JOURNAL OF
Emerging Paradigms
in Computing Systems
JEPCS

Journal of Emerging Paradigms in Computing Systems (JEPCS) | Volume 03, Issue No. 02, 2026

ARTICLE TYPE: Research Article

Article ID: JEPCS-2026-0302-0068

Coverage Degradation under Violated Exchangeability: Calibration and Conformal Prediction for Drug-Drug Interaction Prediction on Structurally Novel Drugs

Bajpayee Singha[1], **Ajay Kumar Sinha**[1,*]

[1] School of Artificial Intelligence, Yanshan University, Qinhuangdao 066004, China

*Corresponding author: ajaysinha@stumail.ysu.edu.cn

Contributing authors: singhabajpayee@stumail.ysu.edu.cn, ajaysinha@stumail.ysu.edu.cn

Received: October 14, 2026 **Accepted:** November 05, 2026

Abstract- Deep learning models for drug-drug interaction (DDI) prediction are increasingly evaluated on drugs absent from training, yet the reliability of their uncertainty estimates in this cold-start setting has not been examined. We present ColdCal, a benchmark that measures calibration and conformal coverage for multi-class DDI type prediction as drug novelty increases. Using DrugBank data covering 1,704 drugs, 191,870 interacting pairs and 86 interaction types, we held out about 20% of drugs under three regimes of increasing structural novelty (random selection and Butina clustering at Tanimoto cutoffs of 0.6 and 0.7) and evaluated pairs containing one or two unseen drugs. A fingerprint-based multilayer perceptron and two graph neural networks, SSI-DDI and DSN-DDI, were combined with temperature scaling, deep ensembles, and three split-conformal procedures over nine splits. In distribution, all models were well calibrated, and conformal sets met the 90% target. For unseen drugs, where exchangeability is violated and the guarantee no longer applies, empirical coverage of least-ambiguous conformal sets fell to 0.43-0.66 when one drug was new and to 0.13-0.44 when both were new, decreasing steadily with structural novelty. Temperature scaling, ensembling and class-conditional calibration recovered little, and adaptive prediction sets restored coverage only by enlarging sets to 34-68 of 86 types for the perceptron. Coverage rose with the new drug's similarity to training drugs for both pair types but stayed below target in every similarity bin. In this benchmark, uncertainty calibrated on known drugs did not transfer to new drugs, which motivates shift-aware recalibration.

Keywords- Drug-drug interaction; uncertainty quantification; conformal prediction; model calibration; cold-start prediction

1. Introduction

Adverse drug-drug interactions (DDIs) are a common and largely preventable cause of harm, particularly in patients who take several medicines at once. Because the number of possible drug pairs grows quadratically with the number of drugs, only a small fraction can ever be studied clinically. Computational models that predict whether, and how, two drugs interact from their chemical structure are therefore an active research area. Early work framed the task as classification over interaction types derived from DrugBank descriptions [1], and later models learned drug representations from interaction networks [2] or directly from molecular graphs [3], [4].

For a newly approved or investigational compound, no interactions have been recorded, so the practically important setting is the cold start. Recent graph models are evaluated in exactly this way: whole drugs, rather than drug pairs, are held out, and performance is reported separately for test pairs that contain one unseen drug (S1) and pairs in which both

drugs are unseen (S2) [3][5]. These studies consistently show that accuracy drops sharply in S1 and further in S2. They report accuracy, however, and say little about whether the models know when they are likely to be wrong.

If such predictions are used to prioritize candidate interactions for expert review, knowing when they are unreliable matters as much as their accuracy. Two families of tools address it. Calibration asks whether a model's stated confidence matches its observed accuracy [6]. Conformal prediction wraps any trained classifier so that it returns a set of plausible labels that contains the true label with a user-chosen probability, for example 90% [7], [8]. This guarantee is distribution-free, but it rests on one assumption: the calibration data and the test data must be exchangeable [7], [8]. A drug that never appeared in training violates that assumption by construction, so the guarantee no longer applies, and theory gives no bound on the resulting loss without further information about the shift [9]. How much empirical coverage degrades in DDI prediction, and whether the degradation depends on how structurally unusual the new drug is, has not been measured. Main contributions of the paper are:

- ColdCal is a benchmark for uncertainty quantification in multi-class DDI type prediction that grades novelty along two axes: structural distance of the held-out drugs (random, Butina-0.6 and Butina-0.7 splits) and pair novelty (S1 and S2), with an in-distribution calibration set and an in-distribution test set as references;
- A systematic evaluation of three models of increasing specificity to the cold-start problem, a fingerprint-based multilayer perceptron, SSI-DDI [3] and DSN-DDI [4], combined with temperature scaling, deep ensembles and three split-conformal procedures over nine splits;
- Evidence empirical coverage, which matches its 90% target in distribution, degrades to 0.13-0.66 on unseen drugs, in proportion to structural novelty, and that standard remedies do not repair it; and
- An analysis linking coverage degradation to the new drug's similarity to the training set, reported separately for S1 and S2 pairs, together with released splits, audit tables and code.
  The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the data, the splitting protocol, the models and the uncertainty methods. Section 4 gives the experimental setup, Section 5 presents and discusses the results, and Section 6 concludes.

2. Related Work

DeepDDI predicted 86 interaction types from structural similarity profiles of the two drugs [1]. Decagon modeled polypharmacy side effects with graph convolutions over a multimodal network of drugs and proteins [2]. SSI-DDI represents each drug as a molecular graph, extracts substructures at several depths with graph attention layers [10] and scores their pairwise interactions with a RESCAL-style relation head [3], [11]. DSN-DDI extends this design with an inter-graph view in which every atom of one drug attends to every atom of the other, and it was designed and evaluated for strong performance on unseen drugs [4].

Random splits over drug pairs place most test drugs in the training set as well, which inflates reported performance. Dewulf et al. distinguished prediction tasks by how much information about each drug is available and showed that cold-start tasks are substantially harder [5]. Inductive protocols that hold out drugs and report S1 and S2 separately are now standard for graph-based DDI models [3], [4]. In molecular property prediction, scaffold and cluster-based splits go further by keeping structurally similar molecules out of the test set [12]. ColdCal combines both ideas: drugs are held out by Butina clustering of circular fingerprints [13], [14], which produces graded structural novelty within the S1/S2 protocol.

Modern neural networks tend to be overconfident, and a single temperature fitted after training is a simple and strong correction in distribution [6]. Deep ensembles improve both accuracy and the quality of predictive uncertainty [15]. Under dataset shift, however, the calibration of all of these methods degrades, and the degradation grows with the size of the shift [16]. To our knowledge, this has not been quantified for DDI prediction on unseen drugs.

Split conformal prediction gives finite-sample, distribution-free coverage when calibration and test points are exchangeable [7], [8]. The least-ambiguous set-valued classifier (LAC) produces the smallest sets on average [17]; adaptive prediction sets (APS) let set size grow with the difficulty of each input [18] and class-conditional (Mondrian) variants target coverage within each class rather than on average [19]. Weighted conformal prediction restores validity under covariate shift when the likelihood ratio between test and calibration distributions is known [20], and more general results bound the coverage loss in terms of how far the data depart from exchangeability [9]. ColdCal supplies a realistic, graded shift on which such methods can be tested.

3. Methodology

3.1. Dataset

We used the DrugBank-derived dataset distributed with SSI-DDI [3], [21]. It contains 1,704 drugs with SMILES strings, 191,870 ordered drug pairs and 86 interaction types; all SMILES were parsed successfully with RDKit [22]. Interaction

Table 1: *Split Statistics Averaged over Three Seeds.*

Split	New drugs	Train pairs	Calib.	ID test	S1	S2	Reassigned	Mean max-Tan.	Fraction < 0.3
Random	341	94,977	18,561	5,258	60,608	7,470	26	0.498	0.228
Butina-0.6	324	100,865	19,710	5,585	54,699	5,707	10	0.347	0.413
Butina-0.7	333	94,681	18,496	5,242	60,865	7,605	9	0.337	0.529

Note: "Reassigned" denotes S1 pairs moved to S2 during the audit. Similarity statistics refer to the unseen drug in S1 pairs.

types are directional (for example, the first drug may increase the effect of the second), so the order of each pair was preserved throughout, and every model receives the two drugs in that order.

3.2. Drug Hold-Out and Novelty Regimes

Following the inductive protocol of SSI-DDI [3], about 20% of drugs were designated new and the remainder old. Training pairs contain two old drugs, S1 pairs contain exactly one new drug, and S2 pairs contain two new drugs. The new drugs were chosen in three ways. In the Random regime they were sampled uniformly, so test drugs are unseen by identity but often have close structural analogues in training. In the Butina-0.6 and Butina-0.7 regimes, drugs were clustered with the Butina algorithm [13] on ECFP4 fingerprints (radius 2, 2,048 bits) [14] at Tanimoto distance thresholds of 0.6 and 0.7, and whole clusters were held out until about 20% of drugs were new, so that test drugs lack close analogues. Each regime was generated with three seeds, giving nine splits in total. Fingerprints were computed with RDKit's AllChem.GetMorganFingerprintAsBitVect (radius 2, 2,048 bits) for the 1,704 drugs sorted by DrugBank identifier. The values 0.6 and 0.7 are Tanimoto distance thresholds, where distance = 1 - Tanimoto similarity, so a drug joins a cluster when its similarity to the cluster centroid is at least 0.4 or 0.3, respectively. The lower-triangular distance list was passed to rdkit.ML.Cluster.Butina.ClusterData (dists, 1704, cutoff, isDistData=True), which returned 892 clusters at 0.6 and 670 clusters at 0.7; each cluster was indexed by its position in the returned list. For each seed $s \in \{0, 1, 2\}$, the cluster indices were shuffled with NumPy's default_rng(s), and whole clusters were taken in the shuffled order. A cluster was skipped if adding it would raise the number of held-out drugs above 1.05×340.8 (20% of 1,704), and selection stopped once at least 0.95×340.8 drugs were held out. This gave 324-325 new drugs under Butina-0.6 and 324-339 under Butina-0.7. In the Random regime, round($0.2 \times 1{,}704$) = 341 drugs were drawn without replacement with default_rng(s).

To quantify how novel each test pair is, we computed for every unseen drug d its maximum Tanimoto similarity to the set of training drugs D_{train}:

$$\nu(d) = \max_{d' \in D_{\text{train}}} \text{Tan}(f_d, f_{d'}) \tag{1}$$

where f_{d} is the ECFP4 fingerprint of drug d. For an S1 pair, the pair novelty is ν of its unseen drug; for an S2 pair, it is the smaller of the two values, so that a pair is treated as being as novel as its most unusual member.

3.3. Calibration, Validation and Reference Sets

A calibration set was carved from the training pairs (15% of pairs, sampled at random with default_rng(s)). Any pair whose removal would have left one of its drugs absent from the remaining training data was returned to training. The calibration set therefore contains no new drugs. This is deliberate: because calibration and training data are exchangeable, any loss of coverage on S1 or S2 can be attributed to drug novelty rather than to a mismatch between calibration and test data. From the remaining training pairs we carved, with the same guarantee, an in-distribution (ID) test set (5%), which shows the coverage obtained when exchangeability holds, and a validation set (5%) used only for early stopping, sampled with pandas random states $1000 + s$ and $2000 + s$, respectively.

An audit of the splits showed that a few drugs labeled old appeared in S1 pairs but in no training pair, so the models never saw them. Pairs in which both drugs were therefore unseen were reassigned from S1 to S2, This affected 9-26 pairs per regime on average. Table 1 also confirms that the Butina regimes create real structural shift: the mean similarity of new drugs to their nearest training drug falls from 0.498 under Random to 0.347 and 0.337, and the two cutoffs differ mainly in the tail, with 41% and 53% of S1 new drugs below a similarity of 0.3.

3.4. Models

All models output 86 logits for an ordered drug pair and were trained with softmax cross-entropy. The **fingerprint MLP (FP-MLP)** concatenates the ECFP4 fingerprints of the two drugs and passes them through two hidden layers. **SSI-DDI** [3] and **DSN-DDI** [4] were run from their official implementations [23], [24], with their original atom featurizers, graph encoders, co-attention modules and RESCAL relation heads. Both were published as triple classifiers trained with negative sampling. To obtain class probabilities comparable with the MLP, we scored each pair against all 86 relation matrices of

Table 2: *Model and Training Settings.*

Setting	FP-MLP	SSI-DDI	DSN-DDI
Input	ECFP4 fingerprints of both drugs (2 × 2,048 bits)	Molecular graphs (55 atom features)	Molecular graphs (55 atom features)
Encoder	Two layers with 1,024 and 512 units, ReLU, and dropout 0.3	4 GAT blocks with 2 heads × 32 features	4 blocks with intra- and inter-graph attention and 2 heads × 64 features
Output head	Linear layer, 86 interaction types	RESCAL, 64 × 64 matrix per type	RESCAL, 128 × 128 matrix per type
Optimizer	AdamW, weight decay 10^{-4}	Adam, weight decay 5×10^{-4}	Adam, weight decay 5×10^{-4}
Learning rate	10^{-3}	10^{-3}, multiplied by 0.96 per epoch	10^{-3}, multiplied by 0.96 per epoch
Batch size	512	512	256
Max. epochs / patience	40 / 6	25 / 4	25 / 4
Runs per split	5 (ensemble members)	1	1

the unchanged RESCAL head and multiplied the scores by a learnable scale, initialized at 5, to form logits. No other part of either architecture was modified. Training settings are listed in Table 2 (Section 4.1).

3.5. Post-Hoc Calibration

In temperature scaling [6], the logits z of a trained model are divided by a scalar $T > 0$ before the softmax:

$$p_T(k \mid x) = \frac{\exp(z_k/T)}{\sum_{j=1}^{86} \exp(z_j/T)} \tag{2}$$

T was fitted on the calibration set by minimizing negative log-likelihood (NLL). For FP-MLP we also built a five-member deep ensemble [15]. All five members were trained on the same split files and used the same training, validation, calibration, ID test, S1 and S2 partitions; they differed only in the random seed that controls weight initialization and the order of mini-batches. The ensemble prediction is the average of the five members' softmax probabilities computed from their unscaled logits. Temperature scaling was never applied to the ensemble or to its members when they were averaged. The ensemble was treated as a single model: its conformal thresholds were fitted on the calibration set using the averaged ensemble probabilities, and all its metrics were computed from those probabilities. The single-model FP-MLP results use member 0 with temperature scaling.

3.6. Split Conformal Prediction

Let $s(x, y)$ be a nonconformity score and let $s_1, \ldots, s_n$ be its values on the n calibration pairs. For a miscoverage level α, the conformal threshold is the $\lceil (n+1)(1-\alpha) \rceil$-th smallest calibration score, where subscripts in parentheses denote order statistics:

$$\hat{q} = s_{(\lceil (n+1)(1-\alpha) \rceil)} \tag{3}$$

and the prediction set for a test pair x is $C(x) = \{k : s(x, k) \leq \hat{q}\}$. When calibration and test pairs are exchangeable, the true type lies in $C(x)$ with probability at least $1 - \alpha$ [7], [8]. For S1 and S2 pairs this assumption does not hold, because the calibration set contains only drugs seen in training, so the guarantee does not apply to them. We therefore do not test the guarantee itself, which remains valid under its assumptions; we measure how far empirical coverage degrades when exchangeability is violated in this way. We used $\alpha = 0.1$ and computed all scores from temperature-scaled probabilities. LAC [17] uses the score

$$s(x, y) = 1 - p_T(y \mid x) \tag{4}$$

APS [18] uses the total probability of all types ranked at or above the true type (non-randomized version):

$$s(x, y) = \sum_{k:\, p_T(k|x) \geq p_T(y|x)} p_T(k \mid x) \tag{5}$$

The class-conditional (Mondrian) variant of LAC [19] computes a separate threshold for each interaction type that has at least 100 calibration pairs (about 53 of the 86 types) and uses the marginal threshold for the remaining types.

3.7. Evaluation Protocol

Algorithm 1 summarizes the procedure applied to each of the nine splits and each model. All uncertainty methods are post hoc, so every model is trained once per split and all methods are evaluated from the same saved logits.

Algorithm 1 ColdCal Evaluation of One Model on One Split

Require: Split containing training, calibration, S1, and S2 pairs; model family M; miscoverage level α
Ensure: Accuracy, calibration, and conformal-prediction metrics on the ID test set, S1, and S2
Move to S2 each S1 pair whose nominally old drug does not occur in the training pairs
Create the ID test and validation sets from the training pairs while retaining every drug in the training set
Train M on the training pairs and retain the epoch with the lowest validation NLL
Compute logits for the calibration, ID test, S1, and S2 pairs
Fit temperature T using the calibration pairs according to Eq. (2)
for each method $m \in \{\text{LAC}, \text{APS}, \text{Mondrian-LAC}\}$ **do**
  Compute calibration scores according to Eq. (4) or Eq. (5)
  Compute conformal threshold $\hat{q}$ according to Eq. (3)
  for each test set $S \in \{\text{ID test}, \text{S1}, \text{S2}\}$ **do**
    Construct prediction set $C(x)$ for every $x \in S$
    Record coverage, mean set size, empty-set rate, and CovGap
  end for
end for
Record accuracy, macro-F1, ECE, NLL, AURC, and coverage within novelty bins according to Eq. (1)
return All evaluation metrics

4. Experimental Setup

4.1. Implementation and Training

All experiments were implemented in Python with PyTorch [23], PyTorch Geometric [24] and RDKit [22], and were run on Kaggle notebooks with an NVIDIA Tesla T4 GPU. The recorded environment was Python 3.13.15, PyTorch 2.11.0 (CUDA 12.8), pandas 2.3.3, NumPy 2.1.3, scikit-learn 1.6.1 and PyArrow 23.0.1; RDKit and PyTorch Geometric were installed from PyPI in October 2026 (the latest releases at that time), and their exact version numbers were not logged. Model weights were initialized with torch.manual_seed($100s + m$) for split seed s and ensemble member m ($m = 0$ for single models). Each model was trained once per split; the three seeds of each regime provide the variability reported as standard deviations. Together with the five ensemble members, this amounts to 63 training runs. The validation set was used only for early stopping and never for calibration. Table 2 lists the training settings.

4.2. Evaluation Metrics

Predictive performance was measured by top-1 accuracy and macro-averaged F1 score. Calibration was measured by the expected calibration error (ECE) [25], computed on top-label confidence with 15 equal-width bins $B_1, \ldots, B_{15}$:

$$\text{ECE} = \sum_{b=1}^{15} \frac{|B_b|}{N} \left|\text{acc}(B_b) - \text{conf}(B_b)\right| \tag{6}$$

where N is the number of test pairs and acc and conf are the accuracy and mean confidence within a bin. For conformal sets we report coverage (fraction of pairs whose true type is in the set), mean set size, the empty-set rate and CovGap, the mean absolute deviation of per-type coverage from 90% in percentage points over types with at least 100 calibration and 30 test pairs [19]. The usefulness of confidence for deciding which predictions to act on was measured by the area under the risk-coverage curve (AURC) [26], in which pairs are answered in order of decreasing confidence. Finally, LAC coverage was computed within four bins of pair novelty (Eq. 1): below 0.3, 0.3-0.4, 0.4-0.5 and at least 0.5, keeping bins with at least 50 pairs, separately for S1 and S2 pairs. All results are means over the three seeds of each regime, with standard deviations.

5. Results and Discussion

5.1. In-Distribution Reference

On the ID test set all three models were accurate and well calibrated (Table 3). Accuracy was 0.950-0.953 for FP-MLP, 0.895-0.903 for DSN-DDI and 0.845-0.852 for SSI-DDI, and ECE after temperature scaling was at most 0.018. Every conformal method met its target: LAC coverage was 0.896-0.907 with about one interaction type per set, and CovGap was 2.1-2.9 points (Table 4). These results confirm that the process behaves as theory predicts when its assumption holds, so the coverage degradation described below can be attributed to drug novelty. The fitted temperatures (1.13-1.49) were all above 1, indicating mild overconfidence even before any shift.

Table 3: *Accuracy and Calibration Results Averaged over Three Seeds.*

Model	Split	Test	Accuracy	Macro-F1	ECE (Raw)	ECE (TS)
FP-MLP	Random	ID	0.953 ± 0.005	0.921	0.015	0.006
		S1	0.687 ± 0.011	0.595	0.206	0.169
		S2	0.433 ± 0.024	0.266	0.375	0.312
	Butina-0.6	ID	0.951 ± 0.005	0.911	0.018	0.006
		S1	0.631 ± 0.012	0.473	0.252	0.201
		S2	0.368 ± 0.029	0.120	0.391	0.298
	Butina-0.7	ID	0.950 ± 0.004	0.918	0.017	0.008
		S1	0.563 ± 0.040	0.366	0.288	0.236
		S2	0.277 ± 0.037	0.072	0.441	0.352
SSI-DDI	Random	ID	0.845 ± 0.017	0.786	0.036	0.017
		S1	0.597 ± 0.009	0.497	0.242	0.183
		S2	0.388 ± 0.026	0.201	0.421	0.355
	Butina-0.6	ID	0.852 ± 0.005	0.785	0.035	0.014
		S1	0.554 ± 0.016	0.410	0.274	0.213
		S2	0.316 ± 0.027	0.103	0.467	0.397
	Butina-0.7	ID	0.852 ± 0.003	0.772	0.032	0.018
		S1	0.500 ± 0.048	0.320	0.303	0.243
		S2	0.281 ± 0.083	0.074	0.486	0.419
DSN-DDI	Random	ID	0.899 ± 0.012	0.781	0.023	0.017
		S1	0.640 ± 0.011	0.518	0.213	0.168
		S2	0.407 ± 0.028	0.213	0.393	0.337
	Butina-0.6	ID	0.895 ± 0.006	0.803	0.027	0.009
		S1	0.601 ± 0.017	0.431	0.246	0.200
		S2	0.370 ± 0.018	0.100	0.429	0.372
	Butina-0.7	ID	0.903 ± 0.004	0.804	0.026	0.015
		S1	0.540 ± 0.049	0.354	0.285	0.224
		S2	0.296 ± 0.051	0.082	0.462	0.386

Note: Values are means over three seeds. Accuracy is reported as mean ± standard deviation. ID denotes the in-distribution test set, and TS denotes temperature scaling.

5.2. Accuracy and Calibration under Drug Novelty

Accuracy fell on S1 and fell further on S2, and both drops grew with structural novelty (Table 3). From the Random to the Butina-0.7 regime, S1 accuracy decreased from 0.687 to 0.563 for FP-MLP, from 0.640 to 0.540 for DSN-DDI and from 0.597 to 0.500 for SSI-DDI. On S2 under Butina-0.7, macro-F1 was only 0.072-0.082, meaning that most interaction types were almost never predicted correctly for pairs of two new drugs. FP-MLP was the most accurate model on S1 in every regime, whereas DSN-DDI was the most accurate on S2 under both Butina regimes.

Confidence did not fall in step with accuracy. ECE rose from at most 0.036 in distribution to 0.21-0.30 on S1 and 0.38-0.49 on S2 before scaling. The reliability diagrams in Figure 1 show that this error is systematic overconfidence: for all three models the S1 and S2 curves lie below the diagonal over almost the whole confidence range, and on Butina-0.7 the predictions made with the highest confidence were correct only about 70-75% of the time on S1 and about 36-43% of the time on S2.

Standard remedies helped little. Temperature scaling, which was fitted on in-distribution data, reduced S1 ECE by only 0.04-0.06 and left it between 0.17 and 0.24; on S2 it remained between 0.30 and 0.42. The five-member ensemble matched but did not improve on temperature scaling: its S1 ECE was 0.170, 0.203 and 0.241 in the three regimes, against 0.169, 0.201 and 0.236 for the scaled single MLP, while its ID accuracy (0.958-0.961) was only slightly higher. (Table 5).

5.3. Coverage Degradation on Unseen Drugs

Empirical conformal coverage fell far below its 90% target as soon as one drug was new, and it fell further as novelty increased (Table 4 and Figure 2). Across the three single models and the ensemble, LAC coverage was 0.43-0.66 on S1 and 0.13-0.44 on S2. The ordering Random > Butina-0.6 > Butina-0.7 and S1 > S2 held for every model without exception. Because LAC sets stayed close to one type per pair, they effectively reduced to the top-1 prediction on new drugs, and coverage followed accuracy.

The ensemble behaved like the single FP-MLP (Table 5). Its LAC coverage was 0.897-0.905 in distribution, 0.429-0.566 on S1 and 0.144-0.291 on S2, and its APS sets were nearly as large as the single model's (32.8-40.7 types on S1 and 50.9-65.9 on S2). Averaging five members therefore did not reduce the coverage degradation.

Class-conditional calibration did not help. Mondrian coverage was slightly below marginal LAC coverage in every S1 and S2 setting, and CovGap rose from 2-3 points in distribution to 27-50 points on S1 and 56-82 points on S2, so entire interaction types lost most of their coverage.

APS behaved differently for each model. For FP-MLP it kept coverage at 0.94-0.97, but only by returning 34-68 of the 86 types per pair, which carries little useful information. For SSI-DDI it reached about 0.83 on S1 and 0.72-0.75 on S2 with 4.7-6.4 types. DSN-DDI with APS was the only combination that met the target on S1 (0.902-0.911), at a cost of

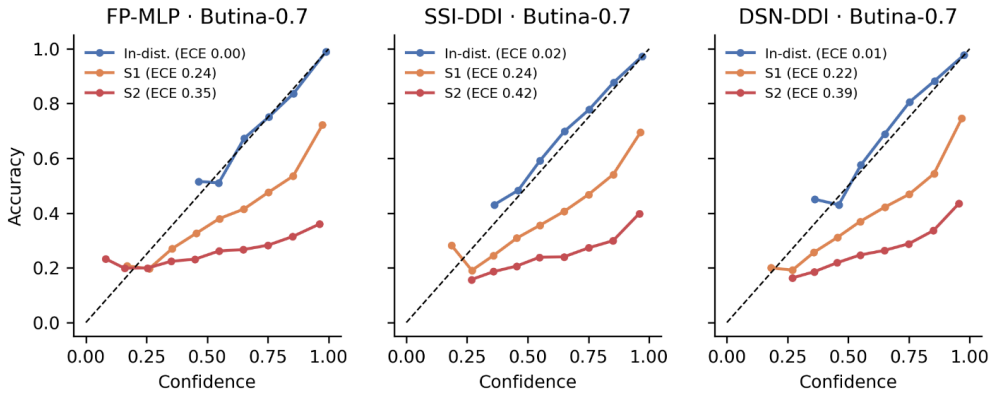


Figure 1: *Reliability diagrams on the Butina-0.7 regime after temperature scaling, pooled over three seeds. Points are accuracy versus mean confidence in ten equal-width bins (bins with at least 30 pairs).*

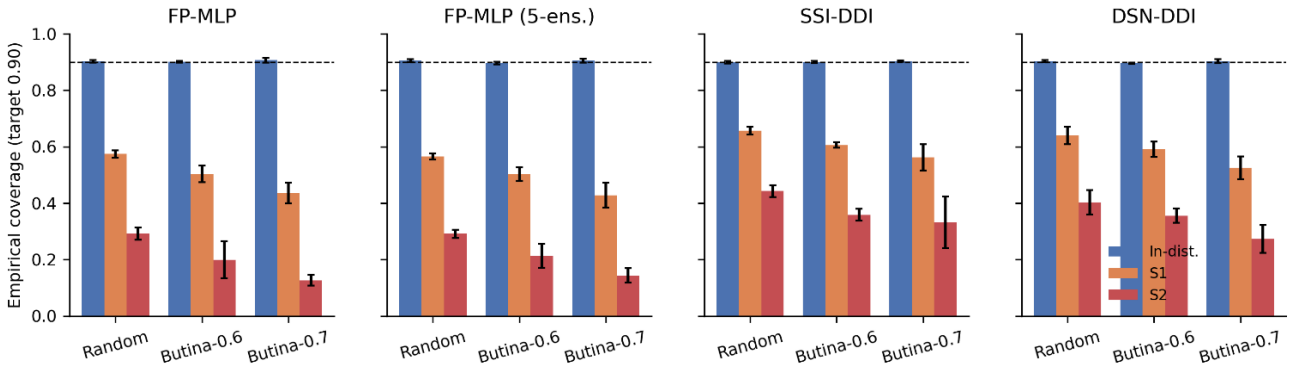


Figure 2: *LAC coverage on the in-distribution test set, S1 and S2 for each model and novelty regime. Bars are means over three seeds and error bars are standard deviations.*

9.2-11.0 types per set, 1.6-1.8 times its in-distribution size; on S2 it still fell short (0.860-0.891). Adaptive sets therefore trade informativeness for coverage, and how favorable that trade is depends strongly on the model.

The empty-set rate revealed a further difference between models. FP-MLP returned empty LAC sets for 7-8% of ID pairs, 27-35% of S1 pairs and 47-63% of S2 pairs: on many new pairs, no type reached the in-distribution threshold. This explains why its LAC coverage was lower than its accuracy, and it also means that an empty set is a usable warning sign for this model. The graph models rarely produced empty sets (at most 1% for SSI-DDI and 14% for DSN-DDI), so this signal is model-dependent.

5.4. Coverage and Structural Similarity

Because S2 pairs are harder by construction and are more common at low similarity, a curve that pools S1 and S2 could confound similarity with pair type. We therefore analyze the two pair types separately (Figure 3). Within each pair type, coverage increased steadily with the similarity of the new drug to its nearest training drug for all three models. Pooled over regimes and seeds, S1 coverage rose from 0.366 below a similarity of 0.3 to 0.668 at 0.5 or above for FP-MLP, from 0.496 to 0.733 for SSI-DDI and from 0.470 to 0.711 for DSN-DDI. On S2 it rose from 0.127 to 0.446, from 0.307 to 0.549 and from 0.263 to 0.523, respectively.

LAC coverage as a function of the maximum Tanimoto similarity of the new drug to the training drugs (Eq. 1), shown separately for S1 pairs (left) and S2 pairs (right). Points are pair-weighted means over the three regimes and three seeds; error bars are standard deviations across splits (bins with at least 50 pairs). The dashed line marks the 90% target.

At matched similarity, S2 coverage was 0.18-0.26 lower than S1 coverage in every bin and for every model, so pair type has an effect beyond the similarity of the most novel drug. Pooling the two pair types had made the similarity effect look slightly stronger: from the lowest to the highest bin, pooled coverage rose by 0.33 (FP-MLP), 0.26 (SSI-DDI) and 0.26 (DSN-DDI), compared with 0.30, 0.24 and 0.24 for S1 alone. The relationship itself, however, holds within each pair type. No bin reached 0.90 for any model, regime or pair type; the highest value, 0.89, came from a bin of only 74 S2 pairs. Two further observations qualify this trend. First, even under the Random regime, S1 pairs whose new drug had a close analogue (similarity of at least 0.5) reached only 0.67-0.75 coverage, so being unseen costs coverage even without

Table 4: *Conformal Prediction Results at 90% Target Coverage.*

Model	Split	Test	LAC coverage	LAC size	Empty	APS coverage	APS size	Mondrian coverage	CovGap
FP-MLP	Random	ID	0.903 ± 0.005	0.92	0.078	0.896	15.5	0.898	2.8
		S1	0.574 ± 0.013	0.73	0.267	0.943	34.1	0.557	37.1
		S2	0.292 ± 0.022	0.53	0.469	0.955	51.4	0.267	69.9
	Butina-0.6	ID	0.901 ± 0.003	0.92	0.079	0.903	14.3	0.898	2.1
		S1	0.504 ± 0.030	0.68	0.319	0.960	39.5	0.468	43.4
		S2	0.200 ± 0.066	0.40	0.603	0.969	64.8	0.151	80.0
	Butina-0.7	ID	0.907 ± 0.008	0.93	0.072	0.904	15.4	0.905	2.1
		S1	0.436 ± 0.037	0.66	0.345	0.962	44.1	0.416	49.6
		S2	0.127 ± 0.020	0.37	0.625	0.964	68.4	0.109	82.3
SSI-DDI	Random	ID	0.899 ± 0.005	1.17	0.001	0.899	3.9	0.900	2.5
		S1	0.657 ± 0.014	1.20	0.003	0.832	4.7	0.653	27.4
		S2	0.443 ± 0.021	1.23	0.004	0.750	5.4	0.439	56.3
	Butina-0.6	ID	0.900 ± 0.004	1.13	0.001	0.897	4.0	0.899	2.6
		S1	0.606 ± 0.010	1.18	0.006	0.830	5.3	0.596	32.8
		S2	0.359 ± 0.021	1.20	0.010	0.740	6.4	0.348	66.9
	Butina-0.7	ID	0.903 ± 0.002	1.15	0.001	0.900	4.0	0.899	2.8
		S1	0.563 ± 0.046	1.22	0.004	0.834	5.5	0.553	37.4
		S2	0.332 ± 0.092	1.24	0.008	0.724	6.4	0.321	66.3
DSN-DDI	Random	ID	0.903 ± 0.004	1.01	0.013	0.903	6.6	0.903	2.6
		S1	0.640 ± 0.031	0.99	0.046	0.904	10.6	0.635	29.4
		S2	0.403 ± 0.044	0.96	0.080	0.891	14.5	0.397	61.1
	Butina-0.6	ID	0.896 ± 0.002	1.00	0.014	0.904	5.4	0.894	2.6
		S1	0.591 ± 0.027	0.96	0.061	0.902	9.2	0.578	37.6
		S2	0.356 ± 0.026	0.91	0.102	0.861	13.0	0.329	72.7
	Butina-0.7	ID	0.903 ± 0.007	1.00	0.014	0.901	6.1	0.902	2.9
		S1	0.525 ± 0.041	0.94	0.080	0.911	11.0	0.513	41.1
		S2	0.273 ± 0.049	0.88	0.137	0.860	14.8	0.257	72.1

Note: Results are averaged over three seeds. LAC coverage is reported as mean ± standard deviation. Size denotes the mean number of interaction types per prediction set out of 86. Empty denotes the fraction of empty LAC sets. Mondrian denotes class-conditional LAC, and CovGap is reported in percentage points.

Table 5: *Performance of the FP-MLP Five-Member Ensemble.*

Split	Test	Accuracy	ECE	LAC coverage	LAC size	Empty	APS coverage	APS size	CovGap
Random	ID	0.961 ± 0.004	0.007	0.905 ± 0.005	0.92	0.083	0.897	13.7	2.7
	S1	0.696 ± 0.016	0.170	0.566 ± 0.010	0.71	0.293	0.943	32.8	38.0
	S2	0.435 ± 0.024	0.337	0.291 ± 0.015	0.52	0.482	0.956	50.9	70.9
Butina-0.6	ID	0.959 ± 0.006	0.006	0.897 ± 0.004	0.91	0.091	0.901	12.2	2.5
	S1	0.641 ± 0.013	0.203	0.503 ± 0.024	0.66	0.338	0.959	37.1	44.2
	S2	0.376 ± 0.025	0.338	0.213 ± 0.042	0.42	0.580	0.966	63.7	79.6
Butina-0.7	ID	0.958 ± 0.002	0.007	0.905 ± 0.007	0.92	0.081	0.902	12.9	2.4
	S1	0.574 ± 0.043	0.241	0.429 ± 0.045	0.62	0.379	0.958	40.7	50.8
	S2	0.287 ± 0.042	0.405	0.144 ± 0.025	0.41	0.594	0.962	65.9	81.8

Note: Results are averaged over three seeds. Accuracy and LAC coverage are reported as mean ± standard deviation. The ensemble uses conformal thresholds fitted to averaged probabilities without temperature scaling, as described in Section 3.5. "Empty" denotes the fraction of empty LAC sets, and CovGap refers to class-conditional LAC.

structural novelty. Second, under Butina-0.7 the S1 trend flattened at the top: pooled over seeds, coverage in the highest bin was equal to or lower than in the 0.4-0.5 bin for all three models, probably because few held-out drugs fall in that bin and they are heterogeneous. Similarity is thus a strong but incomplete predictor of coverage degradation.

5.5. Selective Prediction

Figure 4 shows risk-coverage curves on the Butina-0.7 regime. In distribution, confidence ranked errors almost perfectly (AURC 0.006-0.043). On S1, ranking by confidence still helped: the error rate among the most confident pairs was well below the overall rate, and AURC was 0.286-0.355 against overall error rates of 0.44-0.50. On S2 the ranking carried much less information: even the most confident tenth of the pairs were wrong about half the time or more, and AURC (0.615-0.675) was close to the overall error rate of about 0.70-0.73. Abstaining on low-confidence predictions is therefore a reasonable strategy for S1, but not a reliable one for S2.

5.6. Scope, Implications and Limitations

Within the scope of this benchmark, the main message is that uncertainty estimates validated on known drugs did not transfer to new ones. A nominal 90% prediction set contained the correct interaction type for as few as 13% of pairs of two structurally novel drugs. Because the in-distribution results look excellent, this degradation stays invisible unless

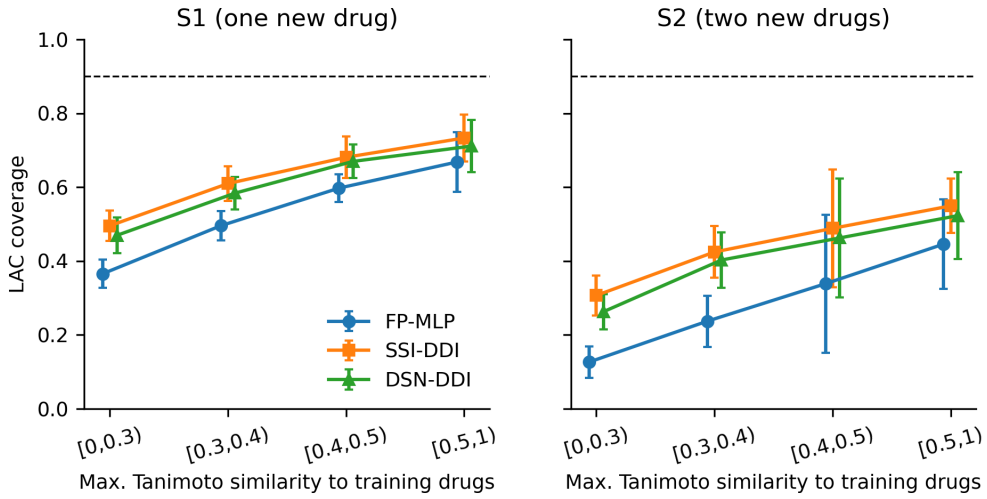


Figure 3: *LAC coverage as a function of the maximum Tanimoto similarity of the new drug to the training drugs (Eq. 1), shown separately for S1 pairs (left) and S2 pairs (right). Points are pair-weighted means over the three regimes and three seeds; error bars are standard deviations across splits (bins with at least 50 pairs). The dashed line marks the 90% target.*

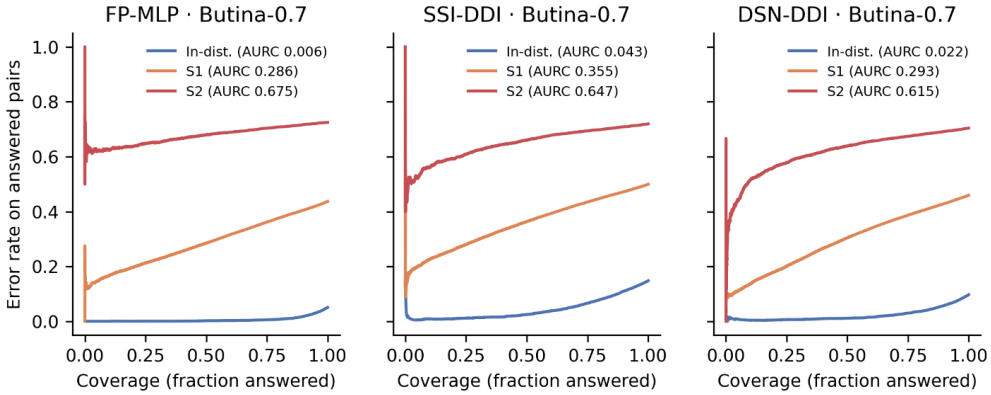


Figure 4: *Risk-coverage curves on the Butina-0.7 regime (temperature-scaled, pooled over three seeds). The error rate among answered pairs is plotted against the fraction of pairs answered, in order of decreasing confidence; the legend gives the AURC of each test set.*

models are evaluated on held-out drugs, ideally on structurally distant ones. We therefore suggest that DDI prediction studies report calibration and coverage on S1 and S2 alongside accuracy.

These experiments concern the prediction of DrugBank interaction types and the behavior of uncertainty estimates on that task. They do not evaluate clinical outcomes, patient-level risk or prescribing safety, and the models and thresholds studied here have not been validated for clinical use.

A model designed for unseen drugs did not solve the problem. DSN-DDI was more accurate than SSI-DDI on S1 and S2 in every regime, yet its calibration error and coverage loss were of similar size. Better representations improve accuracy, but they do not tell the model when it is extrapolating; the difficulty lies in calibrating on data that do not resemble the test drugs. The similarity analysis suggests a remedy: because coverage can be predicted from $\nu(d)$, thresholds could be calibrated within similarity strata or weighted by similarity, in the spirit of weighted conformal prediction [9], [20].

This study has several limitations. It uses a single dataset with 86 interaction types; other sources and binary or multi-label formulations may behave differently. The graph models were trained for at most 25 epochs and with a multi-class cross-entropy objective rather than negative sampling, which likely explains why their ID accuracy (0.85-0.90) was below that of FP-MLP; our conclusions concern relative degradation under novelty, which was consistent across all three models. Three seeds per regime leave some estimates variable (for example, SSI-DDI on Butina-0.7 S2 had a coverage of 0.332 ± 0.092). Novelty was measured by ECFP4 similarity alone, and shift-aware conformal methods were not evaluated.

6. Conclusion and Future Work

We introduced ColdCal, a benchmark for calibration and conformal coverage in DDI type prediction under graded drug novelty. For a fingerprint MLP, SSI-DDI and DSN-DDI, uncertainty estimates that were well calibrated on known drugs degraded substantially on new ones, where exchangeability is violated: empirical conformal coverage at a 90% target fell to 0.43-0.66 for pairs with one new drug and to 0.13-0.44 for pairs with two, and the degradation grew with structural novelty. Temperature scaling, deep ensembles and class-conditional calibration did not repair it, and adaptive sets did so only on S1 and only at a substantial cost in set size. Because coverage tracks similarity to the training drugs within both pair types, our next step is to test similarity-stratified and weighted conformal calibration on ColdCal, and to extend the benchmark to further datasets and longer-trained graph models. ColdCal is a benchmarking study of DrugBank interaction-type prediction and of uncertainty behavior on that task; it is not a validated clinical decision system, and its results should not be used to guide prescribing decisions.

Author Declarations

Author Contributions (CRediT): Bajpayee Singha: Conceptualization; Data Curation; Methodology; Software; Formal Analysis; Investigation; Validation; Visualization; Writing - Original Draft. Ajay Kumar Sinha: Supervision; Project Administration; Writing - Review & Editing. All authors have read and approved the submitted version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The ColdCal split files are openly available on Kaggle at https://www.kaggle.com/datasets/bajpayeesingha/coldcal-splits. The evaluation code and model training scripts are available in the Kaggle notebooks https://www.kaggle.com/code/bajpayeesingha/coldcal-03-ssiddi, https://www.kaggle.com/code/bajpayeesingha/coldcal-04-dsn and https://www.kaggle.com/code/bajpayeesingha/coldcal-05-paper, whose outputs contain the saved model outputs and result tables reported in this paper. The SSI-DDI model outputs are available at https://www.kaggle.com/datasets/bajpayeesingha/coldcal-results-ssiddi. The DrugBank-derived drug–drug interaction dataset is the version distributed with the official SSI-DDI implementation https://github.com/kanz76/SSI-DDI, which is subject to DrugBank's licensing terms. SSI-DDI and DSN-DDI were run from their official public repositories, https://github.com/kanz76/SSI-DDI and https://github.com/microsoft/Drug-Interaction-Research/tree/DSN-DDI-for-DDI-Prediction, respectively.

Acknowledgments: The authors thank Kaggle for providing free GPU computing resources used in the experiments.

Conflicts of Interest: The authors declare no conflict of interest.

Ethics Statement: Not applicable.

Declaration of Generative AI: During the preparation of this manuscript, the authors used Claude (Anthropic) for coding assistance, data analysis support, and drafting and revising the manuscript text. The authors reviewed and verified the resulting content and take full responsibility for the accuracy, integrity, and content of the manuscript.

References

[1] J. Y. Ryu, H. U. Kim, and S. Y. Lee, "Deep learning improves prediction of drug-drug and drug-food interactions," *Proceedings of the National Academy of Sciences*, vol. 115, no. 18, pp. E4304–E4311, 2018, doi: 10.1073/pnas.1803294115.

[2] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, 2018, doi: 10.1093/bioinformatics/bty294.

[3] A. K. Nyamabo, H. Yu, and J.-Y. Shi, "SSI-DDI: Substructure-substructure interactions for drug-drug interaction prediction," *Briefings in Bioinformatics*, vol. 22, no. 6, Art. no. bbab133, 2021, doi: 10.1093/bib/bbab133.

[4] Z. Li, S. Zhu, B. Shao, X. Zeng, T. Wang, and T.-Y. Liu, "DSN-DDI: An accurate and generalized framework for drug-drug interaction prediction by dual-view representation learning," *Briefings in Bioinformatics*, vol. 24, no. 1, Art. no. bbac597, 2023, doi: 10.1093/bib/bbac597.

[5] 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. no. 429, 2021, doi: 10.3390/ph14050429.

[6] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017, pp. 1321–1330.

[7] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World.* New York, NY, USA: Springer, 2005.

[8] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023, doi: 10.1561/2200000101.

[9] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," *The Annals of Statistics*, vol. 51, no. 2, pp. 816–845, 2023, doi: 10.1214/23-AOS2276.

[10] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, 2018.

[11] M. Nickel, V. Tresp, and H.-P. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011, pp. 809–816.

[12] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: A benchmark for molecular machine learning," *Chemical Science*, vol. 9, no. 2, pp. 513–530, 2018, doi: 10.1039/C7SC02664A.

[13] D. Butina, "Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets," *Journal of Chemical Information and Computer Sciences*, vol. 39, no. 4, pp. 747–750, 1999, doi: 10.1021/ci9803381.

[14] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010, doi: 10.1021/ci100050t.

[15] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, vol. 30, Long Beach, CA, USA, 2017.

[16] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems*, vol. 32, Vancouver, Canada, 2019.

[17] M. Sadinle, J. Lei, and L. Wasserman, "Least ambiguous set-valued classifiers with bounded error levels," *Journal of the American Statistical Association*, vol. 114, no. 525, pp. 223–234, 2019, doi: 10.1080/01621459.2017.1395341.

[18] Y. Romano, M. Sesia, and E. J. Candès, "Classification with valid and adaptive coverage," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[19] T. Ding, A. N. Angelopoulos, S. Bates, M. I. Jordan, and R. J. Tibshirani, "Class-conditional conformal prediction with many classes," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[20] R. J. Tibshirani, R. F. Barber, E. J. Candès, and A. Ramdas, "Conformal prediction under covariate shift," in *Advances in Neural Information Processing Systems*, vol. 32, Vancouver, Canada, 2019.

[21] 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, doi: 10.1093/nar/gkx1037.

[22] RDKit Contributors, "RDKit: Open-source cheminformatics software," RDKit. Accessed: Oct. 24, 2026. [Online]. Available: https://www.rdkit.org

[23] A. K. Nyamabo, "SSI-DDI: Official implementation," GitHub repository. Accessed: Oct. 24, 2026. [Online]. Available: https://github.com/kanz76/SSI-DDI

[24] Microsoft, "DSN-DDI: Official implementation (branch DSN-DDI-for-DDI-Prediction)," GitHub repository. Accessed: Oct. 24, 2026. [Online]. Available: https://github.com/microsoft/Drug-Interaction-Research/tree/DSN-DDI-for-DDI-Prediction

[25] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, Vancouver, Canada, 2019.

[26] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, New Orleans, LA, USA, 2019.

[27] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, Austin, TX, USA, 2015, pp. 2901–2907.

[28] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 30, Long Beach, CA, USA, 2017.