Selection by Prediction: Prediction-Assisted Screening with Conformal p-values

Ying Jin [1] Emmanuel J. Candès [1,2]

[1]Department of Statistics, Stanford University [2]Department of Mathematics, Stanford University

The Selection Issue in Predictive Inference

Marginally valid prediction sets (i.e., conformal prediction)

- Training data $\{X_i, Y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, test sample $(X_{n+1}, Y_{n+1}) \sim \mathbb{P}$.
- Marginally valid $(1-\alpha) = .9$ prediction intervals [3] $\widehat{C}(\cdot)$:

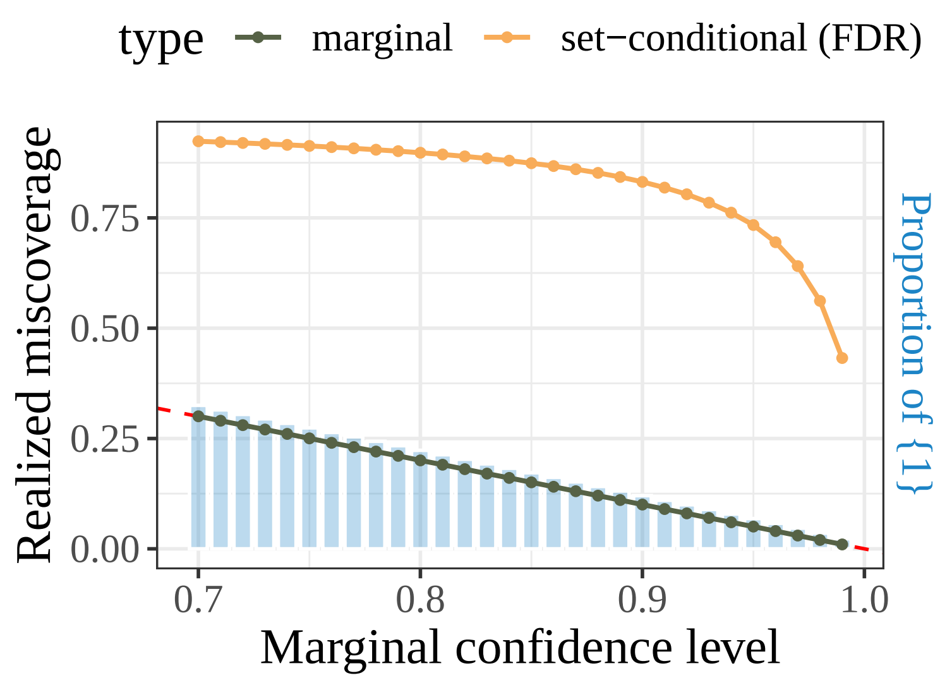
$$\mathbb{P}\big(Y_{n+j} \in \widehat{C}(X_{n+1})\big) \geq 0.9$$

What would practitioner do for a batch of test samples?

- Test samples $(X_{n+j}, Y_{n+j}) \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, $\{Y_{n+j}\}_{j=1}^m$ unobserved.
- Construct and **inspect** .9 prediction intervals $\widehat{C}(X_{n+j})$
- Look at seemingly promising ones
- Does higher prediction intervals mean more promising outcomes?

Calibrated prediction may still be overly confident

- Binary Y in drug discovery dataset, interested in $Y=1$
- $(1-\alpha)$ prediction sets of the form $\{0\}, \{1\}, \{0,1\}$
- Strategy: look at those $\widehat{C}(X_{n+j}) = \{1\}$ -- confident prediction for 1!



- If we look at those with $\widehat{C}(X_{n+j}) = \{1\}$ at all different levels α (x-axis on the left)...
- Over-confident (orange)
- Coverage for those $\{1\}$ is $\leq 50\%$ even when we take $\alpha = 0.01$!!
- Marginal coverage $\neq$ Coverage on selected

This work: Screening with FDR control

- Reliable selection of candidates: output a set $\mathcal{R} \subseteq \{1, \dots, m\}$, s.t.

$$\mathrm{FDR} := \mathbb{E}\left[\frac{\sum_{j=1}^m \mathbb{1}\{j \in \mathcal{R}, Y_{n+j} \leq c_j\}}{1 \vee |\mathcal{R}|}\right] \leq q,$$

- **Efficient allocation of resources in later stages of costly investigation.**
- Drug discovery: virtual screening use ML models to search for promising drugs to proceed to later stages (clinical trials, HTS) [2].
  $\Rightarrow$ Most of the prioritized drugs are truly active
- Job recruitment: automatic search of candidates that suit a position, include talent sourcing (reaching out to candidates) and interviewing.
  $\Rightarrow$ Most of the interviewed candidates are qualified
- Counterfactual inference, healthcare, individual treatment effects, . . .

Our Procedure: Conformal BH

- Step 0. Conformal prediction: Train any model $\widehat{\mu}(\cdot)$ on another fold
- Step 1. Construct any score $V(x,y)$ that is monotone in y
- Step 2. Compute **p-values** to quantify confidence in large outcomes

$$p_j = \frac{\sum_{i=1}^n \mathbb{1}\{V_i < \widehat{V}_{n+j}\} + (1 + \sum_{i=1}^n \mathbb{1}\{V_i = \widehat{V}_{n+j}\}) \cdot U_j}{n+1},$$

for $V_i = V(X_i, Y_i)$, $i = 1, \dots, n$, and $\widehat{V}_{n+j} = V(X_{n+j}, c_j)$, $j = 1, \dots, m$

- Step 3. Apply Benjamini-Hochberg [1] procedure to $\{p_j\}_{j=1}^m$

Informal Theorem (J. and Cand`es '22)

If calibration and test data are i.i.d. or exchangeable, then FDR $\leq q$.

Interesting statistical facts about our p-values

- An interpretation: p_j is the critical confidence level α where the one-sided conformal prediction interval $\widehat{C}(X_{n+j}; 1-\alpha)$ hits c_j. Thus, a smaller p_j means greater confidence in $Y_{n+j} > c_j$
- Controls type-I error for **random** hypothesis $H_j\colon Y_{n+j} \leq c_j$:

$$\mathbb{P}(p_j \leq \alpha \text{ and } j \in \mathcal{H}_0) \leq \alpha, \quad \text{for all } \alpha \in [0,1].$$

- Plugged-in **PRDS** property: $(p_1, \dots, p_{j-1}, p_j^*, p_{j+1}, \dots, p_m)$ is PRDS on $p_j^* := [\sum_{i=1}^n \mathbb{1}\{V_i < \widehat{V}_{n+j}\} + (1 + \sum_{i=1}^n \mathbb{1}\{V_i = V_{n+j}\}) \cdot U_j]/(n+1)$.
- Works even for **random variables** c_j (useful in counterfactual inference)

Dealing with Distribution Shifts

- In reality, calibration data may be different from test samples
  - e.g. Different preference for drugs over time
  - e.g. Shift of talent pool for hiring
  - e.g. Demographic differences in patients for healthcare
- An ongoing extension of this method deals with **covariate shift**
- Needs to adjust for covariate shift in constructing valid p_j
- Multiple testing is more difficult & more interesting!

References

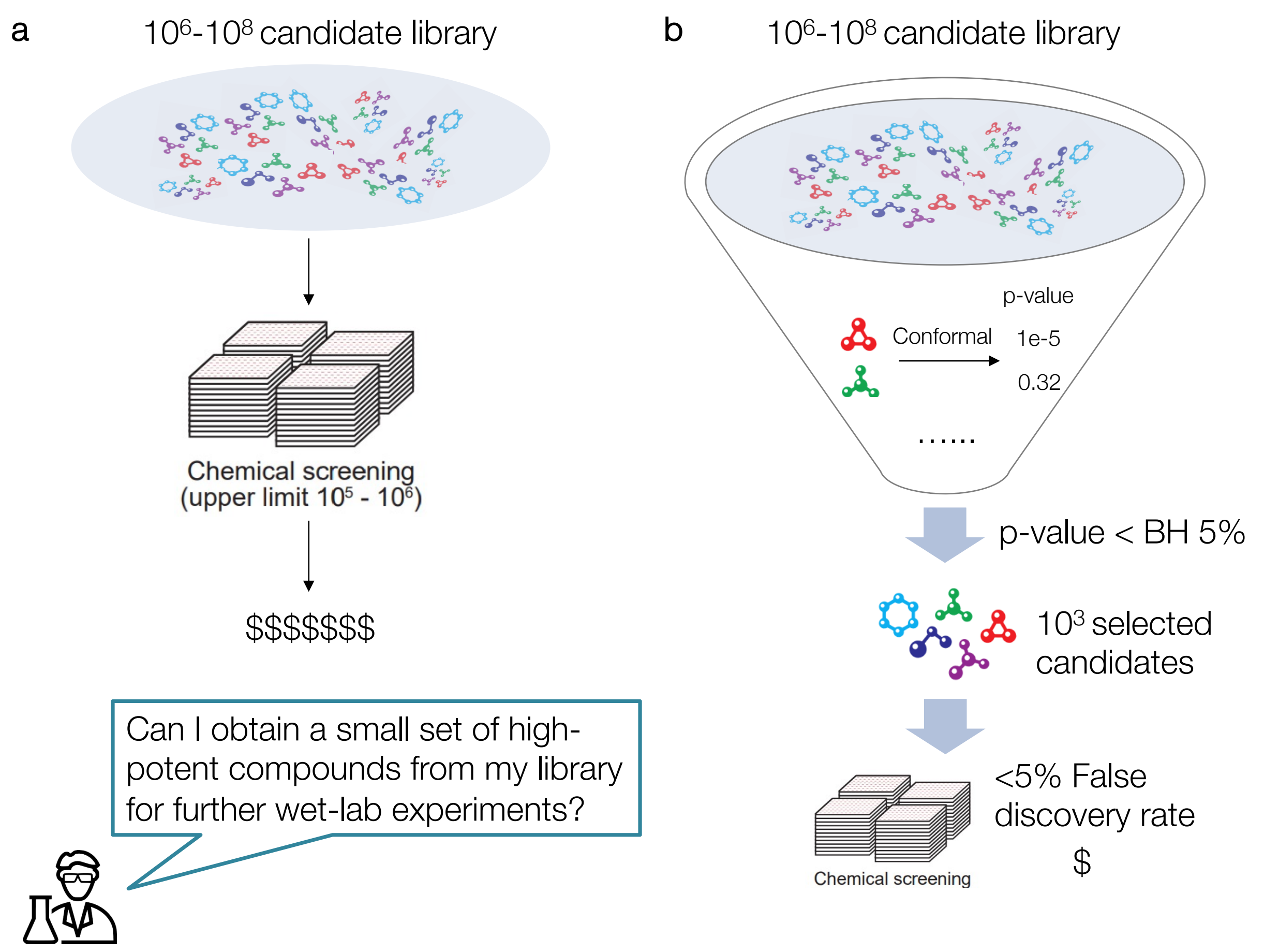
[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289--300, 1995.

[2] Ziwei Huang. *Drug discovery research: new frontiers in the post-genomic era*. John Wiley & Sons, 2007.

[3] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Application to Drug Discovery

- Rigorously guide **ML-driven** search of promising drugs with **FDR control**


a
10^6-10^8 candidate library
Chemical screening
(upper limit 10^5 - 10^6)
$$$$$$$
Can I obtain a small set of high-potent compounds from my library for further wet-lab experiments?
b
10^6-10^8 candidate library
p-value
Conformal
1e-5
0.32
......
p-value < BH 5%
10^3 selected candidates
<5% False discovery rate
$
Chemical screening

Application I: Drug Activity Prediction

Find highly-active drugs to a specific disease target

- **Dataset:** 8K+ drugs for HIV target
- **Overall hit rate:** 3% of them are active to HIV
- **Our goal:** Select a subset s.t. on average, 80% of the selected are active
- **Prediction model:** Deep learning
- **Result:** FDR = 0.196, select 26 drugs on average, 17.4% of active drugs

Application II: Drug-Target Interaction Prediction

Find highly-active drug-target pairs to proceed

- **Dataset:** 18K+ pairs with continuously-valued binding score (regression)
- **Our goal:** Select a subset of pairs s.t. on average, in 80% of pairs, the drug is more active than 80% of training drugs for its target
- **Prediction model:** Deep learning
- **Result:** FDR = 0.194, select 358 pairs, find 8% of qualified pairs
- FDP concentrates tightly around FDR

More forthcoming...