Are Virtual Screening Methods Smarter than KNNs?

Michael Brocidiacono
University of North Carolina at Chapel Hill
mixarcid@unc.edu

Konstantin I. Popov
University of North Carolina at Chapel Hill
kpopov@unc.edu

Alexander Tropsha
Univerity of North Carolina at Chapel Hill
alex_tropsha@unc.edu

Abstract

Structure-based virtual screening (SBVS) is a key workflow in computational drug discovery. SBVS models are assessed by measuring the enrichment of known active molecules over decoys in retrospective screens. However, the usual formula for enrichment cannot estimate model performance on very large libraries. Additionally, current screening benchmarks cannot easily be used with machine learning (ML) models due to data leakage. We propose an improved formula for calculating VS enrichment and introduce the BayesBind benchmarking set composed of protein targets that are structurally dissimilar to those in the BigBind training set. We assess current models on this benchmark and find that none perform appreciably better than a KNN baseline. We publicly release the BayesBind benchmark at https://github.com/molecularmodelinglab/bigbind.

1 Background

Structure-based virtual screening (SBVS) aims to identify purchasable compounds that bind to a protein target. Given the 3D structure of the target's binding site, a large library of compounds is scored according to their predicted ability to bind to a protein target, and the top-scoring molecules are selected for experimental validation. Two popular benchmarks for SBVS campaigns are DUD-E [1] and CASF16 [2], which mix known binders for several proteins with dissimilar "decoy" compounds. VS models are used to select the top fraction C (e.g. 1%) of each set, and the enrichment factor EF_C is computed by the fraction of actives in the selected set divided by the fraction of actives overall.

The $\text{EF}_{1\%}$ is the most commonly reported metric. In real-life virtual screens, however, a much smaller fraction of the VS libraries is selected (effectively, only a couple of compounds out of a library of millions). Computing the EF for such large screens would incur significant computational costs. Another downside of current benchmarks is that they are not associated with any training set. To rigorously evaluate ML SBVS methods, one should ensure that no binding site in the benchmark is too structurally similar to those in the training set [3–6].

2 Methods

The enrichment factor EF_C outlined above is an estimate of the ratio $\frac{P(A|S \geq S_C)}{P(A)}$, where A is the event that a molecule binds to the current target, S is the model's score for the molecule, and S_C is the cutoff score for the screen, chosen such that $P(S \geq S_C) = C$. Using Bayes' Theorem, it is easy to show that $\frac{P(A|S \geq S_C)}{P(A)} = \frac{P(S \geq S_C|A)}{P(S \geq S_C)}$. The latter ratio can easily be estimated by separately scoring a set of active molecules and a set of random compounds. The cutoff score S_C is determined only by the random compounds. The enrichment factor is estimated by the fraction of actives greater than S_C divided by the fraction of random compounds greater than S_C. We denote this value as the "expected" enrichment factor EEF_C to distinguish it from the traditional formula for EF_C, though we emphasize that both these formulas are estimates of the same probability ratio.

Preprint. Preliminary work.

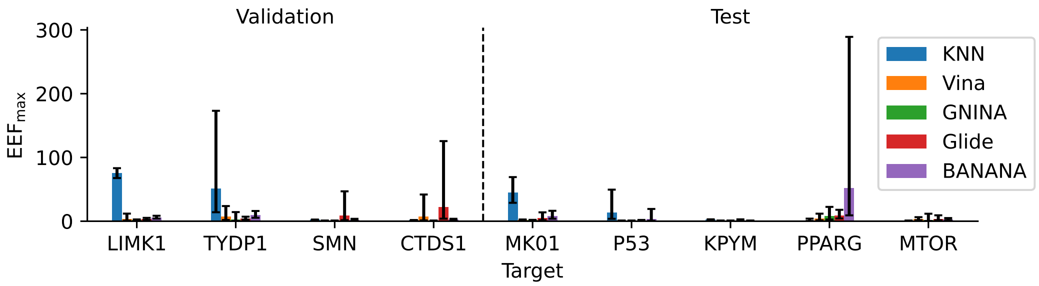

Figure 1: $\mathrm{EEF}_{\mathrm{max}}$ values for all the models on each target in the BayesBind benchmark.

Table 1: Summary results for all the models. Shown are the median $\mathrm{EEF}_{\mathrm{max}}$ (and 95% confidence interval), the $\mathrm{C}_{\mathrm{max}}$ for that median, and the number of targets for which $\mathrm{EEF}_{\mathrm{max}}$ is greater than 10.

Method	Validation			Test		
	Med. $\mathrm{EEF}_{\mathrm{max}}$	$\mathrm{C}_{\mathrm{max}}$	# > 10	Med. $\mathrm{EEF}_{\mathrm{max}}$	$\mathrm{C}_{\mathrm{max}}$	# > 10
KNN	51, CI = [14, 173]	0.07%	2	2.4, CI = [2.0, 3.0]	3.1%	2
Vina	7.0, CI = [1.9, 24]	0.51%	0	2.1, CI = [1.6, 2.7]	9.2%	0
GNINA	1.8, CI = [1.2, 2.8]	4.1%	0	1.6, CI = [1.2, 2.1]	12%	0
Glide	9.2, CI = [1.6, 47]	0.31%	1	3.7, CI = [1.5, 9.3]	0.8%	0
BANANA	6.0, CI = [4.1, 8.5]	1.7%	0	3.4, CI = [0.6, 19]	0.06%	1

The EEF has several advantages. First, it does not assume that all compounds in the random set are inactive. The usual EF formula, by contrast, requires that the fraction of *known* actives in the selected set be equal to the fraction of *total* actives. Additionally, the usual formula requires that the number of compounds in the selected fraction is relatively large in order to get an accurate estimate of the fraction of actives. The EEF formula, however, has the same runtime for all values of C up to $\frac{1}{N}$, where N is the size of the random set. This makes it orders of magnitude faster for small C.

We constructed the BayesBind benchmark for use with this improved metric. Nine protein binding sites were selected from the BigBind [7] validation and test sets according to structural diversity and the number of known active compounds. 10,000 diverse compounds from elsewhere in the validation and test sets were selected for each target's "random" set. Further details about the benchmark construction can be found in A.1.

All the analysis here uses an activity cutoff of 10 µM (pChEMBL value of 5 or more) to define the set of active compounds, though other cutoffs may also be used.

3 Results and Conclusions

We evaluated three popular SBVS methods (Vina [8], GNINA [9], and Glide [10]) on the benchmark, along with a simple neural network (BANANA) that was trained on the BigBind set [7]. We also evaluated a KNN baseline that uses a combination of the ligand and pocket similarity (details in A.2).

For each model and target in the benchmark, we compute the EEF_C for a range of cutoff fractions C. We additionally determined the $\mathrm{EEF}_{\mathrm{max}}$, the maximum enrichment achievable by a model on a particular target. The full $\mathrm{EEF}_{\mathrm{max}}$ results for all models on all targets are shown in figure 1, and median values are shown in table 1. Notably, no method does particularly well overall, and the KNN is the only method that achieves an $\mathrm{EEF}_{\mathrm{max}} > 10$ on more than 1 target.

The failure of popular SBVS methods on this benchmark stands in contrast to the reported success of these methods on DUD-E and in real-life perspective screens. This is likely because DUD-E is mostly composed of well-studied target classes such as GPCRs and kinases, whereas BayesBind uses a more diverse set. Similarly, successful prospective virtual screens are almost always on well-studied target classes [11]. However, we argue that the true potential of SBVS lies in novel targets. It is relatively easy to identify binders for kinases or GPCRs, but it is also not as useful (selectivity profiles are the more difficult challenge in this regime [12, 13]). In contrast, an SBVS method that can truly generalize to novel targets could significantly accelerate early-stage drug discovery. Thus it is important to measure the performance of our methods in this more useful regime.

The field of structure-based virtual screening has a long way to go before any method can be trusted with new targets. We hope that our comrades will use this benchmark to rise to the challenge of making new methods succeed in this area.

References

[1] Michael M. Mysinger, Michael Carchia, John. J. Irwin, and Brian K. Shoichet. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, July 2012. ISSN 0022-2623. doi: 10.1021/jm300687e. URL https://doi.org/10.1021/jm300687e. Publisher: American Chemical Society. 1

[2] Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal of Chemical Information and Modeling*, 59(2):895–913, February 2019. ISSN 1549-9596. doi: 10.1021/acs.jcim.8b00545. URL https://doi.org/10.1021/acs.jcim.8b00545. Publisher: American Chemical Society. 1

[3] Christian Kramer and Peter Gedeck. Leave-Cluster-Out Cross-Validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets. *Journal of Chemical Information and Modeling*, 50(11):1961–1969, November 2010. ISSN 1549-9596. doi: 10.1021/ci100264e. URL https://doi.org/10.1021/ci100264e. Publisher: American Chemical Society. 1

[4] Jincai Yang, Cheng Shen, and Niu Huang. Predicting or Pretending: Artificial Intelligence for Protein-Ligand Interactions Lack of Sufficiently Large and Unbiased Datasets. *Frontiers in Pharmacology*, 11:69, 2020. ISSN 1663-9812. doi: 10.3389/fphar.2020.00069. URL https://www.frontiersin.org/article/10.3389/fphar.2020.00069.

[5] Paul G. Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B. Iovanisci, Ian Snyder, and David R. Koes. Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design. *Journal of Chemical Information and Modeling*, 60(9):4200–4215, September 2020. ISSN 1549-9596. doi: 10.1021/acs.jcim.0c00411. URL https://doi.org/10.1021/acs.jcim.0c00411. Publisher: American Chemical Society.

[6] Mikhail Volkov, Joseph-André Turk, Nicolas Drizard, Nicolas Martin, Brice Hoffmann, Yann Gaston-Mathé, and Didier Rognan. On the Frustration to Predict Binding Affinities from Protein–Ligand Structures with Deep Neural Networks. *Journal of Medicinal Chemistry*, 65(11): 7946–7958, June 2022. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.2c00487. URL https://doi.org/10.1021/acs.jmedchem.2c00487. Publisher: American Chemical Society. 1

[7] Michael Brocidiacono, Paul Francoeur, Rishal Aggarwal, Konstantin Popov, David Koes, and Alexander Tropsha. BigBind: Learning from Nonstructural Data for Structure-Based Virtual Screening, November 2022. URL https://chemrxiv.org/engage/chemrxiv/article-details/637e3b5094ff6027063cc956. 2

[8] Oleg Trott and Arthur J. Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Journal of computational chemistry*, 31(2):455–461, January 2010. ISSN 0192-8651. doi: 10.1002/jcc.21334. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041641/. 2

[9] Andrew T. McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. GNINA 1.0: molecular docking with deep learning. *Journal of Cheminformatics*, 13(1):43, June 2021. ISSN 1758-2946. doi: 10.1186/s13321-021-00522-2. URL https://doi.org/10.1186/s13321-021-00522-2. 2

[10] Richard A. Friesner, Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K. Perry, David E. Shaw, Perry Francis, and Peter S. Shenkin. Glide: A New Approach for Rapid, Accurate Docking and Scoring. *Journal of Medicinal Chemistry*, 47(7):1739–1749, March 2004. ISSN 0022-2623. doi: 10.1021/jm0306430. URL https://doi.org/10.1021/jm0306430. Publisher: American Chemical Society. 2

[11] Hui Zhu, Yulin Zhang, Wei Li, and Niu Huang. A Comprehensive Survey of Prospective Structure-Based Virtual Screening for Early Drug Discovery in the Past Fifteen Years. *International Journal of Molecular Sciences*, 23(24):15961, December 2022. ISSN 1422-0067. doi: 10.3390/ijms232415961. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9781938/. 2

[12] Dahlia R. Weiss, Joel Karpiak, Xi-Ping Huang, Maria F. Sassano, Jiankun Lyu, Bryan L. Roth, and Brian K. Shoichet. Selectivity Challenges in Docking Screens for GPCR Targets and Antitargets. *Journal of Medicinal Chemistry*, 61(15):6830–6845, August 2018. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.8b00718. URL https://doi.org/10.1021/acs.jmedchem.8b00718. Publisher: American Chemical Society. 3

[13] Philip Cohen, Darren Cross, and Pasi A. Jänne. Kinase drug discovery 20 years after imatinib: progress and future directions. *Nature Reviews Drug Discovery*, 20(7):551–569, July 2021. ISSN 1474-1784. doi: 10.1038/s41573-021-00195-4. URL https://www.nature.com/articles/s41573-021-00195-4. Number: 7 Publisher: Nature Publishing Group. 3

[14] H. L. Morgan. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, 5 (2):107–113, May 1965. ISSN 0021-9576. doi: 10.1021/c160017a018. URL https://doi.org/10.1021/c160017a018. Publisher: American Chemical Society. 4

[15] Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, April 2005. ISSN 0305-1048. doi: 10.1093/nar/gki524. URL https://doi.org/10.1093/nar/gki524. 4

[16] Edwin B. Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158):209–212, June 1927. ISSN 0162-1459. doi: 10.1080/01621459.1927.10502953. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1927.10502953. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1927.10502953. 5

A Appendix

A.1 BayesBind Benchmark Construction

The BigBind dataset clusters the protein targets in the dataset according to the pocket similarity; each pair of proteins is in the same split (train, validation, and test) if they are in the same cluster. To create a diverse benchmarking set, we selected a single protein per cluster in the validation and test sets.

When creating the benchmarking set for each target, we clustered all active molecules according to a Tanimoto similarity cutoff of 0.4 on Morgan fingerprints (2048 bits with radius 3) [14]. For each cluster, we selected the active molecule with the median activity (pChEMBL value). This was done to create a diverse set of active molecules. Proteins were rejected from the benchmark if they had less than 150 active molecules after clustering with pChEMBL value less than 6 (1 μM); this was done because we wanted a large number of low-activity molecules, which represent the majority of actives in real-life screens. Unfortunately, we realized only recently that there were a large number of activity values with type “Potency” from ChEMBL; these come from noisy qHTS screens from PubChem, and their values cannot be trusted. After throwing out all such values, we kept the remaining pockets if there remained greater than 30 actives with pChEMBL value greater than 5.

After selecting the active molecules, we created a diverse set of random molecules by clustering all the molecules in the validation and test sets that were not known to bind to any protein in the same pocket cluster and the current target. We defined the random clusters using the same Tanimoto cutoff of 0.4. We selected a random compound from each cluster until we reached 10,000 molecules for each target.

A.2 KNN Baseline

To score a protein-ligand interaction using a KNN, we must first define a single similarity measure over the space of protein and ligand pairs. Fortunately, we already have the Tanimoto similarity over small molecules and several metrics of binding site similarity. Here, we use a pocket-level version of the TM-score [15]. To compute the optimal linear combination of these two features, we recognize that proteins with similar pockets will bind to similar ligands; that is, we expect that high ligand and high pocket similarity will co-occur more than would be expected if they were independent. To quantify this non-independence, we estimate the joint probability distribution $P(L, R)$ for all pairs of datapoints in BigBind, where L is the ligand similarity R is the receptor similarity. We then fit a linear model to predict the ratio $\log \frac{P(L,R)}{P(L)P(R)}$. We use the coefficients on L and R to define the global similarity measure $S = 0.18L + 3.57R$. To run the KNN model (with K=1) on a new protein-ligand pair, we simply find the molecule with the closest S in the BigBind training set and return S as the activity score. Returning the raw similarity score works because BigBind (without SNA) entirely contains molecules with relatively high activity.

When initially evaluating the benchmark, we discovered that the KNN baseline (below) did incredibly well on some targets. Upon further inspection, we realized that there were a couple clear homologs of those targets in the training set. The pocket similarity metric used to split BigBind had given them too low of a similarity score, indicating the need for further research into pocket similarity detection. For expediency, the targets where this was an issue were manually removed.

A.3 Uncertainty Quantification

To compute the uncertainty in our estimates of $\frac{P(S \geq S_C|A)}{P(S \geq S_C)}$, we first assume that the $P(S \geq S_C)$ is accurate due to the large sample size (10,000). By contrast, we use a binomial confidence interval for the estimate of $P(S \geq S_C|A)$ using the Wilson score [16]. Since this interval is divided by $P(S \geq S_C)$, it naturally becomes larger as C decreases.

A.4 Additional Benchmarking Results

The EEF$_C$ values for all C studied are shown in figure A.2

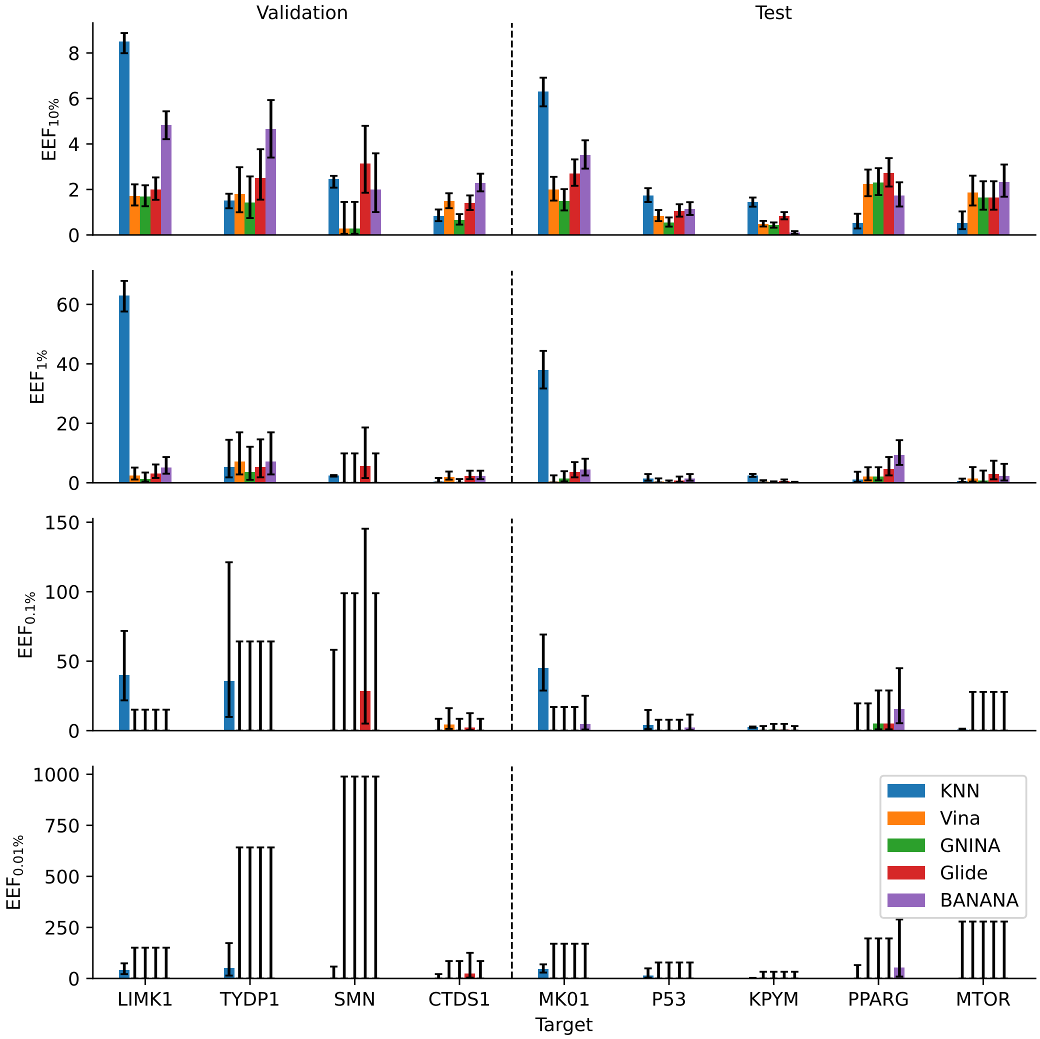

Figure A.2: $EEF_{10\%}$, $EEF_{1\%}$, $EEF_{0.1\%}$, and $EEF_{0.01\%}$ values for all the models on each target in the BayesBind benchmark.