

文献・データ・実験をつなぐ 物理モデル構築の自動化

加藤 祥太

shota@human.sys.i.kyoto-u.ac.jp

京都大学 / 産業技術総合研究所



- ・ 専門：プロセスシステム工学 x データサイエンス x 自然言語処理
- ・ 研究内容
 - ・ 製造プロセスのモデリング、制御
 - ・ 物理モデル構築の効率化

京大院工学研究科
化学工学専攻 修士課程

京大院情報学研究科
システム科学専攻 博士後期課程

京大院・情報学研究科 助教

マンチェスター大学
客員研究員

産総研
招聘研究員

産総研
特定フェロー

2017

2018

2019

2020

2021

2022

2023

2024

2025

2026

JST ACT-X

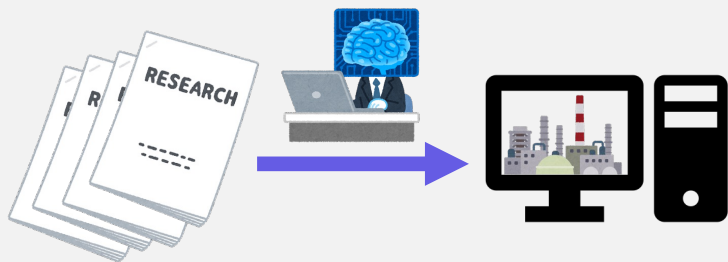
JST さきがけ

JST BOOST

Czochralskiプロセスのグレーボックスモデリング /
グレーボックスモデルに基づく制御手法開発



文献からの物理モデル自動構築AIの開発



データ駆動型
モデル発見手法開発

1. 物理モデル構築の課題と研究全体像
2. 文献からのモデル候補構築：AutoPMoB
3. 数値データからの方程式探索：SINDy
4. 物理知識とAIを用いた候補生成・評価
5. モデルを識別する実験選択と今後の展望

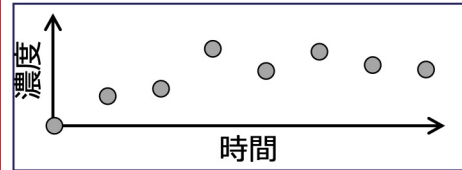
プロセスシステム工学 (Process Systems Engineering, PSE)

- 製造プロセス：石油化学、医薬品、エネルギー産業など（≠ 組み立て）
- PSE とは、製造プロセスの理解・設計・制御を体系的に扱う学問。



石油化学プラント

https://www.ipc-ihl.co.jp/business/oil_plant/index.html



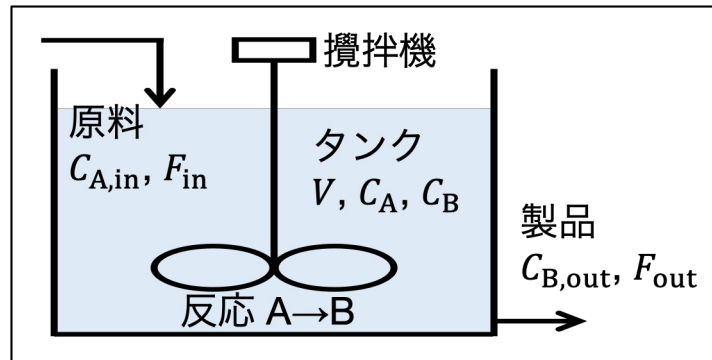
数値データ

$$V \frac{dC_A}{dt} = F_{in}(C_{A,in} - C_A) - rV$$

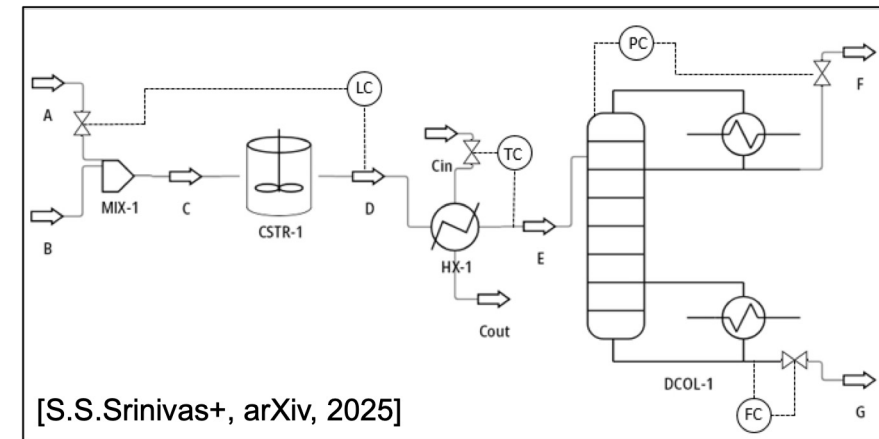
数式

This equation represents the material balance for component A in a CSTR. ...

文



図



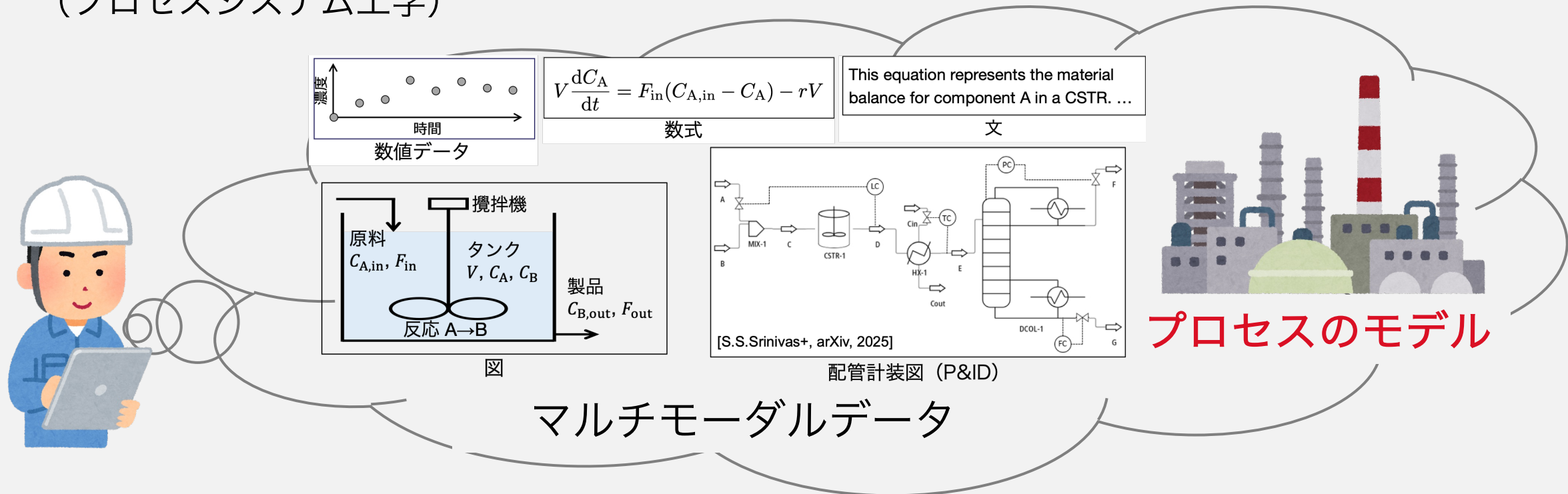
[S.S.Srinivas+, arXiv, 2025]

配管計装図 (P&ID)

PSE ではマルチモーダルデータを統合的に活用している。

プロセスの理解・設計・制御のためのツール：モデル

- 現場のオペレータやエンジニアは、マルチモーダル情報を見つつ、各自の**モデル**に基づいて、プロセスのふるまいを予測している。
- PSE では、計算機上にモデルを作成し、設計や制御に用いる。
(プロセスシステム工学)

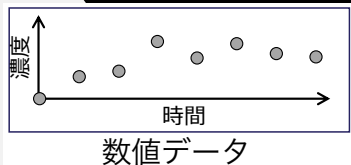


統計モデル

(ブラックボックスモデル)

データと統計的な手法に基づいて構築したモデル

- 線形回帰モデル
- ガウス過程回帰
- ニューラルネットワーク
- ランダムフォレスト
- 勾配ブースティング決定木



ハイブリッドモデル

(グレーボックスモデル)

データ・統計的な手法と科学原理を統合して構築したモデル

- 物理モデル+誤差やパラメータを予測する統計モデル
- 物理モデルの信頼性が高ければ、統計モデルの性能を上回る可能性がある [Geremia et al., Comput. Chem. Eng., 2026]。

物理モデル

(ホワイトボックスモデル)

物理的・化学的原理に基づいて構築したモデル

- 運動方程式
- ナビエ-ストークス方程式
- 熱伝導方程式
- 反応速度方程式
- 状態方程式

$$V \frac{dC_A}{dt} = F_{in}(C_{A,in} - C_A) - rV$$

数式

This equation represents the material balance for component A in a CSTR. ...

文

十分なデータ取得が難しいプロセス産業では、**物理モデル構築が非常に重要**

製造プロセスの物理モデルは、物質・エネルギー収支式に、反応速度式などの関係式を組み合わせ、適切な仮定のもとで構築する。

例) 連続槽型反応器 (continuous stirred tank reactor, CSTR)

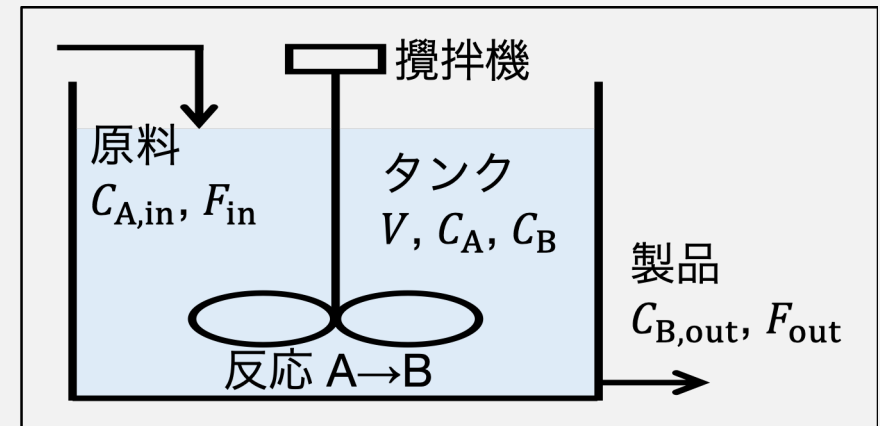
化学工学で用いられる代表的なモデル。不可逆一次反応 $A \rightarrow B$ を対象とし、完全混合・等温・体積一定を仮定する。原料にはAのみが含まれる。

物質収支式 $V \frac{dC_A}{dt} = F_{in} C_{A,in} - F_{out} C_{A,out} - k V C_A$ $V \frac{dC_B}{dt} = -F_{out} C_{B,out} + k V C_A$

アレニウスの式 $k = k_0 \exp \left(\frac{-E}{RT} \right)$

完全混合仮定 $C_{A,out} = C_A$ $C_{B,out} = C_B$

液面一定の仮定 $F_{in} = F_{out}$



チョクラルスキー (CZ) 法による結晶成長プロセス

CZ: Czochralski

- 半導体原料のシリコンインゴットを製造。
- 結晶半径などの複数変数を予測する
微分代数方程式は、数十個の数式を含む。

CZプロセスを表す物理モデルの一部

[Zheng et al., J. Cryst. Growth, 2018; Kato et al., J. Cryst. Growth, 2021]

$$\frac{dT_m}{dt} = \frac{Q_{cm(1)} + Q_{cm(2)} + Q_{cm,bottom} - Q_{me} - Q_m}{m_m C_m} - \frac{T_m}{m_m} \frac{dm_m}{dt}$$

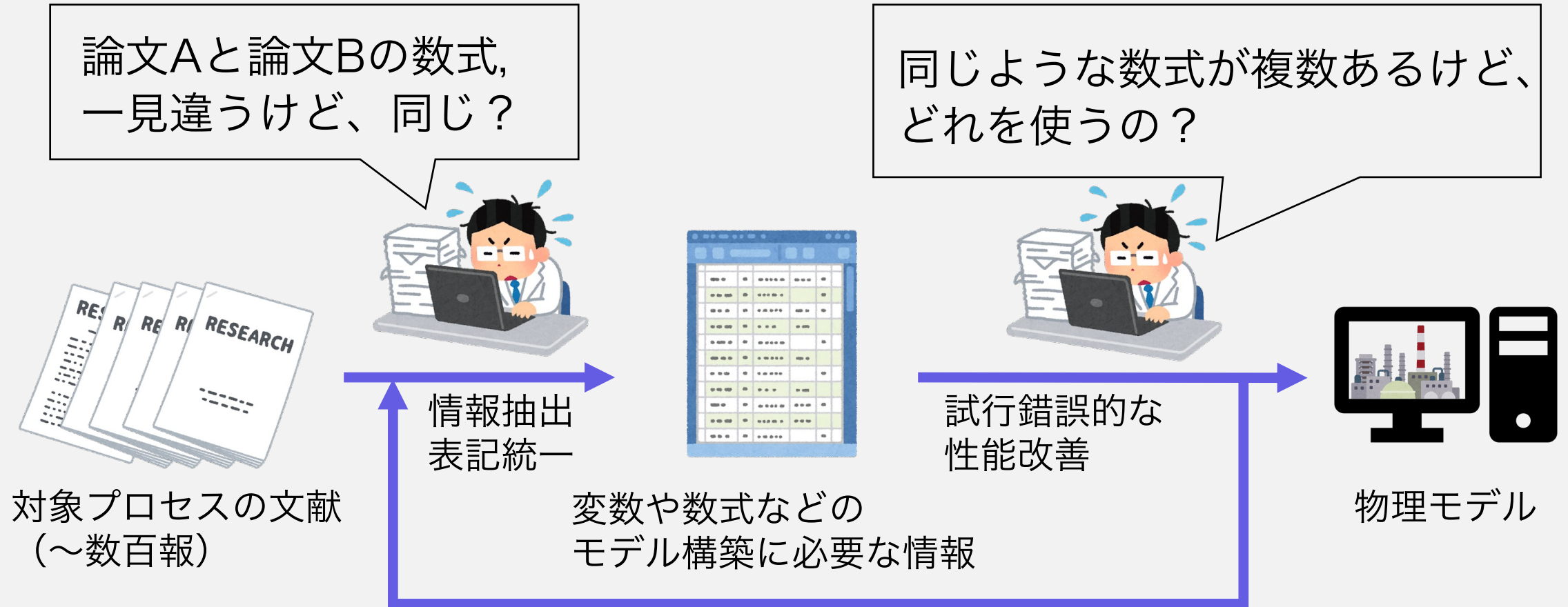
$$Q_{hb} = A_{h,in(3)} F_{hb} \sigma \left(\frac{1}{\varepsilon_h} + \frac{A_{h,in(3)}}{A_b} \left(\frac{1}{\varepsilon_b} - 1 \right) \right)^{-1} (T_{h(23)}^4 - T_b^4)$$

$$\frac{d\varphi}{dt} = \frac{v_c + \psi_1(r_{cry}, \varphi)v_g - v_p}{-\psi_2(r_{cry}, \varphi)} \quad \frac{\partial h_{men}}{\partial r_{cry}} = \frac{a^2(1 - \sin(\varphi_0 + \varphi))^{1/2}}{2^{3/4}r_{cry}^{1/2}(\sqrt{2}r_{cry} + a)^{3/2}}$$

$$\psi_2 = \left(1 - \frac{r_{cry}^2}{r_{c,in}^2} \right) \frac{\partial h_{men}}{\partial \varphi} + \frac{a^2 r_{cry}}{r_{c,in}^2} \sin(\varphi_0 + \varphi)$$



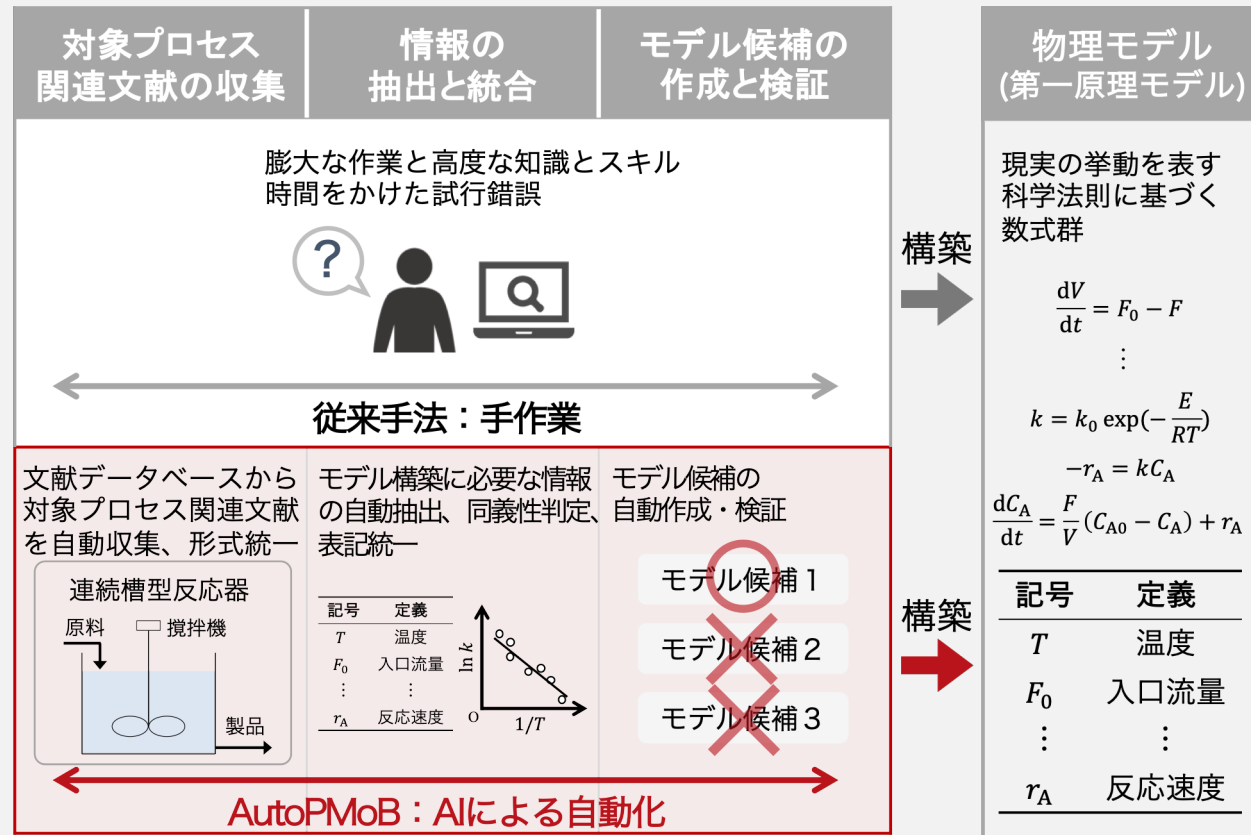
https://www.sumcosi.com/products/process/step_01.html



物理モデル構築には、**膨大な人的・時間的リソースが必要**

文献データ起点のモデリング

文献情報から物理モデルを自動構築する枠組み
Automated Physical Model Builder (AutoPMoB) を提案。
[Kato and Kano, Comput. Aided. Chem. Eng., 2022]



数値データ起点のモデリング

シンボリック回帰 (Symbolic regression) により時系列データから物理モデルを導出する手法をレオロジー分野で実証。
[Sato, Miyamoto and Kato, J. Rheol., 2025]



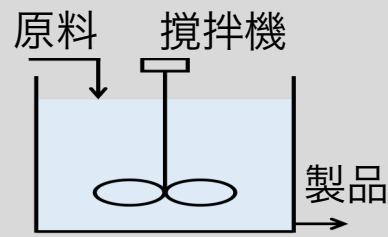
文献・数値データの両方から、物理モデル構築・発見の効率化・自動化に関する研究が進められている。

AutoPMoB の実現に必要な要素技術

対象プロセス関連文献の収集



例：連続槽型反応器



収集した文献の表記統一



モデル構築関連情報の抽出

記号	定義
F_0	入口流量
F_i	供給流量
$\vdots$	$\vdots$

$$-r_A = k_0 \exp\left(-\frac{E}{RT}\right) C_A$$

$$-r = K_0 \exp\left(-\frac{E}{RT_R}\right) C_A^2$$

抽出した情報の表記統一

記号	定義
F_0	入口流量
F_0	供給流量
$\vdots$	$\vdots$

$$-r_A = k_0 \exp\left(-\frac{E}{RT}\right) C_A$$

$$-r_A = k_0 \exp\left(-\frac{E}{RT}\right) C_A^2$$

モデル候補の作成と検証

モデル候補1



$$\frac{dV}{dt} = F_0 - F$$

$\vdots$

$$k = k_0 \exp\left(-\frac{E}{RT}\right)$$

記号	定義
F_0	入口流量
T	温度
$\vdots$	$\vdots$

モデル候補2



$\vdots$

自然言語処理に基づく手法

(Natural language processing, NLP)

化学工学・計算機代数 に基づく手法

現在取り組んでいる範囲

モデル構築関連情報の抽出

記号	定義
F_0	入口流量
F_i	供給流量
$\vdots$	$\vdots$

$$-r_A = k_0 \exp\left(-\frac{E}{RT}\right) C_A$$

$$-r = K_0 \exp\left(-\frac{E}{RT_R}\right) C_A^2$$

抽出した情報の表記統一

記号	定義
F_0	入口流量
F_0	供給流量
$\vdots$	$\vdots$

$$-r_A = k_0 \exp\left(-\frac{E}{RT}\right) C_A$$

$$-r_A = k_0 \exp\left(-\frac{E}{RT}\right) C_A^2$$

モデル候補の作成と検証

モデル候補1



$$\frac{dV}{dt} = F_0 - F$$

$\vdots$

$$k = k_0 \exp\left(-\frac{E}{RT}\right)$$

記号	定義
F_0	入口流量
T	温度
$\vdots$	$\vdots$

モデル候補2



$\vdots$

文献からの物理モデル構築関連情報の抽出

- 物理モデル構築関連の情報として変数と数式を正確に抽出する。

For single-phase reactions, the rate constant is typically a strong function of reaction temperature given by the Arrhenius relation,

$$k = k_0 \exp(-E/RT) \quad (2-63)$$

where k_0 is the frequency factor, E is the activation energy, and R is the gas constant. The expressions in

[Seborg et al., Process dynamics and control, JohnWiley&Sons, 2010]

- 変数記号・数式：表記が通常の単語と異なるため、比較的容易に抽出できる。
→ ルールベース手法 [Kato and Kano, J. Chem. Eng. Jpn., 2025]
- 変数定義：文中の単語列であるため、文の意味を扱う必要がある。
→ NLP技術 [山本ら, 言語処理学会, 2023; Nagayama et al., NLDB, 2024; Kato et al., DAFUSAI, 2024]

BERTモデルを用いた変数定義抽出手法

定義抽出対象の文

T is reactor temperature and ...

1. 文をトークンに分割して変数記号を特殊トークン [target] に変換する。

トークン列

[CLS] [target] is reactor temperature and ...

BERT モデル + 線形層

開始位置の確率

0.02

0.01

0.04

0.32

0.13

0.07

...

終了位置の確率

0.03

0.01

0.05

0.02

0.49

0.02

...

2. 各トークンが定義の開始位置または終了位置である確率を得る。

3. 開始位置 + 終了位置の確率の和が最大となる文字列を定義として抽出する。

抽出される定義

reactor temperature

[Kato+, DAFUSAI, 2024]

- BERT、SciBERT、DeBERTaV3 の変数定義抽出性能を比較。
- 性能は、DeBERTaV3 > SciBERT > BERT。
- 分野によって定義の長さが異なることが性能に影響することを確認。
 - 製造プロセス用にデータを拡張して性能向上 [Nagayama et al., NLDB, 2024]

モデル	正解率	再現率	精度	F1-score
BERT [Devlin+, NAACL-HLT, 2019]	74.4	75.7	67.8	71.4
SciBERT [Beltagy+, EMNLP-IJCNLP, 2019]	79.7	78.5	74.9	76.6
DeBERTaV3 [He+, ICLR, 2023]	81.9	83.3	77.9	80.5

SciBERT：科学関連の論文で学習されたBERT系モデル

DeBERTaV3：ネットワーク構造や学習方法の工夫によって性能を向上させたBERTモデル

- 変数定義：定義抽出と同様にデータを作成して、fine-tune。



→ F : volumetric flow rate



→ λ : the feed flow rate



論文Aの F と論文Bの λ は同義。
表記を統一します。

Fine-tuned ProcessBERT

変数定義の同義性判定用にfine-tune。

[Kato et al., IFAC-PapersOnLine, 2022]

- 数式群：**計算機代数システム**を用いた同義性判定手法を開発。

(Computer algebra system, CAS)



→

$$\begin{aligned} -r_A &= k_0 \exp(-E/(RT)) C_A \\ V \frac{dC_A}{dt} &= q(C_0 - C_A) + r_A V \end{aligned}$$



→

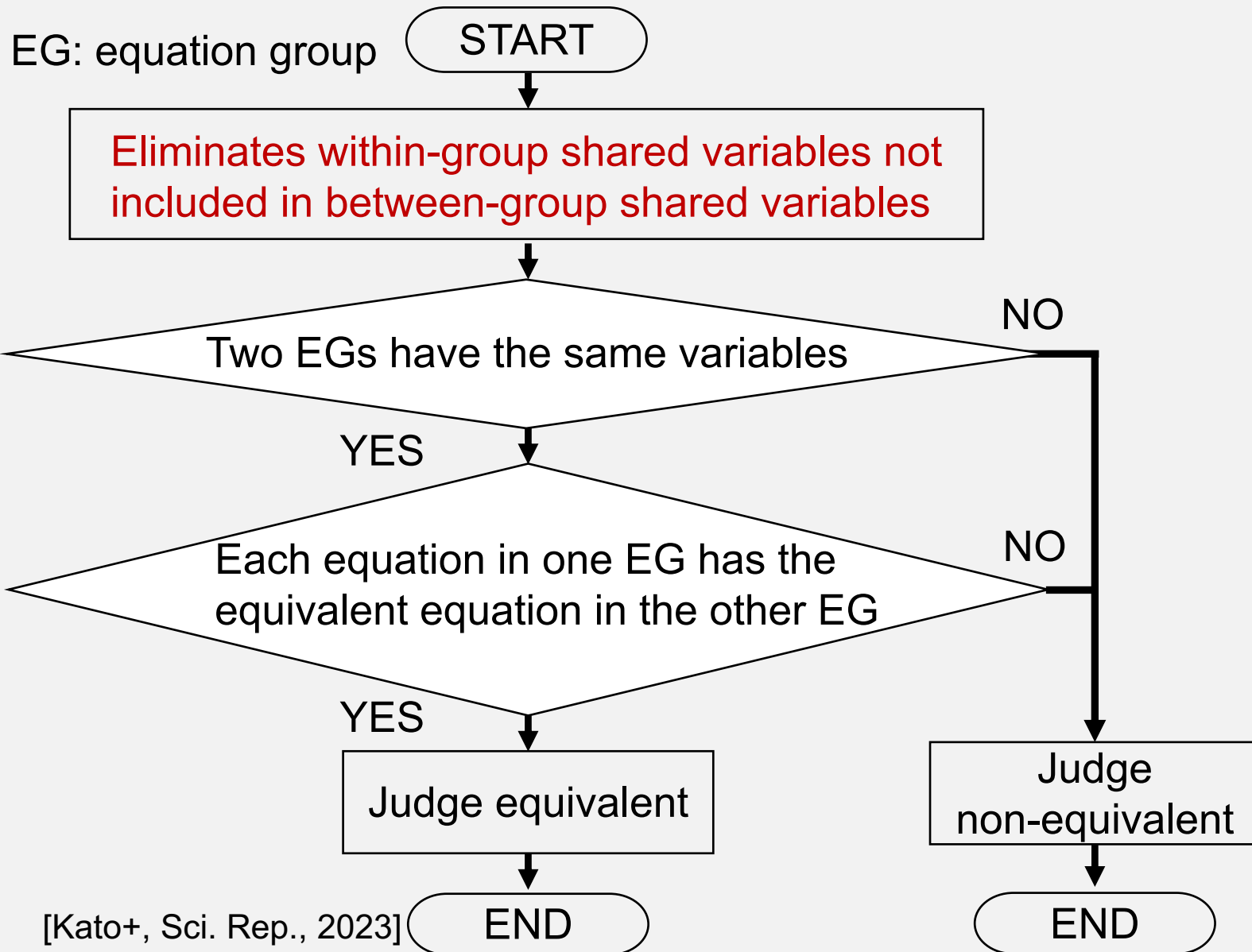
$$\begin{aligned} -r_A &= k C_A \\ k &= k_0 \exp(-E/(RT)) \\ V \frac{dC_A}{dt} &= q(C_0 - C_A) + r_A V \end{aligned}$$



論文Aと論文Bの
数式群は同義です。

CASに基づく

同義性判定アルゴリズム [Kato et al., Sci. Rep., 2023]



	Equation group	Equivalent
E_A	$-r_A = k_0 \exp(-E/(RT)) C_A$	Yes
	$V \frac{dC_A}{dt} = q(C_0 - C_A) + r_A V$	
	$V \rho C \frac{dT}{dt} = wC(T_i - T) + H_r V r_A + UA(T_c - T)$	
E_B	$-r_A = k C_A$	
	$k = k_0 \exp(-E/(RT))$	
	$V \frac{dC_A}{dt} = q(C_0 - C_A) + r_A V$	
E_A	$A \frac{dh}{dt} = q_i - C_v \sqrt{h}$	Yes
	$C_v = C_0 \sqrt{g/g_c}$	
	$P = P_a + \rho gh/g_c$	
E_B	$q = C_0 \sqrt{(P - P_a)/\rho}$	
	$A \frac{dh}{dt} = q_i - q$	
E_A	$-r_A = k C_A$	No
	$k = k_0 \exp(-E/(RT))$	
	$V \frac{dC_A}{dt} = q(C_0 - C_A) + r_A V$	
E_B	$-r_A = k C_A^2$	
	$k = k_0 \exp(-E/(RT))$	
	$V \frac{dC_A}{dt} = q(C_0 - C_A) + r_A V$	
E_B	$V \rho C \frac{dT}{dt} = wC(T_i - T) + H_r V r_A + UA(T_c - T)$	

モデル構築を数式選択問題として定式化する

[Kato and Kano, Comput. Chem. Eng., 2025]

入力

1. ユーザが与える情報（モデル構築目的）：
入力変数（IV）と出力変数（OV）



ユーザ

IV：流量 v_0 , 入口濃度 C_{A0} , 温度 T , 体積 V

OV：濃度 C_A

2. 候補の数式
(文献から収集して表記統一した数式)

$$\begin{aligned} v_0 C_{A0} - v_0 C_A + r_A V &= 0 & k &= k_0 \exp(a/T) \\ v_0 (C_{A0} - C_A) &= kV & k_0 &= 0.1 \\ -r_A &= k C_A & k_0 &= 0.73 \\ -r_A &= k C_A^2 & a &= 10 \end{aligned}$$

出力

4つの条件を満たす数式群

1. 入出力変数を含む。
2. 自由度が指定された入力変数の数に一致する。
3. 不要な変数を含まない。
4. 定数の重複定義がない。

4条件はモデル候補としての必要条件であり、可解性・物理的妥当性は確認が必要。

$$\begin{aligned} v_0 C_{A0} - v_0 C_A + r_A V &= 0 \\ -r_A &= k C_A, \quad k = k_0 \exp(a/T) \\ k_0 &= 0.1, a = 10 \end{aligned}$$

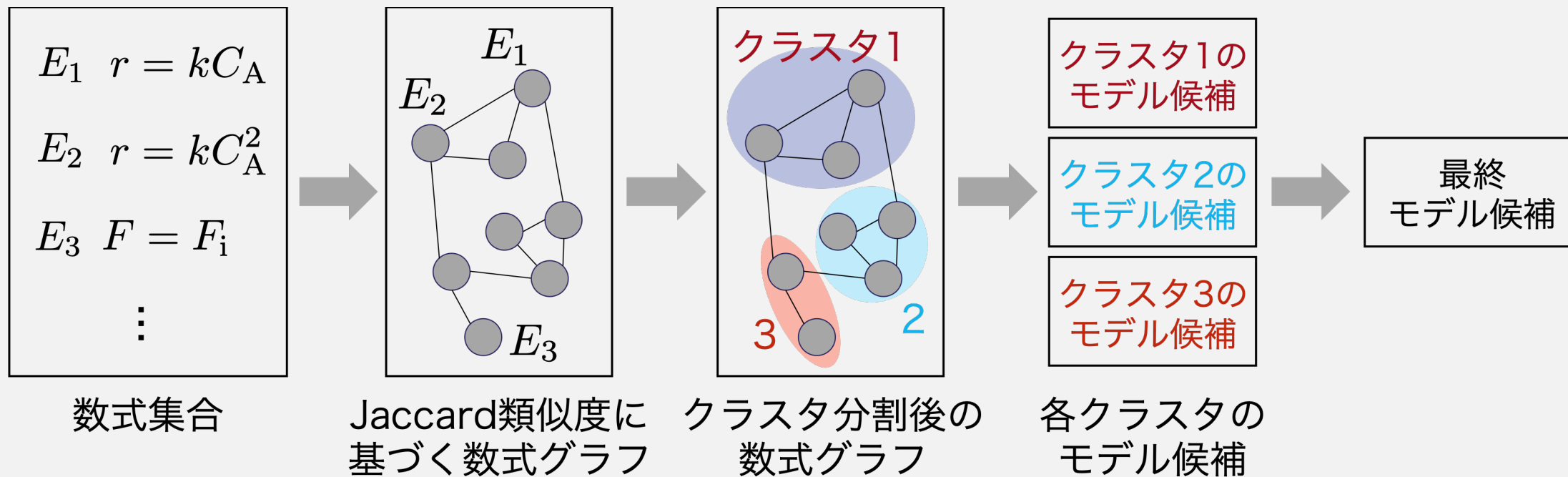
$$\begin{aligned} v_0 (C_{A0} - C_A) &= kV \\ k &= k_0 \exp(a/T), \quad k_0 = 0.73 \end{aligned}$$

共有変数によって数式群の探索を分割する

Q. 4条件を満たすモデル候補をどう見つけるか？

全組合せを列挙せず変数共有関係に基づいて小規模問題へわけろ。

(小規模問題のイメージ：物質収支、エネルギー収支といった異なる観点に基づく分類)



2つのケース (A, B) を対象に、既存手法と提案手法を比較。

例) ケースB [Kato&Kano, Comput. Chem. Eng., 2025]

CSTR 関連文献2報 [Seborg et al., 2010; Bhadriraju et al., 2019] から抽出した29数式。

入力 $S_{IV} = \{T_c\}$, $S_{OV} = \{C_A, T\}$

数式集合 S_e

$$e_1 : r = kC_A,$$

$$e_2 : k = k_0 \exp\left(-\frac{E}{RT}\right),$$

$$e_3 : \frac{d(\rho V)}{dt} = \rho F_i - \rho F, \quad e_4 : F = F_i,$$

$$e_5 : V \frac{dC_A}{dt} = F(C_{A0} - C_A) - V k C_A,$$

$$e_6 : Q = U A (T_c - T),$$

$$e'_1 : \frac{dC_A}{dt} = \frac{F}{V} (C_{A0} - C_A) - k_0 \exp\left(-\frac{E}{RT}\right) C_A^2,$$

$$e'_2 : \frac{dT}{dt} = \frac{F}{V} (T_0 - T) - \frac{\Delta H}{\rho c_p} k_0 \exp\left(-\frac{E}{RT}\right) C_A^2 + \frac{Q}{\rho c_p V},$$

$$e'_3 : F = 1.39 \times 10^{-3}, \quad e'_4 : k_0 = 2.35 \times 10^3,$$

正解モデル：1,024モデル

※パラメータも1数式とカウント。

パラメータの組み合わせだけ

正解モデルがある。

評価指標

再現率 (Recall) 、精度 (Precision) 、

F1値、計算時間

- 再現率 1.0 を維持、計算時間は約 300 倍高速化 (4,669 s → 14 s) 。
精度は 12.5 倍改善 (0.02 → 0.25) し、不要な候補モデルを大幅削減。
- 閾値 0.15 では、クラスタが大きくなり、精度・計算時間がやや悪化。
※ Jaccard閾値: グラフのエッジ密度を制御 (低い → エッジ多 → クラスタ大)

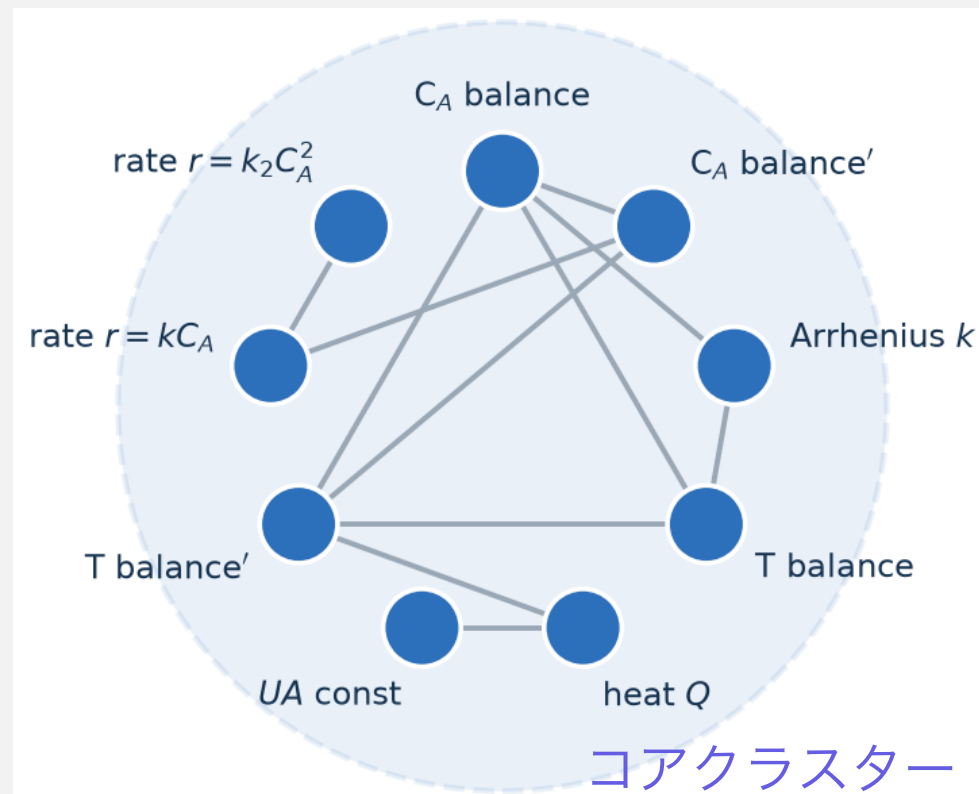
手法	Jaccard閾値	再現率	精度	F1	計算時間 [s]
段階探索法	—	1.00	0.02	0.04	4,620
段階探索法 + フィルタリング [Kato and Kano, ESCAPE, 2025]	—	1.00	0.07	0.13	4,669
提案手法	0.25	1.00	0.25	0.40	14
提案手法	0.15	1.00	0.17	0.29	28

ケースBで探索の分割が機能した理由

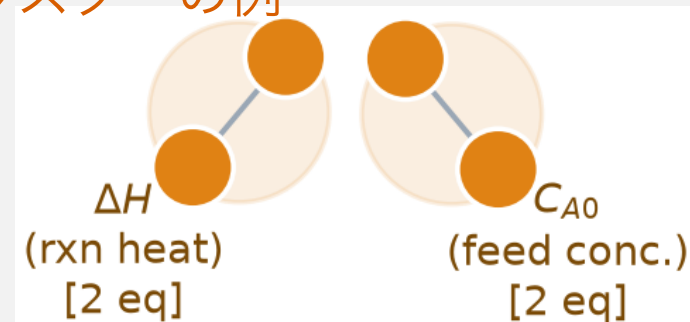
- 10個の互いに独立したクラスターを得た。
- コアクラスターは、すべての支配方程式、残りのクラスターは、パラメータ定義に対応。

重要な知見

- 1,024個の正しいモデルが、クラスターごとの選択の組み合わせとして構築可能。
- クラスターごとに独立して探索することが可能になり、探索空間が劇減した。



残りクラスターの例

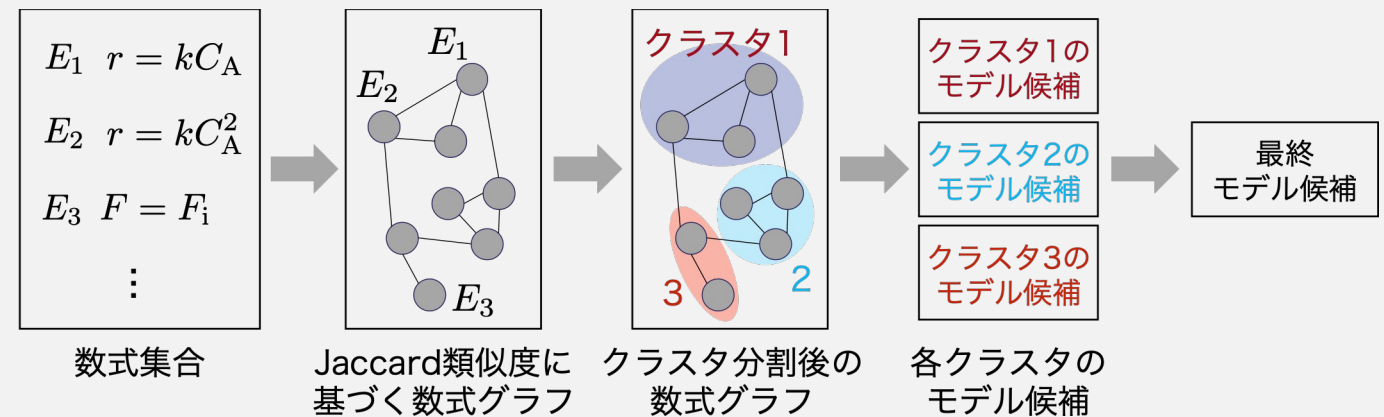


モデル候補作成の成果と課題

- 物理モデル構築を4条件を満たす数式群を見つける方程式組み合わせ問題として定式化した。
- 数式に含まれる変数の共有関係を利用する物理モデル構築手法を提案した。
- ケースBでは、再現率 1.0 を維持しつつ、計算時間を300倍 (4,669 s $\rightarrow$ 14 s) 短縮し、精度を12.5倍改善 (0.02 $\rightarrow$ 0.25) し、不要な候補モデルを大幅削減した。
- ケースAでは、計算時間を50倍 (238 s $\rightarrow$ 5s) 短縮したが、再現率、精度ともに低下した (それぞれ, 1.0 $\rightarrow$ 0.80, 0.36 $\rightarrow$ 0.33) 。

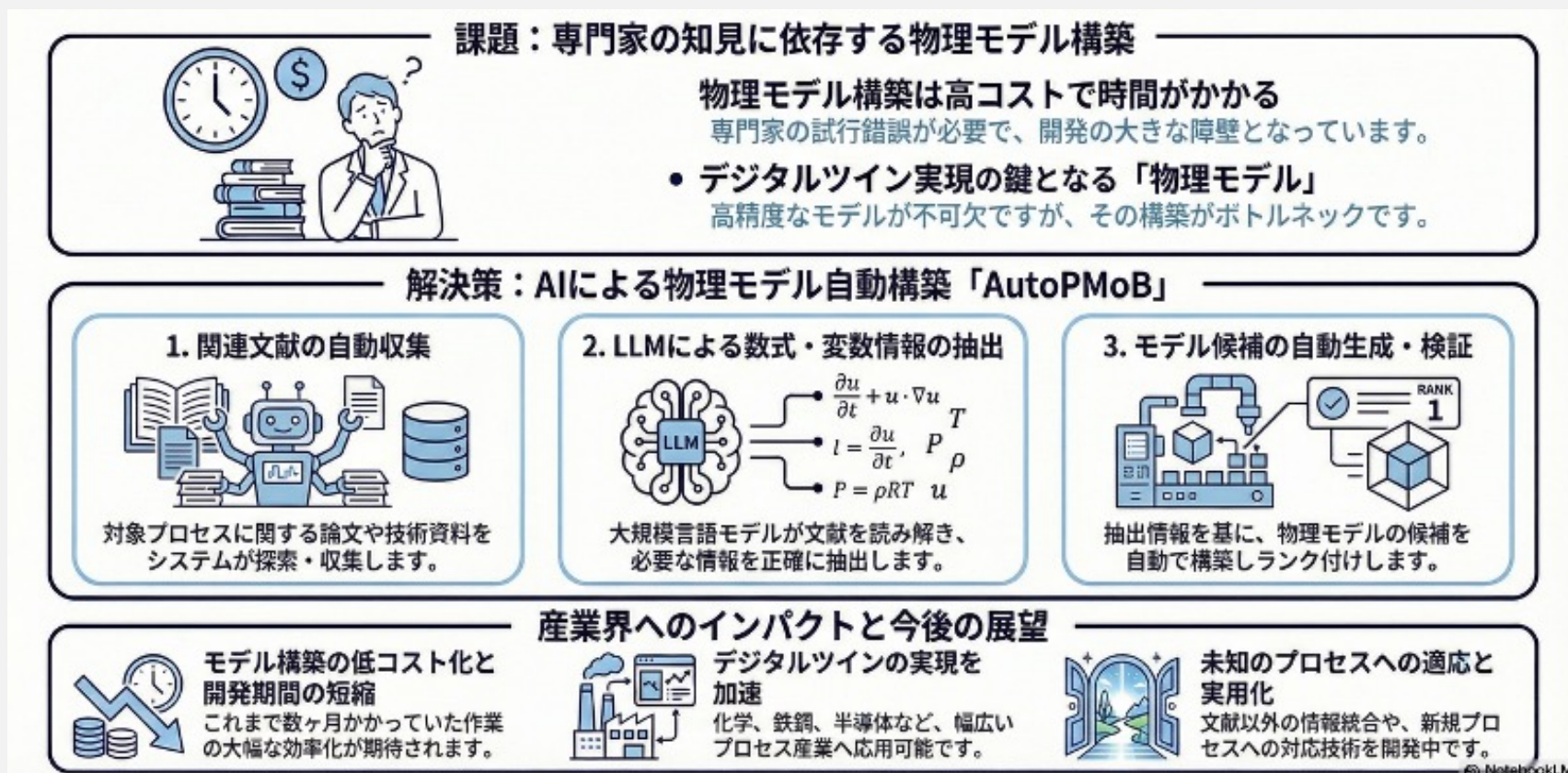
今後の展望

- パラメータの自動調整法の確立
- 数値解法との統合による候補モデルのさらなる絞り込み
- より多様な事例への適用と汎化性の検証

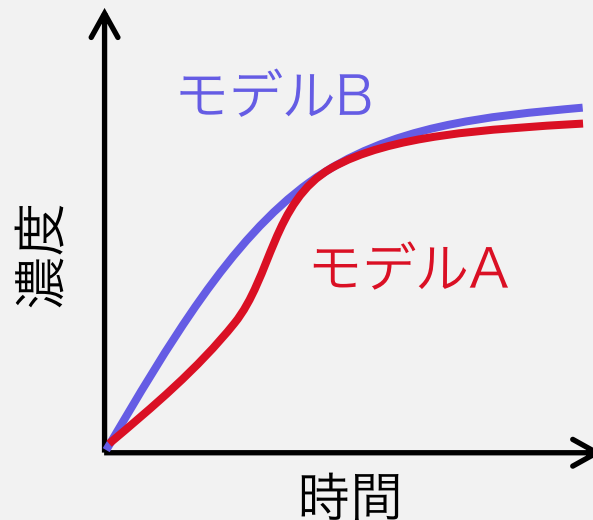
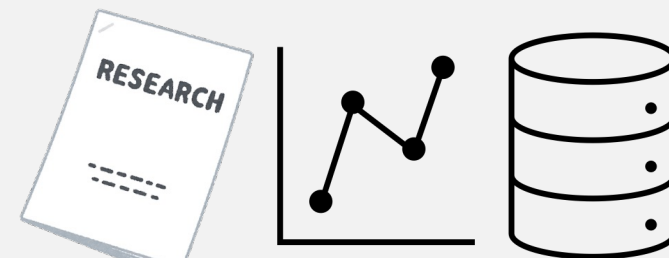


LLMを用いたAutoPMoBのプロトタイプ

- LLM に基づく物理モデル自動構築システムのプロトタイプを作成。
 - 第20回 ICT イノベーションにて優秀研究賞を受賞（発表者：古荘）
<https://ict-nw.i.kyoto-u.ac.jp/ict-innovation/20th/award/> [古荘ら、人工知能学会全国大会, 2026]



モデル候補の評価と数値データの利用



✓ モデル構築が高コスト
→ AutoPMoB で自動化

✗ 統一的なモデルの
評価手法が未確立

✗ 表現したい現象の
データが少ない



物理モデルは得られるけど、どう評価するの？
文献データに加えて、実データも使いたいな…

パラメータ同定と、数式構造の探索

- パラメータ同定では、式の構造を与えて係数を求める。
- シンボリック回帰では、数値データから式の構造も探索する。

パラメータ同定	数式構造の探索
式の構造を与える	使う項も候補から選ぶ
$dx/dt = a x + b u$	$1, x, u, x^2, xu, u^2, \dots$
未知の係数 a, b を推定	残す項と係数を推定

関連研究：数式を探索する二つの方法

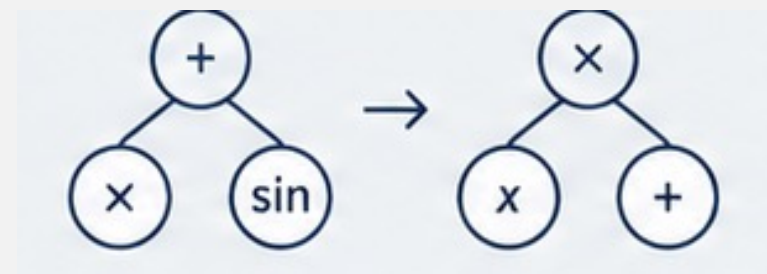
- SINDyは、用意した候補関数から残す項を選ぶ。
- 進化的探索は、式を生成・変異させながら候補を選ぶ。

候補関数から選ぶ：SINDy



候補ライブラリを設計
少数の項と係数を推定

式を生成・変更する：進化的探索

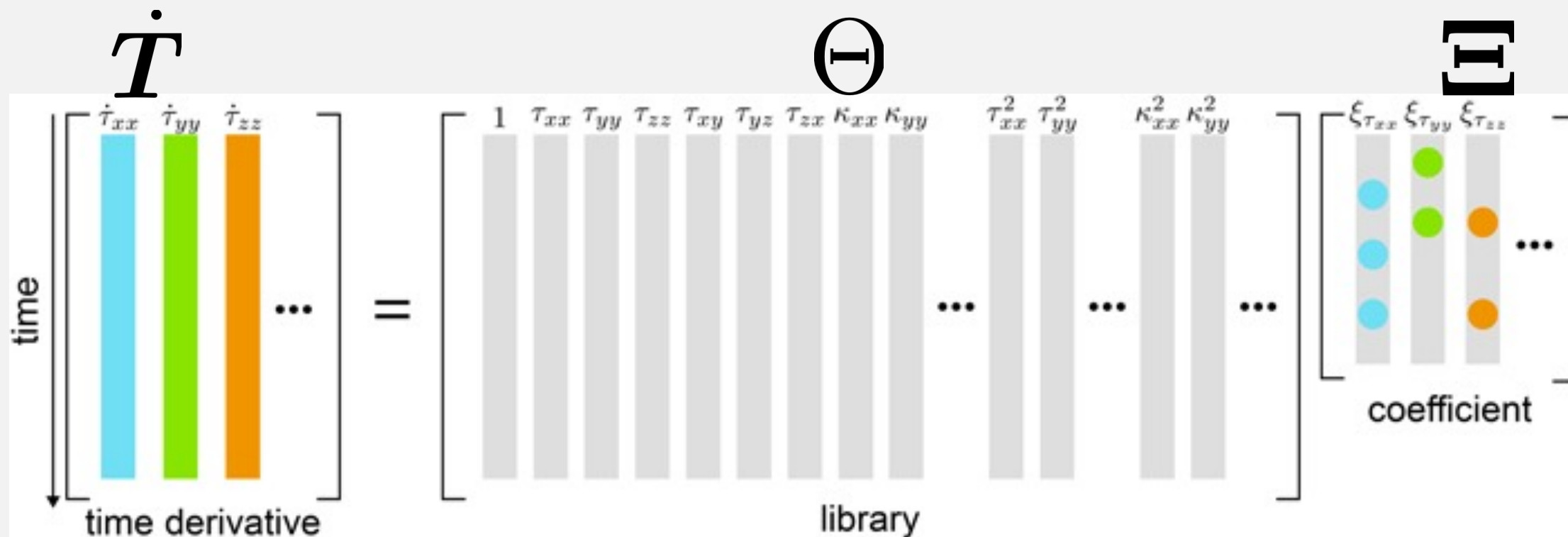


演算子と式の構造を指定
生成・変異・選択を反復

式の探索方法によって、事前に設計する内容が異なる。

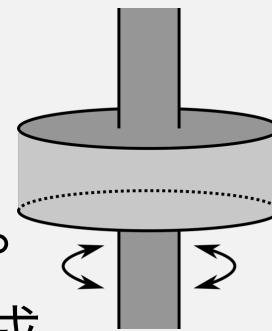
SINDy の概要：スパースな方程式同定手法

- Sparse Identification of Nonlinear Dynamics [Brunton et al., PNAS, 2016]
 - 時系列データにスパース回帰を適用して、支配方程式（物理モデル）を得る手法。
 - 正しいモデルを得るには、関数ライブラリ Θ の設計と、最適化手法の選定が重要。



関数ライブラリの例：レオロジーの構成方程式

- 対象系：粘弾性測定装置（レオメータ）
- 剪断速度 $\kappa(t)$ を与え、剪断応力 $\tau(t)$ を計測する装置。
- Giesekus モデル によるシミュレーションでデータ生成。



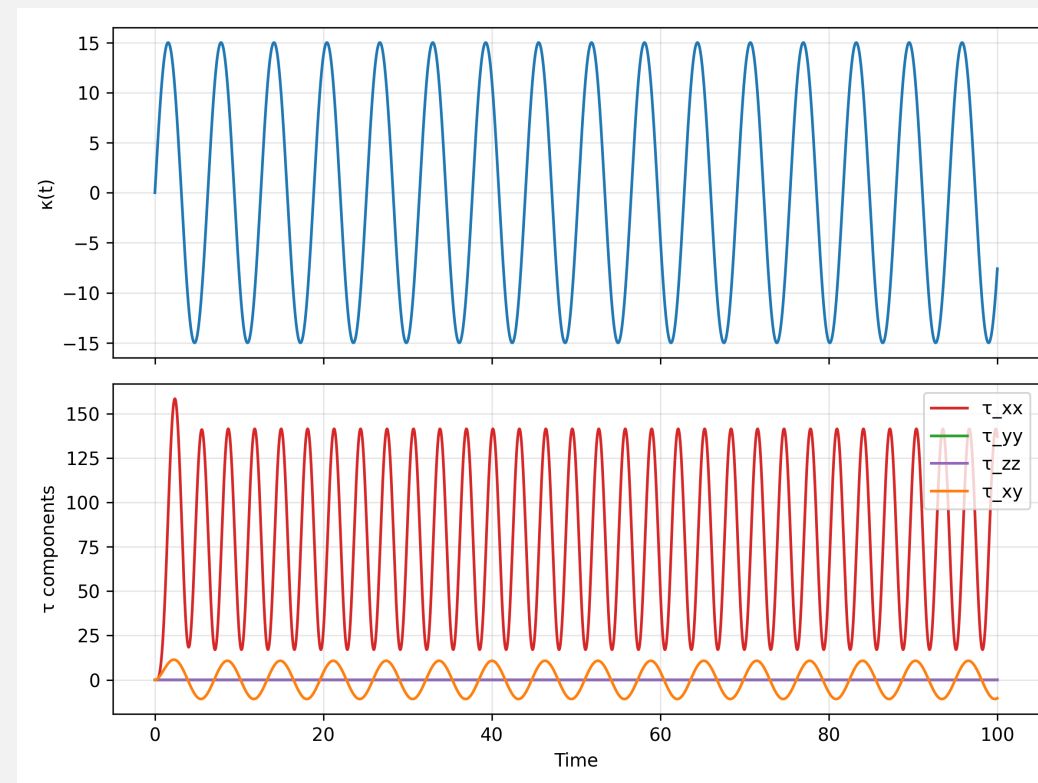
$$\dot{\tau}_{xx} = -\tau_{xx} - \frac{1}{2}(\tau_{xx}^2 + \tau_{xy}^2) + 2\tau_{xy}\kappa_{xy}$$

$$\dot{\tau}_{yy} = -\tau_{yy} - \frac{1}{2}(\tau_{yy}^2 + \tau_{xy}^2)$$

$$(\dot{\tau}_{zz} = 0)$$

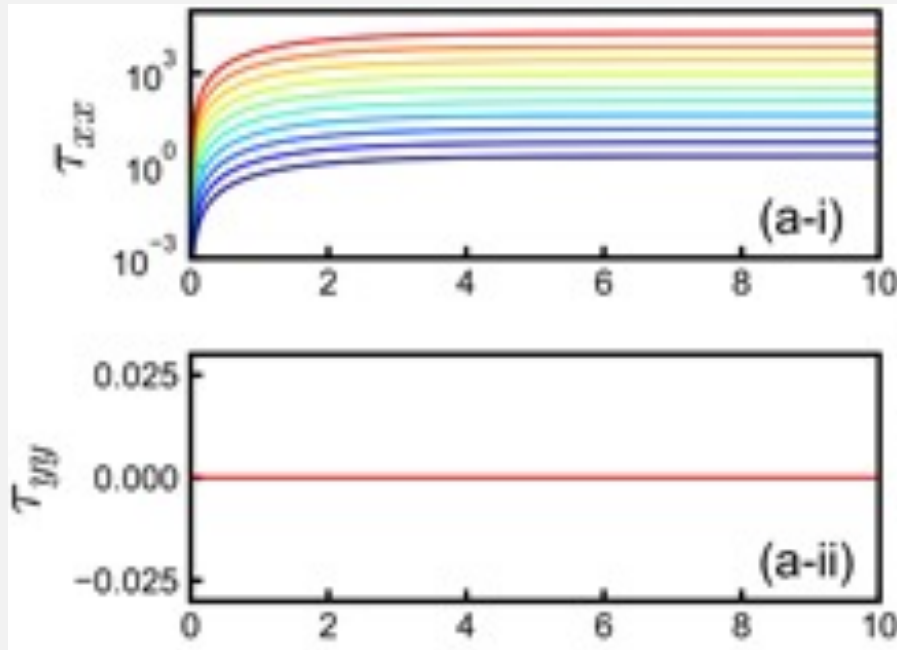
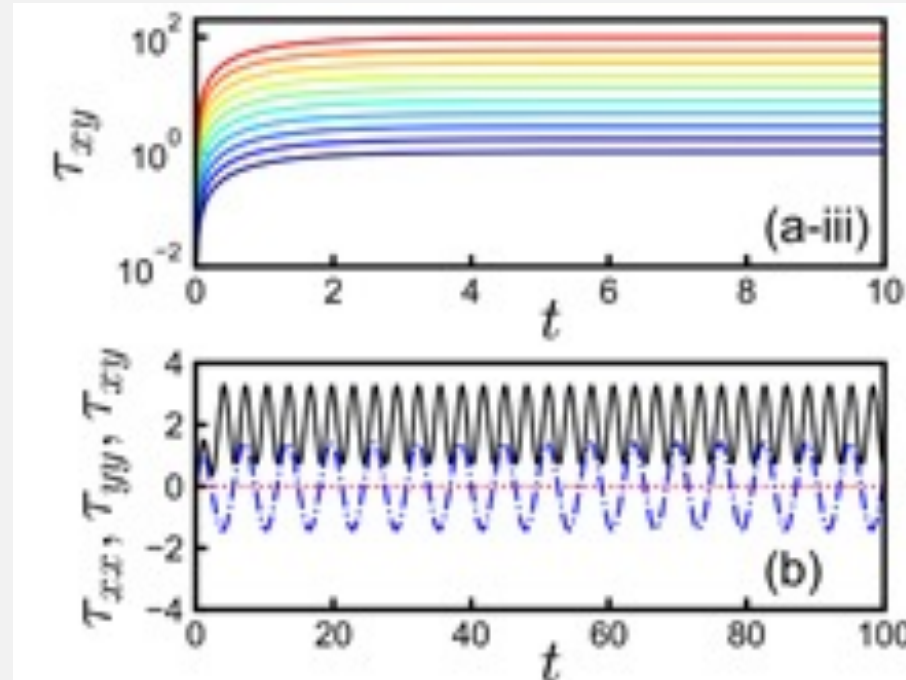
$$\dot{\tau}_{xy} = -\tau_{xy} + \frac{1}{2}(\tau_{xx} + \tau_{yy})\tau_{xy} + \tau_{yy}\kappa_{xy} + \kappa_{xy}$$

- 例：2次の多項式ライブラリを用いると、4変数のときの要素数は15。



Rheo-SINDy：使用する応答データ

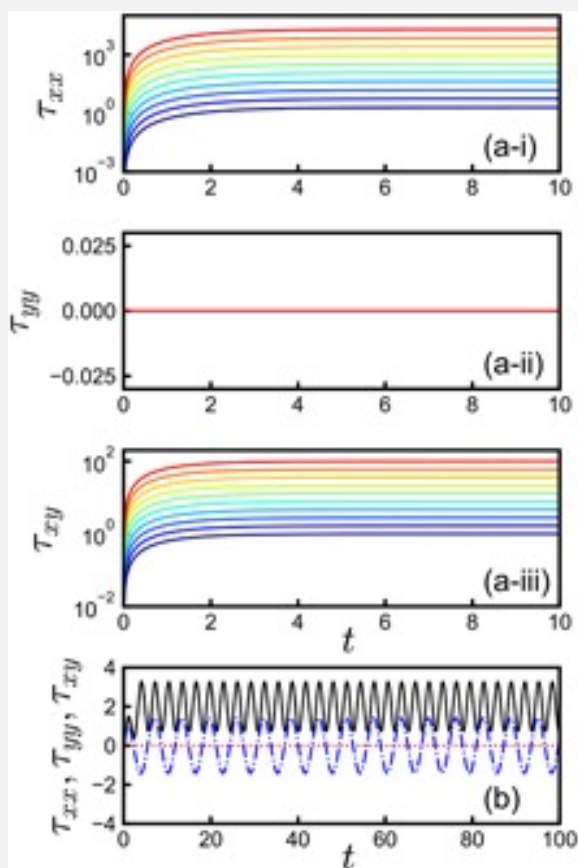
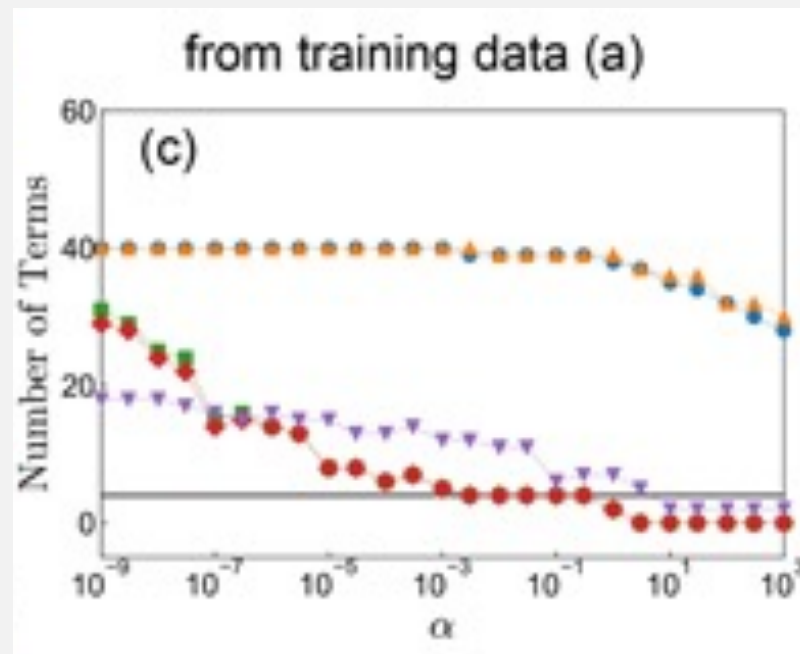
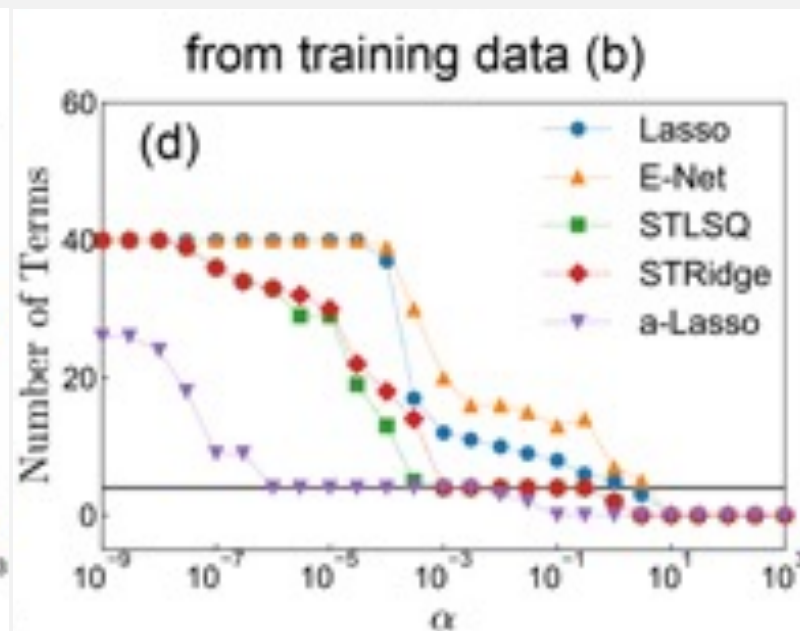
- 応力成分の時系列から、レオロジーの構成方程式を同定する。
- 立ち上がる応答 (a) と振動する応答 (b) を比較する。



どの応答がデータに含まれるかが、方程式の同定に関わる。

Rheo-SINDy : 最適化手法・設定の影響

- 正則化パラメータ α を変えると、方程式に残る項数が変化する。
- 同じデータでも、最適化手法と設定によって残る項が異なる。



Rheo-SINDy：同定された方程式

- データと最適化手法・設定の組合せによって、得られる式が変わる。
- 残った項数だけでなく、同定された方程式を確認する。

data	method	α	equations
(a)	STRidge	3×10^{-3}	$\dot{\tau}_{xx} = -1.000\tau_{xx} + 2.000\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = 0$ $\dot{\tau}_{xy} = -1.000\tau_{xy} + 1.000\kappa_{xy}$
(a)	a-Lasso	1×10^{-3}	$\dot{\tau}_{xx} = -0.998\tau_{xx} + 1.995\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = 0$ $\dot{\tau}_{xy} = 0$

data	method	α	equations
(b)	STRidge	3×10^{-3}	$\dot{\tau}_{xx} = -1.000\tau_{xx} + 2.000\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = 0$ $\dot{\tau}_{xy} = -1.000\tau_{xy} + 1.000\kappa_{xy}$
(b)	a-Lasso	1×10^{-6}	$\dot{\tau}_{xx} = -1.000\tau_{xx} + 2.000\tau_{xy}\kappa_{xy}$ $\dot{\tau}_{yy} = 0$ $\dot{\tau}_{xy} = -1.000\tau_{xy} + 1.000\kappa_{xy}$

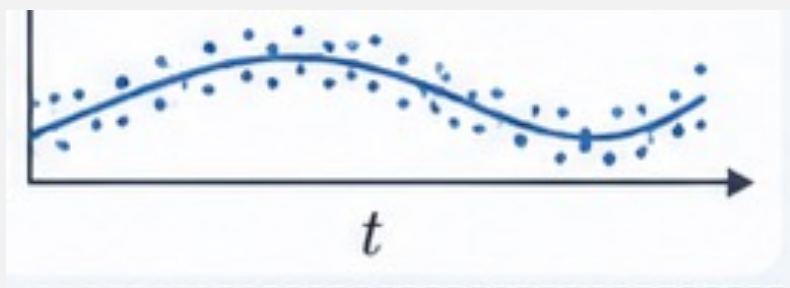
データ・ライブラリ・最適化手法の組み合わせが重要。

関連研究：ノイズ・少数データへの対応

- Weak SINDy：時間微分を直接求めず、積分を用いて式を同定する。
- Ensemble-SINDy：データの再標本化で同定を繰り返し、結果を集約する。

Weak SINDy

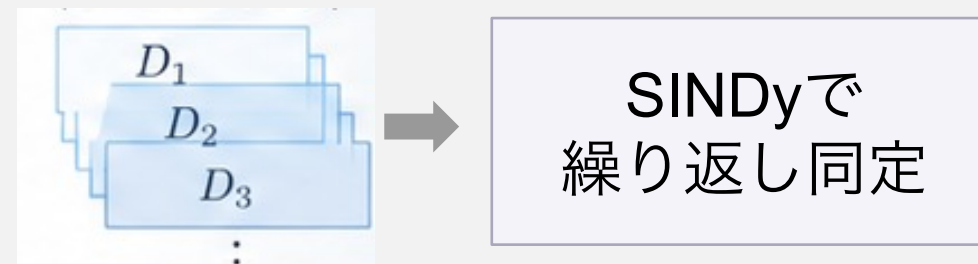
微分推定の影響を減らす



観測データを積分形式の
回帰問題に変換

Ensemble-SINDy

再標本化と集約で不確かさを扱う



項の採用確率や
予測のばらつきを確認

- **関数ライブラリと最適化手法を適切に選択することが重要。**
 - 関数ライブラリ：どの変数・関数を使うか？相互作用は入れるか？
 - 要素が不十分：真の項が含まれないので、正しいモデルを得られない。
 - 要素が過剰：無関係な候補が増加し、スパース回帰が不安定化する。
 - 最適化手法の選択肢：どの手法を使うか？全部試してチューニング？
 - 例：least absolute shrinkage and selection operator (Lasso)、Elastic-Net (E-Net)、adaptive-Lasso (a-Lasso) など。

Q. 関数ライブラリや最適化手法をどう選択すればよいか？

A. 専門家が設計する。

→ **大規模言語モデルの知識**を活用できるのでは？

二つのモデル構築における専門知識

- AutoPMoBとSINDyでは、出発点と探索する対象が異なる。
- 知識が必要な工程を特定し、LLMの利用方法を考える。

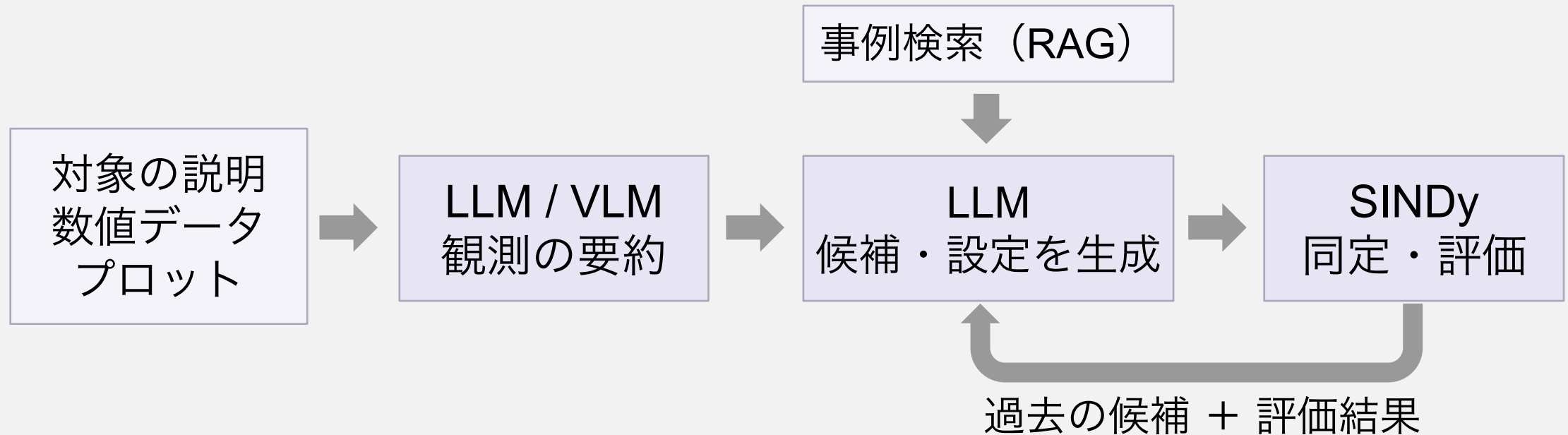
	AutoPMoB	SINDy
出発点	文献中の数式・変数定義	数値データ・候補関数
探索対象	数式の組合せ	方程式の項と係数
専門知識	変数の意味・数式と同義性	変数・候補関数・物理制約
残る検証	数值的に解けるか・対象に合うか	対象の挙動を表すか

LLMの利用目的を、モデル構築の具体的な工程で定める。

[Kato and Kano, Comput. Chem. Eng., 2025; Sato, Miyamoto, and Kato, J. Rheol., 2025]

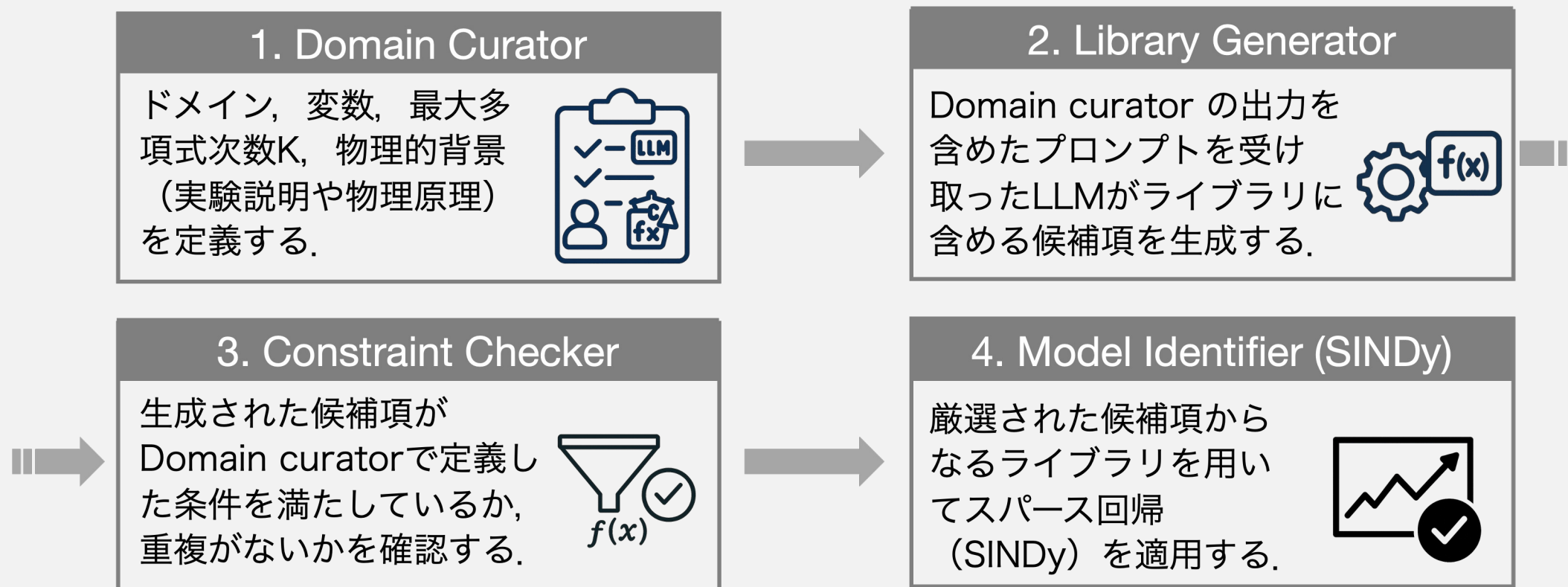
関連研究：Al-Khwarizmi

- 対象の説明・数値データ・プロットを、LLMとVLM (vision language model) で説明する。
- 事例検索と過去の試行を使い、ライブラリと最適化器を選ぶ。



LLMがSINDyを適用するための設計を支援する研究

- 物理制約に着目して、LLMでライブラリに含める項を生成（選択）する。
- 物理制約をプロンプトで明示、後段で検証し、有用な項を選定する。

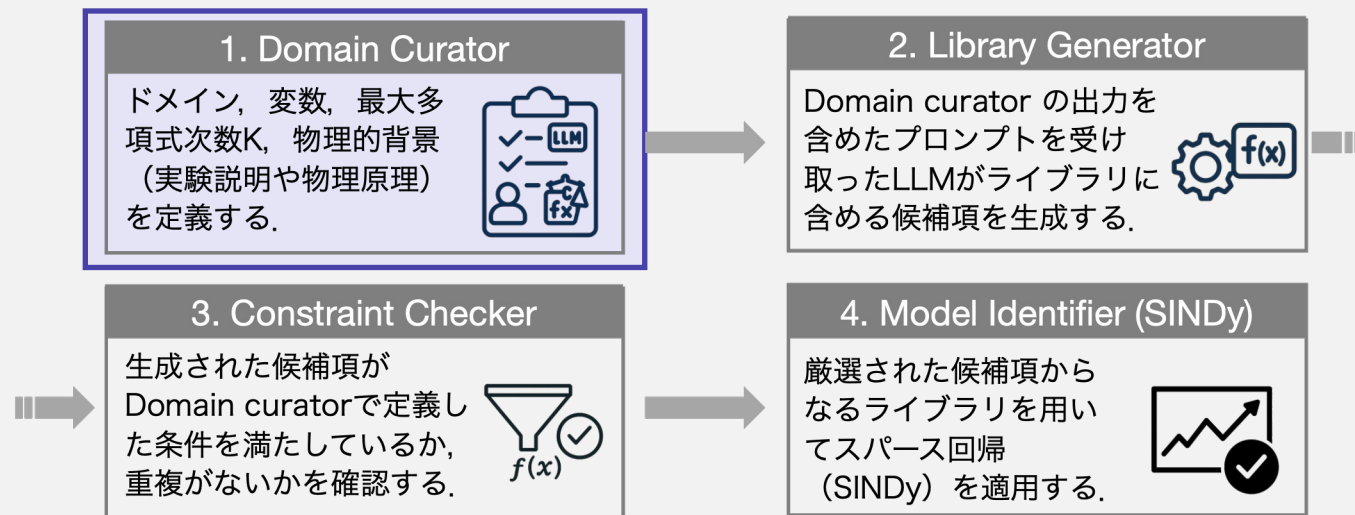


```
## Domain: {system type}
### State Variables
- {state variable}: {variable description}
- ...
### Input Variables
- {input variable}: {variable description}
- ...
### Constraints
- Maximum polynomial degree: {max degree}
- Allowed operators: {allowed operators}
### Experimental Background
{experimental background}
```

物理知識を3つの要素で構造化してプロンプトに明示する。

- Domain：対象系・実験条件
- Variables：状態変数・入力変数
- Constraints：物理的に満たすべき条件

レオロジー専門家の知見に基づいてプロンプトを設計したが、将来的には、文献情報の活用により自動化する計画。



- ...
Physical constraints for constitutive equations:
- Spatial symmetry sign-reversal symmetry): Under shear reversal ($\kappa_{xy} \rightarrow -\kappa_{xy}$), the constitutive equations must respect the corresponding symmetry of the physical system.
- As a consequence, even-order terms of the velocity gradient appearing alone (e.g., κ_{xy}^2 , κ_{xy}^4 without coupling to stress) are not physically admissible.
- The velocity gradient κ_{xy} may only appear in symmetry-consistent combinations, such as convective coupling terms with stress components.
...

You are an expert in scientific machine learning and sparse identification of nonlinear dynamics (SINDy). Your task is to generate candidate library terms for discovering governing equations from time-series data.

{output by Domain Curator}

Task

Generate a comprehensive list of candidate library terms that could appear in the governing equations for this system.

Guidelines

1. Include fundamental terms: Constants (1), linear terms, and polynomial products up to degree {max degree}
2. ...

Expected Output Format

Return a JSON object with: ...

Variables Available

{variable list}

Generate the candidate library terms now.

1. Domain Curator

ドメイン, 変数, 最大多項式次数K, 物理的背景 (実験説明や物理原理) を定義する.



2. Library Generator

Domain curator の出力を含めたプロンプトを受け取ったLLMがライブラリに含める候補項を生成する.



3. Constraint Checker

生成された候補項が Domain curator で定義した条件を満たしているか, 重複がないかを確認する.



4. Model Identifier (SINDy)

厳選された候補項からなるライブラリを用いてスパース回帰 (SINDy) を適用する.



Domain: {system type}

State Variables

- {state variable}: {variable description}

- ...

Input Variables

- {input variable}: {variable description}

- ...

Constraints

- Maximum polynomial degree: {max degree}

- Allowed operators: {allowed operators}

Experimental Background

{experimental background}

Constraint Checker と SINDy

3. Constraint Checker

- 物理制約はプロンプトで指示。

Checker は下記の決定的処理を行う。

① 次数制約の検証

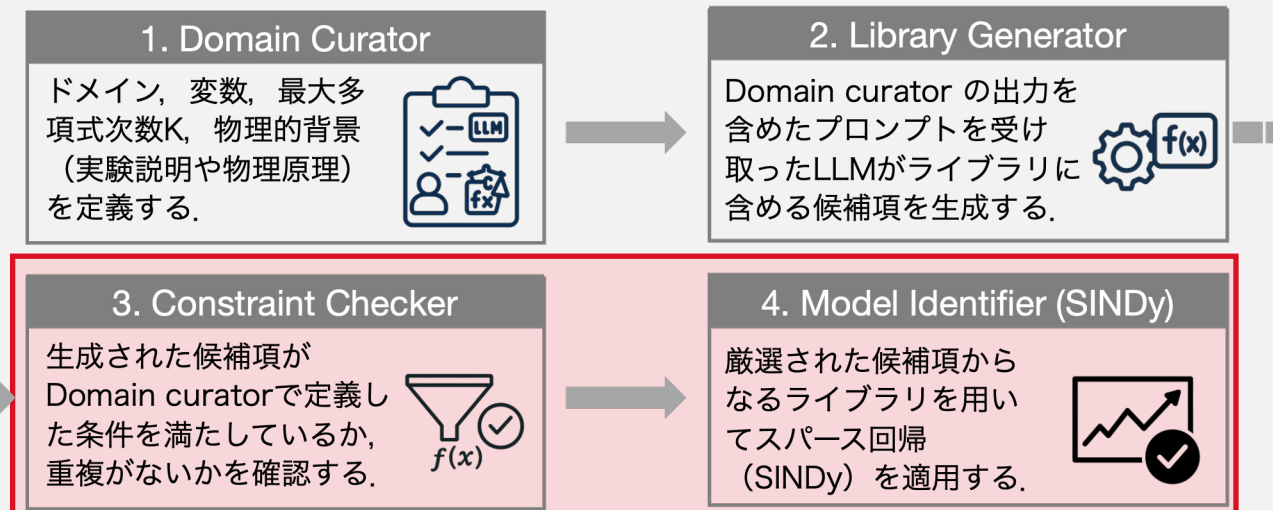
各候補項の多項式次数 $\leq K$ を確認。

② 積の可換性に基づく正規化

$\tau_{xy} \cdot \tau_{xx}$ と $\tau_{xx} \cdot \tau_{xy} \rightarrow$ 一意表現に変換。

③ 重複除去

同一候補が複数回生成されていたら統合。



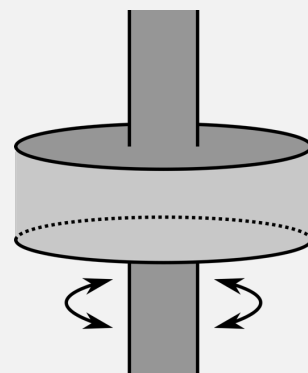
4. Model Identifier (SINDy)

$$\xi = \arg \min_{\xi} \|\dot{X} - \Theta \xi\|_2^2 + \lambda \|\xi\|_2^2$$

- 最適化器：STLSQ (sequentially thresholded least squares), $\lambda = 0.05$, $\theta = 0.1$
 - 閾値 θ 未満の係数を逐次的にゼロ化。変数の数が収束するまで反復する。

検証設定：Giesekusモデルによる数値データ

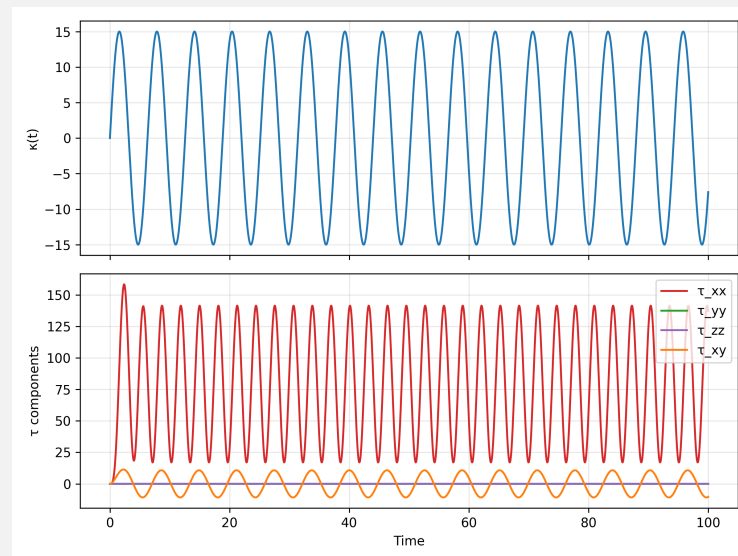
- Giesekus モデル を真のモデルとし、数値データから真のモデルを得られるかを検証。



$$\dot{\tau}_{xx} = -\tau_{xx} - \frac{1}{2}(\tau_{xx}^2 + \tau_{xy}^2) + 2\tau_{xy}\kappa_{xy}$$

$$\dot{\tau}_{yy} = -\tau_{yy} - \frac{1}{2}(\tau_{yy}^2 + \tau_{xy}^2) \quad \dot{\tau}_{zz} = 0$$

$$\dot{\tau}_{xy} = -\tau_{xy} + \frac{1}{2}(\tau_{xx} + \tau_{yy})\tau_{xy} + \tau_{yy}\kappa_{xy} + \kappa_{xy}$$



- 入力信号：振動せん断: $\kappa_{xy}(t) = \dot{\gamma}_0 \cos(\omega t)$ 、 $\omega = 0.5 \text{ rad/s}$, $\dot{\gamma}_0 = 2.0 \text{ s}^{-1}$
- 数値積分：前進Euler法 ($\Delta t = 10^{-4} \text{ s}$)、シミュレーション時間: $T = 100 \text{ s}$
- サンプリング： $\Delta t_s = 0.01 \text{ s}$ で等間隔抽出 ($N_{\text{full}} = 10,000$ 点)
- 評価指標：真の項集合とSINDyで得られた項集合に対する適合率 (Precision) ・ 再現率 (Recall) ・ F1スコアおよび、データに対するRMSE。

比較する3手法：ライブラリの作り方

- 物理制約をプロンプトで与える条件と、与えない条件を比較する。
- LLMを使わず、全多項式を列挙するbaselineも評価する。

手法	ライブラリの作成方法
PGL-SINDy	物理制約を与えてLLMで生成
one-shot	物理制約なしでLLMに生成を指示
baseline	LLMを使わず全多項式を列挙

専門知識の提示が、候補項と同定結果にどう影響するか。

- プロンプトに重要な項を明示すると、元の SINDy から性能向上。
 - 物理制約なしでの候補項生成 (one-shot) は元の SINDy と性能が同じ。
 - 重要な項を提示しないと、物理的推論に基づいて項を削除できず。
 - 今回のケースの場合、専門知識が不足していた可能性がある。

データ数 N (Full=10,000, 500, 200) とノイズの大きさ (SNR=20, 25dB) を変化させた結果

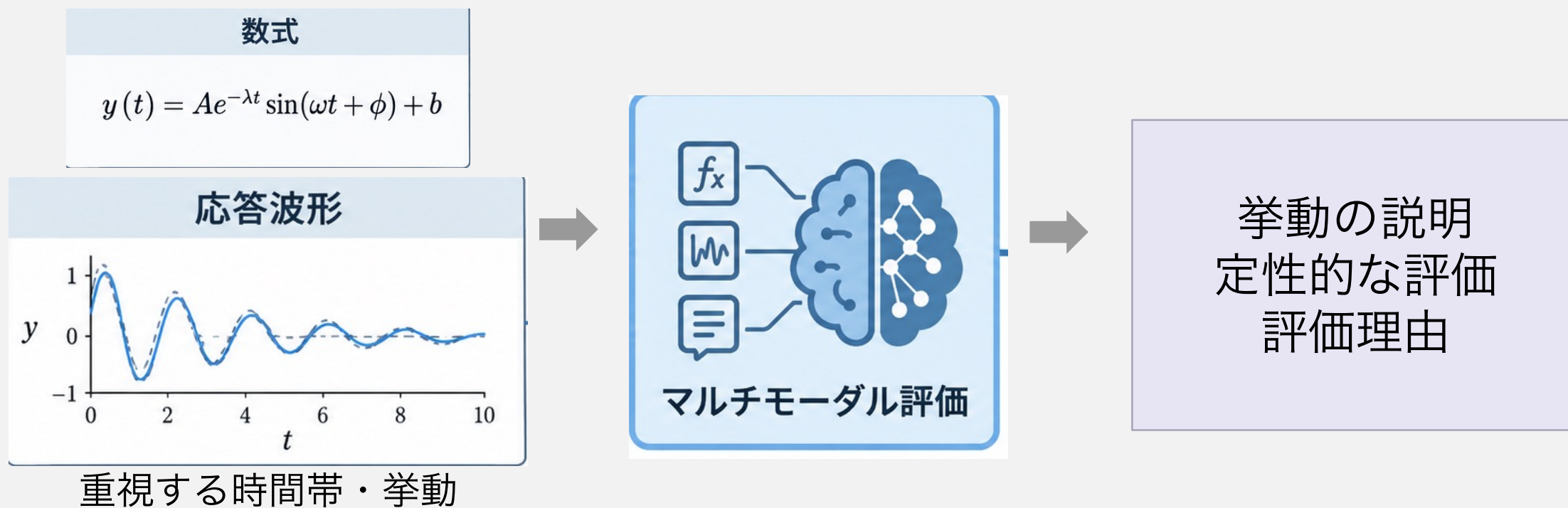
条件	F1 スコア			RMSE		
	提案手法	one-shot	baseline	提案手法	one-shot	baseline
Full	0.846	0.720	0.720	1.66×10^{-4}	3.28×10^{-4}	3.28×10^{-4}
$N=500$	0.615	0.455	0.455	2.68×10^{-3}	1.03×10^{-2}	1.03×10^{-2}
$N=200$	0.615	0.455	0.455	2.68×10^{-3}	1.03×10^{-2}	1.03×10^{-2}
SNR=20dB	0.800	0.720	0.720	8.47×10^{-3}	8.47×10^{-3}	8.47×10^{-3}
$N=500$, SNR=25dB	0.615	0.522	0.522	5.43×10^{-3}	1.15×10^{-2}	1.15×10^{-2}

LLM を用いた SINDy 効率化まとめ

- 専門家依存の SINDy の関数ライブラリ構築において、LLM が有用かを検証した。
- ドメイン専門家が設計した物理制約プロンプトを用いた LLM、ライブラリ構築のみを指示した LLM、多項式ライブラリを用いた SINDy の性能を比較した結果、
 - プロンプトに重要な項を明示することで性能が向上した。
 - 高性能な LLM を用いても、物理的推論は困難だった。
- 多様なプロセスでの検証、異なる LLM・プロンプトを用いた実験などは、今後の課題。

構想：マルチモーダル情報によるモデル評価

- 数値誤差・複雑度に加え、プロット・数式・自然言語も使う。
- 専門家が重視する挙動を、評価理由として扱う。



数値誤差と、画像・言語に基づく判断を組み合わせる。

構想：自然言語の実験条件を含む最適化

- 数値に加えて自然言語で示す条件もパラメータとして利用する。
- まずは、**指定可能な実験条件が限定**されているときに、次に行うべき実験条件を提示する方法を開発する。

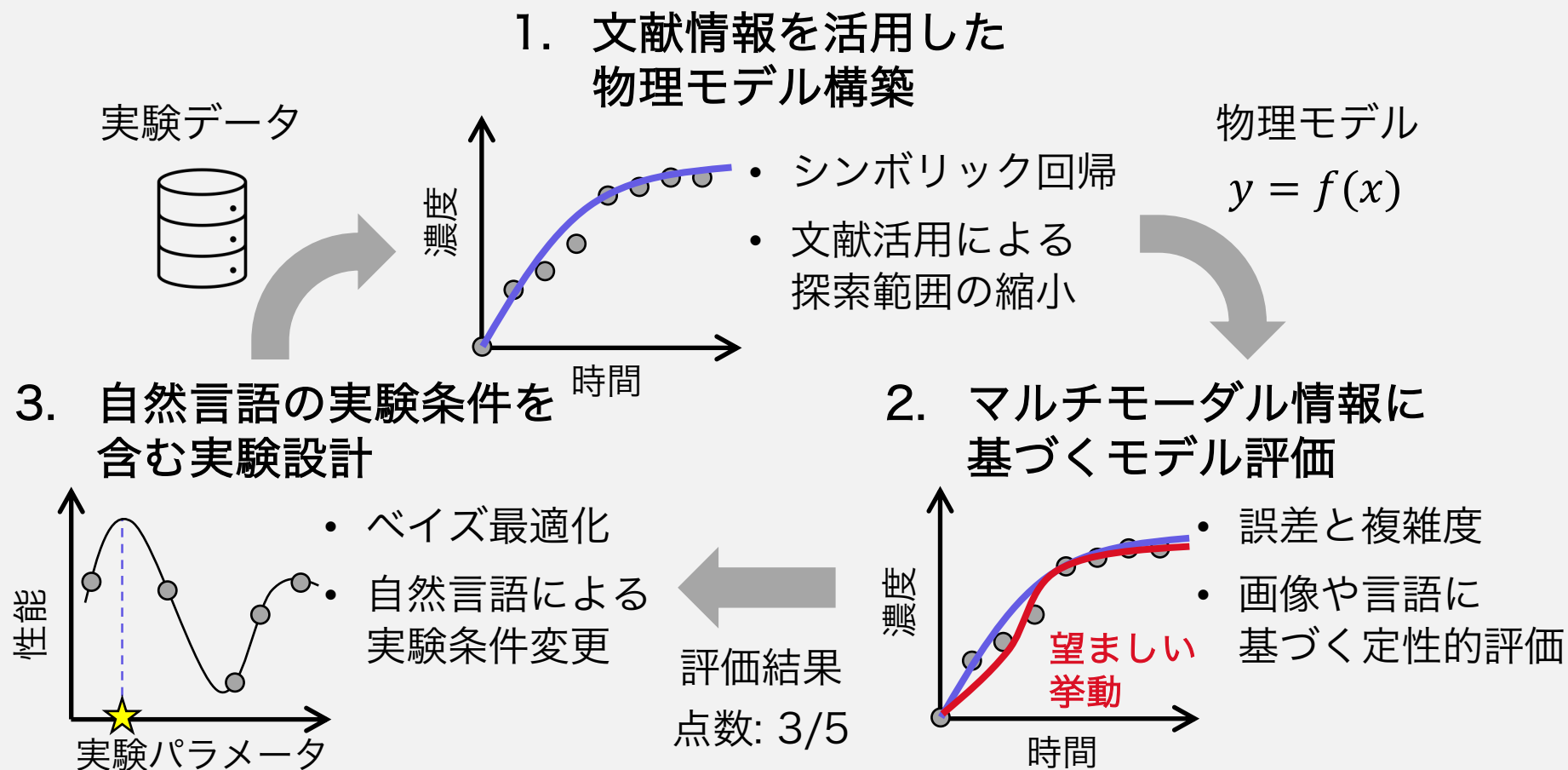
使用・変更可能な実験条件

- 従来：パラメータ（種類、値域などを含む）
- 本研究：パラメータと、パラメータに含まれない実験条件（測定条件、変数の増減など）



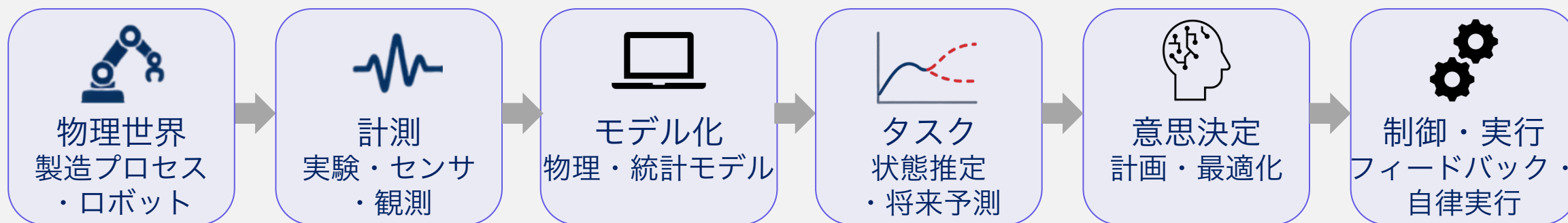
全体構想：モデル構築・評価・実験設計の接続

- 少数データ下でも物理モデルを発見するシステムの開発を目指す。



まとめ：文献・データから、使えるモデルへ

- 文献情報からモデル候補を構築し、数値データから方程式を探索。
- 物理知識とモデル評価を用いて、追加実験とモデルの更新に接続。
- 目的に応じて、文献や数値データ、知識といった情報と、AI・計算技術などのツールを活用する手法を開発しています。



活用するツール



LLM



VLM



シンボリック回帰



RAG/エージェント



基盤モデル

本発表に含まれる研究成果は、JST ACT-X JPMJAX23C5、さがけJPMJPR25T1、JSPS 科研費 JP21K18849、JP26K17699、公益財団法人高橋産業経済研究財団の助成を受けて実施したものです。