

MINLP for regularized symbolic regression with applications to data-driven modeling of critical minerals processes



Dimitrios Fardis¹, Radhakrishna Tumbalam Gooty^{2,3}, Nick Sahinidis^{1,4}

¹School of Chemical & Biomolecular Engineering, Georgia Institute of Technology

²National Energy Technology Laboratory (NETL)

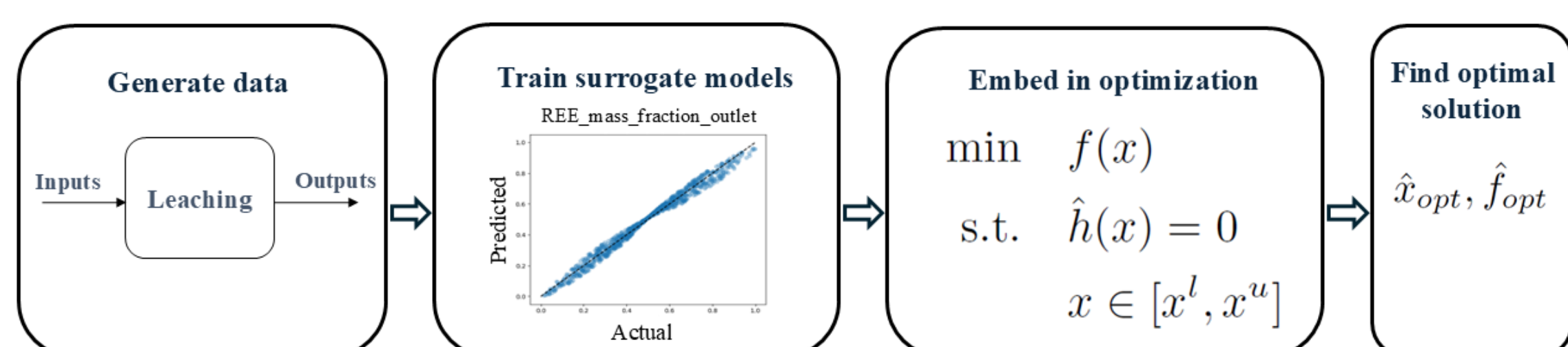
³NETL Support Contractor

⁴H. Milton School of Industrial and Systems Engineering, Georgia Institute of Technology



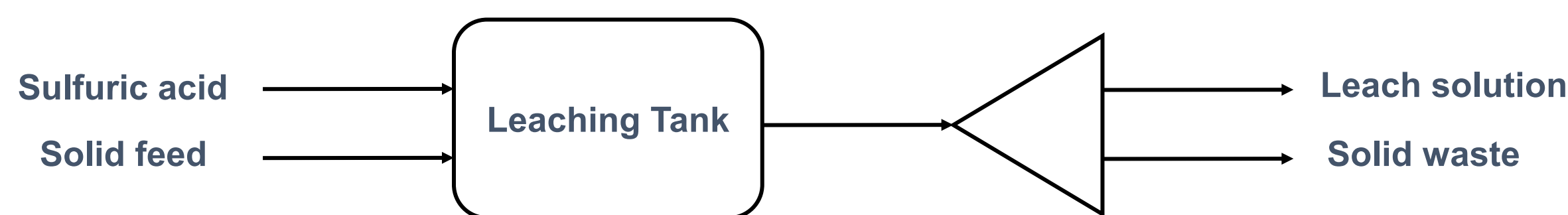
Data-driven modeling and optimization

What data-driven techniques are the most appropriate to model and optimize critical mineral processes?



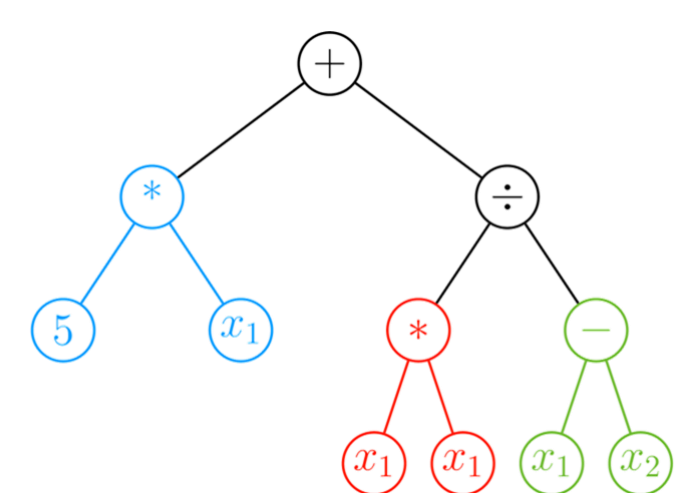
ALAMO sparse quadratic models balanced accuracy and efficiency in the best way compared to lasso, ANN, kriging, RBF techniques (Fardis et al., 2025).

Leaching process



MINLP for symbolic regression

$$5x_1 + \frac{(x_1)^2}{x_1 - x_2}$$



- Symbolic regression (SR) as disjunctive program (Cozad, 2014)
- STAR algorithm: randomized rounding to solve relaxed MINLP (Sarwar, 2022)

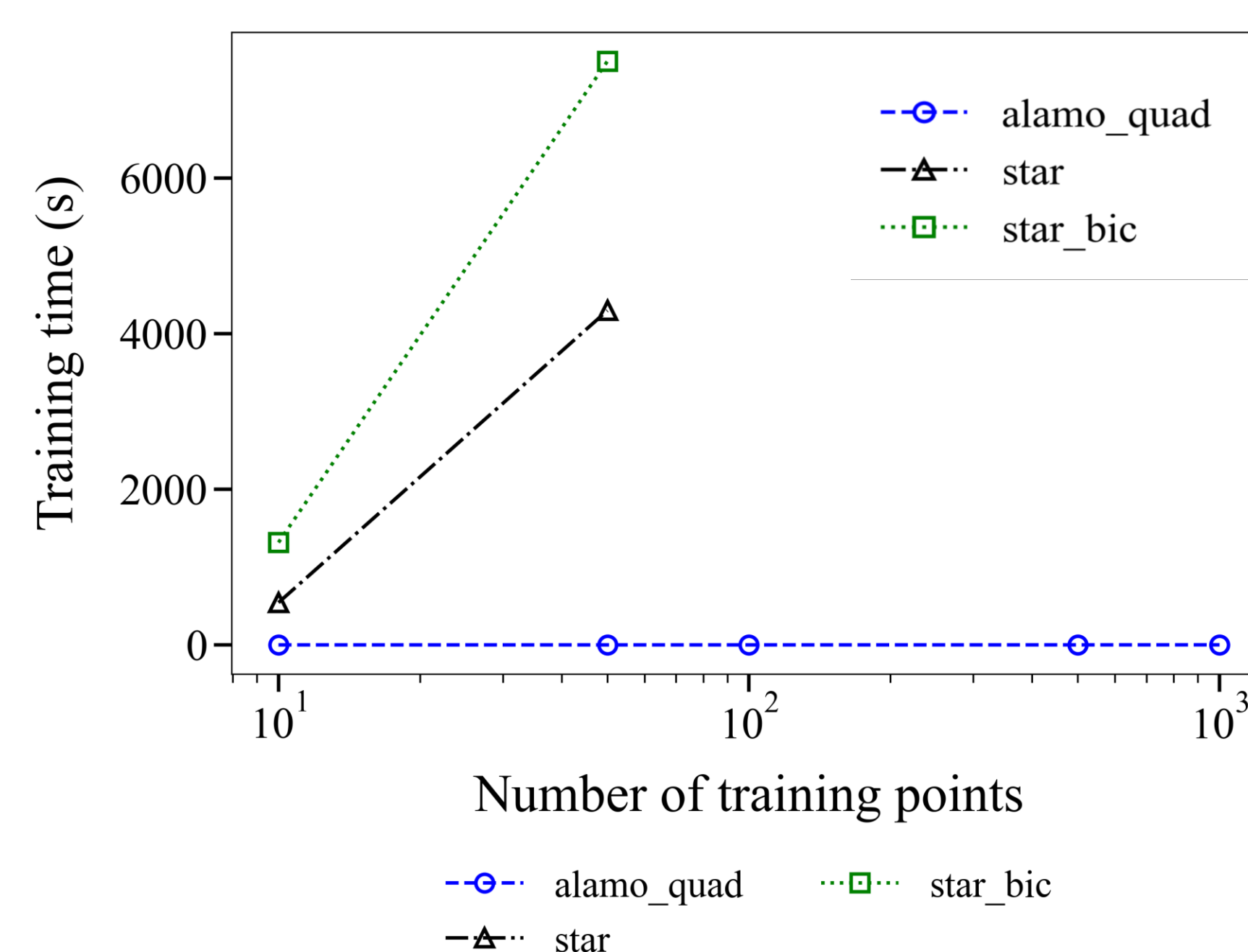
Regularized symbolic regression

- Complexity as the depth L of tree
- Minimize the Bayesian Information Criterion (BIC) in the MINLP:

$$BIC = n_{data} \ln \frac{SSR}{n_{data}} + L \ln n_{data}$$

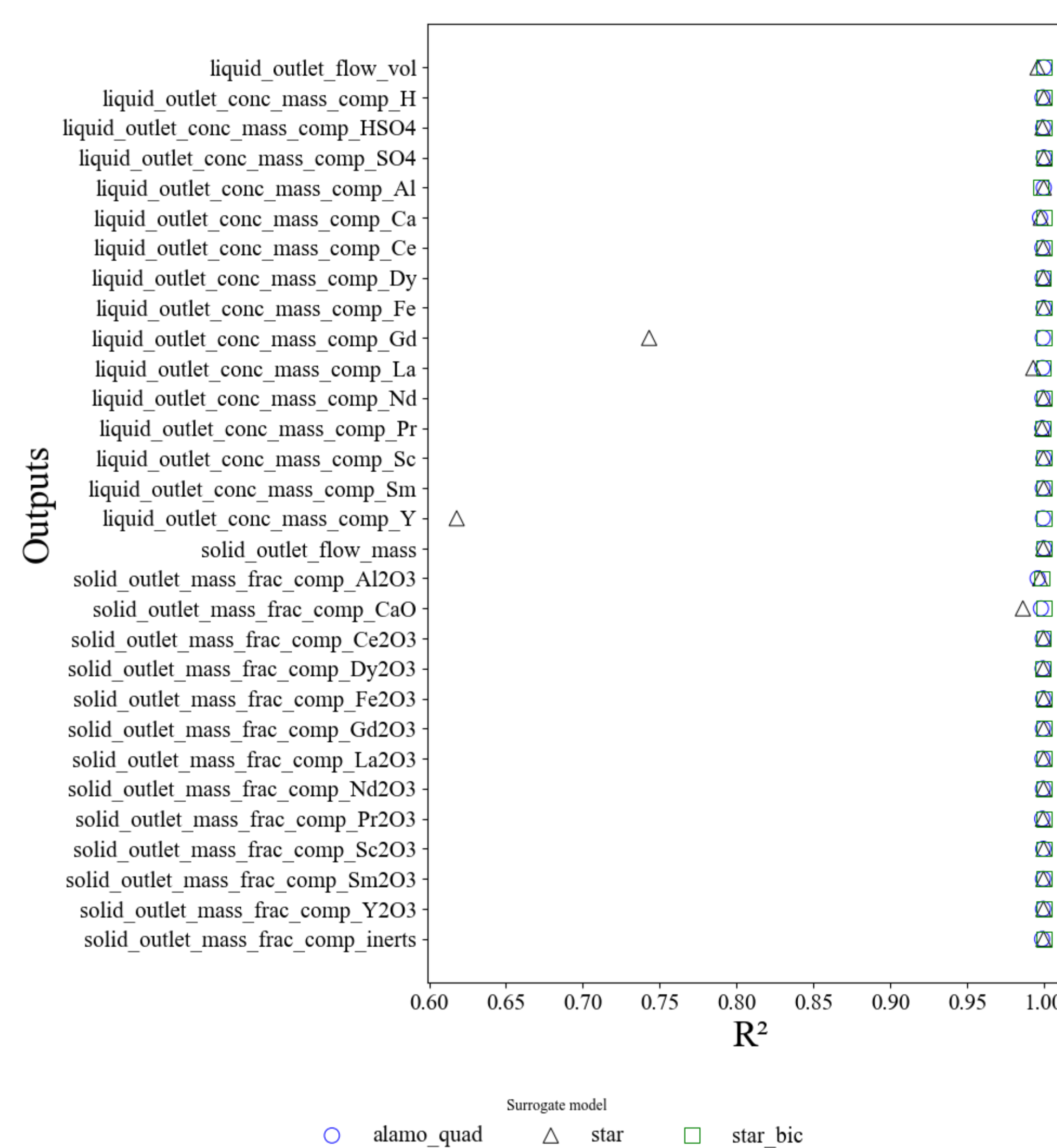
Evaluation of data-driven models

Training times



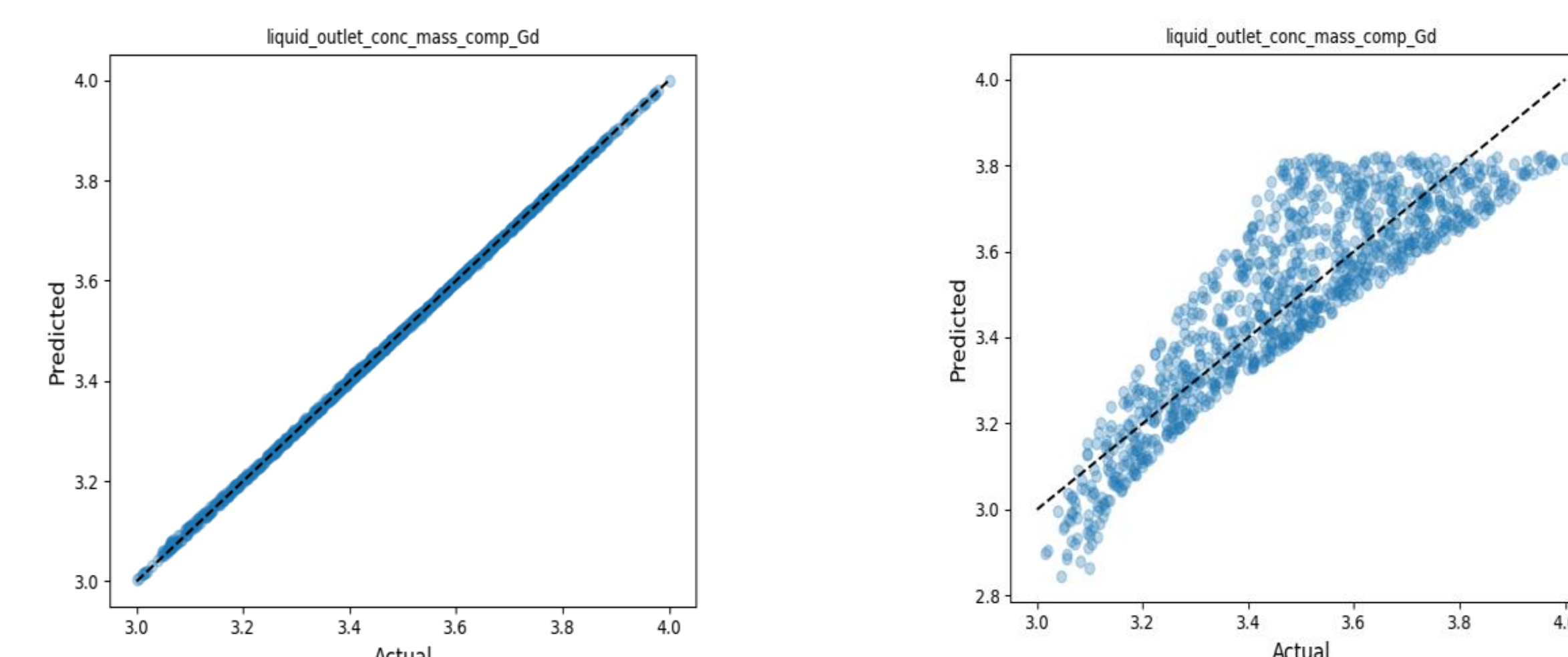
- Training times are 4-5 orders of magnitude larger for SR models.
- Regularization with BIC doubles the training time for SR.

R² values



- Regularized SR models fit the data really well, even when trained with 10 points.

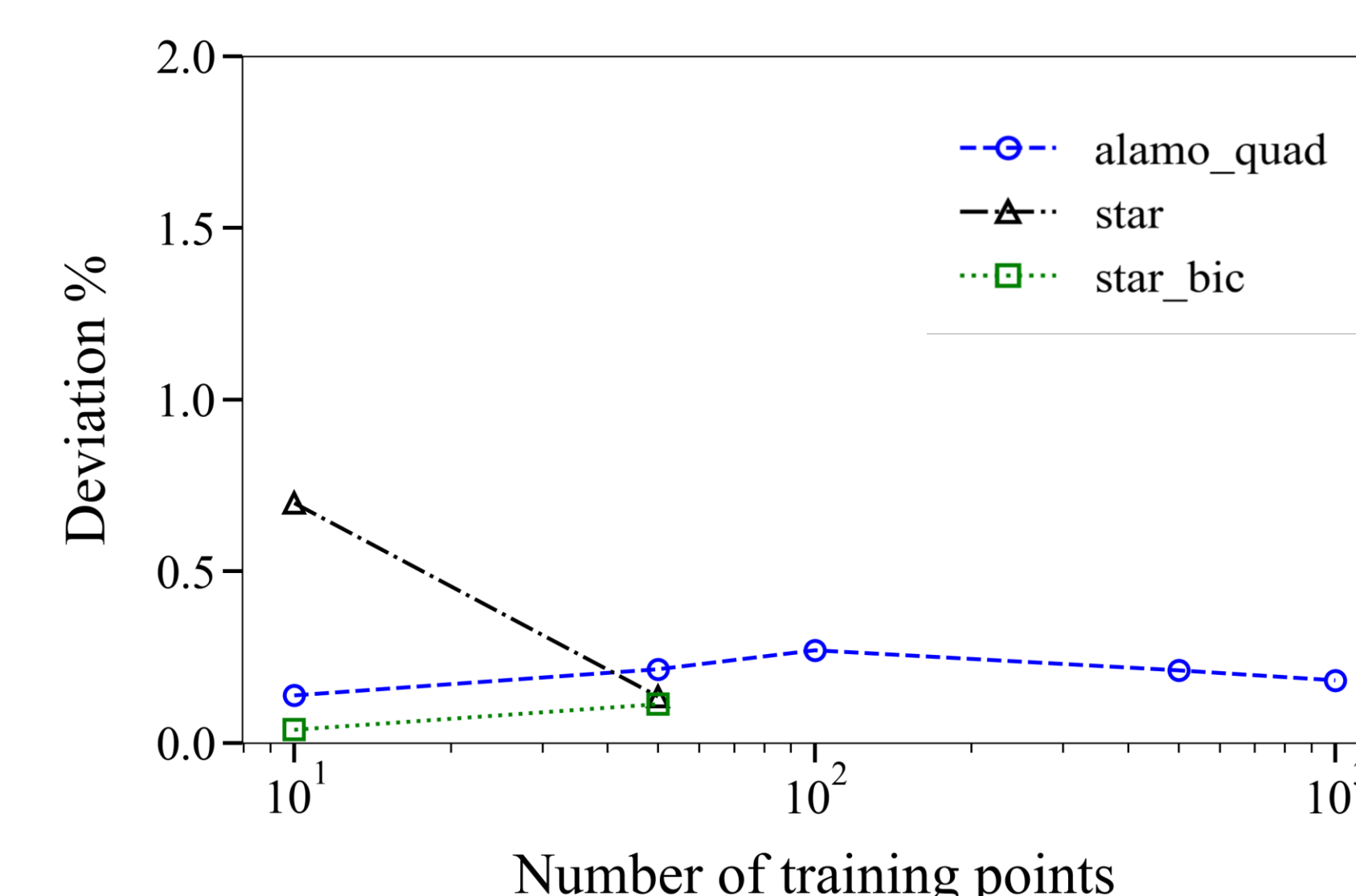
Regularized vs. Non-regularized STAR models



a) Regularized

b) Non-regularized

Optimization results



- Correct solution point with little deviation in the objective.
- Regularized SR models trained with 10 points provide the lowest deviation.

Conclusions

- SR finds accurate models without making assumptions about their form (useful for law discovery).
- SR is efficient for moderate numbers of training points.
- **Regularized SR prevents overfitting.**
- Future work will address various model selection criteria and complexity metrics (number of nodes, number of complexity-weighted operators, etc.).

Disclaimer: This presentation was prepared as an account of work sponsored by an agency of the United States Government, in part, through a site support contract. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Acknowledgements: This effort was funded by the U.S. Department of Energy's Process Optimization and Modeling for Minerals Sustainability (PrOMMIS) Initiative, supported by the Hydrocarbons and Geothermal Energy Office.

