Conformal Prediction for Uncertainty Estimation in Drug-Target Interaction Prediction

Morteza Rakhshaninejad   Morteza.Rakhshaninejad@UGent.be
Mira Jürgens   Mira.Juergens@UGent.be
Nicolas Dewolf   Nicolas.Dewolf@UGent.be
Willem Waegeman   Willem.Waegeman@UGent.be
Department of Data Analysis and Mathematical Modeling, Ghent University, Belgium

Editor: Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

Abstract

Accurate drug–target interaction (DTI) prediction with machine learning models is essential for drug discovery. Such models should also provide a credible representation of their uncertainty, but applying classical marginal conformal prediction (CP) in DTI prediction often overlooks variability across drug and protein subgroups. In this work, we analyze three cluster-conditioned CP methods for DTI prediction, and compare them with marginal and group-conditioned CP. Clusterings are obtained via nonconformity scores, feature similarity, and nearest neighbors, respectively. Experiments on the KIBA dataset using four data-splitting strategies show that nonconformity-based clustering yields the tightest intervals and most reliable subgroup coverage, especially in random and fully unseen drug–protein splits. Group-conditioned CP works well when one entity is familiar, but residual-driven clustering provides robust uncertainty estimates even in sparse or novel scenarios. These results highlight the potential of cluster-based CP for improving DTI prediction under uncertainty.

Keywords: Conformal Prediction, Drug–target Interaction, Uncertainty Quantification, Cluster-conditioned Prediction, Prediction Intervals

1. Introduction

Drug–target interaction (DTI) prediction is a central task in computational drug discovery, enabling the identification of candidate compounds likely to bind to specific protein targets (Öztürk et al., 2018). Machine learning models have significantly improved the scalability and accuracy of DTI screening by using chemical and biological features to predict binding affinities. However, most existing approaches provide only point estimates of interaction strength, without quantifying the uncertainty of their predictions. This lack of uncertainty estimation can mislead downstream validation, undermining the reliability of experimental decisions. To address this, uncertainty quantification (UQ) methods have emerged as crucial tools, particularly those that offer predictive intervals for regression and prediction sets for classification alongside point estimates. Among these, conformal prediction (CP) provides a distribution-free framework for generating calibrated prediction intervals with finite-sample coverage guarantees (Vovk et al., 2005; Angelopoulos and Bates, 2021). In regression tasks, CP methods wrap around any predictive model to output prediction intervals that contain the true label with a user-specified probability $1 - \alpha$, under the assumption of data exchangeability. While marginal CP ensures coverage on average

across all test points, it can fail to capture heterogeneous error behavior across subpopulations, such as distinct drug or protein families. Specifically, marginal coverage satisfies $P(y \in \widehat{C}(\mathbf{x})) \geq 1 - \alpha$, where the probability is taken over the joint distribution of test inputs and labels. However, this guarantee does not imply that all subgroups receive adequate coverage. In contrast, conditional coverage aims to satisfy $P(y \in \widehat{C}(\mathbf{x}) \mid \mathbf{x}) \geq 1 - \alpha$ for every input $\mathbf{x}$—a much stronger criterion that is generally impossible to achieve without further assumptions. Therefore, to improve practical utility in the presence of data heterogeneity, we investigate methods that approximate conditional coverage by conditioning on meaningful subgroups.

In this study, we investigate the application of a cluster-conditioned conformal prediction (CCP) framework to provide more accurate and personalized uncertainty estimates in DTI regression. Building on recent advances in class-conditional CP (Ding et al., 2023), we propose two clustering-based variants—one based on nonconformity score distributions (CCP-NC) and the other on clustering input features (CCP-FC)—as well as a nearest-neighbor–based variant (CCP-NN). A key novelty of our approach is that we construct calibration sets by conditioning on both components of a drug–target interaction—rather than treating the interaction as a single atomic unit—allowing prediction intervals to more accurately reflect uncertainty arising from either side. We benchmark these CCP methods against standard Marginal CP (MCP) (Vovk et al., 2005) and Group-Conditioned CP (GCP) (Vovk et al., 2003), where drug and protein identities are used as group labels. Our experiments use the KIBA dataset (Tang et al., 2014), with a gradient boosting regressor and feature representations combining molecular fingerprints, chemical descriptors, amino acid composition, and transformer-based embeddings. To evaluate robustness and generalization, we implement four data-splitting strategies that progressively increase the dissimilarity between training and test sets. We introduce CCP framework for DTI regression, comprising three variants: CCP-NC, CCP-FC, and CCP-NN. These models are benchmarked against MCP and GCP across all four data-splitting strategies, including the challenging New Drug–Protein split. To improve coverage at the subgroup level, we propose a joint configuration strategy that tunes both the number of clusters and the calibration set ratio. Finally, we evaluate all methods using global coverage, prediction interval width, and subgroup-level coverage gap.

The remainder of the paper is organized as follows: Section 2 reviews prior work on CP and its applications in computational biology. Section 3 describes the dataset, predictive model, data-splitting strategies, and CP methodologies used in this study. Section 4 presents experimental results comparing different CP variants across various evaluation criteria. Section 5 concludes the paper and outlines directions for future work.

2. Related Work

CP estimates a threshold on conformity scores to construct prediction intervals or set-valued predictions, ensuring that the error rate does not exceed a user-specified level α in finite samples (Vovk et al., 2005). In the original transductive scheme, the underlying predictive model is re-trained for every candidate label of the test point (Saunders et al., 1999). The more practical split/inductive variant (Papadopoulos et al., 2002; Angelopoulos and Bates, 2021) instead trains a single predictive model on a proper-training set and reserves a dis-

joint calibration set to estimate the $(1 - \alpha)$-quantile of a nonconformity score. The choice of NCS—which quantifies how atypical an observation is—controls the efficiency of the resulting intervals and has spurred extensive research on score design, normalization, and learning (Kato et al., 2023; Bellotti, 2020). Split-conformal guarantees hold only marginally, i.e. on average over the population; particular subgroups may be under- or over-covered. Mondrian, or label-conditional, CP addresses this by calibrating within each predefined group to achieve exact conditional coverage (Vovk et al., 2003), but the per-group quantile becomes unstable when few calibration samples are available, leading to very wide intervals. Recent theoretical work also shows that, for heteroskedastic data, normalizing nonconformity scores by estimated noise levels can achieve approximate conditional validity without requiring explicit group splits (Dewolf et al., 2023).

Classical CP methods ensure marginal validity but do not guarantee conditional validity, a stricter and often more desirable property that tailors prediction regions to individual covariates. Prior work has shown that constructing non-trivial prediction sets with exact conditional validity from finite samples is impossible without making strong assumptions on the data distribution (Vovk, 2012; Lei and Wasserman, 2013; Foygel Barber et al., 2021). Consequently, most research focuses on achieving approximate conditional validity while retaining distribution-free guarantees (Gibbs et al., 2025). A common relaxation involves group-conditional guarantees (Ding et al., 2023; Jung et al., 2022), where coverage is ensured for a predefined set of groups. Another line of work addresses covariate shift, relaxing the exchangeability assumption, and imposes conditional guarantees with respect to a fixed family of shifts (Tibshirani et al., 2019). However, this is distinct from conditional conformal prediction under exchangeability, and may miss shifts tied to the predictive model's uncertainty. Efforts to broaden the covariate shift space (Gibbs et al., 2025) risk conservatism and unnecessarily wide intervals, as noted by Hore and Barber (2023). Alternative approaches, such as localised CP (Guan, 2023), weigh the residuals on the calibration set by a similarity function between test and calibration samples, although such methods can struggle in high-dimensional feature spaces. Recent clustered variants merge groups with similar score distributions before calibration—balancing validity and efficiency—and have proved effective in image recognition and property-valuation tasks (Ding et al., 2023; Hjort et al., 2024; Lim and Bellotti, 2021). Other work pursues softer relaxations such as local or normalized coverage (Angelopoulos and Bates, 2021). Together these advances shift the focus from global to equitable, subgroup-aware uncertainty quantification.

In computational biology, CP has also seen emerging applications. Boger et al. (2025) applied CP principles to protein structure and homology detection, enabling robust pre-filtering for structural alignment and achieving state-of-the-art performance in enzyme classification without retraining predictive models. In drug discovery, Laghuvarapu et al. (2023) proposed CP-based prediction sets for molecular properties under covariate shift, significantly reducing the coverage gap—by over 35%—compared to standard CP methods. Within DTI prediction, most studies focus on point estimates of binding affinity (Öztürk et al., 2018; Iliadis et al., 2024), while CP remains underexplored. A notable exception is Oršolić and Šmuc (2023), who proposed a local conformal approach using dynamically constructed calibration sets based on neighborhood similarity. However, their method does not leverage cluster-based conformal predictors, which our experiments show to be more effective in controlling the subgroup-level coverage gap. Thus, the application of CP to DTI

	Drugs	Targets	Morgan fingerprints	Mordred descriptors	Amino acid composition	ProtBERT embeddings	Binding affinity
Drug Target Interactions	d_1	t_1	1024 features	1505 features	20 features	1024 features	$y(d_1, t_1)$
	d_1	t_2	1024 features	1505 features	20 features	1024 features	$y(d_1, t_2)$
	d_n	t_n	1024 features	1505 features	20 features	1024 features	$y(d_n, t_n)$

Figure 1: Schematic of dataset structure: each interaction is a unique drug–target pair (d_i, t_j) with binding affinity $y(d_i, t_j)$.

remains largely unexplored, particularly in terms of generating calibrated prediction intervals with conditional coverage guarantees at the level of individual drugs and proteins—a gap this work aims to address.

3. Materials and Methods

3.1. Dataset and Feature Representations

We constructed our dataset by integrating multiple feature modalities for drugs and protein targets to capture the chemical and biological variability of kinase-inhibitor interactions. Specifically, SMILES strings (Weininger, 1988) were used to represent small-molecule inhibitors d_i, and primary amino acid sequences were used to represent protein targets t_j. Each drug-target interaction is defined as a unique pair (d_i, t_j). Binding affinities (K_d) were transformed into dimensionless values using $y(d_i, t_j) = -\log_{10}(K_d/10^9)$, where $y(d_i, t_j)$ denotes the transformed affinity between drug d_i and target t_j. We evaluated our models exclusively on the KIBA benchmark dataset (Tang et al., 2014), which comprises 2068 small-molecule inhibitors, 229 protein kinase targets, and 118,036 recorded drug–target interactions.

Each drug was represented using Morgan fingerprints with 1024 features, which encode molecular substructures using circular fingerprints with a radius of 2 (Rogers and Hahn, 2010), and Mordred descriptors, comprising 1505 features that capture a comprehensive set of physicochemical, topological, and constitutional molecular properties (Moriwaki et al., 2018). Protein targets were encoded using Amino Acid Composition, a 20-feature representation that quantifies the frequency of each amino acid within the sequence, and ProtBERT embeddings, which consist of 1024-dimensional transformer-based representations trained on protein sequences (Elnaggar et al., 2021). Motivated by their complementary strengths, we combined four feature types—Morgan fingerprints, Mordred descriptors, Amino Acid Composition, and ProtBERT embeddings—to capture diverse chemical and biological properties at both local and global levels. The resulting drug–target representations were concatenated into a feature matrix $X \in \mathbb{R}^{N \times 3573}$, where N is the number of interactions. Figure 1 illustrates the structured representation of our dataset, showing the feature modalities used for each pair of interactions.

3.2. Predictive Model

To estimate the binding affinity of the drug-target, we used the gradient boosting regressor (Friedman, 2001), a widely used ensemble learning technique recognized for its efficiency, scalability, and ability to model complex non-linear relationships in tabular data. Its integrated feature selection and regularization mechanisms make it particularly suitable for high-dimensional tasks such as DTI prediction. The trained model maps each input feature vector $\mathbf{x}(d_i, t_j) \in \mathbb{R}^{N \times 3573}$, representing a drug–target pair, to a continuous output $\hat{y}(d_i, t_j) = f_\theta(\mathbf{x}(d_i, t_j))$, where f_θ denotes the gradient boosting regressor model and θ represents the learned parameters. We configured the model with squared error as the loss function, 500 boosting stages, a learning rate of 0.05, and a maximum tree depth of 6. The implementation was carried out using the `GradientBoostingRegressor` class from the `scikit-learn` Python library (version 1.3.0).

3.3. Dataset Splitting Strategies

To evaluate the robustness and generalization capability of our model, we employed four dataset splitting strategies. Each strategy ensures that the model is evaluated under different distributional constraints, ranging from favorable overlap to strict disjointness. In every case, the dataset is partitioned into three disjoint subsets: a training set ($\mathcal{D}_{\text{train}}$) for model fitting, a calibration set ($\mathcal{D}_{\text{cal}}$) for CP, and a test set ($\mathcal{D}_{\text{test}}$) for final evaluation. All strategies maintain a 50%–25%–25% split across training, calibration, and test data. The four splitting strategies are defined as follows. In the Random Split, interaction-level random splitting is performed: 50% of all interaction samples are assigned to the training set, and the remaining 50% are then split equally into calibration and test sets. Since calibration and test samples are drawn from the same pool, they are exchangeable by construction. This split imposes the least restriction and assumes that drugs and proteins are evenly distributed across all subsets. The Drug Split and Protein Split strategies follow the same mechanism. In the Drug Split, 50% of unique drugs are randomly selected, and all their corresponding interactions are assigned to the training set. The remaining 50% of drugs are randomly divided into two equal groups to form the calibration and test sets. Similarly, in the Protein Split, 50% of unique proteins are used for training, and the remaining proteins are split equally into calibration and test sets. In both cases, interactions are assigned based on entity membership, ensuring that no drug or protein appears in more than one subset, while the other side of the interaction (proteins or drugs) may be shared. Calibration and test sets are constructed from a common pool of interactions involving unseen entities, preserving exchangeability. Finally, the New Drug–Protein Split represents the most challenging scenario. Both the drugs and proteins in the calibration and test sets are completely unseen during training. Specifically, 50% of drugs and 50% of proteins are randomly selected to form the training set, which includes only interactions where both entities belong to the training pool. The remaining interactions—comprising combinations of unseen drugs and proteins—form the held-out pool, which is then equally divided into calibration and test sets. This split ensures exchangeability but may result in a reduced number of available samples due to its stringent filtering criteria.

A visual illustration of the dataset splitting strategies is shown in Figure 2. The left panel represents the drug–target interaction matrix with binding affinity values. The center panel
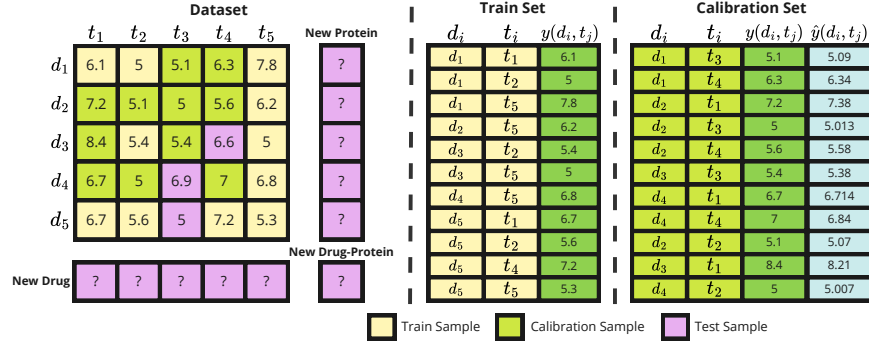
Figure 2: Illustrative example of dataset preparation for model training and conformal prediction. The drug–target interaction matrix shows samples used for training, calibration, and testing.

highlights the subset of interactions used to train the predictive model, while the right panel shows the calibration set samples, whose affinities are predicted using the trained model and used for conformal quantile estimation. Each split strategy defines how these sets are constructed and reflects a different level of generalization difficulty, from the permissive Random Split to the stringent New Drug–Protein Split.

3.4. Problem statement

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ be a labeled dataset of N samples, where each $\mathbf{x}_i \in \mathbb{R}^{N \times 3573}$ is a feature vector representing an interaction between drug d_i and target t_j, and $y_i \in \mathbb{R}$ is the corresponding transformed binding affinity label. We consider a point prediction function $f_\theta : \mathbb{R}^{N \times 3573} \to \mathbb{R}$, which in our case is realized by the `GradientBoostingRegressor` model described in Section 3.2. Our goal is to construct a predictive interval $\widehat{C}(\mathbf{x})$ for a new test input $\mathbf{x}$, such that the probability of $\widehat{C}(\mathbf{x})$ containing the true value y is approximately $1 - \alpha$. We typically partition the dataset of N samples into three disjoint subsets $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{cal}}, \mathcal{D}_{\text{test}}$ (as we discussed in section Section 3.3) of respective sizes $n_{\text{train}}, n_{\text{cal}}, n_{\text{test}}$, with $n_{\text{train}} + n_{\text{cal}} + n_{\text{test}} = N$. Denoting the samples by $(\mathbf{x}_i, y_i)$, we train the model f_θ on $\mathcal{D}_{\text{train}}$. The calibration set $\mathcal{D}_{\text{cal}}$ is used to compute nonconformity scores and estimate the relevant quantile, which determines how wide the prediction interval must be to ensure the desired coverage level $1 - \alpha$. Finally, $\mathcal{D}_{\text{test}}$ is reserved for evaluation.

A key component of CP is the choice of a nonconformity (or conformity) measure, which quantifies how unusual an observed label y_i is relative to the model's prediction $f_\theta(\mathbf{x}_i)$. A common choice in regression is the absolute residual (Papadopoulos et al., 2011):

$$s_i = |y_i - f_\theta(\mathbf{x}_i)|. \tag{1}$$

However, if the data is heteroscedastic (i.e., the noise level varies across the input space), the efficiency of prediction intervals may degrade. In such cases, a normalized nonconformity

score can be used (Papadopoulos et al., 2002):

$$s_i = \frac{|y_i - f_\theta(\mathbf{x}_i)|}{\sigma_\theta(\mathbf{x}_i)}, \tag{2}$$

where $\sigma_\theta(\mathbf{x}_i)$ is an estimate of the local uncertainty for $\mathbf{x}_i$. In our case, since we applied a Box–Cox transformation (Box and Cox, 1964) to normalize the input binding affinity values, we assume that the transformed data exhibit homoscedastic behavior. Consequently, we adopt the absolute residual form in Equation (1) as our nonconformity score throughout this study. After computing the nonconformity scores s_i on the calibration set, we sort them in non-decreasing order: $s_{(1)} \leq s_{(2)} \leq \cdots \leq s_{(n_{\text{cal}})}$, and select the conformal quantile threshold:

$$\hat{q} \;=\; s_{(\lceil (1-\alpha)\,(n_{\text{cal}}+1)\rceil)}, \tag{3}$$

which corresponds to the $(1-\alpha)$-quantile of the empirical distribution of the nonconformity scores s_i. The construction of the prediction interval $\widehat{C}(\mathbf{x})$ depends on the choice of the nonconformity function. If the absolute residual is used as the nonconformity score (see Equation (1)), the prediction interval is defined as

$$\widehat{C}(\mathbf{x}) \;=\; [\, f_\theta(\mathbf{x}) - \hat{q}, \;\; f_\theta(\mathbf{x}) + \hat{q}\,]. \tag{4}$$

Alternatively, if the normalized residual is used (see Equation (2)), incorporating an estimate of local uncertainty $\sigma_\theta(\mathbf{x})$, the prediction interval becomes

$$\widehat{C}(\mathbf{x}) \;=\; [\, f_\theta(\mathbf{x}) - \hat{q} \cdot \sigma_\theta(\mathbf{x}), \;\; f_\theta(\mathbf{x}) + \hat{q} \cdot \sigma_\theta(\mathbf{x})\,]. \tag{5}$$

Under appropriate conditions—most notably, exchangeability between the calibration and test samples—this interval satisfies the approximate coverage guarantee $P\big(y \in \widehat{C}(\mathbf{x})\big) \approx 1 - \alpha$. In what follows, we describe CP approaches. These methods differ in how the calibration set is partitioned (if at all) and whether the coverage guarantee is interpreted globally or conditionally (e.g., within a group or cluster).

3.4.1. MARGINAL CONFORMAL PREDICTION

Marginal Conformal Prediction (MCP) does not condition on any subgroup or feature partition (Papadopoulos et al., 2002). It uses a single global calibration set to compute prediction intervals. This approach represents the simplest form of CP, as it assumes no additional structure in the data beyond the exchangeability assumption. We partition the dataset $\mathcal{D}$ into $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{cal}}$, and $\mathcal{D}_{\text{test}}$, and train the regression model f_θ on $\mathcal{D}_{\text{train}}$ to predict binding affinity values $y(d_i, t_j)$ from the concatenated feature vector $\mathbf{x}(d_i, t_j)$. This training process is applied separately for each dataset splitting strategy described in Section 3.3. Next, we compute the nonconformity scores for each calibration sample $(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{cal}}$, using the absolute residual defined in Equation (1). The scores are then sorted in non-decreasing order, and the conformal threshold $\hat{q}$ is selected as the $(1 - \alpha)$-quantile.

For any test sample $(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{test}}$, where $\mathbf{x}_i \in \mathcal{X}$, the model produces a point prediction $\hat{y}_i = f_\theta(\mathbf{x}_i)$ and constructs a prediction interval, as defined in Equation (4). Under the exchangeability assumption, this procedure guarantees that the true label y_i will fall within the predictive interval $\widehat{C}(\mathbf{x}_i)$ with probability at least $1 - \alpha$, i.e., $P(y \in \widehat{C}(\mathbf{x})) \geq 1 - \alpha$. This

approach provides marginal coverage across all test samples without conditioning on groups or structure. While simple and broadly applicable, it produces uniformly wide intervals (unless scaled by input variance), making it insensitive to underlying data heterogeneity.

3.4.2. Group-Conditioned Conformal Prediction

Group-Conditioned Conformal Prediction (GCP) is based on the concept of Mondrian CP (Vovk et al., 2005), which extends the standard marginal approach by conditioning on known group structures in the data. The goal is to obtain valid coverage within each group, rather than across all test instances. Let $g(\mathbf{x})$ denote a group assignment function that maps each input $\mathbf{x} \in \mathcal{X}$ to one of G distinct groups, indexed by $g \in \{1, \ldots, G\}$. After training the model f_θ on $\mathcal{D}_{\text{train}}$, we calculate nonconformity scores separately within each group of the calibration set. In contrast to MCP, which applies a single global threshold, GCP divides the calibration samples into finer-grained categories. In our DTI application, natural groupings emerge on the basis of drug and protein identities. Specifically, we consider each drug and each protein as defining a Mondrian category, allowing the prediction intervals to adapt to the uncertainty associated with individual compounds or targets. This group-based structure is especially important in data-splitting strategies such as drug split, protein split, or new drug–protein split (see Section 3.3), where the challenge of generalization varies according to which molecular entities are held.

For each calibration sample $(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{cal}}$, we compute a nonconformity score s_i. These scores are grouped using a group function $g(\cdot)$, which can correspond to the drug identity, protein identity, or both. For each group g, we collect the corresponding nonconformity scores and sort them: $s_{(1)}^g \leq s_{(2)}^g \leq \cdots \leq s_{(n_g)}^g$, where n_g is the number of calibration samples assigned to group g. The group-specific quantile is then computed as $\hat{q}_g = s_{(\lceil (1-\alpha)(n_g+1) \rceil)}^g$. At test time, for a given interaction $(d_{\text{test}}, t_{\text{test}})$, we determine whether the test drug or test protein is represented in the calibration set. The quantile threshold $\hat{q}_{g^*}$ is selected based on the availability of the test drug and protein in the calibration set. If both the test drug and test protein appear in the calibration set, their corresponding groups are merged, and a combined quantile is computed from the union of their nonconformity scores. If only the test drug is present, the quantile threshold is taken from its respective group. Similarly, if only the test protein appears, the quantile is based on its group. If neither the drug nor the protein is observed in the calibration set, a global quantile computed from all calibration scores is used instead. This logic is formalized as:

$$
\hat{q}_{g^*} = \begin{cases}
\hat{q}(\{s_i : d_i = d_{\text{test}} \vee t_i = t_{\text{test}}\}, 1-\alpha), & \text{if } d_{\text{test}}, t_{\text{test}} \in \mathcal{D}_{\text{cal}}, \\
\hat{q}(\{s_i : d_i = d_{\text{test}}\}, 1-\alpha), & \text{if } d_{\text{test}} \in \mathcal{D}_{\text{cal}}, \ t_{\text{test}} \notin \mathcal{D}_{\text{cal}}, \\
\hat{q}(\{s_i : t_i = t_{\text{test}}\}, 1-\alpha), & \text{if } t_{\text{test}} \in \mathcal{D}_{\text{cal}}, \ d_{\text{test}} \notin \mathcal{D}_{\text{cal}}, \\
\hat{q}_{\text{global}}, & \text{otherwise.}
\end{cases}
\tag{6}
$$

The final prediction interval for test input $\mathbf{x}$ is given by $\widehat{C}(\mathbf{x}) = [\, f_\theta(\mathbf{x}) - \hat{q}_{g^*}, \ f_\theta(\mathbf{x}) + \hat{q}_{g^*} \,]$. Assuming exchangeability holds within each group, this procedure yields the following group-conditional coverage guarantee (Vovk et al., 2003): $P\left(y \in \widehat{C}(\mathbf{x}) \mid g(\mathbf{x}) = g\right) \geq 1 - \alpha$, for all $g \in \{1, \ldots, G\}$.

Compared to MCP, this method provides more fine-grained, context-aware intervals that adapt to variability between groups. However, its effectiveness depends on having a sufficient number of calibration samples in each group to ensure stable quantile estimation (Ding et al., 2023). In the DTI setting, this is especially important when generalizing to novel drugs or targets not seen in training, as group-conditioned thresholds allow the conformal predictor to keep coverage within these molecular categories.

3.4.3. CLUSTER-CONDITIONED CONFORMAL PREDICTION: NONCONFORMITY-BASED CLUSTERING

In some applications, explicit grouping variables may be unavailable, or the number of samples within each group may be too small to reliably estimate group-specific quantiles. In such cases, we can partition the calibration set in an unsupervised manner using clustering, which allows us to identify latent subpopulations that may exhibit different noise or uncertainty characteristics. This gives rise to CCP, which leverages data-driven clusters to construct calibration sets and potentially improve local coverage properties. In the Nonconformity-Based Clustering variant (CCP-NC), clustering is performed directly on the nonconformity scores of the calibration samples (Ding et al., 2023). After training the predictive model f_θ on $\mathcal{D}_{\text{train}}$, we compute nonconformity scores for each calibration sample $(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{cal}}$. These scores are then used as input to a clustering algorithm, k-means, which assigns each sample to one of K clusters via a mapping $\kappa : \mathcal{X} \to \{1, \ldots, K\}$. To enable adaptive thresholding, the calibration set $\mathcal{D}_{\text{cal}}$ is partitioned into two disjoint subsets (Hjort et al., 2024). The first subset is used to perform clustering, grouping drugs and proteins based on the distribution of their nonconformity scores. The second subset is then used to compute cluster-specific quantiles, which are used for constructing the final prediction intervals. The splitting ratio is controlled by a parameter $\gamma \in \{0.25, 0.5, 0.75\}$, with higher γ values allocating more data to clustering and lower values favoring quantile estimation.

We then perform two distinct clustering operations over the residuals in the calibration set. First, for each drug d_i, we construct an empirical cumulative distribution function (ECDF) over its nonconformity scores and extract the 10th to 90th percentiles to form a fixed-length embedding vector. Using these embeddings, we apply k-means clustering with Euclidean distance to assign each drug to a unique drug cluster. Similarly, for each protein t_j, we extract percentile-based ECDF embeddings from its nonconformity scores and apply a second k-means clustering to assign each protein to a unique protein cluster. As a result, each calibration sample (d_i, t_j) inherits a pair of cluster memberships $(\kappa_{\text{drug}}(d_i), \kappa_{\text{protein}}(t_j))$. This process is illustrated in Figure 3, where nonconformity score matrices are transformed into percentile-based feature vectors for clustering, and the resulting cluster indices are appended to the calibration set. The number of clusters n_{clusters} and the calibration set splitting ratio γ are selected jointly to optimize subgroup coverage gaps. As introduced in Section 3.5, the Mean Absolute Coverage Gap (MACG) quantifies the deviation from the target coverage level for drugs and proteins. To identify the best configuration, we evaluate all combinations of $\gamma \in \{0.25, 0.5, 0.75\}$ and $n_{\text{clusters}} \in \{1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$, resulting in 33 possible settings. The optimal pair $(\gamma, n_{\text{clusters}})$ is selected by minimizing the
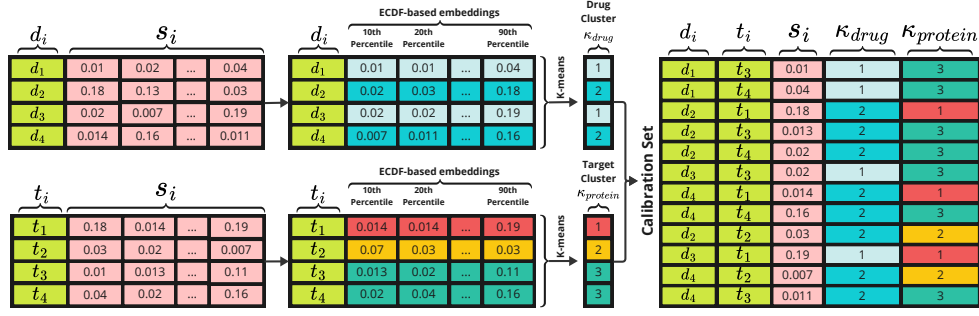
Figure 3: Two-level clustering in CCP-NC. **Left:** Nonconformity scores grouped by drug and protein. **Middle:** ECDF embeddings (10th–90th percentiles) extracted. **Right:** k-means assigns cluster indices κ_{drug}, $\kappa_{\mathrm{protein}}$ for cluster-specific quantile computation.

average MACG across drug and protein subgroups:

$$\min_{\gamma\in\{0.25,0.5,0.75\},\, n_{\mathrm{clusters}}\in[1,50]} \left(\frac{\mathrm{MACG}_{\mathrm{drug}}(\gamma, n_{\mathrm{clusters}}) + \mathrm{MACG}_{\mathrm{protein}}(\gamma, n_{\mathrm{clusters}})}{2} \right) \quad (7)$$

This strategy enables an optimal trade-off between reliability and subgroup-level validity, yielding accurate prediction intervals across heterogeneous biological subgroups.

Using the second portion of the calibration set, we compute cluster-specific nonconformity thresholds for each drug and protein cluster. For a test interaction (d_i, t_j), we denote the cluster assignments by $\kappa^*(d_i)$ and $\kappa^*(t_j)$, respectively. The corresponding quantile thresholds are computed from the nonconformity scores s_k of calibration samples as follows:

$$\hat{q}_{\kappa^*(d_i)} = \hat{q}(\{s_k : \kappa_{\mathrm{drug}}(d_k) = \kappa^*(d_i)\}, 1 - \alpha), \quad (8)$$

$$\hat{q}_{\kappa^*(t_j)} = \hat{q}(\{s_k : \kappa_{\mathrm{protein}}(t_k) = \kappa^*(t_j)\}, 1 - \alpha). \quad (9)$$

If a drug or protein in the test set belongs to a cluster not represented in the quantile estimation subset, we default to the global threshold $\hat{q}_{\mathrm{global}}$. The final quantile threshold $\hat{q}_{\mathrm{c}}$ is selected based on the availability of cluster-level information:

$$\hat{q}_{\mathrm{c}} = \begin{cases} \hat{q}(\{s_k : \kappa_{\mathrm{drug}}(d_k) = \kappa^*(d_i) \vee \kappa_{\mathrm{protein}}(t_k) = \kappa^*(t_j)\}, 1 - \alpha), & \text{if } d_{\mathrm{i}} \text{ and } t_{\mathrm{j}} \in \mathcal{D}_{\mathrm{cal}}, \\ \hat{q}_{\kappa^*(d_i)}, & \text{if only } d_{\mathrm{i}} \in \mathcal{D}_{\mathrm{cal}}, \\ \hat{q}_{\kappa^*(t_j)}, & \text{if only } t_{\mathrm{j}} \in \mathcal{D}_{\mathrm{cal}}, \\ \hat{q}_{\mathrm{global}}, & \text{otherwise.} \end{cases} \quad (10)$$

This logic ensures that the prediction interval for a test input $\mathbf{x} = (d_i, t_j)$ is informed by the most relevant cluster-level statistics available. If both clusters are represented in the calibration data, we pool their nonconformity scores to compute a combined threshold.

Otherwise, we rely on the available side, or default to the global threshold when necessary. The prediction interval is then formed as:

$$\widehat{C}(\mathbf{x}) = [f_\theta(\mathbf{x}) - \hat{q}_{\text{c}}, \ f_\theta(\mathbf{x}) + \hat{q}_{\text{c}}].$$ (11)

Provided the nonconformity scores within each cluster are approximately exchangeable, CCP aims to ensure valid coverage at the cluster level, i.e., $P(y \in \widehat{C}(\mathbf{x}) \mid \kappa(\mathbf{x}) = k) \gtrsim 1 - \alpha$ for all $k \in \{1, \ldots, K\}$.

By incorporating structural information derived from drug and protein nonconformity scores patterns, this method offers more personalized and locally calibrated prediction intervals. The use of ECDF-based clustering allows to better capture heteroscedasticity and latent subpopulation effects, especially under challenging generalization scenarios such as drug and protein disjoint splits. Overall, unlike Ding et al. (2023), (i) we build separate clusters for drugs and for proteins rather than clustering labels alone, and (ii) the calibration set for a test pair is the intersection of its drug-cluster and protein-cluster, so the resulting quantile adapts to uncertainty contributed by either component.

3.4.4. CLUSTER-CONDITIONED CONFORMAL PREDICTION: FEATURE-BASED CLUSTERING

In contrast to Nonconformity-Based Clustering, this variant, called CCP with Feature-Based Clustering (CCP-FC), groups calibration samples according to their original input feature representations. The goal is to leverage biologically or chemically meaningful similarities embedded in the feature space to enable more tailored uncertainty quantification. The feature vectors are the same as those described in Section 3.1, encompassing molecular fingerprints and descriptors for drugs, and sequence-based or embedding features for proteins. Let $\mathbf{x}_i$ denote the feature vector for the drug–target interaction (d_i, t_j). We define a clustering function $\kappa_{\text{feat}} : \mathcal{X} \to \{1, \ldots, K\}$ that assigns each calibration sample to one of K clusters based on similarity in feature space, using k-means clustering. Following the same protocol as in Nonconformity-Based Clustering, we split the calibration set $\mathcal{D}_{\text{cal}}$ into two disjoint subsets: one used for clustering based on input feature vectors, and the other for computing cluster-level quantiles. The split ratio is determined by a hyperparameter $\gamma \in \{0.25, 0.5, 0.75\}$, which balances clustering fidelity and quantile estimation reliability. Feature-based clustering is applied separately for drugs and proteins, as illustrated in Figure 4: each drug d_i is represented by a concatenation of Morgan fingerprints and Mordred descriptors, then assigned to a cluster $\kappa_{\text{drug}}(d_i)$ via k-means; each protein t_j is encoded using amino acid composition and ProtBERT embeddings, then assigned to a cluster $\kappa_{\text{protein}}(t_j)$ through k-means. Each calibration interaction (d_i, t_j) is thus mapped to a cluster pair $(\kappa_{\text{drug}}(d_i), \kappa_{\text{protein}}(t_j))$. These indices are then used for quantile estimation. Cluster-specific quantiles are computed using Equations (8) and (9), the combined test-time threshold $\hat{q}_{\text{c}}$ is determined by Equation (10), and the final prediction interval is constructed using Equation (11). To determine the best number of clusters n_{clusters} and the optimal γ, we follow the same procedure as described in Section 3.4.3.
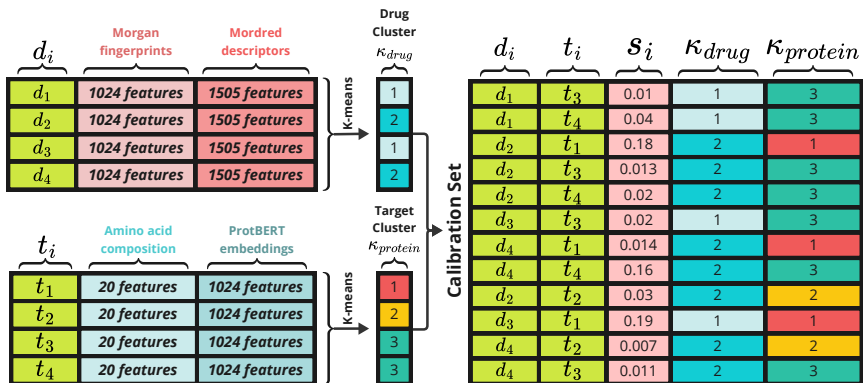
Figure 4: Two-level feature-based clustering in CCP-FC. Drugs d_i and proteins t_j are embedded using molecular and sequence features, respectively, and clustered via k-means to assign κ_{drug} and $\kappa_{\mathrm{protein}}$, used for grouping calibration scores and estimating cluster-specific quantiles.

3.4.5. CLUSTER-CONDITIONED CONFORMAL PREDICTION: NEAREST NEIGHBOR-BASED CALIBRATION

This variant, referred to as CCP with Nearest Neighbor-Based Calibration (CCP-NN), eliminates the use of fixed global or cluster-based partitions by dynamically constructing a local calibration set for each test sample based on feature similarity (Oršolić and Šmuc, 2023). The main idea is to tailor the quantile estimation to the most relevant neighborhood of each test instance, allowing for finer-grained adaptation to local uncertainty patterns. Let $\mathbf{x}_i$ denote the feature vector corresponding to a test drug–target pair (d_i, t_j). We define a dynamic local calibration set $\mathcal{D}_{\mathrm{cal}}(\mathbf{x}) \subset \mathcal{D}_{\mathrm{cal}}$, constructed by identifying calibration interactions involving compounds and targets most similar to d_i and t_j, respectively. Specifically, we compute the Tanimoto similarity (Rogers and Hahn, 2010) between the test compound d_i and all compounds in the calibration set based on their Morgan fingerprints, as well as between the test protein t_j and all calibration proteins using binary feature profiles (e.g., amino acid composition or domain annotations). We then select the top 20 most similar neighbors for both d_i and t_j. Finally, we extract all interaction pairs (d_k, t_l) from $\mathcal{D}_{\mathrm{cal}}$ such that d_k belongs to the top 20 similar compounds and t_l belongs to the top 20 similar targets, yielding a neighborhood-specific subset of calibration interactions:
$$\mathcal{D}_{\mathrm{cal}}(\mathbf{x}) = \left\{ (\mathbf{x}_k, y_k) \in \mathcal{D}_{\mathrm{cal}} \;\middle|\; d_k \in \mathrm{NN}_{20}^{\mathrm{compound}}(d_i), \; t_k \in \mathrm{NN}_{20}^{\mathrm{target}}(t_j) \right\}.$$

From this dynamically constructed neighborhood $\mathcal{D}_{\mathrm{cal}}(\mathbf{x})$, we collect the associated non-conformity scores s_k for each calibration sample interaction and sort them in non-decreasing order. The conformal quantile $\hat{q}$ is then computed according to Equation (3), using the number of scores in the local set. The resulting prediction interval for $\mathbf{x}$ is formed via Equation (11). This method offers highly adaptive and personalized calibration, particularly useful in heterogeneous or sparse interaction domains where global exchangeability assumptions are weak. By relying on similarity-based neighborhoods derived from domain-

relevant metrics, it enables localized uncertainty quantification while still preserving conformal validity assumptions within the dynamic calibration subsets.

3.5. Evaluation Metrics

CP provides a principled approach for constructing calibrated prediction intervals with finite-sample validity guarantees. To assess the performance of CP methods across different splitting and calibration set constructing strategies, we report several evaluation metrics that quantify both calibration (validity) and informativeness (efficiency). The most fundamental validity metric is the empirical coverage probability, which measures the proportion of test samples whose true binding affinities fall within the predicted intervals:

$$\text{Coverage}_{(1-\alpha)} = \frac{1}{n_{\text{test}}} \sum_{k=1}^{n_{\text{test}}} \mathbb{I}\big(y_k \in \widehat{C}_{(1-\alpha)}(\mathbf{x}_k)\big), \tag{12}$$

where $\widehat{C}_{(1-\alpha)}(\mathbf{x}_k)$ denotes the prediction interval for sample k at coverage level $1 - \alpha$, and $\mathbb{I}\{\cdot\}$ is the indicator function. This metric reflects how well the empirical coverage aligns with the nominal target level. To assess efficiency, we compute the average width of the prediction intervals:

$$\text{Width}_{\text{avg}} = \frac{1}{n_{\text{test}}} \sum_{k=1}^{n_{\text{test}}} (\text{upper}_k - \text{lower}_k), \tag{13}$$

where upper_k and lower_k are the endpoints of the interval for test instance k. Narrower intervals imply more confident and informative predictions, provided that validity is maintained. To evaluate fairness and subgroup-wise calibration, we use the Mean Absolute Coverage Gap (MACG), which quantifies the deviation from the desired coverage within specific groups:

$$\text{MACG}_{\text{subgroup}} = \frac{1}{D} \sum_{i=1}^{D} \left|\text{Coverage}_i - (1 - \alpha)\right|, \tag{14}$$

where D is the number of unique drugs, proteins, or clusters (depending on the subgroup type), and Coverage_i is the empirical coverage for subgroup i, computed as:

$$\text{Coverage}_i = \frac{1}{|\mathcal{I}_i|} \sum_{k \in \mathcal{I}_i} \mathbb{I}\big(y_k \in \widehat{C}(\mathbf{x}_k)\big), \tag{15}$$

with $\mathcal{I}_i$ being the index set for subgroup i (e.g., all test samples involving drug i). We report the following metrics: MACG per Drug, which measures calibration gaps for individual compounds; MACG per Protein, which measures calibration gaps across different target proteins; and MACG per Cluster (applicable in Cluster-Conditioned Prediction), which evaluates how well CCP achieves localized calibration across learned drug and protein clusters. In addition to numeric metrics, we also use reliability diagrams and coverage versus coverage plots to visualize empirical coverage across varying coverage levels $(1 - \alpha)$, providing further insight into the calibration quality of conformal predictors.

Table 1: Regression performance on the dataset using different data splitting strategies.

Splitting Strategy	RMSE	R^2
Random Split	0.0207	0.5729
Drug-based Split	0.0225	0.3821
Protein-based Split	0.0256	0.4185
New Drug–Protein Split	0.0266	0.1465

4. Experiments and Results

This section presents results of evaluation of the CP methods introduced in Section 3.4. We evaluated each method under the four dataset splitting strategies introduced in Section 3.3. All CP models were built upon the same predictive model architecture described in Section 3.2, but were trained separately using the training sets obtained from each splitting strategy. This ensured that the corresponding test sets remained completely unseen during model training. Below, we report the results of regression using the trained models on their respective test sets. Performance is measured using the Root Mean Squared Error (RMSE) and the coefficient of determination (R^2) for each splitting strategy. As shown in Table 1, the predictive model achieves the best performance under the Random Split, while generalization becomes progressively more challenging in the Drug, Protein, and especially the New Drug–Protein Split, as reflected by increasing RMSE and decreasing R^2 values.

We begin by assessing global coverage performance, as defined in Equation (12). Figure 5 compares observed coverage against the expected coverage level for each CP method across the dataset splitting strategies. The observed deviations from the ideal coverage line may not necessarily indicate violations of the exchangeability assumption. Instead, they could partly be attributed to the use of a fixed test set shared across all CP methods. This shared test set might introduce consistent coverage deviations within each split, causing multiple methods to exhibit similar patterns of overcoverage or undercoverage. In the Random Split, both CCP-FC and CCP-NC closely follow the ideal line, while MCP deviates slightly at the 90% coverage. GCP and CCP-NN are unstable with observed coverages exceeding the nominal levels. For the Drug Split, none of the methods achieve consistent coverage, but CCP-FC and CCP-NC are closest to the 95% target. All methods are overconfident at 90% and 95%. In the Protein Split, CCP-NC achieves reliable coverage at both 90% and 95%. MCP and CCP-NN perform well at 95%, whereas GCP and CCP-FC fall short. Under the New Drug–Protein Split, all methods yield reasonably accurate coverage, though GCP and CCP-NN show the weakest performance in this challenging scenario.

We next assess the informativeness of CP methods by comparing their mean prediction interval widths (Figure 6). Narrower intervals indicate more confident and precise predictions, while wider intervals reflect higher uncertainty needed to maintain coverage guarantees. CCP-NC produces the tightest intervals in both the Random and New Drug–Protein Splits at 90% and 95% coverage. In the Random Split, this sharpness stems from stable residual patterns across randomly partitioned calibration and test sets, resulting in lower nonconformity scores and tighter quantiles. Prediction intervals tend to be narrower when model residuals are small—i.e., when $f_\theta(\mathbf{x})$ closely approximates the true label y.
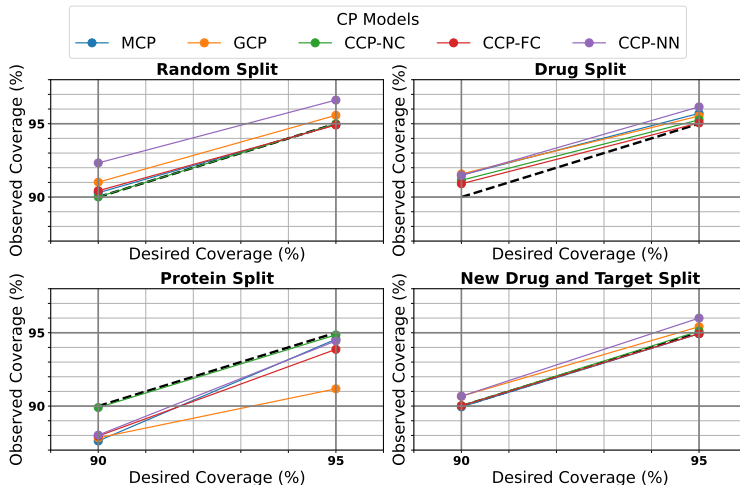
Figure 5: Observed vs. expected coverage across four dataset splitting strategies. Colors represent the six CP methods; the black diagonal indicates ideal coverage.

Additionally, in homoscedastic settings where the data noise is relatively uniform, nonconformity scores vary less across the input space, resulting in more efficient thresholds. In the more challenging New Drug–Protein Split, CCP-NC maintains superior performance by leveraging its adaptive thresholding mechanism (Equation (10)) to condition on both drug and protein clusters. This strategy enables CCP-NC to select efficient quantiles even under data scarcity by pooling nonconformity scores from structurally similar clusters. By contrast, GCP attains narrower intervals in the Drug and Protein Splits through its group-based thresholding scheme (Equation (6)), which adapts intervals based on previously seen drugs or proteins. However, when both entities in a test pair are unseen—as in the New Drug–Protein Split—GCP becomes less reliable, whereas CCP-NC continues to perform robustly due to its cluster-based generalization.

To assess subgroup-level coverage, we report MACG at the drug and protein levels (Figure 7). Each subplot corresponds to a data-splitting strategy, comparing five CP methods across coverage levels. Bars show mean MACG over test compounds or targets, with error bars denoting the STD of absolute coverage gaps. Lower values indicate smaller subgroup-level gaps and reduced uncertainty. CCP-NC consistently achieves the lowest MACG in both the Random and New Drug–Protein Splits, for both drug and protein subgroups, demonstrating its robustness in both well-behaved and challenging generalization settings. In the Drug Split, GCP yields the lowest MACG at both the drug and protein levels. This reflects the strength of GCP's entity-level conditioning: when a test drug is unseen but the associated protein is known, GCP computes quantiles using calibration interactions involving the same protein. Since there are only 229 unique proteins in the dataset—compared to 2068 drugs—each protein appears more frequently, resulting in more reliable quantile estimates. Consequently, GCP performs well when conditioning on the well-represented protein side of the interaction. In contrast, CCP-NC outperforms other methods in the
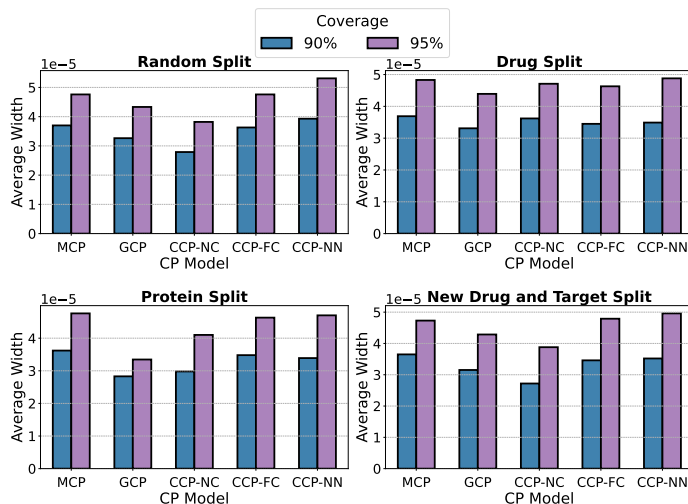
15

Figure 6: Mean prediction–interval width ($\times 10^{-5}$) produced by five CP models—MCP, GCP, CCP-NC, CCP-FC, and CCP-NN—across four data–splitting strategies.

Protein Split, where the test proteins are unseen and the drugs are known. Due to the large number of unique drugs and their sparser individual representation, GCP's conditioning on specific drugs becomes less stable. However, CCP-NC leverages clustering of nonconformity score distributions to group similar drugs and proteins, allowing it to pool calibration data across structurally or behaviorally similar entities. This generalization via cluster-level quantiles helps CCP-NC maintain low MACG even when facing data sparsity and unseen proteins. Overall, while GCP performs well in settings where one entity is well-represented and directly observed in the calibration set (e.g., Drug Split), CCP-NC is more effective in scenarios where both entities may be sparse or unseen, as it generalizes via clustering. These complementary behaviors are evident in the Protein Split, where CCP-NC achieves the best MACG at both the drug and protein levels—except at the 90% level for drugs, where its performance slightly trails.

The results presented above for CCP-NC and CCP-FC correspond to their best-performing configurations, selected from a grid of 33 combinations of calibration set split ratios $\gamma \in \{0.25, 0.50, 0.75\}$ and cluster counts $n_{\mathrm{clusters}} \in \{1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$. For each setting, we evaluated performance using the combined MACG across both drug and protein subgroups, as defined in Equation (7), and selected the configuration that minimized this value. Figure 8 illustrates the combined MACG values achieved by CCP-NC and CCP-FC across all tested parameter combinations, covering different coverage levels and dataset splitting strategies. Each subplot corresponds to a specific pair of coverage level and data split. Colored lines represent different calibration ratios γ, while the horizontal axis denotes the number of clusters. Red stars highlight the optimal configuration that produced the lowest combined MACG. For reference, baseline performance is indicated by two horizontal lines: a blue dashed line for MCP and a green dash-dotted line for GCP.
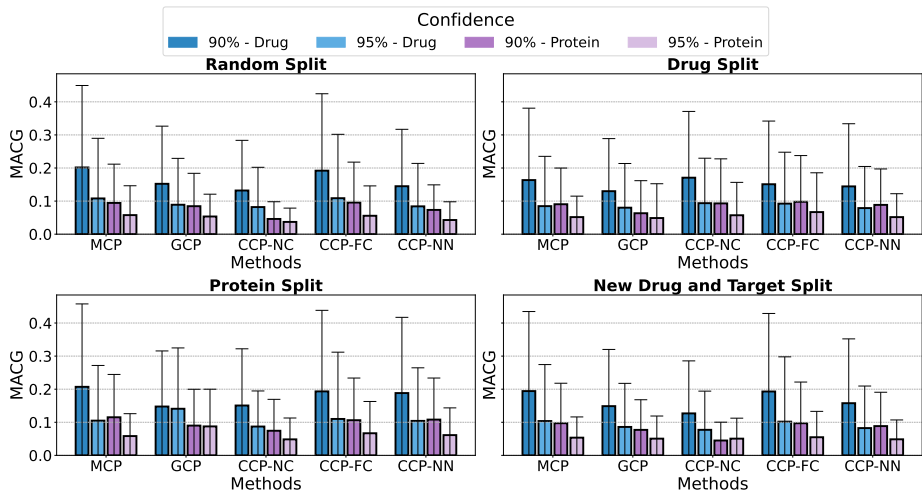
16

Figure 7: MACG (mean ± STD) computed at the drug and protein levels in the test set. Within each split, bar height gives the MACG across all compounds or targets for five CP methods. Error bars represent standard deviations across subgroups.

5. Conclusion

Based on the obtained results, it is evident that the choice of dataset splitting strategy has a substantial impact on the performance of CP methods. Among all methods, CCP-NC demonstrates better performance in terms of uncertainty quantification, consistently producing tighter prediction intervals and achieving lower subgroup-level coverage gaps (MACG) in both the Random Split and the most challenging New Drug–Protein Split. This highlights its robustness under both favorable and generalization-heavy settings. In the Drug Split, GCP outperforms other methods in terms of interval width, benefiting from its ability to condition directly on well-represented proteins. In the Protein Split, GCP yields narrower intervals overall, yet its subgroup-level calibration is less stable: CCP-NC performs better at the 90% coverage level, whereas GCP performs slightly better at 95%. Our findings suggest that clustering-based conformal predictors—particularly those leveraging nonconformity score distributions like CCP-NC, offer a better solution for reducing uncertainty in drug–target interaction datasets. By capturing structure across both drugs and proteins, CCP-NC provides a data-efficient approach to generating tighter prediction intervals. These insights may extend to other interaction-centric domains, where samples depend on two or more structured entities, making residual-driven clustering a powerful tool for uncertainty quantification.
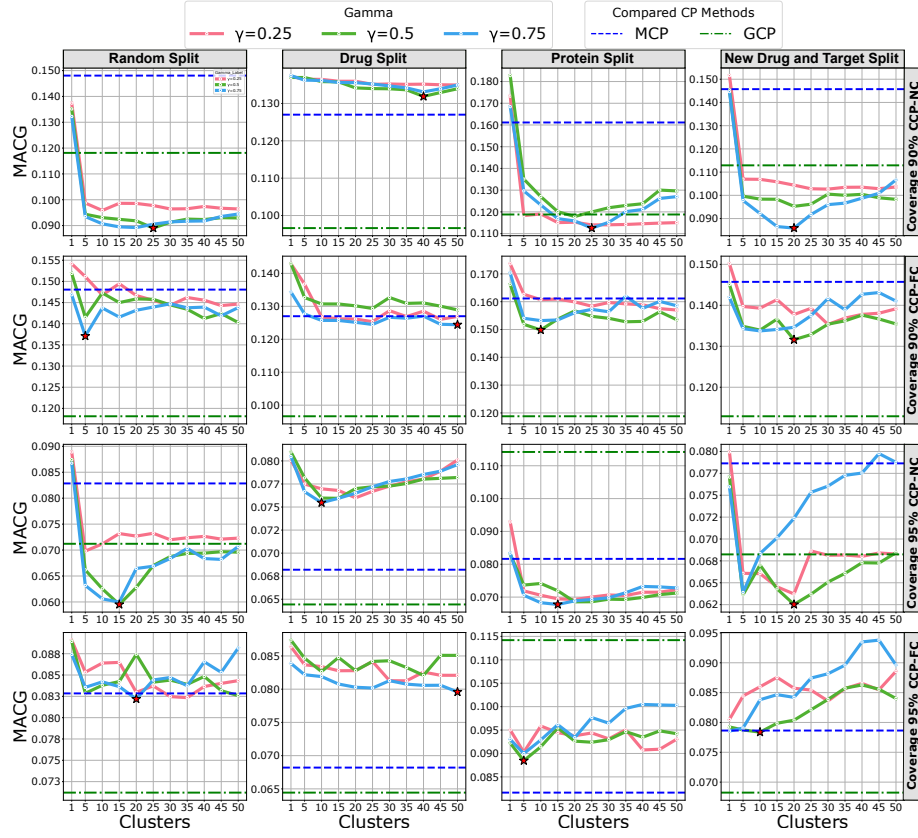
Acknowledgments

17

Figure 8: Combined MACG values for CCP-NC and CCP-FC across cluster counts, calibration set ratios (γ), desired coverage, and split strategies. Each subplot shows MACG variation with clustering parameters; colored curves indicate γ values. Red stars mark the best configurations. Blue dashed and green dash-dotted lines show MCP and GCP baselines.

References

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anthony Bellotti. Constructing normalized nonconformity measures based on maximizing predictive efficiency. In *Conformal and Probabilistic Prediction and Applications*, pages 41–54. PMLR, 2020.

Ron S Boger, Seyone Chithrananda, Anastasios N Angelopoulos, Peter H Yoon, Michael I Jordan, and Jennifer A Doudna. Functional protein mining with conformal guarantees. *Nature Communications*, 16(1):85, 2025.

George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2):211–243, 1964.

Nicolas Dewolf, Bernard De Baets, and Willem Waegeman. Conditional validity of heteroskedastic conformal regression. *arXiv preprint arXiv:2309.08313*, 2023.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36:64555–64576, 2023.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: towards cracking the language of life's code through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:7112–7127, 2021.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf008, 2025.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Anders Hjort, Jonathan P Williams, and Johan Pensar. Clustered conformal prediction for the housing market. *Proceedings of Machine Learning Research*, 230:1–21, 2024.

Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: randomization enables local guarantees. *arXiv preprint arXiv:2310.07850*, 2023.

Dimitrios Iliadis, Bernard De Baets, Tapio Pahikkala, and Willem Waegeman. A comparison of embedding aggregation strategies in drug–target interaction prediction. *BMC bioinformatics*, 25(1):59, 2024.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *arXiv preprint arXiv:2209.15145*, 2022.

Yuko Kato, David MJ Tax, and Marco Loog. A review of nonconformity measures for conformal prediction in regression. *Conformal and Probabilistic Prediction with Applications*, pages 369–383, 2023.

Siddhartha Laghuvarapu, Zhen Lin, and Jimeng Sun. Codrug: Conformal drug property prediction with density estimation under covariate shift. *Advances in Neural Information Processing Systems*, 36:37728–37747, 2023.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1): 71–96, 07 2013. ISSN 1369-7412.

Zhe Lim and Anthony Bellotti. Normalized nonconformity measures for automated valuation models. *Expert Systems with Applications*, 180:115165, 2021.

Hirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Mordred: a molecular descriptor calculator. *Journal of cheminformatics*, 10(1):4, 2018.

Davor Oršolić and Tomislav Šmuc. Dynamic applicability domain (dad): compound–target binding affinity estimates with local conformal prediction. *Bioinformatics*, 39(8):btad465, 2023.

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pages 345–356. Springer, 2002.

Harris Papadopoulos, Vladimir Vovk, and Alex Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

C. Saunders, A. Gammerman, and V. Vovk. Transduction with confidence and credibility. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, page 722–726, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of chemical information and modeling*, 54(3):735–743, 2014.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pages 475–490, Singapore, Nov 2012. PMLR.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. *Technical Report*, 2003.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.