MOLECULAR FINGERPRINTS REMAIN AN IMPORTANT REPRESENTATION OF SMALL MOLECULES FOR BIOCHEMISTRY

Anonymous authors
Paper under double-blind review

1 INTRODUCTION

Drug–drug interactions (DDIs), also known as polypharmacy, refer to the potential non-additive effects of multiple co-administered medications and are regularly encountered in clinical contexts. A considerable amount of effort has been made on tackling his problem and the much progress has been achieved thanks to cutting-edge machine learning (ML) methods, which take many forms: molecular similarity, (bio)chemical structure, knowledge graph, and even natural language processing (NLP) (Lin et al., 2023).

Moreover, as recent empirical studies demonstrate, molecular fingerprints appear to encode sufficient information such that simple architectures built on top of them, such as multilayer perceptron (MLP), can achieve competitive performances compared to methods that are much more computationally expensive. Our work hence supplements well a recent benchmark paper (Xia et al., 2023) focusing on FP for downstream tasks as that work considered only one drug at a time.

In this study, we join this latest trend by providing a more fine-grained analysis on the performance of a class of simple models over a very popular molecular fingerprinting system. Our main contributions are as follows:

- Provide more detailed ablation studies on the use of Morgan FP to predict DDIs.
- Evaluate the performance of the Morgan FP method under more challenging settings.
- Connect the problem with recent empirical studies regarding the merits of deep learning vs boosted trees vs foundation models.

2 RESULTS AND DISCUSSION

We used an augmented version of the standard benchmarking dataset, DrugBank 5.0 (Wishart et al., 2018), first described by Long et al. (2022). All models were implemented using PyTorch and PyTorch Geometric. For the transductive setting, we randomly partitioned the dataset using $8:1:1$ train/validation/test split. For the inductive settings we used the ratio of masked drugs $k = 0.2$.

Hyperparameters	Accuracy	Weighted F_1	AUROC	Total Time (s)
Morgan MLP				
$R = 2$, $D = 2048$ (default)	96.15	96.13	99.81	1,286
$R = 3$, $D = 2048$	95.97	95.95	99.79	1,620
$R = 4$, $D = 2048$	96.13	96.12	99.80	1,841
$R = 3$, $D = 4096$	96.05	96.04	99.80	1,008
$R = 4$, $D = 4096$	96.28	96.26	99.82	1,217
SSI-DDI	95.07	95.04	99.23	15,120

Table 1: Radius and dimension of embedding space; the default value in `rdkit` is $R = 2$, $D = 2048$; depth of MLP has been set to 4

# layers	Accuracy	Weighted F_1	AUROC	Total Time (s)
Morgan MLP				
1	91.04	90.92	96.53	1,099
4 (default)	96.15	96.13	99.81	1,286
6	95.85	95.84	99.81	1,452
8	95.96	95.95	99.86	2,008
16	96.57	96.54	99.86	3,900
SSI-DDI	95.07	95.04	99.23	15,120

Table 2: Number of MLP layers; for R and D we used the default value as in Table 1

Method/Setting	Accuracy	Weighted F_1	Total Time (s)
Morgan MLP I_1	**62.95**	**63.29**	**1,367**
SSI-DDI I_1	62.49	62.61	12,060
Morgan MLP I_2	**37.04**	24.81	**1,701**
SSI-DDI I_2	35.46	**35.94**	10,920

Table 3: Inductive settings: under I_1 exactly 1 drug in an pair in the test set was unseen, under I_2 both were unseen

In this work we investigated two sets of hyperparameters that could have an impact on the performance of Morgan FP models, namely the radius and number of bits, as well as the number of MLP layers. Table 1 showed that using larger r and number of bits made only a very small difference. Compared to the competitive baseline, however, all Morgan FP configurations were comfortably ahead. This is especially the case for total time elapsed until convergence (rightmost column) — the Morgan FP models spent only a small fraction of time compared to the baseline to reach superior performances. The last point was also illustrated by Table 2, except that when the Morgan FP model had only a single-layer MLP the performance was much worse.

The inductive settings, as expected, turned out to be much harder. Here the advantage of Morgan FP over SSI-DDI was less pronounced in terms of performance metrics, but the time-saving factor was still very sizable. However, we needed to point out that the figures were much smaller than in the relevant literature (Nyamabo et al., 2021; 2022; Zhang et al., 2025), which we believe this was not due to overfitting since when we added dropouts to our MLP the results were comparable, but rather the nuanced difference between multi-class and multi-label classification *and* the way other researchers generated their negative samples for addressing dataset imbalance: since they treated *any* unseen combinations as negative rather using real world evidence. One way to mitigate the effect could therefore be to exclude interaction types and drugs below a certain threshold number (e.g. 30).

Future work should also focus on further dissecting the advantages and disadvantage of FP models with respect to graph-based methods. One important area would be to repeat the same experiments using other kinds of molecular FPs as well as combinations thereof. Interestingly, there has been a recent work (Zhang et al., 2023) on using a combination of an ensemble of 4 different kinds of fingerprints (*excluding* Morgan FP). Given that some FP methods, such as the PubChem fingerprint, yield features that are biochemically meaningful (cf. Adamczyk and Ludynia (2024)), it could be helpful to treat them as a tabular dataset and use tabular learning method for downstream tasks (Xu et al., 2021; Grinsztajn et al., 2022; McElfresh et al., 2023; Holzmüller et al., 2024; Hollmann et al., 2025).

Finally, yet another promising line of work would be to integrate FP features with knowledge graph-based approaches where biochemical pathways are explicitly modelled (cf. Gonzalez-Cavazos et al. (2023)), because unlike molecular graphs these graphs encode physiologically relevant information that can help elucidate *mechanisms* of DDI.

MEANINGFULNESS STATEMENT

For small molecules, it may be difficult to find a singular representation that is good for any kind of downstream tasks. Our suggestion is that which representation to choose depends heavily on the nature of a task: for DDIs, which are inherently biochemical in nature, expert-curated substructure features may be better due to the fact that it is not so much molecular similarity but rather participation in biochemical pathways that really matters.

REFERENCES

Adamczyk, J. and Ludynia, P. (2024). Scikit-fingerprints: Easy and efficient computation of molecular fingerprints in Python. *SoftwareX*, 28:101944.

Gonzalez-Cavazos, A. C., Tanska, A., Mayers, M., Carvalho-Silva, D., Sridharan, B., Rewers, P. A., Sankarlal, U., Jagannathan, L., and Su, A. I. (2023). DrugMechDB: A Curated Database of Drug Mechanisms. *Scientific Data*, 10(1):632.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326. Publisher: Nature Publishing Group.

Holzmüller, D., Grinsztajn, L., and Steinwart, I. (2024). Better by default: Strong pre-tuned MLPs and boosted trees on tabular data.

Lin, X., Dai, L., Zhou, Y., Yu, Z.-G., Zhang, W., Shi, J.-Y., Cao, D.-S., Zeng, L., Chen, H., Song, B., Yu, P. S., and Zeng, X. (2023). Comprehensive evaluation of deep and graph learning on drug–drug interactions prediction. *Briefings in Bioinformatics*, 24(4):bbad235.

Long, Y., Pan, H., Zhang, C., Hy, T.-S., Kondor, R., and Rzhetsky, A. (2022). Molecular fingerprints are a simple yet effective solution to the drug–drug interaction problem.

McElfresh, D. C., Khandagale, S., Valverde, J., C, V. P., Ramakrishnan, G., Goldblum, M., and White, C. (2023). When Do Neural Nets Outperform Boosted Trees on Tabular Data?

Nyamabo, A. K., Yu, H., Liu, Z., and Shi, J.-Y. (2022). Drug–drug interaction prediction with learnable size-adaptive molecular substructures. *Briefings in Bioinformatics*, 23(1):bbab441.

Nyamabo, A. K., Yu, H., and Shi, J.-Y. (2021). SSI–DDI: substructure–substructure interactions for drug–drug interaction prediction. *Briefings in Bioinformatics*, 22(6):bbab133.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., and Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082.

Xia, J., Zhang, L., Zhu, X., Liu, Y., Gao, Z., Hu, B., Tan, C., Zheng, J., Li, S., and Li, S. Z. (2023). Understanding the Limitations of Deep Models for Molecular property prediction: Insights and Solutions.

Xu, H., Kinfu, K. A., LeVine, W., Panda, S., Dey, J., Ainsworth, M., Peng, Y.-C., Kusmanov, M., Engert, F., White, C. M., Vogelstein, J. T., and Priebe, C. E. (2021). When are Deep Networks really better than Decision Forests at small sample sizes, and how? arXiv:2108.13637 [cs, q-bio, stat].

Zhang, M., Gao, H., Liao, X., Ning, B., Gu, H., and Yu, B. (2023). DBGRU-SE: predicting drug–drug interactions based on double BiGRU and squeeze-and-excitation attention mechanism. *Briefings in Bioinformatics*, 24(4):bbad184.

Zhang, Z., Liu, F., Shang, X., Chen, S., Zuo, F., Wu, Y., and Long, D. (2025). ComNet: A Multiview Deep Learning Model for Predicting Drug Combination Side Effects. *Journal of Chemical Information and Modeling*, 65(2):626–639. Publisher: American Chemical Society.

A METHOD

A.1 DATA

First, prescription records were extracted from electronic health records (EHRs) from a very large health insurance claims data based in the United States. From these, the identity of drug, start date and the end date could be specified. Then, whenever a pair of two drugs overlapped in their period of validity (i.e. between start and end) it was deemed to be "safe" — without adverse combination effects and treated as an negative example. This is different from all other existing methods in the literature using oversampling methods (e.g. SMOTE-ENN used by Zhang et al. (2023)) designed to address the problem of data imbalance (Lin et al., 2023) since they all make the assumption that all drug–drug–relation triples are negative samples.

A.2 LEARNING TASK

We formulate the DDI problem as a multi-class classification task. The dataset is presented by $\left\{d_i^{(1)}, d_i^{(2)}, r_j\right\}_{i=1}^{N}$, where $r_j \in \mathcal{C} = [1, K] \bigcap \mathbb{N}$ is the type of DDI. The task is then to learn a parameterized function

$$\Phi_\theta \colon D \times D \mapsto \mathcal{C}, \tag{1}$$

where D is the set of drugs. An alternative setup is multi-label classification, where the function Φ_θ is given by

$$\Phi_\theta \colon D \times D \times \mathcal{C} \mapsto \{0, 1\}. \tag{2}$$

The difference between multi-class and multi-label classification is that the under the latter some pairs of drugs may have more than one type of interaction.

A.3 TRANSDUCTIVE VS INDUCTIVE LEARNING

Transductive refers to the usual setting where all the drugs encountered at test time have already been seen during training and validation, whereas *inductive* means that at test time there are drugs that have not been observed during training and validation, hence requiring the algorithm to be able to generalize. Formally, let D be the set of all drugs, and D_{test} the set of drugs held out for testing with $|D_{\text{test}}| \,/\, |D| = k$. The inductive 1 ($I_1$) setting is where the test set containing drug pairs with 1 seen and 1 unseen drug, whereas the inductive 2 (I_2) setting considers only those pairs with 2 unseen drugs.

A.4 MORGAN FINGERPRINT

The FP we used in this study was the Morgan fingerprint as implemented by RDKit (**?**). The fingerprint was generated by the eponymous algorithm (**?**), which collects at each atom the substructure (subgraph) induced by its r-hop neighbours, where r is the radius. In this way, it is closely related to the Weisfeiler–Lehman algorithm, but also takes into consideration chemically relevant features such as electric charge and membership in rings structure(s). See Figure 1 for an example. While the aforementioned step produced the counts (i.e. a histogram) of the subgraphs, in practice they are mapped to a 1-dimensional vector of binary bits by a hash function. Both the radius and the number of bits were hyperparameters we investigated.

A.5 MODEL TRAINING

We chose the multi-class cross-entropy over each minibatch (size $B = 256$) as our objective and used the rectified Adam (RAdam) optimizer to minimize it, with an early stopping $\Delta = 5 \times 10^{-3}$ and a tolerance of 4 epochs.

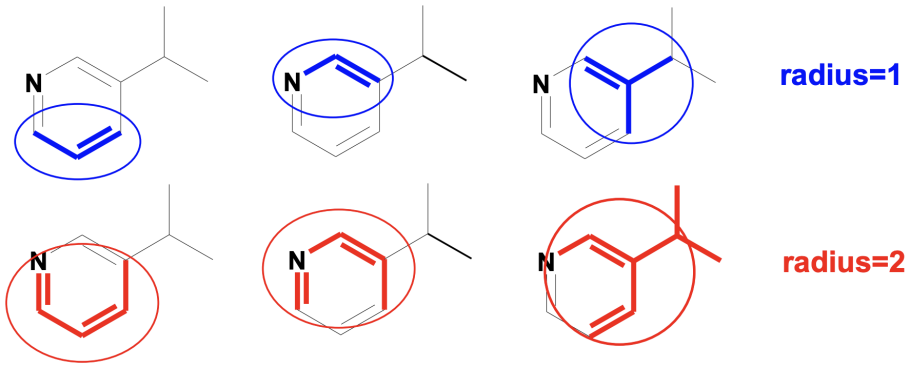

Figure 1: The Morgan algorithm of molecular fingerprinting. From **?**.