Ranking targets with desirability functions and latent variable models

22 Sept 2022

Stanley E. Lazic, PhD
stan.lazic@prioris.ai
@stanlazic

Prioris.ai
When predictions matter

Problem

- Integrating diverse data is key to identifying and ranking targets.
- How best to do it?

Target	Assay1	Type	Tissue Expression	ML prediction	Patented
T1	37.81	Enzyme	374	0.79	Yes
T2	2.11	Ion channel	25690	0.09	No
T3	28.39	Structural	3287	0.41	No
...					

Problem

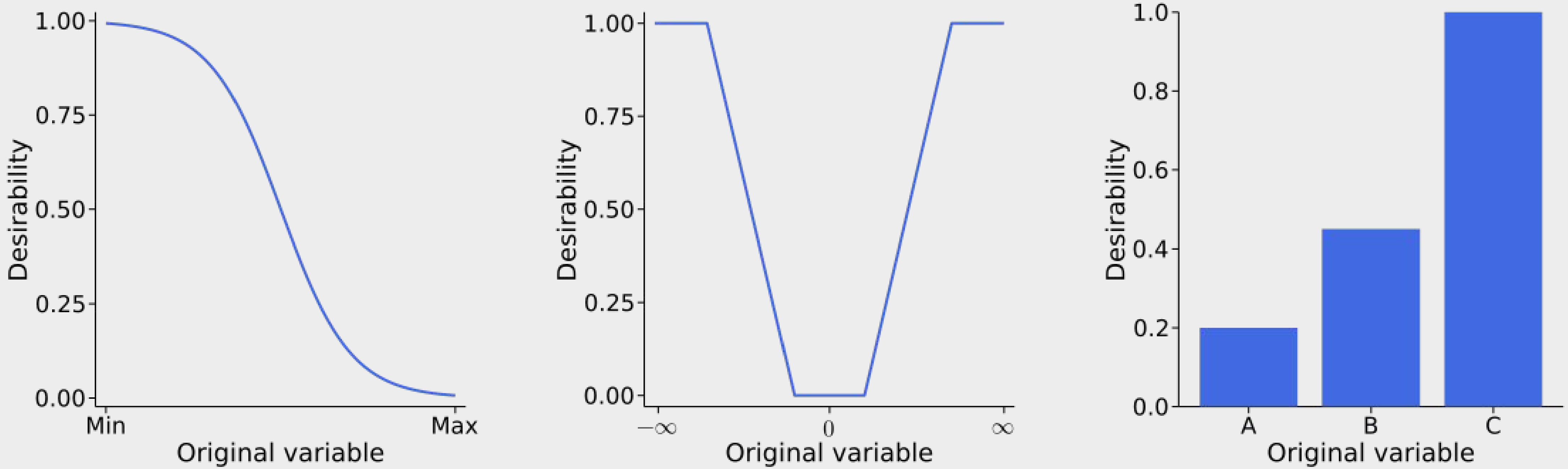
- Integrating diverse data is key to identifying and ranking targets.
- How best to do it?

Target	Assay1	Type	Tissue Expression	ML prediction	Patented
T1	37.81	Enzyme	374	0.79	Yes
T2	2.11	Ion channel	25690	0.09	No
T3	28.39	Structural	3287	0.41	No
...					

- Filter? Ignores uncertainty and all variables treated equally.

A solution: desirability functions

- Map data (assay values, target properties, etc.) to a common scale from 0 to 1 by how well they meet criteria or have useful properties.



Lazic SE (2015). Ranking, selecting, and prioritising genes with desirability functions. *PeerJ* 3:e1444.

A solution: desirability functions

- Calculate the overall (weighted) desirability for each target.
- Weights are set according to a variable's relevance.

Target	Assay1 (w = 1.0)	Type (w = 0.5)	Tissue Expression (w = 0.95)	ML prediction (w = 0.2)	Patented (w = 0.1)	*D*
T1	0.62	0.99	0.2	0.89	0.2	**0.45**
T2	0.98	0.99	0.83	0.02	1.0	**0.70**
T3	0.77	0.5	0.71	0.52	1.0	**0.68**
...						

A solution: desirability functions

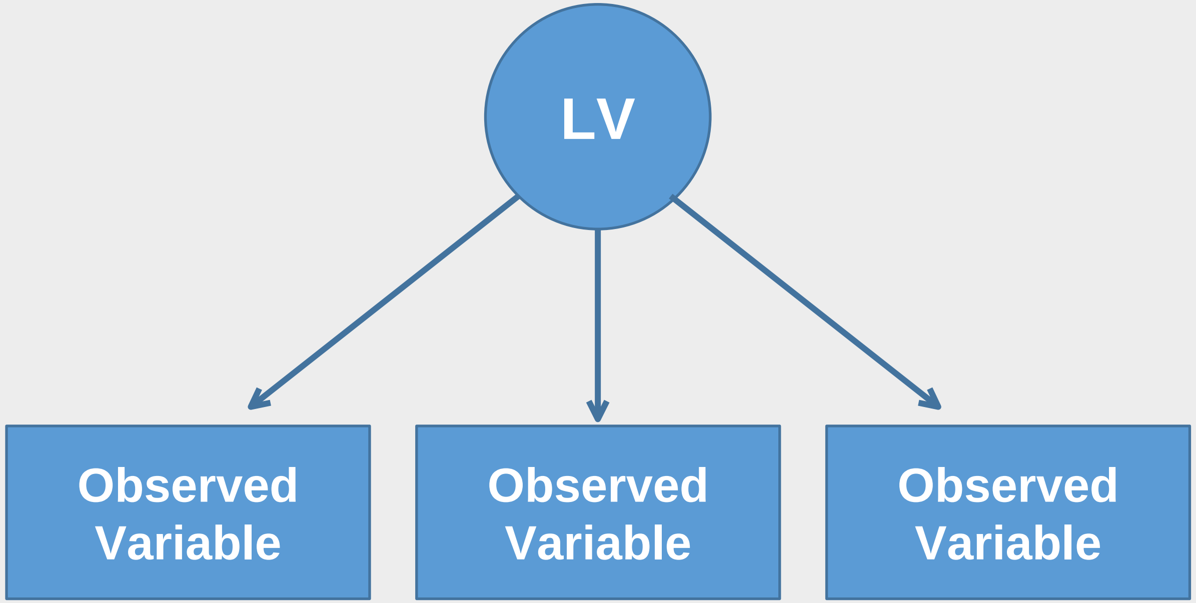
- Calculate the overall (weighted) desirability for each target.
- Weights are set according to a variable's relevance.

Target	Assay1 (w = 1.0)	Type (w = 0.5)	Tissue Expression (w = 0.95)	ML prediction (w = 0.2)	Patented (w = 0.1)	*D*
T1	0.62	0.99	0.2	0.89	0.2	**0.45**
T2	0.98	0.99	0.83	0.02	1.0	**0.70**
T3	0.77	0.5	0.71	0.52	1.0	**0.68**
...						

- How certain are we that T2 is better than T3?
- Is a weighted geometric mean the best way to combine values?
- What about missing values?

Latent variable models

- Treat each targets' suitability as a latent variable.
- Estimate suitability based on observed data using a Bayesian latent variable model → provides probabilistic estimates for each target.



Based on Item Response Theory models

- Used in psychometrics to estimate people's latent ability or knowledge.
- Rows are people.
- Columns are items/questions.
- Entries in table are correct/incorrect answers.
- Bonus: can also estimate the latent difficultly of each question.

Person	Q1	Q2	Q3	Q4	...
P1	1	0	1	1	
P2	0	1	0	1	
P3	1	0	0	1	
...					

Adaptations

- Desirability scores are not binary, but continuous values between zero and one.
- "Discrimination" parameters are not estimated, but fixed, and equal to the variable weights.

Model details

- y = data matrix
- w = fixed weights (one for each variable).
- t = index for target (1 to number of targets).
- v = index for weights (1 to number of variables).
- θ = latent suitability parameters (one for each target).
- d = "difficulty" parameter.

$$\mu_{t,v} = w_v\,(\theta_t - d_v)$$

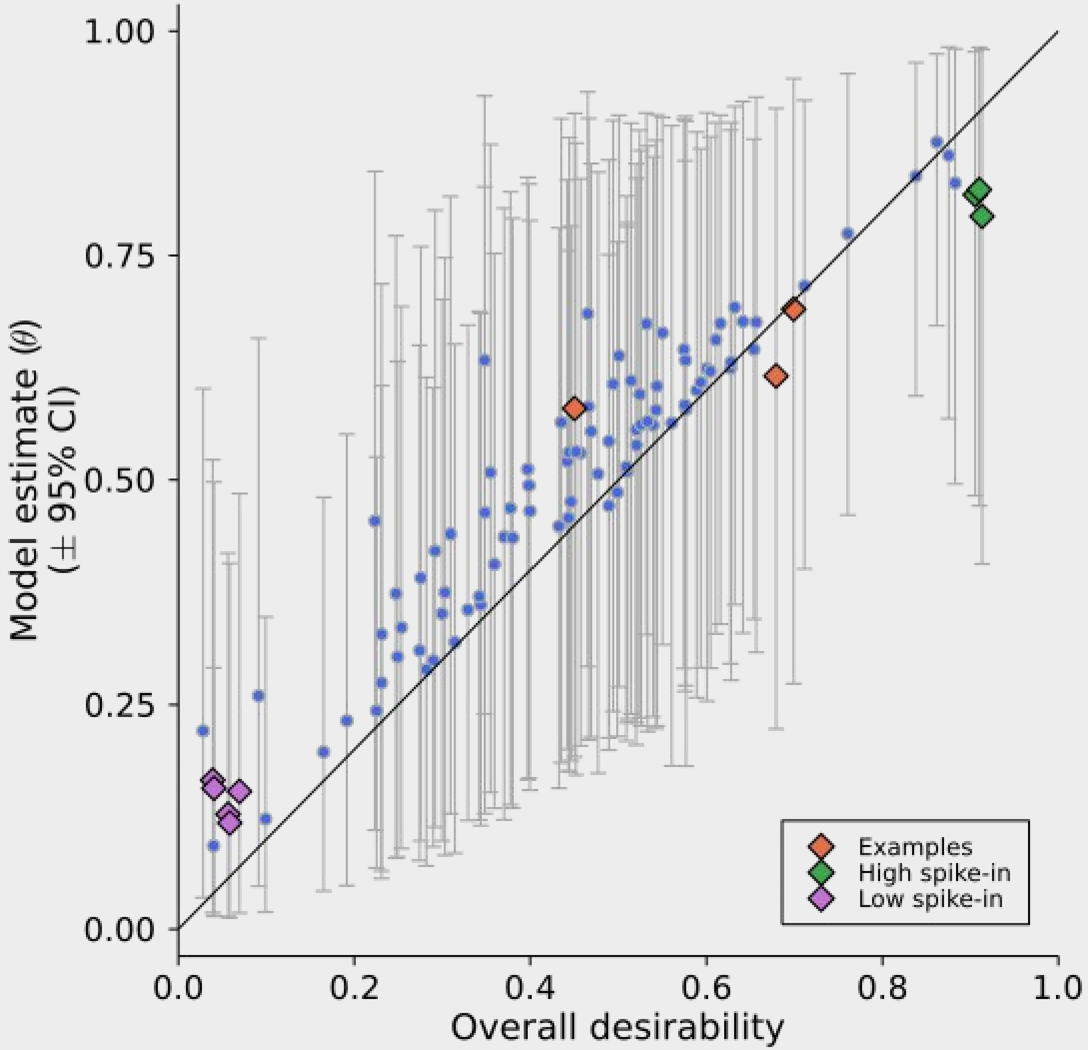
$$P\left(y_{t,v} = 1 \mid \theta, d\right) = \frac{1}{1 + e^{-\mu_{t,v}}}$$

Implementation in Julia and Turing.jl

```
@model model_def(y, w; N_targs = size(y, 1), N_items = size(y, 2)) = begin
    # define priors
    θ ~ filldist(Normal(0, 3), N_targs)
    ϕ ~ filldist(Truncated(Normal(0, 5), 0.01, Inf), N_targs)
    d ~ filldist(Normal(0, 3), N_items)

    for t = 1:N_targs
        for v = 1:N_items
            μ = invlogit(w[v] * (θ[t] - d[v]))

            # transform parameters & enforce constraints
            A = μ * ϕ[t]
            B = (1.0 - μ) * ϕ[t]
            A = A <= 0 ? 0.001 : A
            B = B <= 0 ? 0.001 : B

            y[t, v] ~ Beta(A, B)
        end
    end
end
```

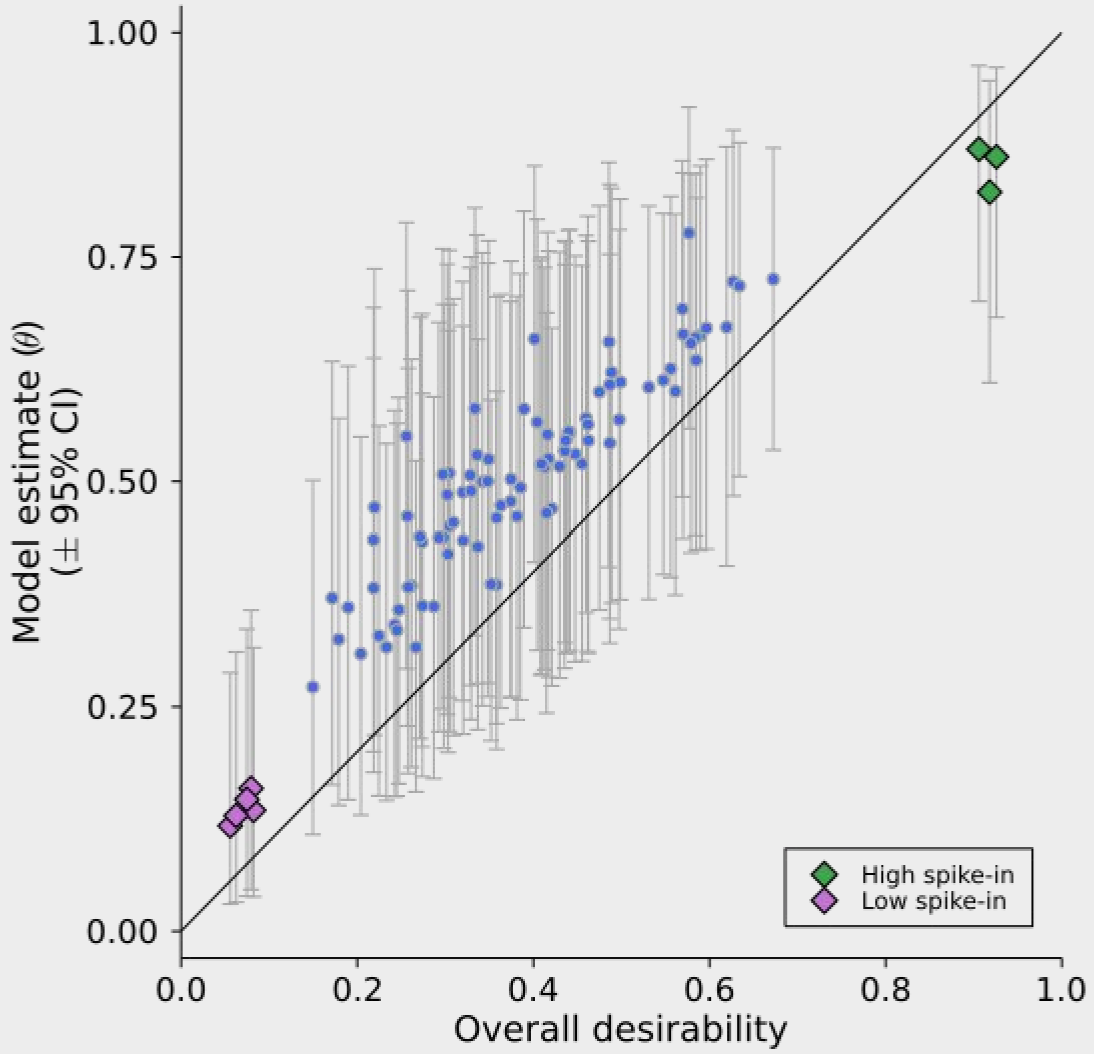
Compare LV & geomean: simulation results

- Simulated random values [0, 1] for 100 targets and **5** variables.
- Simulated “spike-ins” with all low or all high desirability values.



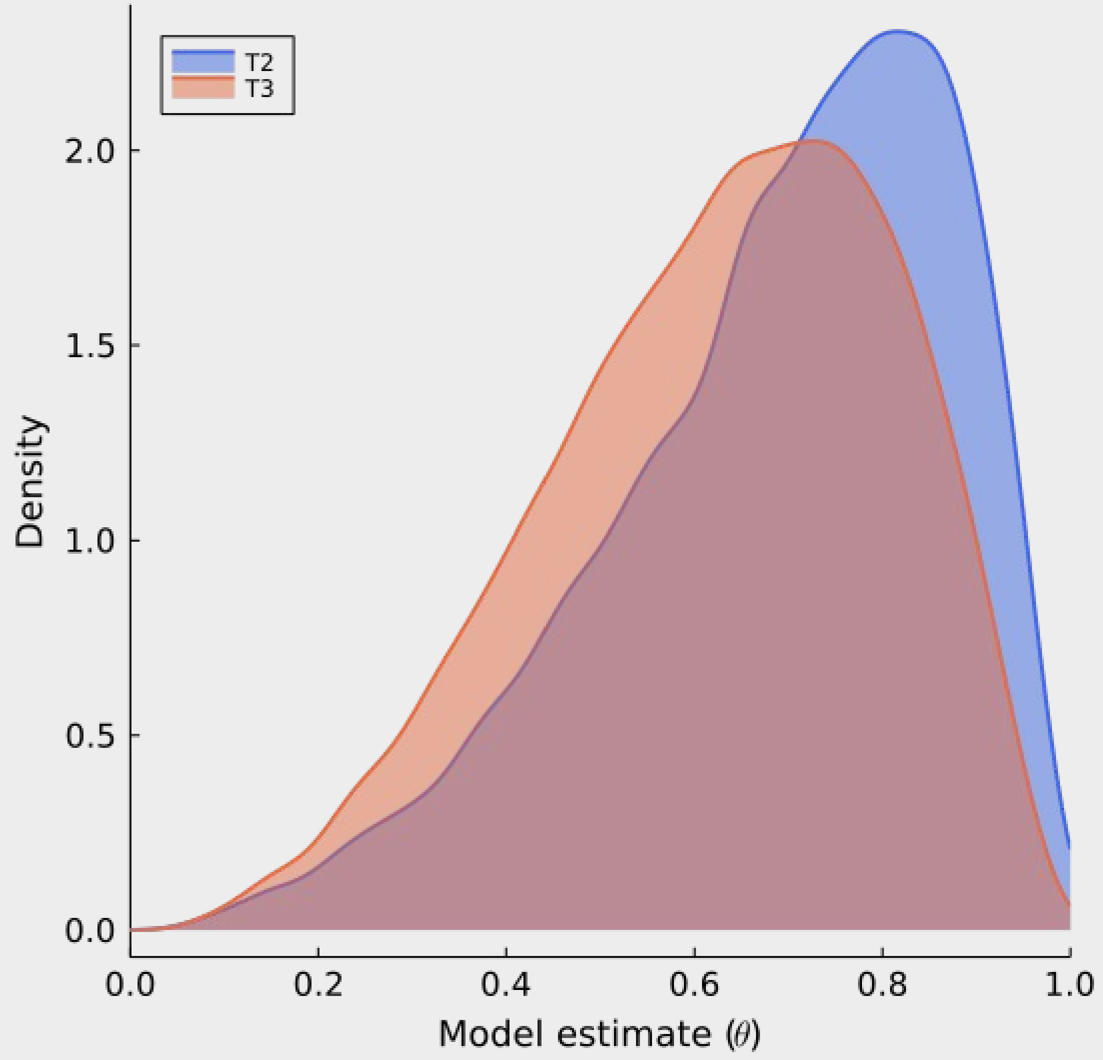
Compare LV & geomean: simulation results

- Simulated random values [0, 1] for 100 targets and **15** variables.
- Simulated “spike-ins” with all low or all high desirability values.

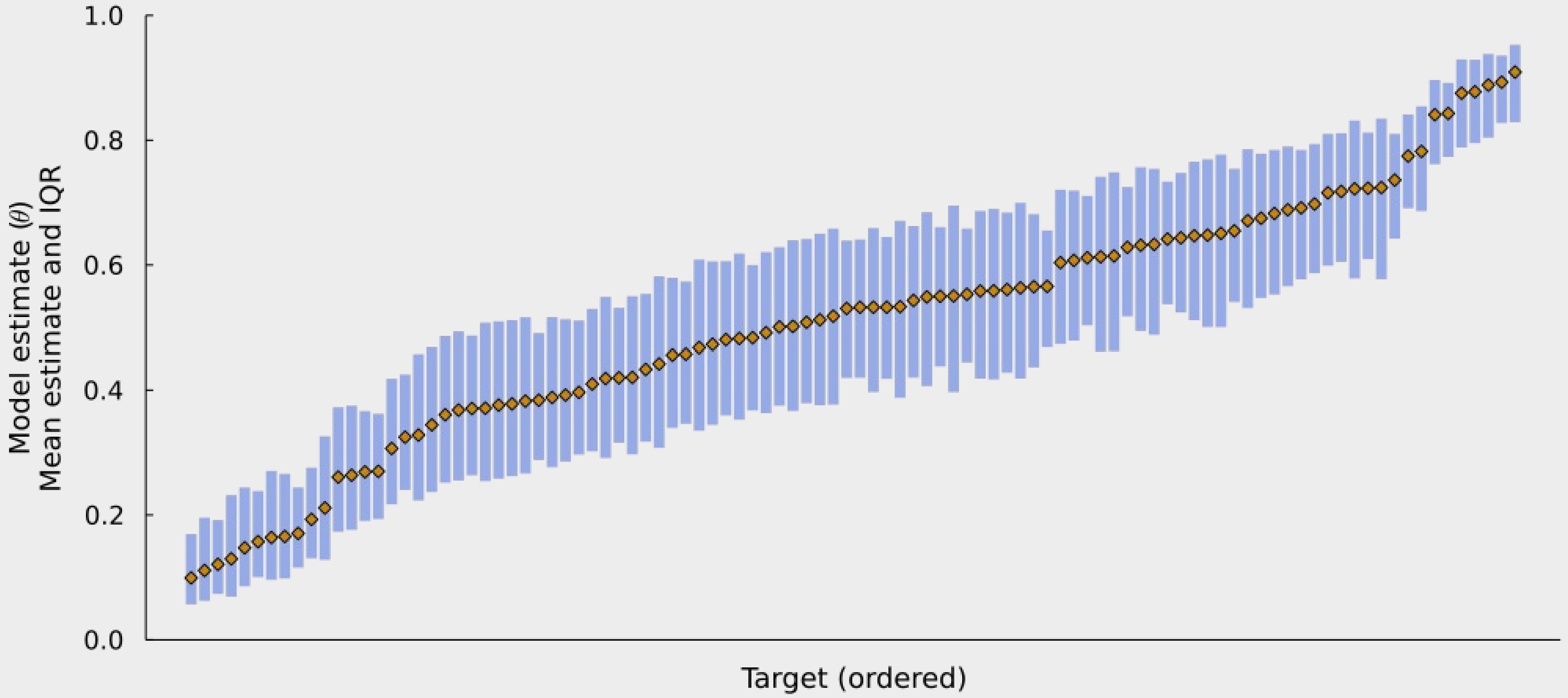


Differentiating between targets

- T2 and T3 had overall desirability scores of 0.70 and 0.68.
- The mean model predictions are 0.69 and 0.63
- P(T2 > T3) = 0.61

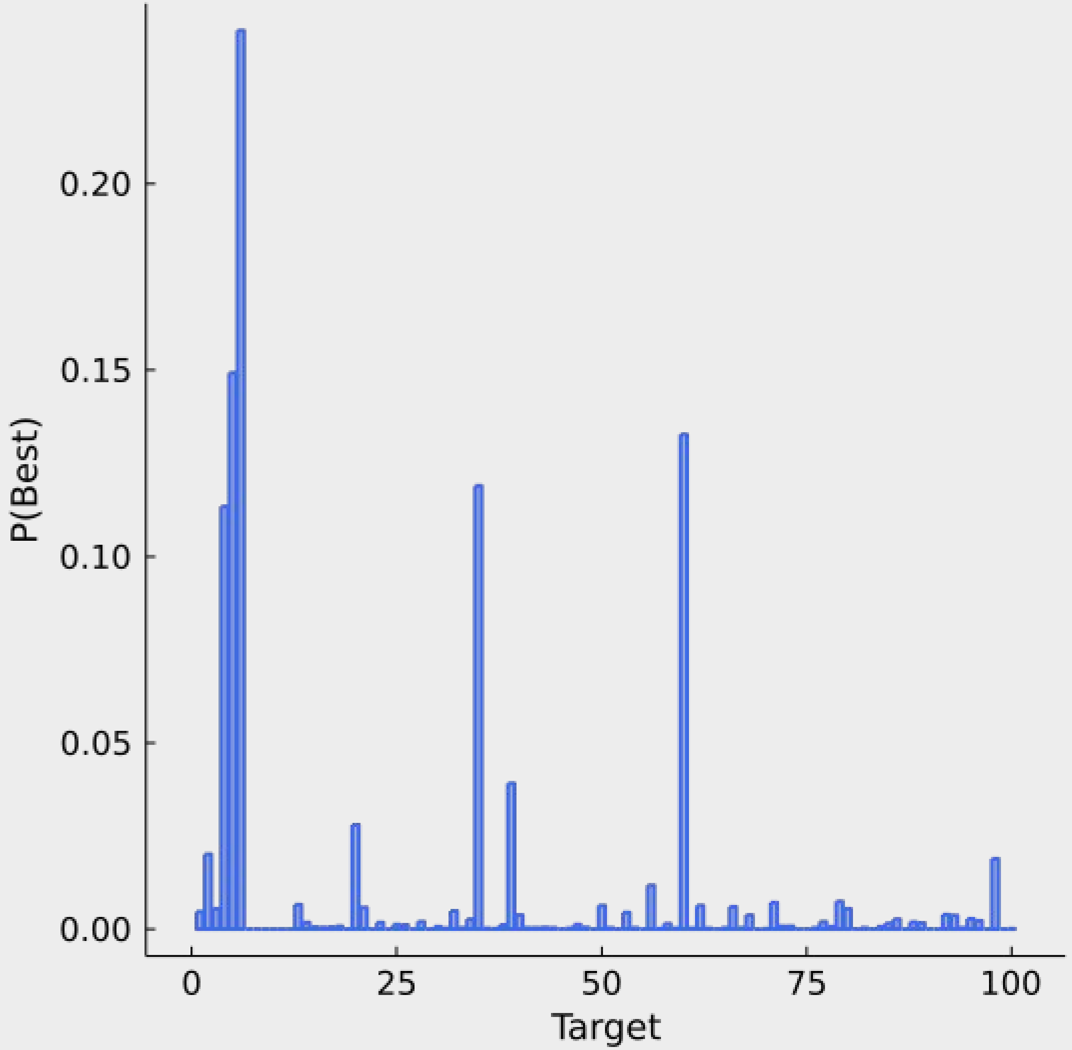


Ranked estimates of target suitability



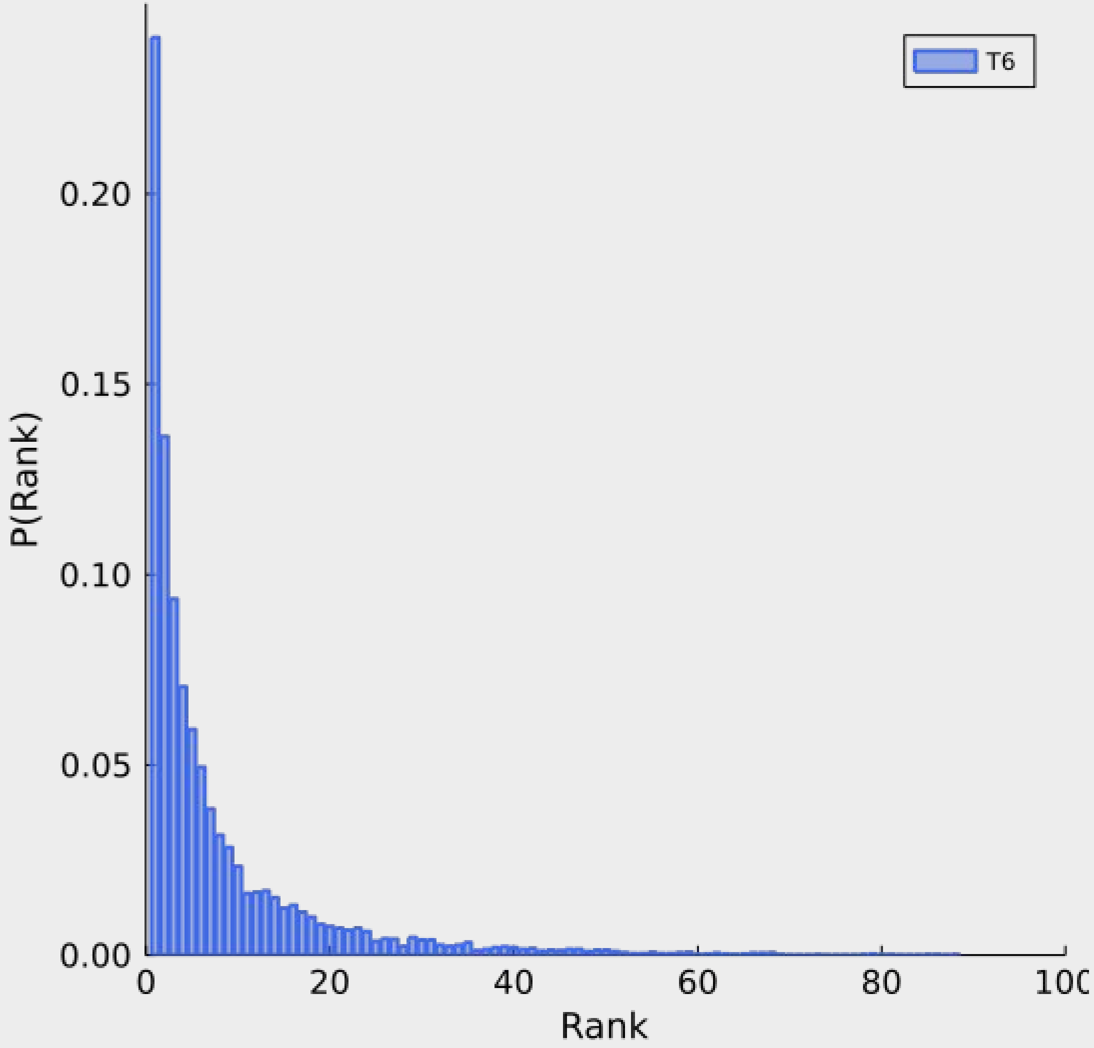
What's the best target?

- P(Best) = 0.24 for T6.



What’s the uncertainty in the ranking

- P(Rank = 1) = 0.24
- P(Top 10) = 0.77



Lazic SE, Edmunds N, Pollard CE (2018). Predicting drug safety and communicating risk: benefits of a Bayesian approach. *Toxicological Sciences* 162(1):89–98.

Missing data

- Use multiple imputation to generate several data sets.
- Run analysis on each data set.
- Combine distributions from each analysis.

Summary

- Latent variable models are an acceptable alternative to the geometric mean for calculating overall desirability/suitability scores:
  - They provide uncertainty in the overall scores, which can help rank targets,
  - And they can easily handle missing data.

Resources

- Lazic SE (2015). Ranking, selecting, and prioritising genes with desirability functions. *PeerJ* 3:e1444
  https://doi.org/10.7717/peerj.1444

- desiR R package on CRAN
  https://cran.r-project.org/web/packages/desiR/index.html

- DesirabilityScores.jl on Github (WIP)
  https://github.com/stanlazic/DesirabilityScores.jl

Acknowledgments

- Gabriel Phelan