Frontiers in Bioinformatics

TYPE Review
PUBLISHED 24 February 2025
DOI 10.3389/fbinf.2025.1507448

OPEN ACCESS

EDITED BY
Natasa Miskov-Zivanov,
University of Pittsburgh, United States

REVIEWED BY
David Alfredo Medina Ortiz,
University of Magallanes, Chile
Yannan Bin,
Anhui University, China

*CORRESPONDENCE
Andigoni Malousi,
✉ andigoni@auth.gr

RECEIVED 07 October 2024
ACCEPTED 30 January 2025
PUBLISHED 24 February 2025

CITATION
Papangelou C, Kyriakidis K, Natsiavas P, Chouvarda I and Malousi A (2025) Reliable machine learning models in genomic medicine using conformal prediction.
Front. Bioinform. 5:1507448.
doi: 10.3389/fbinf.2025.1507448

Reliable machine learning models in genomic medicine using conformal prediction

Christina Papangelou[1], Konstantinos Kyriakidis[2], Pantelis Natsiavas[3], Ioanna Chouvarda[2] and Andigoni Malousi[1,4]*

[1]School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece, [2]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, United States, [3]Institute of Applied Biosciences, Center for Research and Technology Hellas, Thessaloniki, Greece, [4]GENeTres Research Group, Center for Interdisciplinary Research and Innovation, Thessaloniki, Greece

Machine learning and genomic medicine are the mainstays of research in delivering personalized healthcare services for disease diagnosis, risk stratification, tailored treatment, and prediction of adverse effects. However, potential prediction errors in healthcare services can have life-threatening impact, raising reasonable skepticism about whether these applications have practical benefit in clinical settings. Conformal prediction offers a versatile framework for addressing these concerns by quantifying the uncertainty of predictive models. In this perspective review, we investigate potential applications of conformalized models in genomic medicine and discuss the challenges towards bridging genomic medicine applications with clinical practice. We also demonstrate the impact of a binary transductive model and a regression-based inductive model in predicting drug response as well as the performance of a multi-class inductive predictor in addressing distribution shifts in molecular subtyping. The main conclusion is that as machine learning and genomic medicine are increasingly infiltrating healthcare services, conformal prediction has the potential to overcome the safety limitations of current methods and could be effectively integrated into uncertainty-informed applications within clinical environments.

KEYWORDS

conformal prediction, machine learning, genomic medicine, uncertainty estimate, reliable predictions, perspective review

1 Introduction

Artificial Intelligence (AI)-based models are having transformative impact on high-risk predictions made for personalized medicine applications (Hamet and Tremblay, 2017). Genomic medicine as a cornerstone of precision medicine has the potential to revolutionize healthcare for rare diseases and cancer through robust and reliable personalized diagnosis, risk stratification, and tailored treatment solutions (Brittain et al., 2017). However, prediction errors can have life-threatening impact, raising reasonable skepticism on whether these applications are reliable and have clear practical benefit in routine clinical practices.

The main sources of prediction errors are the stochasticity and complexity of the models, the different data collection/curation protocols, and domain shifts that result in data falling outside training distributions (Stacke et al., 2020). Besides, even optimized models and data curation protocols do not guarantee reliability and robustness. Prediction errors can also be the result of aleatoric uncertainties reflecting the irreducible variability which often arises from the inherent randomness of the data that cannot be controlled or predicted. In addition, incomplete domain knowledge and limited data cause epistemic uncertainties that often result in model inadequacies and inevitably poor performance. These challenges underscore the need of rigorous uncertainty quantification to assess and mitigate the risks associated with prediction errors. Particularly in clinical applications that most often have no tolerance for errors, enhancing the safety and reliability of AI-based predictions is a prerequisite towards informed decision-making, improving also trustworthiness and broader acceptance of AI-driven healthcare solutions in practice.

1.1 Uncertainty quantification methods

To mitigate the risks associated with singleton predictions, where each output corresponds to a single data instance, several uncertainty quantification methods have been employed in healthcare machine learning (ML) applications to calibrate outcomes into distributional predictions (Chen and Guestrin, 2016). Bayesian inference is a statistical approach that uses Bayes' theorem to combine prior knowledge of a model with the observed data. Bayesian inference interprets probabilities as degrees of belief and helps in estimating and managing uncertainty in predictions. Specifically, Bayesian methods such as Monte Carlo dropout (Hammersley, 2013), variational inference (Blei et al., 2017), and Dempster-Shafer Theory (DST) (Xiao, 2020), along with non-Bayesian techniques like deep ensembles (Lakshminarayanan et al., 2017), softmax calibration, and selective classification (Geifman and El-Yaniv, 2019), leverage prior distributions and posterior inference to provide a probabilistic framework for estimating uncertainty in predictions. Other techniques, including Fuzzy systems (Karaboga and Kaya, 2019), Rough Set Theory (RST) (Pawlak, 1998), and Imprecise Probability (Augustin et al., 2014), have also been applied to manage uncertainty of ML models in healthcare applications (Vidhya and Shanmugalakshmi, 2020; Ahmed P et al., 2020; Giustinelli et al., 2022). These estimate the probabilistic distributions rather than deterministic outcomes, enhancing both decision-making and model interpretability—critical factors in healthcare settings for decision support and risk mitigation. Distribution-free uncertainty quantification techniques implement an alternative approach as a general framework offering rigorous statistical guarantees for black-box models, thereby reducing uncertainty in the decision-making process. Conformal Prediction (CP) provides a particularly effective and versatile distribution-free approach for statistically quantifying uncertainty (Angelopoulos and Bates, 2021). Unlike traditional prediction methods, CP generates prediction sets with guaranteed error rates, rather than point estimates. CP operates under the assumption of independent and identically distributed random variables (i.i.d.), with an emphasis on exchangeability.

1.2 Conformal prediction: principles and frameworks

Conformal prediction was initially proposed by Vladimir Vovk (Vovk et al., 2005), and later expanded by Vovk and Shafer (Shafer and Vovk, 2008). CP provides a structured framework for quantifying the uncertainty of model predictions. Depending on the class labels, CP estimates prediction intervals for regression problems and a set of classes for classification problems, guaranteeing coverage of the actual value with a predefined confidence level. CP leverages the concept of how "unusual" a new sample is relative to prior observations to produce reliable confidence levels. Therefore, CP uses past experience to determine accurate confidence levels in new predictions (Shafer and Vovk, 2008). Operating under the assumption of independent and identically distributed random variables (i.i.d.) or exchangeability, CP ensures that the order of observations does not impact their joint distribution. This makes CP particularly valuable in real-world biomedical applications in which making assumptions about the underlying data distribution may be challenging or unrealistic. CP is defined as a mathematical framework that can be used with any ML model to produce reliable predictions with high probability and user-defined error rates (Vovk et al., 2005; Shafer and Vovk, 2008).

Given a set of training data D, with n instances $\{(x_i, y_i), \ldots, (x_n, y_n)\}$, where x_i is a feature vector and y_i is the true label of the ith sample, with K labels in Y, the objective is to predict the label $y_{n+1} \in Y$ for a new sample with feature vector x_{n+1}. In classification problems, all possible classes of a new instance are tested and the probability of a prediction to be the correct one for each class is quantified. To this end, a *non-conformity* score α_i is calculated, which is based on the underlying ML algorithm and indicates how strange an instance is compared with other instances. A simple example of a *non-conformity* score is the *1-predicted probability of the true class*, otherwise called inverse probability. For a new instance the *non-conformity* score is estimated for each possible label and a *p-value* for each possible label is calculated to evaluate the *non-conformity* score of the new instance against all other scores. In a regression analysis framework, CP transforms point predictions to intervals that contain the true value with a level of guarantee defined by the user. To compute the *non-conformity* score for every sample in the training set, we measure how different the observed value is compared to the model's prediction. Detailed information about the mathematical framework can be found in the Supplementary File.

Figure 1, illustrates the two primary CP frameworks, the Transductive CP (TCP) and Inductive CP (ICP). TCP, also known as the full version of CP, uses all available data to train the model, resulting in highly accurate and informative predictions (Vovk et al., 2005; Gammerman and Vovk, 2007). In TCP, after selecting a suitable non-conformity function, the features of a new instance x_{n+1} are added to the dataset, and assuming its class y_{n+1}, the model is retrained K times, where K represents the number of possible classes (e.g., two for binary classification). For each retrained model, the *non-conformity* scores and *p-values* are calculated to determine whether the new instance conforms to the existing data. This process creates prediction regions but is computationally intensive, making it ideal for small datasets or online applications. In contrast, ICP addresses the computational inefficiency of TCP by splitting the training dataset into two subsets: a smaller training set and a

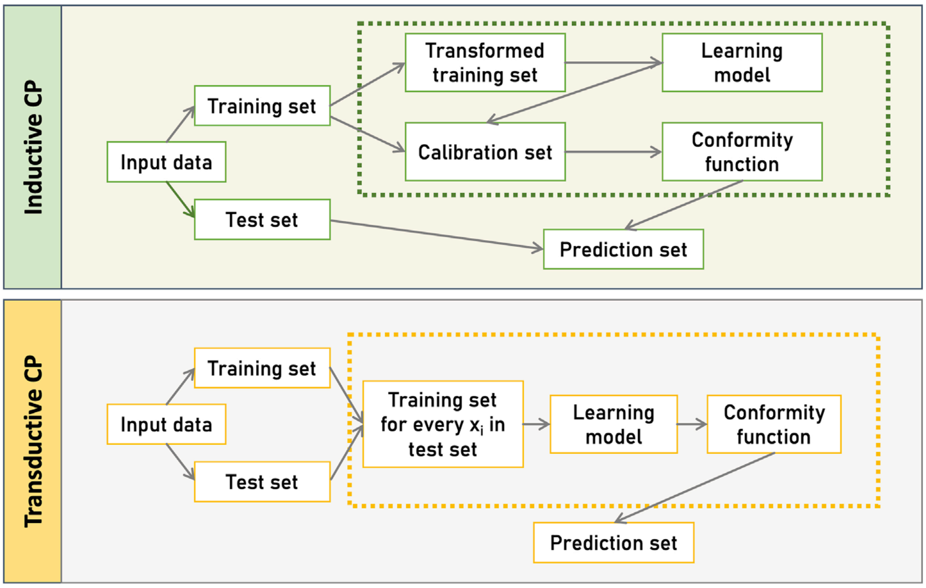

FIGURE 1
CP-based frameworks. Inductive CP (ICP), involves a static split of the dataset into training and calibration sets, using a single model for all predictions, while Transductive CP (TCP), builds a new model for each test instance, offering potentially more tailored predictions at the cost of increased computational effort.

calibration set (Papadopoulos, 2008). The training set trains the model only once, while the calibration set is used exclusively to compute *p-values* for new test instances. Although ICP sacrifices some flexibility, it provides unbiased predictions and is well-suited for large datasets, such as those encountered in multi-omics analyses. The efficiency of ICP depends on factors like dataset size and quality, the underlying ML algorithm, and the chosen non-conformity measure. Both frameworks offer robust methods for creating reliable prediction intervals while accommodating different computational constraints.

1.3 Parameterization and evaluation

Conformal predictors enhance the reliability of black-box models by generating prediction sets that reflect uncertainty in high-risk applications. The evaluation of the conformalized model generally concerns adaptivity, size, and coverage of the prediction intervals. As Angelopoulos (Angelopoulos and Bates, 2021) proposed, a model's adaptivity can be assessed by the size of the prediction sets, with larger sets indicating higher uncertainty and more challenging predictions, while smaller sets signify easier ones. Adaptivity is closely linked to the model's conditional coverage, which ensures that the true label falls within the prediction region at a defined confidence level for any subset of the test set. Marginal coverage is achievable, but conditional coverage requires consistency across subsets of the test data. Angelopoulos et al. recommended the size-stratified coverage (SSC) to measure adaptivity and suggested verifying conditional coverage by repeating the framework with different combinations of calibration and test sets (Angelopoulos and Bates, 2021). In a similar vein, Park et al. proposed the meta-XB, a meta-learning approach designed for cross-validation-based CP that focuses on reducing the average size of prediction sets while ensuring formal calibration for each task (Park et al., 2023).

Given that CP can be applied across various prediction models, its effectiveness hinges on three critical parameters, the *non-conformity* function, the size of the calibration set within an ICP framework, and the underlying model. The effectiveness of the *non-conformity* function depends on how well it aligns with the underlying ML model, influencing the accuracy and reliability of prediction intervals (Vovk et al., 2005). The calibration set size is equally significant, as larger sets can enhance model coverage at the expense of increased computational costs. Recent innovations, such as scaling methods (Abad et al., 2022) and CP extensions for efficient transformer inference (Abad et al., 2022), address these challenges. Lastly, the choice of the underlying model profoundly impacts CP outcomes, as the model's predictive performance directly affects the *non-conformity* measure and, consequently, the quality of prediction intervals. Selecting well-calibrated and high-performing models tailored to specific applications is essential in building effective conformal prediction frameworks.

1.4 Distribution shift

A major concern in ML applications is the deviation of the properties and distribution of the new, unseen data compared to those of the training set. The so-called distribution shift is frequently observed in real-world predictive models, when the joint distribution of inputs and outputs differs between training and test stages. Covariate shift occurs when there is a discrepancy between the distributions of input points in the training and the test datasets, even though the conditional distribution of output values given input points remains consistent (Sugiyama et al., 2007). The weighted CP proposed by Tibshirani et al. (2019) can handle covariate shift by weighting each *non-conformity* score by a probability that is proportional to the likelihood ratio of the new data distribution to those used to build the model. The maximum mean discrepancy (MMD), proposed by Borgwardt et al. (2006), is a kernel-based statistical test used to determine whether two samples are drawn from different distributions. In contrast to typical measures like Kolomogorov-Smirnov test that can only be applied in vectors, MMD is applicable in multivariate data that are frequently met in genomic data analyses. The null hypothesis in MMD statistical test states that there is no difference between the distributions of the two datasets and therefore that the datasets are drawn from the same distribution.

The label or prior probability shift, refers to a shift in the distribution of class variables. A variation of CP named Mondrian CP (MCP) can remedy this difference between the train and validation samples. In MCP, each class is evaluated independently to determine the confidence of assigning an instance to that class. Predictions for the calibration set produce *non-conformity* scores for each class. MCP ensures controlled error rates by categorizing training sets based on features or their combinations and defining significance for each category (Vovk et al., 2005; Boström et al., 2021). It compares non-conformity scores only within the same category, making it suitable for poorly distributed datasets. Label Conditional Mondrian Conformal Prediction (LCMCP) is a specific case of MCP where the category of each instance is determined by its label. Under the same scope, Bostrom et al. proposed the Mondrian conformal regressors handling the range of the prediction interval (Boström and Johansson, 2020).

Recent work suggests CP as an effective framework that can handle distribution shifts. Cai et al. utilized an Inductive Conformal Anomaly Detection (ICAD) approach for online detection of distribution shifts on high-dimensional data with low computational cost and efficiency (Cai et al., 2021). Hernandez et al. demonstrated the robustness of conformalized models in predicting the activities of novel molecules on cancer cell lines, offering valuable insights for drug discovery under strong distribution shifts (Hernandez-Hernandez et al., 2024). However, in real world applications, distribution shifts are commonly encountered with unexpected results in model performance. For large scale datasets, black-box model architectures or hidden distribution shifts, predictions must undergo careful examination before being applied in clinical decision making. To prove this Kasa et al. examined how those shifts affect CP and concluded that the performance degrades and the coverage guarantees are frequently violated, highlighting the challenges and the need for further elaboration on these issues (Kasa and Taylor, 2023).

2 Conformal modelling in genomic medicine

2.1 Current application landscape in biomedicine

In principle, CP coupled with any traditional learning model can be used to address uncertainty in a wide range of scientific domains. In medical applications, it is crucial for any predictive model to generate predictions tailored to each individual patient rather than relying on generalizations from a broader population. Hence the definition of the confidence intervals for individual predictions is critical especially when these models are adopted in clinical environments (Vazquez and Facelli, 2022). In such clinical applications CP is used to intuitively express the uncertainty of a prediction and to facilitate the model's transparency and robustness (Lu et al., 2022a). For example, CP has been employed in medical imaging applications for subgroup analysis, distribution shift estimation, and for the elimination of prediction errors in safety-critical applications (Lu et al., 2022b; Millar et al., 2024). Using microscopic biopsy images Olsson et al. implemented an effective CP-based model for diagnosis and grading of prostate cancer (Olsson et al., 2022). Additionally, Kapuria et al. proved that, using CP, clinicians can make informed decisions and minimize the risk of colorectal cancer polyps misdiagnosis (Kapuria et al., 2024). In non-cancer applications, CP was used by Lu et al. to develop a deep learning model for grading the severity of spinal stenosis in lumbar spine MRI (Lu et al., 2022a) and Wieslander et al. combined deep learning methods with CP to predict tissue sub-regions using hierarchical identification on rat lung slides (Wieslander et al., 2020).

In preclinical settings, CP has been applied in drug discovery, mainly to predict the biological activity of compounds based on their chemical structure. CP-based methods have been used as an alternative approach to traditional QSAR models, to predict target-ligand binding that are enriched with uncertainty estimates (Xu et al., 2023; Bosc et al., 2019). For example, Alvarsson et al. used CP on top of random forest models to classify three different ATP transporters (Alvarsson et al., 2021). The authors concluded that the higher the level of confidence the larger the prediction interval or set of predictions, and they suggested CP as an effective method for drug discovery applications. Toccaceli et al. demonstrated the application of an Inductive Mondrian Conformal Predictor to predict the biological activities of chemical compounds by addressing challenges such as the large number of compounds, the high dimensionality of the feature space, the sparseness, and the class imbalance (Toccaceli et al., 2017). In the same context, CPSign proposed a conformal predictor that is applied to chemical descriptors for chemoinformatics modeling (McShane et al., 2023) while several other applications in biomolecular design proposed sophisticated methods to handle covariate shift, enabling the computation of distribution-free prediction intervals (Fannjiang et al., 2022; Laghuvarapu et al., 2024). Similar CP approaches have been extensively applied in modeling chemical compound toxicity (Forreryd et al., 2018; Fagerholm et al., 2022; Geylan, 2021; Zhang et al., 2021).

2.2 Existing conformalized models in genomic medicine

Despite their widely recognized contribution to medical imaging and drug discovery, conformal predictors have not been sufficiently used in joint applications of genomics and medicine. Genomic medicine is an emerging medical discipline and a rapidly evolving field of predictive modeling applications. In areas such as oncology, pharmacology, rare or undiagnosed diseases, and infectious diseases genomic medicine has a transformative impact on improving medical decisions, and advancing medical knowledge, and healthcare delivery.

In the field of genomic medicine only few CP uncertainty-aware models have been reported in the literature. Ianevski et al. used patient-derived single-cell transcriptomic data to train a gradient boosting model that prioritizes multi-targeting therapeutic compounds for stratified cancer treatment (Ianevski et al., 2023). In this *ex vivo* drug testing methodology the conformalized model was built using subclone-specific differentially expressed genes and helped to filter out predictions with low conformity scores. Single-cell transcriptomic data was also used by Sun et al. to identify subtypes within the neural stem cell lineage (Sun et al., 2024). In this work, CP is part of a general framework for estimating uncertainty in spatial gene expression predictions and is applied to calculate the calibration score that links the cell-centric variability to the prediction error.

In a different setting, Sun et al. proposed a method to address personalized genetic risk assessment for complex diseases that relies on a Mondrian cross-conformal prediction model to estimate the confidence bounds of the polygenic risk score prediction (Sun et al., 2021). The proposed method showed that using the predicted risk of each individual to classify as a case or control is more clinically relevant than group-wise assignments to high-risk or low-risk groups based on an arbitrary selection of the extreme scoring samples.

On the protein level, conformal predictions have recently been employed as an effective approach to detect protein homologies enabling the discovery of new proteins with likely desirable functional properties (Boger et al., 2024). The method provides statistical guarantees of the homology searches of a query protein against a lookup database -instead of protein pairs- and functional annotations by leveraging the vast amount of protein structures produced by algorithms such as Alphafold (Jumper et al., 2021). The proposed conformalized protein mining method has potentially significant implications in genomic medicine including drug repurposing utilizing proteins with unique and desirable features, the development of therapeutic enzymes or monoclonal antibodies for personalized disease treatment and engineering proteins for enhanced stability, activity, or binding affinity, creating more effective therapeutics.

In pharmacogenomics, prediction error estimates have been employed in a CP model to predict drug sensitivity and prioritize drugs using gene expression levels of cancer cell lines (Lenhof et al., 2024). The prediction outcomes show substantial improvement of CP prediction accuracy and highlight the importance of developing more sophisticated methods that incorporate multi-omics data, to address not only monotherapies but also combinatorial drug delivery.

2.3 Pitfalls and challenges

To advance clinical applications, genomic medicine models must deal with a variety of uncertainty-inducing and safety-critical issues that are mainly caused by the inherent complexity and variability of the biological systems, the inter-individual heterogeneity in genetic profiles, environmental exposure, and lifestyle as well as the non-linearity of the interactions within the patients data. Uncertainty manifests in various steps of genomic analysis, and particularly for ML applications has different dimensions. Uncertainties might involve the ambiguity, complexity, or deficiency of the data, as well as the unpredictability of the models. It is important to understand the dimensions of uncertainty, however, it is also important to recognize that uncertainty is not always problematic (Barlow-Stewart, 2018). Uncertainty estimates can help acknowledge the complexity of molecular events and account for the data variability in a model recalibration.

By definition, CP estimates uncertainty when making personalized decisions and leverages the evidence linking each individual's genetic makeup to zero-tolerance applications such as medical decision-making, diagnosis, risk assessment, and treatment strategies. In this context, CP-enriched models can greatly benefit from the availability of massive amounts of trainable multi-omics data derived from high-throughput sequencing technologies and they can in turn contribute to improved generalizability and calibration of rare events of the learning models. However, ML applications that do not integrate uncertainty measures face significant challenges. The inability to quantify uncertainty can result in overconfident predictions, which pose risks in high-stakes scenarios like personalized diagnostics and tailored treatment planning. Without uncertainty estimates, models may struggle to convey the reliability of their predictions, leading to potential errors in decision-making. Furthermore, such models often fail to adapt to novel scenarios, especially under distributional shifts or when encountering rare events. The lack of uncertainty measures also limits trust and transparency in clinical contexts, where interpretability and confidence in the model's outputs are of paramount importance (Chua et al., 2023; Begoli et al., 2019).

2.4 Potential applications in genomic medicine

CP can be an essential component for a much wider range of genomic medicine applications combining predictive modelling and high-risk decision-making. Genomic ML applications with clinical relevance can greatly benefit by uncertainty estimates in the following fields.

2.4.1 Variant calling and prioritization

The diagnosis and disease risk assessment in genomic medicine is most often based on the presence of genetic variants. In next-generation sequencing studies, genetic variants are detected by complex deep neural network architectures, e.g., DeepVariant (Poplin et al., 2018) and DeepSNV (Gerstung et al., 2012). However, accurate variant calling is not a straightforward process and is often error-prone, especially for tumor samples with high heterogeneity and low purity, or for genomic regions that are difficult to map (Olson et al., 2022). To be able not to take the risk of a prediction could be of great clinical significance, particularly while trying to distinguish between somatic and germline variants or to prioritize rare variants. In addition to variant calling, prioritizing the detected variants based on their functional effect introduces challenges that can be of clinical relevance when sorting neutral or deleterious variants among those of unknown significance.

2.4.2 Immunotherapy response prediction

In a similar setting, the mutational load of tumor DNA samples, known as tumor mutational burden, is a strong predictor of response to immunotherapy. However, several issues, including the variability of response levels by cancer type and the lack of a standardized method for calculating variant burden, limit the reproducibility and reliability of the predictions. In this context, conformalized learning models are suitable for estimating the uncertainty of the immunotherapy response predictions, and to avoid to take the risk of a prediction in inconclusive cases.

2.4.3 Pharmacogenomics

Besides predicting immunotherapy responses, the genetic makeup is a mainstay of research in pharmacogenomics to tailor therapeutic solutions either by identifying biomarkers of pharmacological response or by developing learning models. ML-based applications develop strategies to prioritize candidate anti-cancer drug compounds, or predict the sensitivity levels of a particular compound, yet out of the context of reliability testing and uncertainty estimates (Adam et al., 2020; Kardamiliotis et al., 2022). Recently, Lenhof et al. developed a conformalized approach that predicts and prioritizes drug sensitivity on cell line-based monotherapy responses, based on gene expression profiles and user-defined certainty levels (Lenhof et al., 2023). Compared to cell lines, patient-derived profiles are preferred in the development of clinical pharmacogenomic models however, they introduce additional complexities that increase the uncertainty and risk of an erroneous prediction. In addition, novel approaches demonstrate the need to integrate multi-omics data in drug response predictions, including mutations, copy number variations and proteomics. Multi-omics data can be particularly informative and, when combined with uncertainty estimates, could facilitate safer predictions and decipher the physical/functional gene-drug interactions. These potential applications collectively demonstrate the need to establish robust genomic medicine frameworks capable of evaluating the predictability in clinical applications and enhancing reproducibility.

2.4.4 Reverse vaccinology

Reverse vaccinology (RV) is a rapidly evolving approach in vaccine development against pathogens that utilizes genome sequences to predict antigens that can elicit strong immune responses. RV workflows include several analysis steps (Trygoniaris et al., 2024) in which ML models are often used to predict B-cell and T-cell epitopes based on the pathogen's genomic and proteomic features (Clifford et al., 2022). In addition, predictive models are used to assess and prioritize vaccine candidates based on factors like antigenicity, immunogenicity, conservation across strains, and homology to host proteins to avoid autoimmune reactions (Ong et al., 2020). A critical step in RV is the integration of 3D modelling algorithms to predict the folded 3D structure of the vaccine construct and the development of multi-epitope vaccines. Considering the poor quality of the training data and the difficulties in experimental screening, being able to quantify uncertainties in each ML-based analysis would greatly advance model calibration and validation (Goodswen et al., 2023). In this context, CP models can be particularly useful in validating ML predictions by ensuring that the specified coverage probability is maintained across different datasets and pathogen-host application scenarios.

2.4.5 Antimicrobial resistance

Antimicrobial resistance (AMR) is a serious public health threat that is responsible for prolonged hospitalizations and more than one million deaths per year (Murray et al., 2022). The availability of millions of whole genome sequencing data annotated with diverse AMR phenotypes enabled the development of ML methods that predict AMR using pathogens features, mainly genomic variability (Kim et al., 2022; Nguyen et al., 2018) and biochemical information (Kavvas et al., 2020). However, the reliability of the predictions is subjected to several confounding factors, e.g., biased sampling and poor genome assembly quality due to increased contamination rates, poor coverage and low read depth. Erroneous predictions of AMR against antibiotic compounds can be life-threatening and therefore uncertainty guarantees in either supervised classification (sensitivity/resistance prediction) or regression problems (quantification of the minimum inhibitory concentration values) can be particularly valuable. In this context, conformalized models can be important preventive measures offering safer clinical decision making, while also helping in deciphering the molecular mechanisms underlying AMR.

In this study, we rigorously explore the potential of conformal predictors in genomic medicine and demonstrate their pivotal role in yielding more reliable predictions using three application scenarios. Specifically, we evaluated CP-enriched models on a binary classification, on a multi-class classification problem under distribution shift and a regression-based application aiming to gain further insights into how conformalized predictive modeling can be practically integrated into genomic medicine. The study discusses further the strengths and challenges and highlights the main issues that should be addressed in order to unequivocally ensure patient safety when pivotal decisions are delegated to clinically deployed AI systems.

3 Experimental setting and results

To practically assess the applicability of CP in genomic medicine we sought to examine how ML models can benefit from conformalized predictions in two example classification problems and one regression problem. The objective was to cover both binary and multi-class predictions, small and larger datasets, different application domains and both inductive and transductive frameworks. First, a TCP-based pharmacogenomic learning model was implemented to demonstrate the impact of conformal predictors in tailoring personalized therapeutic decisions. Transcriptomic profiles of rheumatoid arthritis and Crohn's disease patients undergoing infliximab treatment were used to estimate the uncertainty of the drug sensitivity predictions (Figure 2). In the multi-class setting, an inductive conformal predictor was built to assess the diagnostic predictions for patients with different transcriptional subtypes of diffuse large B-cell lymphomas (Figure 2). Finally, in the regression setting, an inductive conformal predictor was used to predict the pharmacological response of cancer cell lines to afatinib. Both classification models used publicly available gene expression datasets deposited in Gene Expression Omnibus under the accession IDs GSE42296 for rheumatoid arthritis and Crohn's disease (Mesko et al., 2013) and GSE181063 for diffuse large B-cell lymphoma samples. To

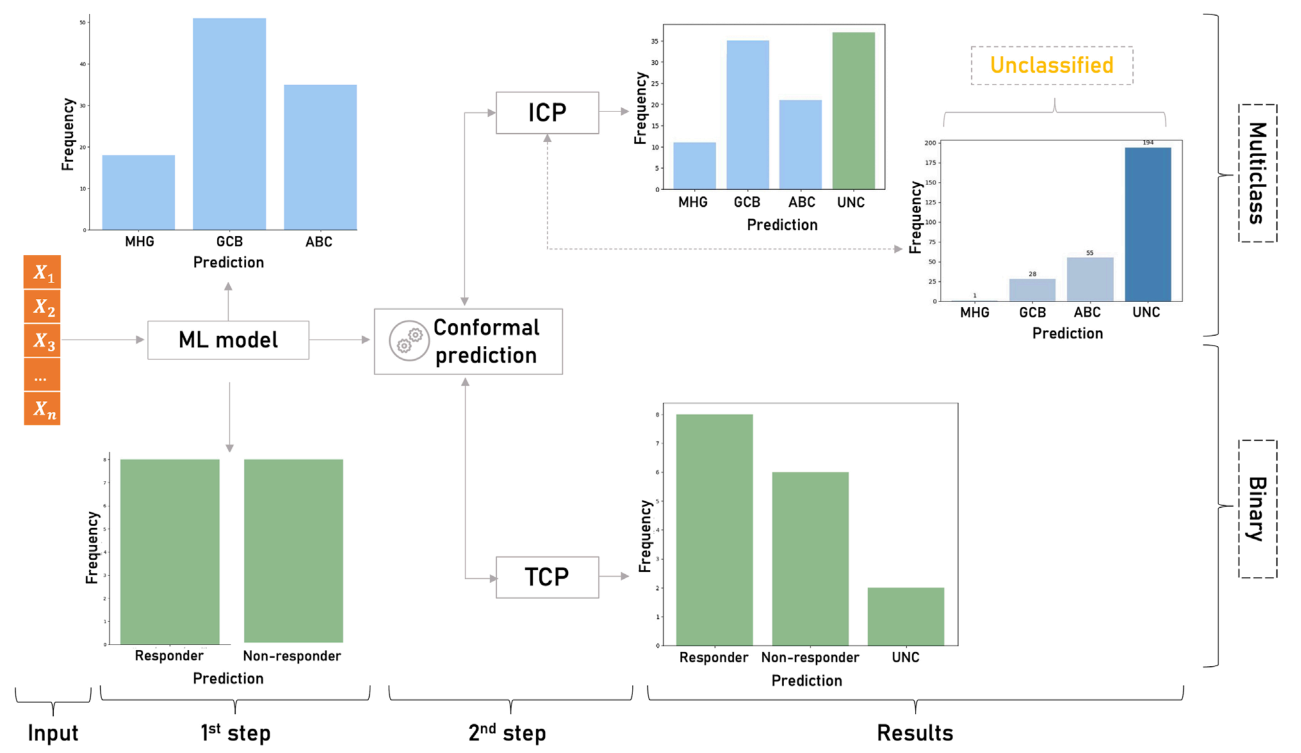

FIGURE 2
Overview of the classification study design. First, the ML model produces singleton predictions for the binary (responder/non-responder patients to infliximab) and the multiclass (MHG, GCB and ABC subtypes of diffuse large B-cell lymphomas) problems. Singleton predictions are given without any indication of their accuracy and reliability. In the second step, CP is applied on the results of the ML models to estimate the uncertainty of each prediction. In the binary classification, TCP produces a prediction region that contains the true class with high probability and detects the uncertain (UNC) predictions. In the multiclass setting ICP identifies unreliable predictions among the samples classified by the non-conformalized model.

train the regression model, data from the Genomics of Drug Sensitivity in Cancer (GDSC) database was used. In all use cases, we applied MRMR (Maximum Relevance - Minimum Redundancy) feature selection and statistical tests to assess the validity of the i.i.d. assumption (Peng et al., 2005). It should be noted that although the application scenarios address real-world research problems, the prediction results are not intended to produce novel research findings as this is out of the scope of this perspective review.

3.1 Responder prediction to infliximab

In their study, Mesko et al. correlated the pharmacological response of rheumatoid arthritis and Crohn's disease patients to infliximab using their transcriptomic profiles (Mesko et al., 2010). The study includes 44 Crohn's disease patients and 34 rheumatoid arthritis patients of which 40 responders and 38 non-responders. Affymetrix Human Gene 1.0 ST array quantified the expression levels of each sample in 33,297 target probes. The objective was to identify subsets of genes that can act as drug sensitivity biomarkers. In our experiment we sought to compare non-conformal and conformalized models in the binary setting using an ML model and a TCP framework to estimate the uncertainty of the model. TCP was selected as a favorable framework because it avoids the extra split for the calibration set which is preferable for small sample sizes. To evaluate TCP, we utilized the *empirical coverage* (Angelopoulos and Bates, 2021), which measures the frequency of the true class within the prediction region. We then assessed the error rate threshold, ensuring it did not exceed the specified significance level of the conformal predictor.

Following the preprocessing step, we trained an SVM model on the top 100 genes with the highest discriminative power according to MRMR. For the 20% randomly selected patients included in the test set, the model yielded 87% accuracy (AUC = 0.9), optimized by a grid-based parameter tuning (Figure 2). By setting the significance level to 95% and defining the inverse probability, $1 - p(y_i|x_i)$, as the *non-conformity measure* conformal predictions resulted in a 2.25% error rate compared to 12.5% of the SVM model without CP (Table 1). Two out of the 16 test cases were marked as uncertain requiring further evaluation by an expert physician. In this case, CP eliminated the misclassified samples by sorting out ambiguous cases, while for half of them the ML model alone made erroneous singleton predictions. The use of CP in this use case succeeded to reduce wrong predictions and to identify those cases that are hard to classify and should be forwarded for manual assessment.

Concerning the singleton predictions, TCP identified eight non-responders and six patients who will respond to infliximab. This group of patients is correctly classified to the actual class with an error rate of 5%. Moreover, for two wrong predictions of the non-conformalized SVM model, CP flagged one of them as uncertain, which identifies this patient as a difficult-to-classify case (Table 1).

TABLE 1 Performance of the non-conformal and the conformalized binary and multiclass models (95% confidence interval).

Experiment	Non-conformal model		Conformalized model			Comparison
	Model	Error rate	E. Coverage	Error rate	UNC rate	Error detection
TCP (Binary)	SVM	12.50%	93.75%	6.25%	12.50%	50.00%
ICP (Test)	XGBoost	16.25%	95.20%	4.80%	35.60%	70.50%
ICP (Validation)	XGBoost	16.35%	96.70%	3.3%	38.18%	86.82%

For this patient, the treatment decision should be made by an expert. Overall in these personalized therapeutic decisions, CP can stand alongside the physicians to flag the difficult-to-predict patient cases for further manual data curation and closer treatment monitoring, thereby improving the decision-making time and minimizing the risk of wrong interventions.

3.2 Predicting molecular subtypes of diffuse large B-cell lymphoma

In the multi-class use case we used CP as a diagnostic predictor to classify patients with diffuse large B-cell lymphoma based on the distinct transcriptional profiles of their tumor cells. Diffuse large B-cell lymphoma is the most common hematological malignancy characterized by highly heterogeneous molecular signatures. Approximately 80% of the lymphomas are curable using R-CHOP combination therapy yet, there is a biologically heterogeneous group of patients that differs in terms of their clinical characteristics and prognostic factors (Painter et al., 2019). Therefore to enable precise patient stratification in clinical trials, we first have to distinguish patients who are likely to respond to R-CHOP alone from patient groups who may benefit from emerging therapies based on the molecular heterogeneities of the disease (Lacy et al., 2020).

So far, diffuse large B-cell lymphoma patients are classified based on the Cell of Origin (COO) in the activated B-cell like type (ABC) and the germinal center B-cell like (GCB) subtypes. Recently, Sha et al. proposed a new distinct molecular subtype with aggressive clinical behavior called molecular high-grade B-cell lymphoma (MHG) (Sha et al., 2019; Sha et al., 2015). Patients of this subtype tend to not respond to R-CHOP therapy, despite the similarity with the GCB subtype, and they may benefit from either intensified chemotherapy or new targeted therapies. Clinical trials require the identification of the COO to personalize therapeutic interventions and to decipher the mechanisms of the disease pathogenesis.

In this experiment we built an inductive version of the CP model on a gene expression dataset of 1,311 samples extracted from formalin-fixed, paraffin-embedded biopsies (GEO Data series: GSE181063). The RNA samples include 345 ABC, 517 GCB, and 170 MHG molecular subtypes except for 278 patients who were not classified in any of the three classes and characterized as unclassified (Figure 2). Illumina's HumanHT-12 WG-DASL V4.0 beadchip array quantified the expression levels of each sample in 29,377 target probes. Following a data cleansing and quality control step 20 probes were selected by the MRMR algorithm to build the training feature set. The multi-class model was trained by XGboost (Chen and Guestrin, 2016), the hinge loss function was applied as *non-conformity* measure in the ICP model and the empirical coverage was used to evaluate the conformal predictor.

XGBoost has a classification error of 16.25% on 10% of randomly selected patients. The conformalized XGBoost model resulted in 4.8% subtype classification error using 95% confidence level. In addition, the inductive predictor flagged, 37 patients (35.6%) as uncertain that are distributed in the following prediction regions: {MHG, GCB} = 8, {MHG, ABC} = 2, {GCB, ABC} = 19, {MHG, GCB, ABC} = 8. Non-singleton predictions involve mainly GCB samples that are most often misclassified as ABC samples. MHG has a clearly separable profile being transcriptionally closer to the GCB subtype. For eight patients the conformalized model was not able to exclude any prediction region. However, the ICP model managed to avoid the misclassification of 12 out of the 17 wrong predictions of the XGBoost model alone. The results reinforce the reliability of the prediction regions, as they detect the wrong assessments of the basic algorithm and give a better view of the difficult examples, while at the same time, they limit the range of possible classes to facilitate the final expert decision.

Concerning the 278 unclassified patients, although there is no class assignment ICP provides singleton predictions for 30.2% of the samples {MHG} = 1, {GCB} = 28, {ABC} = 55 (Figure 2). The remaining are ambiguous cases involving either two classes {MHG, GCB} = 1, {MHG, ABC} = 9, {GCB, ABC} = 135, or all three {MHG, GCB, ABC} = 49. Among the 194 uncertain cases, most of them involve double predictions (145 cases) of GCB and ABC classes, which is also inline with the principal component analysis in Figure 3. Both single and double predictions provide insights beyond what a non-conformalized learning approach alone can offer and can be useful in preventing erroneous predictions that are of major importance in clinical decision-making.

To evaluate the reliability of the prediction regions on unseen data we sought to examine the fundamental exchangeability assumption on an external diffuse large B-cell lymphoma gene expression dataset (GEO Data series: GSE117556). The gene expression profiles were produced from 789 RNA samples extracted from formalin-fixed, paraffin-embedded biopsies using Illumina's HumanHT-12 WG-DASL V4.0 beadchip array. To assess the level of distribution shift we applied the MMD measure and compared the produced probability distributions of the two datasets. Figure 4 shows the distribution shift between the two datasets. We computed an MMD statistic of 0.0011 and performed a permutation test to determine the p-value, which was found to be 0.017. Since this p-value is less than the significance level a = 0.05, we reject the null hypothesis that the two

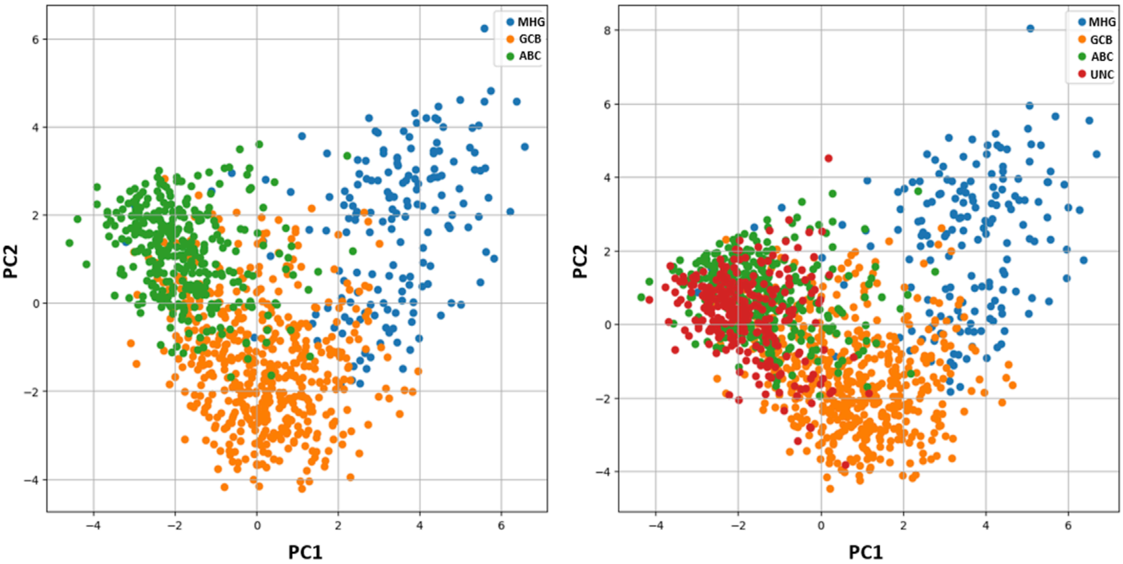

FIGURE 3
Principal Component Analysis (PCA). Left: PCA analysis of the samples without unclassified cases (UNC), revealing a small overlap between subtypes ABC and GCB. Right: PCA analysis of the dataset including unclassified cases (UNC), shows the UNC class overlapping with subtypes ABC and GCB.

datasets are generated from the same distribution. The robustness of the conformalized model in the distribution-shifted data was examined by estimating the classification performance of the ICP model on the external data. For a 95% significant level, the ICP model resulted in 26 misclassifications while the XGboost model alone failed to correctly classify 129 samples, out of the 789 samples. The ICP model flagged 112 out of the 129 misclassified samples as uncertain cases to be further assessed by clinical experts. The erroneous predictions of the conformalized model, limit by 80% the risk of failure on data under distributional shift. However, there is an increased number of double predictions that mainly involve the GCB and ABC samples accounting totally for 225 samples and 77 cases for triple predictions. On the contrary, the MHG class does not show significant overlap with the GCB and ABC samples, accounting totally for 2 and 14 cases, respectively.

The results indicate that MHG has a separable transcriptional signature while deeper investigation is needed to accurately discriminate the GCB and ABC molecular profiles. Despite its ability to protect from false predictions, the conformalized model is more conservative than the XGboost model in making accurate singleton predictions (482 in total, of which {MHG} = 32, {GCB} = 295, {ABC} = 155). In other words CP effectively minimizes the risk of underestimating uncertainties at the expense of a lower number of definitive assessments.

Overall, in this generalizability test the ICP model achieved 96.6% empirical coverage on the unseen dataset. In principle, we proved the ability of our conformal classifier to generalize to data with different distributions, a significant advantage in medical applications where data heterogeneity is a common issue. These results, shown in Table 1, demonstrate the robustness of the conformal classifiers in handling data with varying distributions, but also the need to promote the adoption of CP-based frameworks in genomic medicine to be able to draw safer and more definitive conclusions.

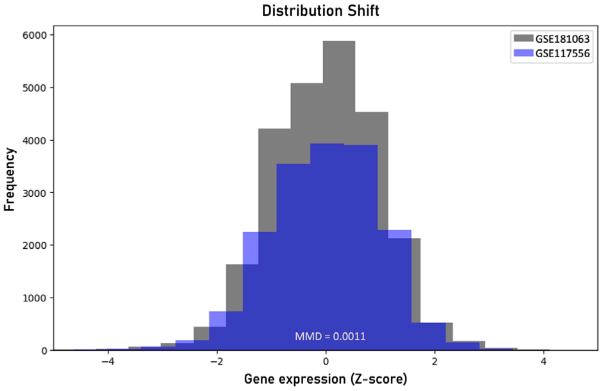

FIGURE 4
Visualization of the distribution shift between the two diffuse large B-cell lymphoma datasets based on the maximum mean discrepancy (MMD).

3.3 Predicting pharmacological response of cancer cell lines to afatinib

In the regression use case, we implemented an ICP approach to predict the resistance of cancer cell lines to afatinib, an antineoplastic agent that is used to treat locally advanced and metastatic non-small cell lung cancer. Instead of categorizing samples into binary classes drug responses are quantified based on continuous drug concentrations that caused inhibition of 50% cell viability (IC50), with higher IC50 values indicating greater resistance. In this study, we sought to evaluate the scalability and robustness of CP uncertainty-aware regression models in predicting IC50 values using gene expression levels of cancer cell lines. The model was trained on a dataset of 765 cancer cell lines, each one including the expression levels of 17,613 genes and

TABLE 2 Performance of the ICP regression model.

	Significance level	α-Quantile	E. Coverage
1	0.95	6.98	95.5%
2	0.90	6.77	92.6%
3	0.85	6.63	90.6%

the corresponding IC50 values were recorded 72 h after treatment. The pre-processing steps, including outlier management and feature selection, refined the dataset to 677 cell lines with IC50 ranging between 0.00316 and 675, and identified 10 significant genes, meeting the i.i.d. assumption. A Random Forest (RF) algorithm was employed as the baseline regression model, the absolute deviation from the ground truth and the predicted value was used as non-conformity measure in the ICP model and the empirical coverage was used to evaluate the performance of the conformal predictor. The predicted IC50 values of the RF regressor had a mean squared error (MSE) of 14.35 and an R-squared value of 0.84 for 20% of the cell lines included in the test data.

As expected, the model exhibits significant deviation from the ground truth, mainly due to the heterogeneity of cancer types included in the dataset. To address this issue and better capture the variability in the data, we incorporated an inductive conformal predictor into the decision-making process. The ICP framework was employed to quantify the uncertainty of point predictions, providing a range within which the IC50 values are likely to fall. This approach aimed to reduce MSE and to provide a more precise and reliable estimation of the drug response for each cell line. By setting the significant level of 90%, and computing the non-conformity scores for the calibration set (20% of the training data), we found that at least 90% of the examples in the calibration set have a deviation value from the true IC50 value below the 6.77. The cutoff was set to 6.77 as it reflects the challenges faced by the baseline algorithm in accurately predicting new cases. With this value, we constructed the predicted range by adding and subtracting this value to every RF prediction. The conformal model constructed prediction ranges that contained the true IC50 value for 92.6% of the test set. When the same process was repeated with significance levels of 85% and 95%, the model achieved coverage rates of 90.6% for an a-Quantile of 6.63% and 95.6% for an a-Quantile of 6.98, as shown in Table 2. These results highlight the scalability of CP and its ability to meet user-defined coverage levels. Additionally, the a-Quantile in each case defined the size of the prediction intervals, with larger ranges corresponding to higher desired coverage levels and smaller ranges to lower ones.

Overall, the regression conformal model effectively mitigated the inaccuracies of the baseline predictions by replacing individual point estimates with prediction intervals that achieve 92.6% coverage. This improvement is particularly significant in clinical settings, where constraining the index value to a high probability interval provides more actionable information than a single estimate with substantial potential deviation. Furthermore, instances where the true value falls outside the prediction interval serve as important indicators for further investigation, alerting experts for unusual cases that may require additional scrutiny.

4 Discussion

As AI is increasingly adopted into real-world problems the trustworthiness of ML applications in clinical environments is progressively acknowledged. However, denying taking a prediction risk when confronted with unusual cases is still not part of the mainstream procedures when building a model. CP is a powerful tool for estimating uncertainties as it combines favorable features such as i.i.d. assumption, the model-agnostic mode of application, and the adjustable prediction regions. CP addresses reliability concerns that often arise when dealing with imbalanced datasets, insufficient conditional coverage, and domain shifts (Mehrtens et al., 2023). Particularly in the genomics era, CP can overcome domain shifts caused by overlooking the heterogeneity introduced during data acquisition processes or data themselves, e.g., differences in the prevalence of a phenotype across populations. Coupled with larger or new representative calibration datasets under domain shift, CP provides adequate flexibility to keep coverage guarantees.

Another important feature is that CP can effectively lie on the top of both Deep Learning (DL) and simpler ML models. The fundamental basis is that CP helps to quantify and communicate the model's uncertainty effectively depending on the underlying model's predictions. Traditional ML models typically deal with lower-dimensional features and simpler decision boundaries. These models typically provide clear decision rules or margins, which CP can straightforwardly translate into probabilistic measures of uncertainty. In contrast, DL operates on high-dimensional spaces with complex decision boundaries, capable of capturing intricate patterns and relationships in the data, which CP can use to generate more detailed and refined prediction intervals. The complexity of DL models allows CP to handle a wider range of applications in DL, such as image processing (Rouzrokh et al., 2024), natural language processing (Randl et al., 2024) graphs or big data models (Norinder and Norinder, 2022; Park, 2022). Thus, CP can adapt to the nature of the underlying model, utilizing the strengths of both traditional ML and DL to enhance the interpretability and trustworthiness of the predictions.

On the other hand, interpretability is another major concern, particularly in complex and black-box deep learning models. CP has been acknowledged for its ability to provide guaranteed prediction sets and intervals that can be easily understood and communicated, offering a clear way to measure uncertainty. In addition, the minimal assumptions about the data distribution enhance interpretability by avoiding strict and probably unrealistic assumptions. However, the extent to which a conformalized prediction is interpretable partially depends on the interpretability of the underlying models themselves. For example, rule-based models, and decision trees offer a straightforward interpretation of their predictions contrary to deep neural networks and non-linear gradient boosting methods.

Although integrating CP into AI models seems compelling, there are a few limitations to be considered. Distribution-free uncertainty quantification methods, such as CP, are gaining interest among researchers due to their ability to provide reliable uncertainty estimates without assuming specific data distributions. CP ensures that, on average, can cover the correct class with a certain probability (marginal coverage). However, CP cannot provide guarantees for individual instances or structured subgroups of the data (conditional coverage). Practically, a conformalized model with a 90% marginal coverage guarantee ensures that the predictive set covers the correct class with 90% probability on average. This does not mean that each prediction covers the actual class with 90% probability for each individual instance or subgroup of the data. This limitation is crucial especially when dealing with specific subsets of data of particular interest, such as rare disease cases or minority classes in classification tasks. In such cases, researchers must be cautious when interpreting predictions, especially in scenarios where precise classification for individual instances is crucial.

In addition, challenges such as class imbalance, variance, and distribution shifts between training and validation data must be examined. These issues are mainly resolved by recalibrating the data using various combinations of attributes and classes with new data or by the adjustment of the existing calibration dataset with weights. Still, obtaining new and especially rare data to train the model with, in real-world scenarios can be challenging.

A reasonable question is how informative a conformal classifier can be in a binary classification setting when an uncertain prediction contains all the possible labeling options. Krstajic et al. question the utility of CP frameworks in binary classification scenarios (Krstajic, 2021). They reasonably wonder why someone should choose CP when a good binary classification model is built and how is it possible to include as correct coverage the predictions in which CP identifies both classes. In this work, we sought to highlight another aspect that is related to the detection of erroneous cases of the underlying model. In high-risk genomic medicine predictions, when specialists want to rely on the predictions of an ML model it is important to give them all the possible views of these predictions. For example, relying on a good binary classification model without any other guarantee of the resulting prediction may be a deterrent to incorporating such models into clinical decision-making. As we proved in the applications of this study, CP managed to detect the erroneous predictions of the underlying algorithm and classify them as uncertain cases. In clinical terms, these cases are translated as difficult to classify and consequently, the decision is risky to be taken by the ML model. In these cases, the contribution of an expert is necessary to avoid any misconduct.

In our point of view, the behavior of the conformal predictors can be a good step forward in bridging the trust between the medical community and the predictive modelling applications, since the latter can work side by side with the experts in the clinical decision-making process as a powerful and informative tool leaving the final decision to be deployed by experts in ambiguous cases. Additionally, a singleton prediction mathematically guarantees a safe decision with high confidence.

Overall, while conformal prediction offers valuable insights and uncertainty estimates for high-stake decision-making processes, it comes with several limitations and challenges. Ensuring the reliability of a prediction requires addressing issues such as distributional shifts in feature variables and labels, as well as the availability of representative calibration data. The need for a separate calibration set, may lead to data inefficiency by reducing the data available for training. This constraint can be particularly challenging in scenarios where data is limited or costly to obtain. Cross-validation-based CP or integrating calibration within the training phase can help optimize data usage, ensuring that model performance is not significantly compromised. Another critical challenge lies in selecting an appropriate confidence level, as it usually requires domain expertise and directly influences the practical utility of the prediction sets. If the confidence level is too low, the resulting prediction sets may be overly narrow, potentially excluding the correct outcomes. Conversely, an excessively high confidence level can lead to overly broad intervals, reducing their interpretability and practical usefulness. To address this, adaptive techniques such as empirical tuning based on validation performance or automated selection using Bayesian optimization can be employed. These methods enable dynamic confidence level adjustments, improving both model interpretability and decision-making accuracy. Although solutions such as recalibration and careful dataset management exist, they may not always be feasible, particularly in settings with limited data availability or where rare conditions are involved. Moreover, the effectiveness of conformal prediction depends on the quality and representativeness of the training data. Insufficient or biased datasets can lead to unreliable predictions, especially in cases of class imbalance or rare events. Strategies like data augmentation, synthetic data generation, and active learning can help mitigate these limitations by enhancing model robustness. Despite these challenges, the interpretability and robustness of conformal prediction make it a promising tool in domains such as healthcare, where the consequences of incorrect decisions can have a life-threatening impact and the ethical use of the models is mandatory. In-depth research and practical applications will be essential to address these challenges and to fully leverage the potential of conformal prediction in real-world scenarios. It is anticipated that as ML and genomic medicine are progressively infiltrating healthcare environments, CP will support more sophisticated approaches and enhance the range of uncertainty-informed multi-omics applications in clinical environments.

Author contributions

CP: Methodology, Writing–original draft, Writing–review and editing, Formal Analysis, Software, Visualization. KK: Methodology, Writing–review and editing, Conceptualization, Validation. PN: Investigation, Validation, Writing–review and editing. IC: Investigation, Supervision, Validation, Writing–review and editing. AM: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2025.1507448/full#supplementary-material

References

Abad, J., Bhatt, U., Weller, A., and Cherubin, G. (2022). Approximating full conformal prediction at scale via influence functions. doi:10.48550/arXiv.2202.01315

Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B., and Goldenberg, A. (2020). Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precis. Oncol.* 4, 19. doi:10.1038/s41698-020-0122-1

Ahmed P, K., and Acharjya, D. P. (2020). A hybrid scheme for heart disease diagnosis using rough set and cuckoo search technique. *J. Med. Syst.* 44, 27. doi:10.1007/s10916-019-1497-9

Alvarsson, J., McShane, S. A., Norinder, U., and Spjuth, O. (2021). Predicting with confidence: using conformal prediction in drug discovery. *J. Pharm. Sci.* 110, 42–49. doi:10.1016/j.xphs.2020.09.055

Angelopoulos, A. N., and Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. doi:10.48550/arXiv.2107.07511

Augustin, T., Coolen, F. P., De Cooman, G., and Troffaes, M. C. (2014). *Introduction to imprecise probabilities*. John Wiley and Sons, 591.

Barlow-Stewart, K. (2018). The certainty of uncertainty in genomic medicine: managing the challenge. *J. Healthc. Commun.* 3, 1–4. doi:10.4172/2472-1654.100147

Begoli, E., Bhattacharya, T., and Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nat. Mach. Intell.* 1, 20–23. doi:10.1038/s42256-018-0004-1

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* 112, 859–877. doi:10.1080/01621459.2017.1285773

Boger, R. S., Chithrananda, S., Angelopoulos, A. N., Yoon, P. H., Jordan, M. I., and Doudna, J. A. (2024). Functional protein mining with conformal guarantees. *Nat. Commun.* 16, 85. doi:10.1038/s41467-024-55676-y

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22, e49–e57. doi:10.1093/bioinformatics/btl242

Bosc, N., Atkinson, F., Felix, E., Gaulton, A., Hersey, A., and Leach, A. R. (2019). Large scale comparison of qsar and conformal prediction methods and their applications in drug discovery. *J. cheminformatics* 11, 4–16. doi:10.1186/s13321-018-0325-4

Boström, H., and Johansson, U. (2020). "Mondrian conformal regressors," in *Conformal and probabilistic prediction and applications* Cambridge, United Kingdom: Proceedings of Machine Learning Research (PMLR), 114–133.

Boström, H., Johansson, U., and Löfström, T. (2021). "Mondrian conformal predictive distributions," in *Conformal and probabilistic Prediction and applications (PMLR)*, 24–38.

Brittain, H. K., Scott, R., and Thomas, E. (2017). The rise of the genome and personalised medicine. *Clin. Med.* 17, 545–551. doi:10.7861/clinmedicine.17-6-545

Cai, F., Ozdagli, A. I., Potteiger, N., and Koutsoukos, X. (2021). "Inductive conformal out-of-distribution detection based on adversarial autoencoders," in *2021 IEEE international conference on omni-layer intelligent systems (COINS)* IEEE, 1–6.

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Chua, M., Kim, D., Choi, J., Lee, N. G., Deshpande, V., Schwab, J., et al. (2023). Tackling prediction uncertainty in machine learning for healthcare. *Nat. Biomed. Eng.* 7, 711–718. doi:10.1038/s41551-022-00988-x

Clifford, J. N., Høie, M. H., Deleuran, S., Peters, B., Nielsen, M., and Marcatili, P. (2022). Bepipred-3.0: improved b-cell epitope prediction using protein language models. *Protein Sci.* 31, e4497. doi:10.1002/pro.4497

Fagerholm, U., Alvarsson, J., Hellberg, S., and Spjuth, O. (2022). Validation of predicted conformal intervals for prediction of human clinical pharmacokinetics. *bioRxiv*, 2022–2111. doi:10.1101/2022.11.10.515917

Fannjiang, C., Bates, S., Angelopoulos, A. N., Listgarten, J., and Jordan, M. I. (2022). Conformal prediction under feedback covariate shift for biomolecular design, *Proc. Natl. Acad. Sci. U. S. A.*, 119, e2204569119, doi:10.1073/pnas.2204569119

Forreryd, A., Norinder, U., Lindberg, T., and Lindstedt, M. (2018). Predicting skin sensitizers with confidence—using conformal prediction to determine applicability domain of gard. *Toxicol. Vitro* 48, 179–187. doi:10.1016/j.tiv.2018.01.021

Gammerman, A., and Vovk, V. (2007). Hedging predictions in machine learning. *Comput. J.* 50, 151–163. doi:10.1093/comjnl/bxl065

Geifman, Y., and El-Yaniv, R. (2019). "Selectivenet: a deep neural network with an integrated reject option," in *International conference on machine learning* Cambridge, United Kingdom: Proceedings of Machine Learning Research (PMLR), 2151–2159.

Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., et al. (2012). Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* 3, 811. doi:10.1038/ncomms1814

Geylan, G. (2021). Training machine learning-based qsar models with conformal prediction on experimental data from dna-encoded chemical libraries.

Giustinelli, P., Manski, C. F., and Molinari, F. (2022). Precise or imprecise probabilities? evidence from survey response related to late-onset dementia. *J. Eur. Econ. Assoc.* 20, 187–221. doi:10.1093/jeea/jvab023

Goodswen, S. J., Kennedy, P. J., and Ellis, J. T. (2023). A guide to current methodology and usage of reverse vaccinology towards *in silico* vaccine discovery. *FEMS Microbiol. Rev.* 47, fuad004. doi:10.1093/femsre/fuad004

Hamet, P., and Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism* 69, S36–S40. doi:10.1016/j.metabol.2017.01.011

Hammersley, J. (2013). *Monte Carlo methods*. Springer Science and Business Media.

Hernandez-Hernandez, S., Guo, Q., and Ballester, P. (2024). Conformal prediction of molecule-induced cancer cell growth inhibition challenged by strong distribution shifts. *bioRxiv*. doi:10.1101/2024.03.15.585269

Ianevski, A., Nader, K., Bulava, D., Giri, A. K., Ruokoranta, T., Kuusanmaki, H., et al. (2023). Single-cell transcriptomes identify patient-tailored therapies for selective co-inhibition of cancer clones. *bioRxiv* (06), 2023. doi:10.1038/s41467-024-52980-5

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with alphafold. *nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Kapuria, S., Minot, P., Kapusta, A., Ikoma, N., and Alambeigi, F. (2024). A novel dual layer cascade reliability framework for an informed and intuitive clinician-ai interaction in diagnosis of colorectal cancer polyps. *IEEE J. Biomed. Health Inf.* 28, 2326–2337. doi:10.1109/jbhi.2024.3350082

Karaboga, D., and Kaya, E. (2019). Adaptive network based fuzzy inference system (anfis) training approaches: a comprehensive survey. *Artif. Intell. Rev.* 52, 2263–2293. doi:10.1007/s10462-017-9610-2

Kardamiliotis, K., Karanatsiou, E., Aslanidou, I., Stergiou, E., Vizirianakis, I. S., and Malousi, A. (2022). Unraveling drug response from pharmacogenomic data to advance systems pharmacology decisions in tumor therapeutics. *Future Pharmacol.* 2, 31–44. doi:10.3390/futurepharmacol2010003

Kasa, K., and Taylor, G. W. (2023). Empirically validating conformal prediction on modern vision architectures under distribution shift and long-tailed data.

Kavvas, E. S., Yang, L., Monk, J. M., Heckmann, D., and Palsson, B. O. (2020). A biochemically-interpretable machine learning classifier for microbial gwas. *Nat. Commun.* 11, 2580. doi:10.1038/s41467-020-16310-9

Kim, J. I., Maguire, F., Tsang, K. K., Gouliouris, T., Peacock, S. J., McAllister, T. A., et al. (2022). Machine learning for antimicrobial resistance prediction: current practice, limitations, and clinical perspective. *Clin. Microbiol. Rev.* 35, e0017921–21. doi:10.1128/cmr.00179-21

Krstajic, D. (2021). Critical assessment of conformal prediction methods applied in binary classification settings. *J. Chem. Inf. Model.* 61, 4823–4826. doi:10.1021/acs.jcim.1c00549

Lacy, S. E., Barrans, S. L., Beer, P. A., Painter, D., Smith, A. G., Roman, E., et al. (2020). Targeted sequencing in dlbcl, molecular subtypes, and outcomes: a haematological malignancy research network report. *Blood, J. Am. Soc. Hematol.* 135, 1759–1771. doi:10.1182/blood.2019003535

Laghuvarapu, S., Lin, Z., and Sun, J. (2024). Codrug: conformal drug property prediction with density estimation under covariate shift. *Adv. Neural Inf. Process. Syst.* 36. doi:10.48550/arXiv.2310.12033

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. neural Inf. Process. Syst.* 30. doi:10.48550/arXiv.1612.01474

Lenhof, K., Eckhart, L., Rolli, L.-M., Volkamer, A., and Lenhof, H.-P. (2023). Reliable anti-cancer drug sensitivity prediction and prioritization.

Lenhof, K., Eckhart, L., Rolli, L.-M., Volkamer, A., and Lenhof, H.-P. (2024). Reliable anti-cancer drug sensitivity prediction and prioritization. *Sci. Rep.* 14, 12303. doi:10.1038/s41598-024-62956-6

Lu, C., Angelopoulos, A. N., and Pomerantz, S. (2022a). "Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets," in *International conference on medical image computing and computer-assisted intervention* (Springer), 545–554.

Lu, C., Lemay, A., Chang, K., Höbel, K., and Kalpathy-Cramer, J. (2022b). Fair conformal predictors for applications in medical imaging. *Proc. AAAI Conf. Artif. Intell.* 36, 12008–12016. doi:10.1609/aaai.v36i11.21459

McShane, S. A., Norinder, U., Alvarsson, J., Ahlberg, E., Carlsson, L., and Spjuth, O. (2023). Cpsign-conformal prediction for cheminformatics modeling. *J. Cheminform.* 16, 75. doi:10.1186/s13321-024-00870-9

Mehrtens, H., Bucher, T., and Brinker, T. J. (2023). "Pitfalls of conformal predictions for medical image classification," in *International workshop on uncertainty for safe utilization of machine learning in medical imaging* (Springer), 198–207.

Mesko, B., Poliska, S., Váncsa, A., Szekanecz, Z., Palatka, K., Hollo, Z., et al. (2013). Peripheral blood derived gene panels predict response to infliximab in rheumatoid arthritis and crohn's disease. *Genome Med.* 5, 59–10. doi:10.1186/gm463

Mesko, B., Poliskal, S., Szegedi, A., Szekanecz, Z., Palatka, K., Papp, M., et al. (2010). Peripheral blood gene expression patterns discriminate among chronic inflammatory diseases and healthy controls and identify novel targets. *BMC Med. genomics* 3, 15–13. doi:10.1186/1755-8794-3-15

Millar, A. S., Arnn, J., Himes, S., and Facelli, J. C. (2024). "Uncertainty in breast cancer risk prediction: a conformal prediction study of race stratification," in *Studies in health technology and informatics* 310, 991–995. doi:10.3233/SHTI231113

Murray, C. J., Ikuta, K. S., Sharara, F., Swetschinski, L., Aguilar, G. R., Gray, A., et al. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *lancet* 399, 629–655. doi:10.1016/s0140-6736(21)02724-0

Nguyen, M., Brettin, T., Long, S. W., Musser, J. M., Olsen, R. J., Olson, R., et al. (2018). Developing an *in silico* minimum inhibitory concentration panel test for klebsiella pneumoniae. *Sci. Rep.* 8, 421. doi:10.1038/s41598-017-18972-w

Norinder, U., and Norinder, P. (2022). Predicting amazon customer reviews with deep confidence using deep learning and conformal prediction. *J. Manag. Anal.* 9, 1–16. doi:10.1080/23270012.2022.2031324

Olson, N. D., Wagner, J., McDaniel, J., Stephens, S. H., Westreich, S. T., Prasanna, A. G., et al. (2022). Precisionfda truth challenge v2: calling variants from short and long reads in difficult-to-map regions. *Cell genomics* 2, 100129. doi:10.1016/j.xgen.2022.100129

Olsson, H., Kartasalo, K., Mulliqi, N., Capuccini, M., Ruusuvuori, P., Samaratunga, H., et al. (2022). Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. *Nat. Commun.* 13, 7761. doi:10.1038/s41467-022-34945-8

Ong, E., Wang, H., Wong, M. U., Seetharaman, M., Valdez, N., and He, Y. (2020). Vaxign-ml: supervised machine learning reverse vaccinology model for improved prediction of bacterial protective antigens. *Bioinformatics* 36, 3185–3191. doi:10.1093/bioinformatics/btaa119

Painter, D., Barrans, S., Lacy, S., Smith, A., Crouch, S., Westhead, D., et al. (2019). Cell-of-origin in diffuse large b-cell lymphoma: findings from the UK's population-based haematological malignancy research network. *Br. J. Haematol.* 185, 781–784. doi:10.1111/bjh.15619

Papadopoulos, H. (2008). "Inductive conformal prediction: theory and application to neural networks," in *Tools in artificial intelligence (Citeseer)*.

Park, H. (2022). Providing post-hoc explanation for node representation learning models through inductive conformal predictions. *IEEE Access* 11, 1202–1212. doi:10.1109/access.2022.3233036

Park, S., Cohen, K. M., and Simeone, O. (2023). Few-shot calibration of set predictors via meta-learned cross-validation-based conformal prediction. *IEEE Trans. Pattern Analysis Mach. Intell.* 46, 280–291. doi:10.1109/tpami.2023.3327300

Pawlak, Z. (1998). Rough set theory and its applications to data analysis. *Cybern. and Syst.* 29, 661–688. doi:10.1080/019697298125470

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. pattern analysis Mach. Intell.* 27, 1226–1238. doi:10.1109/tpami.2005.159

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., et al. (2018). A universal snp and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* 36, 983–987. doi:10.1038/nbt.4235

Randl, K., Pavlopoulos, J., Henriksson, A., and Lindgren, T. (2024). Cicle: conformal in-context learning for largescale multi-class food risk classification. 7695–7715. doi:10.18653/v1/2024.findings-acl.459

Rouzrokh, P., Faghani, S., Gamble, C. U., Shariatnia, M., and Erickson, B. J. (2024). Conflare: conformal large language model retrieval. *arXiv Prepr. arXiv:2404.04287*. doi:10.48550/arXiv.2404.04287

Sha, C., Barrans, S., Care, M. A., Cunningham, D., Tooze, R. M., Jack, A., et al. (2015). Transferring genomics to the clinic: distinguishing burkitt and diffuse large b cell lymphomas. *Genome Med.* 7, 64–13. doi:10.1186/s13073-015-0187-6

Sha, C., Barrans, S., Cucco, F., Bentley, M. A., Care, M. A., Cummin, T., et al. (2019). Molecular high-grade b-cell lymphoma: defining a poor-risk group that requires different approaches to therapy. *J. Clin. Oncol.* 37, 202–212. doi:10.1200/jco.18.01314

Shafer, G., and Vovk, V. (2008). A tutorial on conformal prediction. *J. Mach. Learn. Res.* 9. doi:10.48550/arXiv.0706.3188

Stacke, K., Eilertsen, G., Unger, J., and Lundström, C. (2020). Measuring domain shift for deep learning in histopathology. *IEEE J. Biomed. health Inf.* 25, 325–336. doi:10.1109/jbhi.2020.3032060

Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.* 8.

Sun, E. D., Ma, R., Navarro Negredo, P., Brunet, A., and Zou, J. (2024). Tissue: uncertainty-calibrated prediction of single-cell spatial transcriptomics improves downstream analyses. *Nat. Methods* 21, 444–454. doi:10.1038/s41592-024-02184-y

Sun, J., Wang, Y., Folkersen, L., Borné, Y., Amlien, I., Buil, A., et al. (2021). Translating polygenic risk scores for clinical use by estimating the confidence bounds of risk prediction. *Nat. Commun.* 12, 5276. doi:10.1038/s41467-021-25014-7

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Adv. neural Inf. Process. Syst.* 32. doi:10.48550/arXiv.1904.06019

Toccaceli, P., Nouretdinov, I., and Gammerman, A. (2017). Conformal prediction of biological activity of chemical compounds. *Ann. Math. Artif. Intell.* 81, 105–123. doi:10.1007/s10472-017-9556-8

Trygoniaris, D., Korda, A., Paraskeva, A., Dushku, E., Tzimagiorgis, G., Yiangou, M., et al. (2024). Vaccinedesigner: a web-based tool for streamlined multi-epitope vaccine design. *bioRxiv*, 2024–2103. doi:10.1101/2024.03.20.585850

Vazquez, J., and Facelli, J. C. (2022). Conformal prediction in clinical medical sciences. *J. Healthc. Inf. Res.* 6, 241–252. doi:10.1007/s41666-021-00113-8

Vidhya, K., and Shanmugalakshmi, R. (2020). Modified adaptive neuro-fuzzy inference system (m-anfis) based multi-disease analysis of healthcare big data. *J. Supercomput.* 76, 8657–8678. doi:10.1007/s11227-019-03132-w

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, 29. Springer.

Wieslander, H., Harrison, P. J., Skogberg, G., Jackson, S., Fridén, M., Karlsson, J., et al. (2020). Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images. *IEEE J. Biomed. health Inf.* 25, 371–380. doi:10.1109/jbhi.2020.2996300

Xiao, F. (2020). Generalization of dempster–shafer theory: a complex mass function. *Appl. Intell.* 50, 3266–3275. doi:10.1007/s10489-019-01617-y

Xu, Y., Liaw, A., Sheridan, R. P., and Svetnik, V. (2023). Development and evaluation of conformal prediction methods for qsar. *ACS Omega* 9 (27), 29478–29490. doi:10.1021/acsomega.4c02017

Zhang, J., Norinder, U., and Svensson, F. (2021). Deep learning-based conformal prediction of toxicity. *J. Chem. Inf. Model.* 61, 2648–2657. doi:10.1021/acs.jcim.1c00208