SymReg

JOSEF RESSEL CENTER FOR
SYMBOLIC REGRESSION

# Comparing Shape-Constrained Regression Algorithms for Data Validation

**Florian Bachinger**

**University of Applied Sciences Upper Austria**
**Johannes Kepler University, Linz**

UNIVERSITY
OF APPLIED SCIENCES
UPPER AUSTRIA

# Motivation

MIBA Frictec GmbH – R&D Department

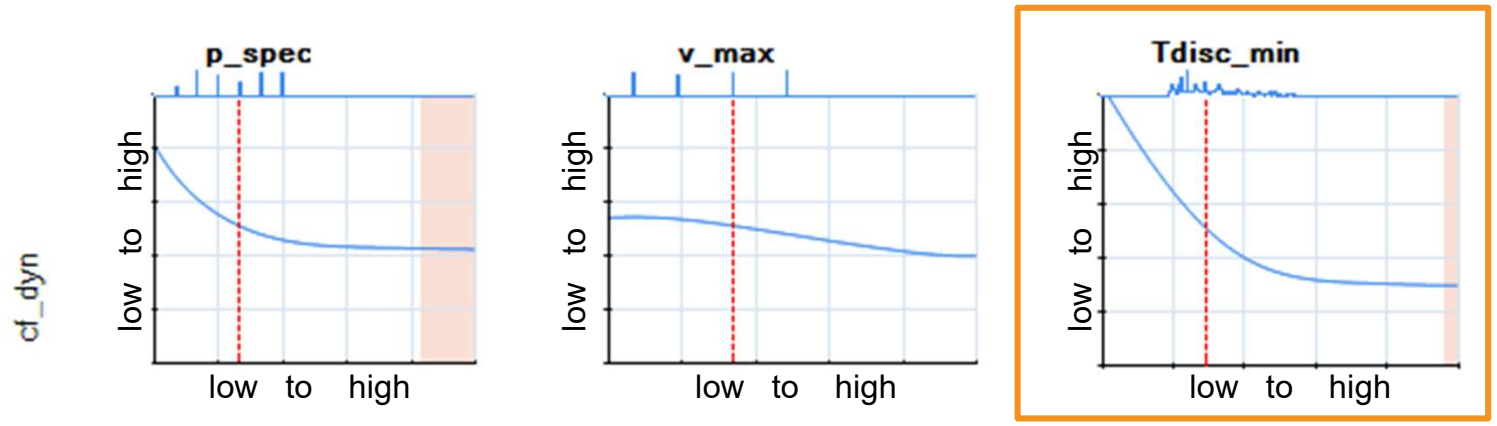
Requires automated data validation:

- Large volumes of data
- Unknown data quality
- Costly to validate by experts
- Existing rule-based approaches not sufficient
  (complex interactions)

SymReg

# Prior Knowledge – Shared Behavior

- Characteristics of new material are **unknown**
- Shared behavior **is known**

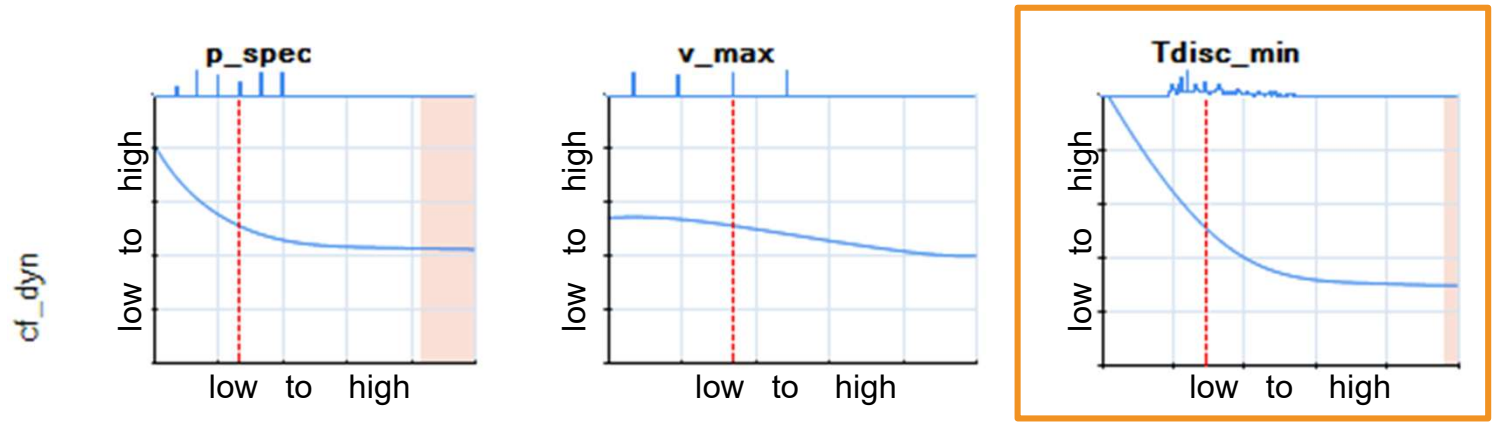



↑T → ↓μ $_{dyn}$ ➡ monotonic decreasing

SymReg

# Prior Knowledge – Shared Behavior

- Characteristics of new material are **unknown**
- Shared behavior **is known**

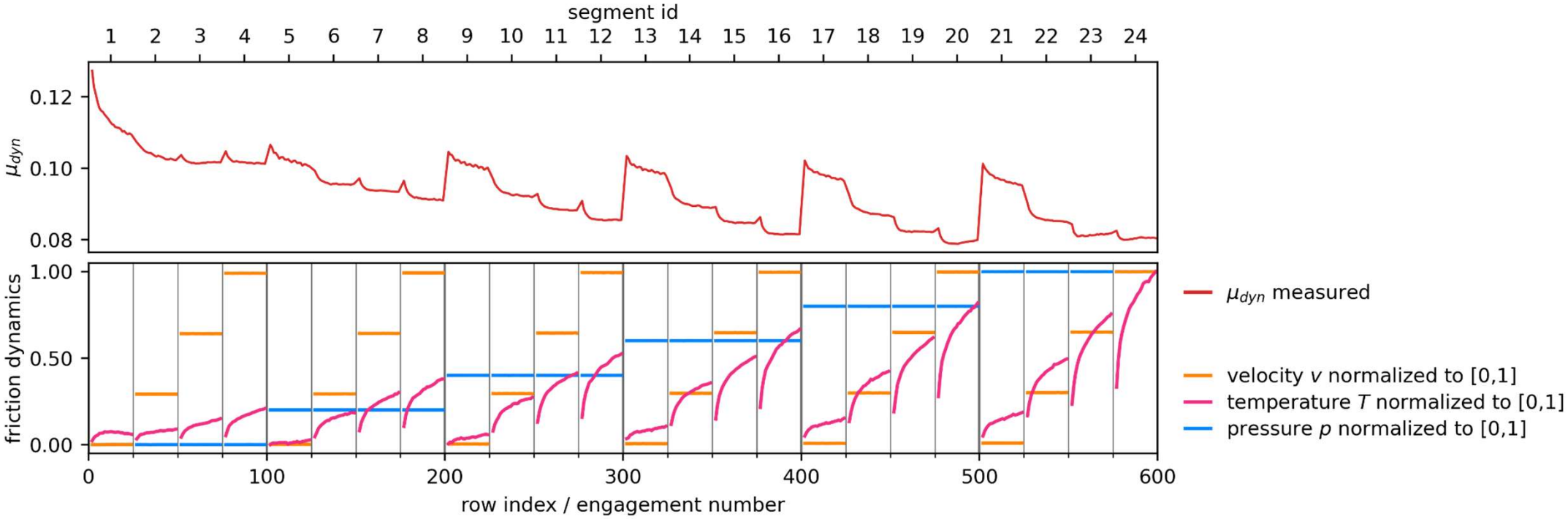



↑T → ↓$\mu_{dyn}$ ➡ monotonic decreasing

More precisely ➡ $\frac{\partial \mu_{dyn}}{\partial T} \leq 0 \wedge \frac{\partial^2 \mu_{dyn}}{\partial T^2} \geq 0$

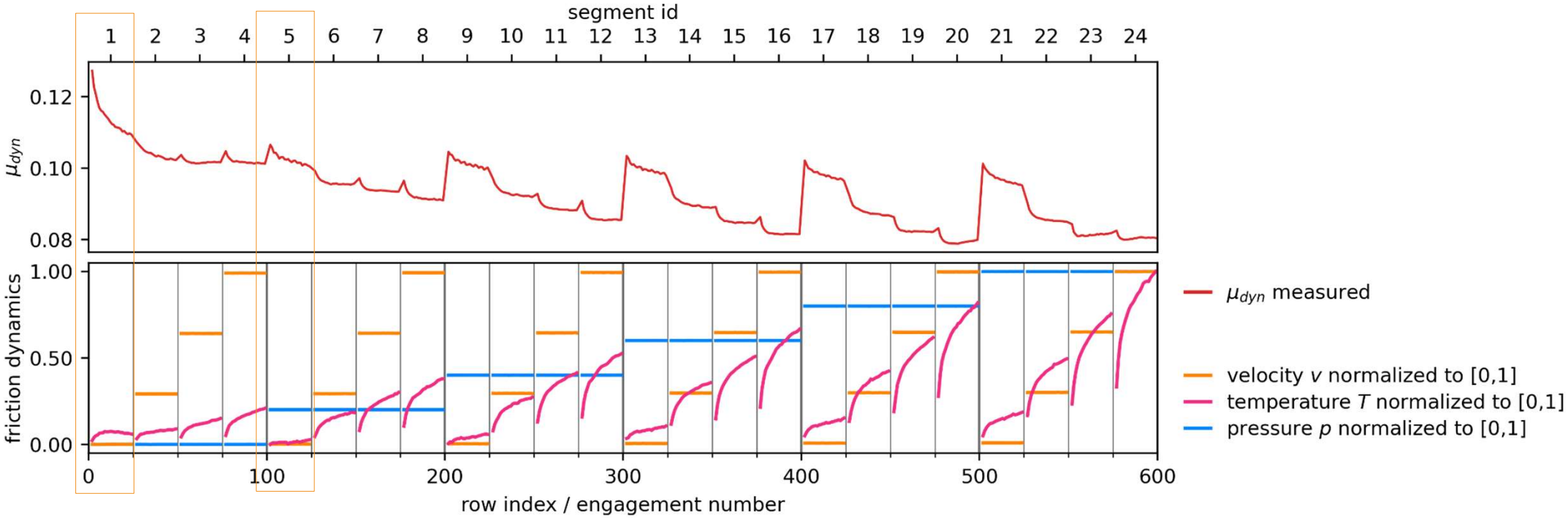
SymReg

# A Detailed Example




SymReg

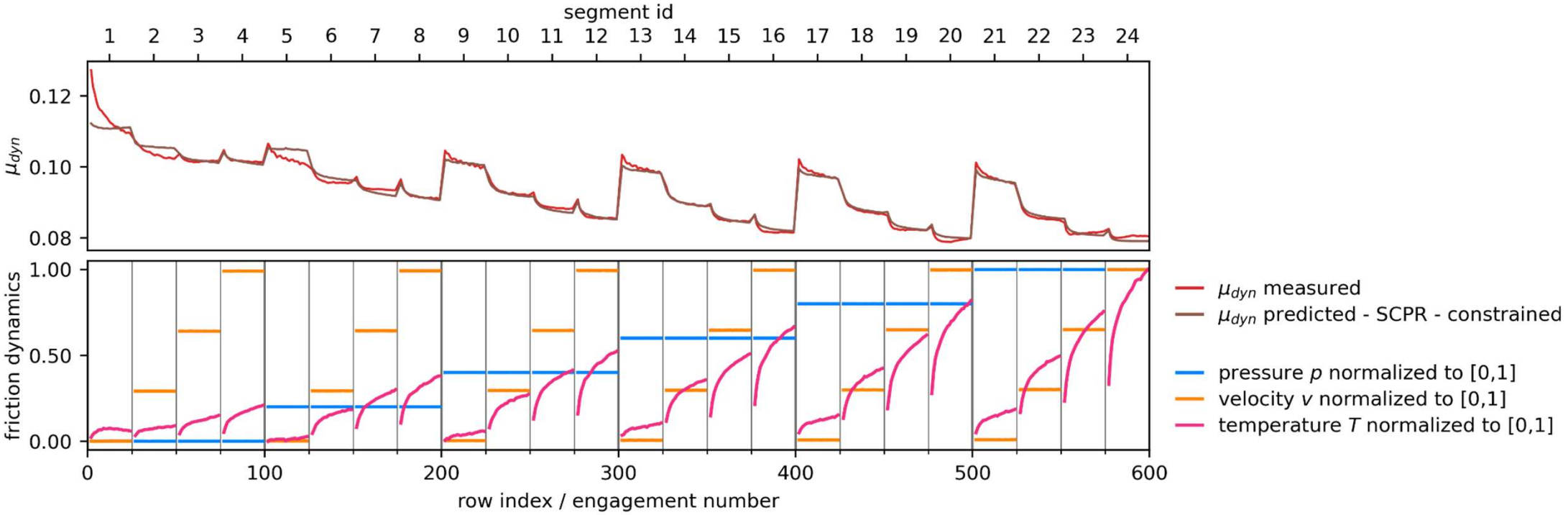
# A Detailed Example




Segments do not comform to expected behavior

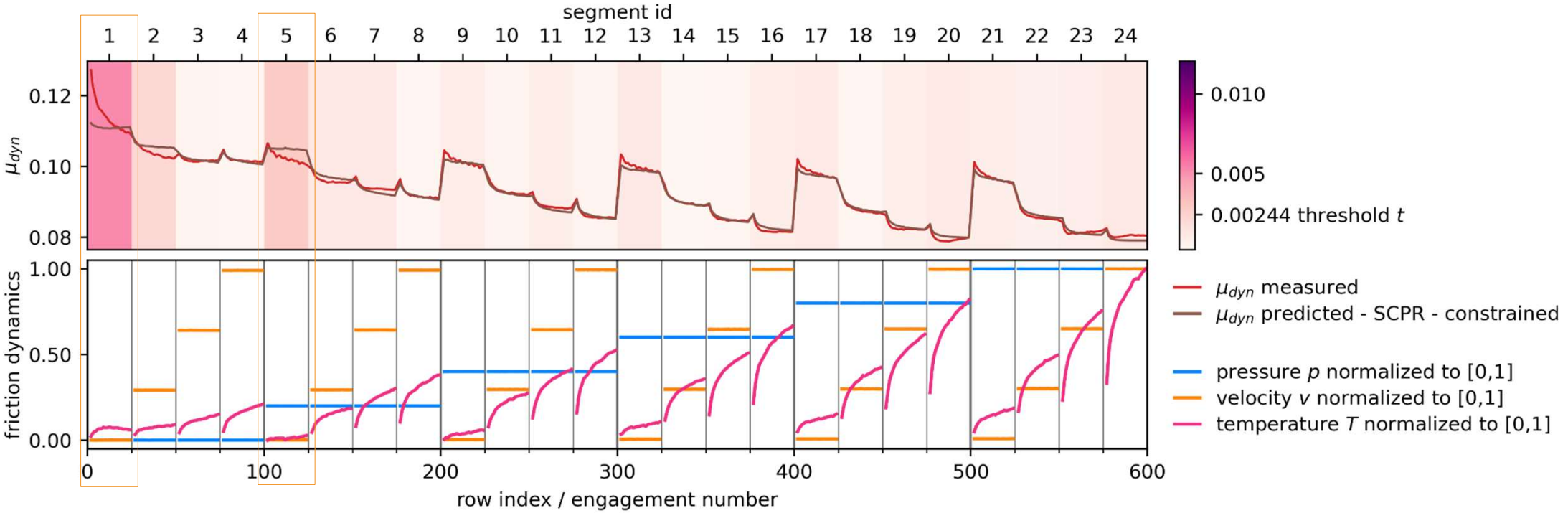
SymReg

# A Detailed Example




SymReg

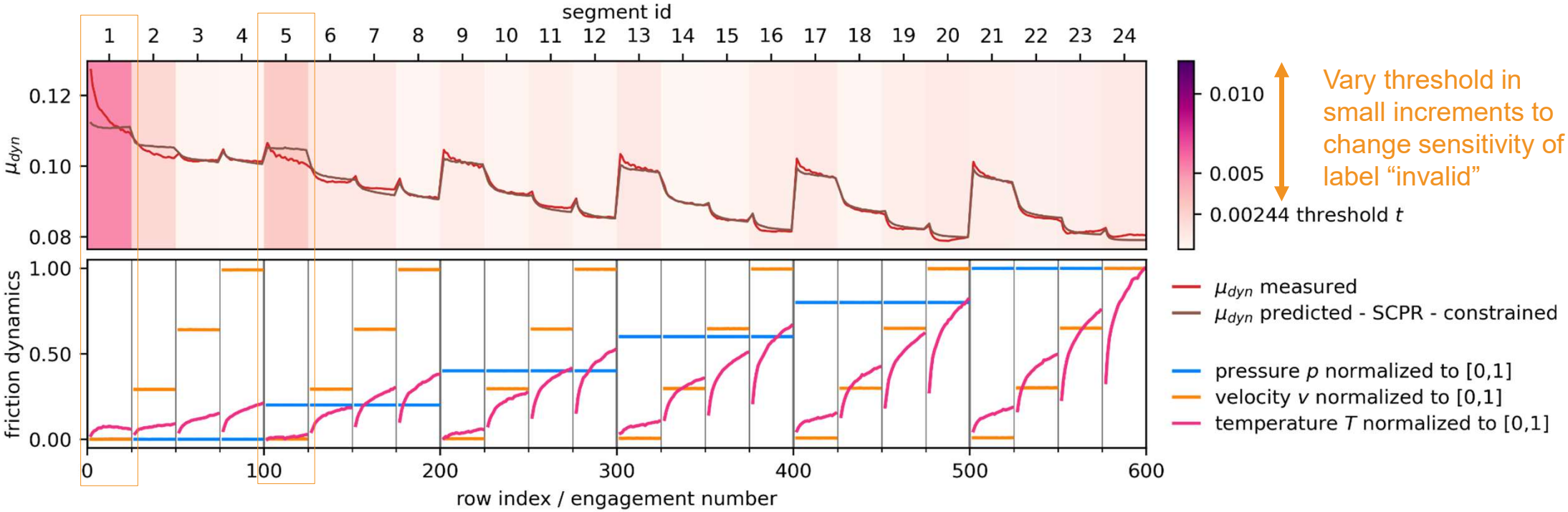
# A Detailed Example




Segments do not comform to expected behavior

SymReg

# A Detailed Example

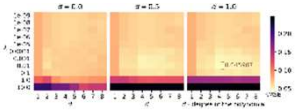


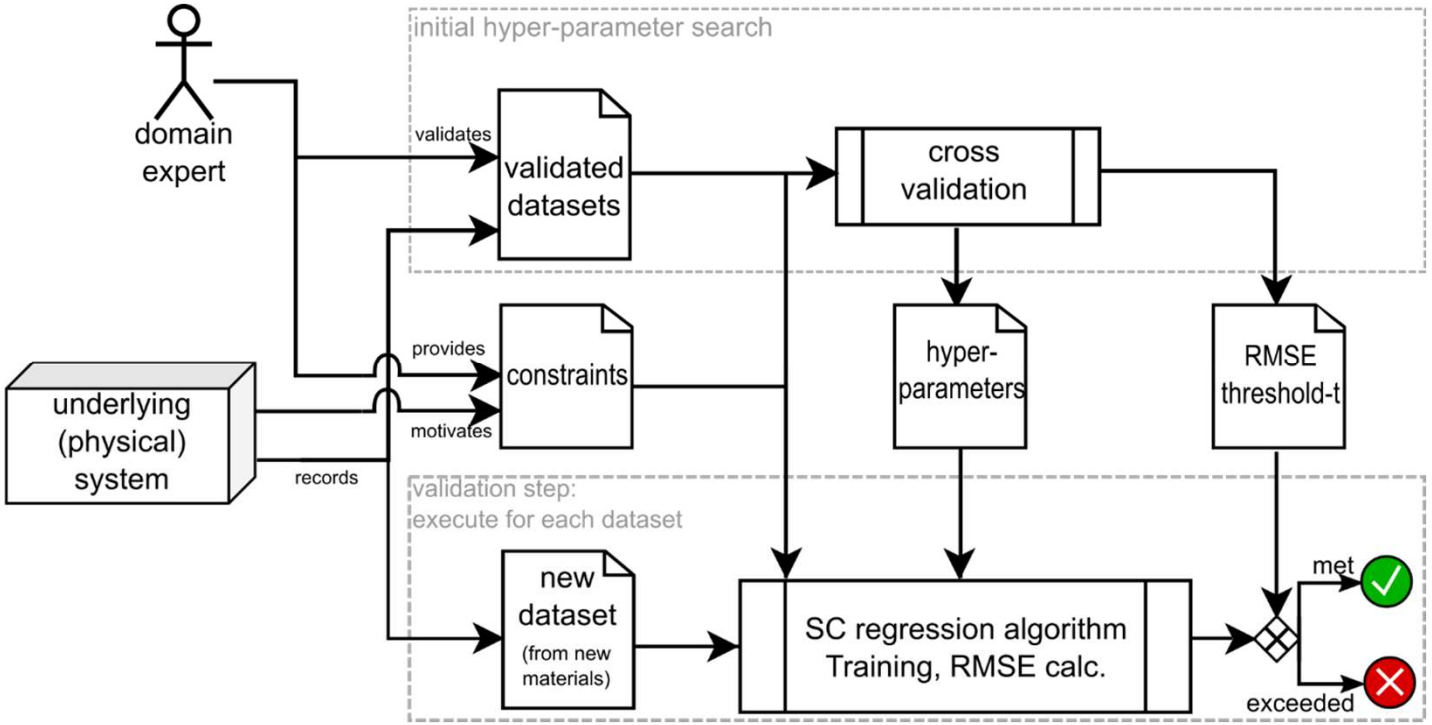

Vary threshold in small increments to change sensitivity of label “invalid”

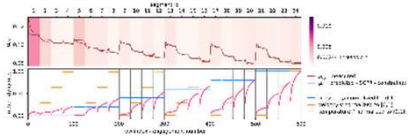
Segments do not comform to expected behavior

SymReg

# Methodology





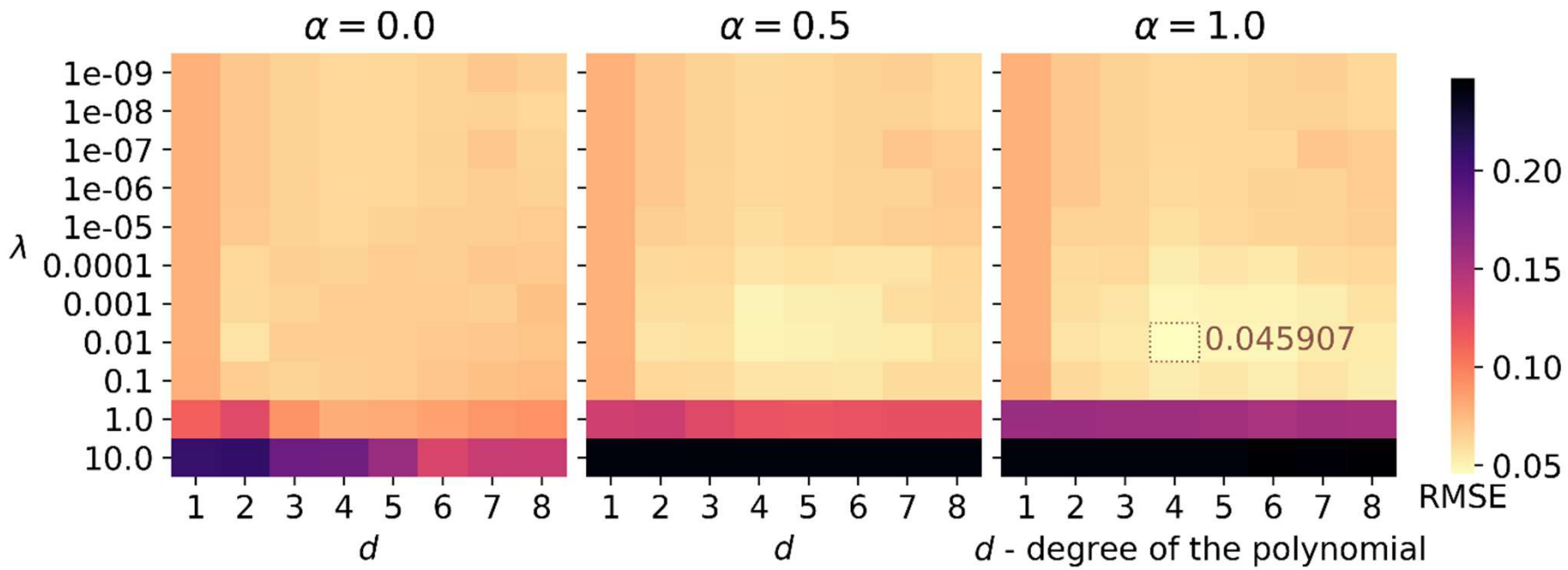
SymReg

## Research Question

How do different shape-constrained regression algorithms compare for data validation?

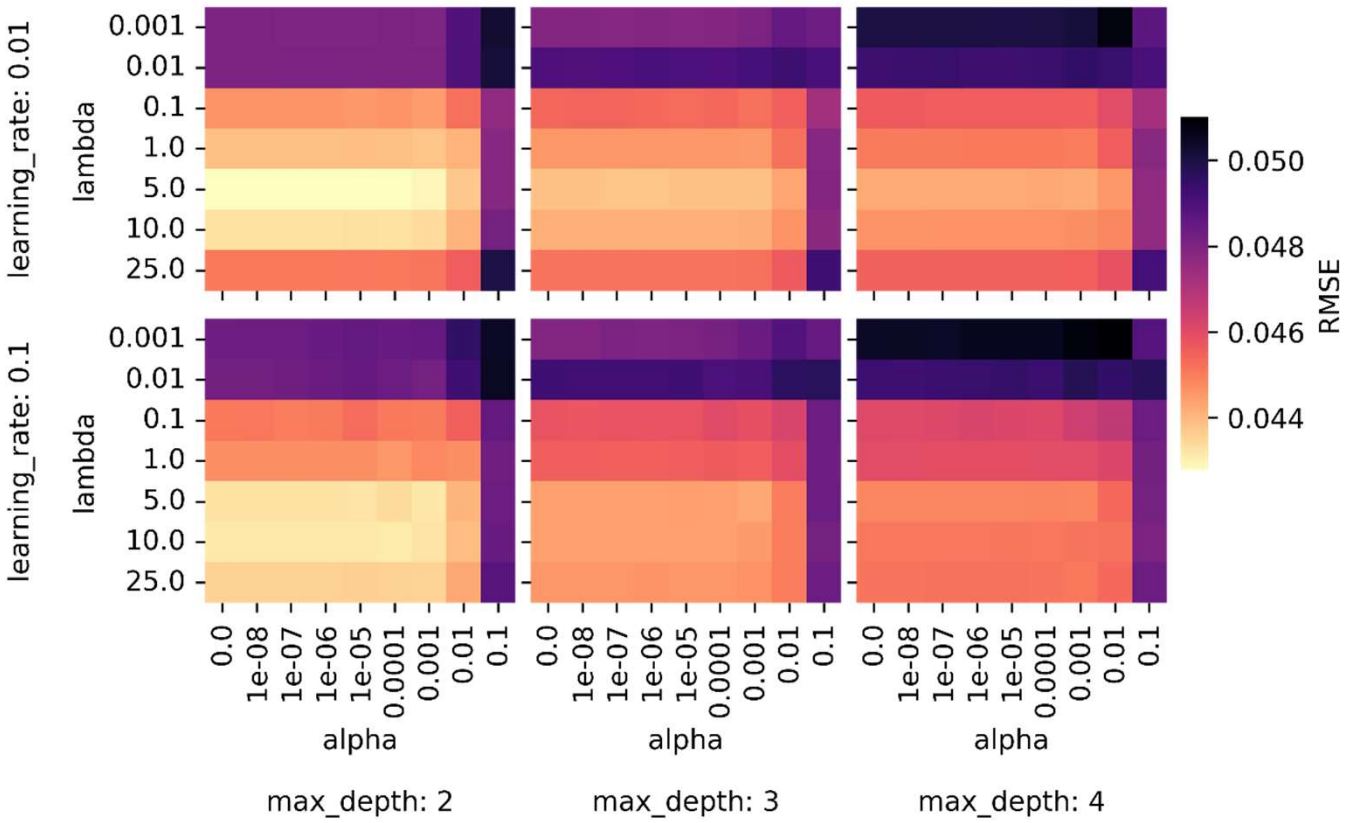
SymReg

# SCPR – Grid Search Results



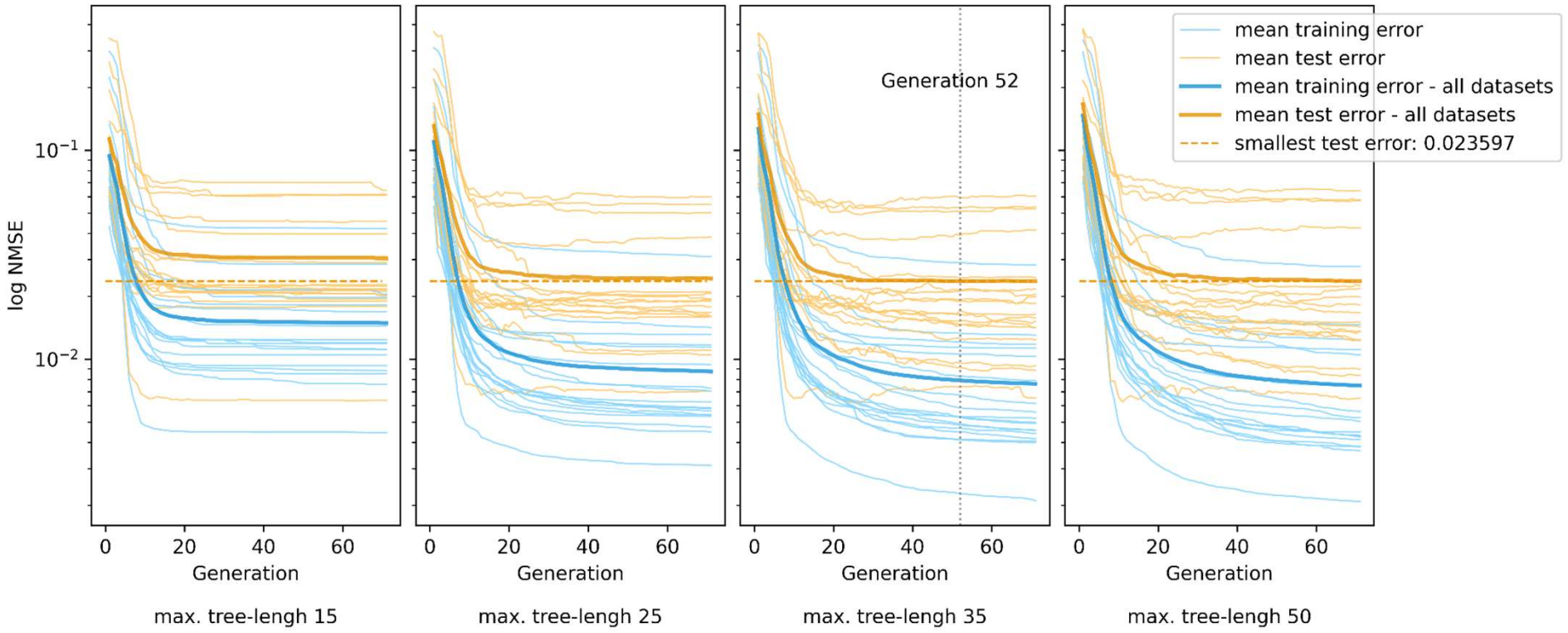

SymReg

# XGBoost - Grid Search Results



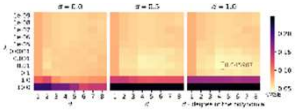

SymReg

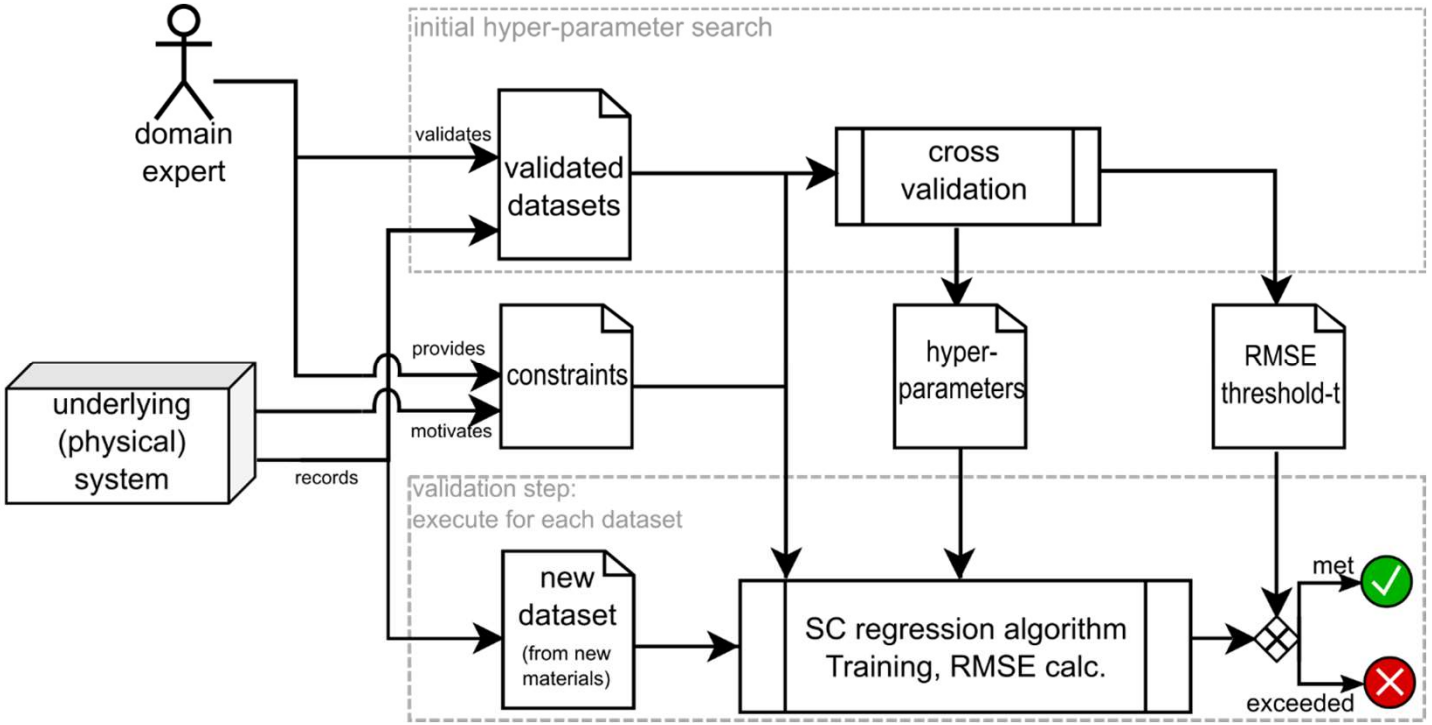
# SC SymReg – Grid Search Results

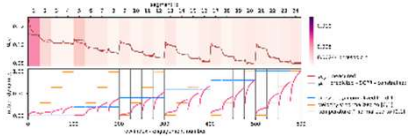



SymReg

# Methodology

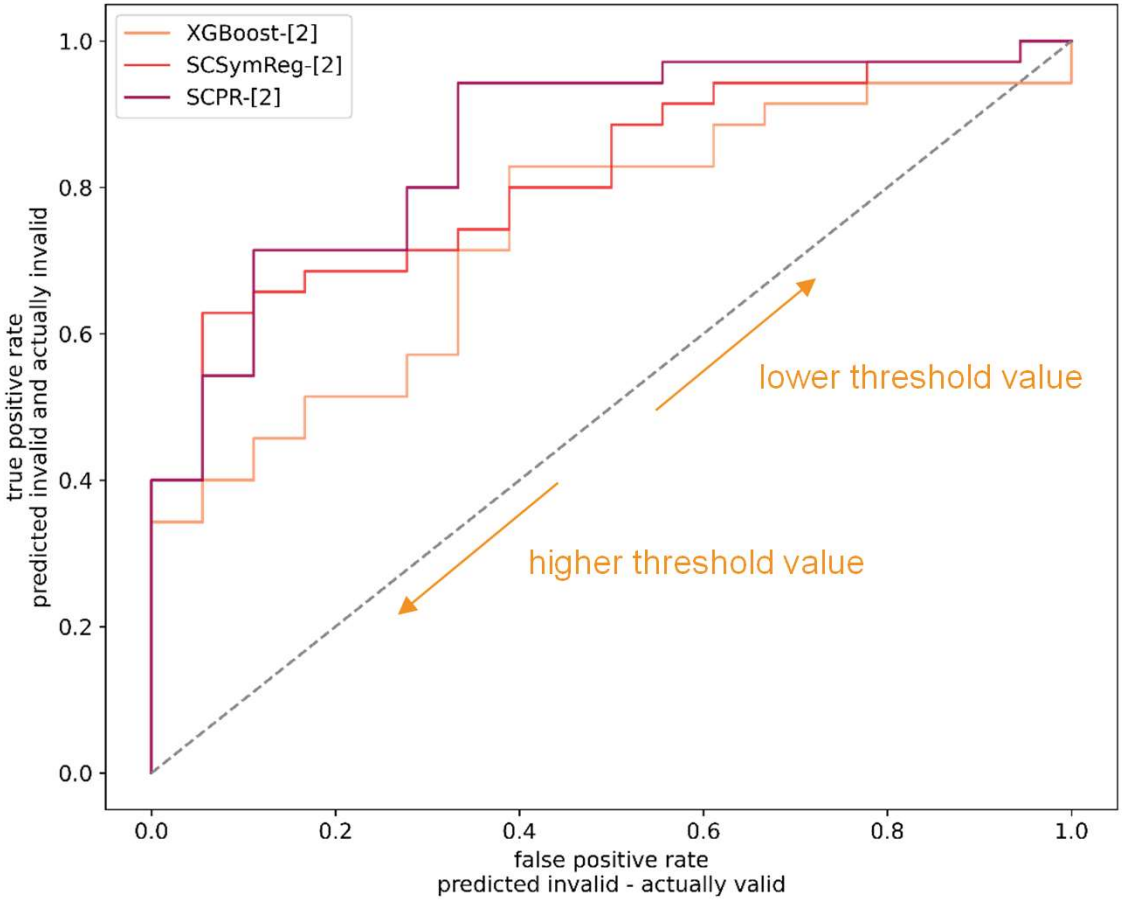


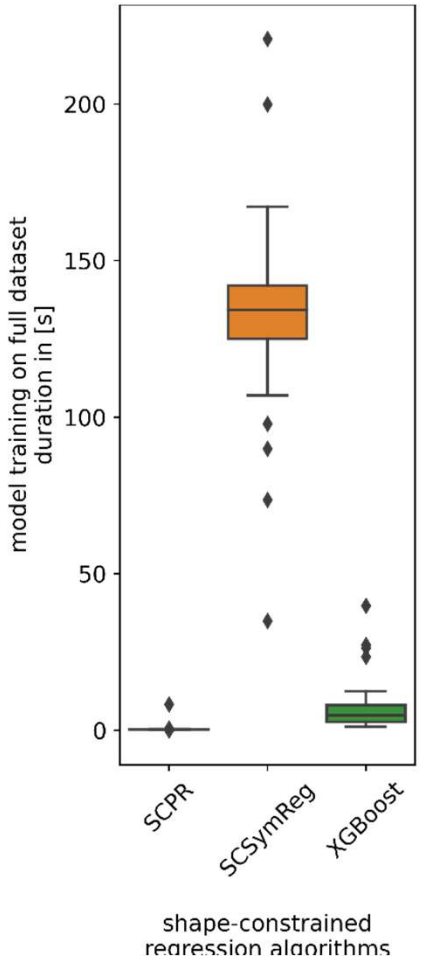


SymReg

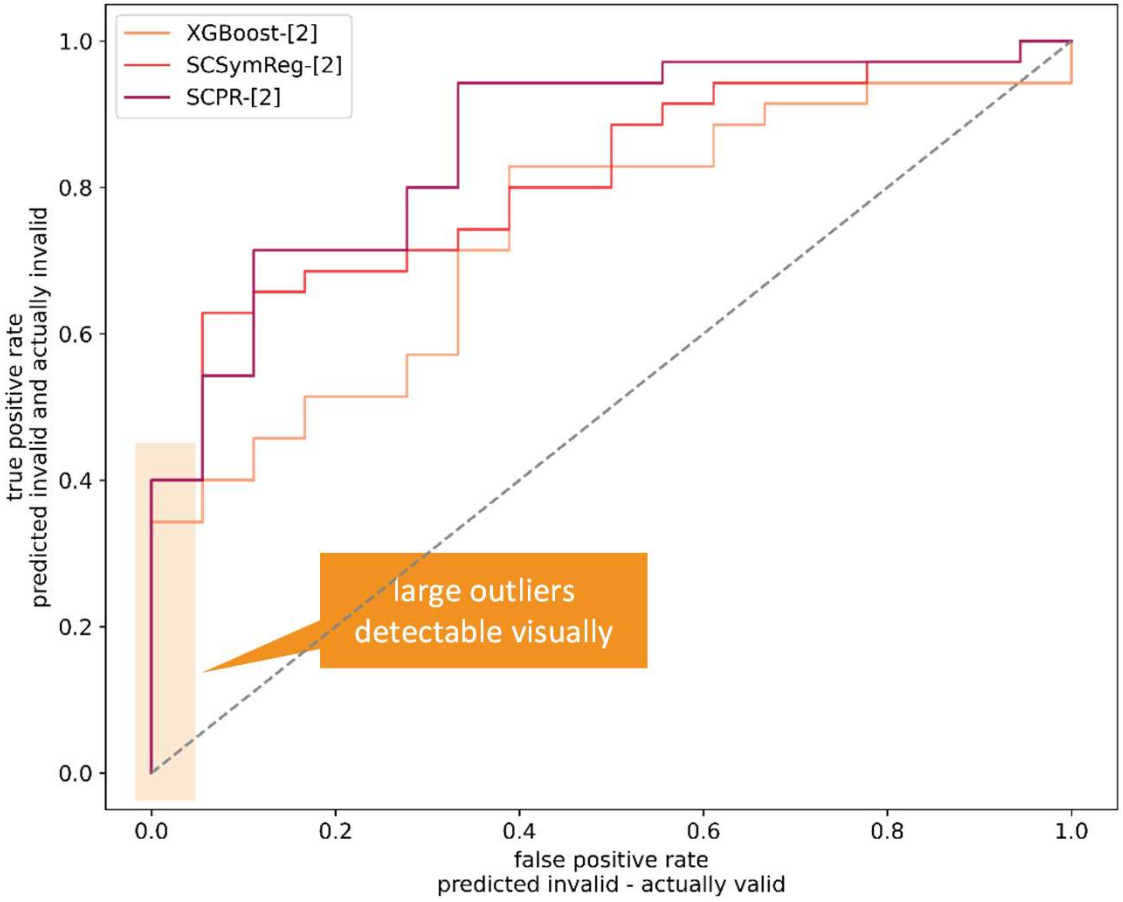
# Classification Results

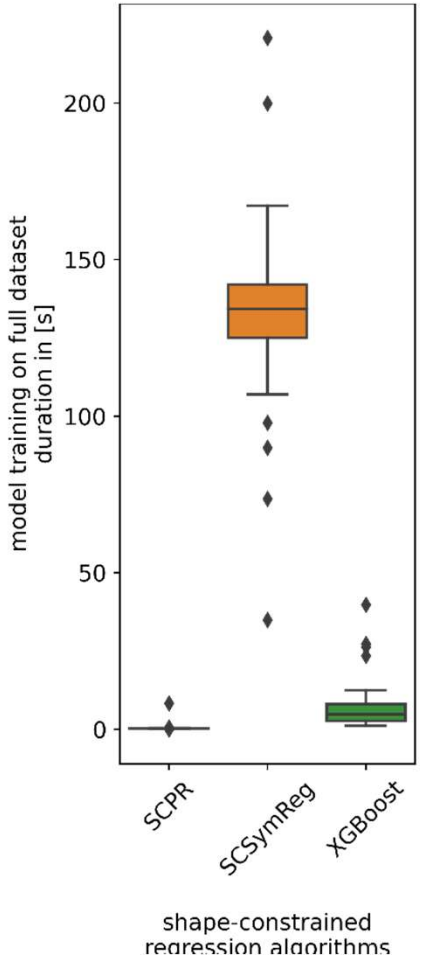



SymReg

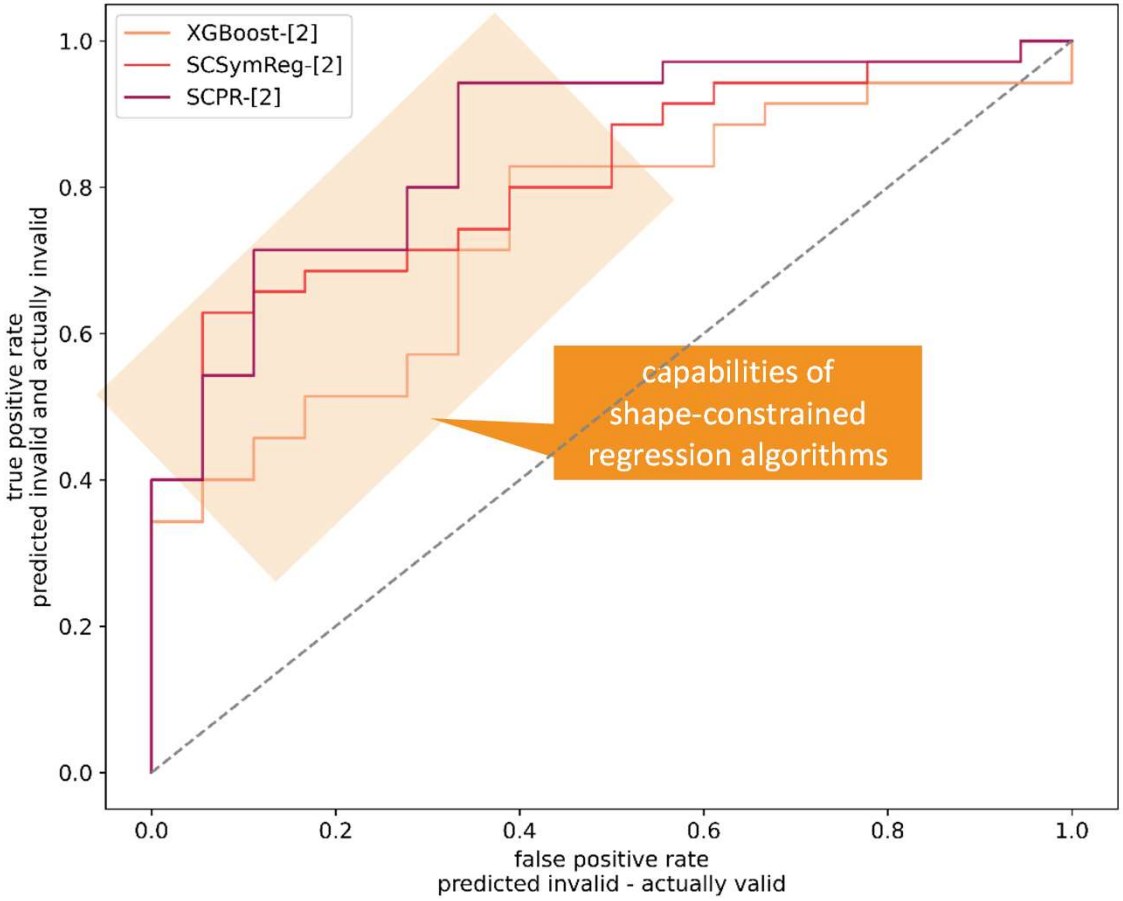
# Classification Results




SymReg

# Classification Results




| Algorithm | Constraints |
|---|---|
| XGBoost | `X = df[['p','v','T']]`<br>`monotone_constraints= "(-1,0,-1)` |
| SC SymReg | $\forall_{v,p,T} \;\; v \in [0,1] \wedge p \in [0,1] \wedge T \in [0,1]$ $\Longrightarrow$ $\Big(0 \leq \mu_{dyn} \leq 1 \wedge \frac{\partial \mu_{dyn}}{\partial v} \in [-0.01, 0.01] \wedge \frac{\partial \mu_{dyn}}{\partial p} \leq 0 \wedge \frac{\partial^2 \mu_{dyn}}{\partial p^2} \geq 0 \wedge \frac{\partial \mu_{dyn}}{\partial T} \leq 0 \wedge \frac{\partial^2 \mu_{dyn}}{\partial T^2} \geq 0\Big)$ |
| SCPR | |

SymReg

# Summary

- Shape-constrained based data validation is fast, quite accurate and **trustworthy**
- Trust facilitated by a-priori knowledge, provides by experts

- SCPR yields **fastest** and **most accurate** classification performance
- SCPR and SCSymReg support more detailed constraints
- Good also for prediction due to constrained interpolation and extrapolation behavior

SymReg

# Best Grid Search Result

| Dataset Id | SCPR - Test RMSE | XGBoost - Test RMSE | SCSymReg - Test RMSE |
|---|---|---|---|
| **0** | 0,003767943 | 0,003395457 | **0,002159833** |
| **1** | 0,003061775 | 0,003399163 | **0,002328463** |
| **2** | 0,00213171 | 0,00174426 | **0,000916964** |
| **3** | 0,00108188 | 0,001122513 | **0,000778793** |
| **4** | 0,001528168 | 0,001806449 | **0,001036082** |
| **5** | **0,002130702** | 0,003210618 | 0,002586256 |
| **6** | **0,002305453** | 0,0029078 | 0,0023627 |
| **7** | 0,002339358 | 0,001943022 | **0,001121476** |
| **8** | 0,001906535 | 0,002167161 | **0,001297637** |
| **9** | 0,001970015 | 0,002283357 | **0,001520898** |
| **10** | 0,002021891 | 0,002013085 | **0,001170407** |
| **11** | **0,00215385** | 0,00332668 | 0,002627362 |
| **12** | 0,001606115 | 0,00189641 | **0,001455251** |
| **13** | 0,002261588 | 0,001856916 | **0,001442107** |
| **14** | 0,002583194 | 0,002226906 | **0,001275521** |
| **15** | 0,002167599 | 0,002476012 | **0,001706853** |
| **16** | 0,002644073 | 0,002221905 | **0,001909599** |
| **17** | 0,003142569 | 0,002783322 | **0,00212639** |

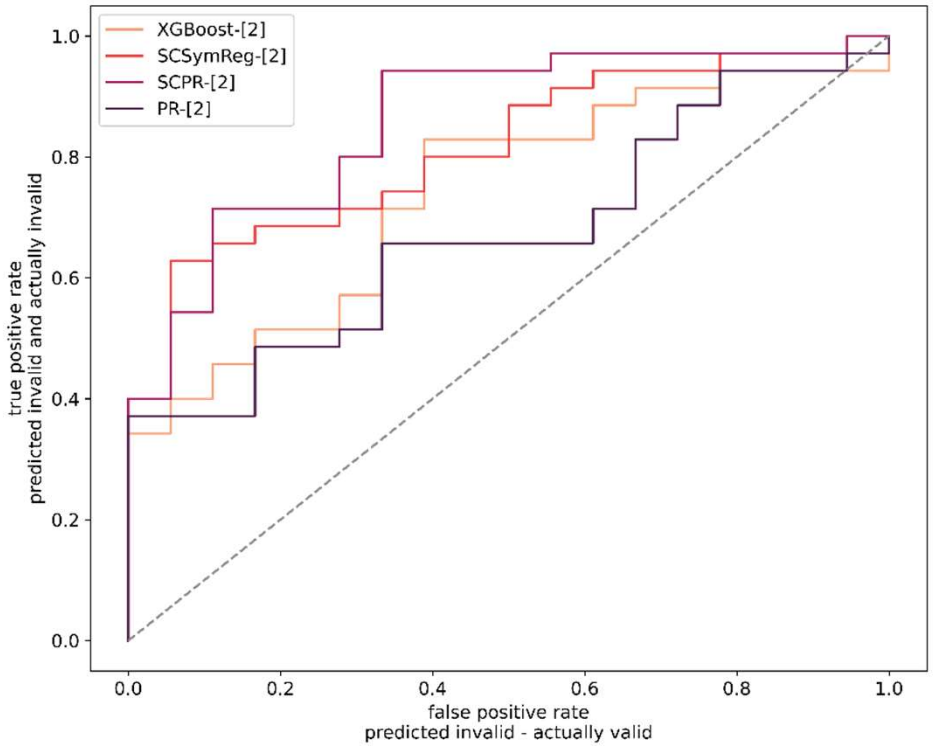
SymReg

# Results on all datasets

| # | actual label | Number of (all occurrences/ correct-/ wrong-predictions) | dataset comment / problem description |
|---|---|---|---|
| 1 | valid | (18/16/ 2) | validated by domain expert |
| 2 | invalid | (35/25/**10**) | all invalid datasets combined |
| 3 | invalid | (16/14/ 2) | physical damage on tested friction package |
| 4 | invalid | (12/ 9/ 3) | test setup issues with effect on results |
| 5 | invalid | ( 7/ 2/ 5) | test setup issues without effect on results |

SymReg

# Comparison

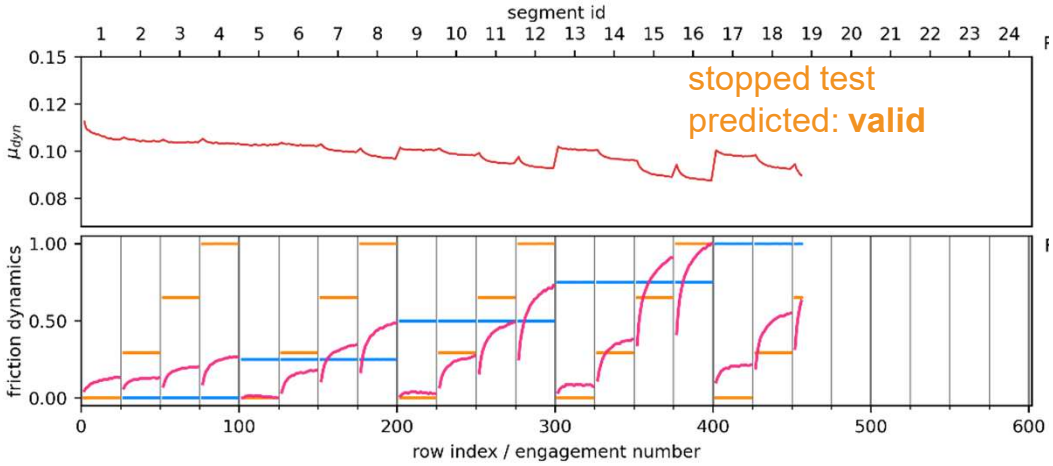


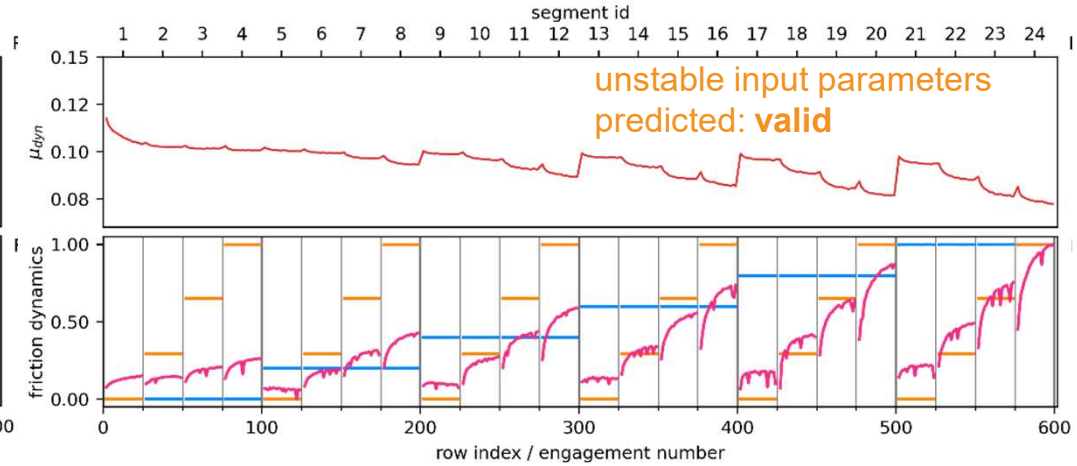

SymReg

# Capabilities

- **validate previously unseen data**
- data stemming from **same (physical) system**
- **using Prior knowledge**

- Cannot detect errors like:



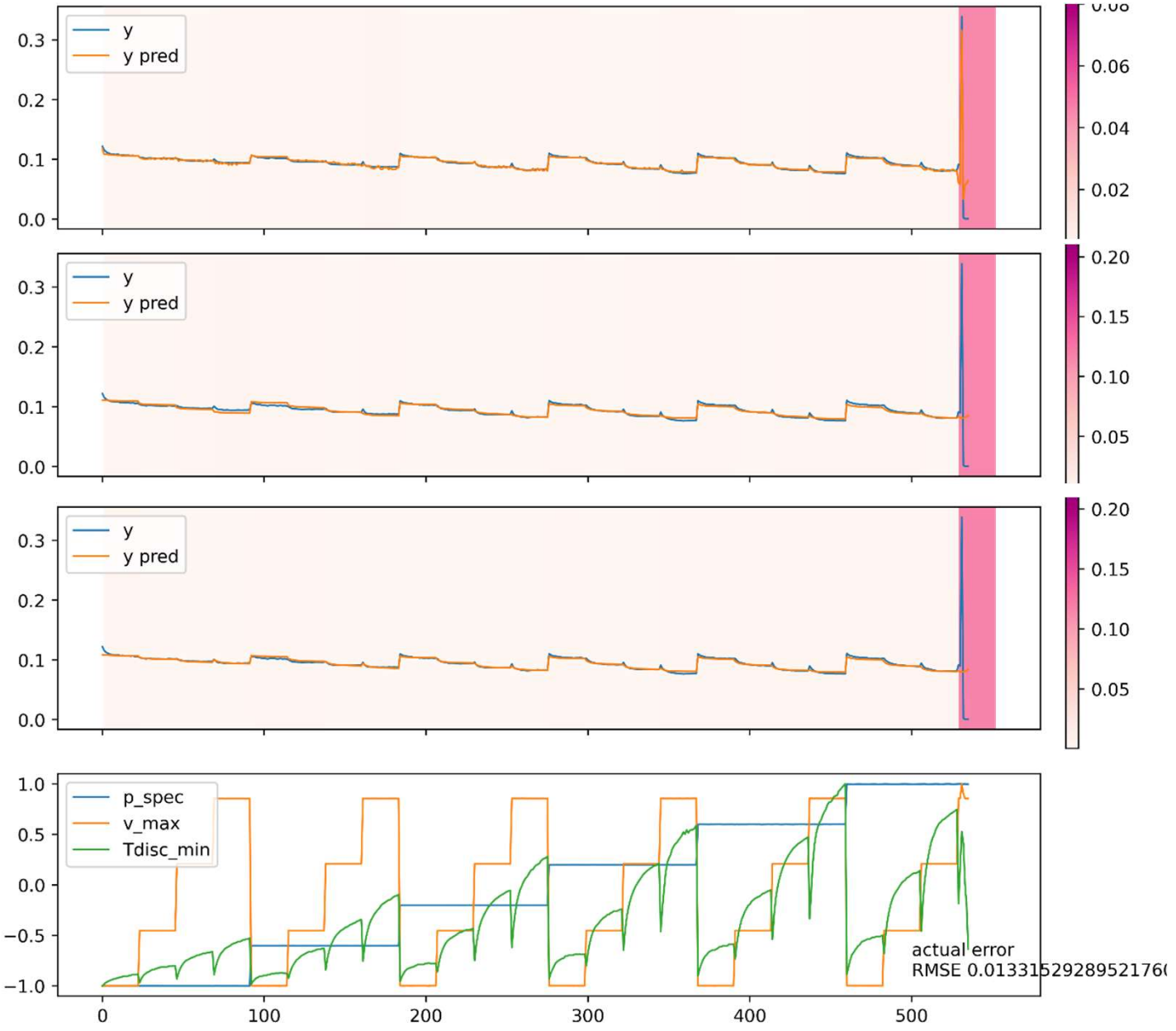




SymReg

XGBoost
RMSE: 0.005469

SC Symreg
RMSE: 0.013499

SCPR
RMSE: 0.013331




SymReg

**Algorithm 1:** Using SCPR for data validation based on RMSE of individual data segments: *scprValidation*

**input** : $X$ : vector of inputs with each $X_i$ of length $n$
$y$ : target values with length $n$
$C$ : the set of constraints
$s_l$ : suitable segment length, depends on application
$\alpha, \lambda, d$ : the best performing algorithm parameters
$t$ : the RMSE threshold applied to each segment
$s_c$ : minimum threshold violations for label *invalid*

**output**: label *valid* or *invalid*

$sRMSE \leftarrow []$ /* RMSE per segment */

$f \leftarrow SCPR(X, y, \alpha, \lambda, d, C)$ /* f - constrained model trained on full dataset */

**for** $0 \leq i \leq \frac{n}{s_l}$ /* iterate over segement ids */ **do**

  $\acute{X} \leftarrow$ subset(data: X, start: $i * s_l$, length: $s_l$)

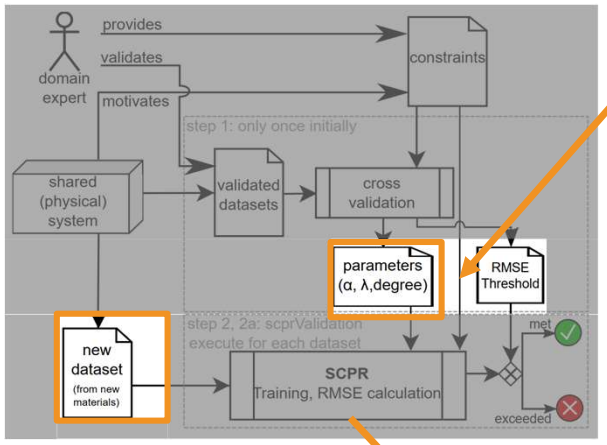
  $\acute{y} \leftarrow$ subset(data: y, start: $i * s_l$, length: $s_l$)

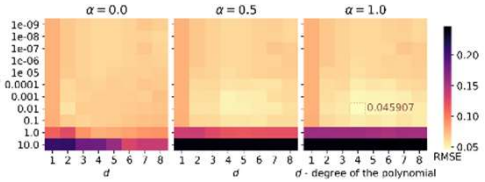
  $sRMSE[i] \leftarrow RMSE(\acute{y}, f(\acute{X}))$ /* Equation 2 */

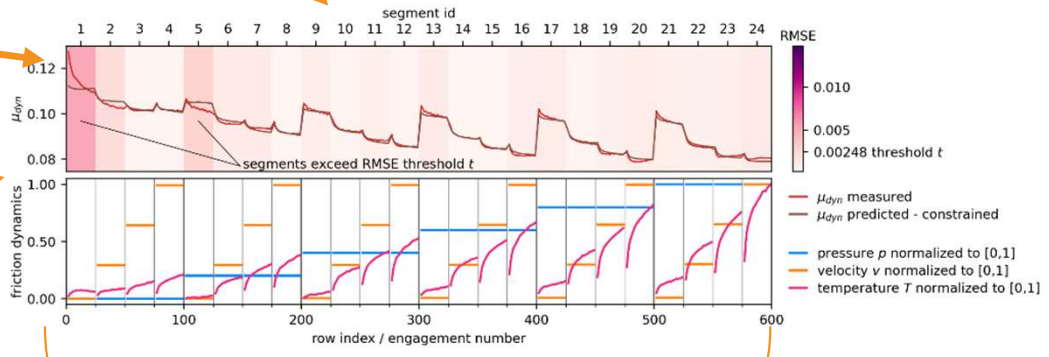
**end**

**if** $count(sRMSE > t) \geq s_c$ **then return** invalid

**else return** valid



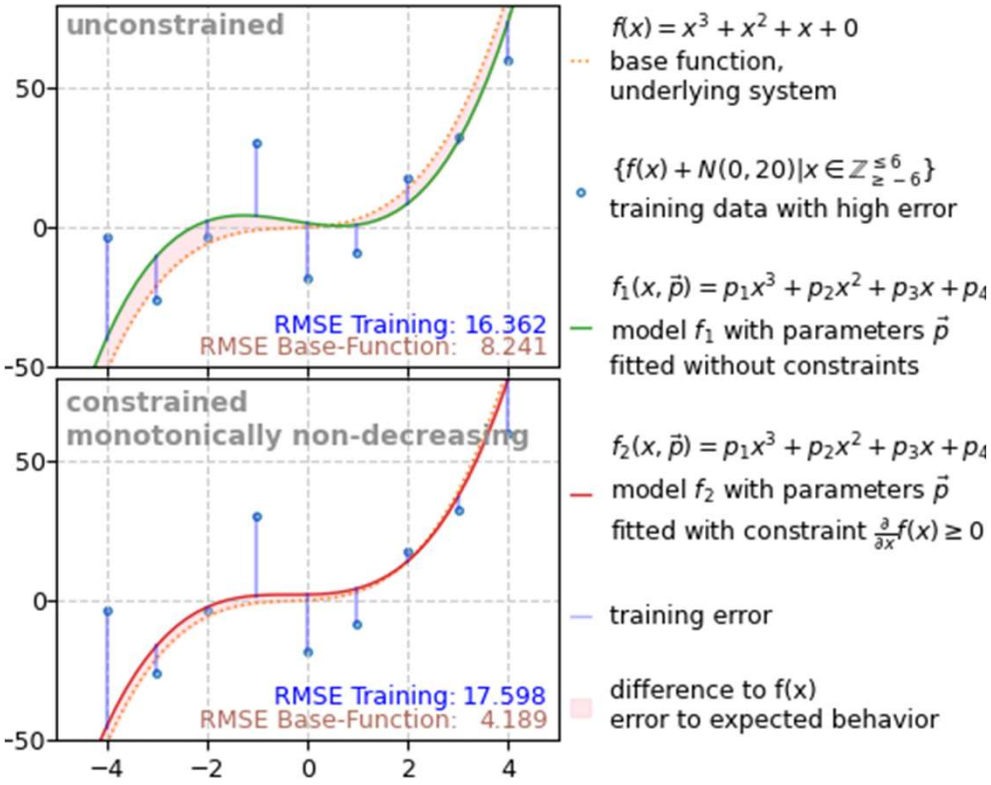




2 -> invalid

SymReg

# Shape-constrained data validation




- $f_2$ is constrained by prior knowledge
- $f_2$ therefore fits closer to $f$
- But $f_2$ has a higher training error
- Meaning higher deviation from our expected system behavior

SymReg