Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Conformal inference for cell type prediction leveraging the cell ontology

Daniela Corbetta, Livio Finos, Ludwig Geistlinger, Davide Risso

Lorenzo Bernardi e la Statistica Sociale — Poster Session: 18 October 2024

Introduction

Cell type annotation–the identification and classification of cell types within a tissue–is essential for **single-cell RNA sequencing** (scRNA-seq) data analysis. This process typically involves training a model on a labeled dataset and using it to predict cell types in new, unlabeled dataset. However, **uncertainty** is often ignored, with only the most probable label assigned to each cell.

To better reflect uncertainty, **prediction sets** can be used instead of single-label predictions.

Integrating the **Cell Ontology**, a directed acyclic graph that organizes cell types hierarchically, further enhances this approach by incorporating biological context. This suggests a **hierarchical notion of uncertainty**: when unsure of the point prediction, provide a broader classification by returning one of its **ancestors** in the ontology.

Conformal inference (Vovk et al., 2005), a statistical approach for generating valid prediction sets independently of the model or data distribution, is ideal here. However, its application in graph-structured problems remains limited.

Goal of the project: Develop a method combining conformal inference with directed acyclic graphs for structure-aligned prediction sets and compare it to split conformal inference (Papadopoulus et al., 2002).

Methods

Let $(X_1,Y_1),\dots,(X_m,Y_m)$ be a set of i.i.d observations, where $X_i \in \mathbb{R}^p$ is a p-dimensional vector of explanatory variables and Y_i is a categorical response variable with K possible classes. Y_i, $i=1,\dots,m$ is known.
Split $(X_1,Y_1),\dots,(X_m,Y_m)$ into two subsets:

- the **calibration set**, $(X_1,Y_1),\dots,(X_n,Y_n)$;
- the **training set**, $(X_{n+1},Y_{n+1}),\dots,(X_m,Y_m)$, used to build a classification model, $\hat{f}$, which estimates class probabilities $\hat{f}(x) \in [0,1]^K$.

Objective: use $\hat{f}$ and the calibration data to construct a prediction set $C(X_{new})$ for a new, unlabelled observation X_{new}, such that

$$P\left(Y_{new} \in C(X_{new})\right) \geq 1-\alpha$$

for a user-chosen error rate α. Methods based on conformal inference are **distribution-free** and provide **finite-sample validity**, assuming that the calibration data are **exchangeable** with the new data.

Split conformal inference

The algorithm of split conformal inference is as follows

```
Algorithm 1 Split conformal inference
Input: Calibration set data $(X_1,Y_1),\dots,(X_n,Y_n)$ and classifier $\hat{f}(x)$
Return: Prediction sets $C(X_{new})$ for test data
for all $(X_i,Y_i)$, $i=1,\dots,n$ do
    Compute the conformal score: $s_i = 1-\hat{f}(X_i)_{Y_i}$
end for
Compute $\hat{q}$, the $\lceil(1-\alpha)(n+1)\rceil/n$ empirical quantile of the conformal scores $\{s_i\}_{i=1}^n$
Form the prediction set: $C(X_{new}) = \{y : \hat{f}(X_{new})_y \geq 1-\hat{q}\}$
```

Graph-based method

- $\hat{y}(x)$: predicted class
- $\mathcal{P}(v)$: set of descendant nodes of v that are leaves of the graph
- $\mathcal{A}(v)$ set of ancestor nodes of v

The algorithm of our graph-based method is as follows

```
Algorithm 2 Graph-based method
Input: Calibration set data $(X_1,Y_1),\dots,(X_n,Y_n)$, a grid of $\lambda$ values $\{\lambda_1,\dots,\lambda_r\}$, and classifier $\hat{f}(x)$
Return: Prediction sets $C(X_{new})$ for test data
for all $\lambda_j,\ j=1,\dots,r$ do
    for all $(X_i,Y_i)$, $i=1,\dots,n$ do
        for all nodes $v$ do
            Compute the scores $g(v,X_i)=\sum_{k\in\mathcal{P}(v)}\hat{f}(X_i)_k$
        end for
        Form the prediction set:
            $C_{\lambda_j}(X_i)=\mathcal{P}(v)\cup\{\mathcal{P}(a): a\in\mathcal{A}(\hat{y}(X_i)),\ g(a,X_i)\leq\lambda_j\}$
        where $v: v\in\mathcal{A}(\hat{y}(X_i)),\ g(v,X_i)\geq\lambda_j,\ v=\arg\min_{u:g(u,X_i)\geq\lambda_j} g(u,X_i)$
        Compute $R_i(\lambda_j)=\mathbf{1}(Y_i\notin C_{\lambda_j}(X_i))$
    end for
    Compute $\hat{R}(\lambda_j)=\frac{1}{n}\sum_{i=1}^n R_i(\lambda_j)$
end for
Set $\hat{\lambda}=\inf\{\lambda:\hat{R}(\lambda)\leq\alpha-(1-\alpha)/n\}$
Form the prediction set for test data:
    $C_{\hat{\lambda}}(X_{new})=\mathcal{P}(v)\cup\{\mathcal{P}(a): a\in\mathcal{A}(\hat{y}(X_{new})),\ g(a,X_{new})\leq\hat{\lambda}\}$
where $v: v\in\mathcal{A}(\hat{y}(X_{new})),\ g(v,X_{new})\geq\hat{\lambda},\ v=\arg\min_{u:g(u,X_{new})\geq\hat{\lambda}} g(u,X_{new})$
```

Data

- Dataset of scRNA-seq data from COVID-19 patients
- Test set: cells of a new patient (1762)
- Reference set: already annotated cells from other patients (5616)

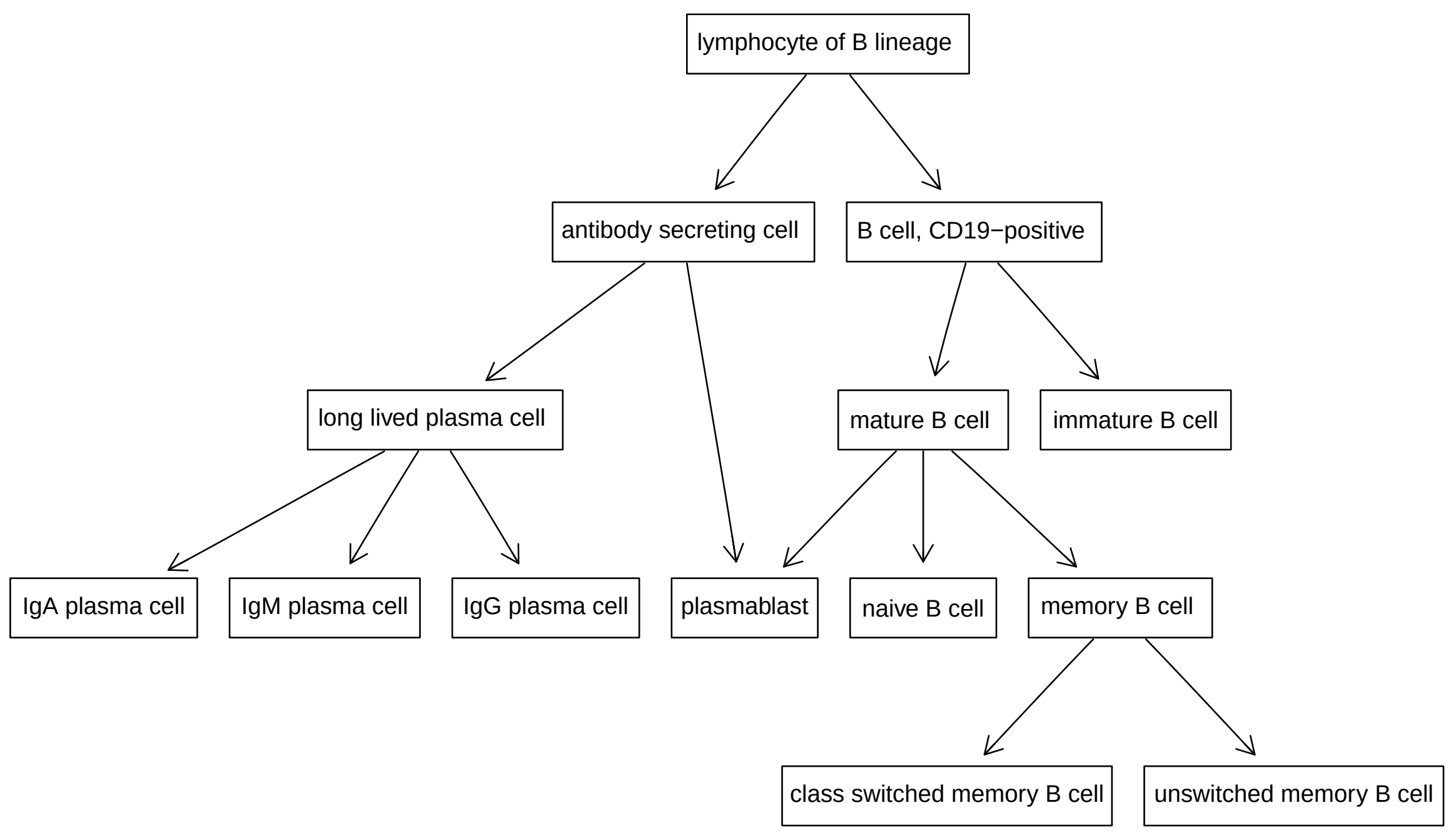

Figure 1. DAG deriving from the cell ontology for the cell types in the COVID dataset.

Results

Split conformal sets and graph-based sets have been compared considering

1. **empirical coverage**
2. **average size** of the resulting prediction sets
3. **homogeneity** of elements within the sets

	Emp. cvg	Avg. size	Avg. dist
Split conformal	0.93	3.39	3.46
Graph-based method	0.92	4.38	2.61

Table 1. Comparison of split conformal and graph-based results for $\alpha = 0.1$.

Graph-based prediction sets allow to gain biological insight:

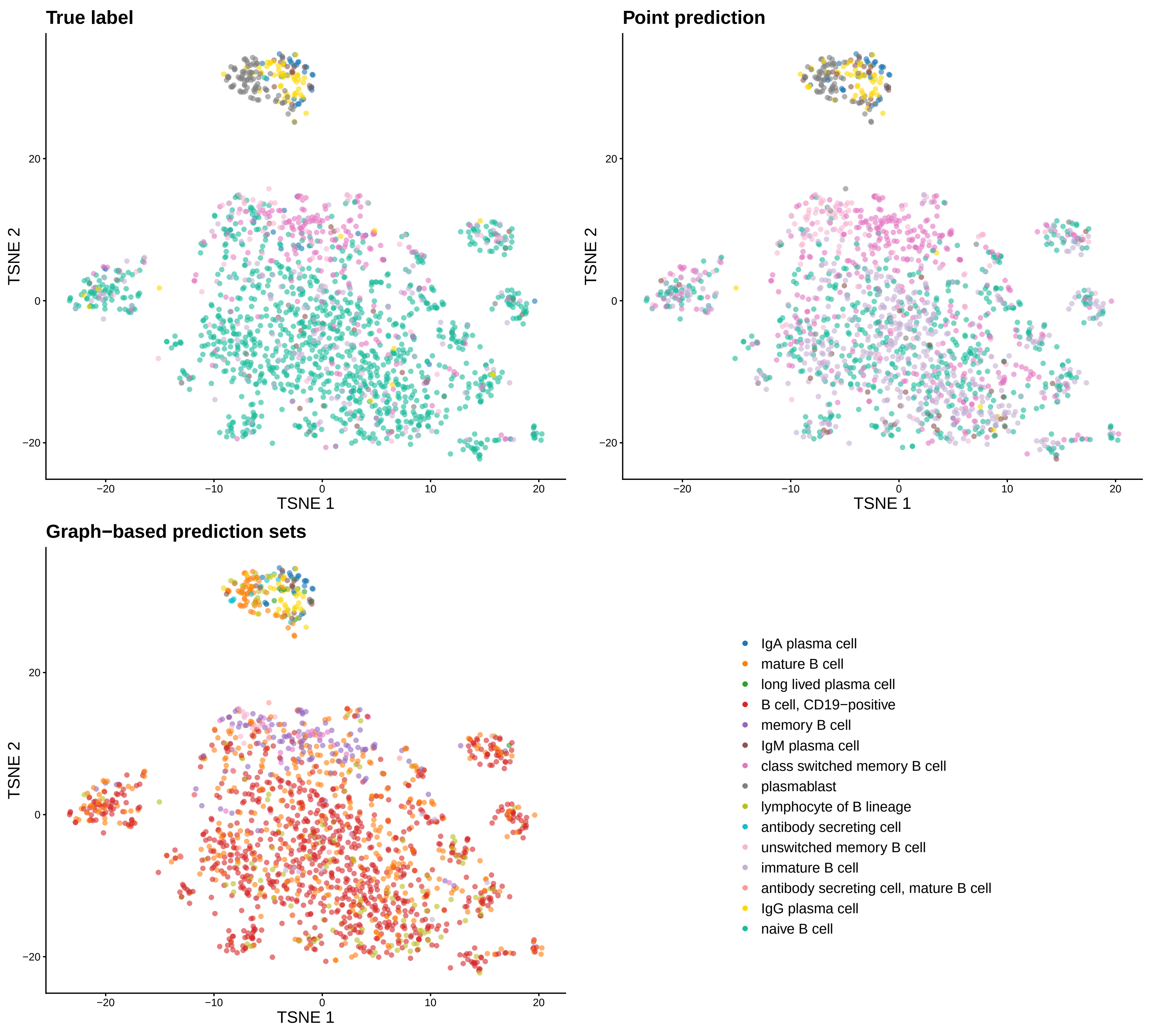

Figure 2. t-SNE representations of cells in the test set. Cells are colored according to their true label (top left), point prediction (top right) and graph-based set (bottom)

Contact information

Daniela Corbetta, PhD student
Department of Statistical Sciences, University of Padua
daniela.corbetta@phd.unipd.it

References

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. (2022). Conformal risk control. arXiv preprint arXiv:2208.02814.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, 345 – 356. Springer.