核融合科学研究所　第11回 研究部セミナー

機械学習の材料科学への応用研究の紹介

草場 穫

核融合科学研究所　学際連携センター　特任助教

目次

1. 自己紹介 - 5 min

2. 機械学習を用いた元素置き換えによる結晶構造予測 - 15 min

3. カーネル平均埋め込みによる材料の表現 - 15 min

4. 関数データのためのベイズカーネル回帰 - 15 min

5. まとめと質疑応答 - 10 min

1. 自己紹介

researchmap

↑統計数理研究所 (東京都　立川市)

経歴

2022年3月:　総合研究大学院大学　統計科学専攻
博士課程修了（統計科学）

~2024年8月:　統計科学研究所
ものづくりデータ科学研究センター　特任研究員

現職

核融合科学研究所　学際連携センター　特任助教

専門

統計科学、機械学習、
マテリアルズインフォマティクス (MI)

2. 機械学習を用いた元素置き換えによる結晶構造予測

2-1. 研究概要

- 結晶構造データベースを元に、機械学習による高速な結晶構造予測手法 (CSPML)を開発した。

- 約2万個の化合物に対する結晶構造予測の数値実験から、結晶系全体の約50~60%が提案手法によって予測可能であると推察された。

論文: Kusaba, Minoru, Chang Liu, and Ryo Yoshida. "Crystal structure prediction with machine learning-based element substitution." Computational Materials Science 211 (2022): 111496.
コード: https://github.com/Minoru938/CSPML

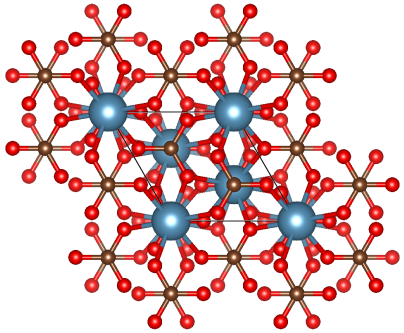
2-2. 結晶構造予測とは？

クエリ組成

$CaCO_3$

予測

$CaCO_3$の最安定構造



- 与えられた化学組成の安定構造を予測する問題である。
- 結晶構造はエネルギーが低いほど、その構造は安定していると言える。
- ある結晶構造のエネルギーはDFT計算によって算出できる。

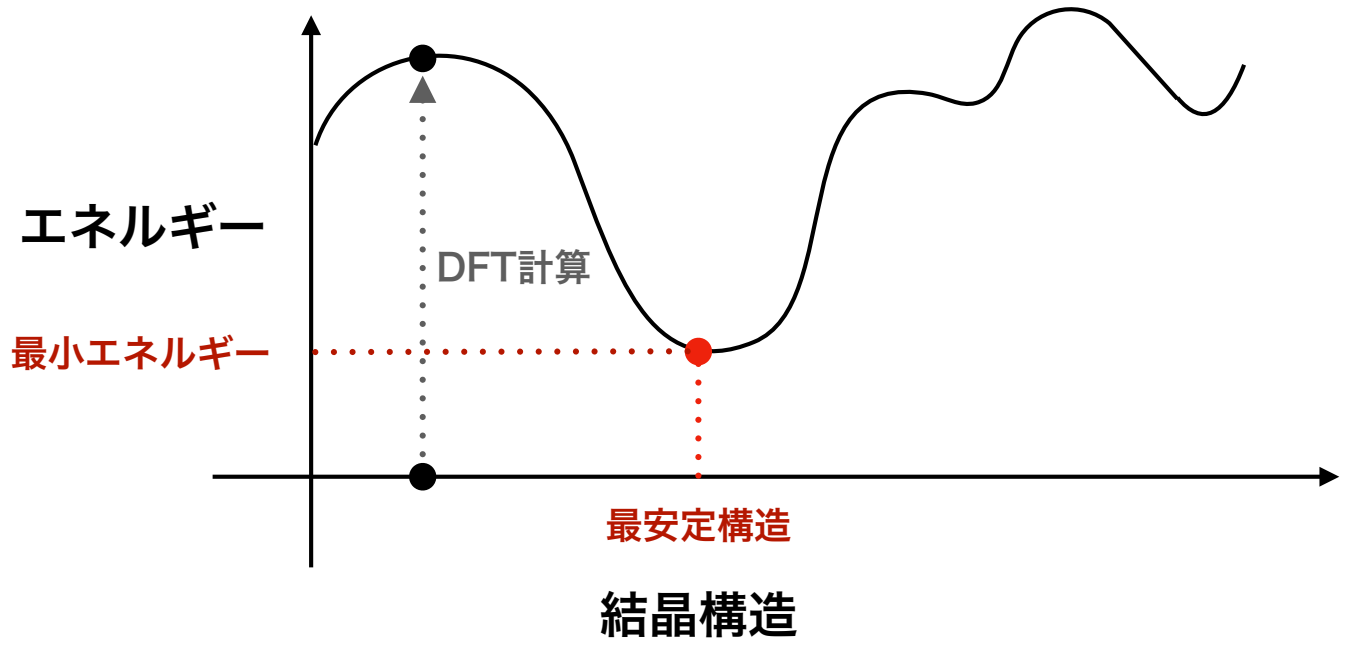
2-3. これまでの結晶構造予測研究

・結晶構造予測は最安定構造を探す問題

・エネルギーが低いほど安定

・ブラックボックス関数の最適化問題

ある系: X_AY_B



主な手法

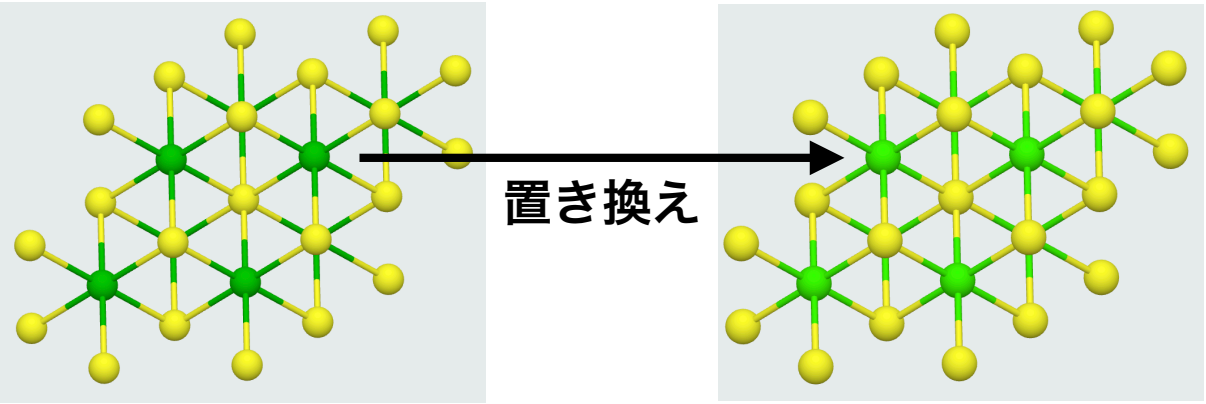
・USPEX (Glass et al. 2006) : 遺伝的アルゴリズム
・CALYPSO (Wang et al. 2012) : Particle Swarm
・CrySPY (Yamashita et al. 2018) : ベイズ最適化
・LAQA (Terayama et al. 2018) : 2次元近似

いずれも多くの
構造最適化を必要とする

結晶構造データを生かした手法が開発できないか?

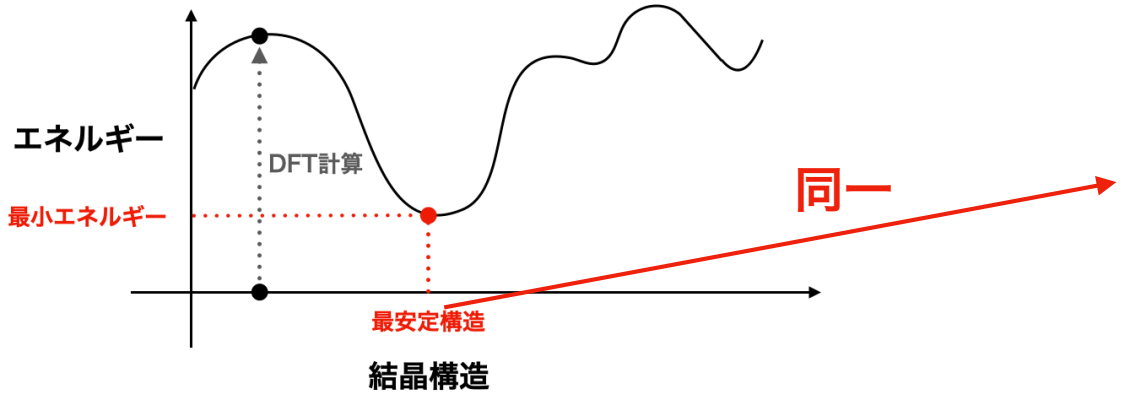
2-4. 結晶構造予測問題に対する別のアプローチ

伝統的な手法：



化学組成：　A_xB_y ⟶ C_xB_y

構造：エネルギー

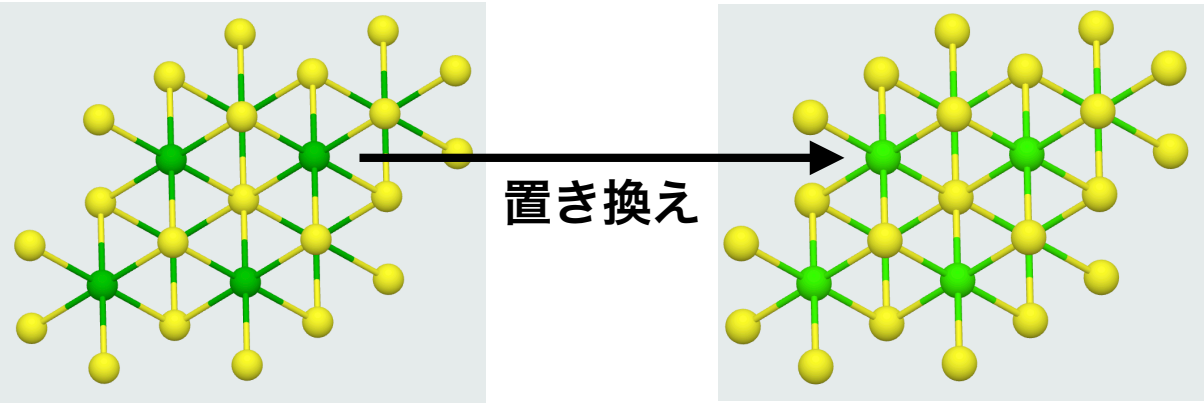


同一

安定構造

2-4. 結晶構造予測問題に対する別のアプローチ

機械学習の応用：



- 近年元素ペアの置き換え可能性を機械学習によってモデル化した元素置き換えベースの結晶構造予測手法が提案されている。

➡構造に変化を与えない元素ペアの共起情報に基づいてモデルを推定

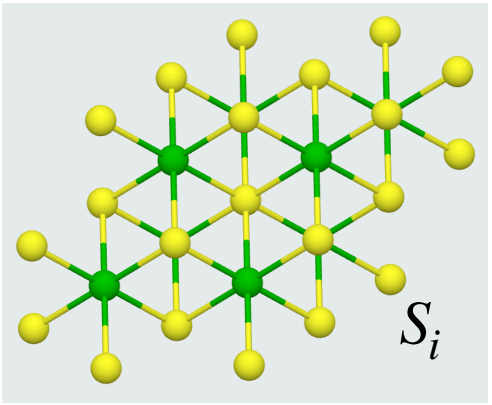
- 負例のデータ、元素レベルの特徴量が未活用
- 単一元素の置き換えのみに対応しており、適用範囲が狭い

出典：Hautier, G., Fischer, C., Ehrlacher, V., Jain, A. & Ceder, G. Data Mined Ionic Substitutions for the Discovery of New Compounds. *Inorg. Chem.*, 50, 656-663 (2011).

Wang, H. C., Botti, S., & Marques, M. A. Predicting stable crystalline compounds using chemical similarity. *npj Computational Materials*, *7*(1), 1-9 (2021).

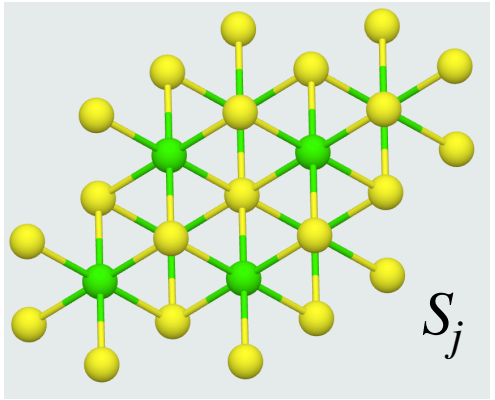
2-5. 提案手法の概要

組成：C_i



類似な結晶構造

組成： C_j



一意的に決定

化合物1： 組成A C_i → 最安定構造A S_i

同じか否か？

化合物2： 組成B C_j → 最安定構造B S_j

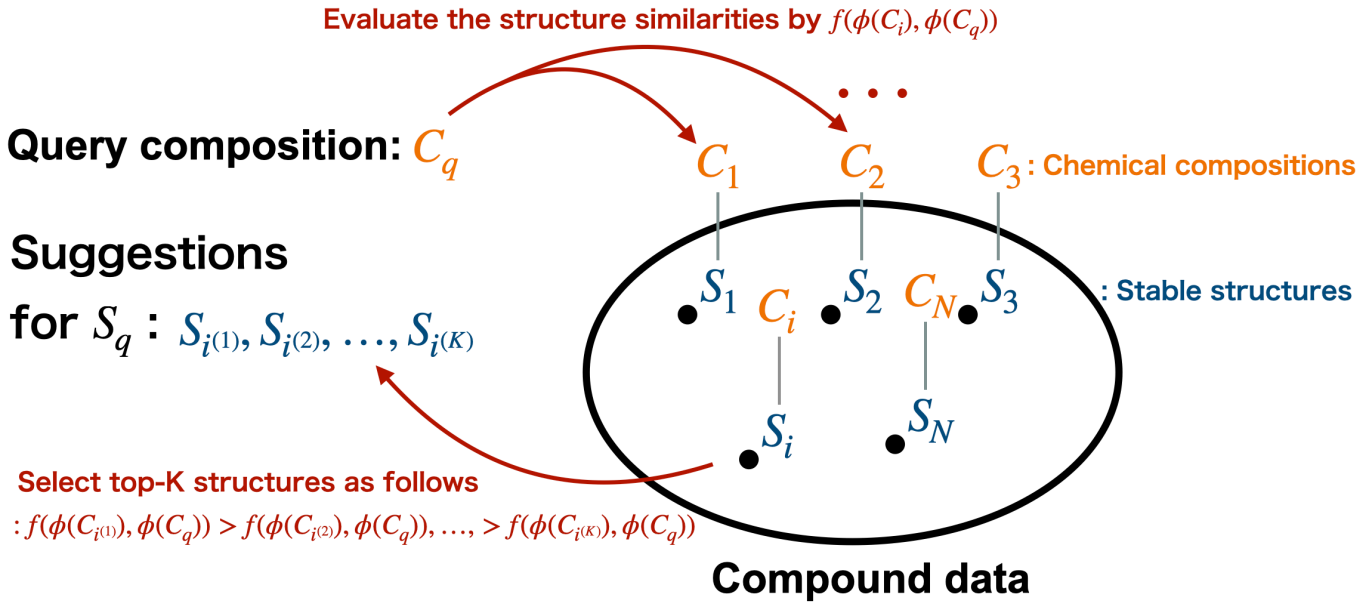
化学組成の記述子：$\phi(C_i)$

1 0

$$Y = f(\phi(C_i), \phi(C_j)); Y \in \{simmilar, dissimilar\}$$

➡上記の f をデータから学習する

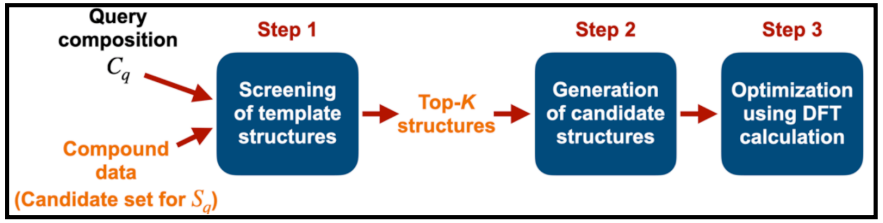
2-5. 提案手法の概要



クエリ組成： C_q

構造既知の化学組成： C_i

化学組成の記述子：$\phi(C_i)$



$$Y = f(\phi(C_q), \phi(C_i))$$

↓

$Y = 1$ に近い順に上位 K 個のテンプレート

$S_{i(1)}, S_{i(2)}, \ldots, S_{i(K)}$ $(C_{i(1)}, C_{i(2)}, \ldots, C_{i(K)})$ を選択する。

↓

クエリ組成と同じになるように元素置き換え

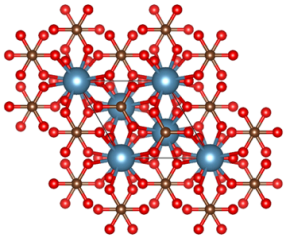
↓

DFT計算による検証

2-6. 提案手法の適応例

True structure

$CaCO_3$ (target)

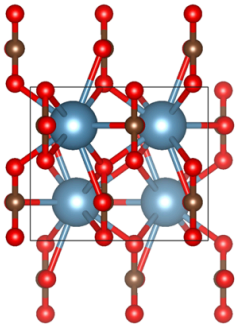


id: mp-3953

Predicted structures

$CaCO_3$ (top 1)

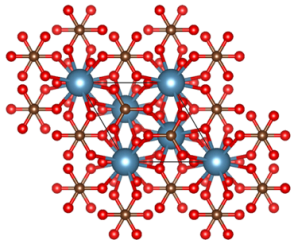
Dissimilarity: 1.825



Template structure:
$YbCO_3$, mp-755213

$CaCO_3$ (top 2)

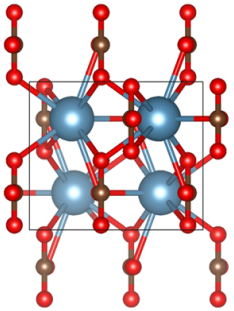
Dissimilarity: 0.077



Template structure:
$NaCO_3$, mp-1120755

$CaCO_3$ (top 3)

Dissimilarity: 1.878



Template structure:
$EuCO_3$, mp-554518

(補足) モデル f の詳細

⬇XenonPy記述子 (290次元)

化学組成ペアの数量化： $(\phi(C_i), \phi(C_j)) = |\phi_{xenonpy}(C_i) - \phi_{xenonpy}(C_j)|$

➡二値分類のNNを適用

NNの構造：

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 290)	0
dense_1 (Dense)	(None, 50)	14550
dropout_1 (Dropout)	(None, 50)	0
dense_2 (Dense)	(None, 50)	2550
dropout_2 (Dropout)	(None, 50)	0
dense_3 (Dense)	(None, 50)	2550
dense_4 (Dense)	(None, 2)	102

Total params: 19,752

適用手法：NNによる二値分類↓

$$Y = f_{NN}(|\phi_{xenonpy}(C_i) - \phi_{xenonpy}(C_j)|); Y \in \{simmilar, dissimilar\}$$

2-7. ベンチマークセットに対する予測実験

- 38個のクエリ組成をベンチマークセットとして選択し、CSPMLによる結晶構造予測を行った。

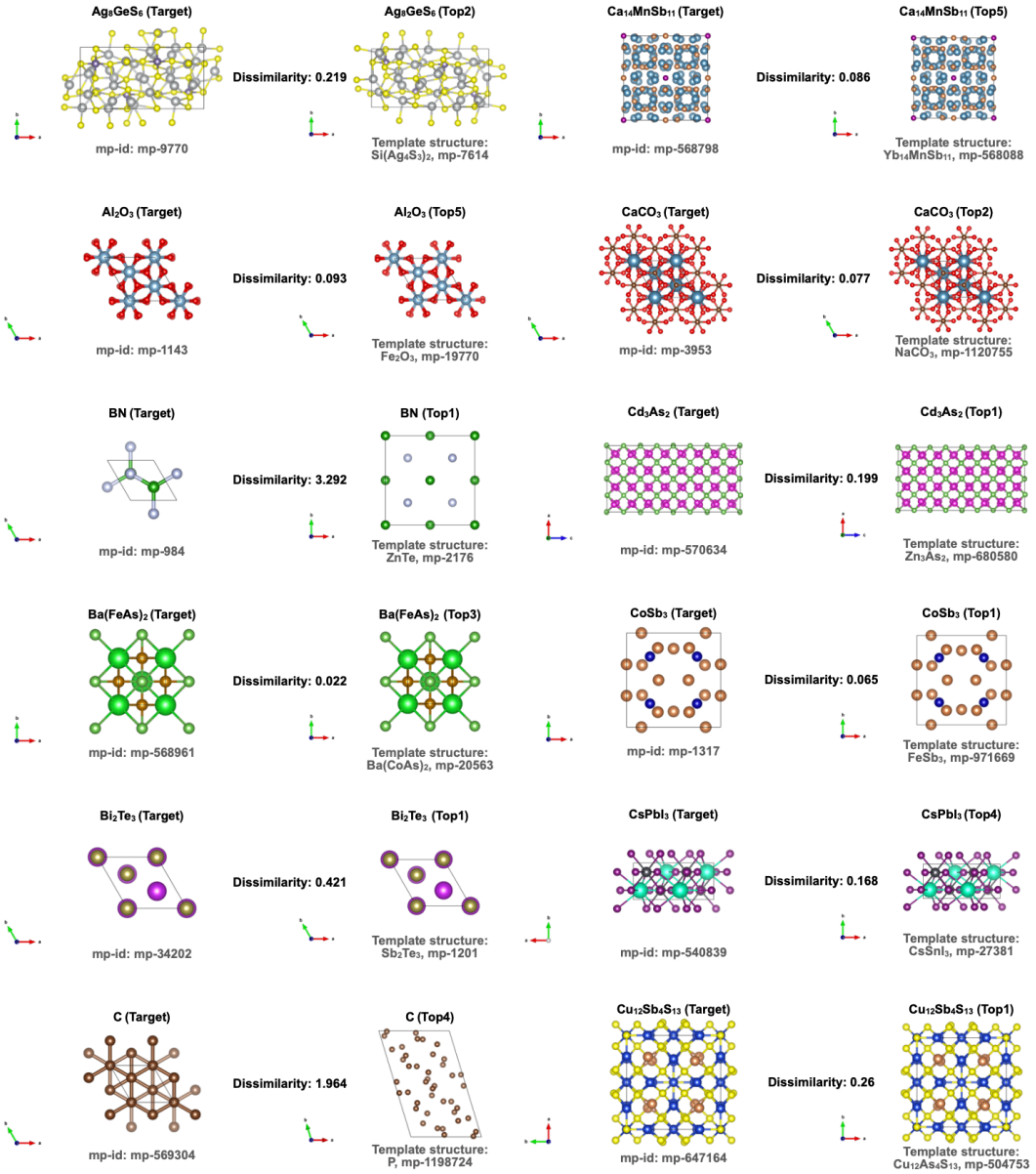
- この実験では、上位5個までのテンプレート構造を提案し、それら全てをDFT計算によって構造緩和した。

The Materials Project

モデルは、Materials Project* (MP)中の安定化合物データを使って訓練された。テンプレ構造としてはMP中の全安定化合物（約３万）を使用した。

[*] Jain, Anubhav, et al. "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation." APL materials 1.1 (2013).

2-7. ベンチマークセットに対する予測実験（結果）



The structures are depicted by VESTA: K. Momma and F. Izumi, "VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data," *J. Appl. Crystallogr.*, 44, 1272-1276 (2011).

2-7. ベンチマークセットに対する予測実験（表）

Composition	Min. dissimilarity of all candidates	Min. dissimilarity of top 5	Rank	Prediction success
Ag_8GeS_6	0.214	0.214	1/34	–
Al_2O_3	0.067	0.093	2/297	✓
BN	1.726	3.292	683/960	–
$Ba(FeAs)_2$	0.091	0.176	9/1424	✓
Bi_2Te_3	0.293	0.293	1/297	✓
C	1.769	1.975	3/87	–
$Ca_{14}MnSb_{11}$	0.083	0.096	2/13	✓
$CaCO_3$	0.054	0.077	3/1000	✓
Cd_3As_2	0.19	0.19	1/297	✓
$CoSb_3$	0.068	0.068	1/1042	✓
$CsPbI_3$	0.129	0.129	1/1000	✓
$Cu_{12}Sb_4S_{13}$	0.24	0.24	1/1	✓
Fe_3O_4	0.216	0.216	1/152	–
GaAs	0	0	1/960	✓
GeH_4	0.383	0.639	22/171	–
La_2CuO_4	0.022	0.022	1/821	✓
Li_3PS_4	0.851	1.216	33/250	–
$Li_4Ti_5O_{12}$	0.282	0.282	1/8	–
$LiBF_4$	0.302	0.592	6/983	–
$LiCoO_2$	0.199	0.207	5/3895	–
$LiFePO_4$	0.113	0.13	2/327	✓
$LiPF_6$	0.046	0.297	6/242	✓
$Mn(FeO_2)_2$	0.022	0.022	1/821	✓
Si	0	2.304	7/87	–
Si_3N_4	0.269	0.269	1/152	–
SiO_2	0.167	0.167	1/1151	–
$SrTiO_3$	0.395	0.643	16/1000	✓
TiO_2	1.015	1.401	20/1151	–
V_2O_5	0.753	1.865	41/85	–
VO_2	0.077	0.077	1/1151	✓
$Y_3Al_5O_{12}$	0.014	0.014	1/49	✓
ZnO	0.006	0.062	5/960	✓
ZnSb	0.316	0.316	1/960	✓
ZrO_2	0.131	0.131	1/1151	✓
$ZrTe_5$	0.039	0.039	1/132	✓

2-7. ベンチマークセットに対する予測実験（総括）

Composition	Min. dissimilarity of all candidates	Min. dissimilarity of top 5	Rank	Prediction success
Ag_8GeS_6	0.214	0.214	1/34	–
Al_2O_3	0.067	0.093	2/297	✓
BN	1.726	3.292	683/960	–
$Ba(FeAs)_2$	0.091	0.176	9/1424	✓
Bi_2Te_3	0.293	0.293	1/297	✓
C	1.769	1.975	3/87	–
$Ca_{14}MnSb_{11}$	0.083	0.096	2/13	✓
$CaCO_3$	0.054	0.077	3/1000	✓
Cd_3As_2	0.19	0.19	1/297	✓
$CoSb_3$	0.068	0.068	1/1042	✓
$CsPbI_3$	0.129	0.129	1/1000	✓
$Cu_{12}Sb_4S_{13}$	0.24	0.24	1/1	✓
Fe_3O_4	0.216	0.216	1/152	–
GaAs	0	0	1/960	✓
GeH_4	0.383	0.639	22/171	–
La_2CuO_4	0.022	0.022	1/821	✓
Li_3PS_4	0.851	1.216	33/250	–
$Li_4Ti_5O_{12}$	0.282	0.282	1/8	–
$LiBF_4$	0.302	0.592	6/983	–
$LiCoO_2$	0.199	0.207	5/3895	–
$LiFePO_4$	0.113	0.13	2/327	✓
$LiPF_6$	0.046	0.297	6/242	✓
$Mn(FeO_2)_2$	0.022	0.022	1/821	✓
Si	0	2.304	7/87	–
Si_3N_4	0.269	0.269	1/152	–
SiO_2	0.167	0.167	1/1151	–
$SrTiO_3$	0.395	0.643	16/1000	✓
TiO_2	1.015	1.401	20/1151	–
V_2O_5	0.753	1.865	41/85	–
VO_2	0.077	0.077	1/1151	✓
$Y_3Al_5O_{12}$	0.014	0.014	1/49	✓
ZnO	0.006	0.062	5/960	✓
ZnSb	0.316	0.316	1/960	✓
ZrO_2	0.131	0.131	1/1151	✓
$ZrTe_5$	0.039	0.039	1/132	✓

- 38個中以下3個のクエリ組成は構造が予測されなかった。$NaCaAlPHO_5F_2$は組成比が一致する化合物が全候補中に無かった。MgB_7と$Ba_2CaSi_4(BO_7)_2$は類似クラスに分類される確率が0.5を超えるものが全候補中に無かった。
- 60% (21/35)の化合物の結晶構造予測が成功した。

・DFTによる構造緩和前の構造非類似度が 0.1以下の場合は100% (11/11)の確率で、構造緩和後の予測が成功しており、0.2以下の場合は94.1% (16/17)の確率で予測が成功していた。

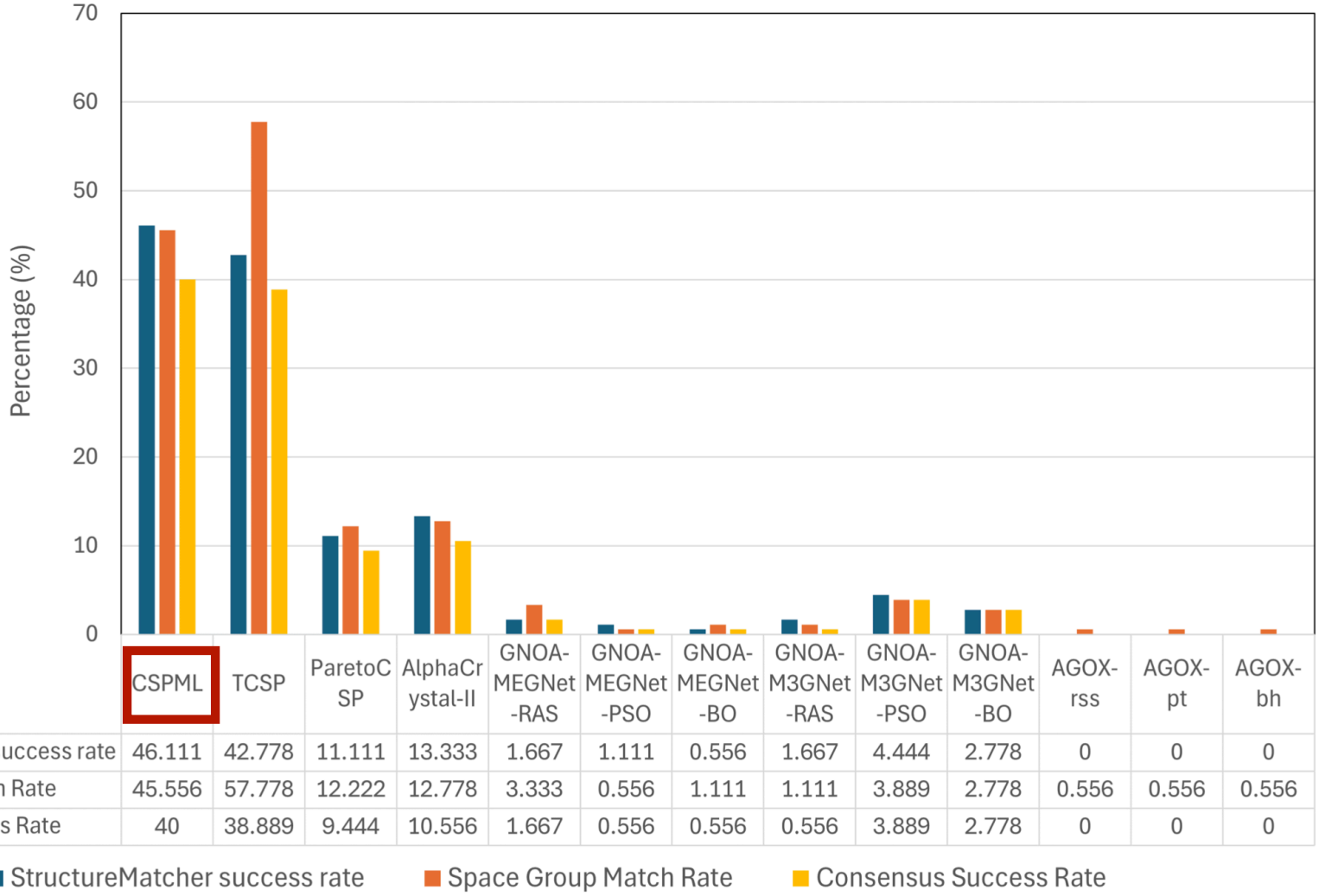
2-8. 提案手法の汎用的な予測能力の検証

- 提案手法の汎用的な予測能力を検証するために、MPからランダムに選択された 約2万個の化合物に対する結晶構造予測の数値実験を行った（構造緩和前）。

τ	All candidates	Top 1	Top 5	Top 10	Top 20	Top 30	Top 50
0.1	51.7%	31.7%	44.8%	47.7%	49.7%	50.5%	51.3%
0.2	68.1%	46.0%	60.7%	63.5%	65.3%	66.0%	66.8%
0.3	76.4%	55.5%	69.8%	72.4%	73.9%	74.5%	75.1%

- 38個のベンチマークセットの結果から、構造非類似度 0.1以下の場合は 100% (11/11) で構造緩和後の構造は同一であり、0.2以下の場合は 94.1% (16/17)で構造緩和後の構造は同一であった。
- よって、提案手法を用いて、結晶系全体の約 51.3% (51.3×1.0) ~ 62.8% (66.8×0.94)が予測可能であると推測される。

2-9. 第三者によるCSPMLの性能検証結果



	CSPML	TCSP	ParetoCSP	AlphaCrystal-II	GNOA-MEGNet-RAS	GNOA-MEGNet-PSO	GNOA-MEGNet-BO	GNOA-M3GNet-RAS	GNOA-M3GNet-PSO	GNOA-M3GNet-BO	AGOX-rss	AGOX-pt	AGOX-bh
StructureMatcher success rate	46.111	42.778	11.111	13.333	1.667	1.111	0.556	1.667	4.444	2.778	0	0	0
Space Group Match Rate	45.556	57.778	12.222	12.778	3.333	0.556	1.111	1.111	3.889	2.778	0.556	0.556	0.556
Consensus Success Rate	40	38.889	9.444	10.556	1.667	0.556	0.556	0.556	3.889	2.778	0	0	0

出典：Wei, Lai, et al. "CSPBench: a benchmark and critical evaluation of Crystal Structure Prediction." arXiv preprint arXiv:2407.00733 (2024).

2-10. 結論

研究結果：

- 構造類似性の予測に基づいた結晶構造予測手法を提案した。
- 結晶構造予測のベンチマークセットを提案した論文では、数ある手法の中で最高性能をマークしたことが報告された。

コード：https://github.com/Minoru938/CSPML

手法の特徴：

- 本手法は検証パートを除いて、DFT計算を必要としないので、高速な予測を提供する。一方、新規の結晶構造を予測できないという欠点がある。
- 本手法は結晶構造データベースが拡充するほど強力となる性質を持つ。

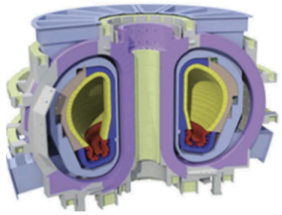
発展研究：

- 準安定構造の予測や構造→組成方向の予測にも対応したより柔軟なCSPMLモデルを現在研究中である。

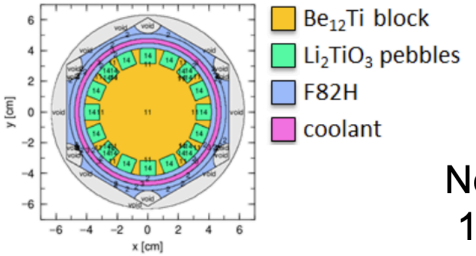
2-11. CSPMLの活用例

NIFS 向井啓祐先生

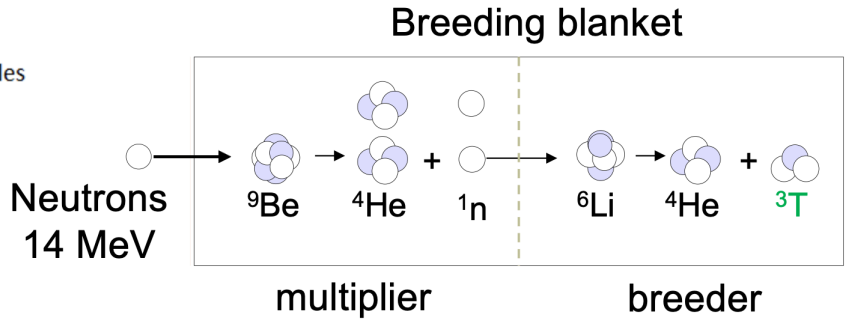
LOCA accident scenario for WCCB blanket



JA-DEMO[1]



Calculation model



- Water-Coled Ceramic Breeding (WCCB) blanket employs ceramic breeder (CB) pebbles and $Be_{12}X$ beyllide blocks
- In-vessel loss of coolant accident (LOCA) is one of the severe accident scenarios [2]

  $Be + H_2O\ (steam) \rightarrow BeO + H_2$

- $Be_{12}X$ beryllides has a good oxidation resistance which can significantly reduce H_2 production [3]
  - This work focuses hybrid ceramic materials to further decrease the H_2 production in LOCA

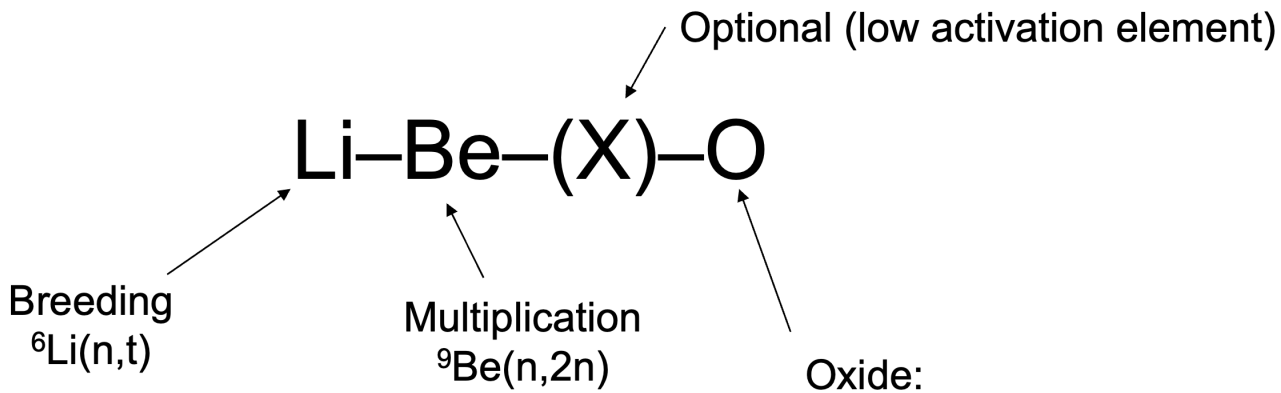
*向井啓祐先生から提供された資料を使用

2-11. CSPMLの活用例

NIFS 向井啓祐先生

Hybrid ceramics

This work focus on Li–Be hybrid ceramics because...



Oxide:

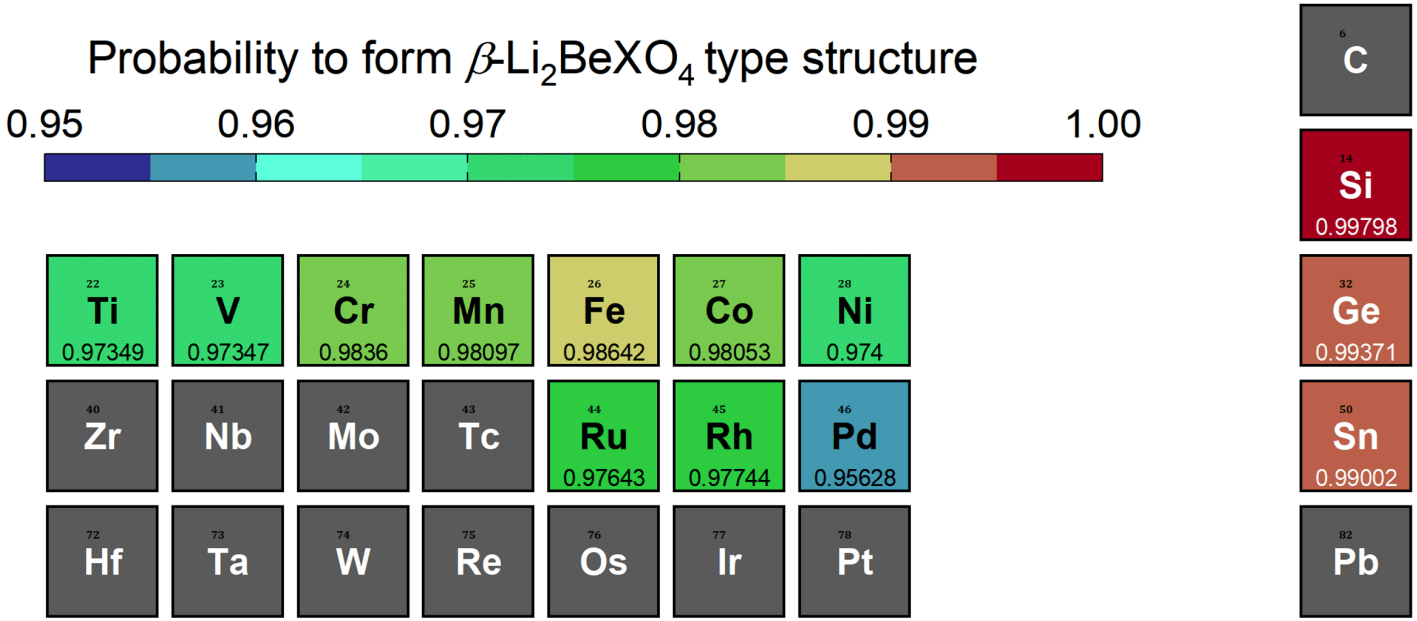
- High melting point
- Possibly significantly reduce H_2 production in LOCA
  (i.e. accident torrelance)

*向井啓祐先生から提供された資料を使用

2-11. CSPMLの活用例

NIFS 向井啓祐先生

CSPML: Li2BeXO4 to form beta Li2BeSiO4 type structure



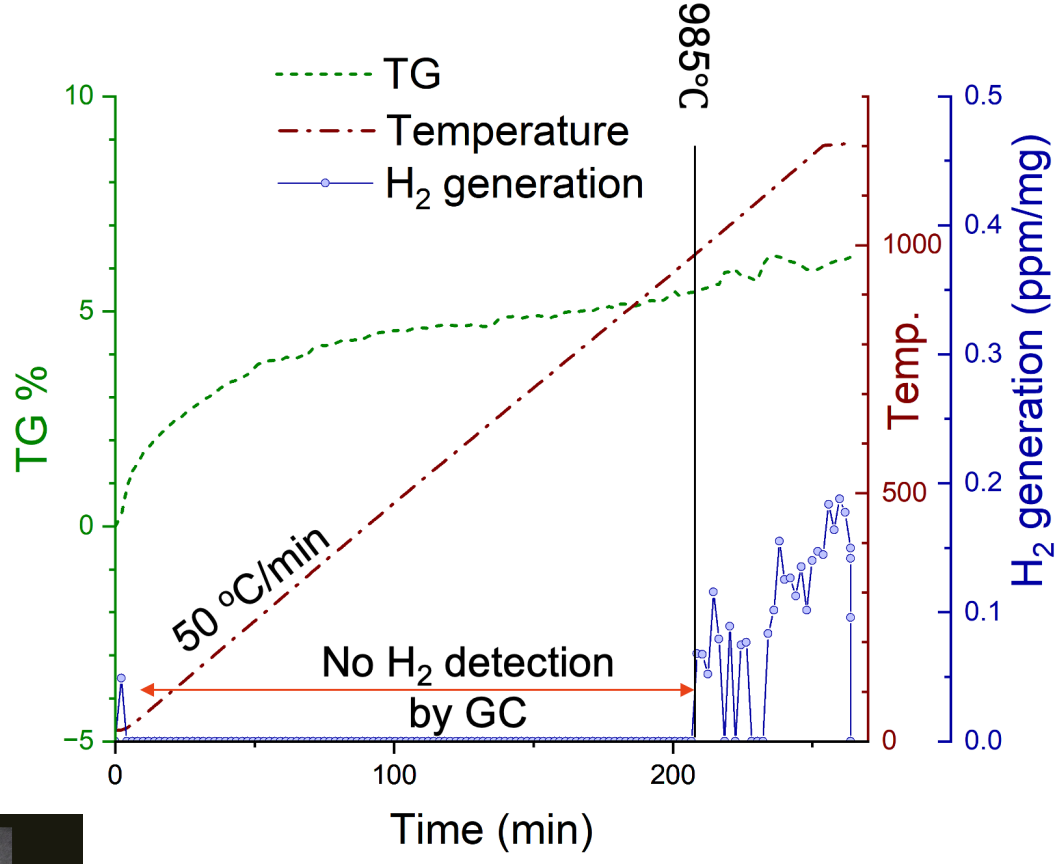
*向井啓祐先生から提供された資料を使用

2-11. CSPMLの活用例

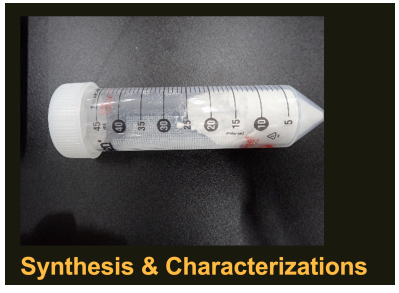
NIFS 向井啓祐先生

Initial result of steam exposure test for Li_2BeSiO_4

Thermogravimetry (TG) + gas chromatography (GC) at up to 1200 °C in Ar-1%H_2O (R.H.) gas flow condition



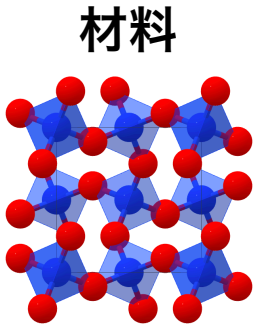
- H_2 generation was not observed T < 985 °C
- Total H_2 production above 985 °C was 3.1 ppm/mg
- Following test for $Be_{12}Ti$ will be conducted In the same condition



*向井啓祐先生から提供された資料を使用

3. カーネル平均埋め込みによる材料の表現

3-1. 背景と研究結果



材料 → ベクトル、グラフなど → 機械学習モデル → 物性

この部分の研究

既存手法の課題点

- 数量化の際に情報損失が起こる
- ベクトル→材料への逆写像が保証されない

研究結果:

- カーネル平均埋め込みに基づいた汎用記述子生成手法を提案した。
- 逆写像を一意に決定できることが保証されている。

論文: Kusaba, Minoru, et al. "Representation of materials by kernel mean embedding." Physical Review B 108.13 (2023): 134107.

コード: https://github.com/Minoru938/KmdPlus

3-2. 材料という情報の特徴

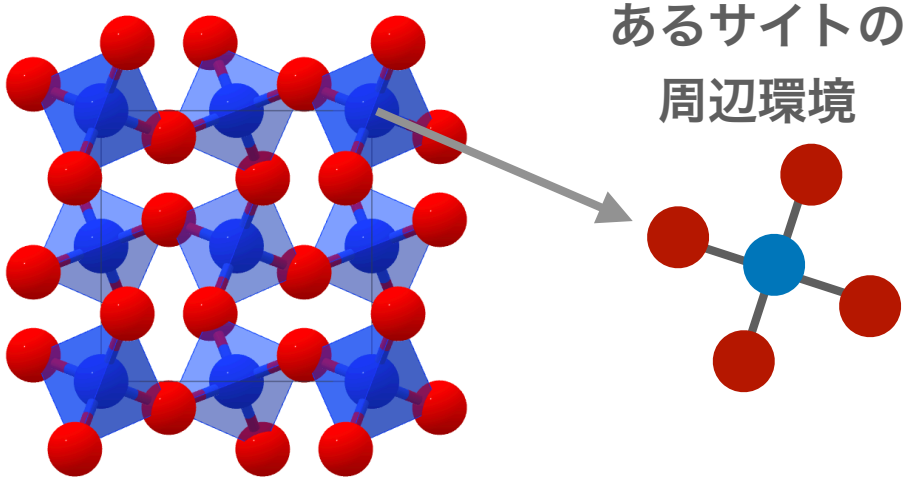
・化学組成

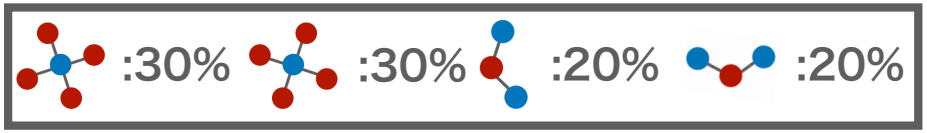
Na_2CO_3

構成元素: Na, C, O

組成比: 2 : 1 : 3

・結晶構造





・ポリマー (繰り返し構造)

~~-A-B-A-C-B-~~

構成要素: A, B, C

相対頻度: 2 : 2 : 1

・材料は、一般に複数の構成要素から構成され、個々の構成要素について特徴量が定義される。

・材料は情報の形式として、ベクトルでは無く確率分布として与えられることが一般的である。

3-3. 材料記述子生成の定式化

材料記述子生成は確率分布をできるだけ情報損失を少なく固定長ベクトルに変換する問題とみなすことができる。

問題設定

・材料 X が N_X 個の構成要素 $x_1, \ldots, x_{N_X}$ からなり、それぞれに重み $w_1, \ldots, w_{N_X}$ が与えられているとする。各構成要素は特徴量 $\lambda_i = \lambda(x_i) \in \mathbb{R}\ (i = 1,\ldots, N_X)$ によって特徴付けられている。

・この時一般に、材料 X の記述子 $\phi(X)$ の要素 $\phi_{(f,\lambda)}(X) \in \mathbb{R}$ は以下のように書ける。

$$\phi_{(f,\lambda)}(X) = f(w_1, \ldots, w_{N_X}, \lambda_1, \ldots, \lambda_{N_X}), \forall \lambda \in \Lambda, f \in F\,.$$

・構成要素数 N_X は材料 X によって異なる。また、記述子 $\phi(X)$ は $x_1, \ldots, x_{N_X}$ の順序交換に対して不変であるべきである。

3-4. これまでの材料記述子

以上のように、要約関数 f は**可変長の混合物系**を扱え、**順序交換に対する不変性**を持つ必要がある。よって、MIの従来研究では f として加重平均、最大最小プーリングなどの**要約統計量**が一般に使われている。

要約統計量の例:

$$\phi_{(\mathrm{mean},\lambda)}(X) = \sum_{i=1}^{N_X} w_i \lambda_i,\ \phi_{(\mathrm{var},\lambda)}(X) = \sum_{i=1}^{N_X} w_i (\lambda_i - \phi_{(mean,\lambda)}(X))^2,$$

$$\phi_{(\mathrm{max},\lambda)}(X) = \min\{\lambda_1, \ldots \lambda_{N_X}\},\ \phi_{(\mathrm{min},\lambda)}(X) = \min\{\lambda_1, \ldots \lambda_{N_X}\}\,.$$

(例)化学組成: Na_2CO_3 → **要約統計量記述子: (8.667, 6, 103.667, 60)**

使用した特徴量 λ : 原子番号, 原子半径
使用した統計量 f : 加重平均 $\phi_{(\mathrm{mean},\lambda)}(X)$, 最小プーリング $\phi_{(\mathrm{min},\lambda)}(X)$

構成元素: Na, C, O
混合比: 1/3, 1/6, 1/2

原子番号: 11, 6, 8
- 加重平均: 11×1/3 + 6×1/6 + 8×1/2 = 8.667
- 最小プーリング: min{11, 6, 8} = 6

原子半径: 186, 70, 60 (pm)
- 加重平均: 186×1/3 + 70×1/6 + 60×1/2 = 103.667
- 最小プーリング: min{186, 70, 60} = 60

3-4. これまでの材料記述子

以上のように、要約関数 f は**可変長の混合物系**を扱え、**順序交換に対する不変性**を持つ必要がある。よって、MIの従来研究では f として加重平均、最大最小プーリングなどの**要約統計量**が一般に使われている。

要約統計量の例:

$$\phi_{(\mathrm{mean},\lambda)}(X) = \sum_{i=1}^{N_X} w_i \lambda_i,\ \phi_{(\mathrm{var},\lambda)}(X) = \sum_{i=1}^{N_X} w_i (\lambda_i - \phi_{(mean,\lambda)}(X))^2,$$

$$\phi_{(\mathrm{max},\lambda)}(X) = \min\{\lambda_1, \ldots \lambda_{N_X}\},\ \phi_{(\mathrm{min},\lambda)}(X) = \min\{\lambda_1, \ldots \lambda_{N_X}\}\,.$$

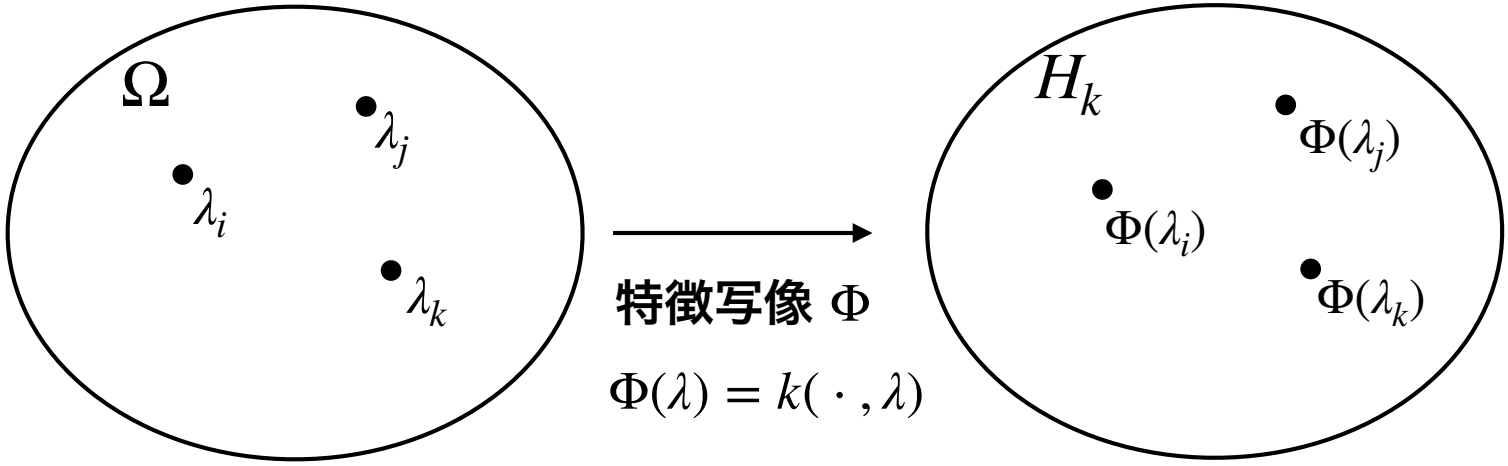
・材料記述子の生成は、**確率分布から固定長ベクトルへの変換**だと見なす事ができるが、要約統計量記述子では確率分布の多峰性など、**高次元モーメントの情報が変換過程で失われてしまう**。

➡機械学習理論であるカーネル平均埋め込みに基づいた、高次元モーメント情報も保存できる材料記述子のクラスを提案する。

3-5. カーネル平均埋め込み

確率変数 $\Lambda_X \sim P_X(\lambda)$

確率ベクトル $\Phi(\Lambda_X)$



元の空間

特徴空間(再性核ヒルベルト空間)

定義: $P_X(\lambda)$のカーネル平均

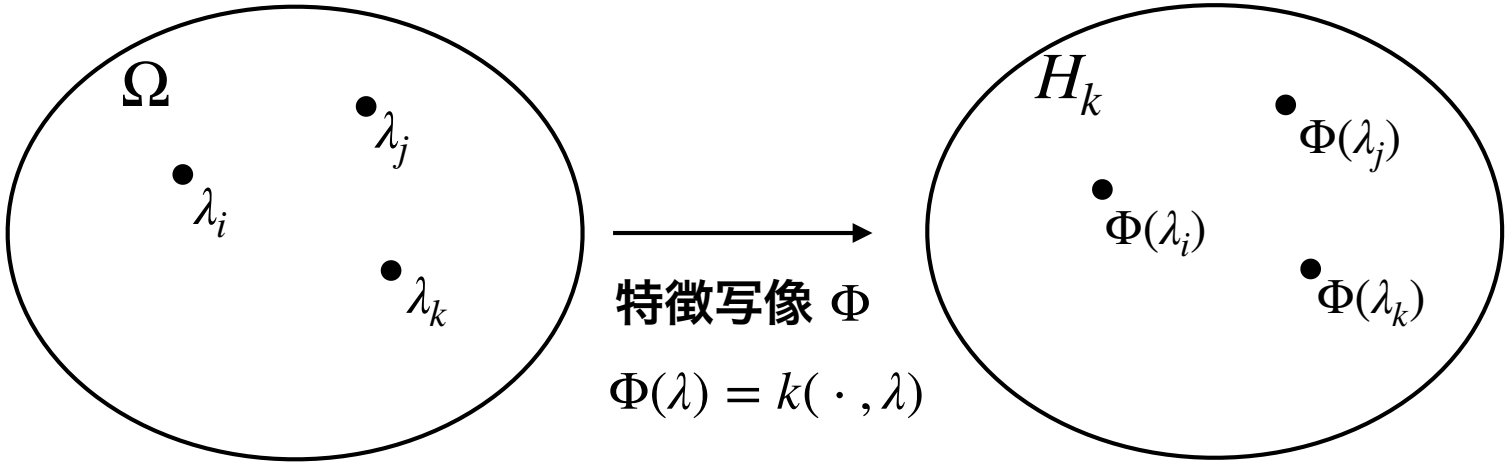
$$m_X(\cdot) := E(\Phi(\Lambda_X)) = E(k(\cdot, \Lambda_X)) = \int k(\cdot, \lambda) \mathrm{d}P_X(\lambda)$$

カーネル平均：確率ベクトル $\Phi(\Lambda_X)$ の平均 $E(\Phi(\Lambda_X))$ だけで表現する。

3-5. カーネル平均埋め込み

確率変数 $\Lambda_X \sim P_X(\lambda)$

確率ベクトル $\Phi(\Lambda_X)$



元の空間

特徴空間(再性核ヒルベルト空間)

定義: $P_X(\lambda)$のカーネル平均

$$m_X(\cdot) := E(\Phi(\Lambda_X)) = E(k(\cdot, \Lambda_X)) = \int k(\cdot, \lambda) \mathrm{d}P_X(\lambda)$$

➡カーネル k が特性的である場合、カーネル平均 $m_X(\cdot)$ は確率 $P_X(\cdot)$ を一意に定める。

$$m_X(\cdot) = m_Y(\cdot) \Leftrightarrow P_X(\cdot) = P_Y(\cdot)$$

3-6. カーネル平均記述子

・材料 X は一般に有限個の構成要素からなるので、確率 $P_X(\lambda)$ は確率質量関数

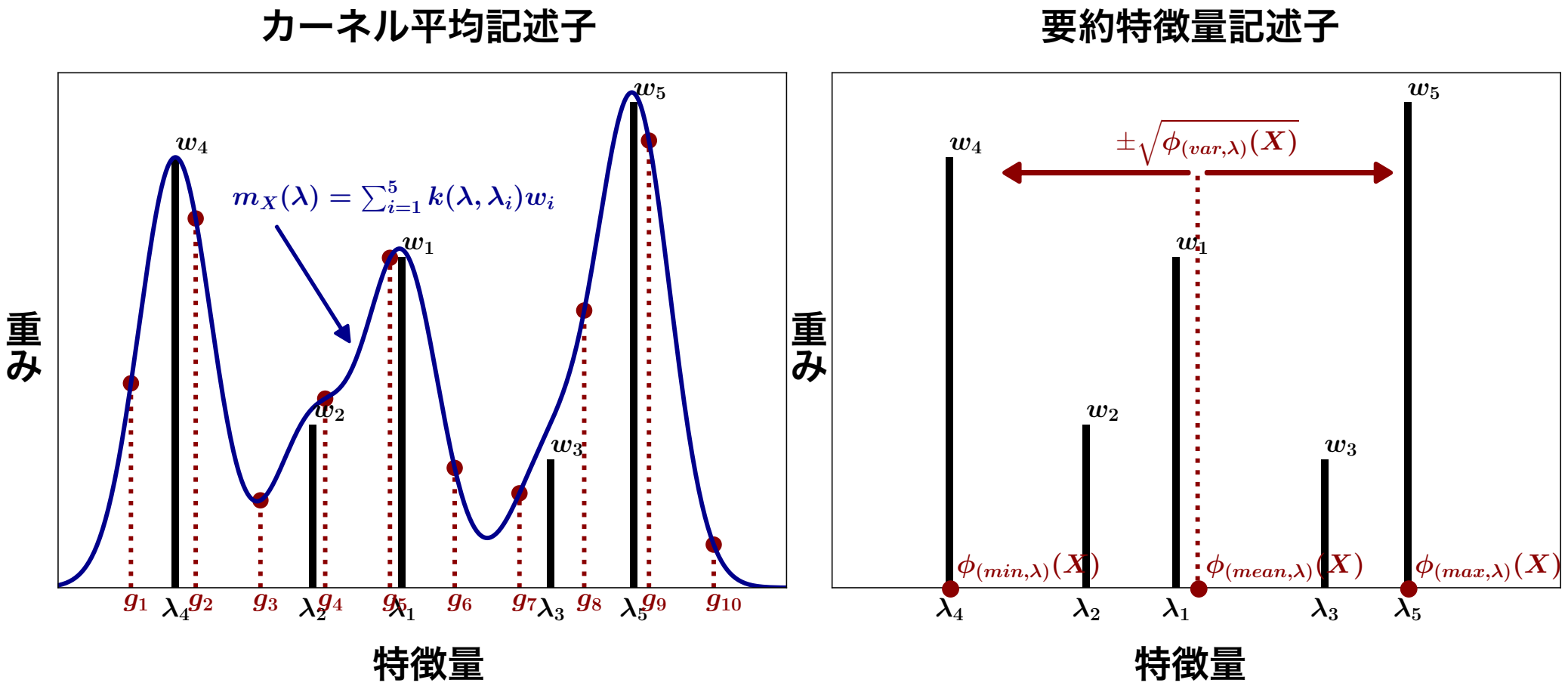
$P_X(\lambda) = \sum_{i=1}^{N_X} \delta(\lambda - \lambda_i) w_i$ のように表現できる。その場合、カーネル平均は以下のように表される。

$$m_X(\lambda) = \sum_{i=1}^{N_X} k(\lambda, \lambda_i) w_i$$

・上式によって、材料 X の分布 $\{(w_i, \lambda_i) \mid i = 1, \ldots, N_X\}$ の情報は、情報損失無くベクトル形式に変換される (k が特性的である場合)。

・しかし、$m_X(\lambda)$ は無限次元のベクトルなので、有限化する必要がある。本手法では λ 上で均等に配置されたグリッド点 $(g_1, \ldots, g_d)$ によって、 $m_X(\lambda)$ を有限化した d 次元ベクトル $(m_X(g_1), \cdots, m_X(g_d))^\top$ を最終的な記述子とする。

3-6. カーネル平均記述子

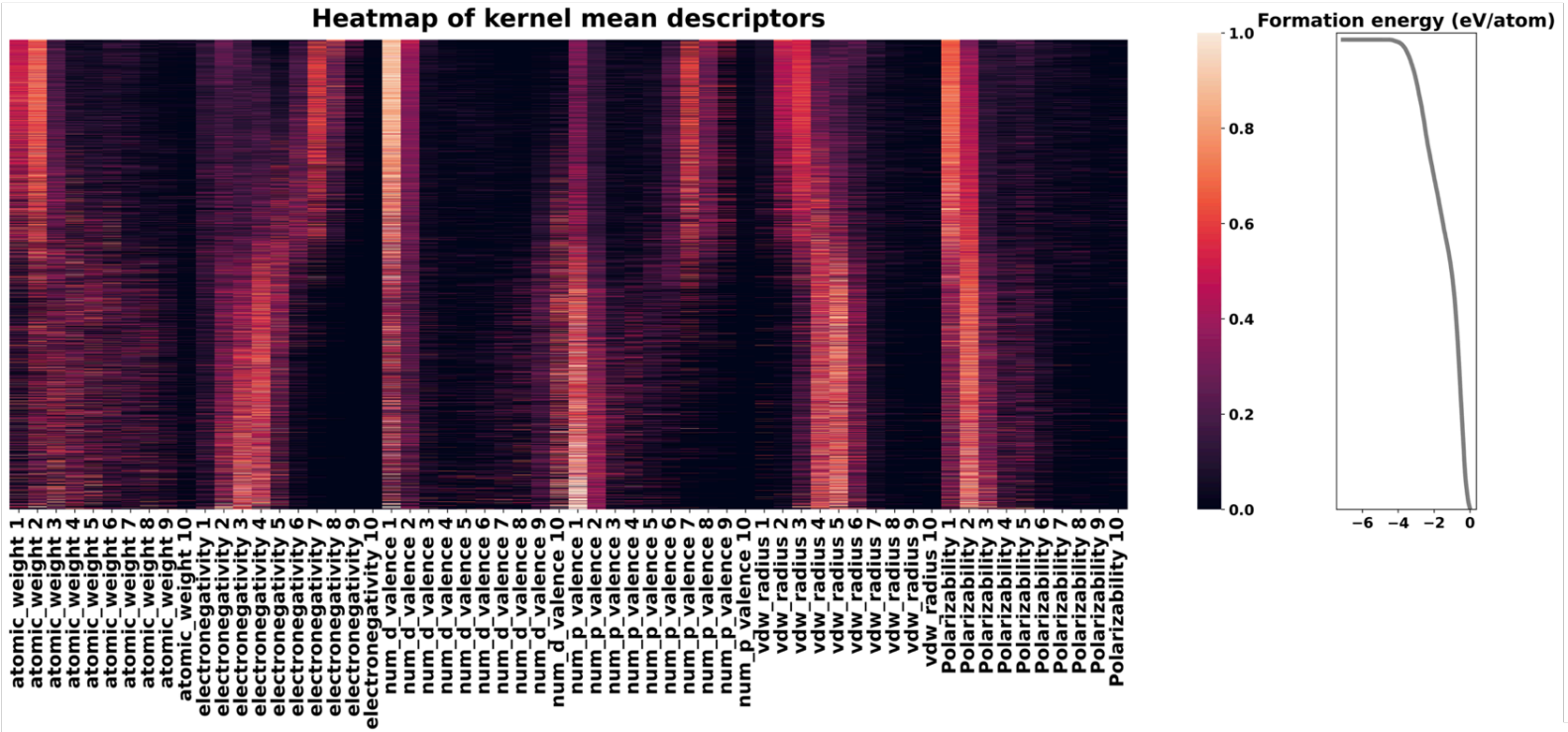


・材料 X の最終的な記述子 $\phi(X)$ は、各特徴量 $\lambda \in \Lambda$ について計算したカーネル平均記述子 $\phi_\lambda(X)$を結合して与えられる。

3-7. カーネル平均記述子の実例

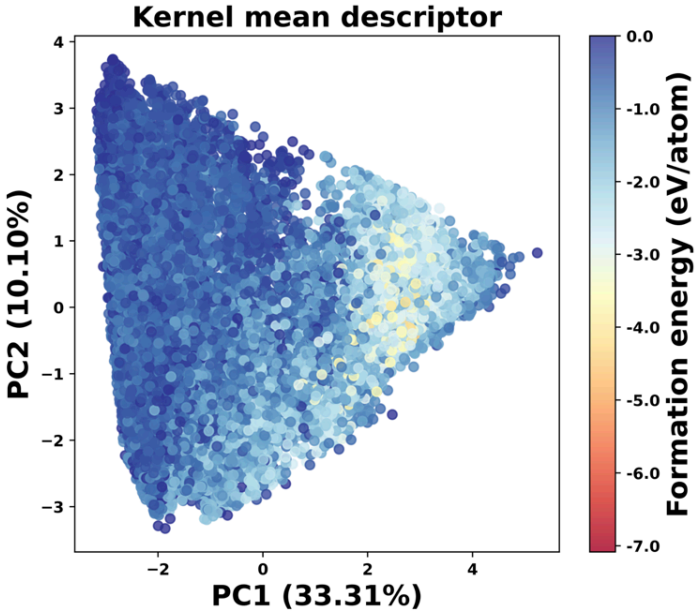
約3.5万個の
化学組成に対して
記述子を生成

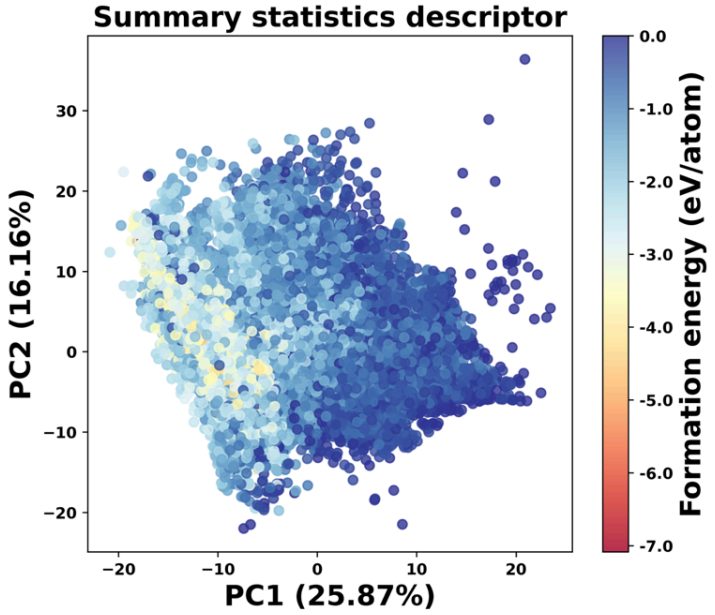
(a)



元素レベルの
記述子: 58次元
グリッド数: $d = 10$
(580次元)

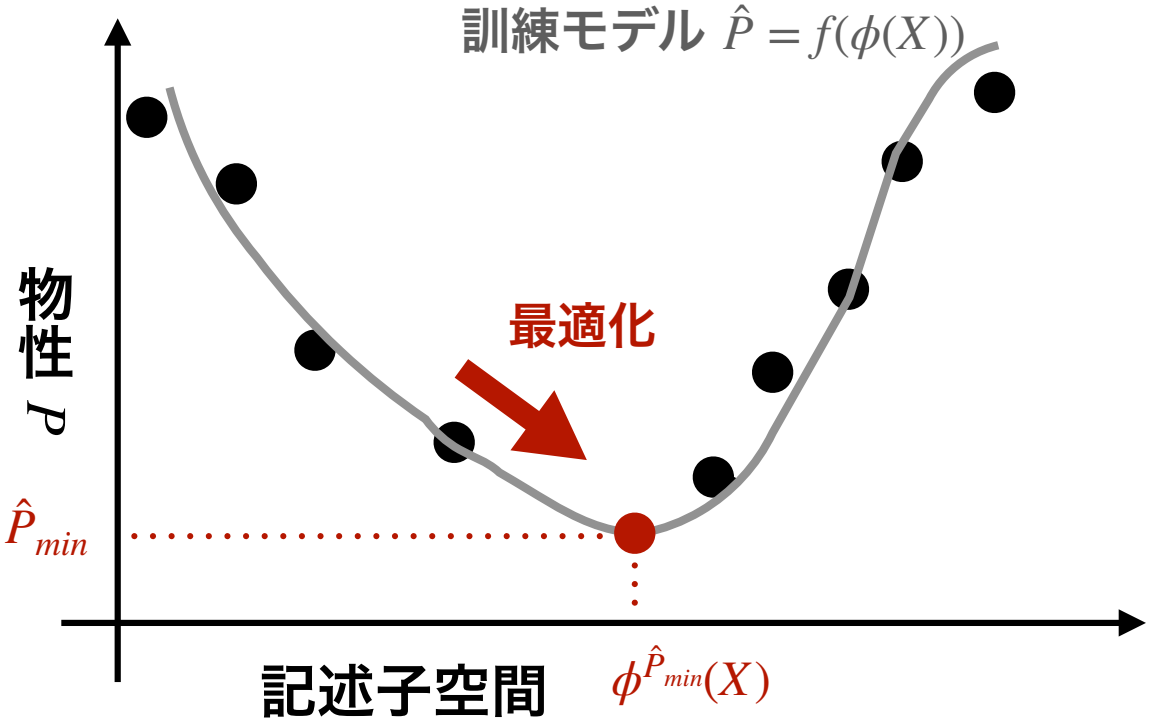
(b)





加重平均,加重
分散,最大, 最
小プーリング
による要約統
計量記述子
(232次元)

3-8. カーネル平均記述子からの材料への逆写像



・実用上、**材料の逆設計**によって望ましい物性を持つと予測される候補材料を得るためには、記述子空間から材料空間への**逆写像手法**が必要である。

・ところが、**要約統計量記述子**では選択される統計量の任意性、加重分散等の変換の非線形性より、**統一的な逆写像フレームワークを作ることができない。**

逆写像 $\phi^{\hat{P}_{min}}(X) \rightarrow X^{\hat{P}_{min}}$ により、候補材料を得る。

3-8. カーネル平均記述子からの材料への逆写像

材料系 X の考えられる全ての構成要素候補を$\{x_1, \ldots, x_N\}$とし、これを構成要素集合と呼ぶ。$\{w_1, \ldots, w_N\}$をこの集合に対する重みとする。各構成要素 x_i に対して K 個の特徴量 $\lambda_i^k = \lambda^k(x_i) \in \mathbb{R}\ (k = 1,\ldots, K)$ が定義されており、各特徴量空間 λ^k 上で d 個 のグリッド点 $(g_1^k, \ldots, g_d^k)$ が用意されているとする。その時、要素が $G_{ij}^k = k(\lambda_i^k, g_j^k)$ である $N \times d$ 行列 G^k によって k 番目の特徴量に対するカーネル平均記述子 $\phi_{\lambda^k}(X)$ は以下のように書ける。

$$\phi_{\lambda^k}(X) = {G^k}^{\top} w, \quad w = (w_1, \ldots, w_N)^{\top}. \quad (1)$$

式 (1) より、材料 X の全ての特徴量を結合した最終的なカーネル平均記述子 $\phi(X)$ は以下のように示せる。

$$\phi(X) = \begin{pmatrix} \phi_{\lambda^1}(X) \\ \vdots \\ \phi_{\lambda^K}(X) \end{pmatrix} = \begin{pmatrix} {G^1}^{\top} \\ \vdots \\ {G^K}^{\top} \end{pmatrix} w = Hw\,. \quad (2)$$

$dK \times N$ 行列 H は、グリッド点、特徴量、カーネルの情報から自動的に計算される値であり、記述子から材料空間への逆写像は、任意の与えられた ϕ^* に対して $\|\phi^* - Hw\|^2$ を最小にする重み $w = w^*$ を推定する問題として定式化できる。

3-8. カーネル平均記述子からの材料への逆写像

$$\begin{aligned} \min_{w} \ & \|\phi^* - Hw\|^2 \\ \text{s.t.} \ & \mathbf{1}^\top w = 1 \\ & w \geq \mathbf{0} \end{aligned} \tag{3}$$

式 (3) は以下のように、二次計画問題の形式に変形できる。

$$\begin{aligned} \min_{w} \ & \frac{1}{2} w^\top H^\top H w - \phi^{*\top} H w \\ \text{s.t.} \ & \mathbf{1}^\top w = 1 \\ & w \geq \mathbf{0}. \end{aligned} \tag{4}$$

式 (4) において、行列 $H^\top H$ がフルランク(i.e., $\mathrm{rank}(H^\top H) = N$)であれば、目的関数は狭義に凸になるので、一意的な最適解が存在する。これは、$H^\top H$ がフルランクであれば、任意のカーネル平均記述子 ϕ^* は、重み w^* を持つ特定の材料 X^* に一意にマッピングされることが保証されることを意味する。

$H^\top H$ がフルランクであるには、$dK \times N$ 行列 H のランクが N である必要があるが、グリッド点数 d はユーザーが調整可能な値であるため、あらかじめ $\mathrm{rank}(H) = N$ を満たすような d を選択することで、一意的な逆写像を保証するカーネル平均記述子を設計することができる。

3-9. 予測実験の結果

(1) 無機化合物の形成エネルギー予測 (eV/atom)

Descriptors	MAE	RMSE	R^2
Kernel mean	**0.0359(±0.0025)**	**0.0590(±0.0069)**	**0.9967(±0.0009)**
Summary statistics	0.0413 (±0.0006)	0.0658 (±0.0070)	0.9959 (±0.0009)

(2) 準結晶材料を形成するための化学組成の予測

	Class	Recall	Precision	F_1	Macro F_1
Kernel mean	QC	0.562 (±0.131)	0.798 (±0.040)	0.653 (±0.106)	**0.790(±0.039)**
	AC	0.662 (±0.050)	0.791 (±0.108)	0.718 (±0.063)	
	Others	1.000 (±0.000)	0.996 (±0.001)	0.998 (±0.001)	
Summary statistics	QC	0.538 (±0.102)	0.765 (±0.071)	0.629 (±0.090)	0.762 (±0.035)
	AC	0.612 (±0.047)	0.727 (±0.122)	0.661 (±0.072)	
	Others	0.999 (±0.001)	0.996 (±0.001)	0.997 (±0.001)	

(3) 高分子材料の特性評価における力場パラメータの使用

Physical properties	Descriptors	MAE	RMSE	R^2
Thermal conductivity $[\times 10^{-2} \cdot \mathrm{W} \cdot \mathrm{m}^{-1} \cdot \mathrm{K}^{-1}]$	Kernel mean	**2.15(±0.18)**	**3.21(±0.37)**	**0.677(±0.067)**
	Summary statistics	2.40 (±0.08)	3.29 (±0.14)	0.662 (±0.017)
Linear expansion coefficient $[\times 10^{-5} \cdot \mathrm{K}^{-1}]$	Kernel mean	**2.14(±0.20)**	**2.94(±0.26)**	**0.597(±0.052)**
	Summary statistics	2.22 (±0.13)	3.03 (±0.21)	0.572 (±0.040)
C_P $[\mathrm{J} \cdot \mathrm{kg}^{-1} \cdot \mathrm{K}^{-1}]$	Kernel mean	72.7 (±7.6)	124.3 (±19.8)	0.966 (±0.011)
	Summary statistics	**65.1(±6.7)**	**98.4(±10.9)**	**0.979(±0.004)**

3-10. 結論

・カーネル平均埋め込みに基づき、高い表現力と一意的な逆写像可能性をもった一般的な材料記述子のクラスであるカーネル平均記述子を提案した。

・カーネル平均記述子は今回紹介した3つの例以外に幅広い応用が考えられる。
例えば、結晶グラフ畳み込みニューラルネットワークモデル中の集約操作として本手法を採用することや、カーネル平均記述子に基づく類似性尺度を使用したより良いデータベース中の類似材料検索システムの構築等が挙げられる。

コード: https://github.com/Minoru938/KmdPlus

4. 関数データのためのベイズカーネル回帰

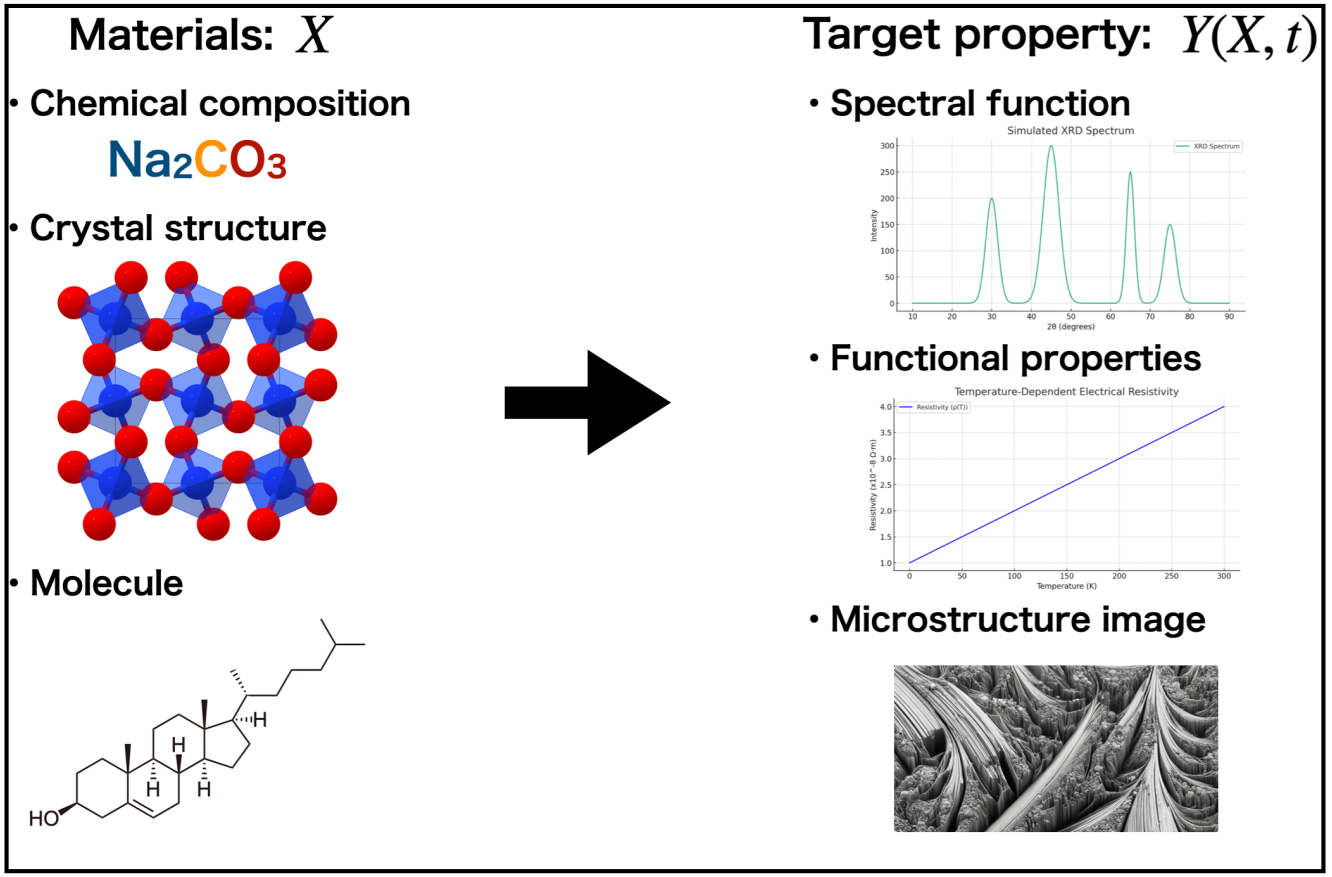
4-1. 研究背景

研究の動機：

・材料の物性は、スカラー値ではなく、スペクトルや分布などの関数値であることが多いが、材料データに関数出力回帰モデルを適用した研究は少ない（関数回帰モデルは、出力が関数形式で表現される回帰モデル）。

・実応用の複雑な問題に耐えうる非線形関数出力回帰モデルの研究が、統計科学分野の中でも少なく、応用分野の研究者にはほとんど知られていないためと考えられる。

・そこで本研究では、カーネル法に基づいた関数出力回帰モデル（KRFD）を提案する。



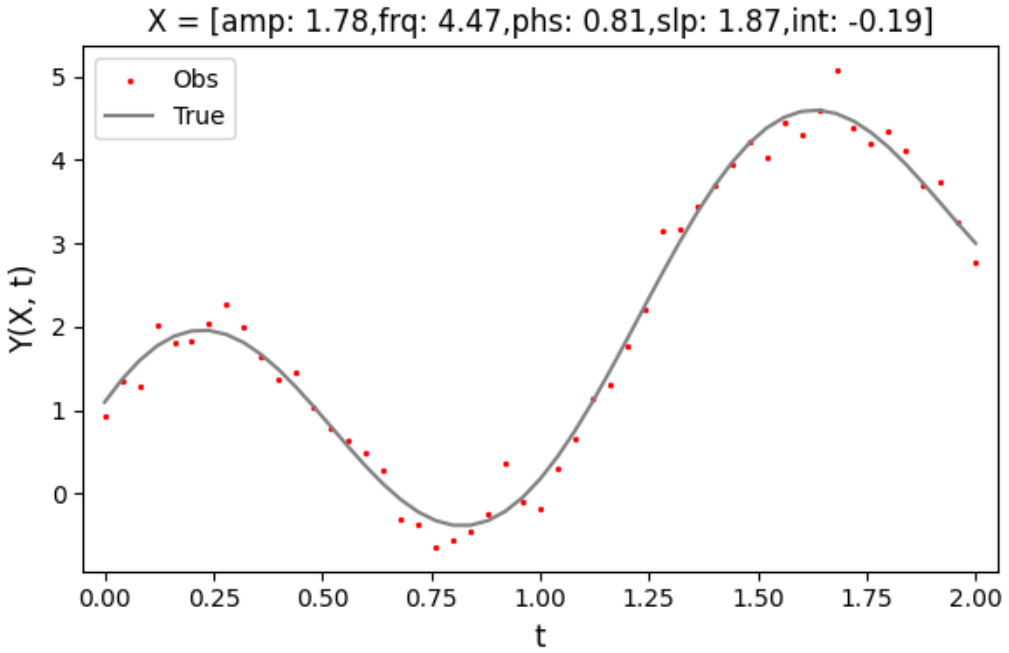
4-1. 研究背景

関数出力回帰モデルの利点：

・関数出力回帰モデルは、測定点ごとにスカラー出力をもつ別々の回帰モデルを独立に当てはめるのに比べ、主に以下3つの理由で、学習効率と汎化性能の向上が期待される;

①関数出力の共分散構造を活用できる
②関数領域全体に渡る自然な正則化をかけることができる
③測定点間で共有されるパターンを共同学習できる。

→さらに、関数出力回帰モデルは滑らかな関数全体を予測するので、一貫性の無い測定点ごとの予測に比べて解釈性が高く、根底にあるデータ生成プロセスについてより多くの洞察を提供することが期待される。



4-1. 研究背景

既存研究：

・関数出力回帰はより広い枠組みである関数データ解析（FDA）の特殊な例と見做せる。

・FDAの中では、離散データを関数データとして関数化する（すなわち、データを連続関数として明示的に表現する）手法や、出力や入力値が関数化されていることを前提とした関数回帰手法は数多く検討されている。一方本研究では、関数化の段階を経ずにベクトル入力値と離散関数データから直接モデルを構築することを目標としている。

・この問題設定に適合する手法は、function-on-scalar 回帰モデル (FSRM) と呼ばれる。FSRMの中では、線型モデルを関数に拡張した関数線型モデル (FLM) が最も一般的であり、数多くの研究が行われてきた。

FLM: $$Y(X,t) = \beta_0(t) + \sum_{j=1}^{p} x_j \beta_j(t) + \epsilon \qquad {}^*X = (x_1, \ldots, x_p)^{\mathrm{T}}$$

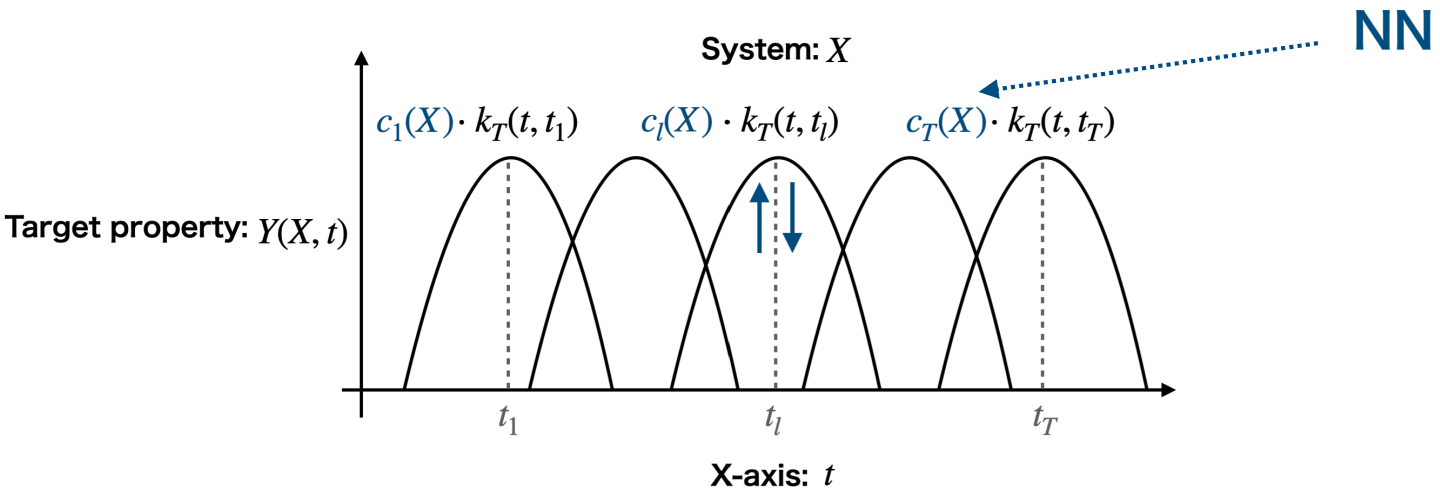
・FLMは入力値 X に対しては線形であり、モデルの表現能力に大きな制約がある。

・X に対する非線形性を持ったFSRMの先行研究は少なく、例としてYaoらによる共変量に2次の項を追加した関数2次回帰モデル （Biometrika, 97, 1, 2010）、Scheiplらによる関数データに加法混合モデルを拡張した関数加法混合モデル（J. Comput. Graph. Stat., 24, 2, 2015）、Luoらによる汎関数普遍近似定理を応用したニューラルネットワーク（NN）ベースのモデル（Biometrics, 79, 4, 2023）が挙げられる。

4-1. 研究背景

既存研究（続き）：

・Iwayamaらは t 空間上に設置したカーネル関数と入力値 X に依存するニューラルネットワークモデル（NN）の線型結合により関数を表現するモデルを提案した (J. Chem. Inf. Model, 62, 20, 2022)。



提案手法（KRFD）の特徴：

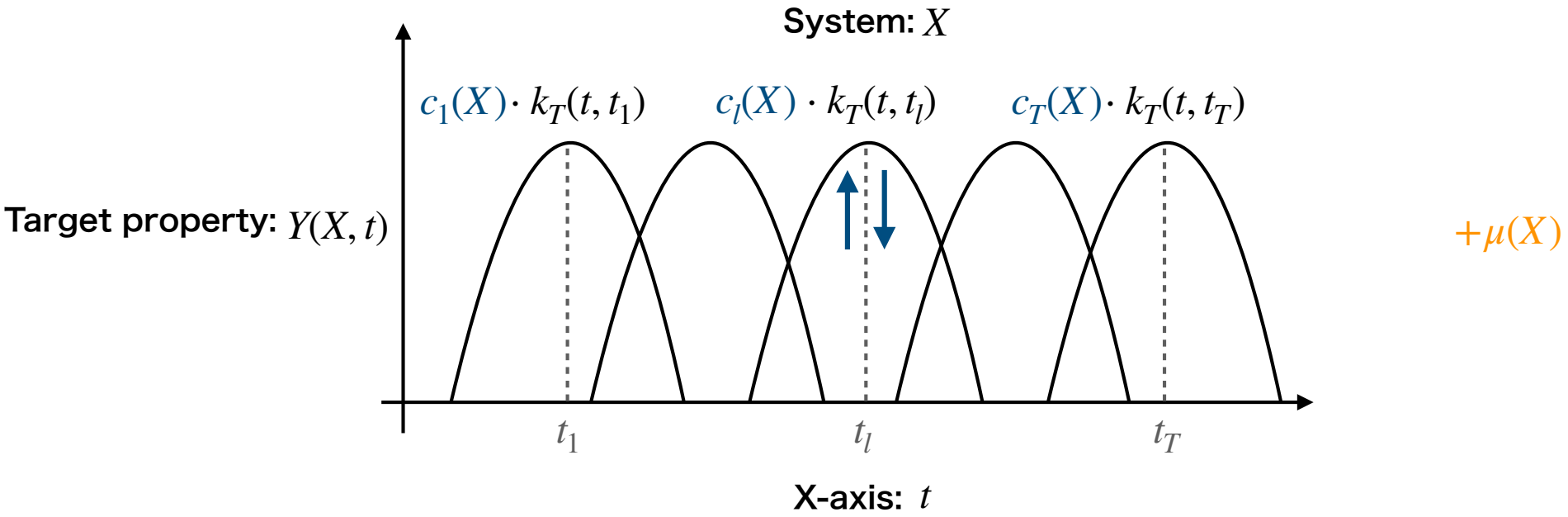
・KRFDはIwayamaらによるんモデルのNN部分をカーネルリッジ回帰（KRR）$c_l(X)=\sum_{k=1}^{N}\theta_{kl}k_G(X,X_k)$

で置き換えたものと見做せる。

・これによりモデルの形式が単純化し、解析的最適解の導出、ベイズ化、モデルの理論的分析が可能になる。

コード: https://github.com/Minoru938/KRFD

4-2. KRFDモデルの概要



Kernel Regression for Functional Data (KRFD)

$$Y(X,t) = f(X,t) + \mu(X)$$

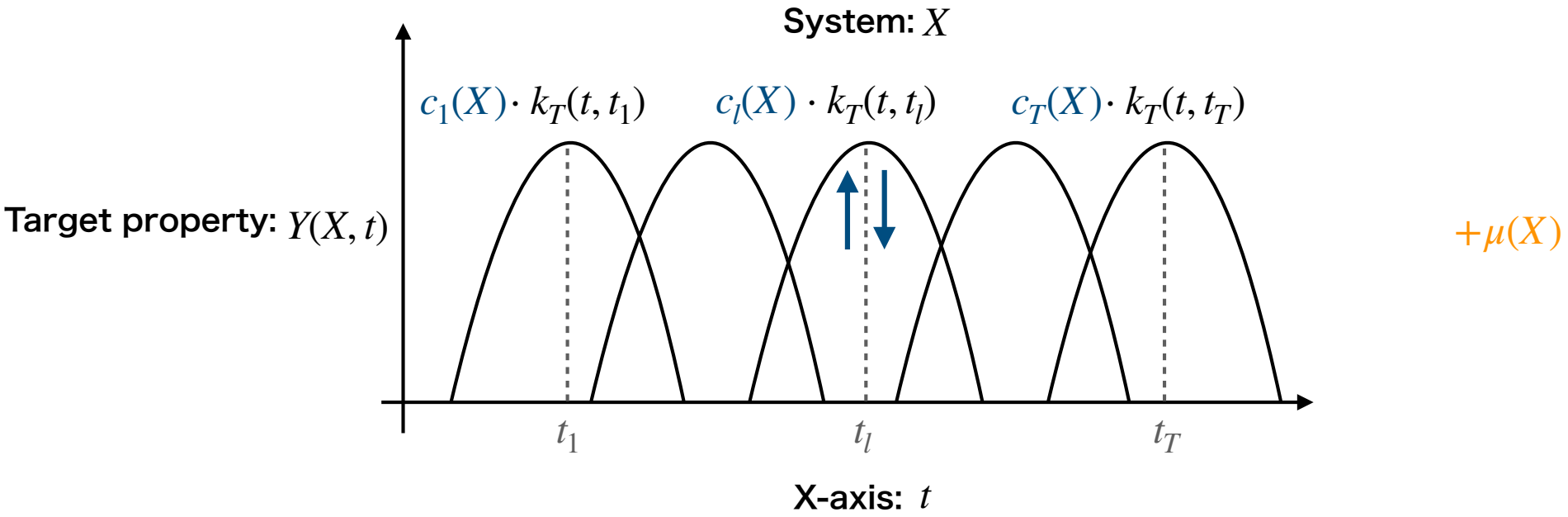
Variation term: $$f(X,t) = \sum_{l=1}^{T} c_l(X) k_T(t,t_l), \quad c_l(X) = \sum_{k=1}^{N} \theta_{kl} k_G(X,X_k) \rightarrow f(X,t) = \sum_{k=1}^{N}\sum_{l=1}^{T} k_G(X,X_k)\theta_{kl}k_T(t,t_l)$$

System-dependent constant term: $$\mu(X) = \sum_{m=1}^{N} c_m k_M(X,X_m)$$

Training data

$$X_i\ (i=1,\ldots,N) \in \mathbb{R}^p$$
$$t_j\ (j=1,\ldots,T) \in \mathbb{R}^q$$
$$Y(X_i,t_j) \in \mathbb{R}$$

4-2. KRFDモデルの概要



Kernel Regression for Functional Data (KRFD)

$$Y(X,t)=\sum_{k=1}^{N}\sum_{l=1}^{T}k_G(X,X_k)\theta_{kl}k_T(t,t_l)\ +\ \sum_{m=1}^{N}c_m k_M(X,X_m)$$

→ベイズモデルとして推定する

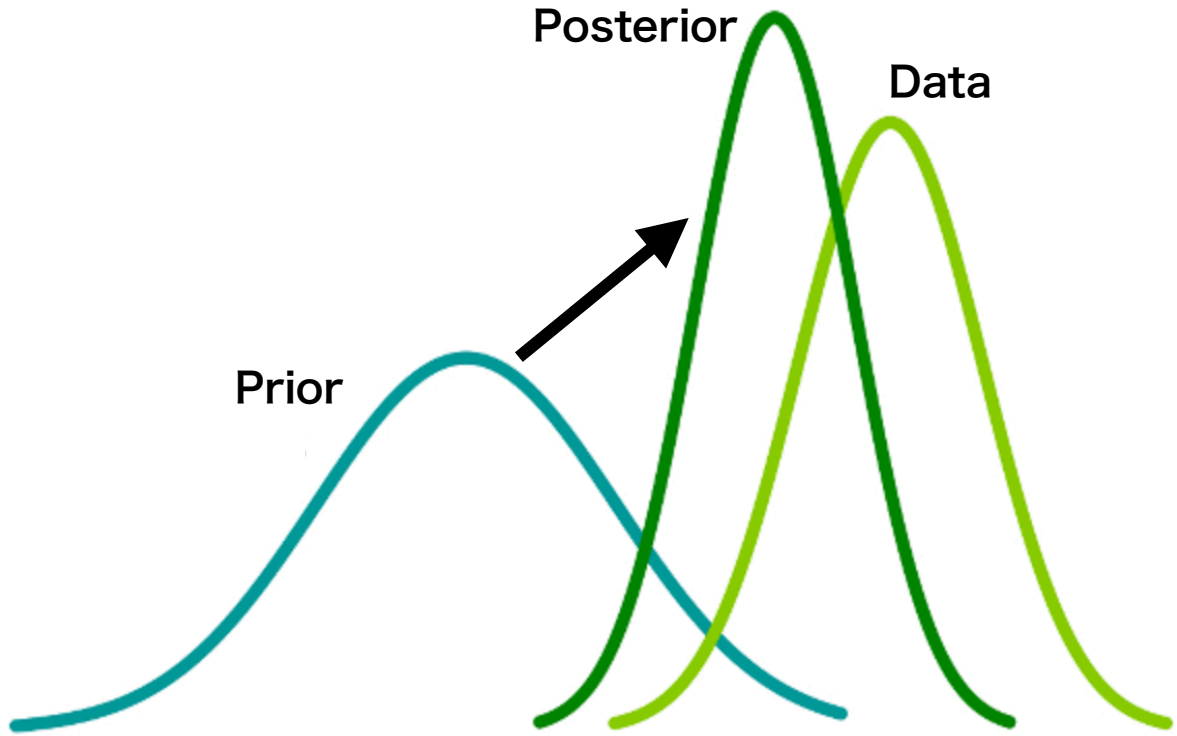
Training data

$X_i\ (i=1,\ldots,N)\in\mathbb{R}^p$

$t_j\ (j=1,\ldots,T)\in\mathbb{R}^q$

$Y(X_i,t_j)\in\mathbb{R}$

4-3. ベイズモデルとは?



Posterior distribution　　Model likelihood

Bayes' theorem: $P(\theta \mid Y) \propto P(Y \mid \theta)P(\theta)$ ← Prior distribution

→ In Bayesian model, model parameters θ are treated as random variables.
Deriving posterior distribution of θ is the goal of Bayesian model.

4-4. KRFDモデルのベイズ推定

Assumption for observed data

$$Y(X_i, t_j) = \sum_{k=1}^{N}\sum_{l=1}^{T} k_G(X_i, X_k)\theta_{kl}k_T(t_j, t_l) + \sum_{m=1}^{N} c_m k_M(X, X_m) + \epsilon_{ij}, \quad \epsilon_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2), \quad (i = 1,\ldots,N,\ j = 1,\ldots,T)\,.$$

Model likelihood

$$p(\boldsymbol{y}\,|\,\boldsymbol{\theta}, \boldsymbol{c}, \sigma^2) = N(\boldsymbol{G} \otimes \boldsymbol{T} \cdot \boldsymbol{\theta} + \boldsymbol{M} \otimes \boldsymbol{1}_T \cdot \boldsymbol{c}, \sigma^2 \boldsymbol{I}_{NT})$$

Prior distributions

$$p(\boldsymbol{\theta}) = N\left(\boldsymbol{0}, \left(2\lambda_a(\boldsymbol{G} \otimes \boldsymbol{T}^2) + 2\lambda_b(\boldsymbol{G}^2 \otimes \boldsymbol{T}) + 2\lambda_c(\boldsymbol{G} \otimes \boldsymbol{T})\right)^{-1}\right)$$

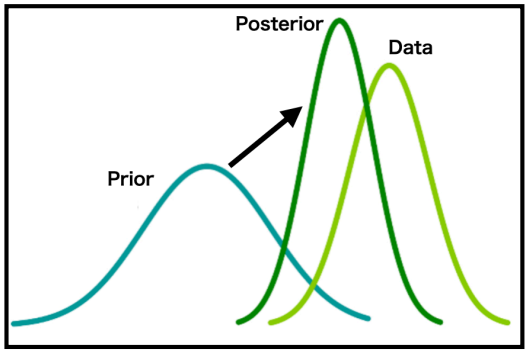
$$p(\boldsymbol{c}) = N\left(\boldsymbol{0}, \left(2T\lambda_d \boldsymbol{M}\right)^{-1}\right)$$

$$p(\sigma^2) = IG(\alpha, \beta)$$ ←prior for the variance of observation noise

↓ By known math formulas for Gaussian distributions

Posterior distributions

→ **Prediction distribution on** (X_{new}, t_{new})



・θ の事前分布はKRFDモデルの関数複雑性に対して X 空間と t 空間別々に正則化できるよう設計した。

4-5. 人工データ上での数値実験

Data generation process: $y(t) = a\sin(bt + c) + dt + e + \epsilon, \quad \epsilon \overset{i.i.d}{\sim} N(0, \sigma^2)$.

a : *amplitude* $\quad$ d : *slope*

b : *frequency* $\quad$ e : *intercept*

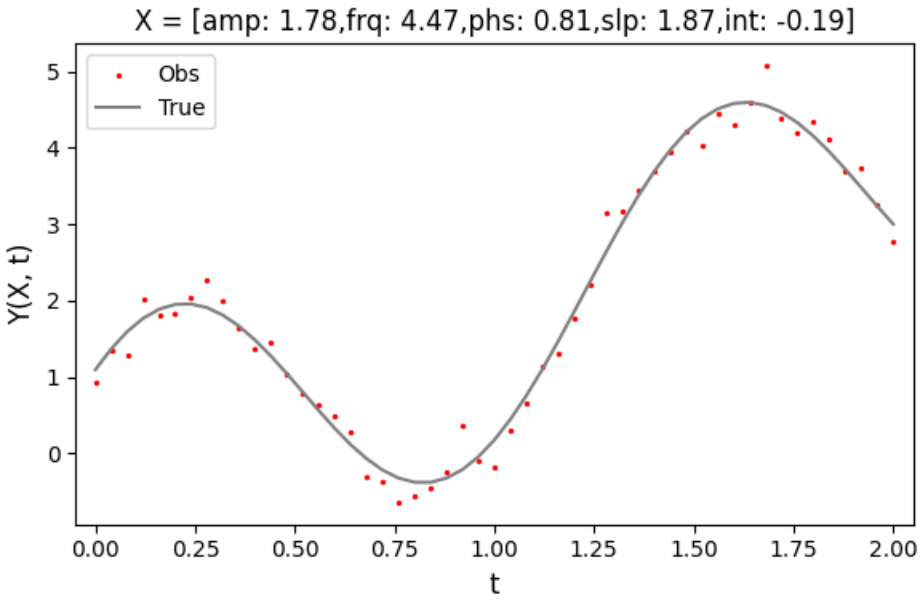
c : *phase* $\quad$ ϵ : *noise*

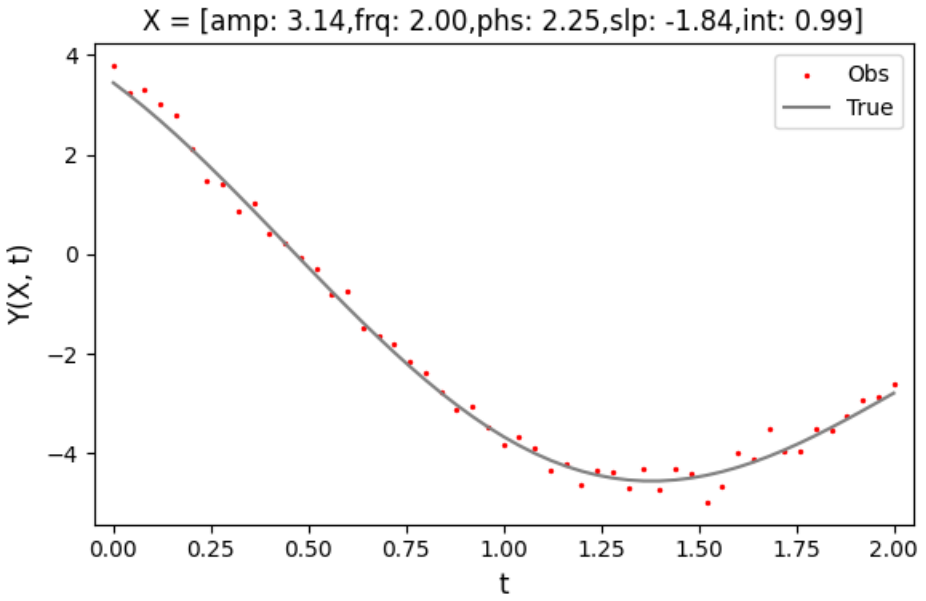
Training data

$X_i = (a_i, b_i, c_i, d_i, e_i)^{\mathrm{T}}, \quad (i = 1, \ldots, N) \in \mathbb{R}^5$

$t_j \ (j = 1, \ldots, T) \in \mathbb{R}$

$Y(X_i, t_j) \in \mathbb{R}$





Task: Can model rediscover true DG process from observed data?

4-5. 人工データ上での数値実験

比較用のモデル：

1. FLM: $Y(X,t) = \beta_0(t) + \sum_{j=1}^{p} x_j \beta_j(t) + \epsilon$ $\qquad {}^*X = (x_1, \ldots, x_p)^{\mathrm{T}}$

$$\downarrow$$

$$\beta_j(t) = \sum_{i=1}^{T} k_T(t_i, t)\theta_i^j$$

→入力値 X に対する非線形性の導入によりどれだけ汎化性能の向上するかを確認

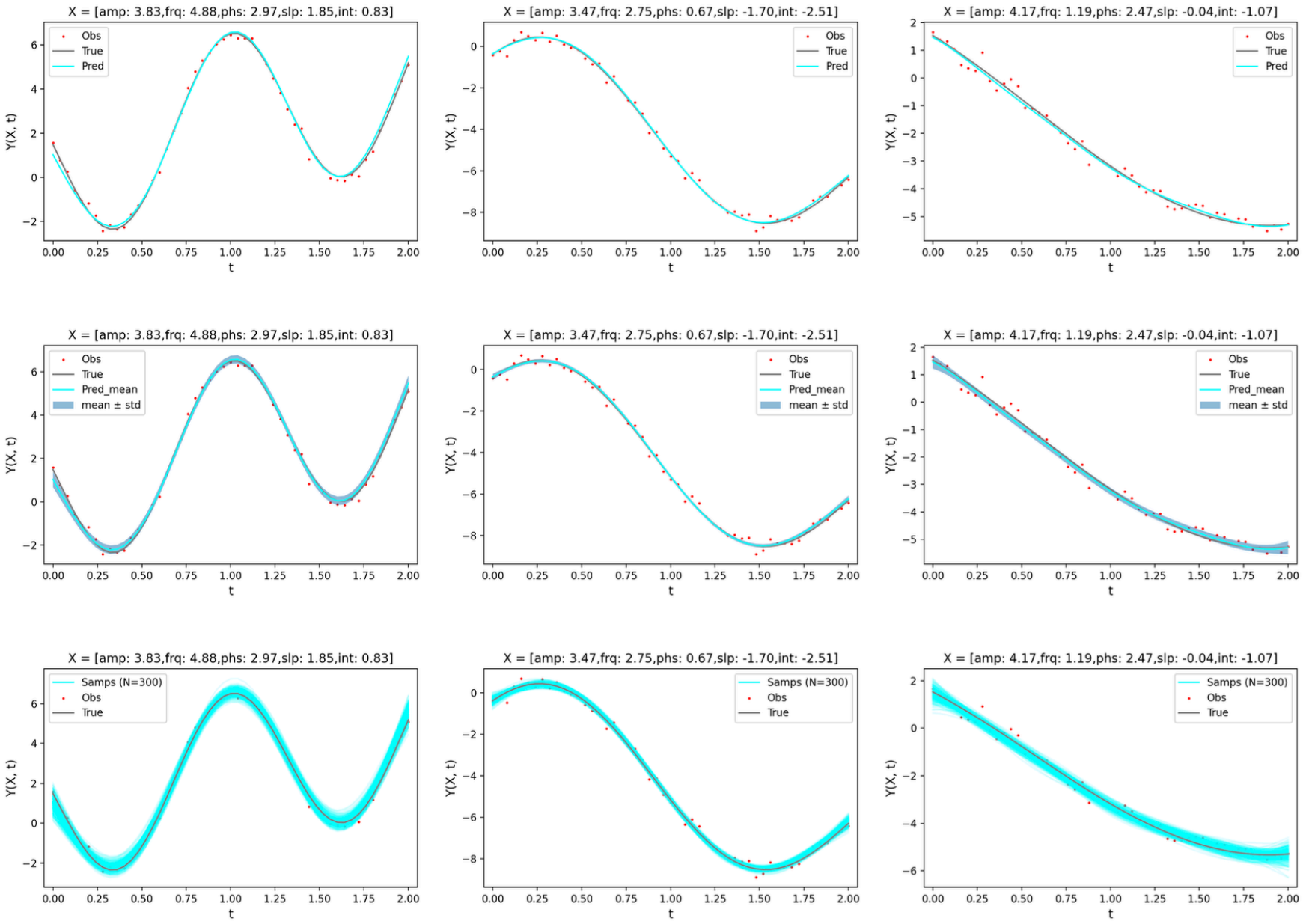
2. KRRs: $Y(X,t_j) = \sum_{i=1}^{N} k_G(X_i, X)\theta_i^j + \epsilon \quad (j = 1,\ldots,T)$

→測定点ごとにスカラー出力をもつ別々のKRRモデルを独立に当てはめる

→共分散構造を活用等により、どれだけ汎化性能の向上するかを確認

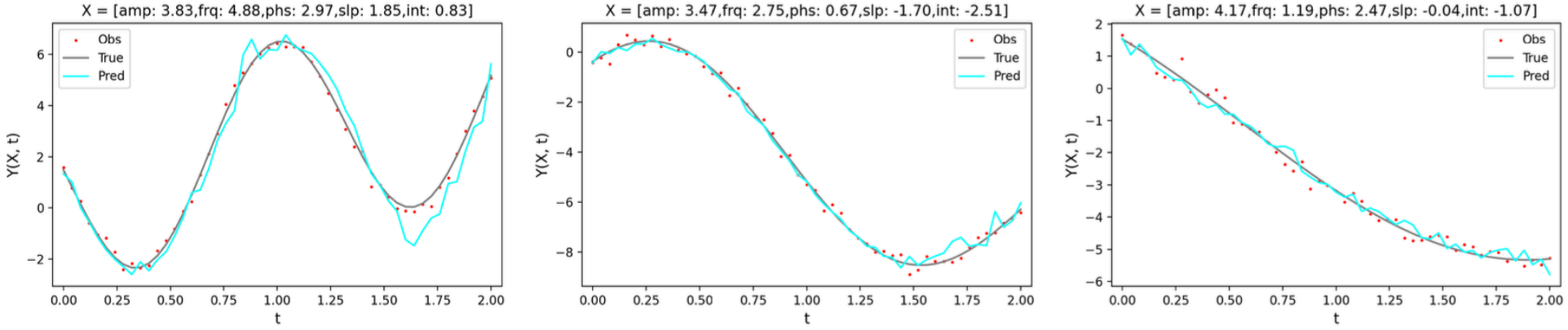
4-5. 人工データ上での数値実験（結果）

(a) KRFD

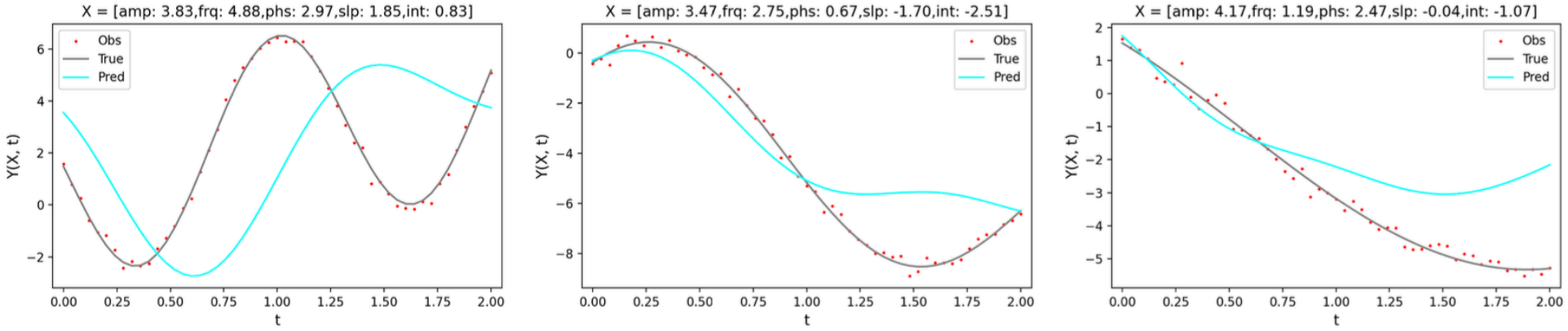


4-5. 人工データ上での数値実験（結果）

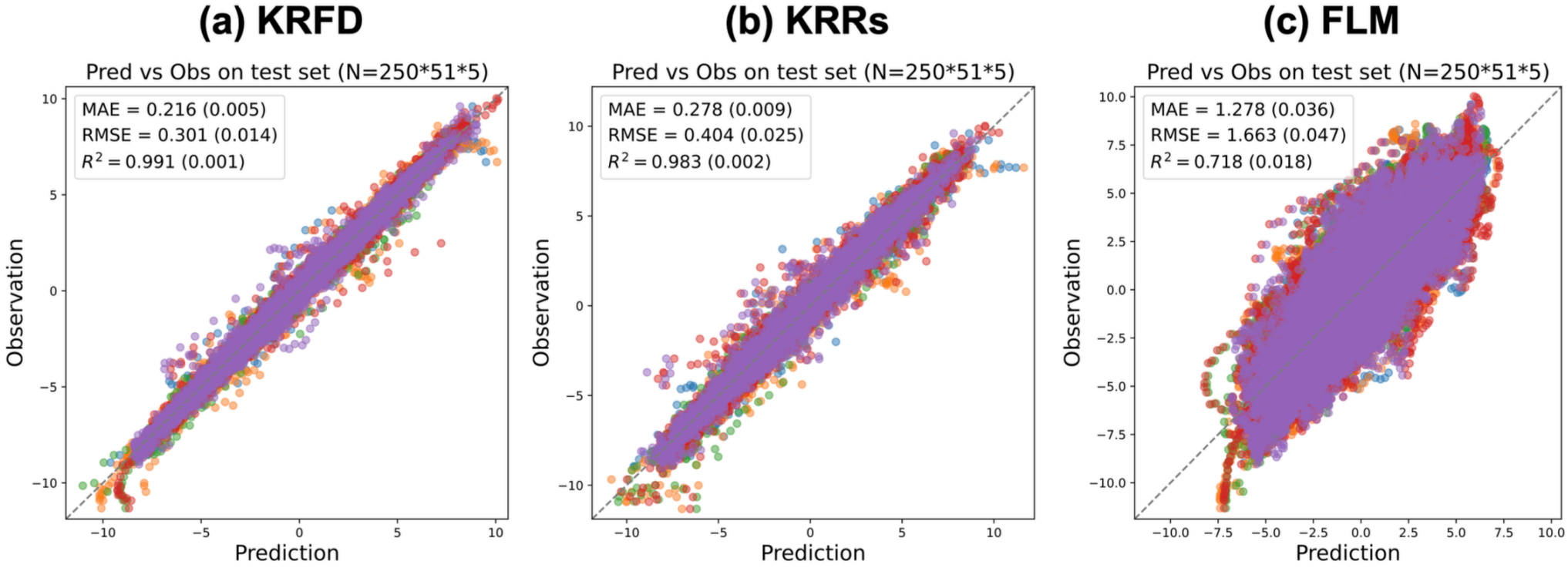
(b) KRRs



(c) FLM



4-5. 人工データ上での数値実験（結果）



Models	MAE	RMSE	R^2
KRFD	0.216 ($\pm$0.005)	0.301 ($\pm$0.014)	0.991 ($\pm$0.001)
KRRs	0.278 ($\pm$0.009)	0.404 ($\pm$0.025)	0.983 ($\pm$0.002)
FLM	1.278 ($\pm$0.036)	1.663 ($\pm$0.047)	0.718 ($\pm$0.018)

4-6. 材料データ上での数値実験

タスク: 安定金属化合物の状態密度（DOS）を化学組成情報のみから予測する。

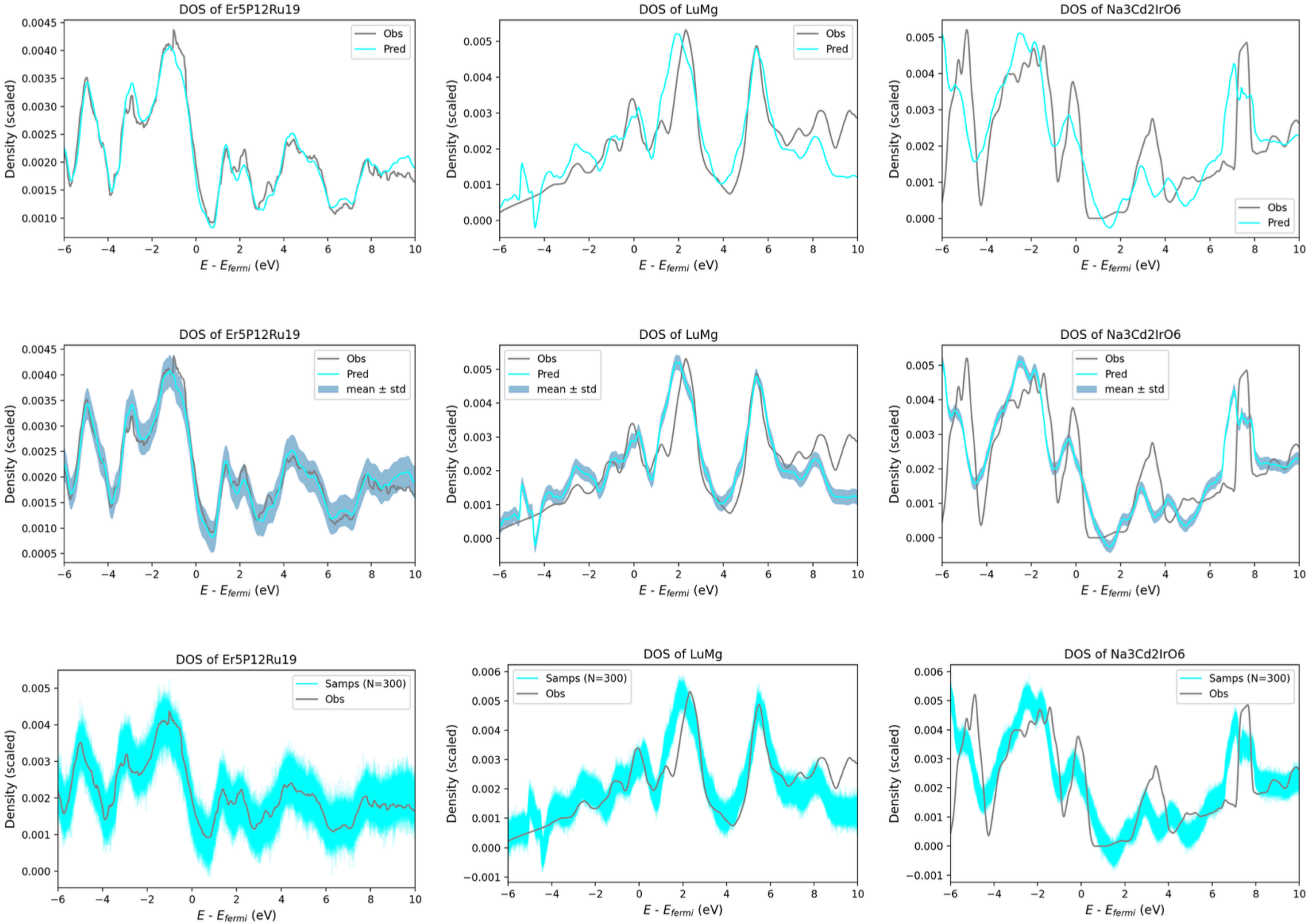
The Materials Project

Materials Project 中でDOSデータが登録されている全安定金属化合物（約1.3万個）をデータとして使用した。

化学組成はカーネル平均記述子により、580次元のベクトルに変換した

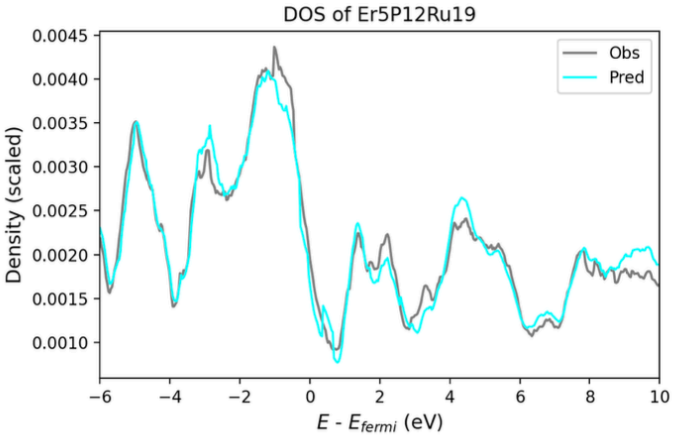
4-6. 材料データ上での数値実験

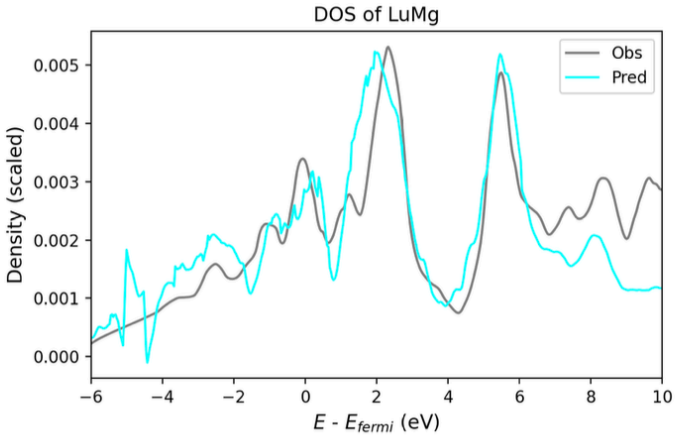
(a) KRFD

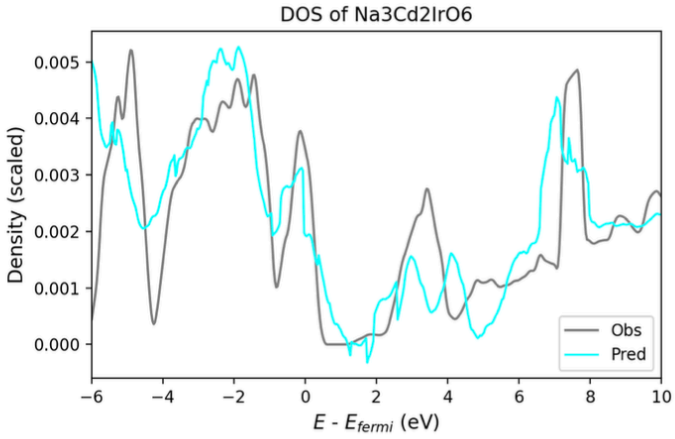


4-6. 材料データ上での数値実験

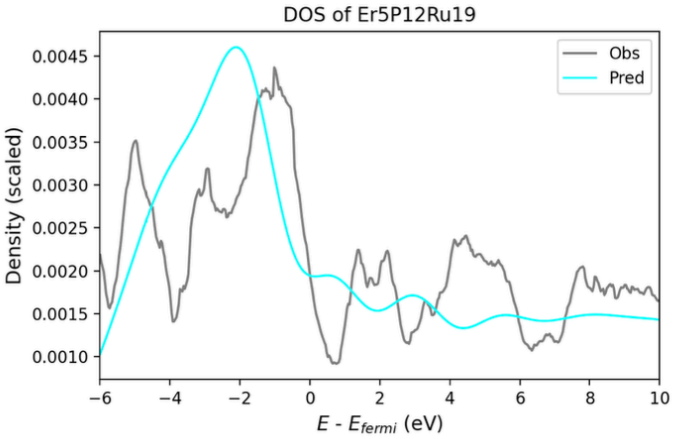
(b) KRRs

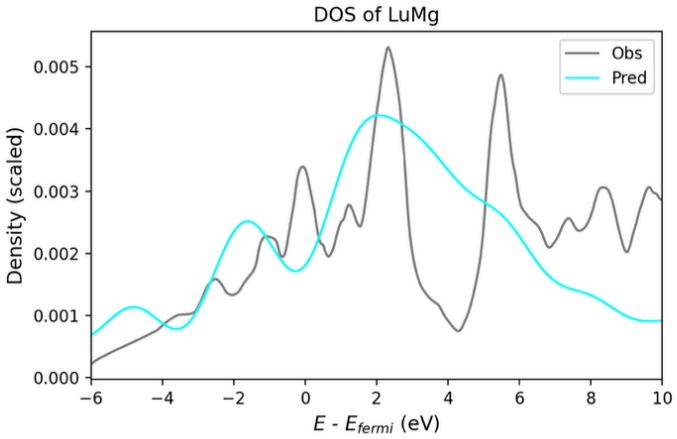


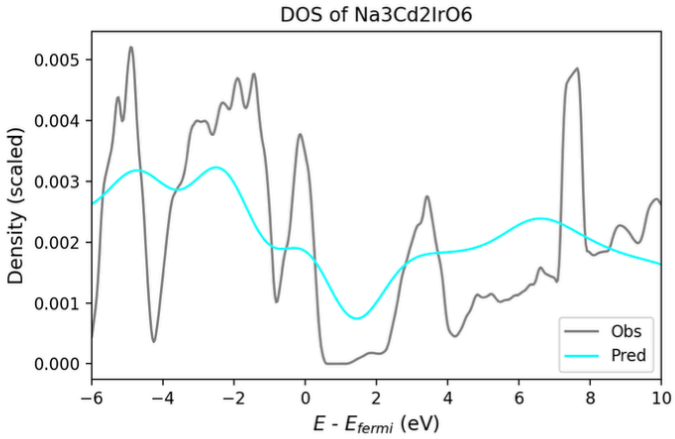




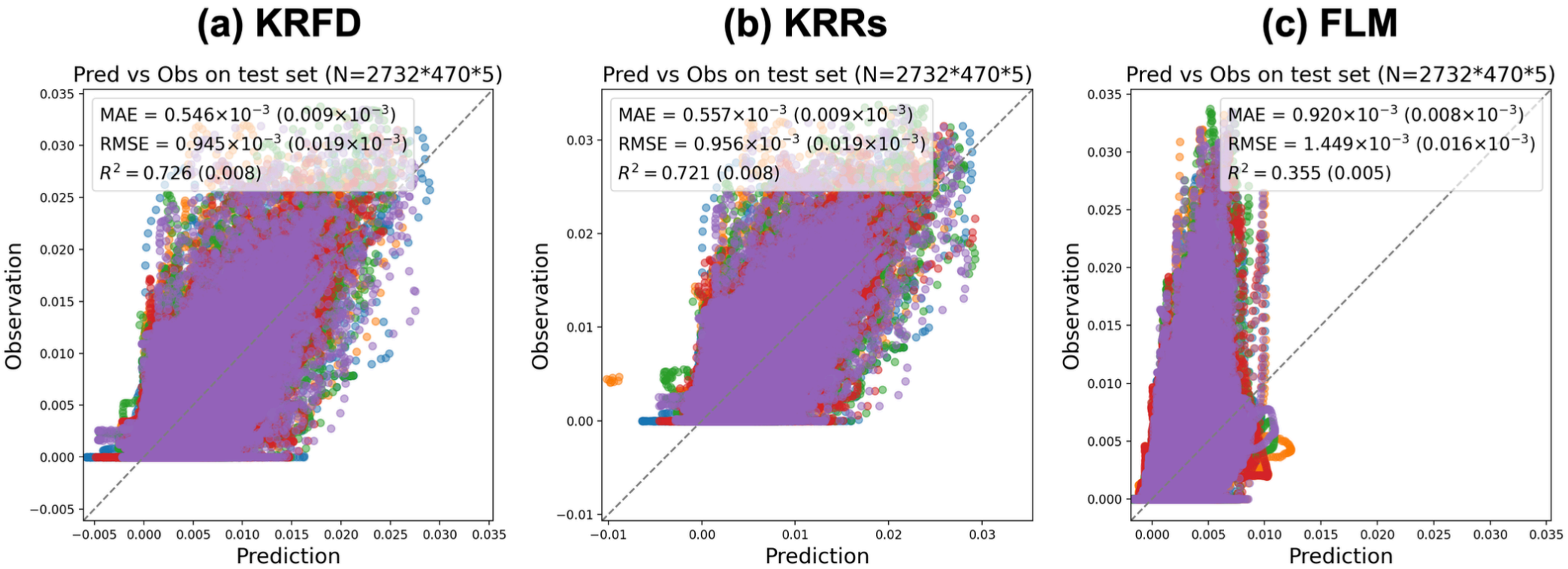
(c) FLM





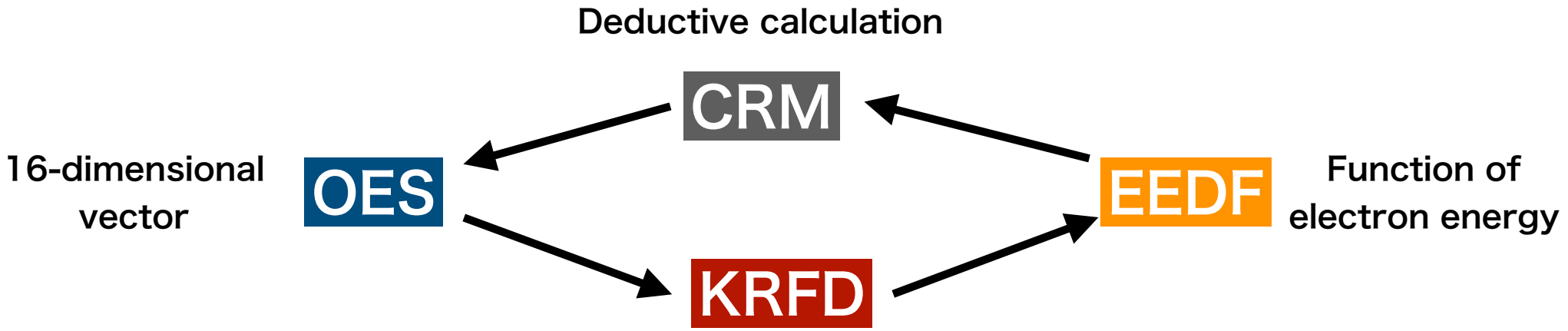


4-6. 材料データ上での数値実験



Models	MAE	RMSE	R^2
KRFD	0.546×10^{-3} $(\pm 0.009 \times 10^{-3})$	0.945×10^{-3} $(\pm 0.019 \times 10^{-3})$	0.726 (± 0.008)
KRRs	0.557×10^{-3} $(\pm 0.009 \times 10^{-3})$	0.956×10^{-3} $(\pm 0.019 \times 10^{-3})$	0.721 (± 0.008)
FLM	0.920×10^{-3} $(\pm 0.008 \times 10^{-3})$	1.449×10^{-3} $(\pm 0.016 \times 10^{-3})$	0.355 (± 0.005)

4-7. プラズマ科学への応用研究



Terms

- Electron energy distribution function (EEDF)
- Optical emission spectroscopy (OES)
- Collisional radiative model (CRM)

→非侵襲的な診断手法であるOESからEEDFを予測することにより、プラズマに干渉しないEEDFの推定を目指す

出典：Arellano, Fatima Jenina, et al. "Machine learning-based prediction of the electron energy distribution function and electron density of argon plasma from the optical emission spectra." *Journal of Vacuum Science & Technology A* 42.5 (2024).

4-7. プラズマ科学への応用研究

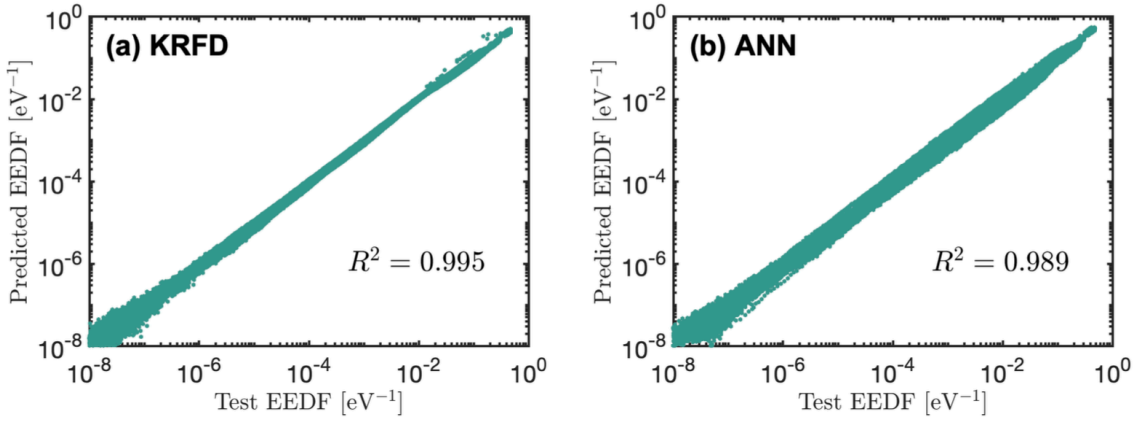


FIG. 2. Comparison of each predicted data point of the EEDF to its corresponding test data point, as predicted using KRFD (a) and ANN (b). A perfect prediction would give a diagonal line with an R^2 of 1.

・実験データでは課題が残ったが、KRFDはシミュレーションデータ上で高い予測精度を記録した。

*ANN: $Y(X, t_j) = NN([X, t_j])$

シミュレーションデータの説明：

・1次元のPIC/MCCシミュレーションを使用して、アルゴンプラズマのEEDFと電子密度（ne）を計算。
・PIC/MCCシミュレーションから得られたEEDFとneのデータを用い、CRMを使用して対応するOESを計算。
・合計108セットのOES, EEDFシミュレーションデータを得た。

出典：Arellano, Fatima Jenina, et al. "Machine learning-based prediction of the electron energy distribution function and electron density of argon plasma from the optical emission spectra." *Journal of Vacuum Science & Technology A* 42.5 (2024).

4-8. 結論

研究結果：

・カーネル法に基づく関数出力回帰モデルを提案した。

コード: https://github.com/Minoru938/KRFD

手法の特徴：

・モデルの形式が単純化であり、解析的最適解の導出、ベイズ化、モデルの理論的分析、係数ベースのモデル解釈を可能にする。

課題点：

・他のカーネル法に基づくモデルと同様にScalabilityに課題がある。行列の低ランク近似、確率的勾配降下法、分割統治法等の適用によりScalabilityを改善させることが考えられる。

その他の特徴：

・モデル形式は、Separable kernels の仮定とRepresenter定理より、再生核ヒルベルト空間から自然に導出される。

・スパースな関数データ（測定地点がシステムごとに異なる）にも対応可能だが、その場合計算コストが大幅に上がる。

・Cilibertoらによるベクトル値関数のRKHS上で定式化されたマルチタスク学習モデル (PMLR, 37，2015) と非常に近いことが明らかになっている。提案モデルはこれらのモデルを関数出力に一般化したものと見做すことができる。

5. 全体まとめ

今日のセミナーで紹介した研究

1. 機械学習を用いた元素置き換えによる結晶構造予測

- 構造類似性の予測に基づいた結晶構造予測手法を提案した。

論文: Kusaba, Minoru, Chang Liu, and Ryo Yoshida. "Crystal structure prediction with machine learning-based element substitution." Computational Materials Science 211 (2022): 111496.
コード: https://github.com/Minoru938/CSPML

2. カーネル平均埋め込みによる材料の表現

- カーネル平均埋め込みに基づいた一般的な材料記述子のクラスを提案した。

論文: Kusaba, Minoru, et al. "Representation of materials by kernel mean embedding." Physical Review B 108.13 (2023): 134107.
コード: https://github.com/Minoru938/KmdPlus

3. 関数データのためのベイズカーネル回帰

- カーネル法に基づく関数出力回帰モデルを提案した。

共同研究の提案等は大歓迎!

コード: https://github.com/Minoru938/KRFD

連絡先: kusaba.minoru@nifs.ac.jp