U.S. DEPARTMENT of ENERGY

PROMMIS
Process Optimization and Modeling for Minerals Sustainability

MINLP for regularized symbolic regression

with applications to data-driven discovery of physical laws

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, Pittsburgh, PA 15236, USA

[3] NETL Support Contractor, Pittsburgh, PA 15236, USA

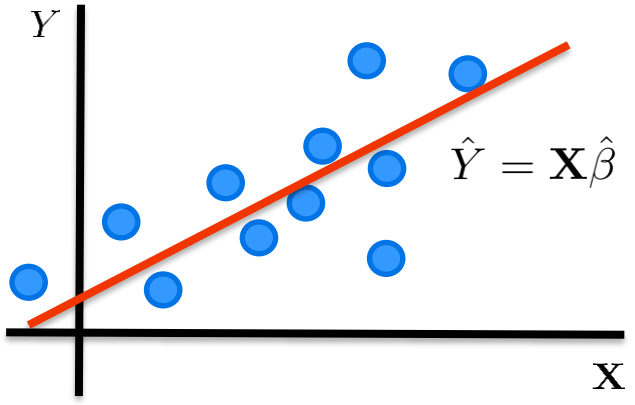
[4] H. Milton School of Industrial and Systems Engineering, Georgia Institute of Technology

NATIONAL ENERGY TECHNOLOGY LABORATORY · BERKELEY LAB · Sandia National Laboratories · Carnegie Mellon University · West Virginia University · University of Notre Dame · Georgia Tech

MODELS FROM DATA

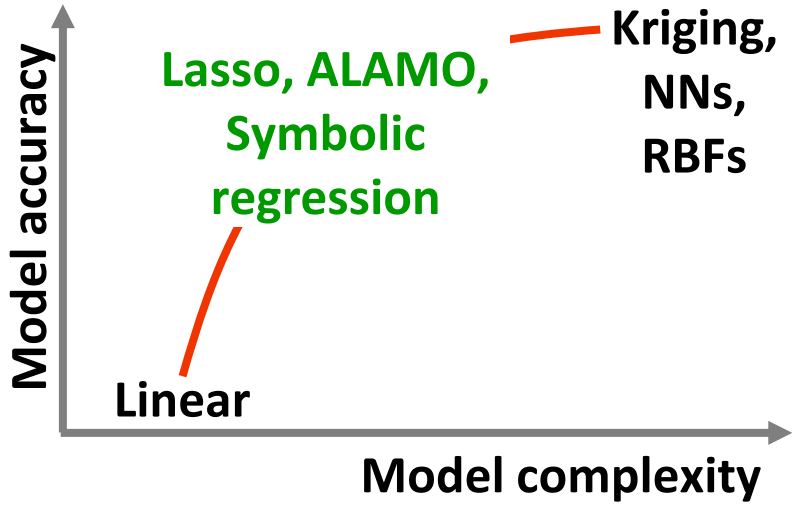
PROMMIS
Process Optimization and Modeling for Minerals Sustainability

- **(Ordinary) Least Squares method[1]**
- **Polynomial models: simple and interpretable**
  - Can be built quickly, necessary for real-time use
  - Not always accurate: too simple or overfitted
- **L1 regularization (Lasso[2]), best subset selection (e.g. ALAMO[3])**



$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- **Neural Networks[4,5], kriging[6,7], radial basis functions[8]**
- **Model complexity-accuracy tradeoff**



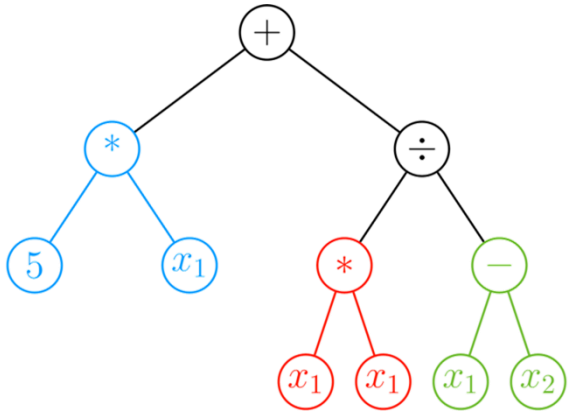
[1]Hastie et al., 2009; [2]Tibshirani, 1996; [3]Cozad et al., 2013; [4]McCulloch and Pitts, 1943; [5]Rosenblatt, 1958; [6]Krige, 1951; [7]Matheron, 1963; [8]Hardy, 1971

TREES AND SYMBOLIC REGRESSION

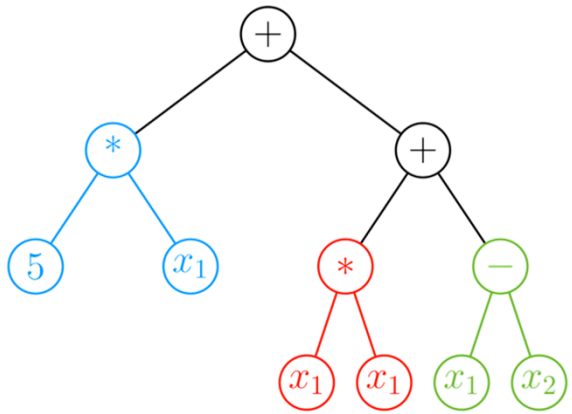
PROMMIS
Process Optimization and Modeling for Minerals Sustainability

- **Any closed-form expression can be represented as a binary tree**
  - Leaves: variables and constants
  - Non-terminal nodes: operators

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



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

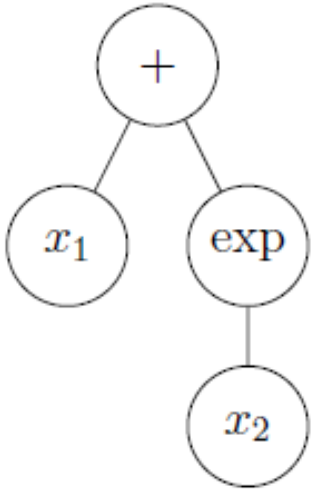


SYMBOLIC REGRESSION

PROMMIS
Process Optimization and Modeling for Minerals Sustainability

- **Supervised machine learning (ML) technique that determines simultaneously**
  - **The functional form of a model**
  - **The values of the regression parameters that best fit the data**
- **Symbolic regression (SR) models are parameterized using a set of operators**
  - **Usual operators are** $+, -, *, \div, \mathrm{sqrt}(\cdot), \exp(\cdot), \log(\cdot), \sin(\cdot), (\cdot)^2$
- **Data-driven modeling with SR**

$$\hat{y} = x_1 + e^{(x_2)}$$

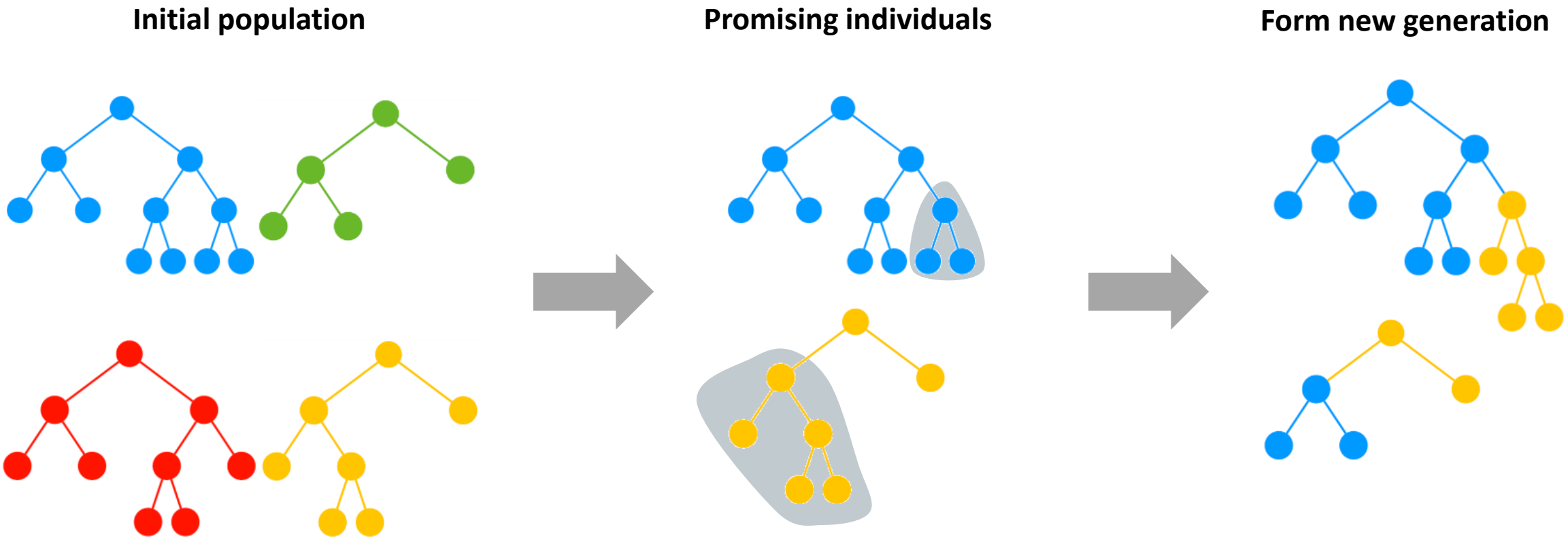


- **Symbolic regression is an emerging field on its own[1,2,3,4]**

Kronberger et al., 2024; Angelis et al., 2023; Makke and Chawla, 2024 ; Dong and Zhong, 2025

GENETIC PROGRAMMING

PROMMIS
Process Optimization and Modeling for Minerals Sustainability



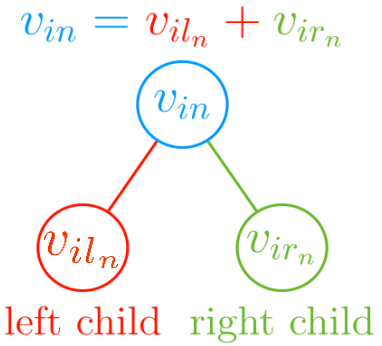
GP (Koza, 1992); Eureqa (Lipson et al., 2009); SRBench (Cava et al., 2021); gplearn (Stephens, 2015); Operon (Burlacu et al., 2020); AIFeynman (Udrescu et al., 2020); PySR (Cranmer, 2023); GP inefficient for SR (Kronberger et al., 2024)

MINLP FOR SYMBOLIC REGRESSION

PROMMIS
Process Optimization and Modeling for Minerals Sustainability

- **SR as a Generalized Disjunctive Program[1,2]**

$$
\begin{aligned}
\min \quad & [\text{error}] \\
\text{s.t.} \quad & \{+\} \vee \{-\} \vee \{*\} \vee \{\div\} \vee \{\text{cst}\} \vee \{x_d\} \vee \ldots \quad \forall \text{ nodes} \\
& [\text{logical constraints}] \\
& v_{in} \in [v_{in}^{\text{lo}}, v_{in}^{\text{up}}] \qquad \forall \text{ nodes } n \text{ and points } i
\end{aligned}
$$



- **Binary variables represent the choice of operators/operands at each node**

$$z_n^+,\ z_n^{\log},\ z_n^{\exp},\ z_n^*,\ z_n^{\text{cst}},\ z_n^{x_1}$$

- **Value-defining constraints formulate operations using big-M constraints**
  - **Defining operation at each non-terminal node, data point, and operator**
  - **Example: Addition operator**

$$
\begin{aligned}
v_{il_n} + v_{ir_n} - v_{in} &\leq \overline{M}_{in}^{+}\,(1 - z_n^{+}) \\
v_{il_n} + v_{ir_n} - v_{in} &\geq \underline{M}_{in}^{+}\,(1 - z_n^{+})
\end{aligned}
$$

[1]Cozad, 2014; [2]Cozad and Sahinidis, 2018; Austel et al., 2017, 2020; Kim et al., 2023

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

CONSTRAINTS AND OBJECTIVE

PROMMIS
Process Optimization and Modeling for Minerals Sustainability

- **Tree-defining constraints formulate logical conditions about the structure of tree**
  - Example: Only one operator/operand can be assigned to a node

$$\sum_{o \in O} z_n^o \leq 1$$

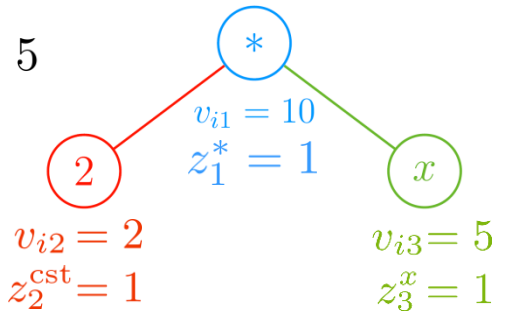
- **Redundancy-eliminating constraints eliminate similar trees**
  - Example: Do not make operations between constants

$$z_{2n}^{cst} + z_{2n+1}^{cst} \leq 1$$

- **The loss function is the SSR in existing MINLP approaches**

$$\text{Loss} = \sum_{i=1}^{n_{data}} (y_i - \hat{y}_i)^2 \quad where \quad \hat{y}_i = v_{i1}$$

Function: $y(x) = 2\,x$
$y = 10$ at point i, $x_i = 5$



RANDOMIZED ROUNDING ALGORITHM

PROMMIS
Process Optimization and Modeling for Minerals Sustainability

- **STAR (Symbolic regression Through Algebraic Representations) algorithm[1]**
  - Solve a continuous relaxation of the MINLP
  - Obtain probabilities of assigning operands and operators to trees' nodes
  - Use randomized rounding to generate a number of expression trees
  - Enhance trees: multiply each input variable with a constant slope and add an intercept
  - Optimize parameters of each tree
  - Evaluate trees

- **PySTAR[2,3] has superior overall predictive performance compared to the state-of-the-art GP methods gplearn, Operon, and PySR**

[1]Sarwar, 2022; [2]Kim et. al, 2025; [3]https://github.com/IDAES/idaes-pySTAR

REGULARIZED SYMBOLIC REGRESSION

PROMMIS
Process Optimization and Modeling for Minerals Sustainability

- **Model selection criteria that penalize complexity**
  - Complexity as the depth L of the expression tree

- **Bayesian Information Criterion (BIC):** $$\mathrm{BIC} = n_{data} \ln \frac{\mathrm{SSR}}{n_{data}} + L \ln n_{data}$$

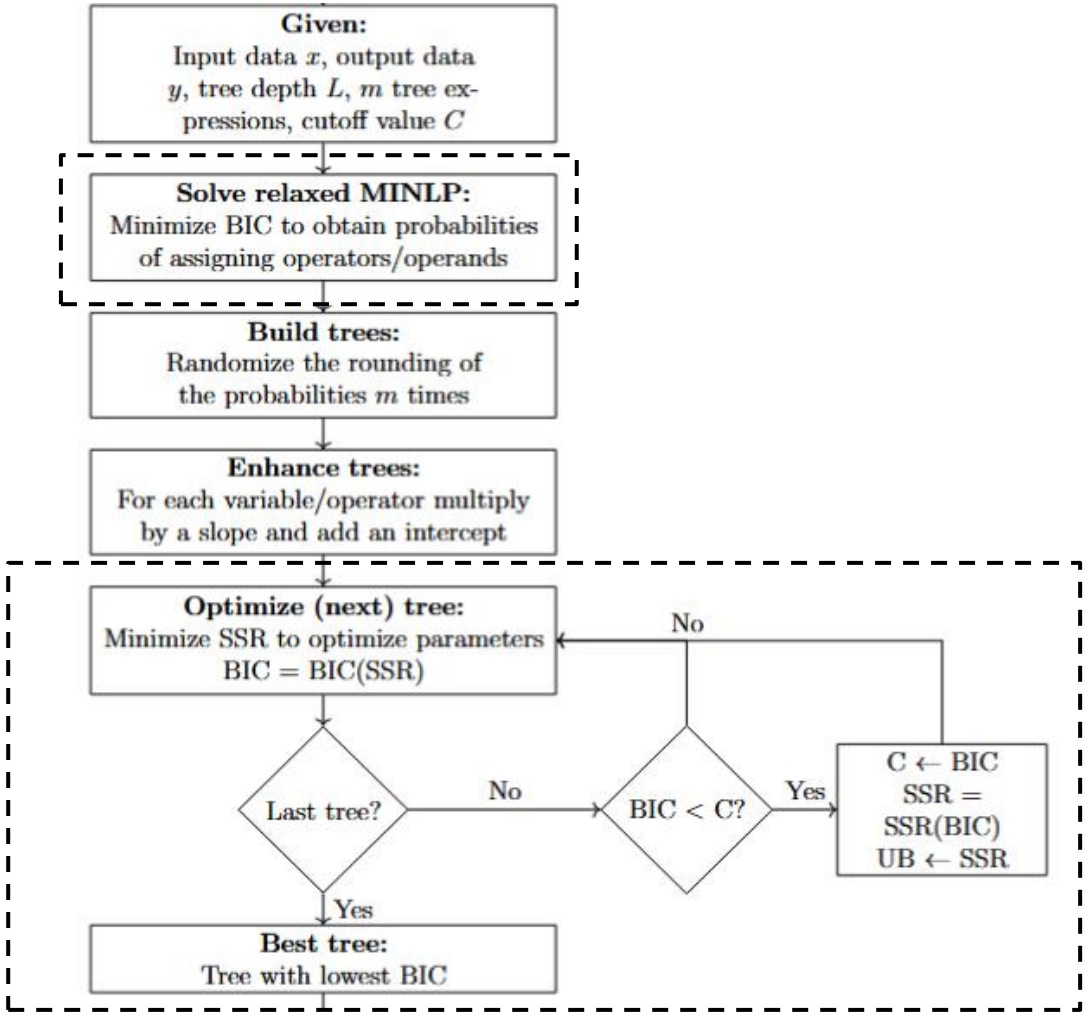
- **The level l_n for node n in the tree is:** $$l_n = \lceil \log_2(n+1) \rceil - 1$$

- **Let z_{no} is the binary variable of assigning the operator or operand $o \in O$ to node $n \in N$. The depth of the tree L is:**

$$L = \max_{n \in \mathcal{N}} l_n \sum_{o \in \mathcal{O}} z_{no}$$

PENALIZE COMPLEXITY USING THE BIC

PROMMIS
Process Optimization and Modeling for Minerals Sustainability



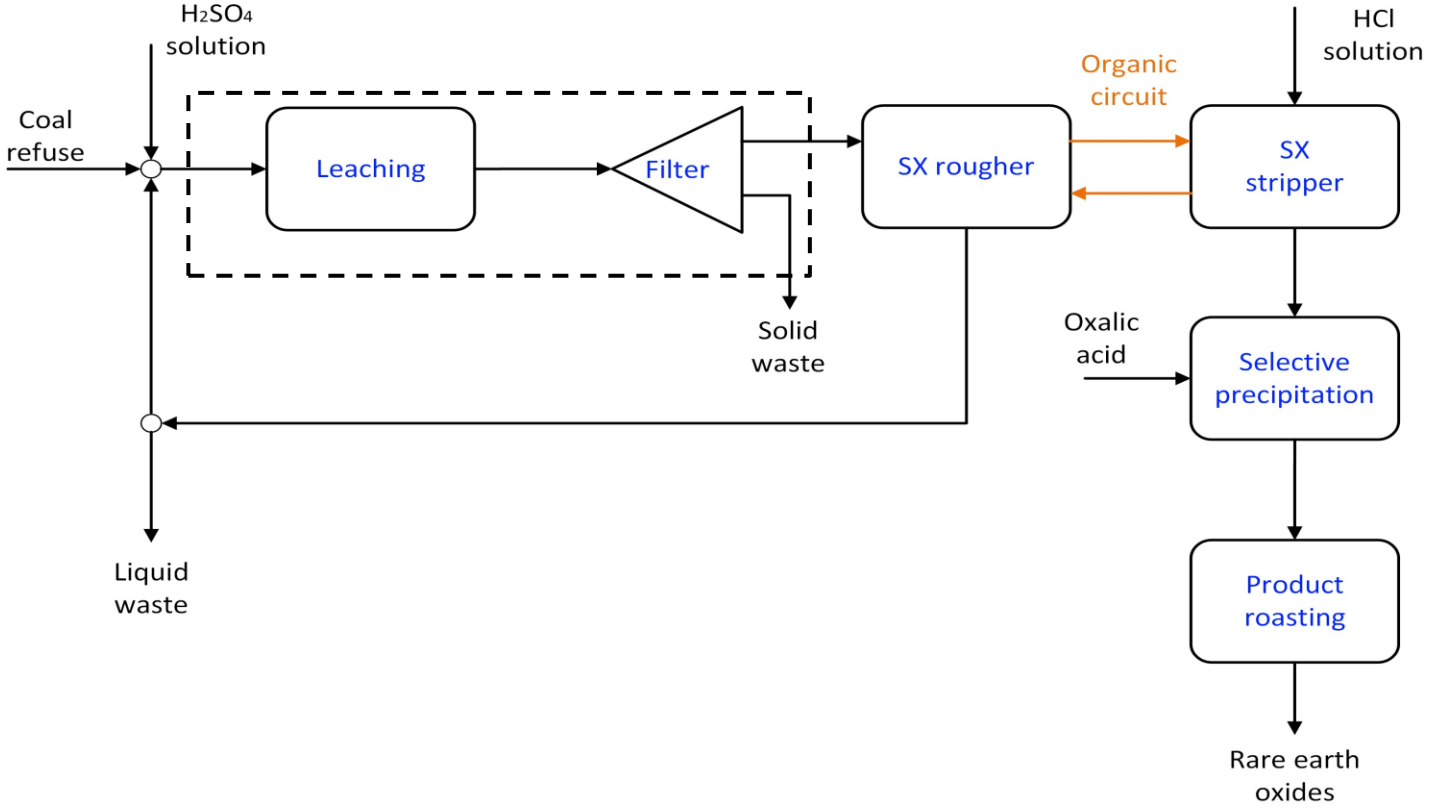
$$\mathrm{BIC} = n_{data} \ln \frac{\mathrm{SSR}}{n_{data}} + k \ln n_{data}$$

$$\mathrm{SSR} = n_{data} \exp\left(\frac{\mathrm{BIC} - k \ln n_{data}}{n}\right)$$

SR FOR CRITICAL MINERALS RECOVERY

PROMMIS
Process Optimization and Modeling for Minerals Sustainability

- **Modeling, simulation and optimization of critical mineral processes**
  - Goal: simplify modeling and optimization of leaching using SR models

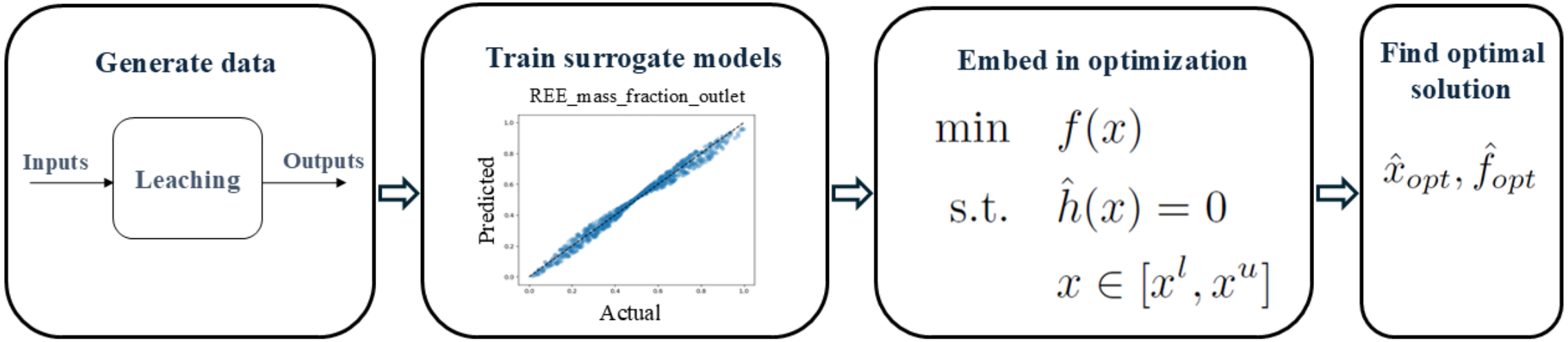


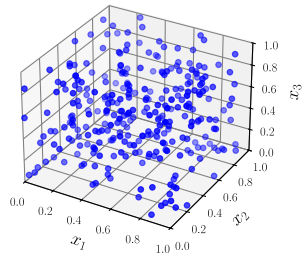
- **Input variables of leaching**
  - Solid feed ranging in [50, 100] lb/hr
  - Concentration of sulfuric acid ranging in [0.025, 0.075] mol/L
- **30 outputs in the liquid and solid outlets**

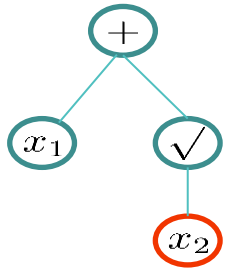
MODELING AND OPTIMIZATION

PROMMIS
Process Optimization and Modeling for Minerals Sustainability

- **Build STAR models for a chemical process**
  - Leaching process of a flowsheet for the recovery of critical minerals[1]



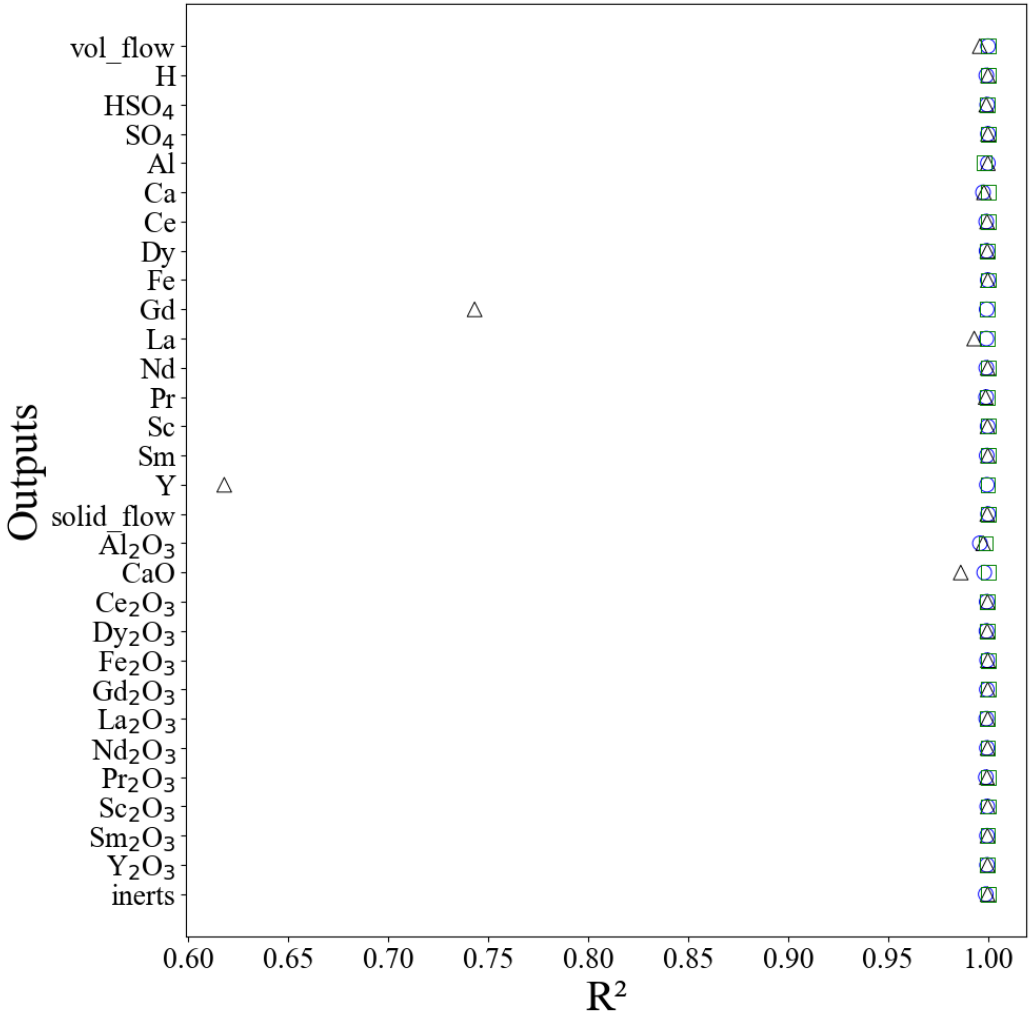


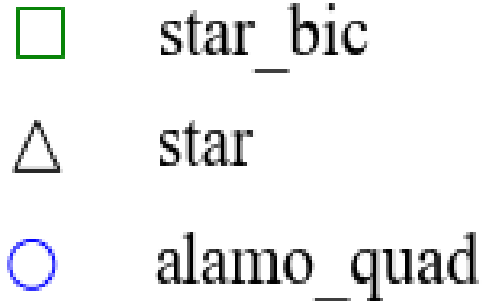


[1]Fardis et al., 2025

EVALUATION OF DATA-DRIVEN MODELS

PROMMIS
Process Optimization and Modeling for Minerals Sustainability



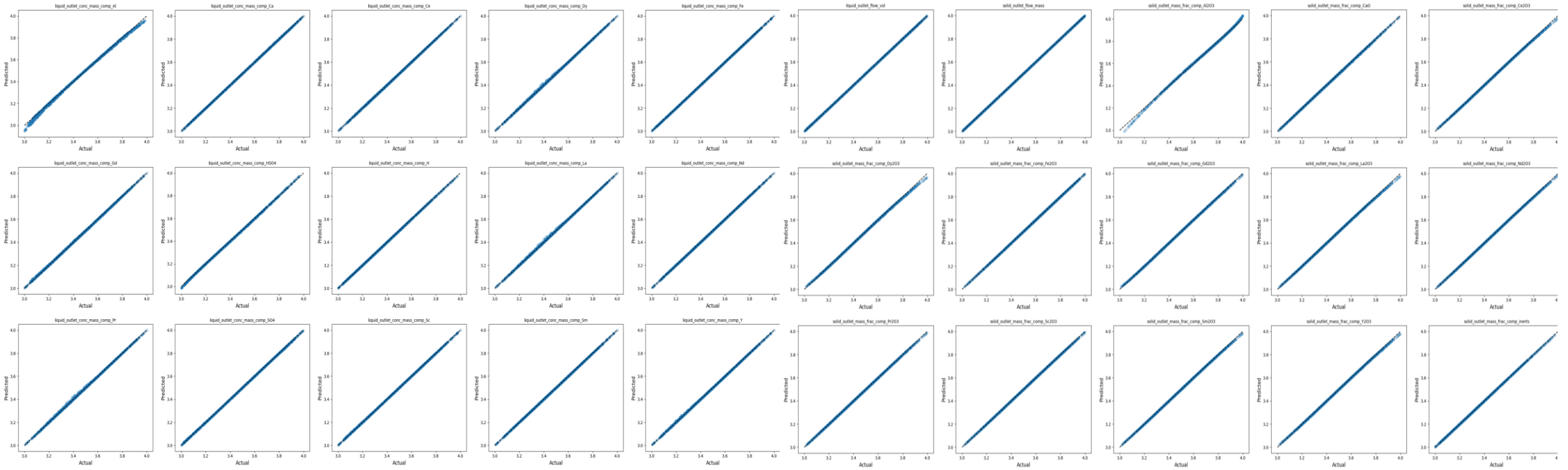


- **ALAMO sparse quadratics balanced accuracy and efficiency in the best way compared to lasso, ANN, kriging, RBF[1]**
- **Regularized SR models are accurate, even when trained with 10 points**

[1]Fardis et al., 2025

REGULARIZED SR MODELS

PROMMIS
Process Optimization and Modeling for Minerals Sustainability



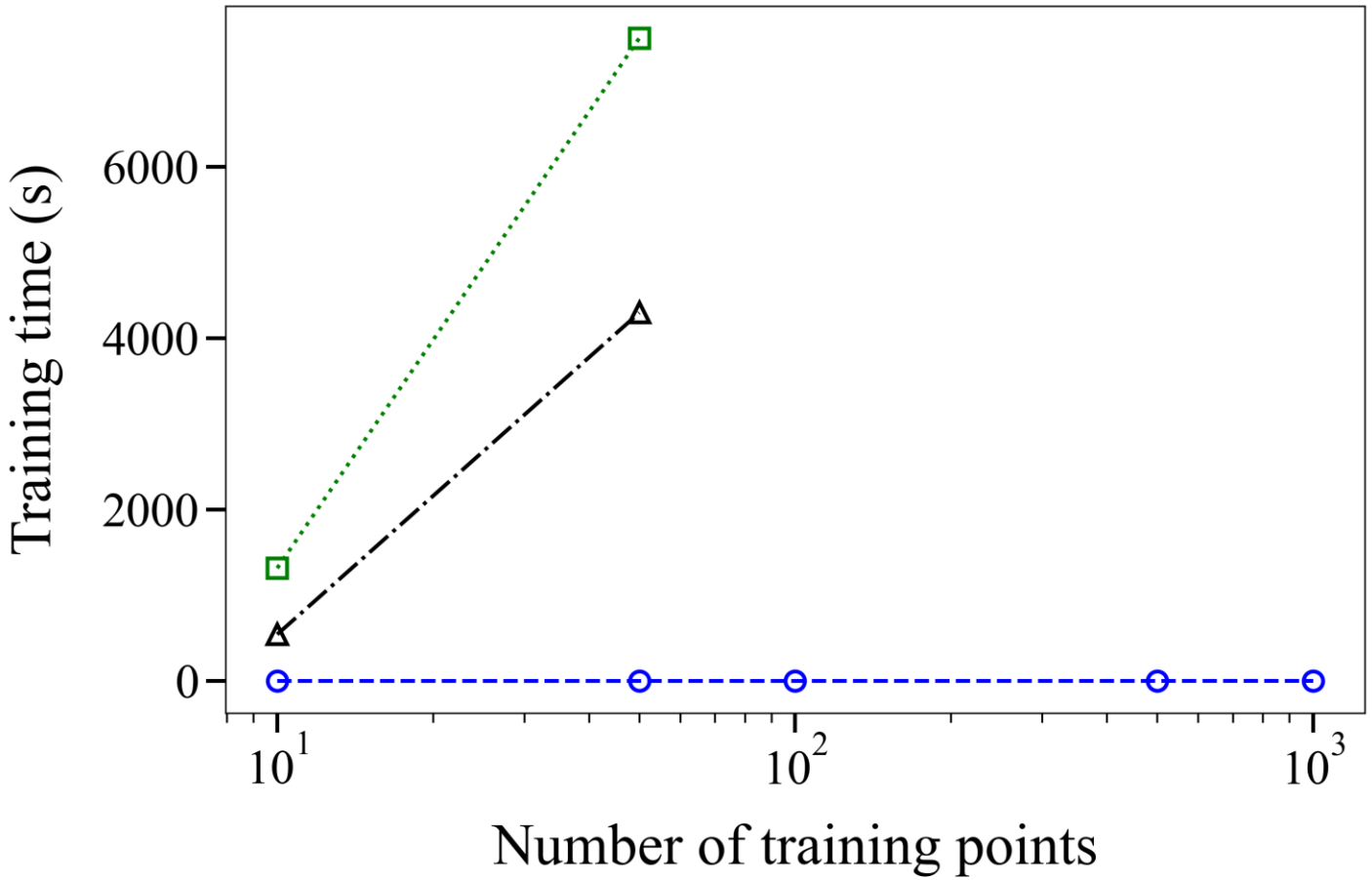
Regularized SR models trained with 10 points achieve

- **Excellent predictive performance**
- **Low model complexity**

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

TRAINING TIMES

PROMMIS
Process Optimization and Modeling for Minerals Sustainability

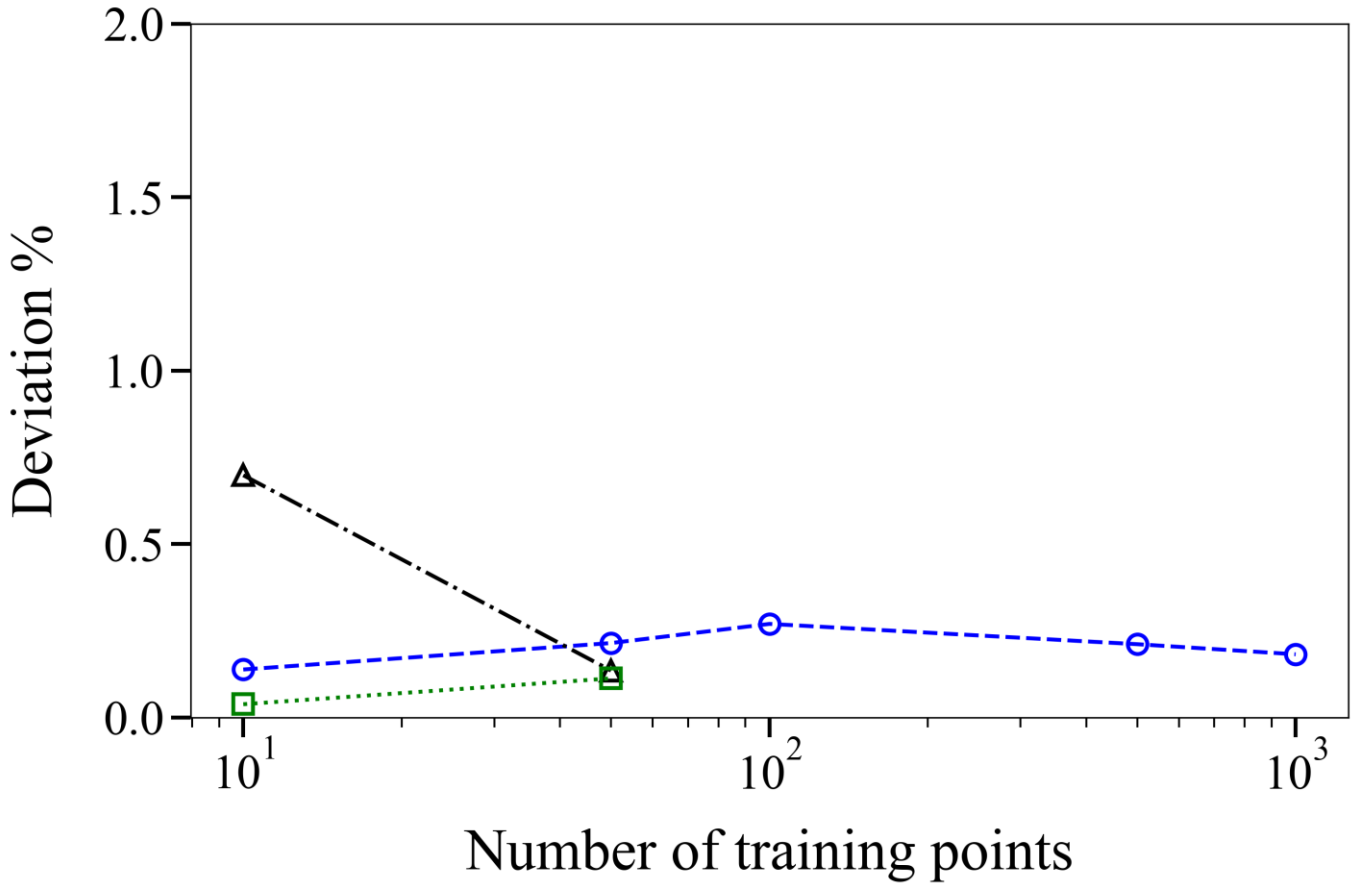


alamo_quad
star
star_bic

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

SURROGATE-BASED OPTIMIZATION

PROMMIS
Process Optimization and Modeling for Minerals Sustainability



- **Correct solution point**
  - 45.36 kg/h solid feed
  - 0.075 M sulfuric acid
- **Little deviation in the objective across methods**
  - EO optimal objective value is 5128.2 mg rare earths per hour
- **Regularized SR models trained with 10 points provide the lowest deviation**

CONCLUSIONS AND FUTURE WORK

PROMMIS
Process Optimization and Modeling for Minerals Sustainability

- **SR is attractive**
  - It provides simple data-driven models
  - It makes no prior assumptions about their functional forms (useful for law discovery)
- **SR efficient for moderate numbers of training points**
- **Regularization can expand the applicability of SR**
  - No overfitting
  - Simple and interpretable models
- **Compare various complexity metrics**
  - Depth, number of nodes, number of operators, number of complexity-weighted nodes
- **Compare various model selection criteria**
  - BIC, AIC, HQIC, etc.

Acknowledgements:

PROMMIS
Process Optimization and Modeling for Minerals Sustainability

This effort was funded by the U.S. Department of Energy's Process Optimization and Modeling for Minerals Sustainability (PrOMMiS) Initiative, supported by the Office of Fossil Energy and Carbon Management's Office of Resource Sustainability.

For questions and comments, please contact our Technical Director, Thomas Tarka (Thomas.Tarka@netl.doe.gov).

Office of Fossil Energy and Carbon Management: Grant Bromhal, Burt Thomas, Gabby Ubay, Morgan Summers.

National Energy Technology Laboratory: Thomas Tarka, Tony Burgard, Miguel Zamarripa, Alison Fritz, Alejandro Garciadiego, Brandon Paul, Anca Ostace, Radhakrishna Gooty, Jinliang Ma, Lingyan Deng, Marcus Holly, Elmira Shamlou, Daison Caballero, Ryan Hughes, Adam Atia.

Sandia National Laboratories: John Siirola, Bethany Nicholson, Michael Bynum, Edna Soraya Rawlings, Emma Johnson.

Lawrence Berkeley National Laboratory: Dan Gunter, Keith Beattie, John Shinn, Oluwamayowa Amusat, Sarah Poon, Ludovico Bianchi.

Carnegie Mellon University: Larry Biegler, Ignacio Grossmann, Carl Laird, Chrysanthos Gounaris, Ana I. Torres, Jason Yao, Christopher Laliwala, Norman Tran.

West Virginia University: Debangsu Bhattacharyya, Akintomiwa Ojo, Arkoprabho Dasgupta.

University of Notre Dame: Alex Dowling, Molly Dougher, Shammah Lilonfe, Dan Laky.

Georgia Tech: Nick Sahinidis, Dimitrios Fardis.

2025 Joint IDAES/CCSI$_2$/PrOMMiS/WaterTAP Technical Team Meeting University of Notre Dame

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