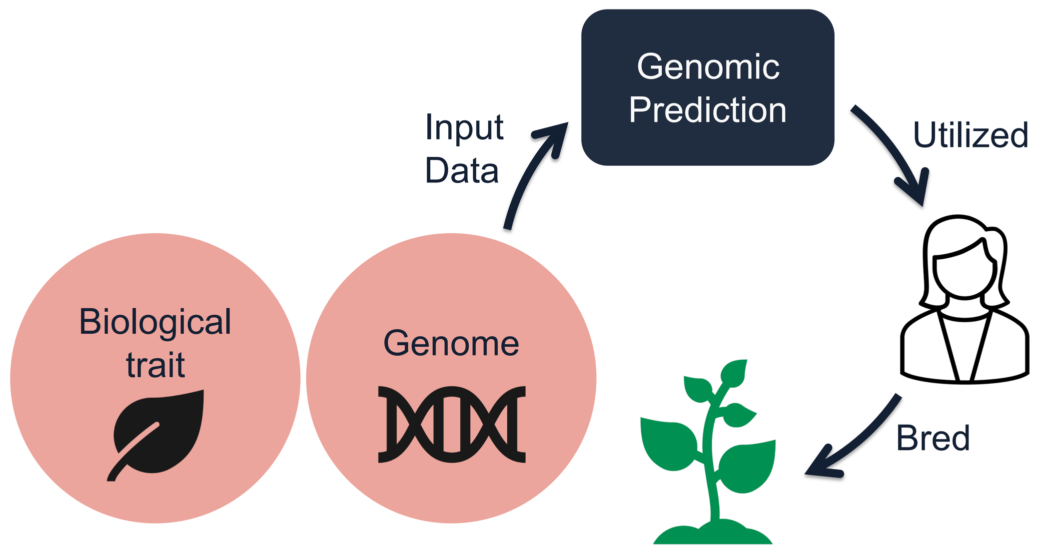
Towards Automating Highly Heritable Phenotype Discovery For Plant Breeding

Ruhana Azam[1,3], Samuel B. Fernandes[2], Mohammed El-Kebir[1], Sanmi Koyejo[3], Alexander E. Lipka[1], Andrew D.B. Leakey[1]

(1) University of Illinois at Urbana-Champaign, (2) University of Arkansas, (3) Stanford University

Motivation

- Plant breeders utilize statistical methods (e.g. Genomic Wide Association Study) to predict highly desirable traits in plants.



- Genomic prediction models are trained on biological traits (e.g. nitrogen per leaf area) which are **expensive and labor intensive** to measure.
- Alternatively, prediction models can be trained using low-cost traits (e.g. hyperspectral data) that are correlated with biological traits of interest. [1]
- **Highly-heritable low-cost traits can improve genomic prediction accuracy of high-cost, low-heritability traits [1]**

Background

What is heritability?

- Heritability is the portion of population variance explained by genetic factors
  - ➢ Ideal to breed for traits reliant on genetic factors.
- Heritability (h^2) is calculated via variance factors from fitting mixed models by using genetic and environmental factors to predict the trait of interest (t).

$$h^2(\sigma_t^2, \sigma_{\text{gene x env}}^2) = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_{\text{gene x env}}^2 + \sigma_{\text{error}}^2}$$

What are synthetic traits?

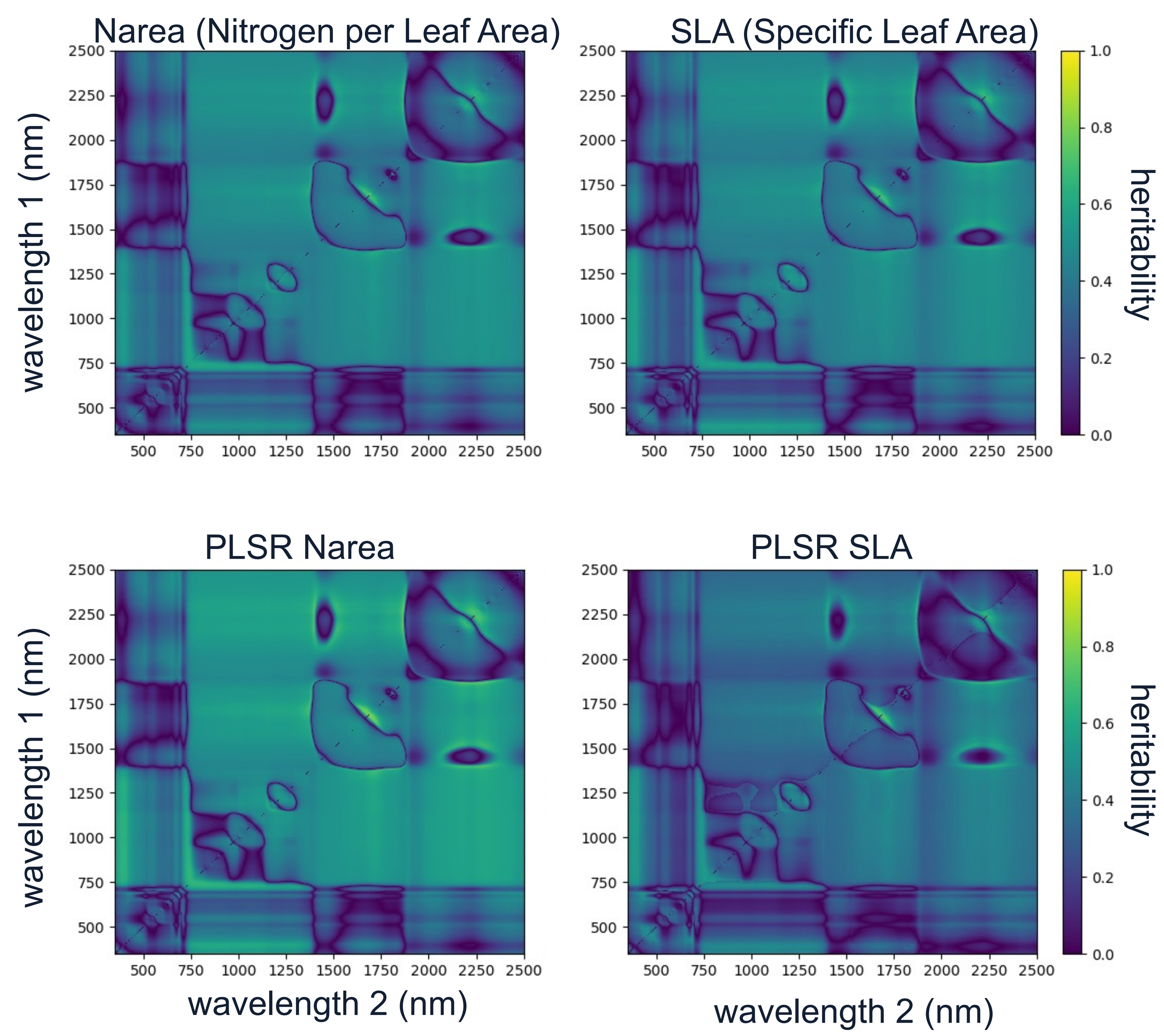
- Synthetic traits are functional combination of multiple low-cost traits (e.g. t_1/t_2 , $t_1 + t_2/t_3$)
- **Search space grows exponentially with function complexity!**

Problem: $\text{argmax}_{t_1 \ldots t_n} h^2(t_1 \ldots t_n)$

How do we discover highly-heritable synthetic traits in large trait spaces?

Experimental Setup

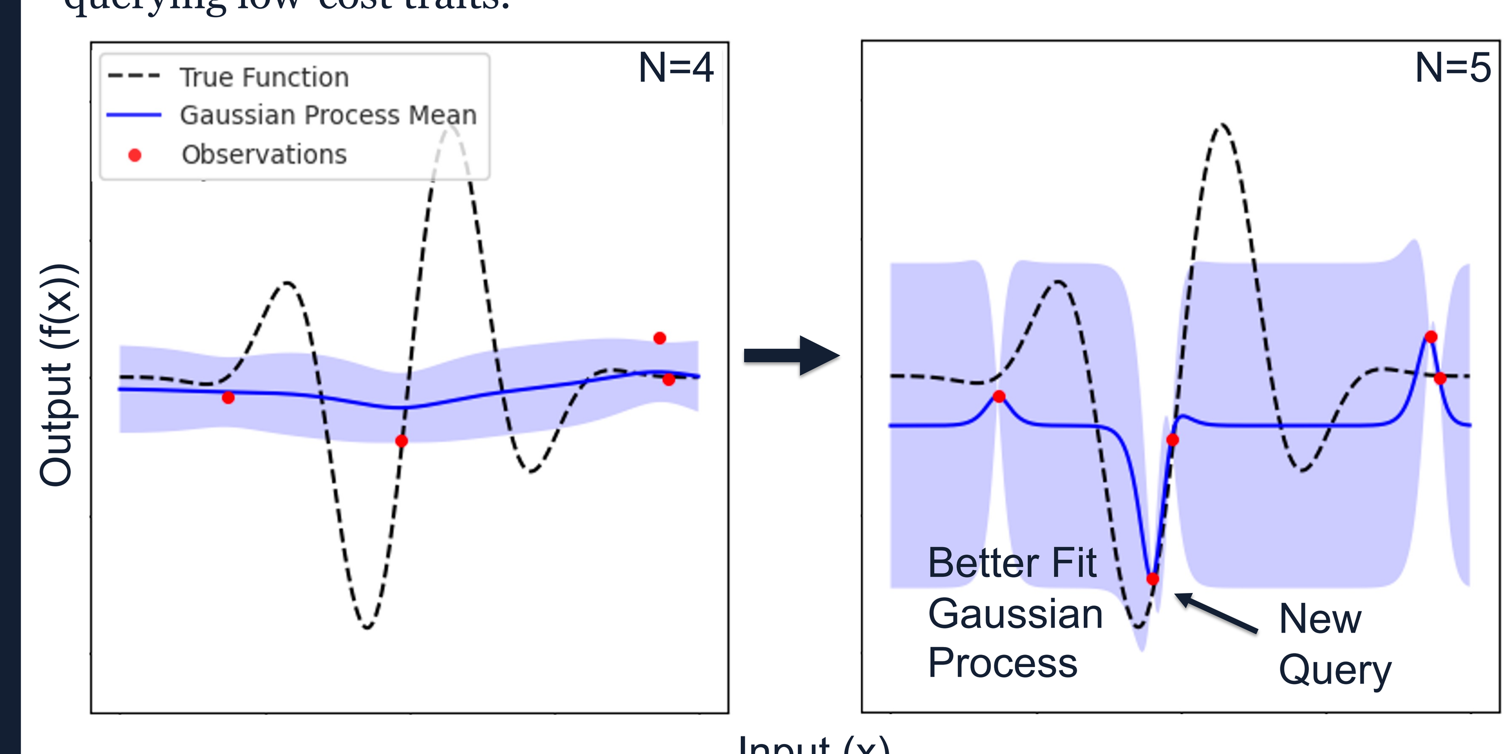
Dataset: 836 Sorghum lines, 2 Locations
Baseline: Grid Search (Brute-Force), Random Search
Evaluation: Number of Queries, Time (CPU: Xeon E7-4870)
Synthetic Trait: Wavelength ratios (t_1/t_2)
Search Space: Smooth in large regions, sparse in others



(*PLSR = Partial Least Squared Regression)

Approach

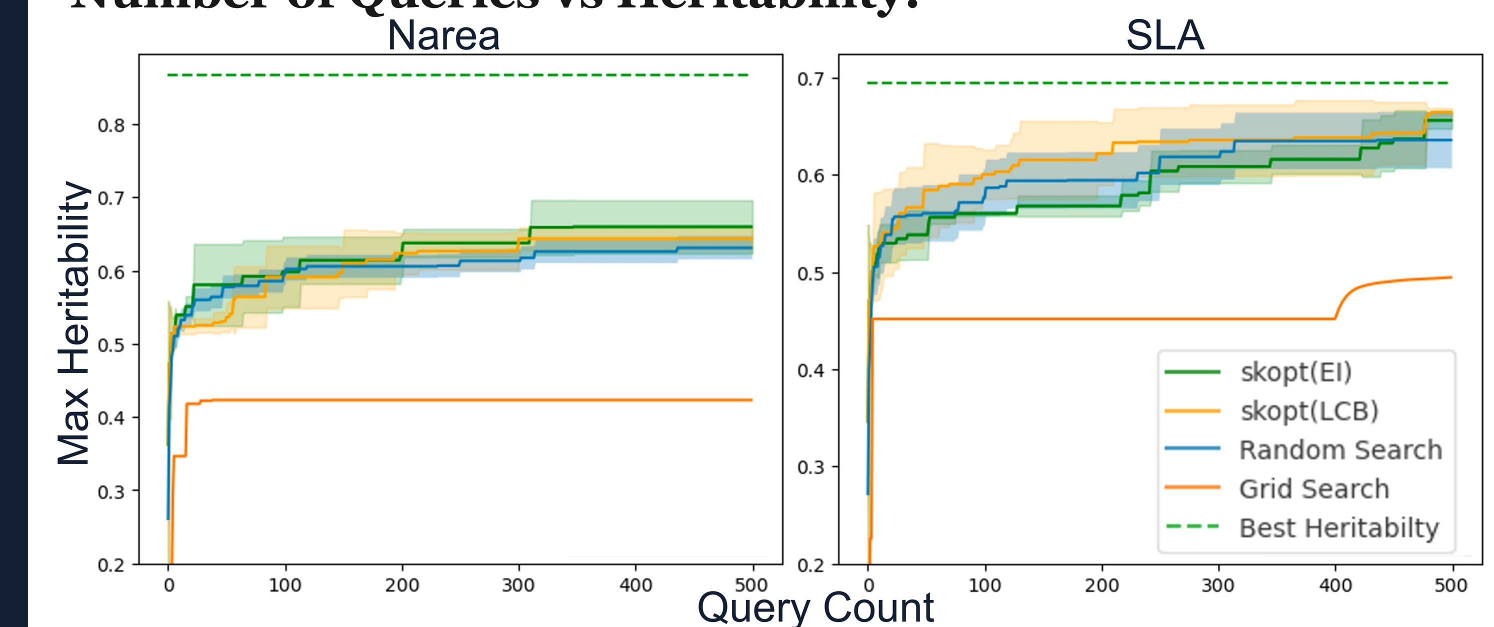
Black box Search with Bayesian Optimization (BO)
Blind to the heritability function initially and learns the function from querying low-cost traits.



1) Uses Gaussian Process as surrogate to model an unknown function
2) Acquisition function, such as expected-improvement (EI), balances issues of **exploration-exploitation** to find the best next query.

Preliminary Results

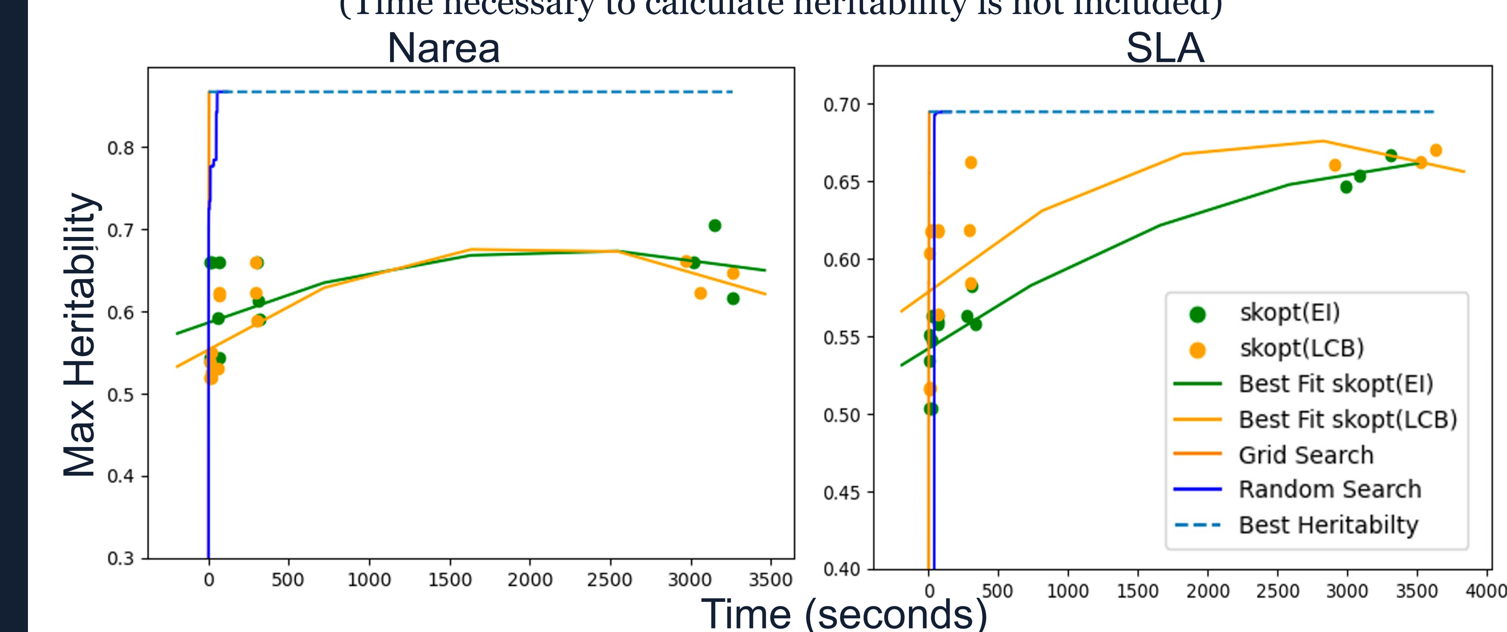
Number of Queries vs Heritability:



➢ Bayesian Optimization methods marginally outperform random search.

Time vs Heritability :

(Time necessary to calculate heritability is not included)



➢ Baseline methods find highly-heritable traits significantly faster than Bayesian Optimization methods.

(Results above are representative of trends shown by other traits tested)

Conclusion

- When searching wavelength ratio (t_1/t_2), random search shows best trade-offs over query number and time.
- Bayesian optimization performs well when comparing number of queries.
- Unclear if random search will be exasperated in larger search spaces.

Future Work:

- ➢ Expand search to larger function classes (e.g. t1 + t2 / t3)
- ➢ How to deal with search problems where search space and query time are both large?

References

[1] Fernandes, S. B., Azam, R. N., Paul, R. E., Yuan, M., El-Kebir, M., Koyejo, S., Lipka, A. E., Leakey, A. (2023). Including High-Throughput Phenotyping Derived Traits in Multi-Trait Genomic Analysis. Presented at the CSSA: Translational Genomics Workshop, Plant and Animal Genome XXI Conference.

ILLINOIS