Classification model for chemical checker database structural and bioactivity similarities

Maths4life report

Teodor Parella Dilmé

Abstract

In this study three different classification models for structural and bioactivity differences between Chemical Checker and Zinc15 ”Non-druglike” databases have been trained. Morgan fingerprints along Signaturizer’s A1 and B4 signatures have been initially computed and found suitable to build 2-fold cross validated classifiers, all using 40000 Zinc15 ”Non-druglike” and 10000 CC random compounds. Gradient boosting has performed best for Morgan fingerprints (ROC AUC = 0.94) and A1 signatures (ROC AUC = 0.87), whilst random forest has performed best with B4 signatures (ROC AUC = 0.93). Loss of classifying performance has been shown in the transition from Morgan fingerprints to A1 signatures. Eventually, information entropy of CC structural 2048 dimension Morgan fingerprints has been computed, and found to be 270.48 Bits.

1 Objectives

Chemical Checker (CC) [1] is a model able to encode valuable information of arround 1 million chemical compounds (structural, targets, biological pathways, cellular assays and clinical data) in a 128 dimension compressed vector for a given field.

An improvement of Chemical Checker aplications was made with Signaturizer [2], which relies in a siamese neural network (SNN) to extended signatures predictions for any given molecule outside the Chemical Checker database.

An external CC database that differs significantly in structural information or bioactivity wants to be found. By doing so, classifiers could be trained to distinguish CC similar molecules. Such molecules are expected to have better Signaturizer prediction due to proximity to SNN training dataset. Overall, by building a CC structural/bioactivity classifier, initially undesired molecules for further prediction could be rejected.

2 Structural signatures distribution

Structural molecular information has initially been computed in the form of Morgan fingerprints (dimension = 2048, information radius = 2) and Signaturizer’s A1 signatures for CC, Chembl [3][4] and Zinc15 [5] databases.

Plotting TSNE multidimensional projections of such molecular descriptors has resulted in no clear set separation. For Morgan fingerprints, highest 3 Tanimoto similarities have been computed for all groups to spot CC nearest neighbours distribution [Fig.1].

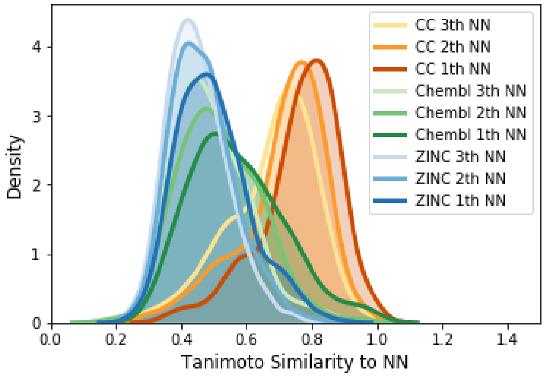


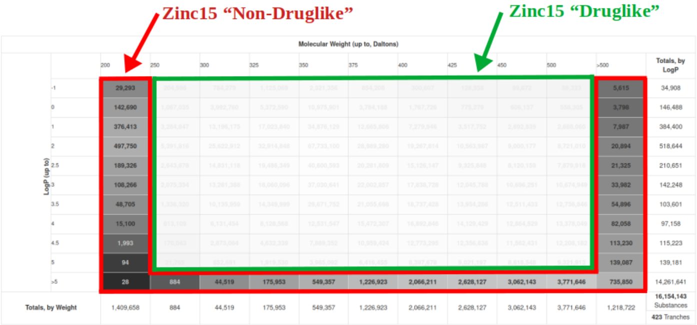
Figure 1: Morgan fingerprints Molecule density distributions as function of Tanimoto similarity for 3 top nearest CC neighbours (NN): highest 3 Tanimoto similarities. Zinc15, Chembl and CC (own comparison excluded) databases have been used.

Zinc15 database has resulted to differ most from CC. Closer clustering from CC fingerprints is due to the fact that assays in a given database are mostly undertaken in structural similar molecule groups (hence many similar molecules are available). Plotting Chembl-Chembl density distributions vs Tanimoto similarity results in similar performance as CC-CC.

12 hour 2-fold crossvalidated AutoSKlearn classificators for 40000 Zinc15 and 10000 CC random compounds have been run for considered descriptors. Resulting metrics have proven the classifiers to be random. Overall, no considerable huge subset from Zinc15 nor Chembl datasets seem to have significant structural diferences with CC. Since this datasets include a huge amount of organic and biological molecules, a small

subset will have to be explored to build the desired classifier.

Ultimately, Zinc15 ”Non-Druglike” compounds have seemed to offer better results, as they offer significant structural differences from CC compounds [Fig.2][Fig.3]. Signaturizer's aplicability domain value distribution for CC, chembl and Zinc ”non-druglike” databases has been computed [Fig.A.1].



Zinc15 “Non-Druglike”

Zinc15 “Druglike”

Molecular Weight (up to, Daltons)

LogP (up to)	200	250	300	325	350	375	400	425	450	500	>500	Totals, by LogP
-1	29,293	204,598	784,279	1,125,069	2,321,356	854,208	300,607	128,558	99,872	88,323	5,615	34,908
0	142,690	1,067,035	3,992,760	5,372,590	10,975,901	3,784,188	1,767,726	775,279	606,137	559,305	3,798	146,488
1	376,413	3,264,847	13,196,175	17,023,840	34,876,129	12,665,806	7,279,946	3,517,752	2,892,839	2,688,060	7,987	384,400
2	497,750	5,391,816	25,622,912	32,914,848	67,733,100	28,989,280	19,267,814	10,563,987	9,000,177	8,721,010	20,894	518,644
2.5	189,326	2,643,678	14,831,118	19,486,349	40,600,593	20,281,609	15,126,147	9,325,848	8,120,159	7,879,918	21,325	210,651
3	108,266	2,075,334	13,281,388	18,060,096	37,030,641	22,002,857	17,838,728	12,045,788	10,696,251	10,674,949	33,982	142,248
3.5	48,705	1,336,320	10,135,959	14,349,999	29,671,752	21,055,698	18,737,428	13,954,286	12,511,433	12,736,846	54,896	103,601
4	15,100	613,109	6,131,454	8,128,568	12,531,547	15,472,307	16,892,846	14,129,429	12,864,529	13,378,049	82,058	97,158
4.5	1,993	170,043	2,873,064	4,632,339	7,889,352	10,959,424	12,773,295	12,356,636	11,562,431	12,208,182	113,230	115,223
5	94	21,765	852,691	1,919,530	3,985,092	6,416,455	8,397,678	9,021,197	8,818,548	9,321,913	139,087	139,181
>5	28	884	44,519	175,953	549,357	1,226,923	2,066,211	2,628,127	3,062,143	3,771,646	735,850	14,261,641
Totals, by Weight	1,409,658	884	44,519	175,953	549,357	1,226,923	2,066,211	2,628,127	3,062,143	3,771,646	1,218,722	16,154,143 Substances 423 Tranches

Figure 2: ”Non-druglike” compounds extracted from Zinc15 database tranches. Three included subsets are low molecular weight compounds, high molecular weight and high reactivity.

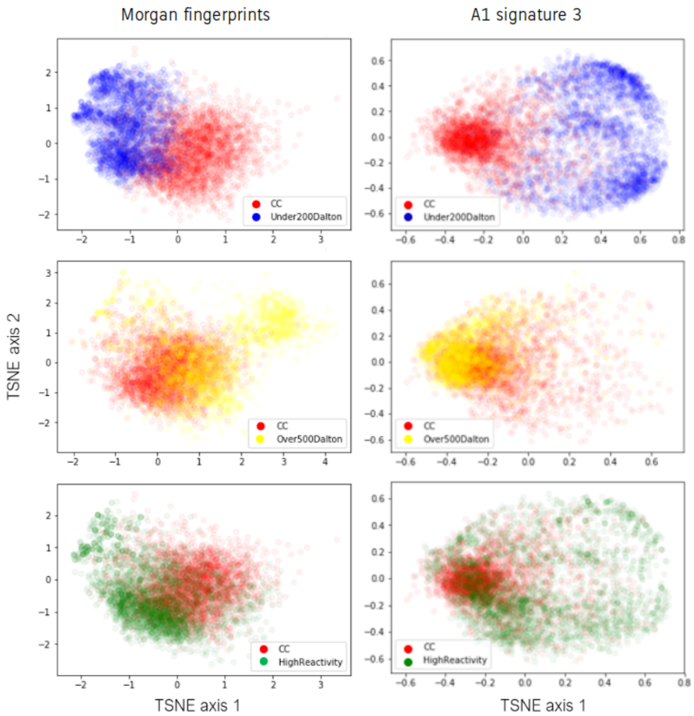


Figure 3: Morgan fingerprint (left column) and Signaturizer's A1 signature (right column) TSNE projections. Comparison between CC and Zinc15 ”non-druglike” three subsets.

In contrast with whole Chembl and Zinc15 compounds picked at random, Zinc15 ”Non-druglike” compounds seem to have direct differences from CC's and hence to be a suitable negative datset to build a classifier for structural separation.

3 Structural similarity classifier

Two major AutoSKlearn [6] tasks have been executed, one for each structural descriptor. In both cases, multiple classification models (over 150) have been trained during 12 hours with 32 cpus and 2-fold crossvalidation in a singular ensemble.

For both tasks, best performance was obtained using Gradient boosting as clasifier [Fig.4].

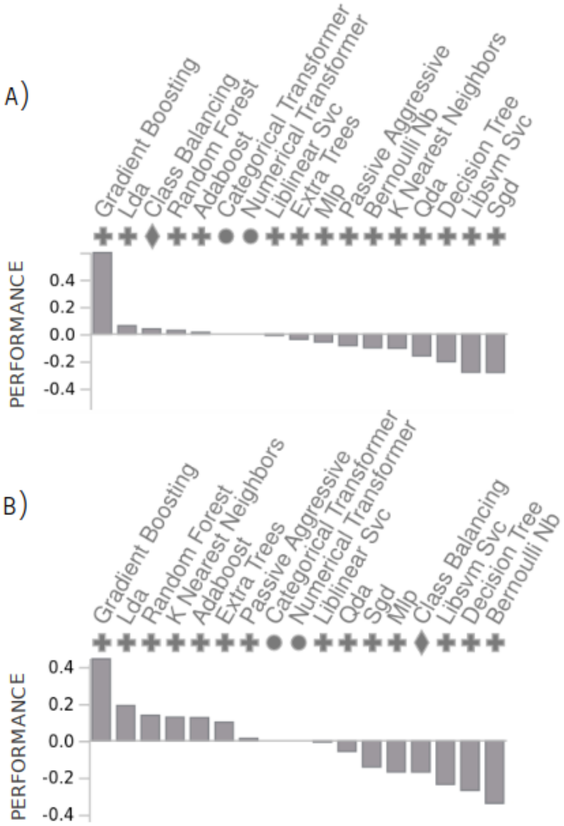


Figure 4: Average performance of AutoSKlearn used Preprocessors (●), Classifiers (✚) and Balancings (♦) for Morgan fingerprints (A) and A1 signatures (B). 10000 CC molecules as positives and 40000 Zinc ”Not-druglike” molecules as negatives were used.

Two new multiple classification trainings, just considering gradient boosting classifiers (best performance in previous run), were performed again in order to get the best hyperparameters for such classifier. Best performing models showed 0.959 accuracy for Morgan fingerprints and 0.912 accuracy for A1 signatures. Hyperparameters were obtained [Fig.5].

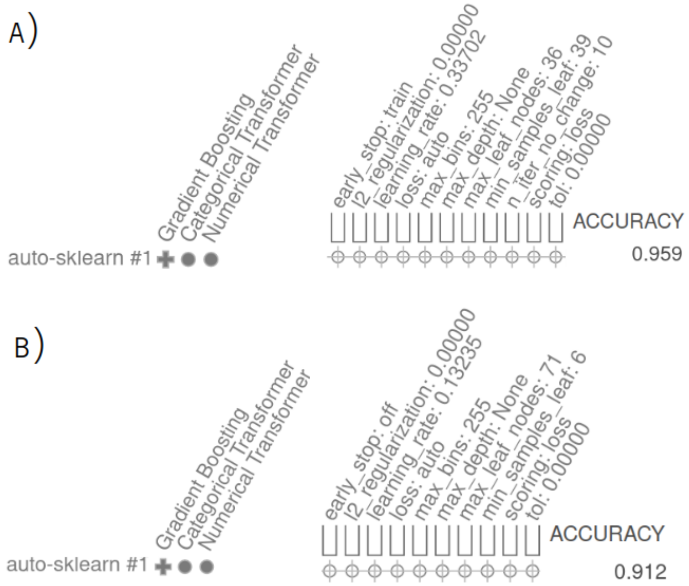


Figure 5: Final best performing AutoSKlearn models hyperparameters and accuracies for Morgan fingerprints (A) and A1 signatures 3 (B), given 10000 molecules CC as positives and 40000 Zinc ”Not-druglike” molecules as negatives.

Corresponding Receiver Operating Characteristic (ROC) curves for the best performing models have been obtained [Fig.8]. Morgan fingerprints offered best area under curve (AUC = 0.94), whilst A1 signatures performed worse (AUC = 0.87). For Precision-Recall curves, PR scores have been computed bor both Morgan fingerprints (PR score = 0.94) and A1 signatures (PR score = 0.88) [Fig.9].

Overall, the loss of classifying performance is clear in the transition form Morgan fingerprints to A1 signature 3. This must be due to the loss of information during Signaturizer’s prediction, but also from the data compression in its SNN training dataset.

4 Bioactivity similarity classifier for signaturizer’s B4 signatures

Since CC database has been taken from Pubchem database molecules with recorded activity, it is a direct question to ask what happens if we compare external databases bioactivity data with CC’s. For doing so, CC B4 space has been considered (target binding information) [Fig.6].

Directly building a classifier with AutoSKlearn out from this signatures results in random forest as best performing model. A second task for only random forest classifiers has been run in order to optimize hyperparameters [Fig.7].

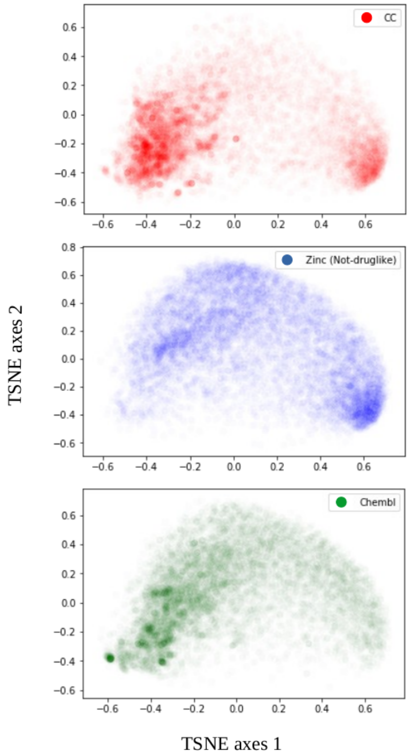


Figure 6: 128 dimension signaturizer’s B4 signature TSNE projection for CC, Zinc15 ”non-druglike” and Chembl libraries (3000 random compounds each).

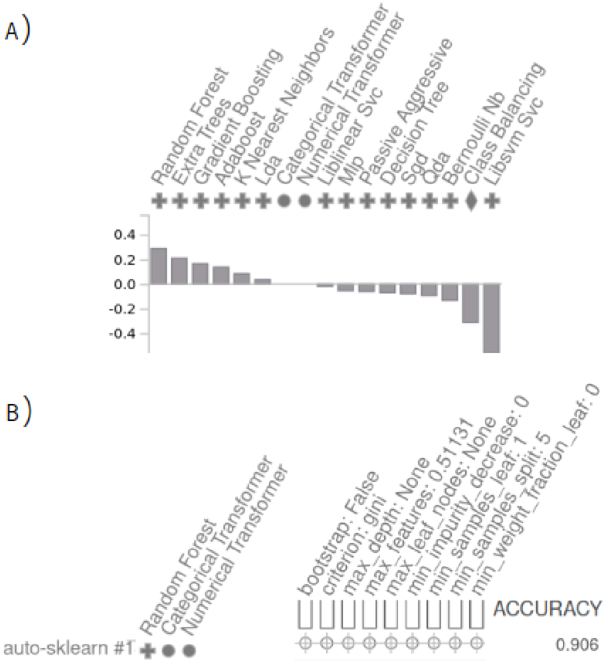


Figure 7: B4 signature average classifier performance (A) and final best performing model hyperparameters (B). 10000 CC moleculesas positives and 40000 Zinc ”Not-druglike” molecules as neg-atives were used.

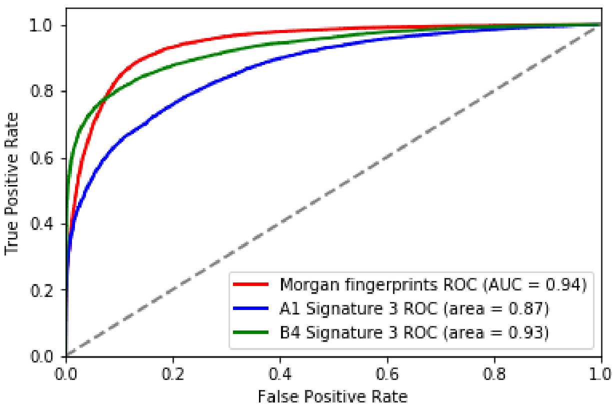


Figure 8: Best performing Gradient Boosting classifiers receiver operating characteristic (ROC) curves for Morgan fingerprints (AUC = 0.94), A1 signatures (AUC=0.87) and B4 signatures (AUC=0.93).

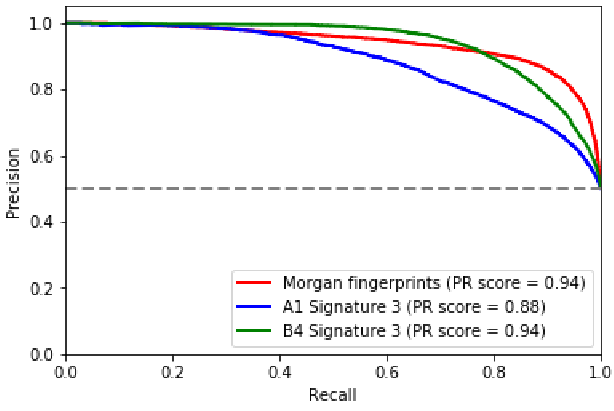


Figure 9: Best performing models precision-recall curves for Morgan fingerprints (pr score = 0.94), A1 signatures (pr score = 0.87) and B4 signatures (pr score = 0.94).

Using B4 signatures results in a slightly worse overall classifier than Morgan fingerprints, but performance is best if high precision is desired. A1 signatures decrease classifying performace from other two molecular descriptors.

5 CC entropy analysis

[7] An interesting study arises from the study of information entropy (H) of the CC. As a reminder, entropy is defined as

$$H(\vec{x}) = \sum_{i=1}^{n} P(\vec{x_i}) log_b\big(P(\vec{x_i})\big) \quad , \tag{1}$$

where $\vec{x} = (\vec{x_1}, ..., \vec{x_n})$ englobes all possible variable value combinations. Parameter b defines the unit of information (bits for $b = 2$). In the case of using n dimension Morgan fingerprints, $\vec{x_i}$ is one of the possible combinations of n zeros and ones.

By considering individual Morgan fingerprints variables statistically independent between them (i.e. $p(x_i^a|x_i^b) = p(x_i^a)$),

$$P(\vec{x_i}) \simeq \prod_{k=1}^{n} P(x_i^k) \quad . \tag{2}$$

The sum over all possible n variables $\vec{x_i}$ can be reexpressed as follows:

$$\sum_{i=1}^{n} = \sum_{x^1=0}^{1} \dots \sum_{x^n=0}^{1} \quad . \tag{3}$$

Hence, inserting 2 and 3 in 1 and expanding results in

$$\begin{aligned}
H(\vec{x}) = \sum_{x^1=0}^{1} \dots \sum_{x^n=0}^{1} \Bigg[\Bigg(\prod_{k=1}^{n} P(x^k)\Bigg) log_b\Bigg(\prod_{l=1}^{n} P(x^l)\Bigg)\Bigg] = \\
\sum_{x^1=0}^{1} \dots \sum_{x^n=0}^{1} \Bigg[\Bigg(\prod_{k=1}^{n} P(x^k)\Bigg) \sum_{l=1}^{n} log_b\Big(P(x^l)\Big)\Bigg] = \\
\sum_{x^1=0}^{1} \dots \sum_{x^n=0}^{1} \Bigg[\sum_{l=1}^{n} log_b\Big(P(x^l)\Big)\Bigg(\prod_{k=1}^{n} P(x^k)\Bigg)\Bigg] = \\
\sum_{l=1}^{n} \Bigg[\sum_{x^1=0}^{1} \dots \sum_{x^n=0}^{1} \Bigg[log_b\Big(P(x^l)\Big)\Bigg(\prod_{k=1}^{n} P(x^k)\Bigg)\Bigg]\Bigg] = \\
\sum_{l=1}^{n} \Bigg[\underbrace{\sum_{x^l=0}^{1} \Bigg[P(x^l) log_b\Big(P(x^l)\Big)\Bigg]}_{=H(x^l)} \underbrace{\sum_{x^1=0}^{1} \overset{x^i \neq x^l}{\dots} \sum_{x^n=0}^{1} \Bigg[\prod_{\substack{k=1 \\ k\neq l}}^{n} P(x^k)\Bigg]}_{=1}\Bigg] = \\
= \sum_{l=1}^{n} H(x^l) \quad .
\end{aligned} \tag{4}$$

That is, if considering Morgan fingerprint variables statistically independent between them, the total entropy for the multivariable Morgan fingerprint is the sum of the individual entropy of each variable.

Morgan fingerprints have been computed for both CC compounds and an equally sized set of Zinc ”non-druglike” compounds. For each set, overall fraction of ones have been considered as probability for each variable. Ultimately, entropy has been computed for each variable and all contributions to entropy have been summed up.

Using $b = 2$, the total entropy for CC library $n = 2048$ dimension morgan fingerprints ends up to be 270.48 bits. That is, in theory the 2048 dimension binary fingerprints could be compressed in 270.1 bits without loosing any information. Entropy for a CC size number of Zinc ”non-druglike” compounds has resulted to be 52.51 bits.

Resulting datasets entropy, CC database proves to have more structural richness than Zinc ”non-druglike” used for building the classifier.

Comparing this results with signaturizer’s signature 3, which is 128 dimension of continuous data (not binary!), makes it clear that data encriptation in binary vectors can not decrease 270.1 bits. Unfortunately, continous data from Signaturizer’s signatures could not enable to directly compare their entropy to discrete binary Morgan fingerprint’s.

6 Results

Overall, successful classifiers for structural CC similarities have been built to discriminate Zinc ”non-druglike” database similar compounds. Such discriminated compounds have shown slightly worse applicability domain in all CC spaces.

Best approach has been using Morgan fingerprints (ROC AUC =0.94, PR score = 0.94). Using compounds bioactivity signaturizer’s B4 signatures has shown worse performance (ROC AUC = 0.93, PR score = 0.94), but higher precision.

Use of signaturizer’s A1 structural signatures has shown a drastic decrease in performance (ROC AUC = 0.87, PR score = 0.88). Such decrease in performance must be due to the information loss in both Signature prediction and Morgan Fingerprint compression to signatures. A study of information entropy has shown CC library entropy to be 270.1 bits, and Zinc ”non-druglike” database 52.51 bits. Such entropies show huge structural richness from CC database, and a more restricted set of molecules for Zinc ”non-druglike”.

Entropy has not been able to be compared between Morgan fingerprints and signaturizer A1 signatures, due to continous variable encriptation from the second ones.

References

[1] Duran-Frigola M. et al. “Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker.” In: *Nat Biotechnol 38* (2020), pp. 1087–1096. URL: https://doi.org/10.1038/s41587-020-0502-7.

[2] Bertoni M. et al. “Bioactivity descriptors for uncharacterized chemical compounds.” In: *NB. Nat Commun 12* (2021), p. 3932. URL: https://doi.org/10.1038/s41467-021-24150-4.

[3] A.P. Bento et al. “‘The ChEMBL bioactivity database: an update’ Nucleic Acids Res. Database Issue”. In: 2014. DOI: 10.1093/nar/gkt103.

[4] A. Gaulton et al. “‘ChEMBL: A Large-scale Bioactivity Database For Chemical Biology and Drug Discovery’ Nucleic Acids Res. Database Issue, 40”. In: 2012. DOI: 10.1093/nar/gkr777.

[5] Sterling and Irwin. “ZINC15”. In: 2015. URL: http://pubs.acs.org/doi/abs/10.1021/acs.jcim.5b00559..

[6] Matthias Feurer et al. “Efficient and Robust Automated Machine Learning”. In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. Curran Associates, Inc., 2015, pp. 2962–2970. URL: https://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf.

[7] Thomas M. Cover and Joy A. Thomas. *”Elements of information theory”*. 2006.

A Appendix

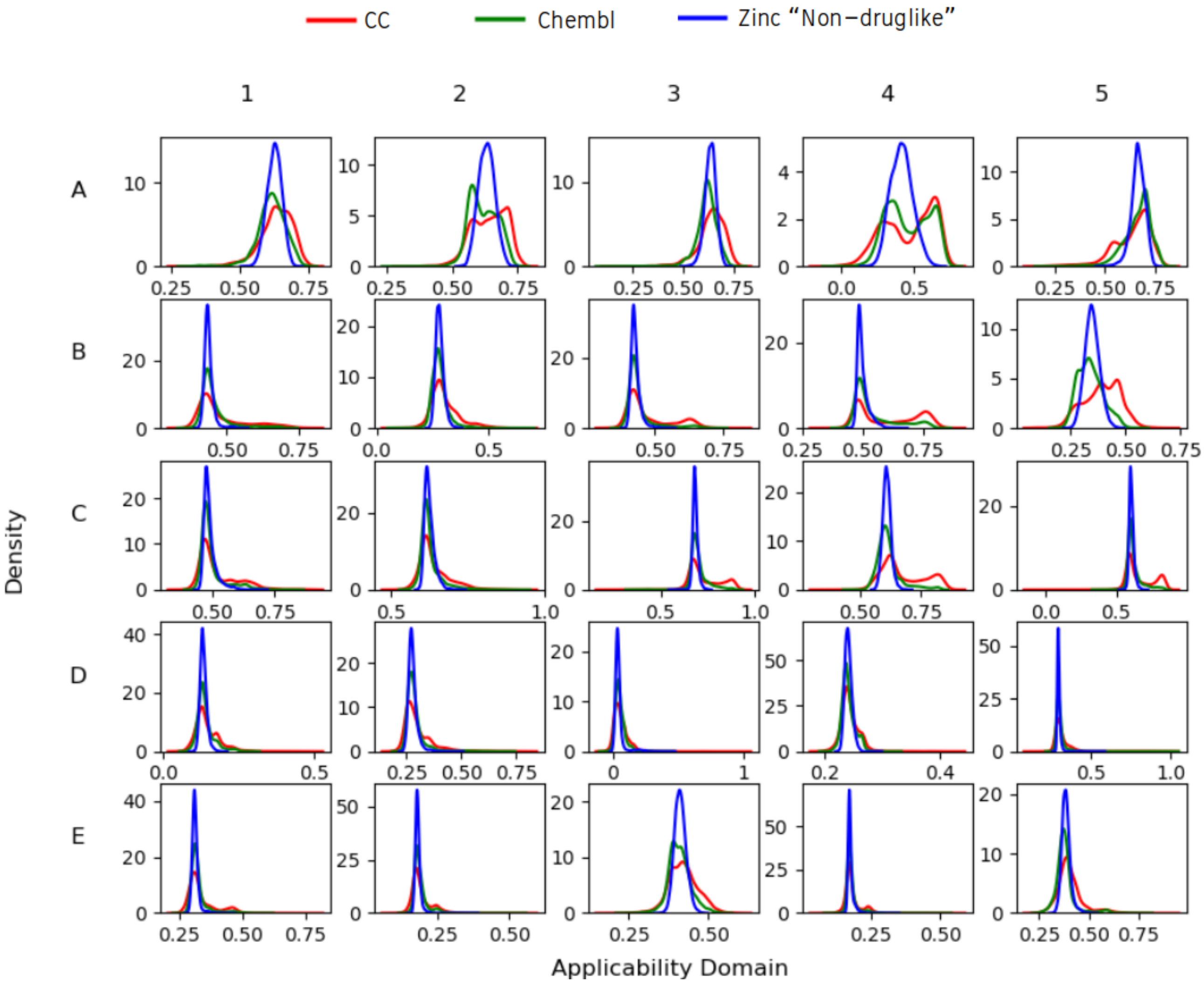


Figure A.1: Distributions for molecular density as function of Signaturizer’s aplicability domain for CC, Chembl and Zinc ”Non-druglike” used databases.