Online Resource 1

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

Steven M. Muskal and Malcolm J. McGregor

Supplementary figures and tables are numbered separately from the main text. All results are for PharmCast version 10.

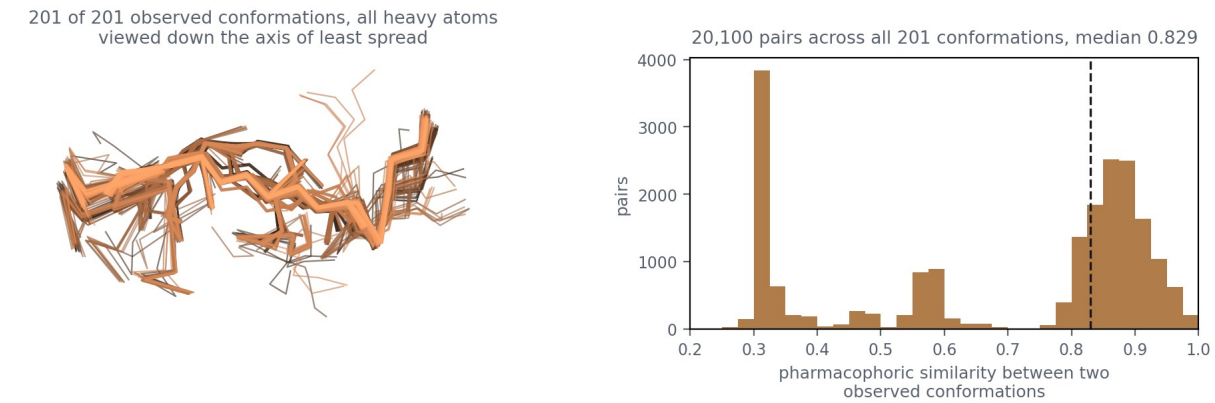


Fig. S1 One loop sequence and every deposited conformation of it. The sequence LGGK appears 217 times across 61 Protein Data Bank entries; the 201 instances sharing the same 25 heavy atoms are drawn superimposed on their backbones, viewed down the axis of least spread, with a maximum backbone RMSD of 2.59 Å. Their 20,100 pairwise comparisons have a median of 0.829. Those 20,100 comparisons fall into three families, near 0.31, near 0.58 and a dominant one near 0.87.

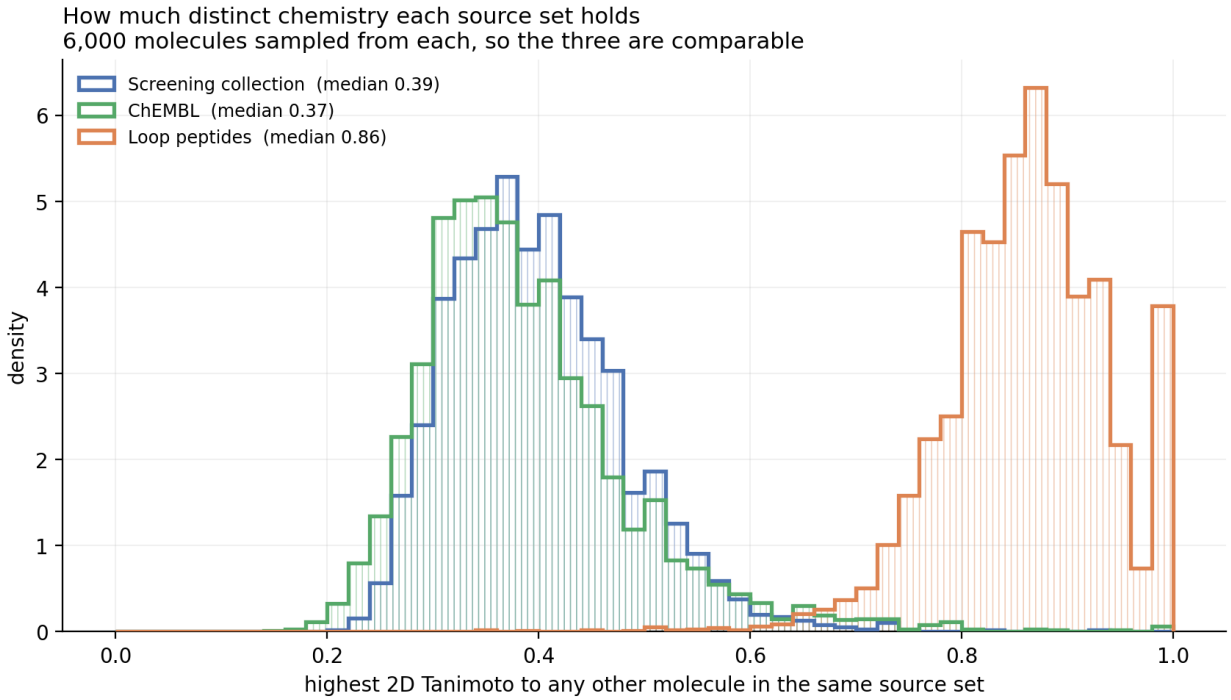


Fig. S2 How much distinct chemistry each source set holds in the version 10 training set. Each distribution uses 6,000 molecules, and each value is a molecule's highest two-dimensional Tanimoto to another molecule in the same source set. Medians are 0.389 for the screening collection, 0.368 for ChEMBL, and 0.864 for loop peptides.

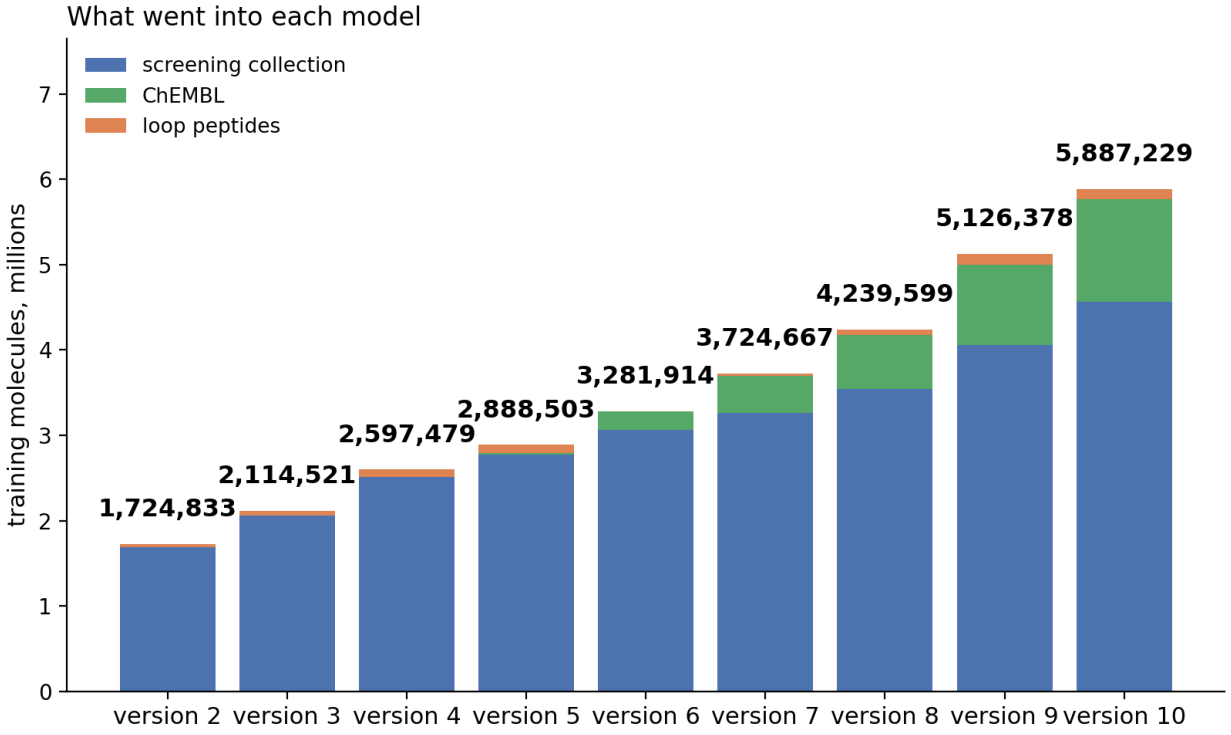


Fig. S3 Training-set composition and total count across the three source sets for each model. PharmCast version 4 is the last model without ChEMBL; PharmCast version 5 was the first to include all three source sets; the series ends at PharmCast version 10.

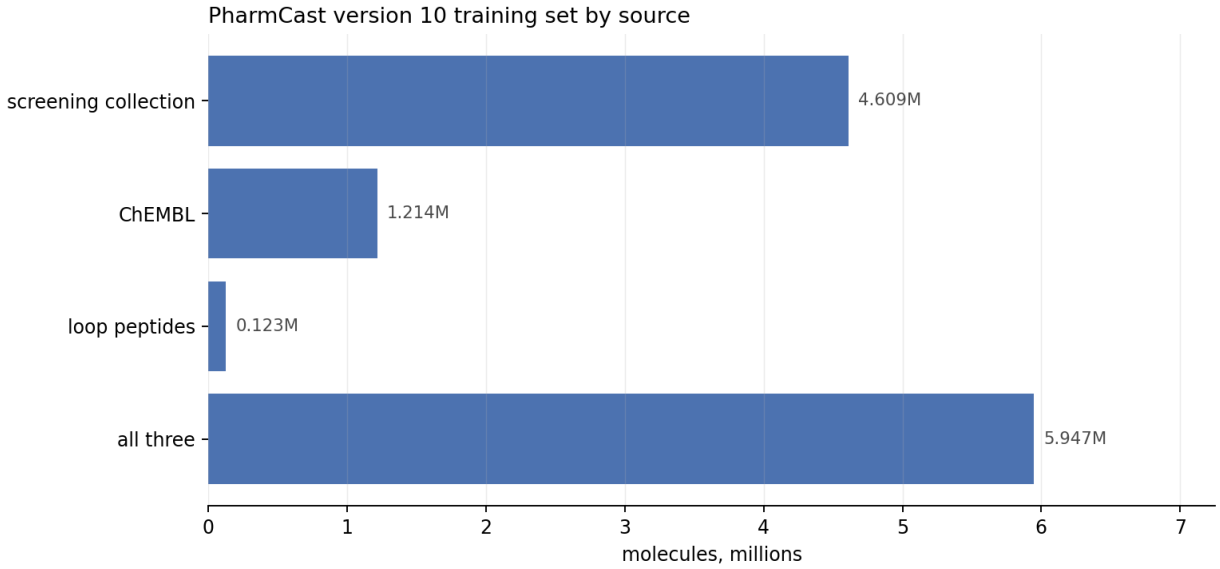


Fig. S4 The PharmCast version 10 training set by source. The screening collection contributes 4.609 million molecules, ChEMBL 1.214 million and the loop peptides 0.123 million, giving 5.947 million in all.

Table S1 Reference fingerprint coverage by source when PharmCast version 10 was trained.

Source	Reference fingerprints	Eligible molecules	Coverage
Screening collection	4,609,488	4,617,292	99.8%
ChEMBL	1,214,214	2,737,190	44.4%
Loop peptides	136,494	100%	
Total	5,960,196	7,490,976	79.6%

Table S2 Composition of the PharmCast version 10 training set by source and by mass. The pooled column counts every molecule available to the model, and the trained column counts those that remained after a molecular-weight-stratified one percent was held back for early stopping; the two differ by that holdback and are not alternative counts of the same population. The three sources cover 74 to 1000 Da between them. The screening collection is the bulk of the training set and is filtered at 600 Da on ingest, so the range above 600 Da is carried by ChEMBL and by the loop peptides.

Source	Pooled molecules	Molecules trained on	Share	Mass range	Median MW
Screening collection	4,609,488	4,563,262	77.51%	74 to 600 Da	344
ChEMBL, activity backed	1,214,214	1,202,183	20.42%	142 to 1000 Da	323
Protein loop peptides	122,994	121,784	2.07%	116 to 988 Da	577
Total	5,946,696	5,887,229	100%	74 to 1000 Da	341

Table S3 Molecules above 600 Da, by mass band, in the PharmCast version 10 training set. Counts come directly from the training set indexes.

Mass band	Molecules	Share of training set	Chemistry
600 to 700 Da	70,873	1.19%	ChEMBL 38,426 and peptides 32,452
700 to 800 Da	16,701	0.281%	peptides
800 to 900 Da	3,652	0.061%	peptides
900 to 1000 Da	162	0.003%	ChEMBL 5 and peptides 157
Total above 600 Da	91,388	1.54%	ChEMBL 38,426 and peptides 52,962

Table S4 Descriptor scope of the version 10 training set, 5th to 95th percentile of a 20,000-molecule sample.

Source	MW	MW median	Donors	Acceptors	Rotatable bonds
Screening collection	256 to 467	344	0 to 3	2 to 7	2 to 8
ChEMBL	217 to 378	323	0 to 3	2 to 7	1 to 8
Loop peptides	387 to 778	575	5 to 12	5 to 11	10 to 24

Table S5 Recall and precision for each pharmacophore, banded by how often the reference calculation sets it. Every one of the 10,549 bits was scored across the whole arm. Prevalence is the share of molecules in which the reference sets that bit, recall is the share of those molecules in which PharmCast also sets it, and precision is the share of molecules where PharmCast sets it and the reference agrees. The last column counts bits the reference sets at least once and PharmCast never recovers. Bits the reference never sets anywhere in an arm are excluded from the recall and precision medians, because neither is defined for them.

Arm	Reference prevalence	Pharmacophores	Share of bits	Median recall	Median precision	Never recovered
Screening collection	below 0.01%	2,037	19.3%	0.00	0.00	796
Screening collection	0.01% to 0.1%	1,864	17.7%	0.38	0.35	28
Screening collection	0.1% to 1%	2,586	24.5%	0.59	0.60	3
Screening collection	1% to 10%	2,410	22.8%	0.72	0.74	0
Screening collection	10% and above	1,652	15.7%	0.94	0.90	0
ChEMBL, 243 to 812 Da	below 0.01%	433	4.1%	0.00	0.00	260
ChEMBL, 243 to 812 Da	0.01% to 0.1%	1,162	11.0%	0.17	0.24	13
ChEMBL, 243 to 812 Da	0.1% to 1%	2,809	26.6%	0.39	0.49	0
ChEMBL, 243 to 812 Da	1% to 10%	3,595	34.1%	0.59	0.65	0
ChEMBL, 243 to 812 Da	10% and above	2,550	24.2%	0.83	0.79	0
Loop peptides	below 0.01%	358	3.4%	0.00	0.00	92
Loop peptides	0.01% to 0.1%	362	3.4%	0.00	0.00	294
Loop peptides	0.1% to 1%	1,240	11.8%	0.05	0.50	417
Loop peptides	1% to 10%	3,229	30.6%	0.37	0.68	55
Loop peptides	10% and above	5,360	50.8%	0.97	0.92	0

Table S6 Sensitivity of the reference fingerprint to conformer sampling depth, on 24 molecules drawn from the flexible tail of the screening collection test set, each carrying 8 or more rotatable bonds and spanning 154 to 977 Da. Each ensemble uses the same ETKDGv3 parameters, the same UFF step and the same pinned fingerprint binary as every corpus in this work, and differs only in how many conformers were requested. The corpus convention is 100. Median bit growth from 100 to 200 conformers is 2.5%. Because the ensemble fingerprint is a bitwise OR, a deeper ensemble can only add bits, and the last column confirms that every fingerprint at every depth is contained in the 200 conformer fingerprint.

Conformers requested	Median reference bits	Tanimoto to the 200 conformer fingerprint	PharmCast against that reference	Molecules contained in the deepest ensemble
25	1,114	0.89	0.81	24 of 24
50	1,171	0.94	0.83	24 of 24
100	1,287	0.98	0.84	24 of 24
200	1,324	1.00	0.84	24 of 24

Table S7 Behavior across homologous series, where a binary Morgan fingerprint of radius 2 can stop distinguishing successive members. Each row compares two adjacent members of one series. The third column counts how many of the 2,048 Morgan bits differ between them, and the fourth how many of the 11 whole molecule descriptors do. In 5 of the 15 pairs the Morgan block is bit for bit identical, and in every one of those the descriptors still separate the pair. The last two columns give the pairwise pharmacophore Tanimoto from the reference calculation and from PharmCast. Reference fingerprints for these molecules were generated by the same worker and the same pinned binary as the training corpora, at 100 conformers.

Series	Adjacent pair	Morgan bits differing	Descriptors differing	Reference Tanimoto	Predicted Tanimoto
biaryl	benzene vs biphenyl	6	7 of 11	0.15	0.08
biaryl	biphenyl vs p-terphenyl	1	7 of 11	0.32	0.44
bipyridine	pyridine vs 2,2'-bipyridine	13	9 of 11	0.15	0.13
bipyridine	2,2'-bipyridine vs 2,2':6',2''-terpyridine	4	9 of 11	0.32	0.27
phenol	phenol vs 4,4'-biphenol	8	10 of 11	0.20	0.23
glycine oligomer	glycine vs Gly2	10	8 of 11	0.04	0.22
glycine oligomer	Gly2 vs Gly3	2	8 of 11	0.22	0.21
glycine oligomer	Gly3 vs Gly4	0	8 of 11	0.37	0.42
glycine oligomer	Gly4 vs Gly6	0	8 of 11	0.39	0.39
4-alkylbenzoic acid	methyl vs ethyl	11	6 of 11	0.64	0.74
4-alkylbenzoic acid	ethyl vs butyl	10	6 of 11	0.62	0.54
4-alkylbenzoic acid	butyl vs hexyl	4	6 of 11	0.53	0.53
4-alkylbenzoic acid	hexyl vs octyl	0	6 of 11	0.61	0.66
4-alkylbenzoic acid	octyl vs decyl	0	6 of 11	0.81	0.79
4-alkylbenzoic acid	decyl vs dodecyl	0	6 of 11	0.71	0.81

Table S8 The individual molecules behind Table S7, with the number of pharmacophores the reference calculation sets, the number PharmCast predicts, and the agreement between them. Median agreement is 0.81 and the weakest is 0.27, on glycine, which carries three reference bits; a Tanimoto over a handful of bits moves a long way on one bit. These molecules sit well below the mass distribution PharmCast was trained on and are included to probe the input representation rather than to characterize the applicability domain, which Table 3 and Table 4 of the main text report.

Series	Molecule	Reference bits	Predicted bits	Tanimoto to reference
biaryl	benzene	2	1	0.50
biaryl	biphenyl	13	12	0.92
biaryl	p-terphenyl	41	27	0.66
bipyridine	pyridine	4	3	0.75
bipyridine	2,2'-bipyridine	27	23	0.61
bipyridine	2,2':6',2''-terpyridine	85	70	0.76
phenol	phenol	15	18	0.83
phenol	4,4'-biphenol	76	77	0.61
glycine oligomer	glycine	3	11	0.27
glycine oligomer	Gly2	47	45	0.77
glycine oligomer	Gly3	205	214	0.80
glycine oligomer	Gly4	540	511	0.81
glycine oligomer	Gly6	1,261	1,298	0.69
4-alkylbenzoic acid	methyl	21	23	0.91
4-alkylbenzoic acid	ethyl	33	31	0.88
4-alkylbenzoic acid	butyl	53	57	0.86
4-alkylbenzoic acid	hexyl	100	107	0.88
4-alkylbenzoic acid	octyl	142	162	0.88
4-alkylbenzoic acid	decyl	175	204	0.83
4-alkylbenzoic acid	dodecyl	246	248	0.84

Table S9 Ranking accuracy against training-set size on fixed populations held out from every model. The first population is 5,000 held-out screening collection molecules reserved from every model's training set; large ChEMBL is 573 real compounds above 600 Da.

Model	Screening collection	Large ChEMBL, n 573
PharmCast version 2	0.908	0.701
PharmCast version 3	0.912	0.710
PharmCast version 4	0.916	0.711
PharmCast version 5	0.917	0.746
PharmCast version 6	0.924	0.811
PharmCast version 7	0.934	0.832
PharmCast version 8	0.934	0.822
PharmCast version 9	0.935	0.817
PharmCast version 10	0.934	0.826

Table S10 Agreement between the reference calculation and PharmCast when both rank the same screen. Queries are ChEMBL compounds with a measured pChEMBL of 7 or better against the target, none present in the version 10 training set. Candidates are the remaining actives for that target together with compounds matched to each query on molecular weight, cLogP, donors, acceptors and rotatable bonds, drawn from the 155,648 Enamine test cases of Table 2. Every query is scored against every candidate by both routes. Spearman is over the whole ranking; the recovery columns ask how much of the reference calculation's top slice appears in PharmCast's. Property matching makes this a hard ranking, because every candidate is the size and shape of the query. The comparison is the reference calculation against PharmCast and nothing else.

Target	Queries	Candidates	Spearman	Reference top 1% recovered	Reference top 10 recovered	Reference's top actives found
VEGFR2 kinase	100	1,123	0.91	0.51	0.51	0.60
MAPK14, p38 alpha	100	1,816	0.91	0.52	0.51	0.63
Thrombin	95	1,427	0.86	0.43	0.41	0.53
Estrogen receptor alpha	100	2,037	0.88	0.61	0.56	0.62
mu opioid receptor	89	1,067	0.94	0.66	0.66	0.72
Mean over targets			0.90	0.54	0.53	0.62