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

Georgia Tech College of Engineering
School of Chemical and Biomolecular Engineering

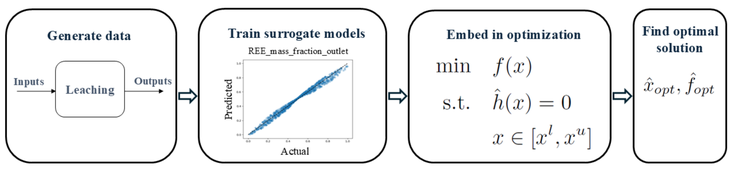
Dimitrios Fardis[1], Radhakrishna Tumbalam Gooty[2,3], Nick Sahinidis[1,4]

[1]School of Chemical & Biomolecular Engineering, Georgia Institute of Technology
[2]National Energy Technology Laboratory (NETL)
[3]NETL Support Contractor
[4]H. Milton School of Industrial and Systems Engineering, Georgia Institute of Technology

PROMMIS
Process Optimization and Modeling for Minerals Sustainability

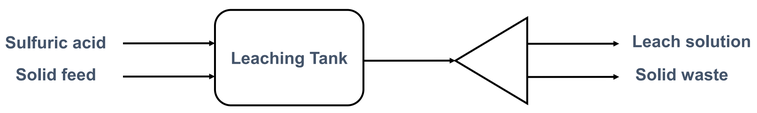
Data-driven modeling and optimization

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



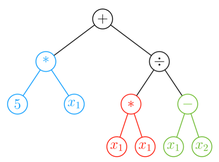
ALAMO sparse quadratic models balanced accuracy and efficiency in the best way compared to lasso, ANN, kriging, RBF techniques (Fardis et al., 2025).

Leaching process



MINLP for symbolic regression

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



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

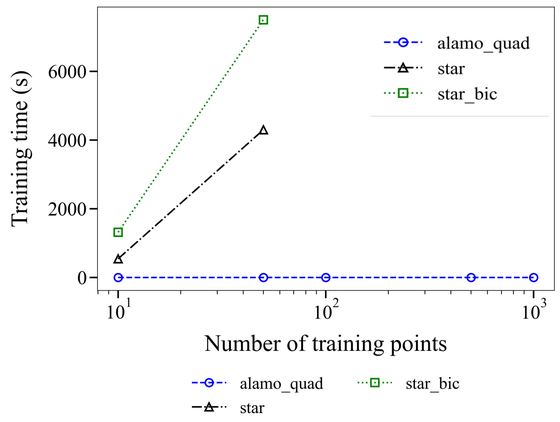
Regularized symbolic regression

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

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

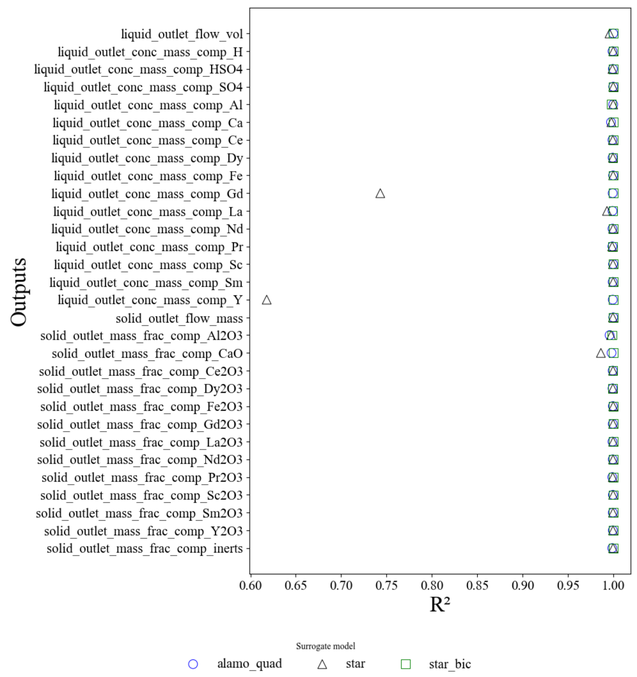
Evaluation of data-driven models

Training times



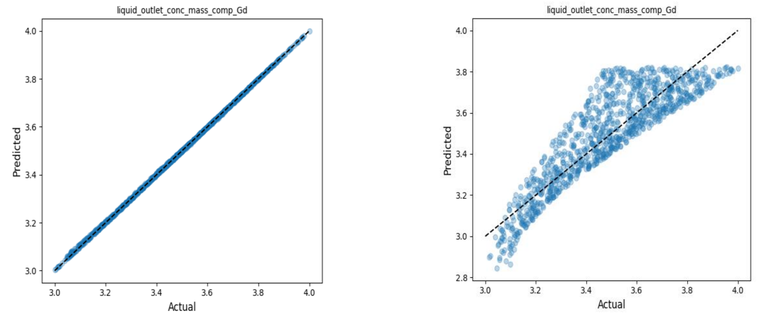
- **Training times are 4-5 orders of magnitude larger for SR models.**
- **Regularization with BIC doubles the training time for SR.**

R^2 values



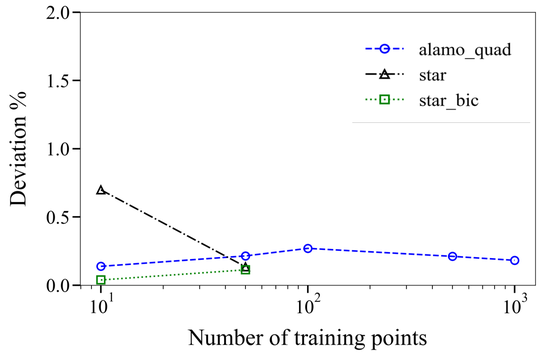
- **Regularized SR models fit the data really well, even when trained with 10 points.**

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.).**

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.

NATIONAL ENERGY TECHNOLOGY LABORATORY · BERKELEY LAB · Sandia National Laboratories · Carnegie Mellon · WestVirginiaUniversity · UNIVERSITY OF NOTRE DAME · Georgia Tech · U.S. DEPARTMENT OF ENERGY