

Structure-Based Drug–Drug Interaction Prediction from Morgan Fingerprints: A Degree-Controlled and Cold-Start Evaluation of a Random Forest Baseline

Shrinivas Wable

`shrinivas.wable_24uds@sanjivani.edu.in`

Sanjivani University <https://orcid.org/0009-0007-6283-6822>

Sayali chimbalkar

Sanjivani University <https://orcid.org/0009-0001-8557-4040>

Research Article

Keywords: drug–drug interaction, molecular fingerprints, negative sampling, cold-start evaluation, Random Forest, calibration, cheminformatics, pharmacovigilance

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11400981/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Structure-Based Drug–Drug Interaction Prediction from Morgan Fingerprints: A Degree-Controlled and Cold-Start Evaluation of a Random Forest Baseline

Shrinivas Baliram Wable¹ and Sayali Chimbalkar²

Department of Artificial Intelligence and Data Science (AIDS), Sanjivani University, Kopergaon, Maharashtra, India

¹ ORCID: 0009-0007-6283-6822 | ² ORCID: 0009-0001-8557-4040

E-mail: shrinivaswable09@gmail.com; chimbalkarsayali@gmail.com

Abstract—Drug–drug interaction (DDI) checkers can only look up combinations that have already been recorded, so they say nothing about pairs that were never studied together, such as those involving newly approved drugs. We ask how much of the apparent performance of a structure-only DDI model survives a careful evaluation. Each drug is represented by Morgan (ECFP) fingerprints and ten physicochemical descriptors computed from PubChem SMILES (1,688 of 1,701 drugs in a DrugBank-derived list resolved), and a drug pair by order-invariant combinations of the two drugs’ features. We first show that a common negative-sampling protocol, which screens random negatives only against the sampled positives, labels $20.9 \pm 0.8\%$ of its “non-interacting” pairs with recorded interactions. We then build two balanced benchmarks of 17,522 pairs whose negatives exclude every recorded pair: one with random negatives and one with degree-matched negatives. On the random-negative benchmark a control that uses only drug identity, with no chemistry, reaches ROC-AUC 0.758, comparable to Random Forests on fingerprints (0.753–0.768). On the degree-matched benchmark the control falls to or below chance, the original concatenated-fingerprint model drops to AUC 0.572, and the order-invariant feature set with descriptors reaches AUC 0.678 (accuracy 62.1%) on random pairs and 0.615 (accuracy 57.6%) when both drugs are unseen in training. Calibration, selective prediction and a ten-pair case panel show that the model misses many recorded interactions (for example Sildenafil + Nitroglycerin) and is best used with an explicit “uncertain” output, which a Streamlit prototype provides. This work is a research prototype and is not intended for clinical use.

Keywords—drug–drug interaction, molecular fingerprints, negative sampling, cold-start evaluation, Random Forest, calibration, cheminformatics, pharmacovigilance.

I. INTRODUCTION

Checking every possible pair of drugs for interactions by hand does not scale, and the problem is worst for newly approved drugs, which by definition have little or no recorded interaction history. Existing interaction checkers are essentially lookup tools: they report combinations that have already been documented in a fixed database, but they say nothing about a pair that has never been recorded, because they contain no model of why interactions occur.

Machine learning offers a complementary route. If a model can associate structural features of two molecules with a recorded interaction, it may generalise to pairs that have never been recorded. Reported performance, however, depends heavily on how the benchmark is built. Three choices matter in particular. First, how non-interacting pairs are generated: a pair that is absent from an interaction database is usually unstudied rather than known to be safe. Second, whether a model can exploit how often each drug appears among interacting pairs instead of anything about its chemistry. Third, whether the test drugs also occur in training, which does not match the situation of a newly approved drug.

Our preliminary experiments (the v1 protocol) reported 65.3% accuracy for a Random Forest on concatenated 256-bit Morgan fingerprints. Revisiting that protocol showed that it suffered from all three weaknesses. This paper therefore asks a narrower and more careful question: how much of the apparent signal in structure-based DDI prediction survives once negatives are

screened against the full interaction list, drug-frequency shortcuts are removed, and test drugs are held out of training?

The contributions are: (i) an audit showing that the v1 negative-sampling protocol labels about one in five of its non-interacting pairs incorrectly (20.9% are recorded interactions); (ii) two balanced benchmarks of 17,522 pairs, one with random and one with degree-matched negatives, together with a drug-identity-only control that uses no chemistry; (iii) evaluation under three regimes (random pairs, one unseen drug, both drugs unseen), repeated over three random splits; (iv) an order-invariant pair representation with physicochemical descriptors, compared with the original concatenated fingerprints, with gradient boosting and logistic regression, and in a feature ablation; and (v) calibration, a three-tier output with an explicit “uncertain” band, selective prediction and a ten-pair case panel, all exposed through a research prototype.

The paper is organised as follows: Section II reviews related work, Section III describes data, features, models and evaluation, Section IV reports results, Section V comments on statistical uncertainty, Section VI discusses implications and limitations, and Section VII concludes.

II. RELATED WORK

Graph neural networks (GNNs) have become the dominant approach for structure-based DDI prediction in recent literature. Gao et al. [1] used automated architecture search rather than hand-designing a GNN, and found the resulting network outperformed manually designed alternatives. Liu et al. [2] combined several different types of drug feature networks using

an attention mechanism and showed that combining multiple feature sources improves accuracy over any single source alone. Chen et al. [3] applied a graph convolutional network directly to molecular structure and demonstrated that learning straight from molecular graphs, rather than from hand-engineered features, works well for this task. Feng and Zhang [4] added attention mechanisms on top of a GNN operating on molecular graphs, which improved both accuracy and interpretability, since attention weights give some indication of which substructures drove a prediction. Lee et al. [5] represents an earlier, pre-GNN approach that analysed interactions using a network of biological information rather than molecular structure directly.

Before graph-based models, feature- and similarity-based methods were widely used. Vilar et al. [6] predicted new interactions by comparing the molecular similarity of a candidate pair with that of known interacting pairs. Ryu et al. [7] trained a deep neural network on structural similarity profiles and reported that it predicted a large number of interaction types. Zitnik et al. [8] modelled polypharmacy side effects with graph convolutional networks over drug-protein networks, which links the interaction problem to the side-effect problem. These studies show that structure alone is informative, which motivates the fingerprint baseline used here.

Less attention has been paid to how DDI predictors are evaluated. Dewulf et al. [9] separate the task of predicting interactions between drugs that already have recorded interactions from cold-start tasks in which one or both drugs are new, and argue that validation schemes must match the intended use. We adopt the same distinction (regimes S1–S3, Section III-F). A second issue is that interaction databases record only positive findings, so a missing pair is unlabelled rather than negative; when negatives are drawn at random, differences in how often drugs occur among positives and negatives can act as a shortcut, which we measure with an identity-only control. Finally, because a clinical user must decide how far to trust a score, we examine calibration [10], [11] and abstention.

This paper does not build a GNN. It establishes a simpler, well-understood baseline with fixed-length fingerprints and tree ensembles, evaluated under controls that a future GNN study should also pass, so that any gain from graph models can be measured against it.

III. METHODOLOGY

Fig. 1 gives an overview of the pipeline; the subsections below describe each step.

A. Data Source and Structures

Interacting pairs come from a DrugBank-derived interaction list [12] (Kaggle dataset “Drug-Drug Interactions” by mghobashy, file `db_drug_interactions.csv`; licence to be confirmed). The file has 191,541 rows over 1,701 distinct drug names. After mapping each row to an unordered pair and removing self-pairs and duplicates, 191,135 unique interacting pairs remain, which is 13.2% of the 1,445,850 possible pairs among these drugs. SMILES strings were looked up by name through PubChem [13] (pubchempy; first match, up to three attempts per name)

and succeeded for 1,688 of 1,701 drugs (99.2%); all 1,688 parsed in RDKit [14]. The 13 drugs without a structure, and pairs involving them, were excluded.

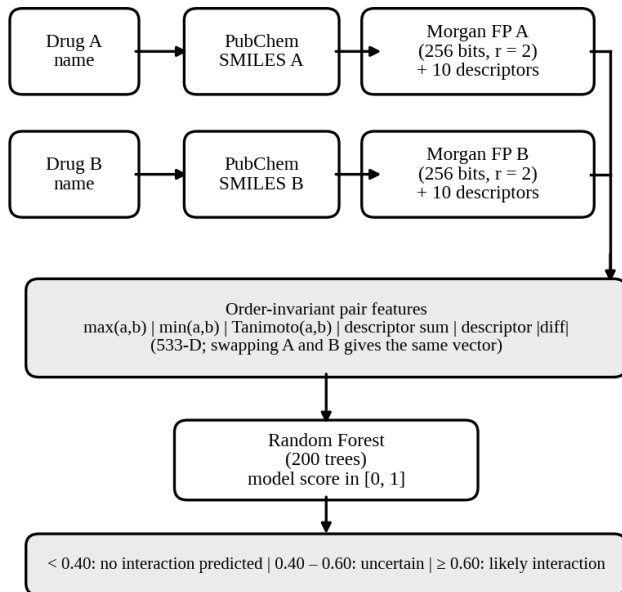


Fig. 1. Prediction pipeline: drug names are resolved to SMILES through PubChem, converted to Morgan fingerprints and descriptors, combined into order-invariant pair features, and scored by a Random Forest. Scores map to three tiers.

B. Audit of the v1 Negative-Sampling Protocol

The v1 protocol sampled 3,000 positive pairs and then drew an equal number of random pairs among the drugs in that sample as negatives, excluding only the sampled positives. We repeated this procedure with five random seeds and checked each resulting “negative” against the full list of 191,135 recorded pairs.

C. Benchmarks

From pairs with both structures available, 15,000 interacting pairs were drawn at random (seed 42). Two balanced benchmarks were built from them, and both exclude every pair in the full list from their negatives. In benchmark B (degree-matched), each positive (a, b) is paired with a negative (a, b’), where b’ is the second drug of another positive, re-drawn until the pair is not a recorded interaction; every drug therefore appears in negatives about as often as in positives. Positives that could not be re-matched (typically drugs that interact with almost every other drug) were dropped, leaving 8,761 of 15,000 (58%). In benchmark A (random), the same 8,761 positives are paired with negatives drawn uniformly at random from pairs of drugs that occur in the positives, again excluding all recorded pairs. A and B thus differ only in how negatives were chosen. Each contains 17,522 pairs (50% positive), with pair orientation randomised.

As a check on the matching, the Spearman correlation between a drug’s number of positive and negative appearances was -0.40 in A and 0.94 in B, and the share of drugs with exactly equal counts rose from 3.9% to 24.4%.

D. Features

For each drug, Morgan (ECFP) fingerprints [15] with radius 2 were computed with RDKit at 256 bits (main experiments) and 2,048 bits (ablation), together with 167 MACCS keys (ablation) and ten physicochemical descriptors: molecular weight, logP, topological polar surface area, hydrogen-bond donors and acceptors, rotatable bonds, rings, aromatic rings, heavy atoms and fraction of sp^3 carbons. Descriptors were z-scored using the mean and standard deviation over all 1,688 drugs (no labels are used). For a pair with bit vectors a and b , the following feature sets were compared.

concat256 is $[a ; b] \in \{0,1\}^{512}$, the $v1$ representation, which depends on pair order. sym256 is $[\max(a,b) ; \min(a,b) ; T(a,b)]$ (513-D), where T is the Tanimoto similarity $|a \wedge b| / |a \vee b|$. sym256+phys adds the sum and absolute difference of the two descriptor vectors (533-D) and is the “enhanced” set. sym2048 is the sym256 construction on 2,048 bits (4,097-D), and sym_all combines it with the same construction on MACCS keys and the descriptor terms (4,452-D). All symmetric sets give the same vector for (A, B) and (B, A) .

E. Models and the Identity-Only Control

A Random Forest (RF) [16] with 200 trees (scikit-learn [17], `random_state = 42`, other settings default), a histogram gradient-boosting classifier (HGB, defaults) and a logistic regression (LR; standardised inputs, L2 penalty, $C = 0.1$) were trained on pair features. The identity-only control uses no chemistry: for each drug it computes the Laplace-smoothed fraction of its training pairs that are positive, $(\text{positives} + 1) / (\text{pairs} + 2)$, uses the smaller and larger of the two drugs’ rates (0.5 for an unseen drug) as inputs, and fits a logistic regression. If a fingerprint model does no better than this control, it has learned how often drugs interact rather than why.

F. Evaluation Regimes and Metrics

Three regimes were used on each benchmark. S1 (random pairs): a stratified random 80/20 split of pairs (14,017 train, 3,505 test); test drugs may occur in training. For the cold-start regimes, 20% of drugs were randomly held out and models were trained only on pairs with neither drug held out. S2 (one unseen drug): test pairs with exactly one held-out drug. S3 (both unseen): test pairs with both drugs held out. Each regime was repeated over three random splits (seeds 42–44) and means are reported. The decision threshold is 0.5. Accuracy, ROC-AUC and the Matthews correlation coefficient (MCC) [18] are reported in the text; balanced accuracy, precision, recall, F1, average precision and Brier score were also computed and are in the supplementary results file.

G. Calibration and Selective Prediction

To assess calibration, RF scores were compared with scores from isotonic regression [10] (3-fold calibration on training

data) using the expected calibration error (ECE, 10 bins) [11] and the Brier score, on benchmark B for S1 and S3 (first repeat). For selective prediction, the model answers only when $|p - 0.5| \geq \delta$; we report coverage, accuracy on answered pairs and the false-negative rate (FNR) among answered interacting pairs, using calibrated scores.

H. Final Model, Case Panel and Prototype

The deployed model is the enhanced RF trained on all 17,522 pairs of benchmark B. Its raw score is mapped to three tiers: below 0.40 “no interaction predicted”, 0.40–0.60 “uncertain”, and 0.60 or above “likely interaction”. A hand-picked panel of ten clinically familiar combinations was scored (none of the ten pairs was among the training pairs; the drugs themselves were). A Streamlit application resolves drug names through a local table of the dataset’s drugs, falls back to live PubChem lookup, and shows the score, tier, the SMILES used with their source, and the Morgan substructures with highest global RF importance present in each drug. The application gives the same result for (A, B) and (B, A) , labels its output a model score rather than a probability, and carries a persistent warning against clinical use.

IV. RESULTS

A. Negative-Sampling Audit

Under the $v1$ protocol, 20.9% (SD 0.8 percentage points over five seeds) of the “non-interacting” pairs were in fact recorded interactions in the full list. This exceeds the overall density of recorded pairs (13.2%), which is plausible because the sampled drugs are biased towards highly connected drugs. A fifth of the negative labels were therefore wrong, in a way that depends on the very drug-frequency effects discussed next.

B. Benchmarks and Main Comparison

TABLE I. DATA AND BENCHMARK STATISTICS.

Quantity	Value
Rows in source list	191,541
Unique unordered interacting pairs	191,135
Distinct drugs	1,701
Drugs with a parsable structure	1,688 (99.2%)
Positive pairs sampled	15,000
Positives kept after degree matching	8,761 (58%)
Pairs per benchmark (A and B)	17,522 (50% positive)
S1 train / test pairs	14,017 / 3,505
Drugs held out for S2 and S3	20% of drugs

Tables II and III give ROC-AUC, and accuracy with MCC, for all five models in all three regimes on both benchmarks (means over three repeats); Fig. 2 plots AUC with repeat-to-repeat standard deviations.

TABLE II. ROC-AUC BY MODEL, BENCHMARK AND REGIME (MEAN OF THREE REPEATS).

	Benchmark A (random negatives)			Benchmark B (degree-matched negatives)		
Model	S1	S2	S3	S1	S2	S3
Identity-only (control)	0.758	0.682	0.500	0.366	0.407	0.500
RF, paper features	0.753	0.710	0.639	0.572	0.567	0.589
RF, enhanced	0.768	0.729	0.663	0.678	0.653	0.615
HGB, enhanced	0.734	0.700	0.636	0.627	0.611	0.586
LR, enhanced	0.659	0.642	0.567	0.560	0.552	0.550

TABLE III. ACCURACY / MCC BY MODEL, BENCHMARK AND REGIME (MEAN OF THREE REPEATS).

	Benchmark A (random negatives)			Benchmark B (degree-matched negatives)		
Model	S1	S2	S3	S1	S2	S3
Identity-only (control)	0.697 / 0.394	0.631 / 0.265	0.511 / 0.000	0.401 / -0.199	0.427 / -0.145	0.518 / 0.000
RF, paper features	0.688 / 0.378	0.655 / 0.310	0.600 / 0.201	0.551 / 0.101	0.541 / 0.084	0.559 / 0.116
RF, enhanced	0.700 / 0.401	0.669 / 0.339	0.619 / 0.239	0.621 / 0.242	0.605 / 0.210	0.576 / 0.150
HGB, enhanced	0.668 / 0.335	0.647 / 0.293	0.599 / 0.198	0.590 / 0.181	0.577 / 0.157	0.550 / 0.099
LR, enhanced	0.614 / 0.228	0.598 / 0.197	0.553 / 0.106	0.545 / 0.090	0.534 / 0.069	0.535 / 0.073

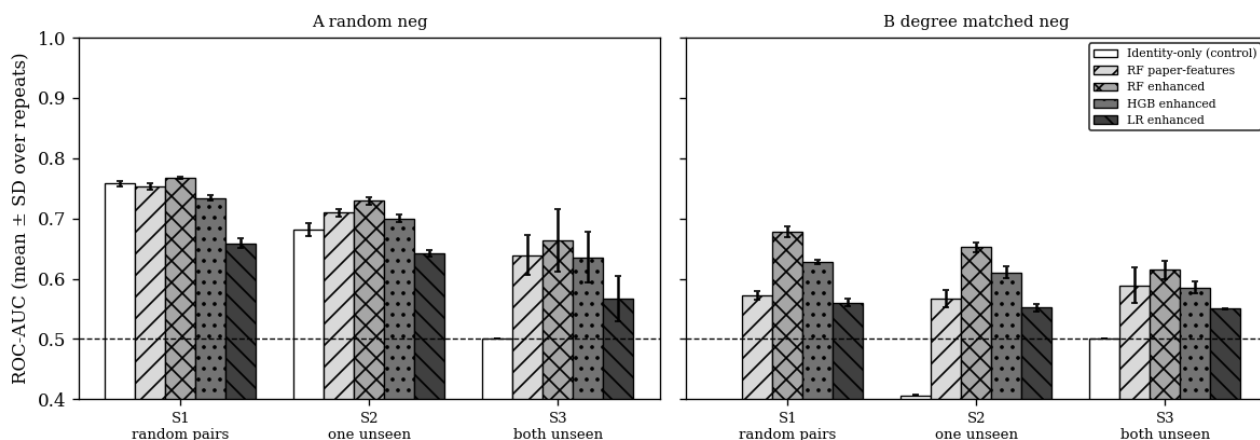


Fig. 2. ROC-AUC (mean $\pm$ SD over three repeats) for each model under the three regimes, on benchmark A (left) and benchmark B (right). Dashed line: chance. The identity-only AUC on benchmark B for S1 (0.366) lies below the plotted range.

Benchmark A (random negatives). The identity-only control, which sees no chemistry, reaches AUC 0.758 and accuracy 69.7% on random pairs. This matches the RF on the original concatenated fingerprints (AUC 0.753, accuracy 68.8%) and is only slightly below the RF on enhanced features (AUC 0.768, accuracy 70.0%). On this benchmark, most of what a fingerprint model appears to learn can therefore be obtained from drug identity alone. When both drugs are unseen the control carries no information (AUC 0.500, since all of its inputs fall back to the prior), whereas the fingerprint models retain some signal (RF enhanced, 0.663). Part of the structural signal is thus real, but it is much smaller than the S1 figures suggest.

Benchmark B (degree-matched negatives). Removing the frequency shortcut changes the picture. The identity-only control falls to AUC 0.366 on random pairs and 0.407 with one unseen drug, which is below chance. This is likely a consequence of the construction of B: each drug has roughly equal numbers of positive and negative pairs overall, so a drug that looks interaction-prone in the training portion tends to have fewer positive pairs left in the held-out portion. The v1

representation also falls sharply, from AUC 0.753 on A to 0.572 on B (accuracy 68.8% to 55.1%), close to chance. The enhanced RF keeps AUC 0.678 and accuracy 62.1% (MCC 0.242) on random pairs.

Cold start. Performance declines as more of the test pair is unseen. For the enhanced RF on B, AUC falls from 0.678 (S1) to 0.653 (S2) to 0.615 (S3), and accuracy from 62.1% to 60.5% to 57.6%. S3 is the closest analogue of a pair of newly approved drugs, and the result there is modest but above chance.

Model class. On the enhanced features the RF ranks first, ahead of gradient boosting and logistic regression, in every regime on both benchmarks. Logistic regression is close to chance on B (AUC 0.550–0.560), which suggests that the signal is non-linear in the pair features.

C. Feature Ablation

TABLE IV. FEATURE ABLATION, RF ON BENCHMARK B (ONE REPEAT): ROC-AUC / ACCURACY.

Feature set (dim.)	S1	S2	S3
concat256 (512)	0.568 / 0.544	0.583 / 0.552	0.555 / 0.526
sym256 (513)	0.652 / 0.606	0.630 / 0.591	0.582 / 0.552
sym256+phys (533)	0.671 / 0.620	0.658 / 0.614	0.615 / 0.568
sym2048 (4,097)	0.635 / 0.600	0.639 / 0.599	0.573 / 0.540
sym_all (4,452)	0.714 / 0.656	0.697 / 0.635	0.647 / 0.611

Replacing concatenation by the order-invariant construction raises AUC by 0.084, 0.047 and 0.027 in S1, S2 and S3 respectively; because sym256 also adds the Tanimoto scalar, the effects of order-invariance and of the similarity term cannot be separated here. Adding the ten descriptors gives a further 0.019, 0.028 and 0.033. Moving from 256 to 2,048 bits does not help the RF consistently (S1 0.635 vs 0.652, S2 0.639 vs 0.630, S3 0.573 vs 0.582). The richest set (2,048-bit Morgan, MACCS keys and descriptors) is best in all regimes (AUC 0.714, 0.697, 0.647), but it comes from a single repeat, was run after the main comparison, and is not the set used by the deployed model.

D. Calibration and Selective Prediction

TABLE V. CALIBRATION OF RF SCORES ON BENCHMARK B (ENHANCED FEATURES).

Regime / scores	ECE	Brier
S1, raw RF	0.0483	0.2307
S1, isotonic	0.0215	0.2278
S3, raw RF	0.0527	0.2405
S3, isotonic	0.0568	0.2389

In S1 the raw RF scores are reasonably calibrated (ECE 0.048; Fig. 3, left), and isotonic regression roughly halves the ECE to 0.022. Under cold start (S3) it does not help (ECE 0.053 raw, 0.057 isotonic). All Brier scores (0.228–0.241) remain close to the 0.25 of an uninformative constant score of 0.5 on balanced data, in line with the modest discrimination reported above.

TABLE VI. SELECTIVE PREDICTION ON BENCHMARK B (CALIBRATED RF, FIRST REPEAT).

δ	Coverage	Accuracy	FNR
S1 (random pairs)			
0.00	100%	61.8%	43.9%
0.10	38.7%	71.6%	29.3%
0.20	11.1%	79.7%	13.3%
S3 (both drugs unseen)			
0.00	100%	57.8%	60.7%
0.05	52.0%	63.1%	63.4%
0.10	20.8%	72.2%	50.0%

Abstaining on pairs whose calibrated score lies within 0.10 of 0.5 leaves 38.7% of random pairs answered, with accuracy 71.6% (61.8% on all pairs) and FNR 29.3% (43.9%). For both-unseen pairs the same rule leaves 20.8% of pairs at accuracy 72.2%, but the FNR is still 50.0%. Estimates at larger δ rest on few pairs (Fig. 3, right) and should be read with caution. Abstention therefore raises accuracy on the answered pairs, but it does not remove the risk of missing real interactions.

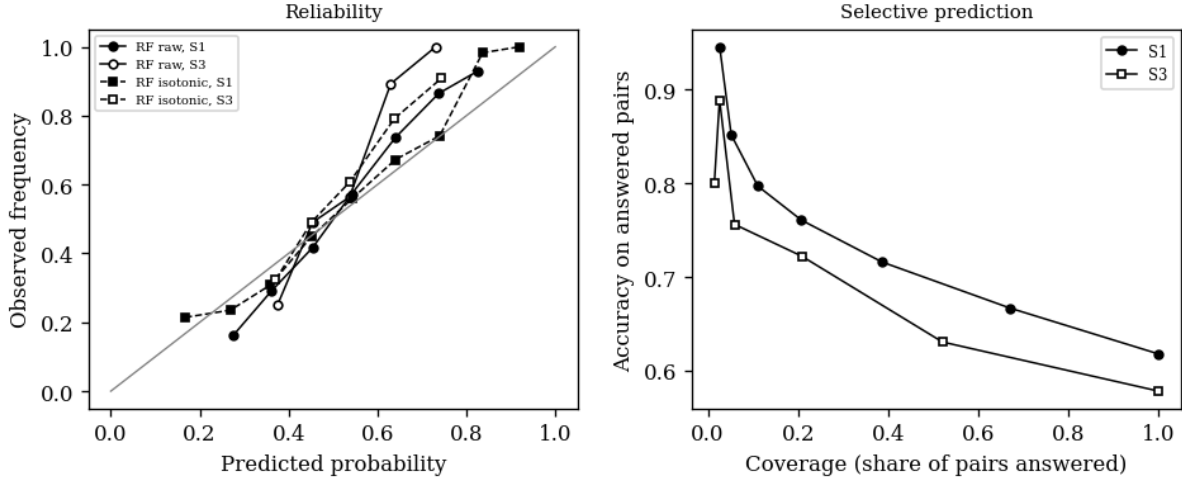


Fig. 3. Left: reliability curves for raw and isotonic-calibrated RF scores on benchmark B, S1 (filled) and S3 (open markers). Right: accuracy on answered pairs against coverage for the calibrated RF (first repeat).

E. Case Panel

TABLE VII. CASE PANEL: SCORES FROM THE FINAL MODEL.

Pair	In list	Score	Tier
Aspirin + Warfarin	Yes	0.525	Uncertain
Simvastatin + Clarithromycin	Yes	0.530	Uncertain
Sildenafil + Nitroglycerin	Yes	0.385	None predicted
Fluoxetine + Tramadol	Yes	0.600	Likely
Clopidogrel + Omeprazole	Yes	0.410	Uncertain
Digoxin + Amiodarone	Yes	0.470	Uncertain
Methotrexate + Trimethoprim	Yes	0.500	Uncertain
Theophylline + Ciprofloxacin	No	0.355	None predicted
Warfarin + Fluconazole	Yes	0.505	Uncertain
Ketoconazole + Simvastatin	Yes	0.645	Likely

Nine of the ten pairs are recorded interactions in the source list. Two of these are flagged as likely (Fluoxetine + Tramadol, Ketoconazole + Simvastatin), six fall in the uncertain band, and one, Sildenafil + Nitroglycerin, is scored below 0.40 and would

be read as “no interaction predicted”. Aspirin + Warfarin, the false negative (37.0%) of our v1 model, now scores 0.525 and is reported as uncertain; the two models and benchmarks differ, so this is not a like-for-like comparison. Theophylline + Ciprofloxacin is not in the list and is also scored below 0.40, although the interaction is widely reported in clinical references, which illustrates the incomplete-label problem. The panel is illustrative only: it has ten pairs, was selected by hand, and the drugs were seen in training, so it is not a performance estimate.

F. Prototype

Fig. 4 shows the prototype for Metformin + Paracetamol, chosen only to illustrate the interface. The model score of 54.5% falls in the uncertain band, so the application advises checking a clinical reference. Metformin was resolved from the local lookup table and Paracetamol, which was not in it, from live PubChem; the SMILES and their sources are displayed for transparency, together with the substructures that carry the highest global importance in the model.

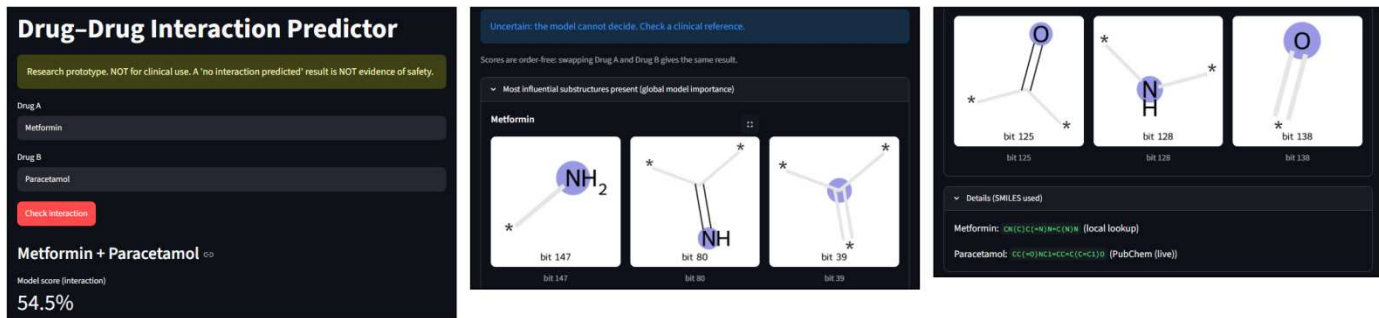


Fig. 4. Streamlit prototype for Metformin + Paracetamol. (a) Persistent warning, inputs and model score. (b) The “uncertain” tier and the highest-importance Morgan substructures present in Metformin (bits 147, 80, 39). (c) Substructures for Paracetamol (bits 125, 128, 138) and the SMILES used, with their sources.

V. STATISTICAL CONSIDERATIONS

For an S1 test set of $n = 3,505$ pairs, the binomial standard error of an accuracy near 0.62 is about 0.008, so test-set sampling alone gives a 95% interval of roughly ± 1.6 percentage points, and an accuracy of 0.621 lies far above the 0.5 chance level ($z \approx 14$). This interval ignores variation from training and from the choice of split. We ran only three repeats, the splits are drawn from the same pool of positives and are not independent, and the repeat-to-repeat SDs in Fig. 2 are therefore rough; we do not report p-values for differences between models. The S1 standard deviations are small, whereas those for S3, whose test sets contain only pairs with two held-out drugs, are visibly larger (up to a few hundredths of AUC), so the S3 estimates are the least certain. The ablation (Table IV) and the calibration and selective-prediction analyses (Tables V–VI) use a single repeat and are descriptive. The enhanced configuration sym256+phys was used in the main comparison and in the deployed model; the ablation was run afterwards on the same benchmark, so its best configuration should be regarded as exploratory and not as a tuned result.

VI. DISCUSSION

A. Interpretation of Results

Three conclusions follow. First, much of the performance of a fingerprint model on a simple random-negative benchmark can

be reproduced from drug identity alone (Table II, benchmark A), so accuracy figures from such benchmarks overstate how much chemistry is being learned. Our v1 result of 65.3% on a smaller sample with weaker negatives should be read in that light. Second, once the frequency shortcut is removed, the pair-level structural signal is real but modest: AUC 0.678 on random pairs and 0.615 when both drugs are unseen. Third, the representation matters more than the model class. Order-invariant pair features with Tanimoto similarity and physicochemical descriptors raised AUC on B (S1) from 0.572 to 0.678 with the same RF, and a richer feature set gave a further gain in the ablation, whereas switching from RF to gradient boosting or logistic regression did not help.

B. Mechanism-Limited Misses

The case panel illustrates why structure alone is not enough. Aspirin + Warfarin is a pharmacodynamic interaction (platelet inhibition together with reduced clotting-factor synthesis), and Sildenafil + Nitroglycerin involves overlapping effects on vascular tone. Neither follows from structural similarity between the two molecules, and the model gives the first an uncertain score and the second a low one. Many other interactions are pharmacokinetic, for example through CYP450 enzyme inhibition, and are also invisible to a model that has no enzyme, dose or protein-binding information. The high-importance bits displayed by the prototype are small and diffuse

(top global importances of about 0.005) and describe what the model uses on average; they are not mechanistic explanations.

C. Limitations

Labels. Pairs absent from the list are unlabelled, not known to be safe. Excluding all recorded pairs from the negatives removes the error found in the v1 protocol, but unrecorded true interactions can remain among the negatives, as the Theophylline + Ciprofloxacin example suggests. Degree matching. Benchmark B keeps 8,761 of 15,000 sampled positives (about 4.6% of all recorded pairs) and drops pairs of highly connected drugs, so it under-represents exactly those drugs. The below-chance identity control on B is a by-product of the construction (Section IV-B) and is used only as a diagnostic. Cold start. Held-out drugs were chosen at random, not by scaffold or therapeutic class, so close analogues of a held-out drug can remain in training and S2 and S3 may still be optimistic for truly novel chemotypes. Structures. SMILES come from the first PubChem match for a name without manual curation; salts, mixtures and large molecules are not treated specially. A 256-bit hash can collide distinct substructures, and descriptor standardisation used all drugs. Scope. The model has no dose, enzyme, protein-binding or patient information, predicts a binary recorded label with no severity, uses 15,000 sampled positives, three repeats and one dataset, and the 0.40/0.60 tier thresholds were not tuned. The deployed model is an uncalibrated RF.

D. Threats to Validity

Internal validity is limited by three repeats, one ablation repeat and the sampling of drug sets. External validity is limited by the single source list, the exclusion of 13 drugs without structures and of unmatched positives, and the absence of an independent test set. Construct validity is limited because a label in the source list does not encode severity, mechanism or clinical relevance: the model predicts a recorded interaction, not clinical risk.

E. Practical Significance and Ethical Considerations

The work shows a complete pipeline from public data (PubChem, RDKit) to a user-facing prototype, and, more usefully, a set of controls that a stronger model should pass. It is a research prototype, complementary to and not a replacement for curated interaction databases and clinical judgment. At the 0.5 threshold the calibrated model misses 43.9% of recorded interactions on random pairs and 60.7% when both drugs are unseen (Table VI, $\delta = 0$), so a low score must never be read as evidence of safety. That miss rate is higher than the 34.0% reported under the v1 protocol, which reflects a harder and cleaner benchmark and not a better model. The prototype therefore carries a persistent warning, reports an explicit uncertain band, shows the SMILES it used, and calls its output a model score and not a probability. The system must not be used for clinical decision-making. No patient data were used in this study.

VII. CONCLUSION AND FUTURE WORK

Controlling how negatives are drawn changes the conclusions of a structure-based DDI study. A common protocol mislabels about 21% of its negatives, and on a random-negative

benchmark a drug-identity-only control matches fingerprint models. On degree-matched negatives with all recorded pairs excluded, an RF on order-invariant fingerprint and descriptor features reaches AUC 0.678 on random pairs and 0.615 when both drugs are unseen, which is above chance but far from sufficient for clinical use. We release this as an honest baseline together with the controls needed to test stronger models.

Priority directions are: graph neural networks operating on molecular graphs, as originally proposed, evaluated against the same controls; scaffold- or class-disjoint cold-start splits; scaling to the full list of 191,135 recorded pairs; curated or severity-labelled data and better treatment of unlabelled pairs; pharmacokinetic features such as CYP450 substrate and inhibitor status; selecting thresholds and calibration on validation data and deploying a calibrated model; the richer feature set from the ablation; and independent external validation, including clinically important pairs such as anticoagulants and CYP450 substrates.

ACKNOWLEDGMENT

The authors thank the faculty of the Department of Artificial Intelligence and Data Science, Sanjivani University, for their guidance and feedback throughout this project.

REFERENCES

- [1] J. Gao, Z. Wu, R. Al-Sabri, B. M. Oloulade, and J. Chen, "AutoDDI: Drug-drug interaction prediction with automated graph neural network," IEEE J. Biomedical and Health Informatics, 2024.
- [2] S. Liu, Y. Zhang, Z. Cui, et al., "Enhancing drug-drug interaction prediction using deep attention neural networks," IEEE/ACM Trans. Computational Biology and Bioinformatics, 2023.
- [3] X. Chen, X. Liu, and J. Wu, "GCN-BMP: Investigating graph representation learning for DDI prediction task," Methods, 2020.
- [4] Y.-H. Feng and S.-W. Zhang, "Prediction of drug-drug interaction using an attention-based graph neural network on drug molecular graphs," Molecules, vol. 27, no. 9, 2022.
- [5] K. Lee, D. Lee, S. Jeon, J. Choi, and J. Kang, "Drug-drug interaction analysis using heterogeneous biological information network," in Proc. IEEE Int. Conf. Bioinformatics and Biomedicine, 2012.
- [6] S. Vilar, E. Uriarte, L. Santana, et al., "Similarity-based modeling in large-scale prediction of drug-drug interactions," Nature Protocols, vol. 9, pp. 2147–2163, 2014.
- [7] J. Y. Ryu, H. U. Kim, and S. Y. Lee, "Deep learning improves prediction of drug-drug and drug-food interactions," Proc. Natl. Acad. Sci. USA, vol. 115, no. 18, pp. E4304–E4311, 2018.
- [8] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," Bioinformatics, vol. 34, no. 13, pp. i457–i466, 2018.
- [9] P. Dewulf, M. Stock, and B. De Baets, "Cold-start problems in data-driven prediction of drug-drug interaction effects," Pharmaceuticals, vol. 14, no. 5, art. 429, 2021.
- [10] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2002, pp. 694–699.
- [11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Machine Learning, PMLR vol. 70, 2017, pp. 1321–1330.
- [12] D. S. Wishart et al., "DrugBank 5.0: a major update to the DrugBank database for 2018," Nucleic Acids Research, vol. 46, no. D1, pp. D1074–D1082, 2018.

- [13] S. Kim et al., "PubChem in 2021: new data content and improved web interfaces," *Nucleic Acids Research*, vol. 49, no. D1, pp. D1388–D1395, 2021.
- [14] G. Landrum, "RDKit: Open-source cheminformatics," <https://www.rdkit.org>.
- [15] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *J. Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010.
- [16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [17] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [18] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, art. 6, 2020.