Model-free selective inference with conformal p-values

and its application to drug discovery

Ying Jin

Department of Statistics

Stanford University

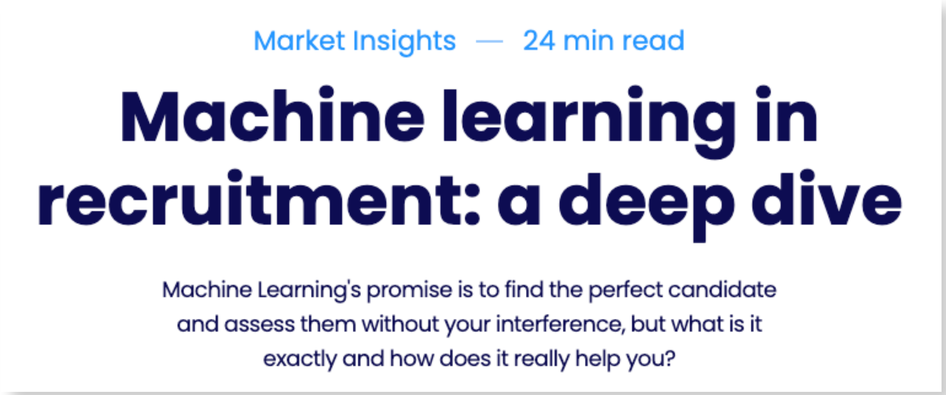
Joint work with Emmanuel Candès

Joint Conference on Statistics and Data Science (JCSDS), Beijing, July 2023

ML prediction assists decision

[VerVoe.com]

[smartdreamers.com]



[HeroHunt.ai]

Job hiring: Who to reach out to?
Who to proceed to interview?

ML prediction assists discovery

[DZone.com]

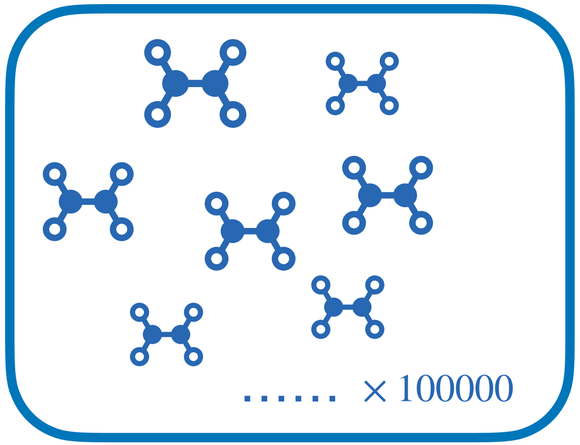
[technologynetworks.com]

Drug discovery: Which molecules/compounds to proceed to physical screening and clinical trials?

Decision and discovery processes

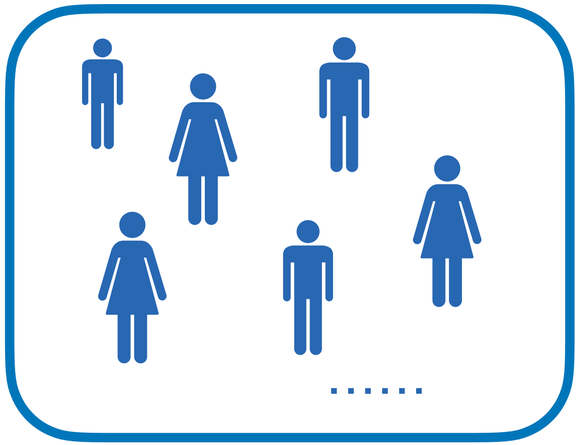
- Find a few interesting cases from a huge pool

Disease (COVID)



Candidate drugs

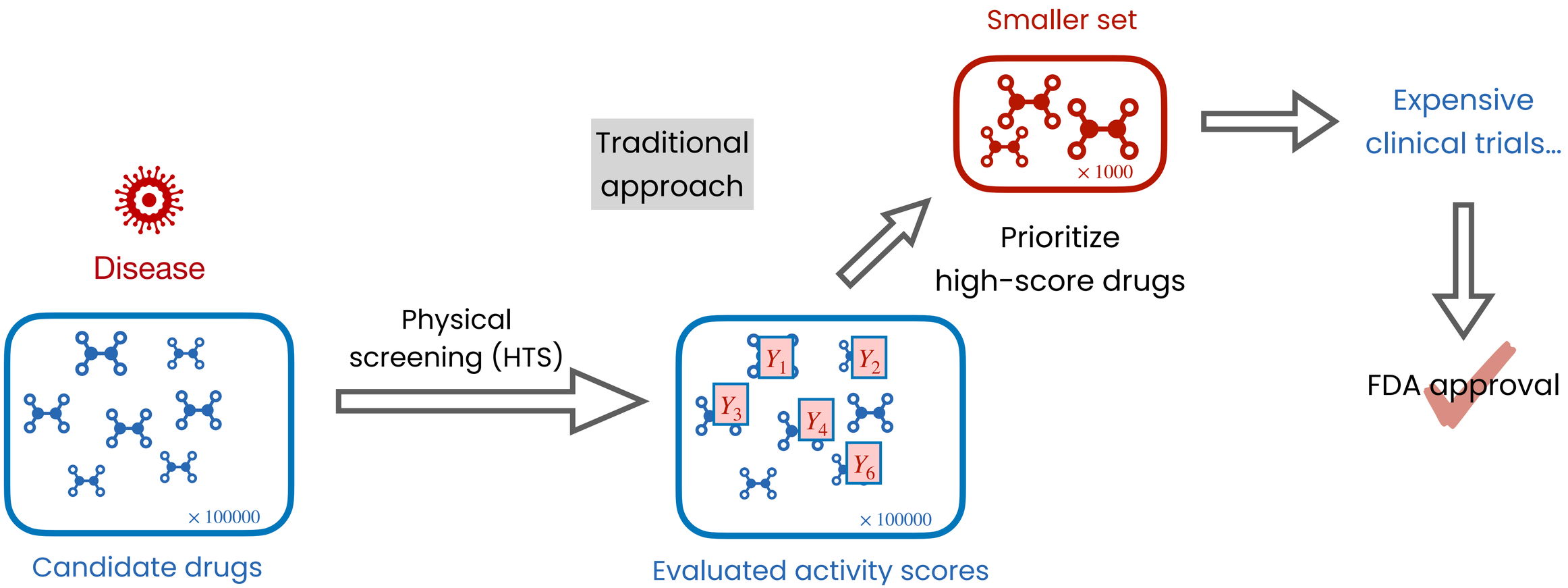
Position



Job applicants

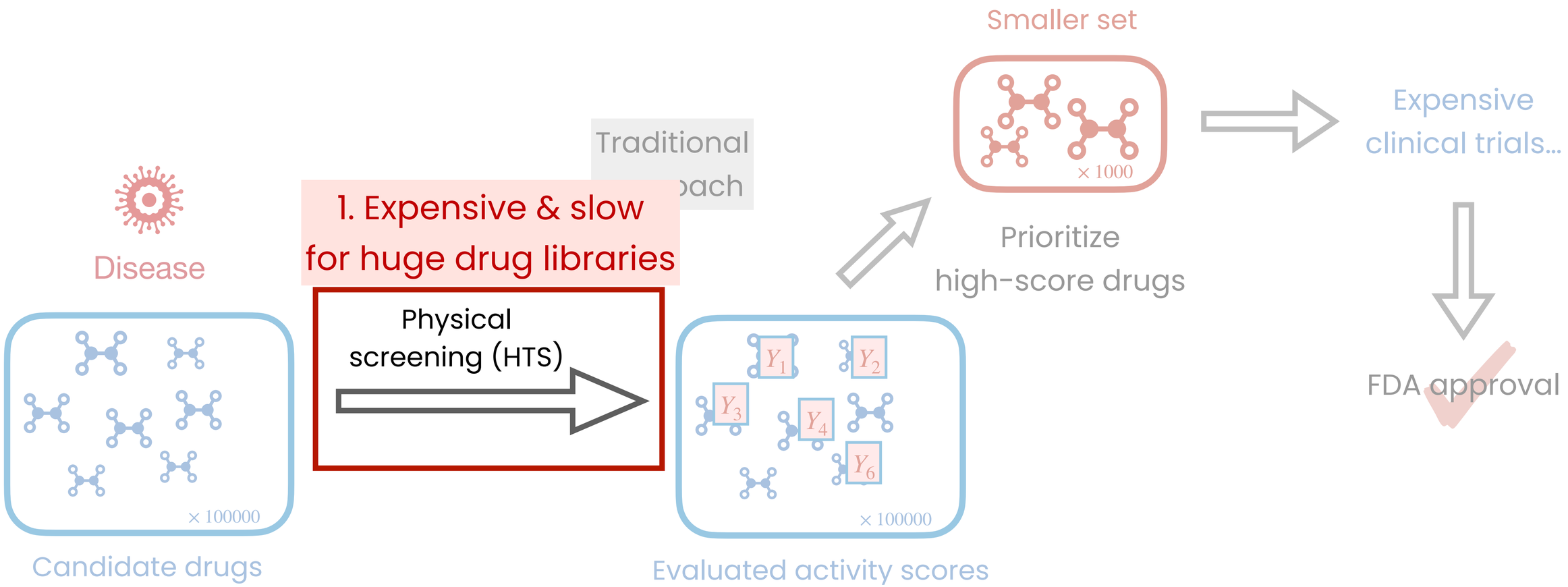
Decision and discovery processes

- Find a few interesting cases from a huge pool



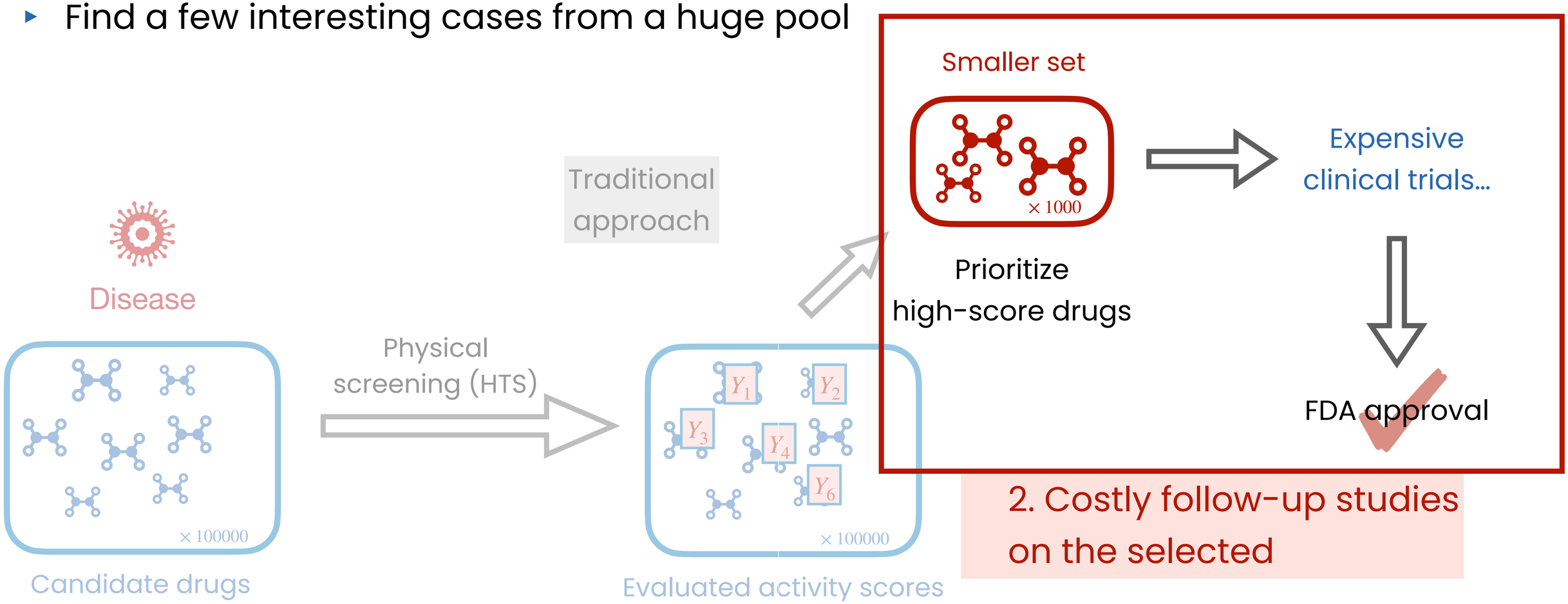
Decision and discovery processes

- Find a few interesting cases from a huge pool



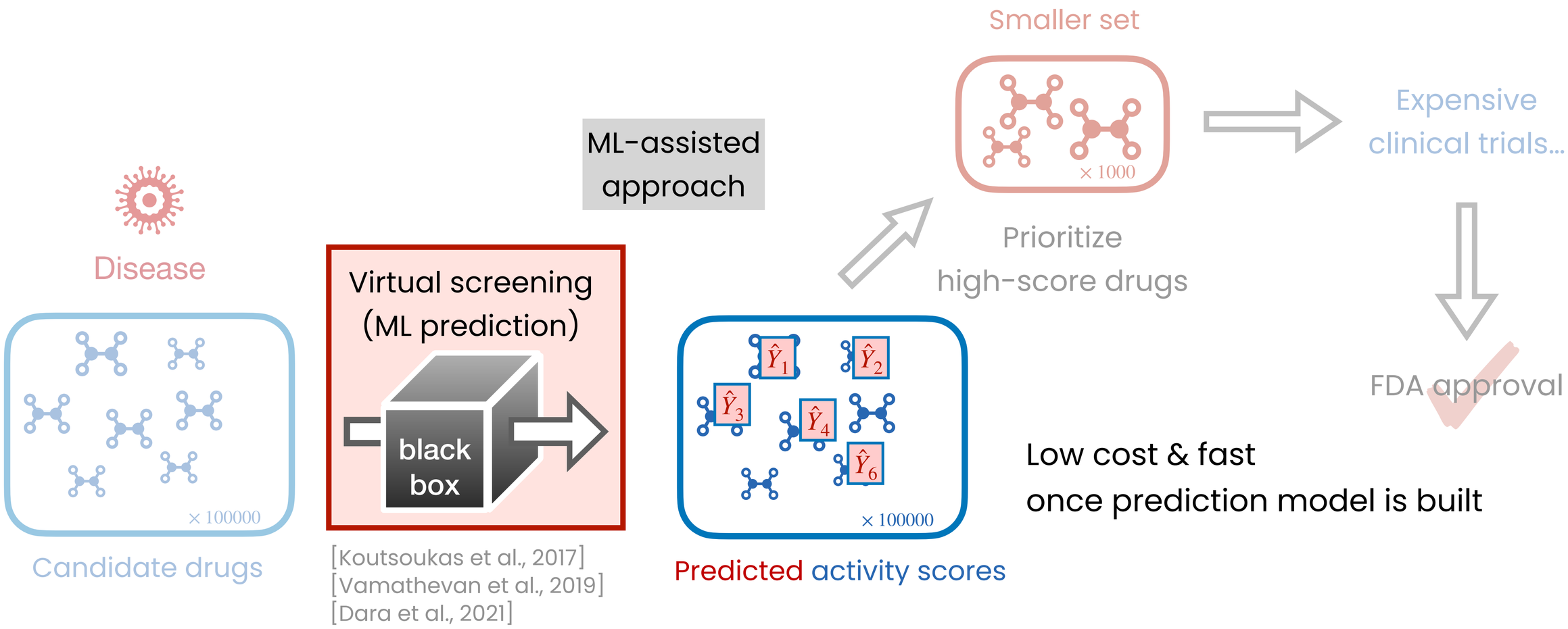
Decision and discovery processes

- Find a few interesting cases from a huge pool



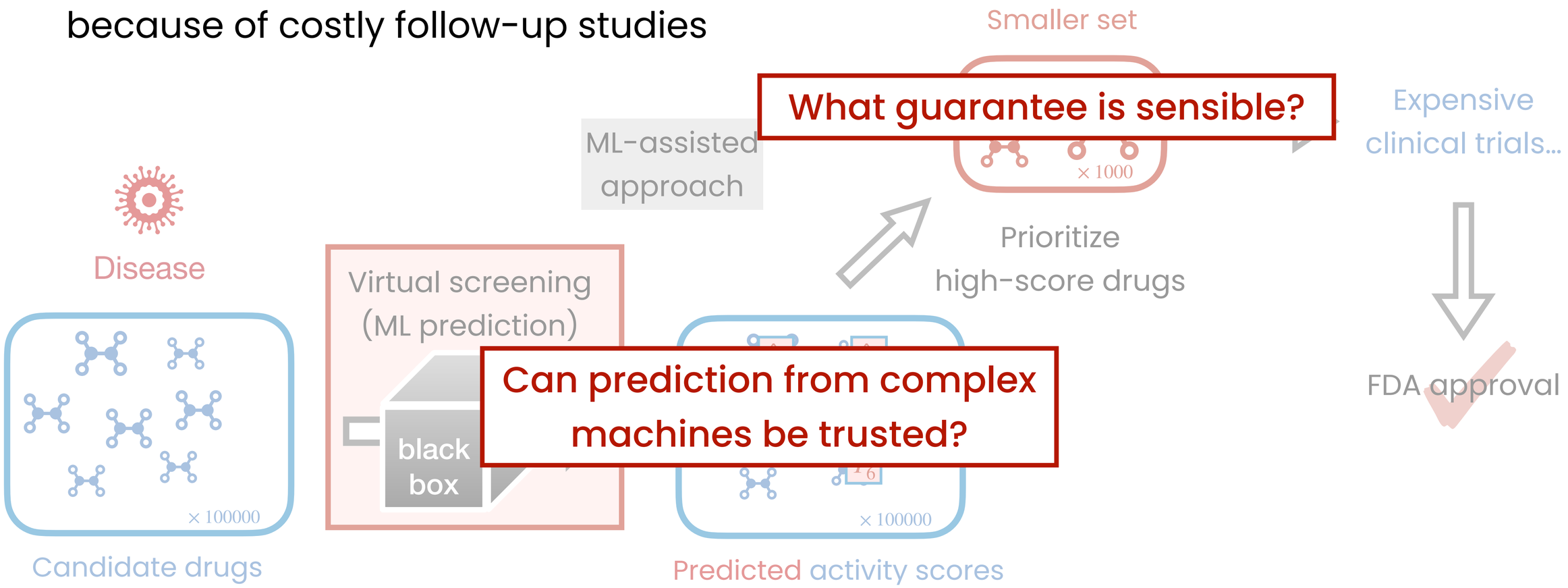
The role of ML in decision and discovery processes

- Find a few interesting cases from a huge pool



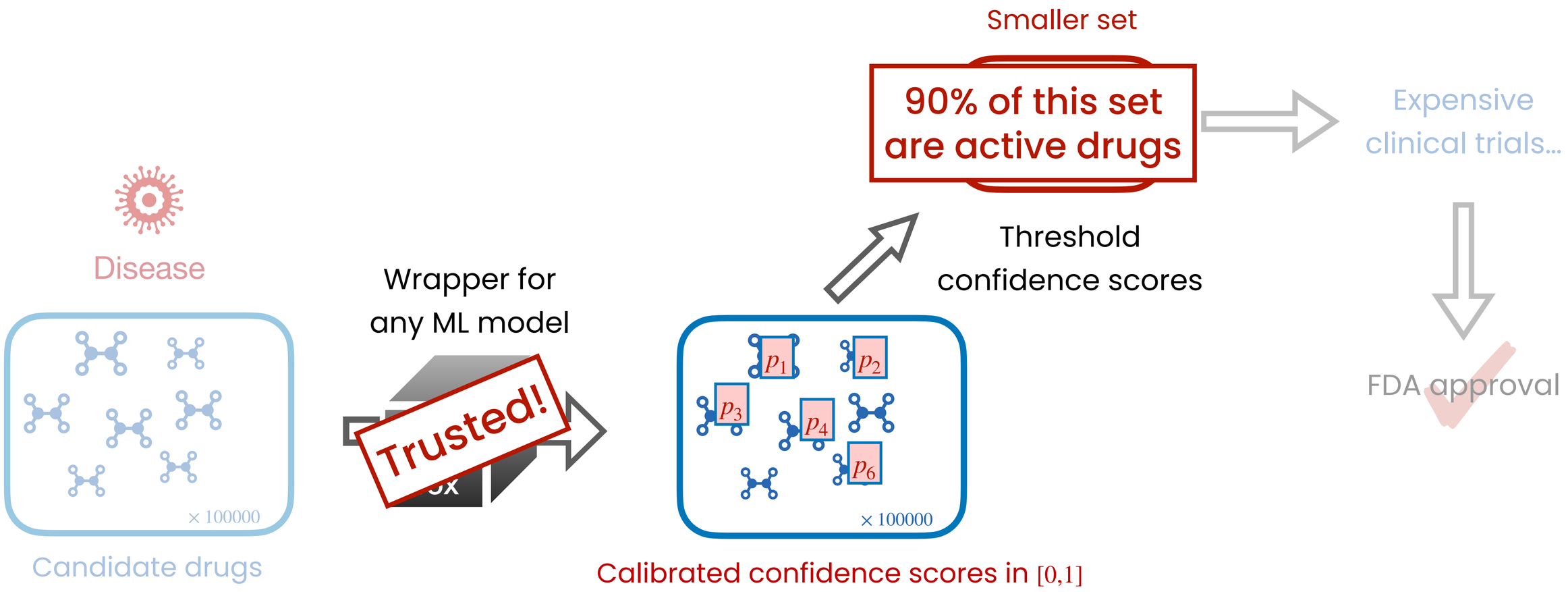
The role of ML in decision and discovery processes

- Error on the selected is concerning because of costly follow-up studies



This work

- Screening with error control on the selected candidates



Mathematical setup

- Any pre-trained ML model $\hat{\mu}: \mathcal{X} \to \mathcal{Y}$
- Training data $\{(X_i, Y_i)\}_{i=1}^{n}$ (already-screened drugs)
- Test samples $\{(X_{n+j}, Y_{n+j})\}_{j=1}^{m}$, only observe covariates $\{X_{n+j}\}_{j=1}^{m}$ (new drugs)
  - $Y \in \{0,1\}$: whether a drug is active for the disease
  - $Y \in \mathbb{R}$: affinity score of a drug for the disease
  - X: physical/chemical structures/properties of the drug

- For now: assume training and test samples are i.i.d. from an unknown distribution
  - Experimentation / Drugs drawn from a diverse drug library
  - Will be relaxed later on to allow for distribution shift

Mathematical setup

- Any pre-trained ML model $\hat{\mu}: \mathcal{X} \to \mathcal{Y}$
- Training data $\{(X_i, Y_i)\}_{i=1}^{n}$ (already-screened drugs)
- Test samples $\{(X_{n+j}, Y_{n+j})\}_{j=1}^{m}$, only observe covariates $\{X_{n+j}\}_{j=1}^{m}$ (new drugs)
  - $Y \in \{0,1\}$: whether a drug is active for the disease
  - $Y \in \mathbb{R}$: affinity score of a drug for the disease
  - X: physical/chemical structures/properties of the drug

- **Goal**: find large outcomes: $Y_{n+j} > c_{n+j}$ for some user-specified thresholds c_{n+j}
  - c_{n+j}: how active a drug should be to be viewed as “interesting”, a known value

Guarantees we seek for

- Interested in large outcomes: $Y_{n+j} > c_{n+j}$ for some user-specified c_{n+j}

- Our goal is to find a subset $\mathscr{R} \subseteq \{1,\ldots,m\}$ as "promising candidates"
- While controlling the false discovery rate (FDR) below some $q \in (0,1)$

$$FDR = \mathbb{E}[FDP], \quad FDP = \frac{\sum_{j=1}^{m} \mathbf{1}\{j \in \mathscr{R}, Y_{n+j} \leq c_{n+j}\}}{1 \vee |\mathscr{R}|}$$

Number of selected but uninteresting units

≈ Number of selected units

[Benjamini and Hochberg, 1995]

- FDR measures the proportion of follow-up resources wasted on uninteresting cases

Our approach: thresholding confidence measure

- Interested in large outcomes: $Y_{n+j} > c_{n+j}$ for some user-specified c_{n+j}

- Build any monotone score function $V(x, y)$, i.e., $y \le y'$ implies $V(x, y) \le V(x, y')$
  - One-sided residual $V(x, y) = y - \hat{\mu}(x)$
  - Fitted cumulative distribution function $V(x, y) = \hat{\mathbb{P}}(Y \le y \mid X = x)$
- Compute $V_i = V(X_i, Y_i)$ for $i = 1,2,\ldots,n$
- Compute test scores $\hat{V}_{n+j} = V(X_{n+j}, c_{n+j})$ for $j = 1,2,\ldots,m$
- Compute confidence measures (p-value in statistics)

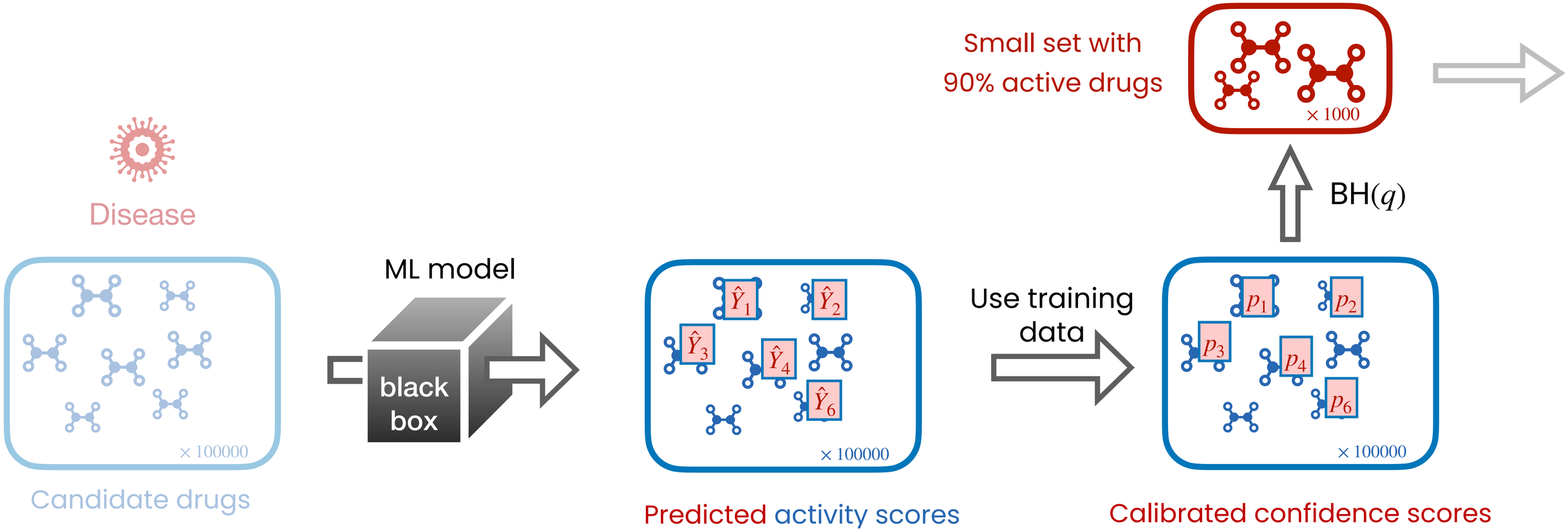
$\approx$ rank of $\hat{V}_{n+j}$ among training scores $\{V_i\}_{i=1}^n$

$$p_j = \frac{\sum_{i=1}^{n} \mathbf{1}\{V_i < \hat{V}_{n+j}\} + U_j}{n+1}, \quad U_j \sim \text{Unif}[0,1]$$

- Get selection set $\mathscr{R}$ by Benjamini-Hochberg procedure applied to $\{p_j\}$ at level q

Our approach: thresholding confidence measure

- Back to the implied pipeline in drug discovery



Interpreting the confidence measure

- Recall: Interested in large outcomes: $Y_{n+j} > c_{n+j}$ for some user-specified c_{n+j}

$$p_j = \frac{\sum_{i=1}^{n} \mathbf{1}\{V_i < \hat{V}_{n+j}\} + U_j}{n+1}, \quad U_j \sim \text{Unif}[0,1]$$

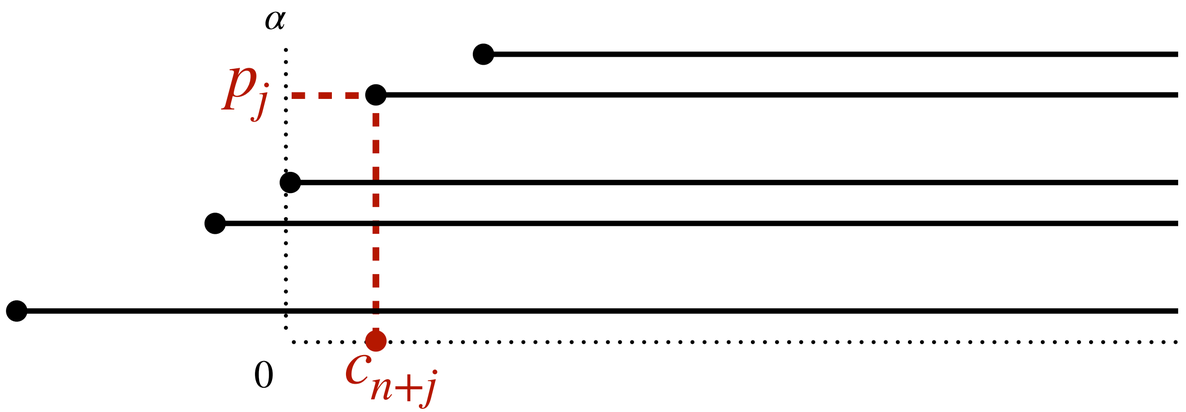
$$p_j \approx \inf\left\{\alpha : c_{n+j} \notin \hat{C}(X_{n+j}; \alpha)\right\}$$

$\hat{C}(X_{n+j}; \alpha)$ is an α-prediction interval for Y_{n+j}, which obeys

$$\mathbb{P}(Y_{n+j} \in \hat{C}(X_{n+j}; \alpha)) \geq 1 - \alpha$$

$\approx$ critical point α such that $\hat{C}(X_{n+j}; \alpha)$ is all larger than c_{n+j}

A smaller p_j means c_{n+j} is smaller than the typical behavior of Y_{n+j}



By monotonicity,

$\hat{C}(X_{n+j}; \alpha) = [\eta(X_{n+j}; \alpha), \infty)$

FDR control with the confidence measure

- Get selection set $\mathscr{R}$ by Benjamini-Hochberg procedure applied to $\{p_j\}$ at level q
  - Set $\mathscr{R} = \{j : p_j \le q\hat{k}/m\}$, where $\hat{k} = \max\left\{k : \sum_{j=1}^{m} \mathbf{1}\{p_j \le qk/m\} \ge k\right\}$

Theorem (J. and Candès, 2022)

If $V(x, y)$ is monotone, the training and test data are i.i.d., and for each j, data in $\{Z_i\}_{i=1}^{n} \cup \{\tilde{Z}_{n+\ell}\}_{\ell \neq j} \cup \{Z_{n+j}\}$ are mutually independent for $Z_i = (X_i, Y_i)$ and $\tilde{Z}_{n+j} = (X_{n+j}, c_{n+j})$, Then for any $q \in (0,1)$, the output $\mathscr{R}$ at level q obeys $FDR \le q$.

- True for random c_{n+j} (will my health risk tomorrow be higher than today?)

Power boosting

- While FDR is controlled for any monotone score $V(x, y)$, some makes it powerful
- If the thresholds are constant $c_{n+j} \equiv c$, a particularly powerful choice is `clipped' score

$$V(x, y) = +\infty \cdot \mathbf{1}\{y > c\} + c \cdot \mathbf{1}\{y \leq c\} - \hat{\mu}(x)$$

- In binary case and $c = 0$, the ideal score is monotone in $\mathbb{P}(Y = 1 \mid X = x)$ (see paper)

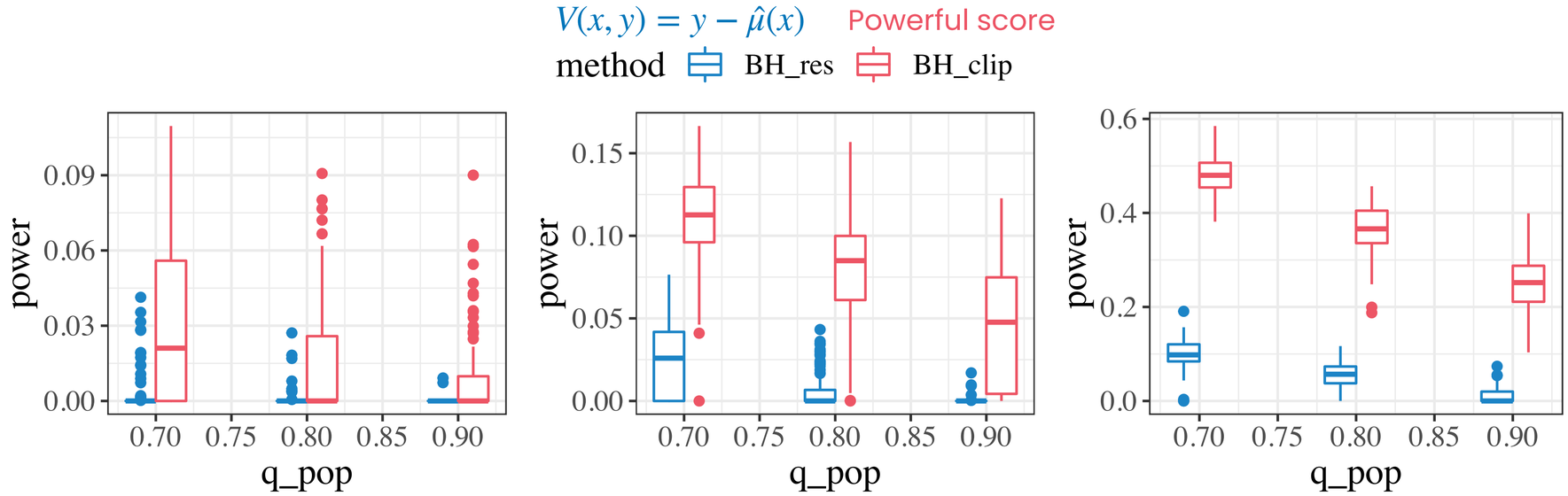
Real application: drug property prediction for HIV

- Binary $Y \in \{0,1\}$: whether the drug interacts with the disease
- The drug library is $n_{tot} = 41127$ in total, use $6:2:2$ split
- Very sparse data: only 3% drugs are active
- Our hope: find a smaller subset to proceed so that $(1-q)$ of the subset are active drugs
- FDR level $q \in \{0.1,0.2,0.5\}$, use a small neural network (can be more complicated)

	Realized FDR			**Power**			$\lvert\mathscr{R}\rvert$		
FDR level	0.1	0.2	0.5	0.1	0.2	0.5	0.1	0.2	0.5
Powerful score	**0.0957**	**0.196**	**0.495**	0.0788	0.174	0.410	26.5	64.2	240
Score $V(x,y) = y - \hat{\mu}(x)$	**0.0989**	**0.196**	**0.494**	0.0766	0.174	0.410	25.8	64.4	239

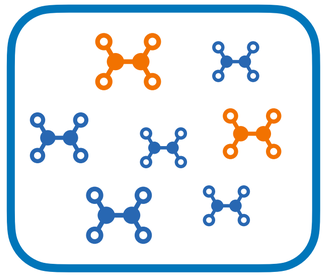
Real application: drug-target-interaction prediction

- Davis dataset, $Y \in \mathbb{R}$ continuous binding affinities, X feature for a drug-target pair
- The drug library is $n_{tot} = 30060$ in total, use $2:2:6$ split
- Set c_{n+j} as the q_{pop}-th quantile of the outcomes in the first training fold with the same binding target as test sample j, where $q_{pop} \in \{0.7, 0.8, 0.9\}$
- FDR level $q \in \{0.1, 0.2, 0.5\}$

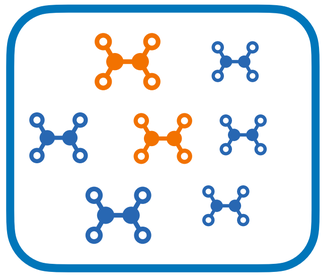


Distribution shifts

- The only assumption for this method to work is i.i.d. data
- Are my evaluated drugs comparable to the unknown drugs?
  - **Yes** if the evaluated ones are drawn without preference from your library



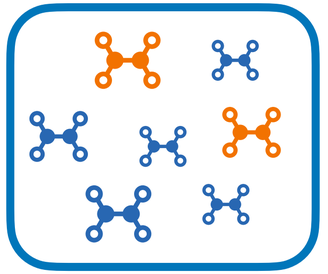
Training drugs



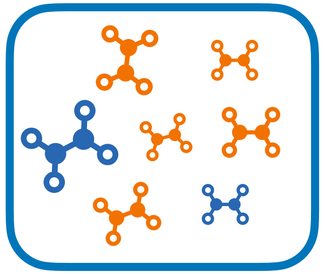
New drugs

Distribution shifts

- The only assumption for this method to work is i.i.d. data
- Are my evaluated drugs comparable to the unknown drugs?
  - **Yes** if the evaluated ones are drawn without preference from your library
  - **No** if you preferred drugs with some specific structures, etc



Training drugs



New drugs

- Similar issues happen in job hiring, health monitoring…
  - Candidates documented last year may differ from current
  - Patients may differ in demographics across hospitals
  - People under treatment may be different than those under control
- Forthcoming: A new procedure **exactly** controlling FDR under covariate shift

Summary

- In prediction-assisted screening problems, FDR can be a sensible measure
- A method that turns any prediction model into a reliable selection procedure
  - Useful if interested in “large” outcomes
  - Builds confidence scores (p-values) upon any prediction model
  - Controls FDR so that your follow-up investigations are well-deserved
- Extension to situations with covariate shifts
  - Some more complicated methodology & theory

arXiv: 2210.01408