of the time overall and 99.63% when the true difference exceeds 0.10. A complete comparison of two molecules from structure alone takes 0.034 ms against 5.0 s for the conventional route. We report where the approximation holds, where it degrades, and a measurement of the reference calculation's own reproducibility that bounds what any surrogate could achieve.

**Keywords:** pharmacophore fingerprint, PharmPrint, PolyPharmPrint, conformer ensemble, virtual screening, scaffold hopping, surrogate model.

## 1. Introduction

A pharmacophore fingerprint encodes the three-dimensional arrangements of chemical features that a molecule is capable of presenting. In the PharmPrint formulation [1,2] the basis is a set of three-point pharmacophores: every legal triangle of three typed features held at three binned distances contributes one bit. Features are assigned from constitution rather

than from a docked pose, and comprise acceptor, donor, negative, positive, hydrophobic, aromatic and other. Distances are binned into six ranges spanning 2.0 to 24.0 Å. The resulting vector has 10,560 bits.

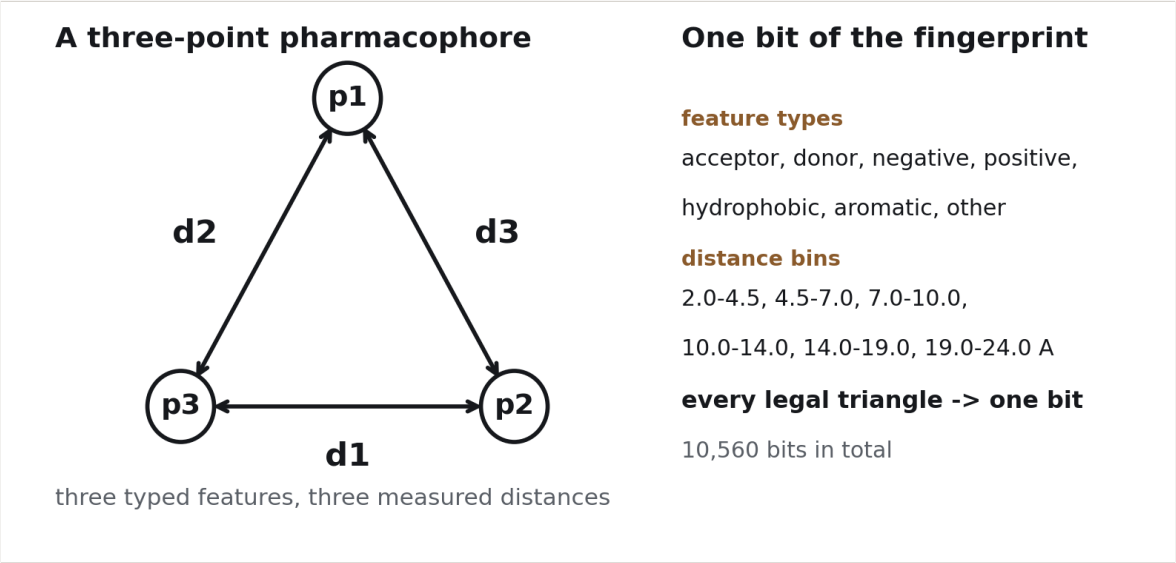


**Figure 1.** The unit of the descriptor. Three typed features held at three measured distances form one three-point pharmacophore; binning the distances makes it a discrete object, and each legal combination of three types and three bins is one bit of the fingerprint.

Because a single conformer presents only one arrangement, the descriptor is made conformationally honest by building an ensemble. Two conventions are in use. The original PharmPrint approach builds many conformers, fingerprints each, and ORs the bits into one combined vector, which answers what a molecule *can* present across its accessible landscape. PolyPharmPrint instead keeps every conformer's fingerprint separate, so that a match is between a specific conformer of the query and a specific conformer of the hit, preserving the matching geometry. This work targets the first convention: one ensemble fingerprint per molecule, ORed over 100 conformers.

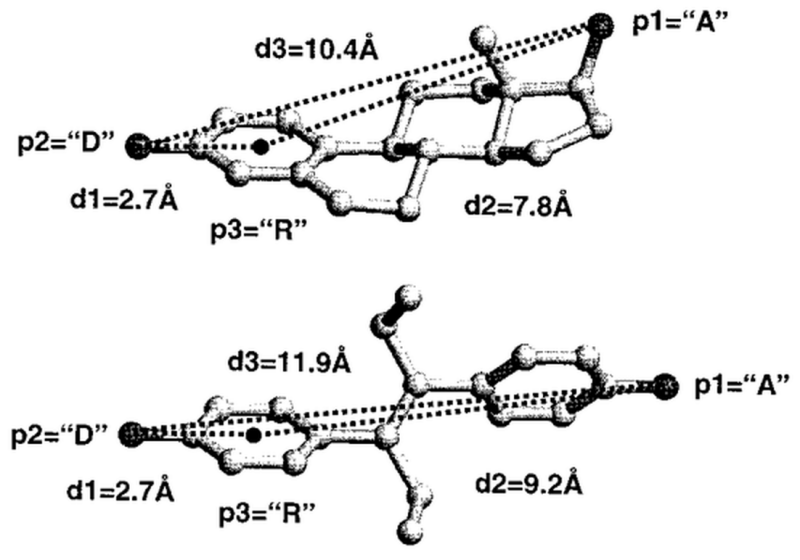


**Figure 2.** A mapping of one three-point pharmacophore onto (i) estradiol (top), the natural ligand, and (ii) diethylstilbestrol (bottom). The hydroxyl groups are both hydrogen bond donors and acceptors (D, A), and the centroid of the aromatic ring is shown (R). The two molecules share no scaffold, yet present the same three features at distances that fall in the same bins, so the same bit is set in both. Reproduced from reference 1.

The ensemble is where the cost lives. Measured over 100-conformer builds, conformer generation and optimisation account for 2.44 s of the 2.48 s total, while the fingerprint calculation itself accounts for 0.043 s. The descriptor is therefore roughly fifty times cheaper than the geometry it is computed on. That imbalance is what motivates a surrogate: if the ensemble fingerprint can be predicted from constitution, the entire conformational stage can be skipped for any application that consumes similarities rather than geometries.

We report such a model. It is not a replacement for the conventional calculation, which remains the arbiter for anything acted upon; it is a filter that allows far more chemistry to be placed in front of that arbiter.

## 2. Methods

### *2.1 Reference calculation*

Ground truth throughout is the conventional route. For each molecule, 100 conformers are generated by distance geometry using experimental torsion-angle preferences (the ETKDG method [3], version 3 [4]), starting from the isomeric line notation for the structure, with hydrogens added and the largest fragment retained. Each conformer is then relaxed against the Universal Force Field [5], as implemented in RDKit [7], and the resulting ensemble is passed through the reference fingerprint implementation. The per-conformer fingerprints are

ORed into one 10,560-bit ensemble vector. The same recipe, with the same fixed seed, is used for every training molecule and every evaluation molecule, so the surrogate is never asked to reproduce a protocol it was not trained on.

### *2.2 Training data*

Training molecules are drawn from a commercial screening collection [8] of 4,612,044 compounds. The collection is broad in scaffold terms, with 135,768 distinct Bemis-Murcko scaffolds in a 200,000 molecule sample of which 87.6% appear exactly once, and simultaneously dense in close analogues, which is characteristic of enumerated catalogue chemistry. The model reported here was trained on 1,192,657 molecules.

### *2.3 Peptide corpus, and the distinction between an observed and a proliferated presentation*

The second corpus is drawn from experimentally determined protein structures [10], and it exists to extend the model beyond catalogue chemistry into a class of molecule that is unquestionably real. The Protein Data Bank [10] was queried at a resolution limit of 2.5 Å and a free R-factor limit of 0.22, which returned **71,018** entries. Of these **70,818** parsed and **69,450** yielded at least one usable loop.

A loop is defined as a run of two to six residues lying between two annotated elements of secondary structure. Extraction found **1,524,535** loop instances across **135,408** chains, a median of fifteen per entry. Each is capped using the real flanking atoms present in the structure rather than with invented groups, so the capped molecule remains a description of something observed. Backbone continuity is enforced by requiring the carbon to nitrogen distance across each junction to fall between 1.0 and 2.0 Å; this check rejected **54,520** candidates that would otherwise have carried chain breaks, alternate conformations or missing density into the corpus, and it is not optional. Collapsing the survivors gives **117,741** distinct sequences and **133,136** distinct capped molecules.

The same sequence frequently appears many times across the Protein Data Bank, and its instances are not equivalent. This is worth stating plainly because the opposite is widely assumed: for short loops, sequence does not fix structure. A loop of a given sequence adopts different backbone geometries in different structures, in different chains of the same structure, and in different crystal forms, and each of those geometries presents a different arrangement of pharmacophoric features. Measured across **1,481** sequences observed three or more times, the median pharmacophoric agreement between two observations of the *same*

sequence is **0.878**, with 18% of comparisons below 0.70 and 19% effectively identical. Sequence identity therefore does not imply pharmacophoric identity, and this spread is signal rather than noise.

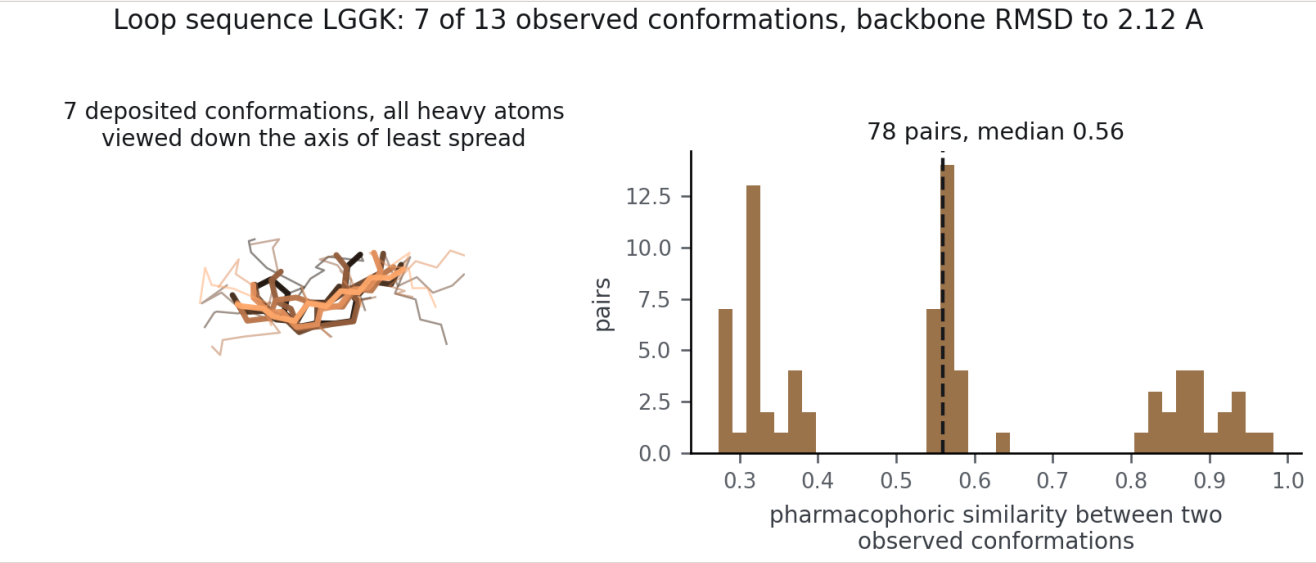


**Figure 3.** Sequence does not determine pharmacophoric presentation. Searching 19,108 loop sequences for the largest divergence between their own observed conformations returns LGGK, shown here as observed in **13** independent Protein Data Bank entries, all of them determined at 1.8 Å or better with a free R-factor at or below 0.20. Left, 7 of those conformations drawn with all heavy atoms and superimposed on the backbone; they disagree by 2.12 Å of backbone root-mean-square deviation. Right, the pharmacophoric similarity between every pair of the observed conformations, 78 pairs, with a median of 0.56, of which 72% fall below 0.70. The distribution separates into distinct populations with gaps between them, so this sequence does not adopt one presentation with noise around it, it adopts several. Across the search, backbone deviations reach 3.1 Å and all-atom deviations 6.3 Å. The conformations drawn are the most pharmacophorically distinct rather than the best resolved, since selecting by resolution returns the same crystal form repeatedly.

The obvious objection is that this variation is an artefact of crystallographic quality: poorly resolved or loosely refined loops would fit badly and appear to differ. That objection is testable, and it fails. Repeating the search over only those instances determined at 1.8 Å or better with a free R-factor of 0.20 or below leaves 9,218 sequences, and the largest backbone divergence found is **3.11 Å**, unchanged from the unrestricted search. All-atom divergence reaches 5.23 Å. The sequence shown in Figure 3 is drawn entirely from that restricted set: its 13 conformations have a median resolution of 1.65 Å, the best at 1.00 Å, and a median free R-factor of 0.189. These are well-determined structures that genuinely disagree.

**Table 1.** The most conformationally divergent loop sequences, restricted to instances at 1.8 Å or better with a free R-factor of 0.20 or below.

| SEQUENCE | CONFORMATIONS | BACKBONE RMSD | ALL-ATOM RMSD | MEDIAN RESOLUTION | MEDIAN R-FREE |
|---|---|---|---|---|---|
| SGGGG | 35 | 3.11 | 3.80 | 1.55 | 0.163 |
| LGGK | 49 | 2.59 | 4.38 | 1.65 | 0.189 |
| GSGL | 34 | 2.50 | 4.05 | 1.54 | 0.185 |
| GGGK | 12 | 2.43 | 5.23 | 1.50 | 0.159 |
| GPSG | 32 | 2.38 | 2.75 | 1.47 | 0.182 |
| AGGK | 9 | 2.36 | 4.32 | 1.77 | 0.196 |
| DLAV | 12 | 2.34 | 3.97 | 1.50 | 0.188 |
| AGAV | 13 | 2.27 | 3.07 | 1.50 | 0.180 |

Two fingerprints can consequently be defined for a peptide, and the difference matters.

The **rigid** fingerprint is computed on a single conformation exactly as it was observed, with no conformer generation of any kind. Every pharmacophore it records was physically present in a crystal structure. It answers what that loop *was* doing in that context, and it is the more conservative object: it cannot contain an arrangement that was never seen.

The **proliferated** fingerprint is the union of two things. It ORs together every observed conformation of that sequence, so that all the crystallographic evidence for the loop is retained, and it additionally ORs in 100 computed conformers of the same capped molecule. The result is a description of what that peptide *can* present, anchored on what it has been seen to present. Because both sources contribute, the effective conformer count exceeds one hundred. The observed geometries are never discarded in favour of the computed ones; the computed ones fill in the accessible space around them.

Repeated sequences are kept rather than deduplicated. How often a loop sequence appears, and in how many distinct geometries, is information about how that sequence actually occupies space, and collapsing it to one representative would discard exactly the evidence the corpus was built to carry.

Peptide ensembles use the identical fingerprint convention and the same conformer recipe as the screening collection, which is the only reason the two chemistries can be trained and evaluated in one model.

### *2.4 Model*

The input is a 2048-bit binary Morgan fingerprint of radius 2 [6], a connectivity descriptor computed from the two-dimensional structure alone. The network is a feedforward multilayer perceptron with two hidden layers of 1024 and 512 units and a 10,560-unit sigmoid output, one unit per pharmacophore bit, trained with binary cross-entropy. Predicted bits are thresholded at 0.5. No three-dimensional information of any kind enters the model at training or inference time.

### *2.5 Evaluation protocol*

Evaluation molecules are those fingerprinted after the training snapshot was taken, so the test set is defined by time rather than by sampling: 150,912 molecules from chunks 2410-2711. Three quantities are reported. Per-molecule fidelity is the Matthews correlation between the predicted and reference bit vectors. Pairwise fidelity is the agreement between the Tanimoto similarity computed from two predicted fingerprints and the same quantity computed from the two reference fingerprints, over 30,000 pairs. Ranking accuracy is the fraction of candidate pairs placed in the correct order, reported as a function of the true similarity difference, since pairs separated by less than the calculation's own noise are not meaningfully rankable by anything.

### *2.6 Domain of applicability*

A critical assumption underlies every result reported here, and it defines what the model is entitled to be asked. Every molecule used for training and for evaluation is one that has actually been made or actually observed: compounds held in a commercial screening collection, and peptides extracted from experimentally determined protein structures. The model has therefore only ever been shown chemistry that is synthetically real, and a degree of synthetic feasibility is built into its training distribution rather than imposed on it afterwards.

An arbitrary two-dimensional structure that has never been synthesised is consequently outside the domain the model was fitted on. Such a molecule is not necessarily excluded, and the model will return an answer for it, but that answer is an extrapolation and should be treated as one. This also disposes of a control that would otherwise be conventional. Comparing against a predictor that has been shown no ligand at all presumes that molecules can be drawn from some unconstrained space and fed to the model, which is not the setting: the relevant question is not whether the model beats knowing nothing, but whether it is doing something a cheap two-dimensional descriptor could already do. Section 3.2 answers that question directly.

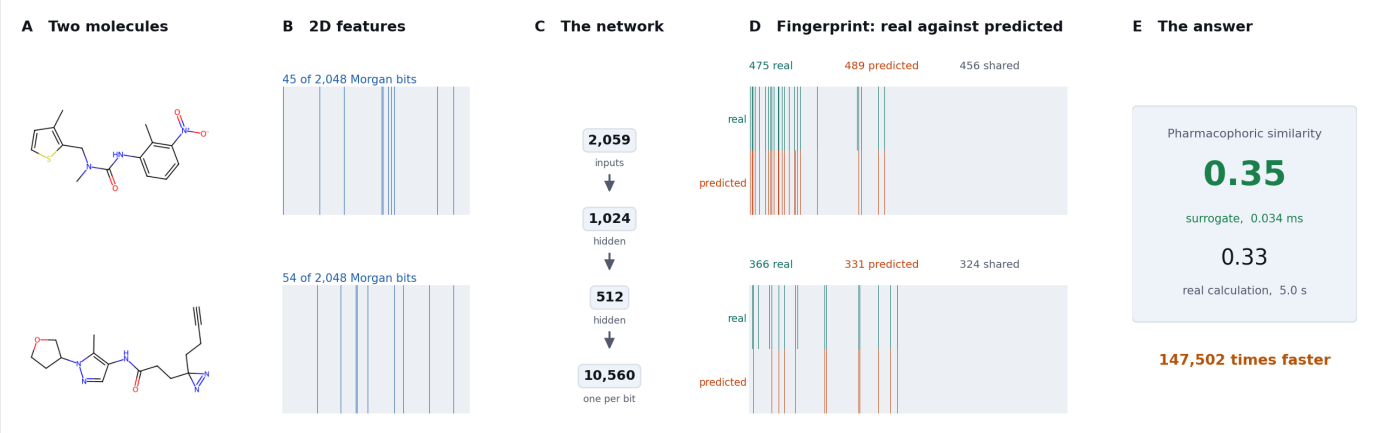


**Figure 4.** Two molecules carried through every stage. Panel A, the input structures. Panel B, the 2D feature vectors that are the model's only input. Panel C, the network. Panel D, each predicted fingerprint against its reference on the same bit axis. Panel E, the pharmacophoric similarity, which is the quantity actually consumed downstream.

# 3. Results

## *3.1 Fidelity*

On 150,912 held-out molecules the median per-molecule agreement between the predicted and reference fingerprints is 0.890.

The quantity that matters more is the similarity between two molecules. Across 30,000 held-out pairs the predicted and reference similarities agree at Pearson 0.964 and Spearman 0.964, with a median absolute error of 0.019 (Figure 2).

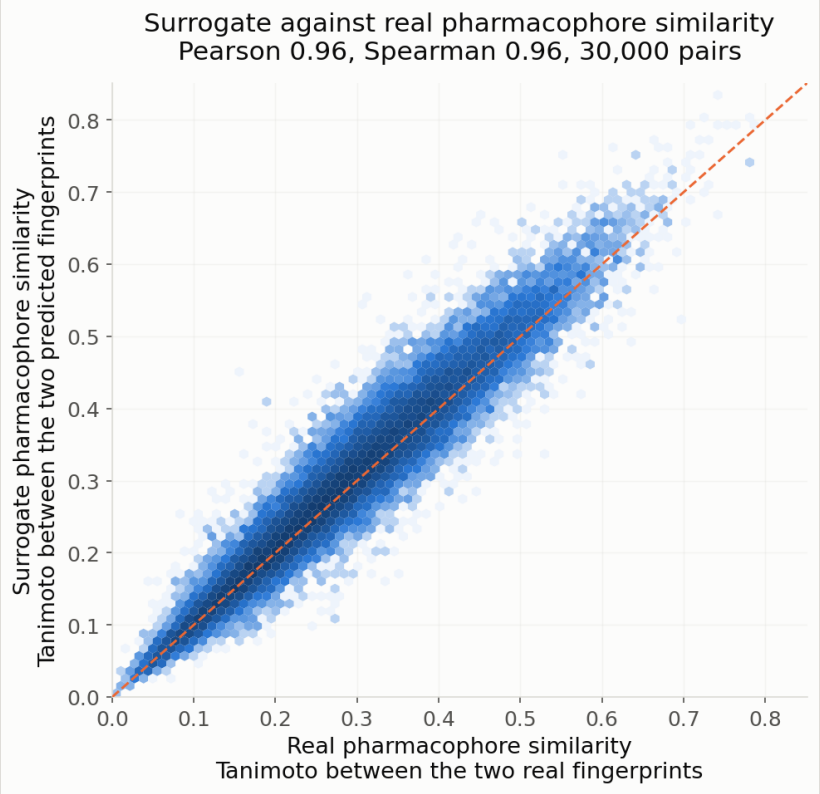


**Figure 5.** Predicted against reference pharmacophoric similarity for 30,000 held-out pairs. The diagonal marks perfect agreement.

### *3.2 The model is not re-deriving two-dimensional similarity*

Since the input is a two-dimensional descriptor, the obvious concern is that the model has learned a monotone function of two-dimensional similarity and is adding nothing three-dimensional. It has not. Across the held-out pairs, the Morgan similarity of a pair correlates with the reference pharmacophoric similarity of the same pair at only 0.24, and the best straight-line predictor of pharmacophoric similarity from Morgan similarity alone carries a median absolute error of 0.081, against PharmCast's 0.019.

The decisive observation is that accuracy does not depend on how alike the two molecules look on paper (Table 2). Pairs that are essentially unrelated in two dimensions are predicted as accurately as pairs that share obvious features. That is the property scaffold hopping requires, and a model that had merely learned two-dimensional similarity could not have it. It is worth adding that the held-out pairs are overwhelmingly dissimilar to begin with: the median pair scores 0.12 and the 99th percentile 0.24, so within this collection "more alike" means "less unlike".

**Table 2.** Prediction accuracy as a function of the two-dimensional Morgan similarity of the pair being compared.

| MORGAN SIMILARITY OF THE PAIR | PAIRS | MEDIAN ERROR | CORRELATION |
|---|---|---|---|
| 0.00 to 0.10 | 7,261 | 0.018 | 0.958 |
| 0.10 to 0.15 | 15,055 | 0.019 | 0.963 |
| 0.15 to 0.20 | 6,288 | 0.020 | 0.964 |
| 0.20 to 1.01 | 1,396 | 0.020 | 0.965 |

### *3.3 Ranking and retrieval*

Most applications consume an order rather than a value. Table 1 reports ranking accuracy as a function of the true similarity difference between the two candidates being compared.

**Table 3.** Ranking accuracy against the true similarity difference, over 2,999,828 comparisons drawn from the held-out set.

| TRUE DIFFERENCE | COMPARISONS | ORDERED CORRECTLY |
|---|---|---|
| any | 2,999,828 | 92.1% |
| greater than 0.05 | 2,308,250 | 98.1% |
| greater than 0.10 | 1,672,454 | 99.63% |
| greater than 0.20 | 719,525 | 99.98% |

In a retrieval setting over 101,444 candidates and 30,000 queries, the true nearest neighbour has a median rank of 5 under the surrogate, and appears within the surrogate's top 100 for 87.5% of queries and within its top 500 for 95.9% (Figure 6). This is the operationally relevant behaviour: the surrogate is used to select a shortlist that the reference calculation then adjudicates.

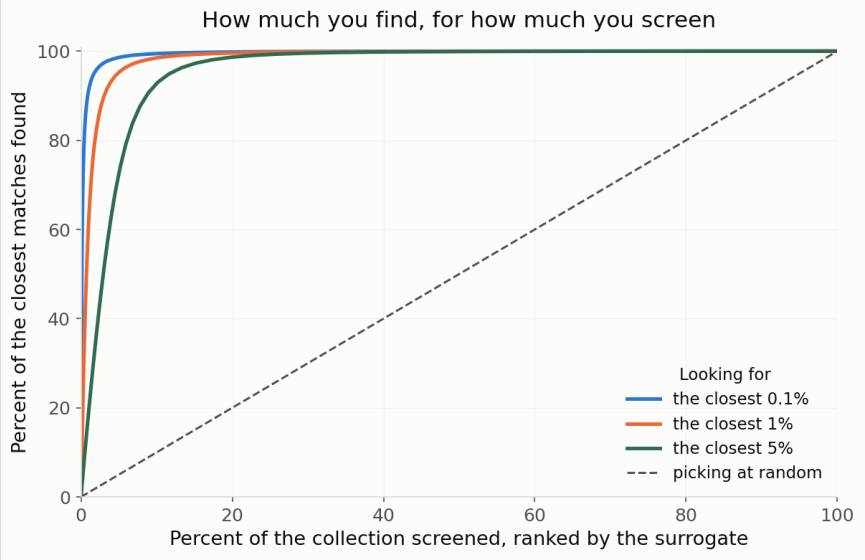


**Figure 6.** Fraction of the genuinely closest chemistry recovered after screening a given fraction of a collection by surrogate score. Left, full range; right, the same curves on a logarithmic depth axis, where selection decisions are made. The dashed line is screening in arbitrary order.

### *3.4 Throughput*

Table 4 gives the cost of each stage. The surrogate removes the conformational stage entirely rather than accelerating it, which is why the speedup is of a different order to what is normally achievable by optimisation.

**Table 4.** Time to obtain one ensemble fingerprint, and to compare two molecules starting from structure alone.

| STAGE | SECONDS PER MOLECULE |
|---|---|
| Conformer generation and optimisation, 100 conformers | 2.44 |
| Reference fingerprint over the ensemble | 0.043 |
| **Conventional route, total** | **2.48** |
| PharmCast, one molecule at a time | 0.000357 |
| PharmCast, batched | 1.07e-05 |

Batched inference sustains 46,516 molecules per second on a single machine, a speedup of 6,942-fold per molecule and 231,694-fold batched. A complete pharmacophoric comparison of two molecules, from two line notations to one similarity value, takes 0.034 ms against 5.0 s conventionally.

### *3.5 Behaviour across chemistries*

Catalogue chemistry is not the only domain of interest. Table 5 reports the composite model pharmcast_sp_v3.pt, trained on **2,056,482** collection molecules together with **58,039** ensemble-enhanced protein loop peptides, evaluated on held-out pairs from three chemistries: the screening collection, loop peptides whose sequences the model never saw, and real ChEMBL compounds above 600 molecular weight, each carrying a measured activity value, which lie outside the size range the collection covers (Figure 7).

**Table 5.** Agreement with the reference calculation across three chemistries, 1,200 held-out pairs each.

| TEST CHEMISTRY | MEDIAN ERROR | CORRELATION |
|---|---|---|
| Screening collection | 0.019 | 0.969 |
| Protein loop peptides | 0.020 | 0.984 |
| Large ChEMBL compounds | 0.071 | 0.615 |
| **All three combined** | **0.027** | **0.893** |

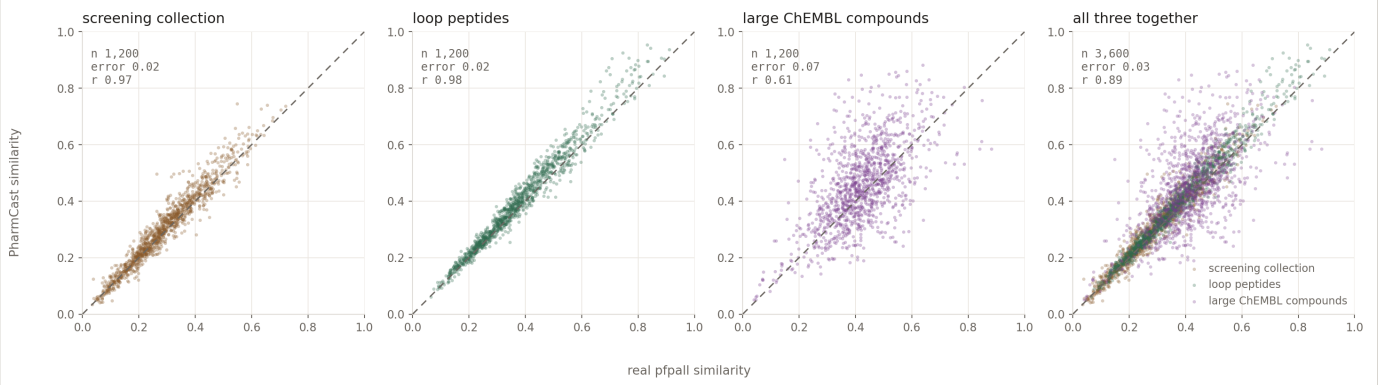


**Figure 7.** Reference against predicted similarity on held-out pairs from three chemistries and then all together. The combined panel should be read as a statement about coverage rather than as a single accuracy, since it blends two chemistries the model handles well with one it handles coarsely.

### *3.6 The limit: molecular size*

The clearest failure mode is size. Table 6 compares performance on catalogue chemistry against real ChEMBL compounds above 600 molecular weight, using an identical protocol on both. The two similarity landscapes have essentially the same spread, 0.114 against 0.118, so the degradation is attributable to size rather than to one task being intrinsically harder.

**Table 6.** The same protocol applied to catalogue chemistry, median 347 molecular weight, and to 573 real compounds above 600, median 638. The two similarity landscapes have comparable spread, 0.114 and 0.118, so the difference is attributable to size rather than to one task being intrinsically harder. All three models are shown because the degradation is a property of the descriptor's difficulty at this size and not of one model.

| MODEL | CATALOGUE ACCURACY | CATALOGUE ERROR | LARGE ACCURACY | LARGE ERROR |
|---|---|---|---|---|
| PharmCast-S, collection only | 90.9% | 0.019 | 72.2% | 0.072 |
| PharmCast-SP, composite | 90.9% | 0.021 | 71.6% | 0.069 |
| PharmCast-SP v3, composite | 91.3% | 0.020 | 71.0% | 0.071 |

Every model loses roughly twenty points of ranking accuracy on the larger compounds, and adding the peptide corpus neither causes nor cures it.

### *3.7 What the reference calculation itself can support*

Any surrogate is bounded by the reproducibility of the quantity it approximates. We measured this directly by rebuilding 149 of the large molecules with a different conformer embedding seed, changing nothing else, and comparing the two independent ground truths. Pairwise similarity reproduces to a median absolute difference of 0.006 at Pearson 0.995. The individual fingerprints are noticeably less reproducible than the similarities computed from them, at a median self-similarity of 0.937 and a median self-agreement of 0.958, which indicates that the conformer-to-conformer variation is largely common-mode and cancels in a comparison.

This bounds the interpretation of Section 3.6. The reference calculation agrees with itself an order of magnitude more tightly than PharmCast agrees with it on large molecules, so that degradation is a genuine limitation of the surrogate and not an artefact of a noisy target.

## 4. Discussion

The intended use follows from the measurements. Where the true difference between two candidates is large, the surrogate orders them almost perfectly and can be trusted to build a shortlist. Where the difference is small, it cannot, and the reference calculation must adjudicate. The economics are favourable because the two regimes are cheap and expensive respectively: a surrogate pass over a multi-million compound collection costs minutes, and the reference calculation is then spent only where it decides the answer.

One failure mode deserves emphasis because it is not visible in any of the tables above. When a surrogate is used as the objective of an optimiser rather than as a filter, the optimiser will eventually exploit its error. In a design campaign that used this model as the scoring function

for a genetic algorithm, a seventeen-fold increase in search budget raised the median surrogate score of the retained designs by 0.20 while raising their median true score by 0.02, and the surrogate's own median error over those designs rose from +0.06 to +0.24. The search had become better at finding molecules the model was wrong about. The practical consequences are that the surrogate should be used to rank and never to decide, that a real rescore is a required stage rather than a formality, and that many survivors rather than a top handful should be carried into it.

## 5. Status of the training corpora

Both corpora are still being built, and the numbers in this report are a measurement of a moving object rather than of a finished one. At the time of writing, **48%** of the 4,612,044 compound screening collection has been fingerprinted by the conventional route, and **52%** of the protein loop corpus, 69,039 of 133,136 distinct capped peptides. Fingerprinting continues on both.

Every result here should therefore be read as a lower bound on what the approach supports rather than as its final performance, and this document is expected to be superseded as the corpora complete. The evaluation protocol is unchanged across rebuilds, and every figure and table is generated from the evaluation files, so a later version is directly comparable to this one.

## 6. Limitations

The model predicts the ORed ensemble fingerprint and therefore cannot support any application that requires per-conformer matching geometry, which is the regime PolyPharmPrint addresses. Performance degrades materially outside catalogue-like chemistry, most clearly with increasing molecular size, and the degradation is a property of the surrogate rather than of a noisy target. Predicted similarities carry a small positive bias arising from a systematic over-prediction of set bits; this does not affect ordering but does affect any absolute threshold. All results reported here are computed on real catalogue compounds and on peptides extracted from experimentally determined structures; no molecules were generated for evaluation purposes.

## References

1. McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569–574.

2. McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprinting. 2. Application to primary library design. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 117–125.
3. Riniker, S.; Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.
4. Wang, S.; Witek, J.; Landrum, G. A.; Riniker, S. Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *J. Chem. Inf. Model.* **2020**, *60*, 2044–2058.
5. Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.
6. Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
7. RDKit: open-source cheminformatics. https://www.rdkit.org
8. Enamine screening collection, June 2026 release. Enamine Ltd. https://enamine.net
9. Enamine building blocks, in-stock catalogue. Enamine Ltd. https://enamine.net
10. RCSB Protein Data Bank. https://www.rcsb.org

All figures and tables in this document are generated from the evaluation files at build time; no value is transcribed by hand.