Model-free selective inference and applications to drug discovery

Ying Jin

Based on join work with Emmanuel Candès, Jure Leskovec, Genentech

Collaborators

Emmanuel Candès
Stanford Stats & Math

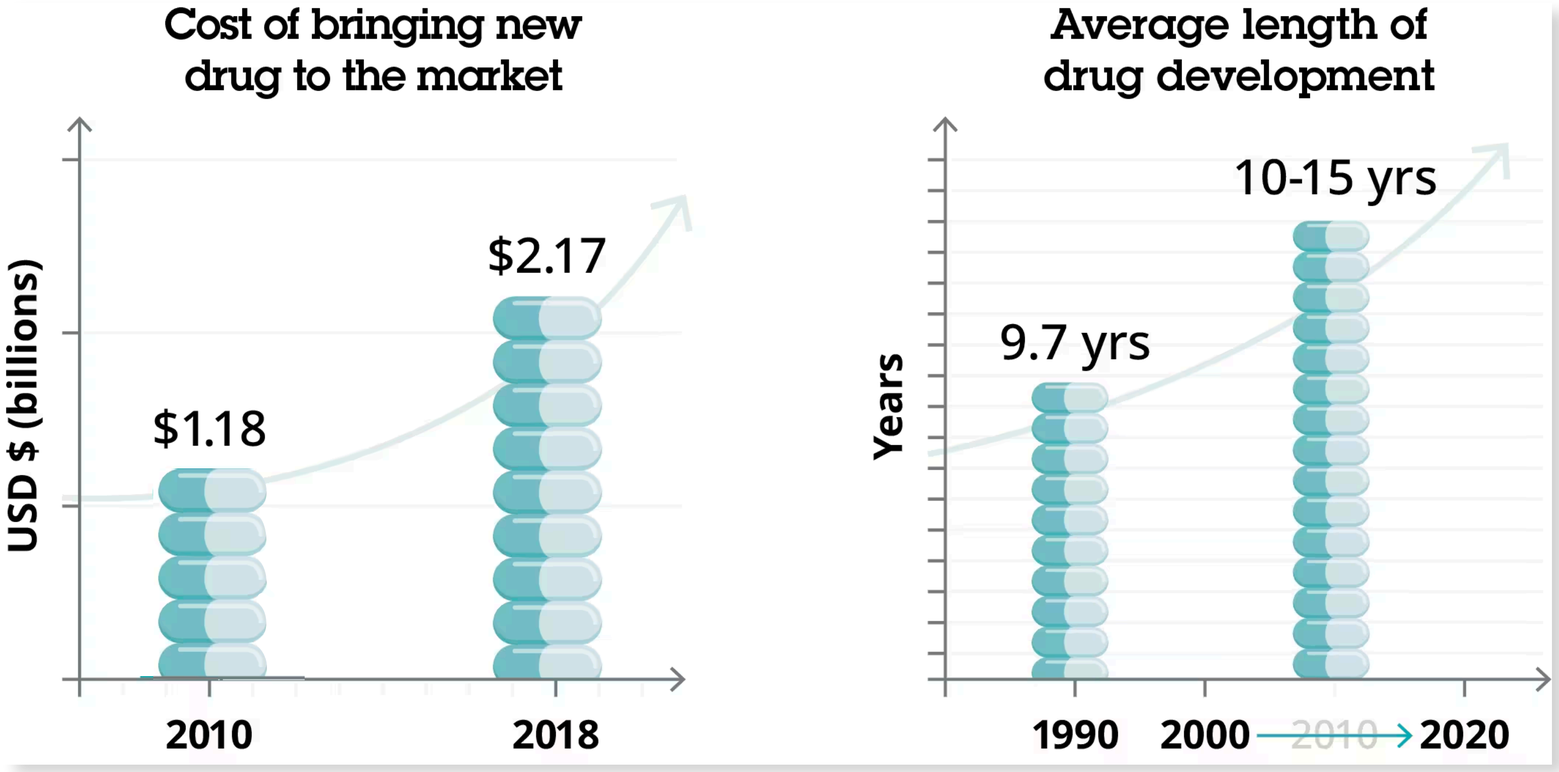
+

Applied work in collaboration with:

Jure Leskovec
Stanford CS

Genentech ML team

Motivation: accelerating drug discovery



[abcam.com]

Can we make drug discovery more *efficient*?

Scientific discovery in the age of AI

Shortcuts to Simulation: How Deep Learning Accelerates Virtual Screening for Drug Discovery

May 11, 2020 14 min read

[DZone.com]

FORBES > INNOVATION

Generative AI Drugs Are Coming

Steve Nouri Forbes Councils Member
Forbes Technology Council
COUNCIL POST | Membership (Fee-Based)

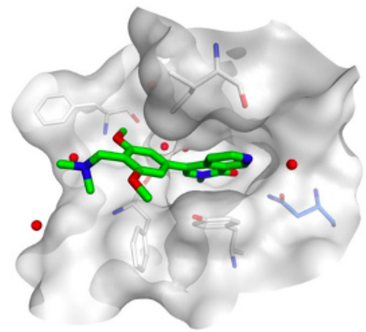
Sep 5, 2023, 07:45am EDT

[forbes.com]

[mckinsey.com]

Promise of AI: *low-cost & fast* drug discovery!

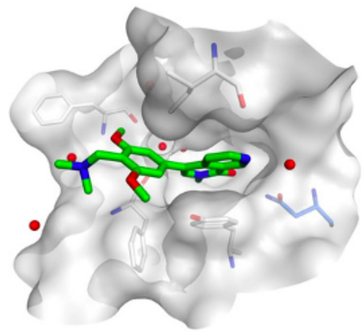
This talk: in search of “interesting/large outcomes”



Want drugs with high binding affinities to a disease target

Which drugs are sufficiently active?

This talk: in search of “interesting/large outcomes”

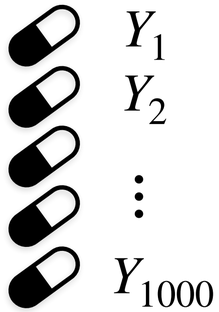


Want drugs with high binding affinities to a disease target

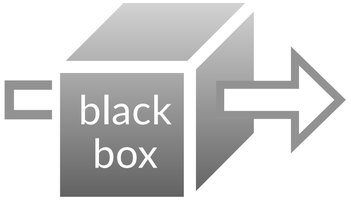
Which drugs are sufficiently active?

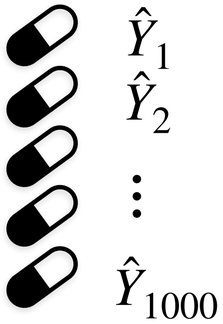
Experiments, clinical trials, …

This talk: in search of “large outcomes”



Traditional approach:
evaluate unknown outcomes





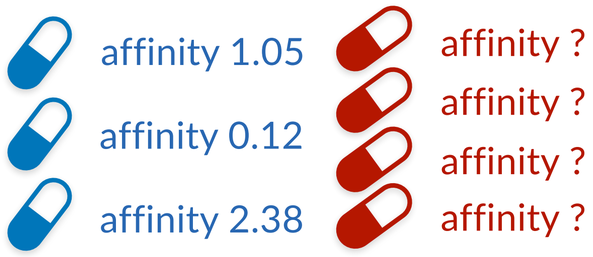
AI-assisted approach:
predict unknown outcomes!

“decisions”

[Koutsoukas et al., 2017; Vamathevan et al., 2019; Dara et al., 2021]

Problem setup

- Any pre-trained prediction model $\hat{\mu}: \mathcal{X} \to \mathcal{Y}$ (independent of training and test data)
  - X physical/chemical feature/amino acids of the drug
  - Y binding affinity
    - $\rightsquigarrow Y \in \{0,1\}$: whether the drug binds to the target
    - $\rightsquigarrow Y \in \mathbb{R}$: how well the drug binds to the target



- Training data $\{(X_i, Y_i)\}_{i=1}^n$ (screened drugs)
- Test samples $\{(X_{n+j}, Y_{n+j})\}_{j=1}^m$ with unknown $\{Y_{n+j}\}_{j=1}^m$ (new drugs)

Goal: find large outcomes $Y_{n+j} > c_{n+j}$ without too many errors

$\rightsquigarrow$ user-specified thresholds c_{n+j} to become 'interesting'

Other applications

Goal: find large outcomes $Y_{n+j} > c_{n+j}$ without too many errors

material design

talent identification

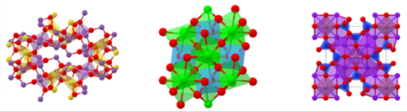
targeted marketing

...

ARTICLE • AI, MATH, AND DATA

Google DeepMind Adds Nearly 400,000 New Compounds to Berkeley Lab's Materials Project

By Lauren Biron
November 29, 2023



[newscenter.lbl.gov]

Microsoft Unveils Predictive Targeting, AI-Based Advertising Tool

Microsoft unveils Predictive Targeting, an AI-based advertising tool enhancing conversion rates, streamlining targeting, and offering flexible audience strategies.

[forbes.com]

HIRING RESOURCES | 9 MIN READ

How Good Machine Learning in Recruitment Can Radically Transform Your Hiring

[VerVoe.com]

Challenges

$\hat{\mu}(X_{n+1})$

$\hat{\mu}(X_{n+2})$

$\vdots$

$\hat{\mu}(X_{n+m})$

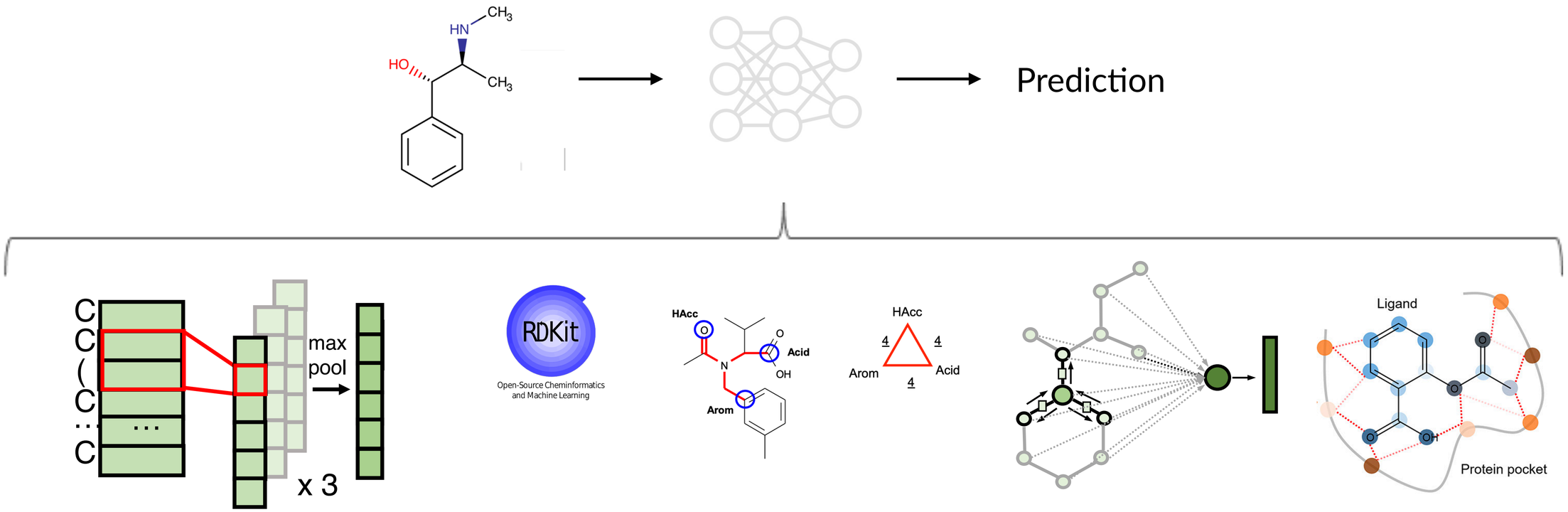
- Quantifying uncertainty in point predictions
- Model-free

  Work for any prediction model

  No modeling assumptions

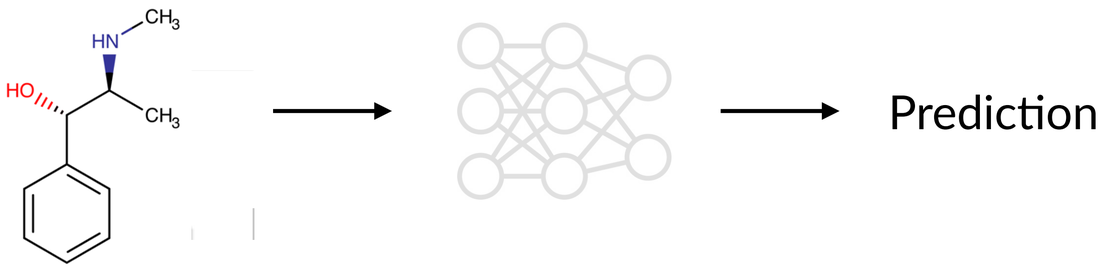
Which drugs are sufficiently active?

The importance of reliability

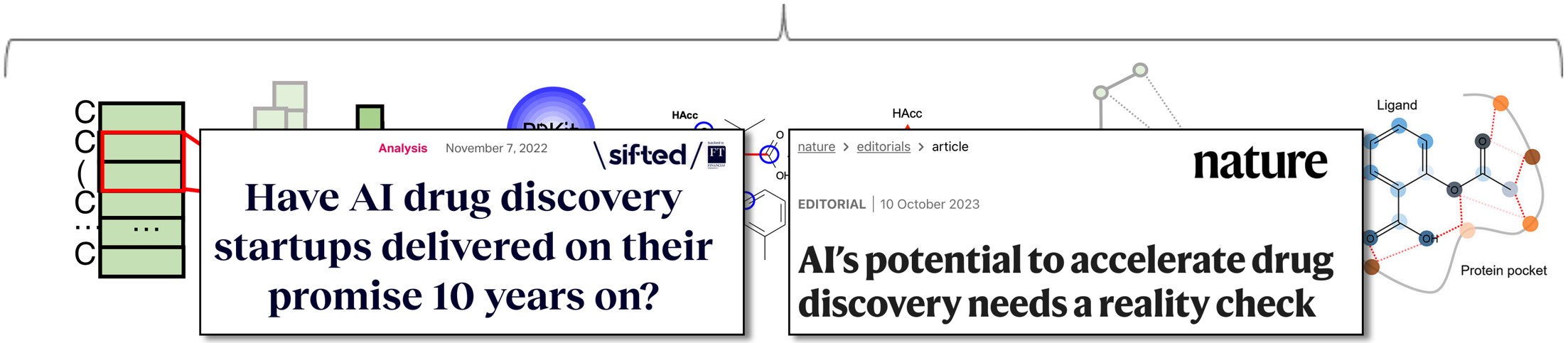


What if AI gives false leads? *Failure of the promise!*

The importance of reliability



Prediction

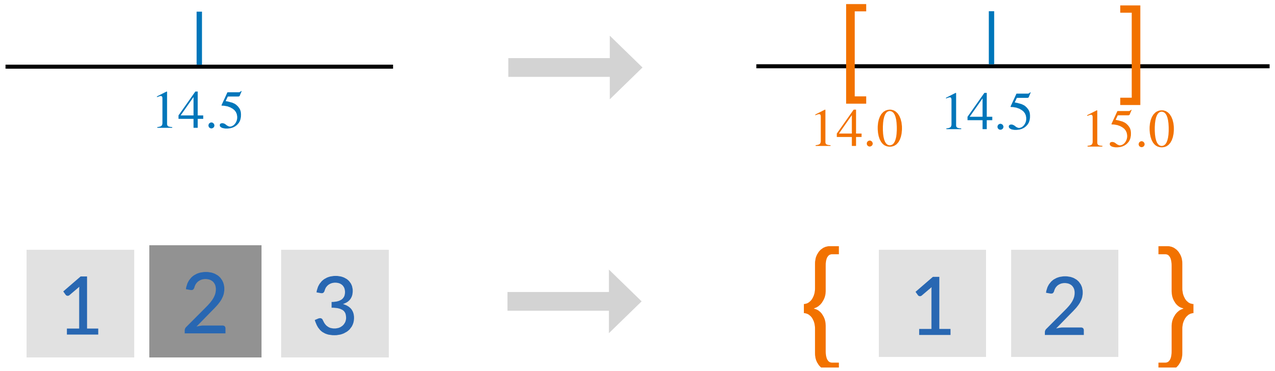


Can we draw discoveries with *few mistakes*?

Conformal prediction: model-free uncertainty quantification

Y_{n+1}?

Point prediction → Prediction set $\hat{C}(X_{n+1})$

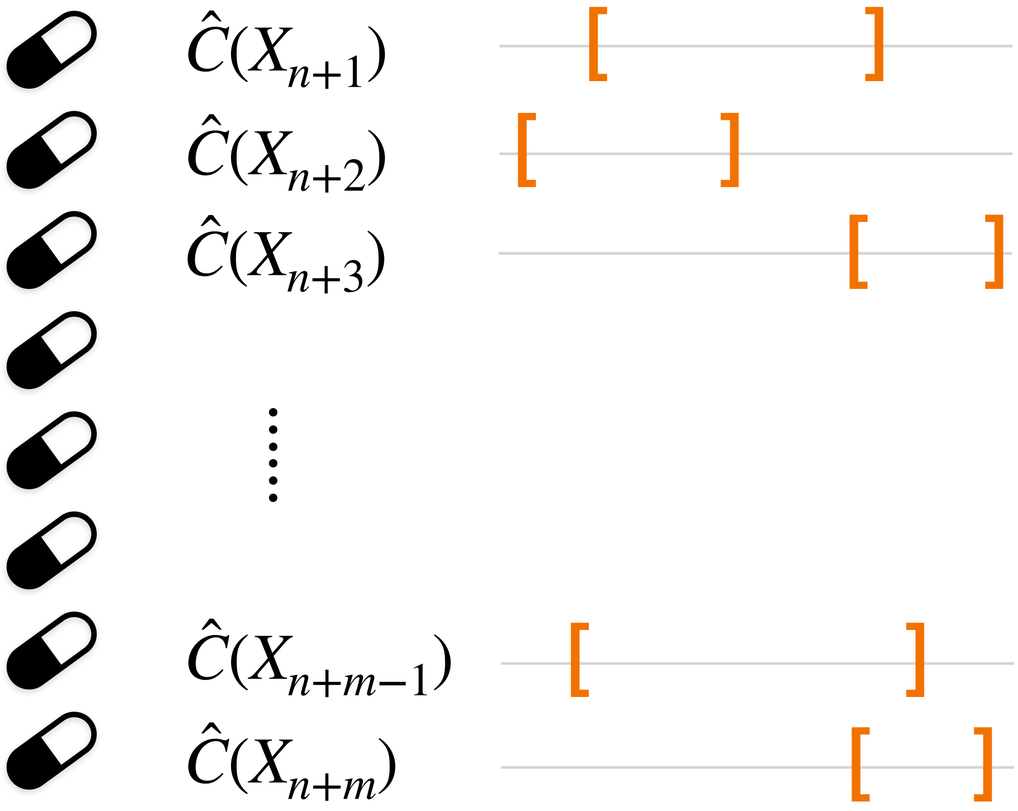


Validity of conformal prediction intervals (PIs) [Vovk et al., 1999]

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}(X_{n+1})\right) \geq 95\,\%$$

⇝ Covers 95% of outcomes <u>no matter prediction model</u>

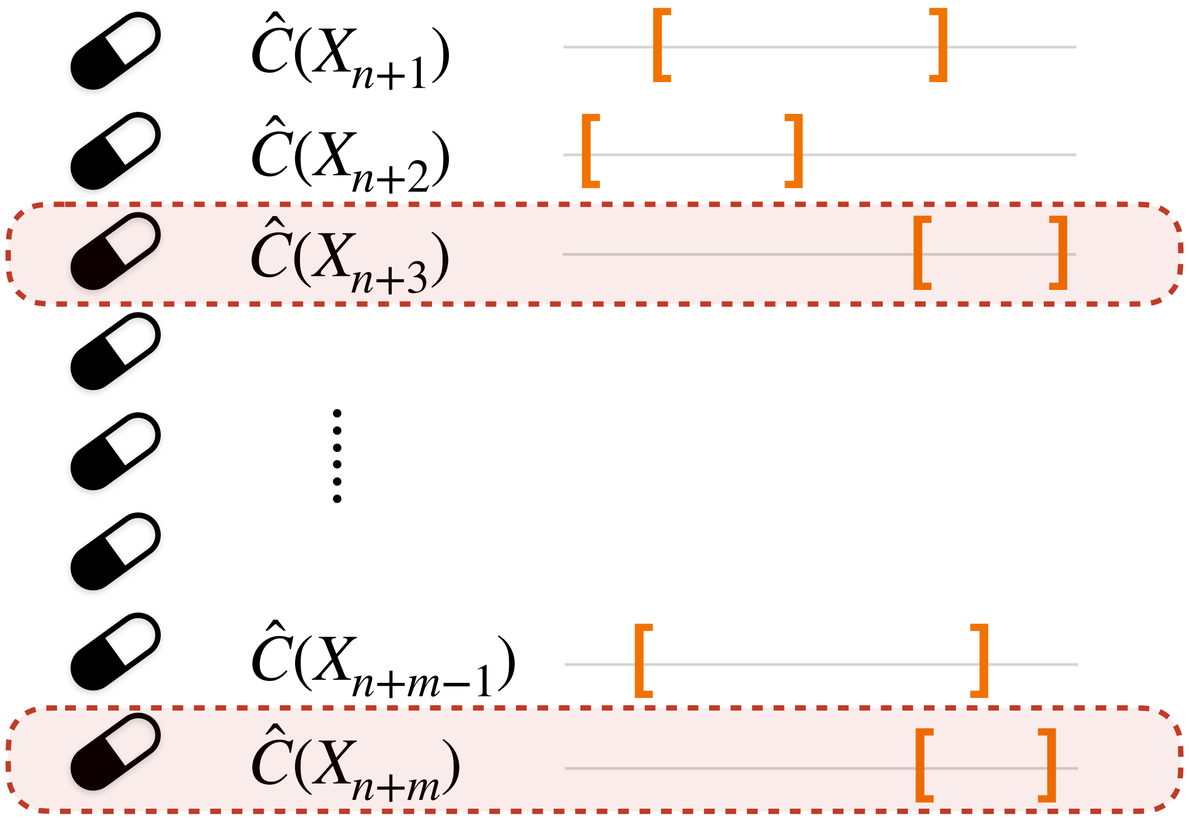
Challenges



- Uncertainty quantification ✓
- Model-free ✓

Which drugs are sufficiently active?

Challenges



- Uncertainty quantification ✓
- Model-free ✓
- Can we use them to find interesting instances (drugs)?

Which drugs are sufficiently active?

“Selective” downstream use of predictive inference

Drug discovery [Svenssen et al., 2017, JCIM]

*[…] **compounds to further screen** can be derived from […] **single class predictions** found at the user-defined confidence level.*

Marketing [redfield.ai/conformal-prediction-for-business]

*[…] interval indicates **strong demand**, the company can **invest more** in advertising […] Conversely, […] suggests **weaker demand**, they can focus on **cost-saving** initiatives.*

Disease diagnosis [Olsson et al., 2022, Nature Communications]

*If the prediction region associated with a point prediction is **too large** […], the corresponding prediction **can be flagged** for human intervention.*

“Selective” downstream use of predictive inference

Drug discovery [Svenssen et al., 2017, JCIM]

*[...] **compounds to further screen** can be derived from [...] **single class predictions** found at the user-defin*

Marketing [redfield.a

[...] interval indic *e in advertising [...] Conversely, [...]* *aving initiatives.*

Practice:

Construct prediction intervals

⇓

Select “interesting” intervals

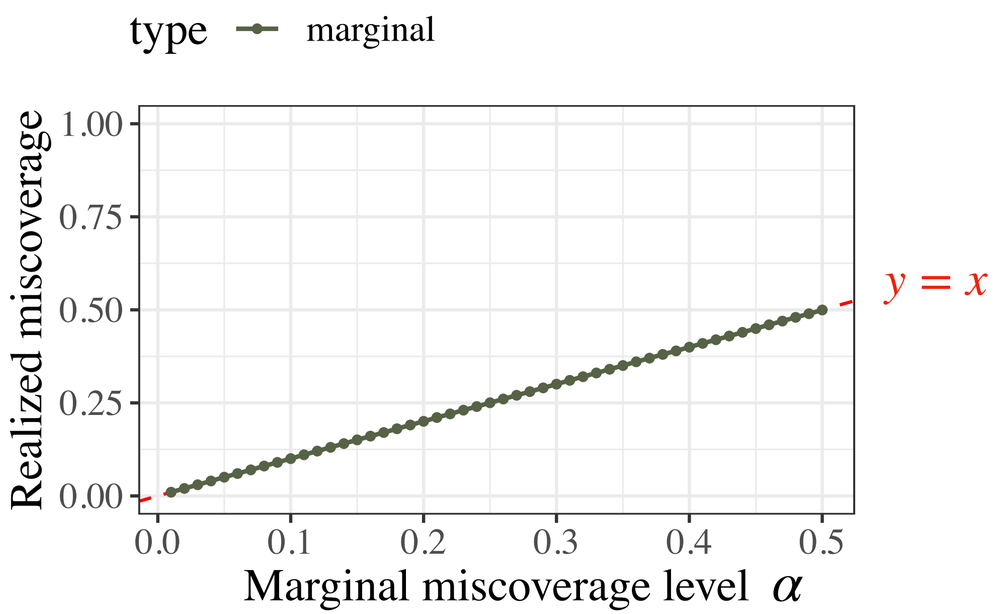
⇓

Strong selection bias problem

Disease diagnosis

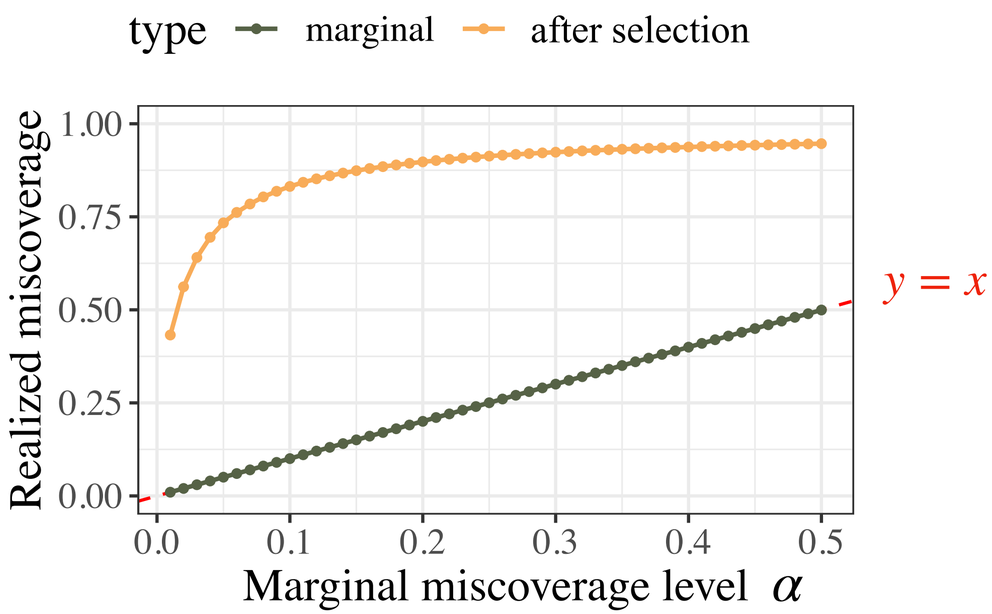
*If the prediction region associated with a point prediction is **too large** [...], the corresponding prediction **can be flagged** for human intervention.*

Evidence in a real drug discovery dataset



Dark: perfect marginal coverage

Evidence in a real drug discovery dataset



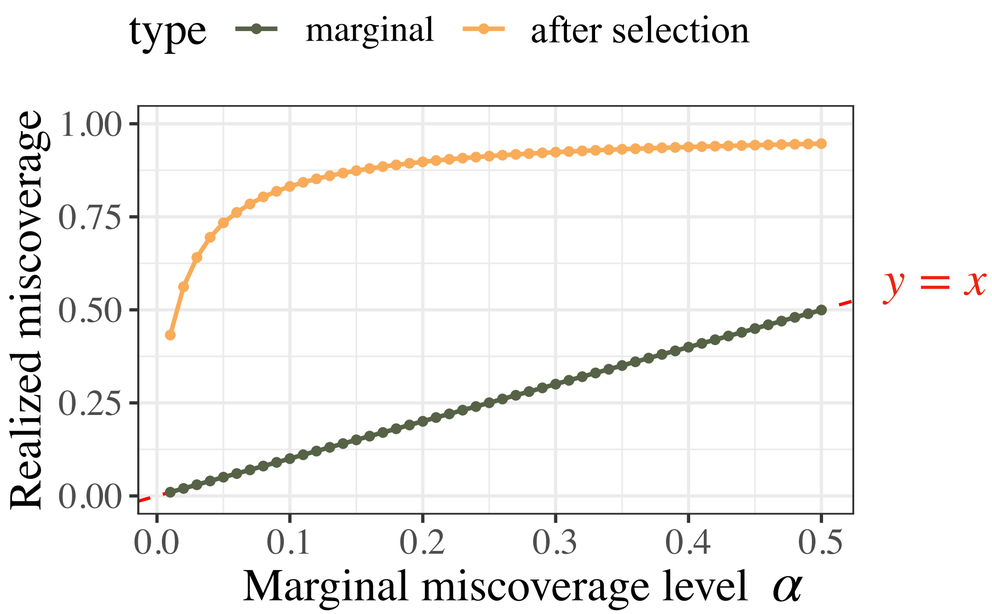
Dark: perfect marginal miscoverage

Orange: miscoverage of those $\hat{C}(X_{n+j}) > c_{n+j}$

Conformal prediction for drug discovery

[Norinder et al., 2014, Svensson et al., 2017, Wang et al., 2022]

Evidence in a real drug discovery dataset



Dark: perfect marginal miscoverage

Orange: miscoverage of those $\hat{C}(X_{n+j}) > c_{n+j}$

Conformal prediction for drug discovery

[Norinder et al., 2014, Svensson et al., 2017, Wang et al., 2022]

1% nominal error, yet >30% error after selection!

This is **the winner's curse** [Soric, 1989]

Inspired a whole field of research: Selective Inference

[Benjamini and Yekutieli, 2005, Berk et al., 2013, Taylor et al., 2014, Fithian et al., 2014; Storey et al, 2003]

Our proposal: select with guarantees

- Find "actionable instances" while controlling fraction of false positive (FDR)

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}], \quad \mathrm{FDP} = \frac{\#\{\text{false discoveries}\}}{\#\{\text{selected instances}\}}$$

[Benjamini and Hochberg, 1995]

- Control of FDR implies
  - Most AI-powered decisions are correct
  - Resource allocation is efficient

Drugs $\rightsquigarrow$ 90% active
Customers $\rightsquigarrow$ 90% responding
Patients $\rightsquigarrow$ 90% benefiting
LLM outputs $\rightsquigarrow$ 90% trustworthy

- Extremely popular notion of error control

Controlling the false discovery rate: a practical and powerful approach to multiple testing

Authors Yoav Benjamini, Yosef Hochberg Total citations Cited by 113748

Part I: Exchangeable/i.i.d. data

Jin, Y. and Candès, E.J., 2023.

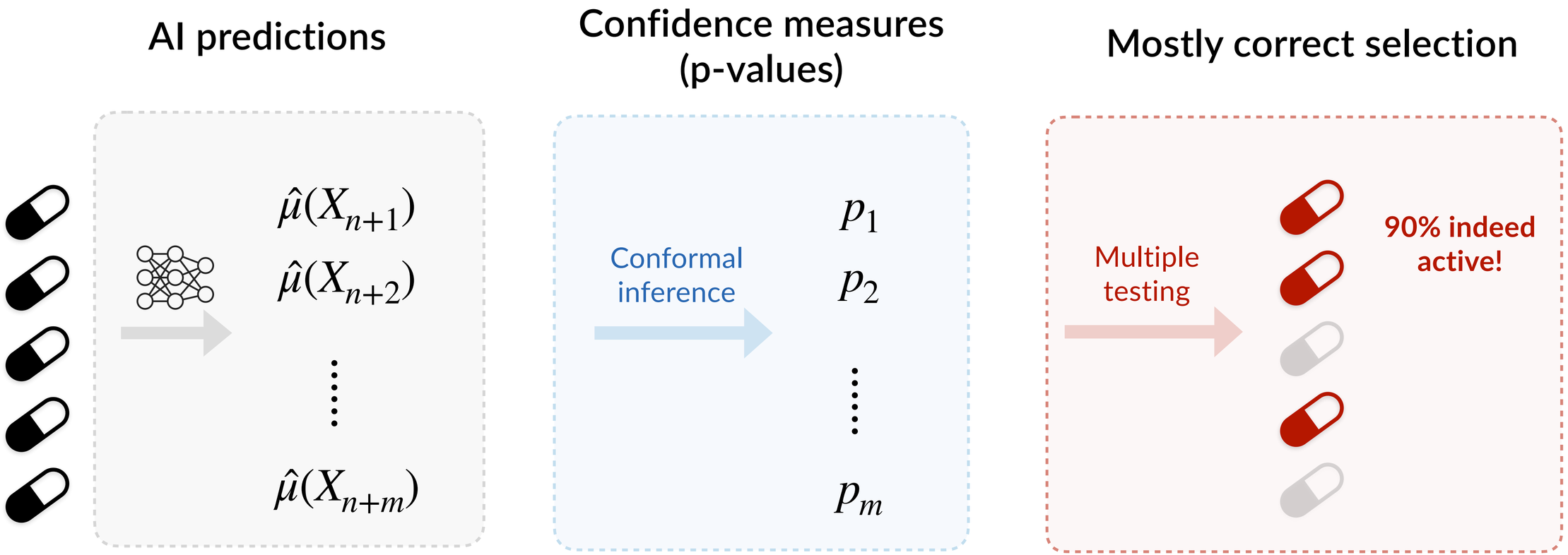
Selection by prediction with conformal p-values.

Journal of Machine Learning Research, *24*(244), pp.1-41.

Exchangeability: for any permutation π of $\{1,\ldots,n+1\}$,

$$\mathbb{P}(V_{\pi(1)} = z_1, \ldots, V_{\pi(n+1)} = v_{n+1}) \equiv \mathbb{P}(V_1 = v_1, \ldots, V_{n+1} = v_{n+1})$$

Model-free selective inference: key strategy



Any black box (model free)

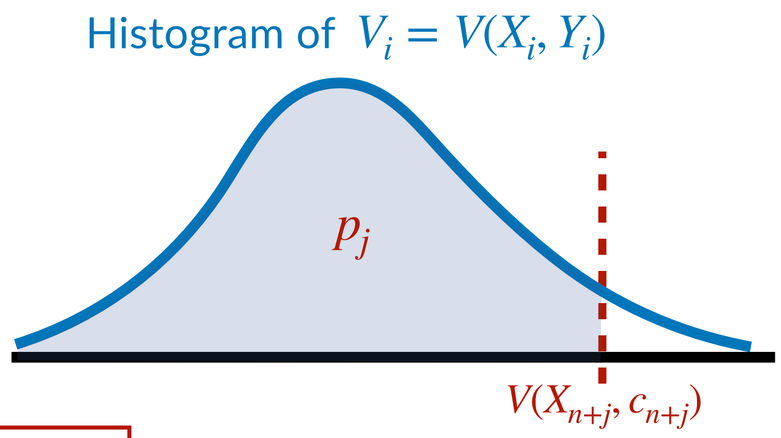
$\mathbb{P}(p_j \leq t,\ Y_{n+j} \leq c_{n+j}) \leq t$

False discovery rate $\leq$ 10%

Conformal p-values

- Monotone conformity score $y \leq y' \Rightarrow V(x, y) \leq V(x, y')$
  - One-sided residual $V(x, y) = y - \hat{\mu}(x)$ [Vovk et al., 2005, Romano et al., 2021]
  - Standardized residual $V(x, y) = [y - \hat{\mu}(x)]/\hat{\sigma}(x)$ [Lei et al., 2018]
  - Fitted cumulative distribution function $V(x, y) = \hat{\mathbb{P}}(Y \leq y \mid X = x)$ [Chernozhukov et al., 2021]

- Training scores $V_i = V(X_i, Y_i), i = 1,2,\ldots,n$
- Test scores $\hat{V}_{n+j} = V(X_{n+j}, c_{n+j}), j = 1,2,\ldots,m$
- Compute p-values



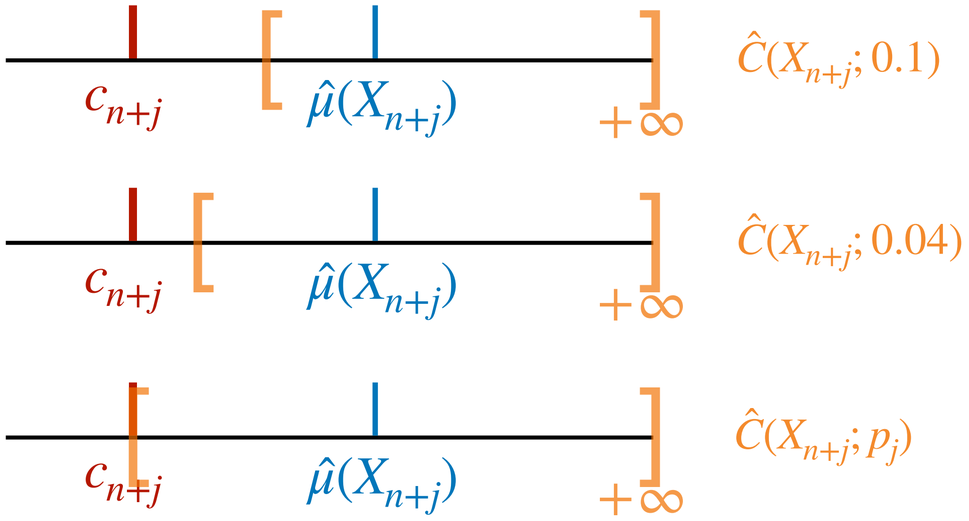
$$\boxed{p_j = \frac{\sum_{i=1}^{n} \mathbf{1}\{V_i < \hat{V}_{n+j}\} + U_j}{n+1}, \quad U_j \sim \text{Unif}[0,1]}$$

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

P-values ⇔ prediction intervals

- With monotone scores, p_j is smallest α such that $\hat{C}(X_{n+j}; \alpha)$ entirely lies above c_{n+j}

Conformal PI with $(1 - \alpha)$ coverage



Setting confidence strength via BH

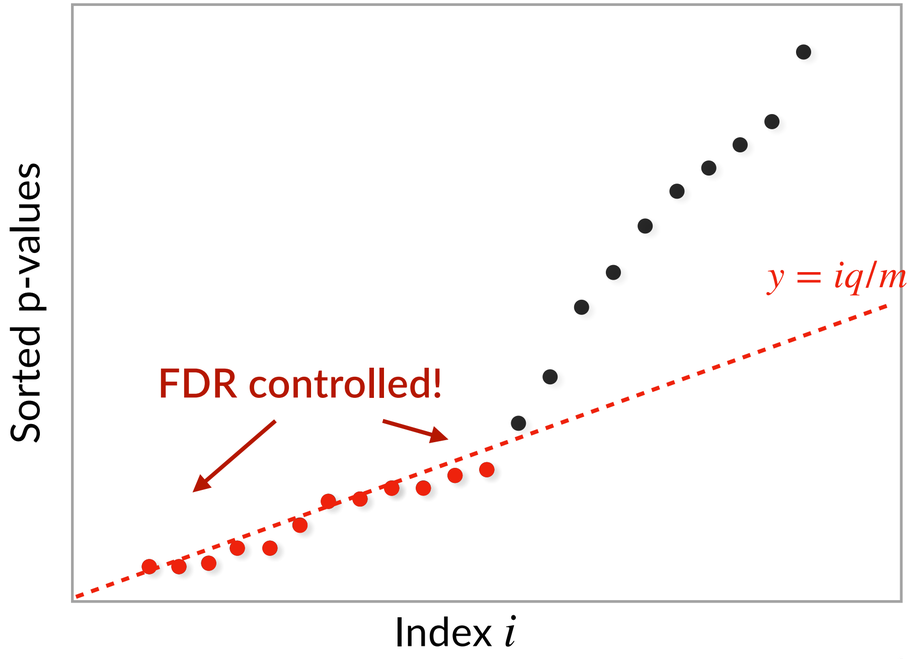
- Rank test samples by p-values / confidence
- Determine a “data-dependent” threshold of p-values

FDR target

Conformal p-values

$$\mathrm{BH}(q; p_1, \ldots, p_m)$$

Benjamini-Hochberg procedure



Model-free FDR control

Theorem (J. and Candès, 2023)

For i.i.d. data and any monotone V, conformal selection at nominal level $q \in (0,1)$ yields

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

Link to complete version

- ✓ Arbitrary prediction model
- ✓ Arbitrary data distribution
- ✓ Random thresholds
- ✓ Dependent data points

Far from classical theory… Why validity?

Why can we ensure model-free error control?

Statistical inference theory: multiple testing for random hypotheses

1. Valid p-values: Well-calibrated for random hypotheses

$$\mathbb{P}(p_j \leq t,\ Y_{n+j} \leq c_{n+j}) \leq t, \quad \forall t \in [0,1]$$

~ Valid p-values from rank test

2. "Multiple testing friendly": P-values are positively dependent

~ 'Good' for BH [Benjamini and Yekutieli, 2001]

Link to positive dependence

How to choose the score?

- Full flexibility: encode preference in choosing V

  $\rightsquigarrow$ procedure selects small $V(X_{n+j}, c_{n+j})$

- If the thresholds are constant $c_{n+j} \equiv c$, a powerful choice is 'clipped' score

$$V(x, y) = \begin{cases} +\infty, & \text{if } y > c \\ c - \hat{\mu}(x), & \text{if } y \leq c \end{cases}$$

  Idea: push training scores $\{V_i\}$ to largest possible

  $\rightsquigarrow$ strictly smaller p-values $\rightsquigarrow$ better power

- For binary outcome with $c = 0$, powerful score should be monotone in $\mathbb{P}(Y = 1 \mid X = x)$

Link to power analysis

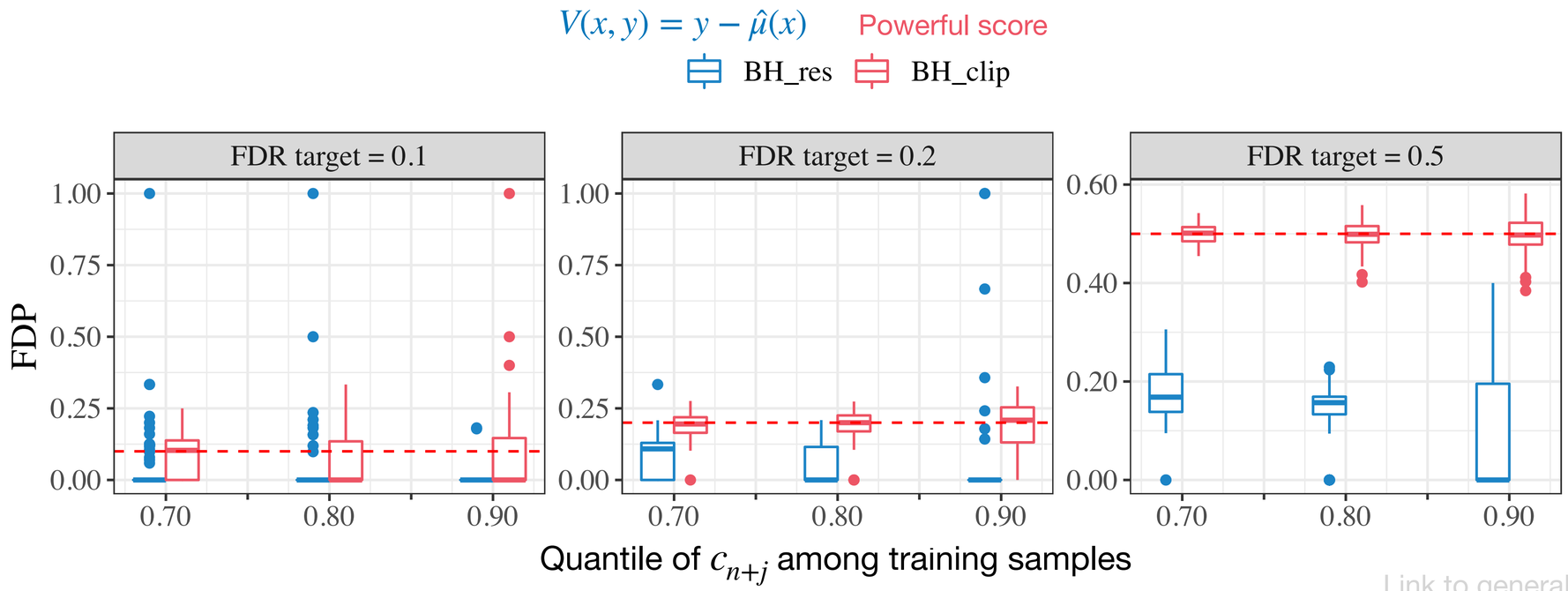
Real data: finding active drugs for HIV with FDR control

- $Y \in \{0,1\}$: whether the drug interacts with the disease
- $n_{tot} = 41127$ in total, $6:2:2$ split
- Very imbalanced data: only **3%** drugs are active
- Goal: select subset with $\approx \{90,80,50\}$ % active drugs

	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
Clipped 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
Naive CP	**0.8315**	**0.8976**	**0.9465**	—	—	—	—	—	—

Real data: finding highly-binding drug-target pairs

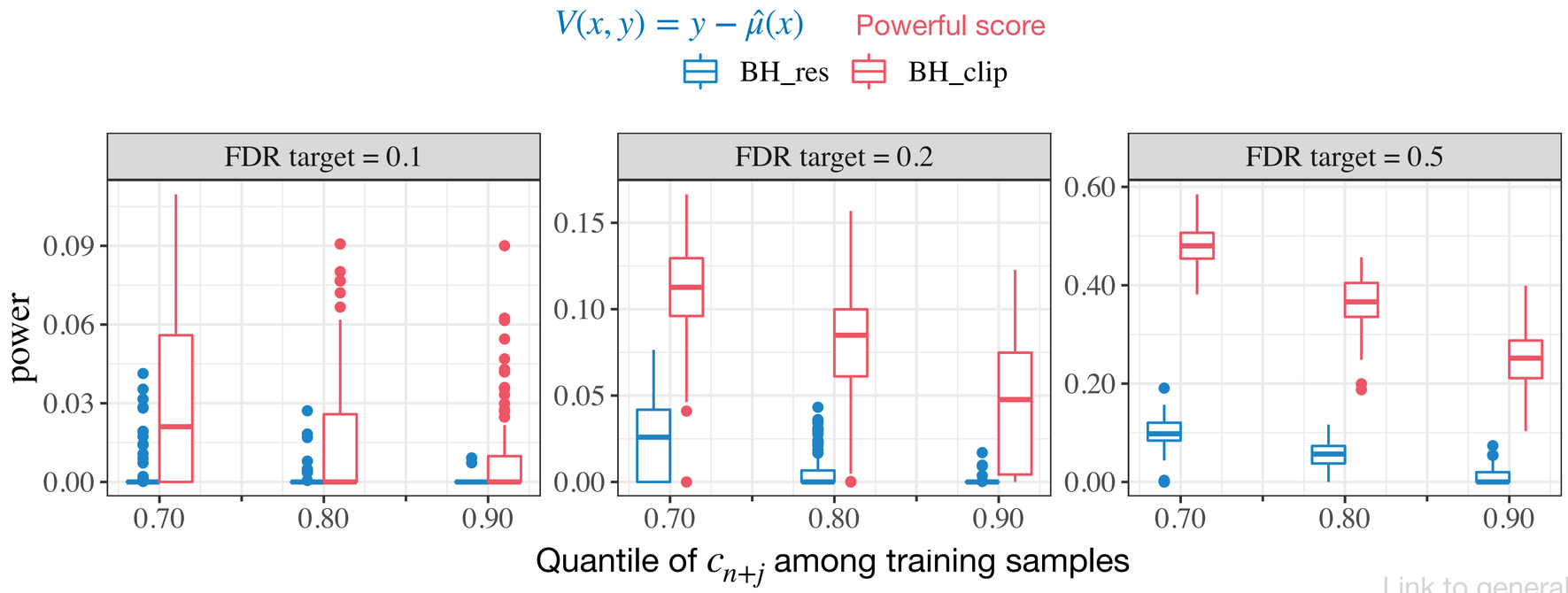
- DAVIS dataset, $Y \in \mathbb{R}$ continuous binding affinities, X feature for drug-target pairs
- $n_{tot} = 30060$ drug-target pairs in total, $2:2:6$ split
- $c_{n+j} = \{0.7, 0.8, 0.9\}$-th quantile of affinities for training pairs with same binding target as j



Link to generalized score

Real data: finding highly-binding drug-target pairs

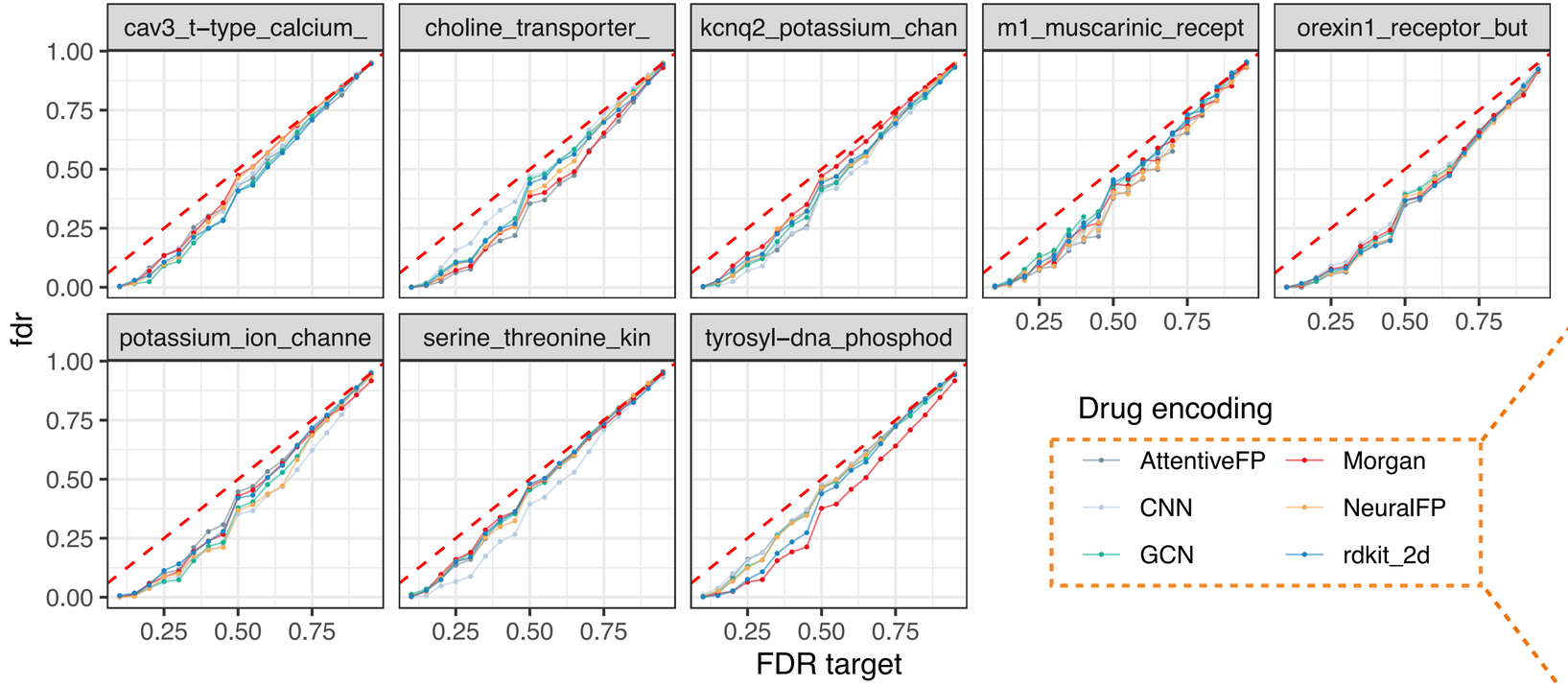
- DAVIS dataset, $Y \in \mathbb{R}$ continuous binding affinities, X feature for drug-target pairs
- $n_{tot} = 30060$ drug-target pairs in total, $2:2:6$ split
- $c_{n+j} = \{0.7, 0.8, 0.9\}$-th quantile of affinities for training pairs with same binding target as j



Link to generalized score

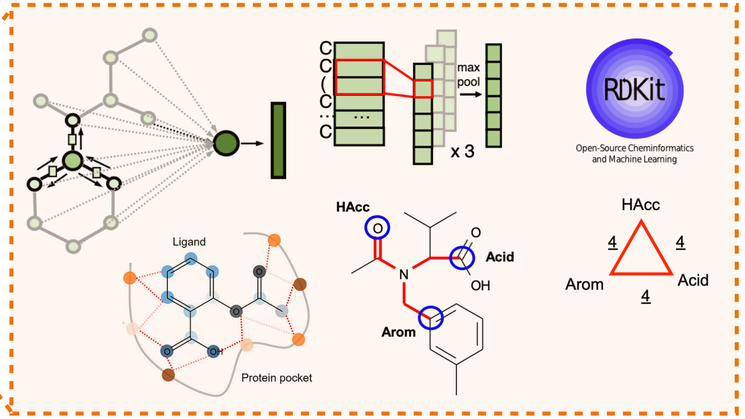
Real data: “needle in the haystack”

- High throughput screening: usually $\approx 0.1\,\%$ active among ~ 100k drugs
- Can narrow down to hundreds of drugs while controlling the FDR



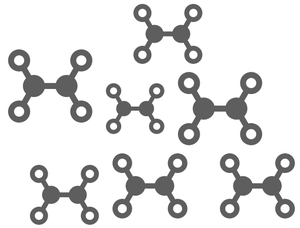
Genentech BIO ONCOLOGY

many other applications

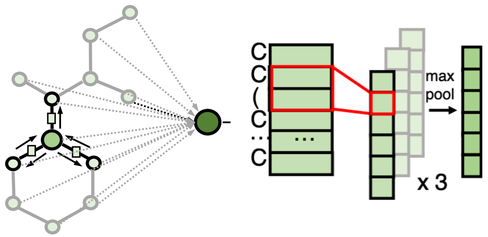


Summary for i.i.d./exchangeable case

Candidate pool

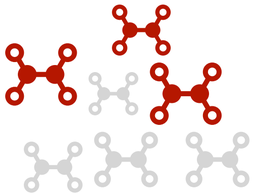


Prediction machine



Conformal p-values ~ confidence measure

Benjamini-Hochberg ~ calibrate threshold



FDR controlled!

- ✓ Arbitrary prediction model
- ✓ Arbitrary data distribution
- ✓ Random thresholds
- ✓ Dependent data points

Part II: Addressing distribution shift

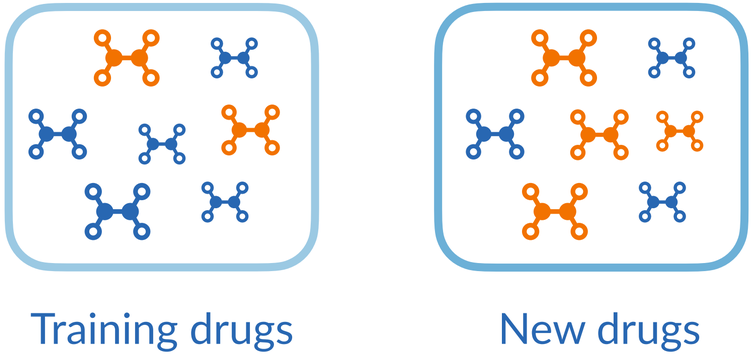
Jin, Y. and Candès, E.J., 2023.

Model-free selective inference under covariate shift via weighted conformal p-values.

arXiv preprint arXiv:2307.09291.

Distribution shift

- Are my evaluated drugs comparable to the unknown drugs?
  - **No** if you preferred drugs with some specific structures, etc



- So far: valid for synthetic-to-synthetic, or well-controlled experiments
- In reality: distribution shift when generating/exploring new drugs
  - ⇝ Similar issues in job hiring, health monitoring, counterfactual inference…

Model-free selective inference under covariate shift

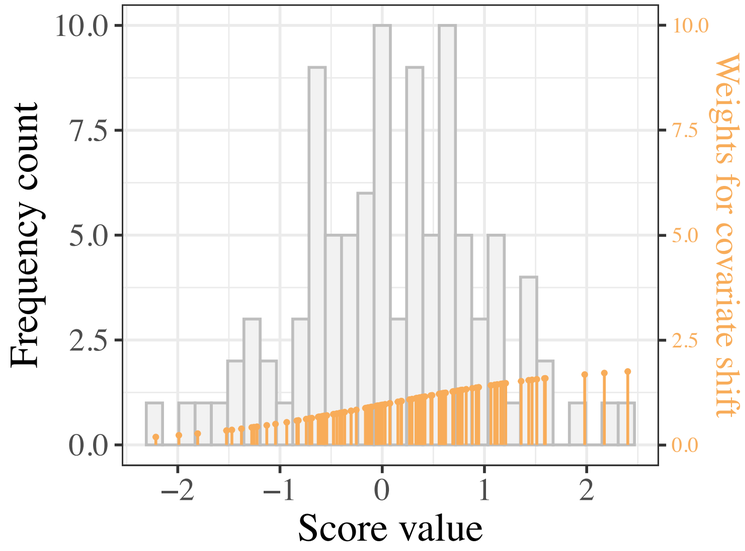
- Test data $\{(X_{n+j}, Y_{n+j})\} \sim \mathbb{Q}$ (unknown)
- Covariate shift: training data $\{(X_i, Y_i)\} \sim \mathbb{P}$ obeying

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(x, y) = w(x)$$

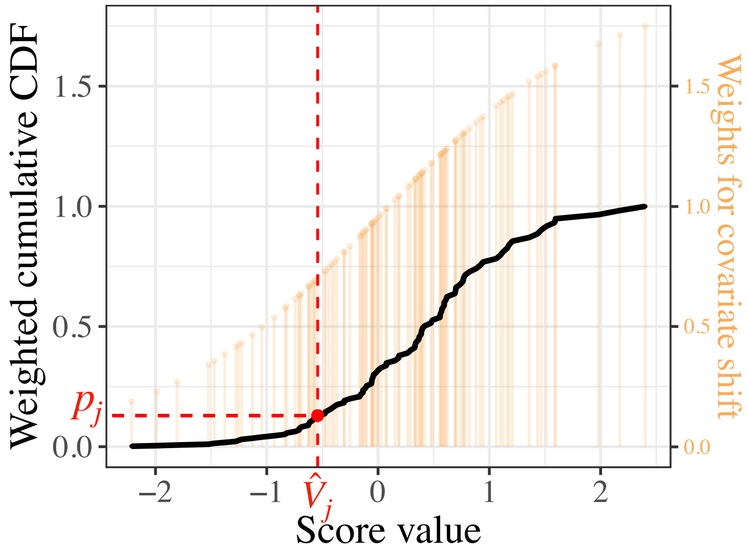
for some (known or estimable) weight function $w\colon \mathcal{X} \to \mathbb{R}^+$ [Sugiyama et al., 2007, Tibshirani et al., 2019]

- **Why? Training data collected by looking at X (drugs, job applicants...)**
- Still want to find test samples $Y_{n+j} > c_{n+j}$ with FDR control

Obtaining valid confidence measures



Histogram of scores and weights in orange



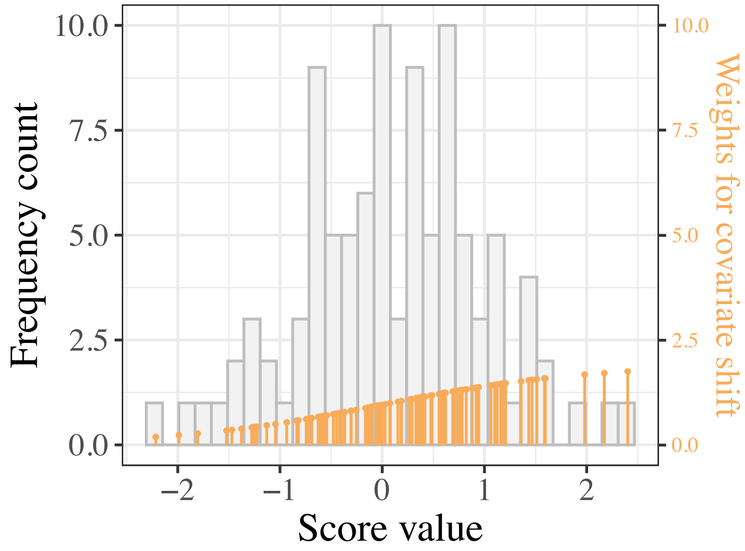
Using weighted ecdf to construct p-values

Weighted conformal p-values

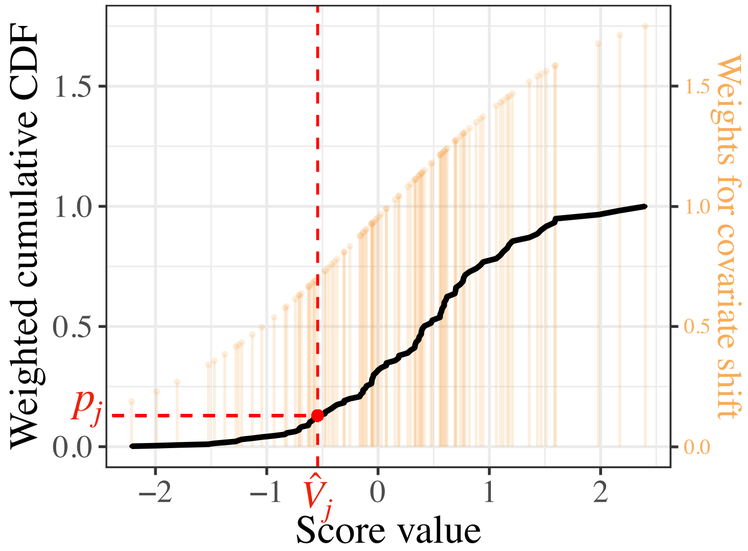
$\approx$ weighted rank of $\hat{V}_{n+j}$ among training scores $\{V_i\}_{i=1}^n$

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

Obtaining valid confidence measures



Histogram of scores and weights in orange



Using weighted ecdf to construct p-values

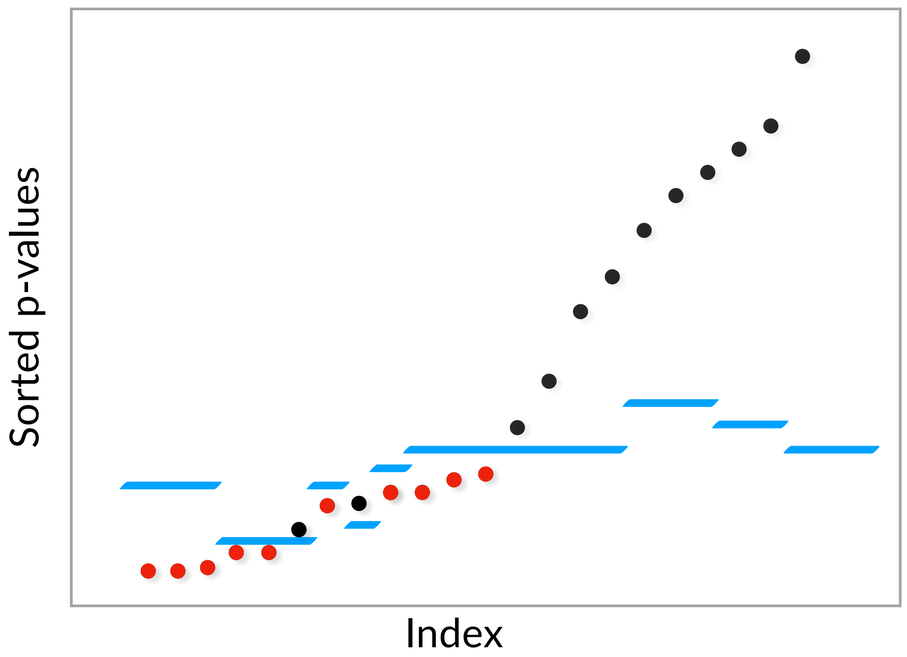
Well-calibrated p-values:

$$\mathbb{P}(p_j \leq t,\ Y_{n+j} \leq c_{n+j}) \leq t, \quad \forall t \in [0,1]$$

~ Valid p-values from <u>weighted</u> rank test

Harnessing difficult dependence by new procedure

Weighted conformal p-values are no longer positively dependent!



Data-dependent evidence levels τ_j

⇝ Extends [Fithian and Lei, 2021]

⇝ Related to e-values [Wang and Ramdas, 2022]

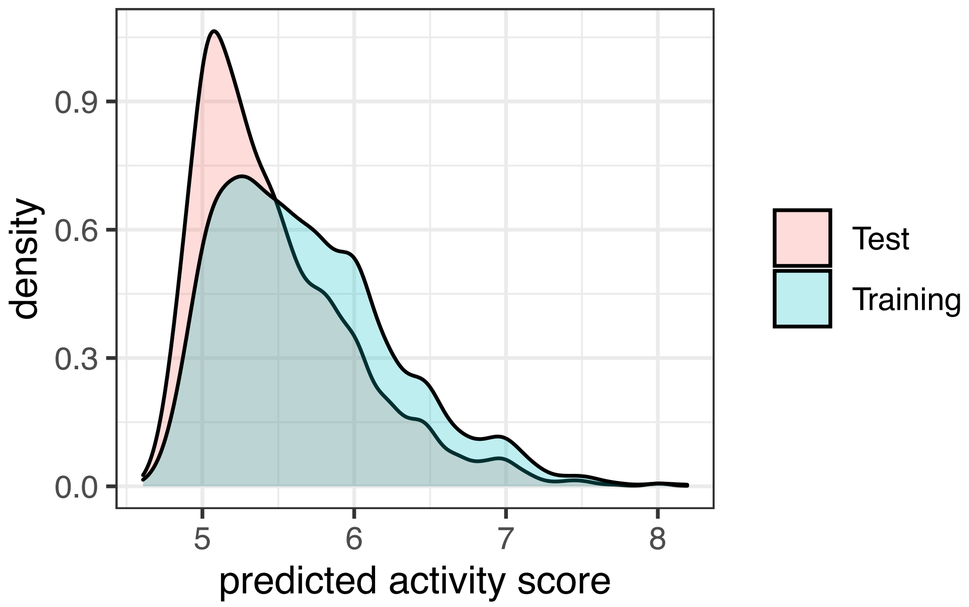
Previously: select if p_j below a common data-dependent level τ

Now: select if p_j below data-dependent level τ_j adapted to each drug

[Detailed method] [Detailed theorem]

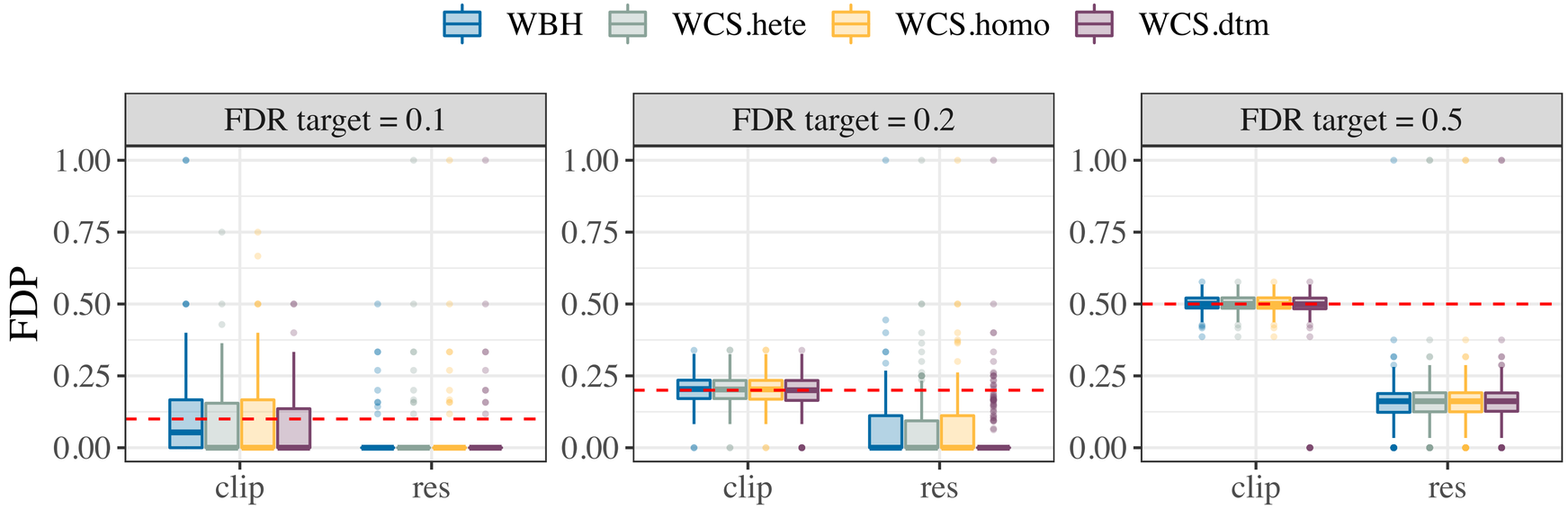
Real data: drug-target-interaction under biased sampling

- DAVIS dataset, $Y \in \mathbb{R}$ continuous binding affinities, X feature for drug-target pairs
- $n_{tot} = 30060$ drug-target pairs in total
- Covariate shift created by preferring high-prediction drugs in training data
- $c_{n+j} = 0.8$-th quantile of affinities for training pairs with same binding target as j



Real data: drug-target-interaction under biased sampling

- DAVIS dataset, $Y \in \mathbb{R}$ continuous binding affinities, X feature for drug-target pairs
- $n_{tot} = 30060$ drug-target pairs in total
- Covariate shift created by preferring high-prediction drugs in training data
- $c_{n+j} = 0.8$-th quantile of affinities for training pairs with same binding target as j



Real applications and shifts

Selection Task (Distribution Shift)

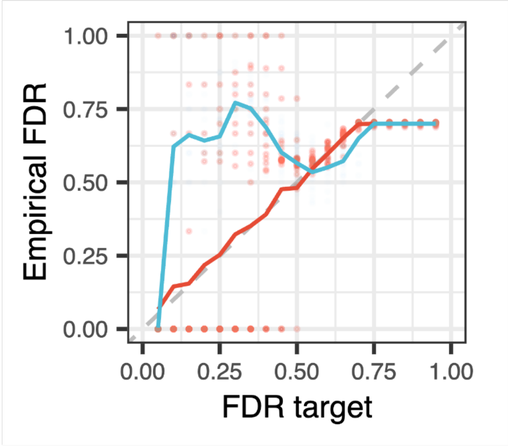
Selection FDR Control

Conformal-Select

Baseline

Covariates: learned representation in the hidden layer of neural nets

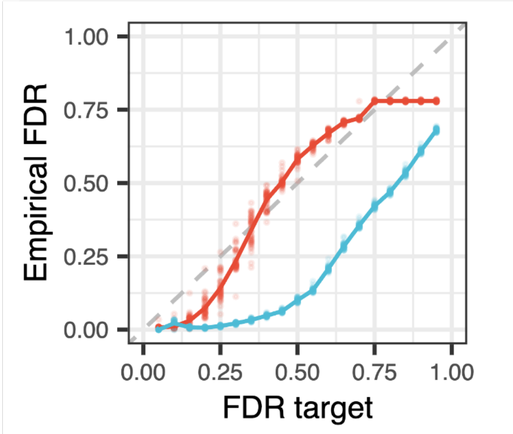
Select gene perturbations with high T-cell proliferation (Uniform)



1: **Gene perturbation selection**

- Experimental setup without shift

Select proteins with high stability (Mutant shift)



2: **Protein stability selection**

- Shift from proteins in four rounds of experiments to single-mutation proteins

Real applications and shifts

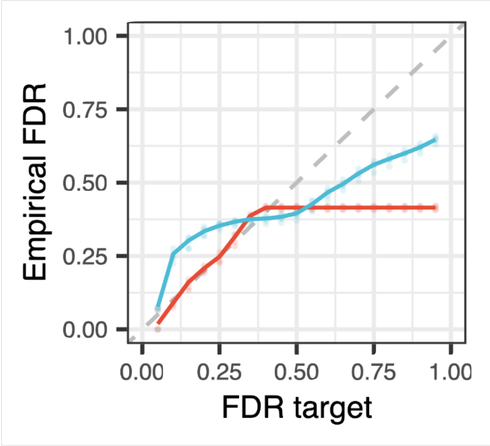
Selection Task (Distribution Shift)

Selection FDR Control

— Conformal-Select
— Baseline

Covariates: learned representation in the hidden layer of neural nets

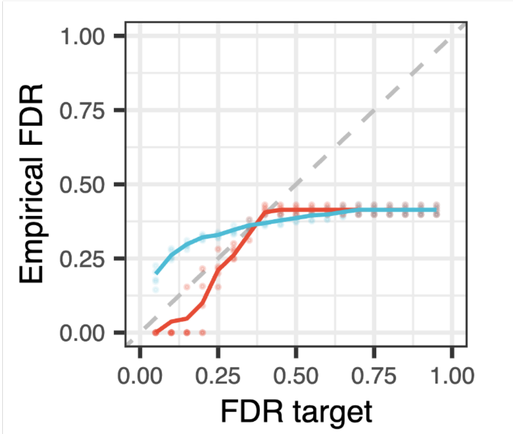
Select compounds with low CYP2C9 inhibition rate (Scaffold shift)



3: **Drug property selection**

- Shift in drug structure (scaffold)

Select clinical trials that meet primary outcome (Temporal shift)

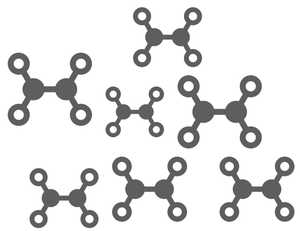


4: **Trial outcome prediction**

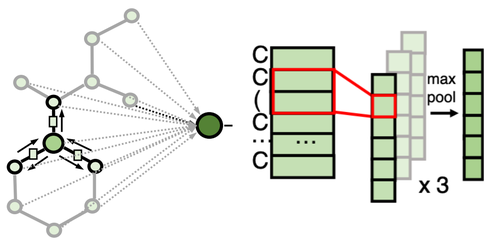
- Shift from earlier to future trials

Summary for covariate shift case

Candidate pool

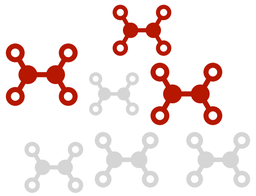


Prediction machine



Conformal p-values ~ confidence measure

New testing method ~ calibrate threshold



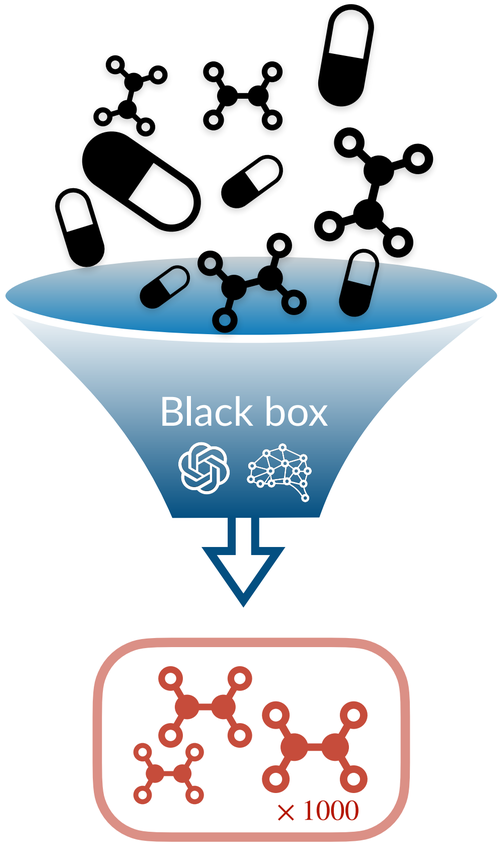
FDR controlled!

- ✓ Arbitrary prediction model
- ✓ Arbitrary data distribution
- ✓ Random thresholds
- ✓ Dependent data points
- ✓ Robust to distribution shift!

Summary

- Controlling FDR is sensible and interpretable
- Novel methods that turn any prediction model into reliable selections
- Can deal with covariate shifts $\rightsquigarrow$ novel testing procedures

Bridge between selective and model-free inference



Focal set with
90% active drugs