



統計的機械学習による新材料創製 産学連携の現状と可能性

吉田 亮^{1,2,3} yoshidar@ism.ac.jp

¹ 情報・システム研究機構 統計数理研究所 データ科学研究系 教授

² 情報・システム研究機構 統計数理研究所 ものづくりデータ科学研究センター センター長

³ 総合研究大学院大学 複合科学研究科 統計科学専攻 教授

ものづくりデータ科学研究センター

2017年7月設立：産学共創の新しいモデルケースを探索・実践



重点分野：マテリアルズインフォマティクス

教授1名，准教授1名，特任研究員等7名（内2名他大学との兼任）



共同研究・学術指導：2021年度 12企業

材料データベース共同開発：2021年度 3大学・21企業 参画者76名



共同研究部門



三菱ケミカル株式会社

ISM-MCCフロンティア材料設計拠点 (2019.10-)



JSR 株式会社

JSR-ISM スマートケミストリーラボ (2020.10-)

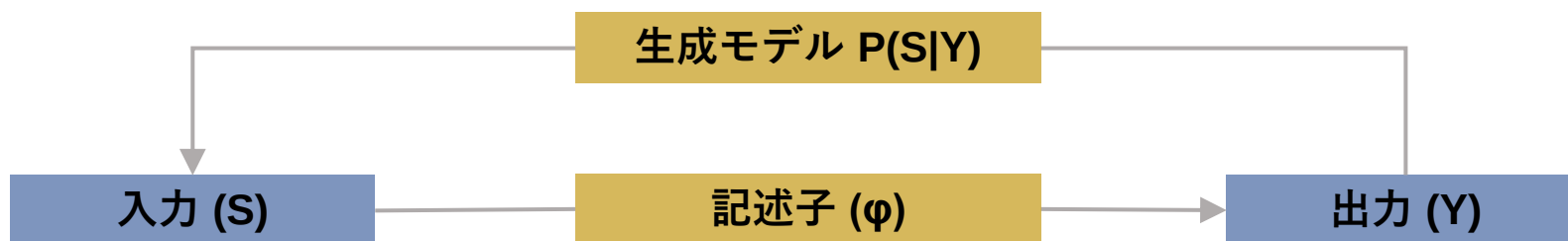
広大な未踏領域から所望の特性
を有する「物質・材料」を発掘

例：有機分子の空間

= span { 10^{60} 個の候補分子 }

マテリアルズインフォマティクスの順問題と逆問題

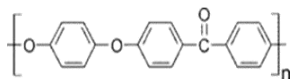
逆問題：所望の Y^* を持つ X を予測（設計・制御） $X = f^{-1}(Y^*)$



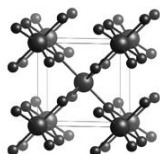
順問題： $Y = f(X) = f(\varphi(X))$ を推定



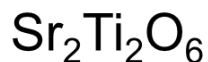
Chemical structure



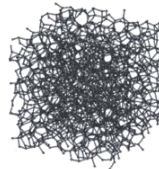
Crystal



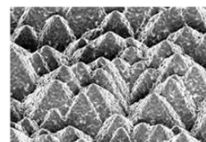
Composition



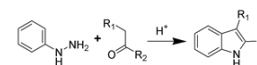
Amorphous



Microstructure

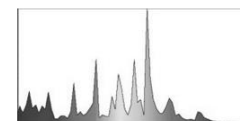


Chemical reaction



Θ, T, P

Spectrum





ハイパーマテリアル
補空間が創る新物質科学

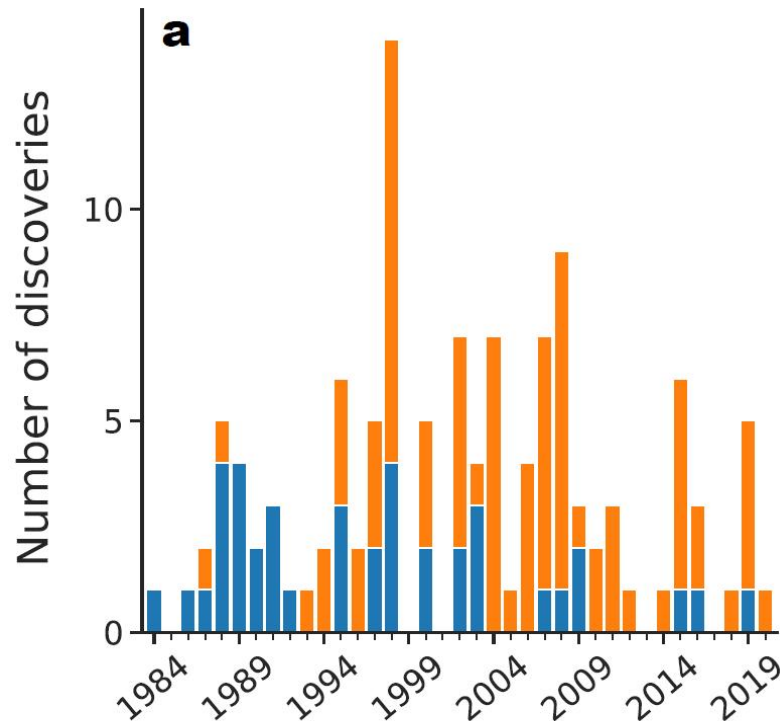
マテリアルズインフォマティクスの適用事例

機械学習による新しい準結晶の発見

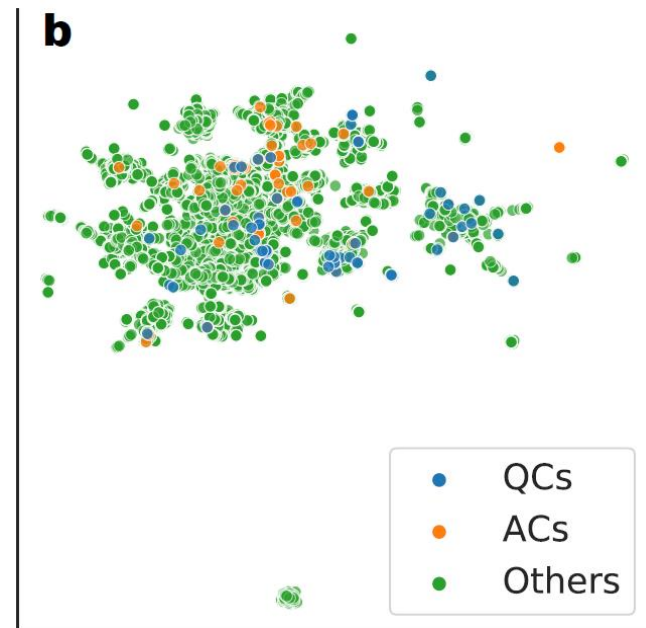
歴史：準結晶の発見

- 1984: Dan Shechtman博士（2011:ノーベル化学賞）による最初の準結晶の発見
- これまでにおよそ100個の準結晶が見つかった。
- 新しい物理現象の発見：電子特性の異常・量子臨界・超伝導など
- 発見のペースが鈍化：約40年の歴史の中で明確な設計原理を確立できなかった。

AIを含む準結晶と近似結晶の発見数

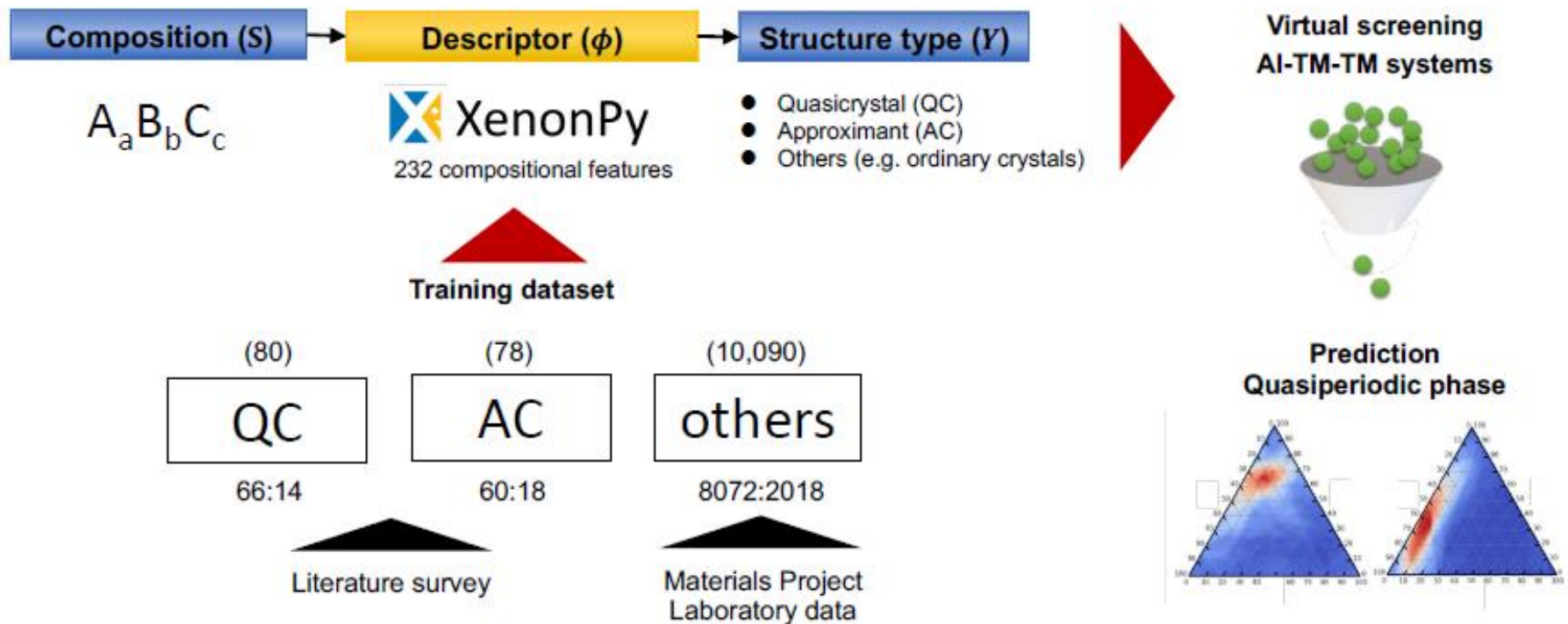


準結晶・近似結晶・結晶の組成分布



機械学習で準結晶の発見を大幅に加速する

化学組成から準結晶・近似結晶・通常の結晶という3クラスのカテゴリを構築する



Liu et al. Machine learning to predict quasicrystals from chemical compositions. *Advanced Materials*. 33(36):e2102507 (2021)

化学組成記述子の設計

Yamada et al. ACS Cent Sci (2019); Liu et al. Adv Mater. 33(36):e2102507 (2021)

S: 化学組成

記述子ベクトルの要素 i



$$\phi_i(S) = f(a, b, c, d, e, \eta_{\text{Ni}}^i, \eta_{\text{Ti}}^i, \eta_{\text{Cu}}^i, \eta_{\text{Fe}}^i, \eta_{\text{Pd}}^i)$$

組成比

元素の特徴量

原子半径, 電気陰性度など

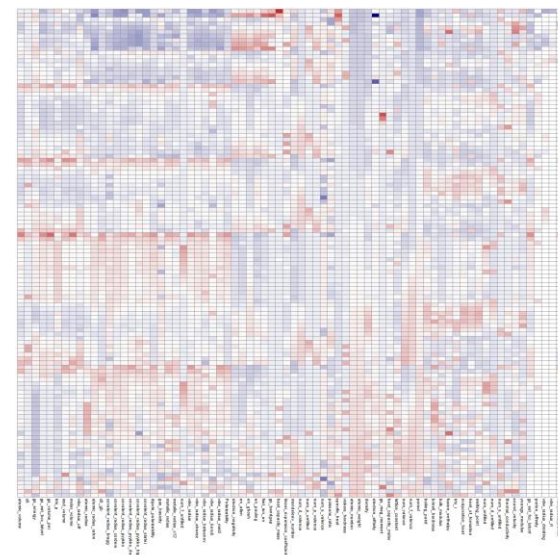
$f(\cdot)$: 加重平均, 加重分散, 最大・最小プーリングなど

58種類の元素記述子

Table S2 in Liu et al. Adv Mater. 33(36):e2102507 (2021)

Feature ID	Description	Unit
atomic_volume	Atomic volume	cm ³ /mol
atomic_weight	Atomic weight	
boiling_point	Boiling temperature	K
bulk_modulus	Bulk modulus	GPa
c6_gb	C_6 dispersion coefficient in a.u. (Gould & Bučko)	a.u.
covalent_radius_cordero	Covalent radius by Cordero et al.	pm
covalent_radius_pyykko	Single bond covalent radius by Pyykko et al.	pm
covalent_radius_pyykko_double	Double bond covalent radius by Pyykko et al.	pm
covalent_radius_pyykko_triple	Triple bond covalent radius by Pyykko et al.	pm
covalent_radius_slater	Covalent radius by Slater	pm
density	Density at 295K	g/cm ³
dipole_polarizability	Dipole polarizability	a.u.
electron_negativity	Pauling electronegativity	
electron_affinity	Electron affinity	eV
en_allen	Allen's scale of electronegativity	eV
en_ghosh	Ghosh's scale of electronegativity	
en_pauling	Pauling's scale of electronegativity	
first_ion_en	First ionisation energy	eV
fusion_enthalpy	Enthalpy of fusion for elements at their melting temperatures	kJ/mol
gs_bandgap	DFT bandgap energy of T=0K ground state	eV
gs_energy	DFT energy per atom (raw VASP value) of T=0K ground state	eV/atom
gs_est_bcc_latent	Estimated BCC lattice parameter based on the DFT volume of the OQMD ground state	

Column: descriptors; Row: elements



References

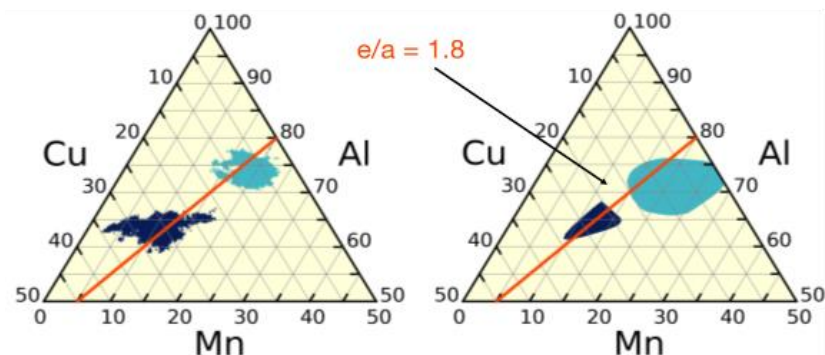
- (1) pymatgen: <http://pymatgen.org/>
- (2) matminer: <https://pypi.python.org/pypi/matminer/>
- (3) CRC handbook

機械学習による準結晶の予測：実証に成功

3元系アルミニウム合金の相図

90% 以上の予測精度

機械学習による予測相図



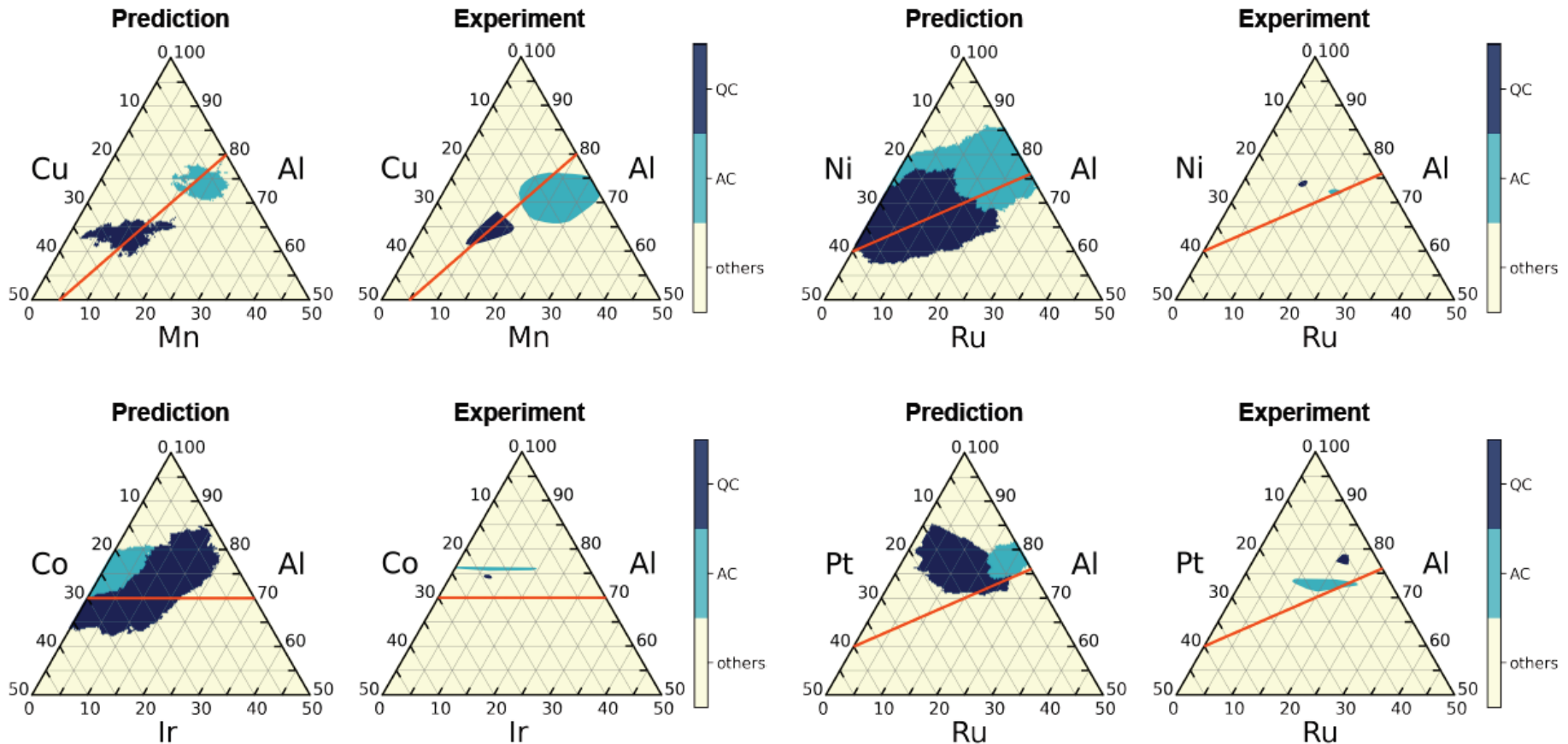
データが存在しないにも関わらず、準結晶相の同定に成功している系が多数見つかった

Liu et al, Advanced Materials 33(36):2170284 (2021).

モデルはヒューム=ロザーリの電子濃度則を学習

Tsai, Sci Technol Adv Mater (2008)

- 経験則: アルミニウム合金: 平均遍歴価電子数 e/a が 1.8 の付近に準結晶・近似結晶相が出現
- 予測された準結晶・近似結晶の領域も $e/a = 1.8$ の付近に出現



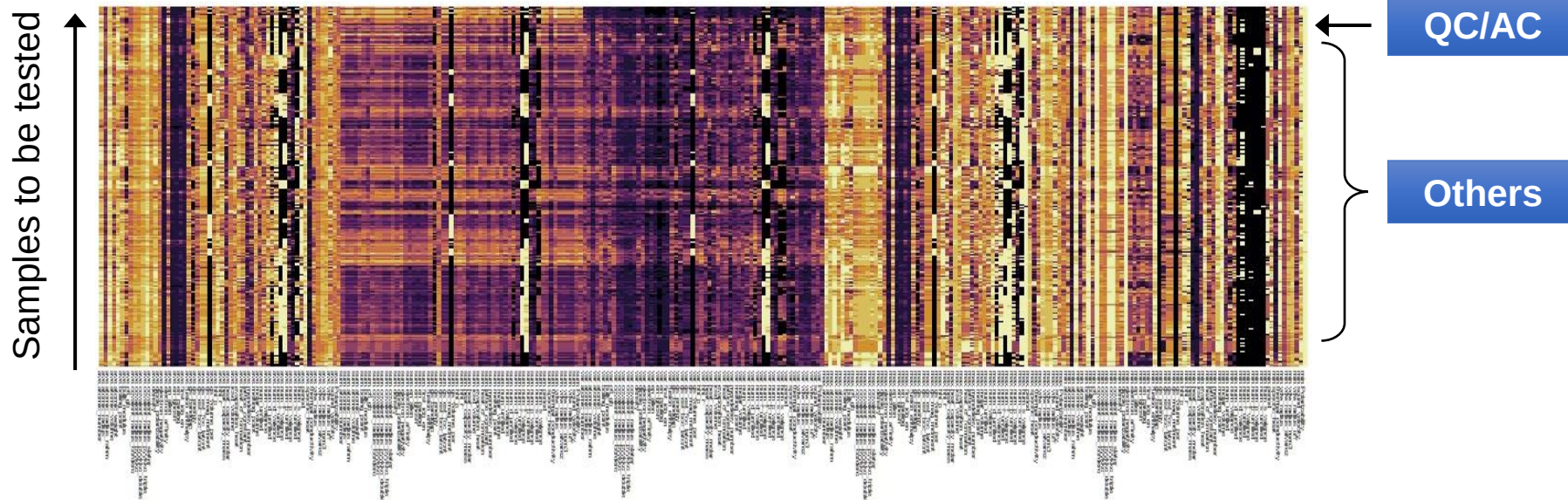
準結晶の形成に関する新しい法則を発見する

問. モデルは何を根拠に準結晶相の有無を判定しているのか？（言語化された知識）

まずは、クラス分布の重なりが小さい特徴量を同定 (MIC: Reshef et al. Science. 2011)

- Hypermaterials vs Others
- Quasicrystals vs Approximants

 XenonPy 232 compositional descriptors



重要特徴量の同定と「知識」の言語化

例1：元素の電気陰性度の加重平均が 0.15 ± 0.01 であることが，QC/ACの必要条件

例2：p軌道の平均価電子数が 0.71 ± 0.06 であることが，QC/ACの必要条件

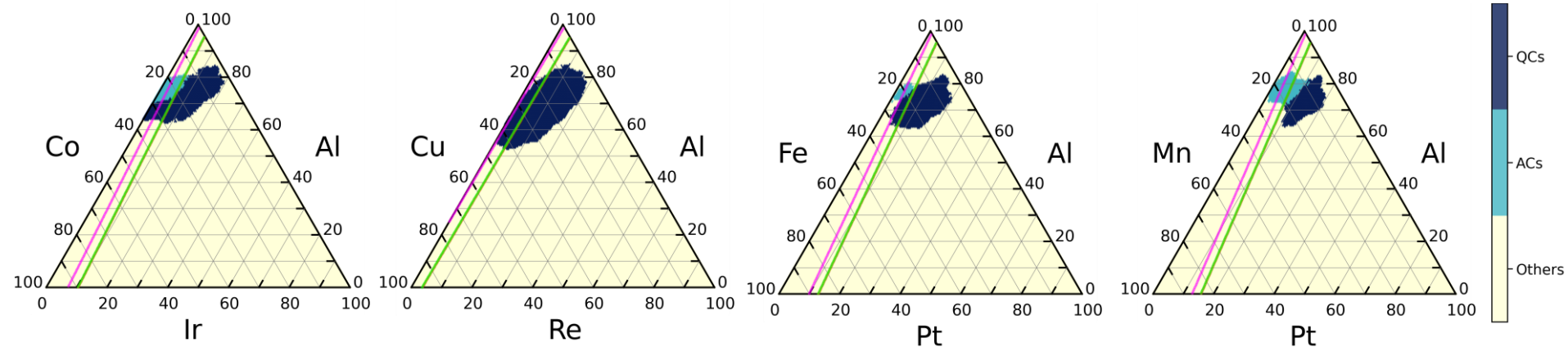
Descriptor ID	Description (unit)	MIC	QC	AC	Others
ave:vdw_radius_uff	Van der Waals radius from the UFF (pm)	0.43	409.05 (3.37)	406.49 (6.81)	382.297 (40.660)
ave:en_ghosh	Ghosh's scale of electronegativity	0.42	0.15 (0.00)	0.15 (0.01)	0.158 (0.021)
ave:first_ion_en	First ionisation energy (eV)	0.41	6.49 (0.09)	6.53 (0.17)	6.843 (0.580)
ave:mendeleev_number	Mendeleev's number	0.41	75.94 (0.36)	75.86 (1.47)	73.323 (4.326)
ave:specific_heat	Specific heat at 20 °C (J/(g mol))	0.40	0.74 (0.02)	0.73 (0.04)	0.658 (0.151)
ave:num_p_valence	Number of filled p valence orbitals	0.40	0.71 (0.06)	0.73 (0.05)	0.569 (0.247)
ave:num_p_unfilled	Number of unfilled p valence orbitals	0.40	3.53 (0.30)	3.63 (0.24)	2.846 (1.237)
ave:heat_capacity_mass	Specific heat capacity at STP ¹ (J/mol-K)	0.40	0.74 (0.02)	0.73 (0.04)	0.655 (0.150)
ave:covalent_radius_cordero	Covalent radius by Cordero et al. (pm)	0.39	126.06 (1.19)	126.38 (2.13)	129.510 (6.311)
ave:vdw_radius	Van der Waals radius (pm)	0.37	189.51 (0.55)	190.67 (1.81)	193.799 (6.328)
ave:gs_energy	DFT energy per atom (raw VASP value) of T=0K ground state (eV/atom)	0.37	-4.57 (0.18)	-4.69 (0.28)	-5.192 (1.118)
ave:thermal_conductivity	Thermal conductivity at 25 °C (W/(m K))	0.36	221.35 (21.74)	201.23 (13.96)	170.722 (60.677)
ave:covalent_radius_slater	Covalent radius by Slater (pm)	0.35	127.92 (1.19)	128.08 (0.93)	130.396 (3.376)
ave:period	Period in periodic table	0.35	3.40 (0.06)	3.52 (0.19)	3.725 (0.553)
var:num_p_valence	Number of filled p valence orbitals (pm)	0.34	0.20 (0.02)	0.20 (0.02)	0.184 (0.071)
ave:num_d_valence	Number of filled d valence orbitals (pm)	0.34	2.30 (0.60)	2.08 (0.52)	3.149 (1.901)
ave:heat_capacity_molar	Molar heat capacity at STP (J/mol-K)	0.34	24.44 (0.10)	24.47 (0.14)	24.807 (0.580)
ave:density	Density at 295K (g/cm ³)	0.34	4.94 (0.32)	5.55 (1.24)	6.700 (3.347)
var:num_p_unfilled	Number of unfilled p valence orbitals	0.34	5.09 (0.60)	4.91 (0.48)	4.602 (1.767)
ave:hhi_p	HerfindahlHirschman Index (HHI) production values	0.33	1810.99 (242.60)	2106.51 (274.67)	2196.878 (706.345)

発見された法則①：平均電気陰性度

経験式 = 組成記述子の目標値を満たす組成比の部分空間

$$\mathcal{S} = \{(c_1, c_2, c_3) \mid c_1 \eta_1^{\text{en}} + c_2 \eta_2^{\text{en}} + c_3 \eta_3^{\text{en}} \simeq 0.15\}$$

ほぼ全てのAl-TM-TMにおいて，経験式の直線は予測されたQC/AC相にオーバーラップ



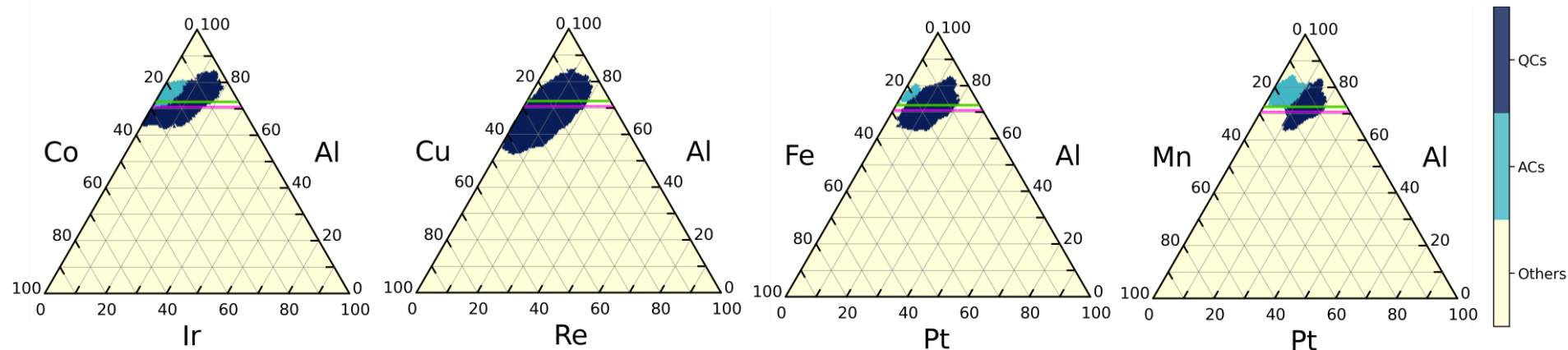
Pink: target = Mean QC; Green: Mean QC/AC

発見された法則②：p軌道の平均価電子数

経験式 = 組成記述子の目標値を満たす組成比の部分空間

$$\mathcal{S} = \{(c_1, c_2, c_3) \mid c_1\eta_1^p + c_2\eta_2^p + c_3\eta_3^p \simeq 0.71\}$$

ほぼ全てのAl-TM-TMにおいて，経験式の直線は予測されたQC/AC相にオーバーラップ



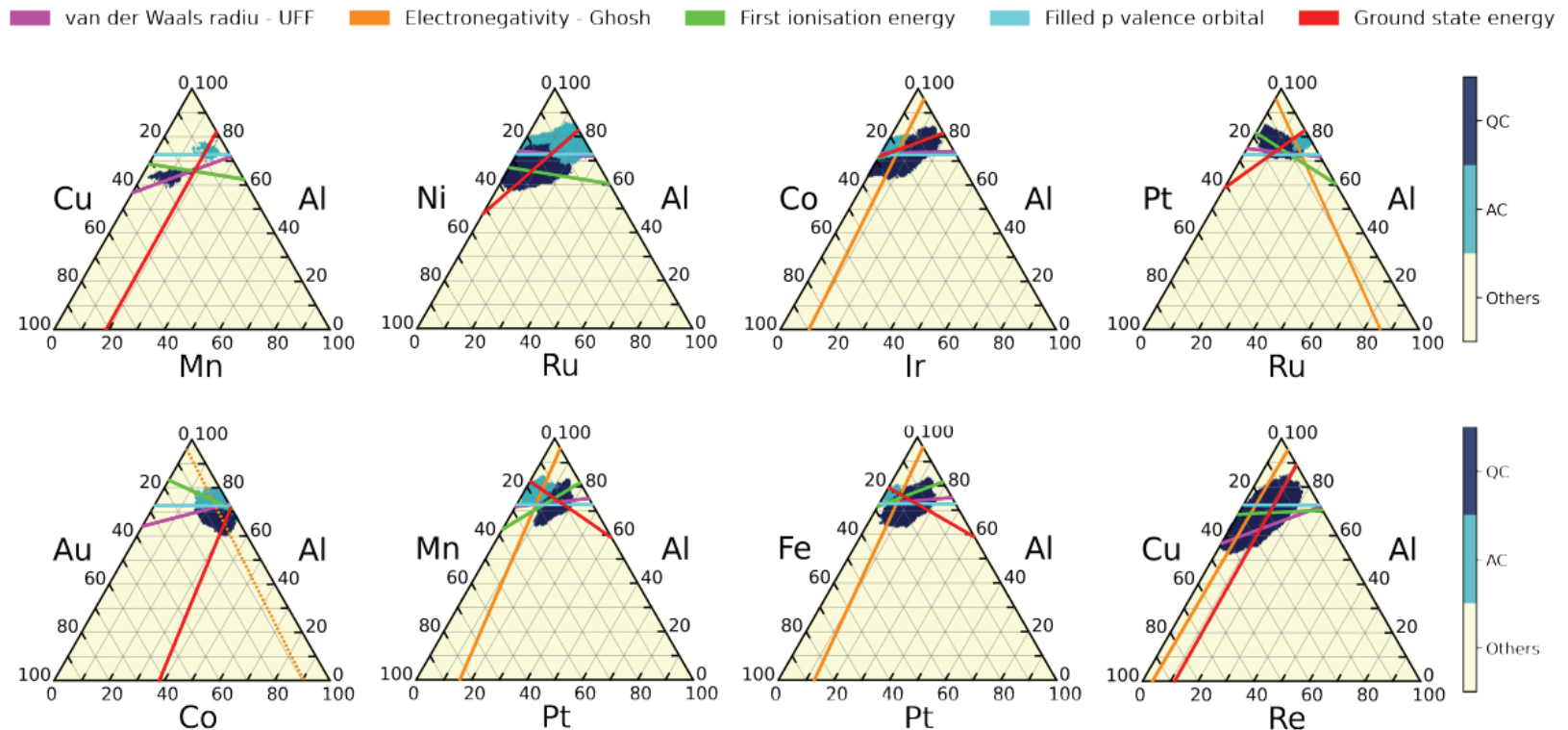
Pink: target = Mean QC; Green: Mean QC/AC

機械学習による隠れた法則の発見

準結晶の形成・安定化に必要な化学組成のルール（五つの数式）を発見

構成元素のファンデルワールス半径や電気陰性度などの分布が満たすべき条件を規定する。

$$\mathcal{S} = \{(c_1, c_2, c_3) | \phi(c_1, c_2, c_3, \eta_1, \eta_2, \eta_3) = \text{target}\}$$



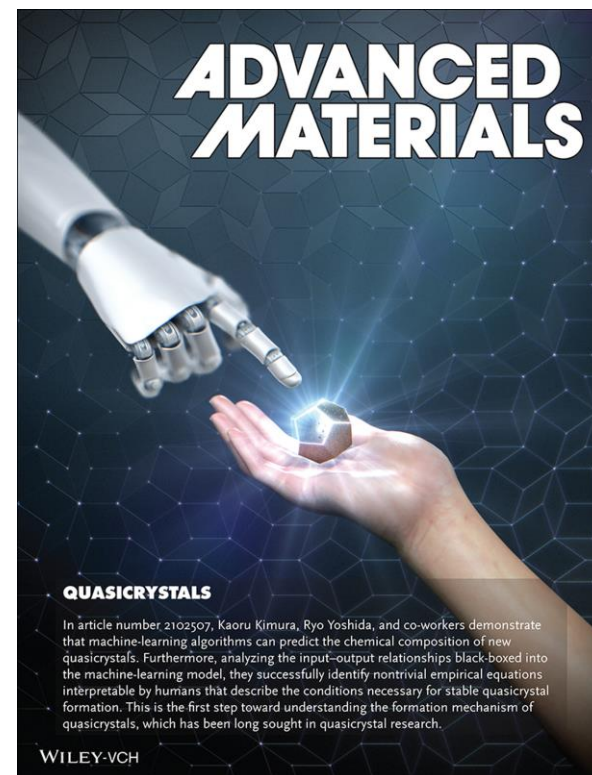
データ科学による「新物質」と「新法則」の発見

準結晶研究の未踏領域を開拓

- 予測能力を実証
- 3個の新規準結晶を発見(東大 木村, 東京理科大 田村)

半導体準結晶, 超伝導準結晶の発見へ

- 現在探索中



Frontpage in Advanced Materials

Liu et al. Machine learning to predict quasicrystals from chemical compositions. Advanced Materials. 33(36):e2102507 (2021)

ものづくりデータ科学研究センター：設立から4年半

産学の価値共創の下，様々な未踏領域を開拓・実証

組成・プロセス

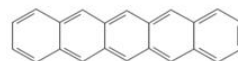


微細組織

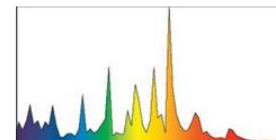


Iwayama et al. (2021) in preparation

化学構造



スペクトル

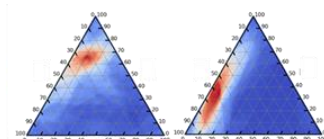


Iwayama et al. (2021) in preparation

化学組成



準結晶

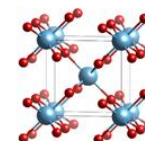


Liu et al. Adv. Mater. 33(36):e2102507 (2021)

化学組成

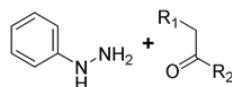


結晶構造・対称性

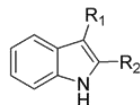


Yamashita et al. Phys Rev Mater 2:013803 (2018)
Yamada et al. ACS Cent. Sci. 5(10):1717-1730 (2019)
Kusaba et al. in preparation
Liu et al. in preparation

反応物

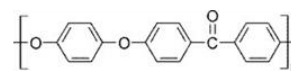


生成物



Guo et al. J. Chem. Inf. Model. 60(10):4474-4486 (2020)
Zhang et al. (2021) in preparation

結晶・高分子



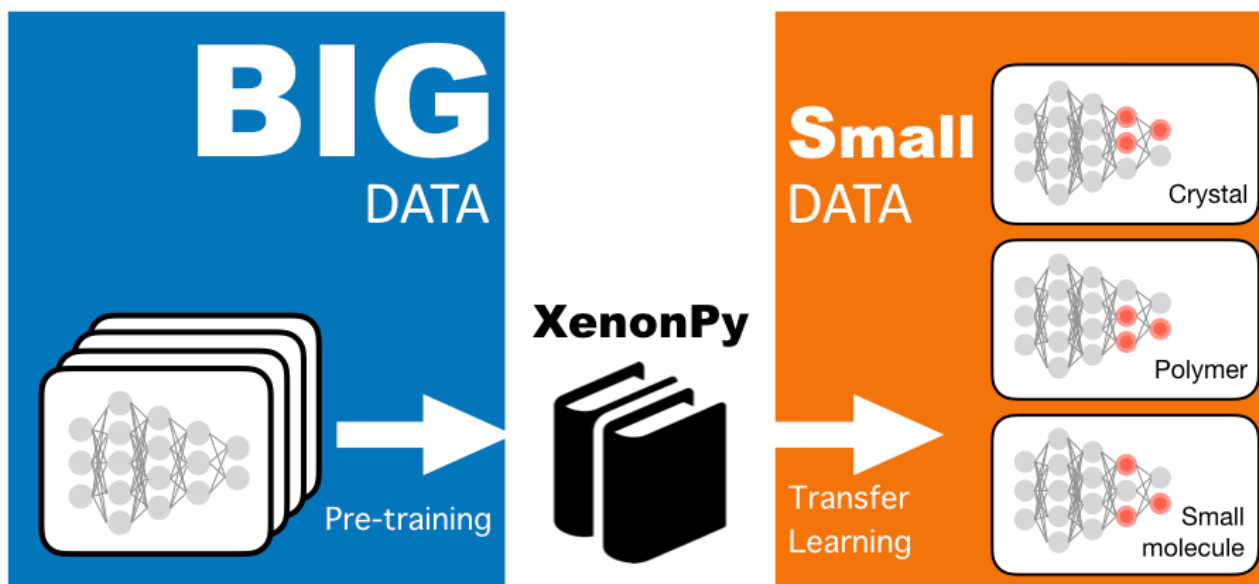
物性

T_g, T_m, C_p

Yamada et al. ACS Cent. Sci. 5(10):1717-1730 (2019)
Ju et al. Phys. Rev. Mater. 5:053801 (2021)
Minami et al. AAI (2021) in press
Ikebata et al. J. Comput. Aided Mol. Des. 31:379-391 (2017)
Wu et al. npj Comput. Mater. 5:66 (2019)
Wu et al. Mol Inform. 39:1-2 (2020)

限られたデータの壁を乗り越える

データ科学の真価を発揮できない数多くの領域を切り拓く



データ駆動型研究に資するデータが不足

短中期的には、大学の研究室や一企業で生産可能なデータが標準的に

コスト

実験・シミュレーション・試料作製・物性評価に要するコストが高い。

ニーズ 多様性

研究者の興味や設計変数（材料種、試料の作製方法など）が多様
コモンデータを創出しようという動きが起きにくい。

インセン ティブ欠如

基礎と応用の垣根が低い。
競合相手に対する情報秘匿の意識が高い。
データを公開するインセンティブが研究者に働きにくい。

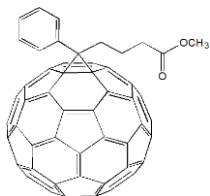
そもそも‘革新的な’材料の周辺にはデータが存在しない

例：有機薄膜太陽電池のドナー分子

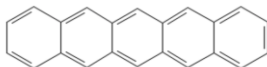
Acceptor = PCBM

Donners = design targets

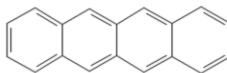
PCBM



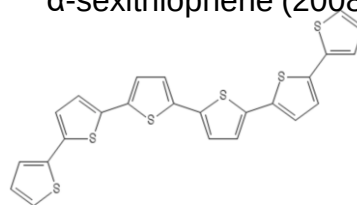
pentacene (2004)



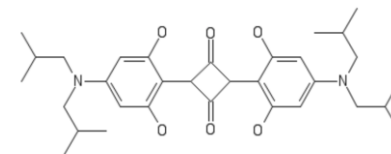
tetracene (2005)



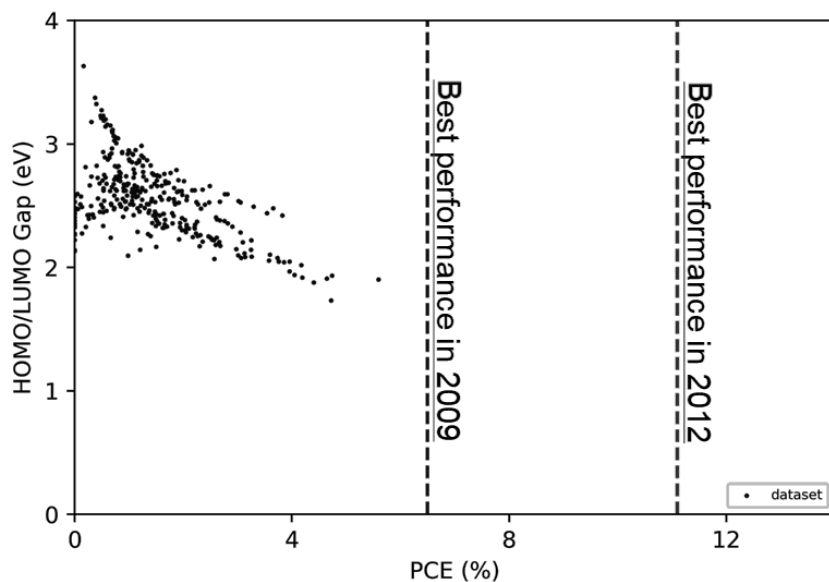
α -sexithiophene (2008)



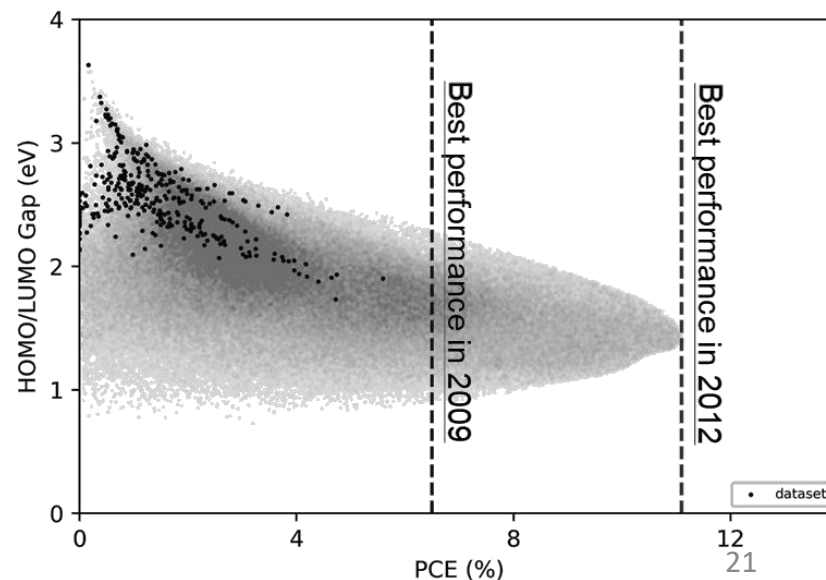
2,4-Bis[4-(*N,N*-diisobutylamino)-2,6-dihydroxyphenyl] squaraine (2009)



2009年の分布



2012年の分布



2018年頃～ 転移学習による解決方策を実践

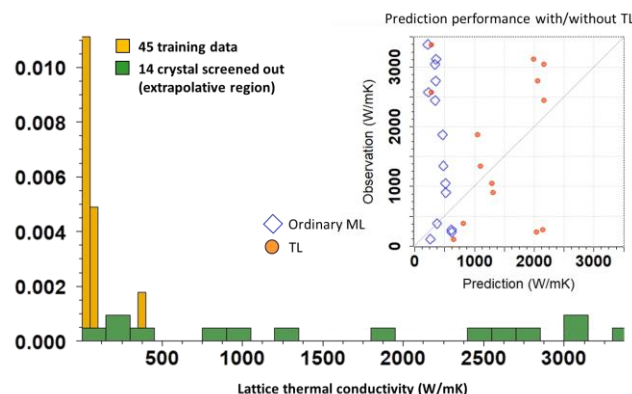
- あるドメイン（元ドメイン）の学習モデルやデータを目標ドメインに利活用するための方法論
- データ量が少なくてフルスクラッチでの学習は難しいが、元ドメインのデータやモデルを巧みに利用することでデータ量の不足を補うことができる。

- 転移モデルが有する外挿性を発見
- 訓練済みモデルデータベース

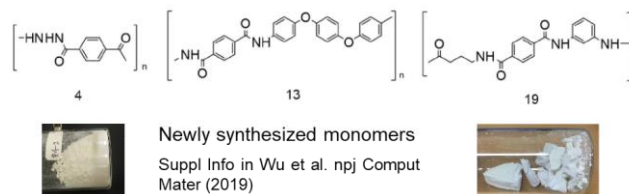
- 高熱伝導ポリマーの合成
- 28点のデータから予測モデル

- 超高熱伝導を持つ無機結晶
- 外挿性の実証

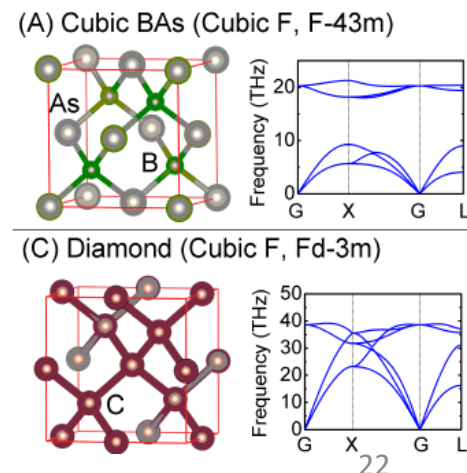
Yamada et al. *ACS Cent Sci* 5:1717-1730 (2019)



Wu et al. *npj Comput Mater* 5:66 (2019)



Ju et al. *Phys Rev Materials* 5, 053801 (2021)



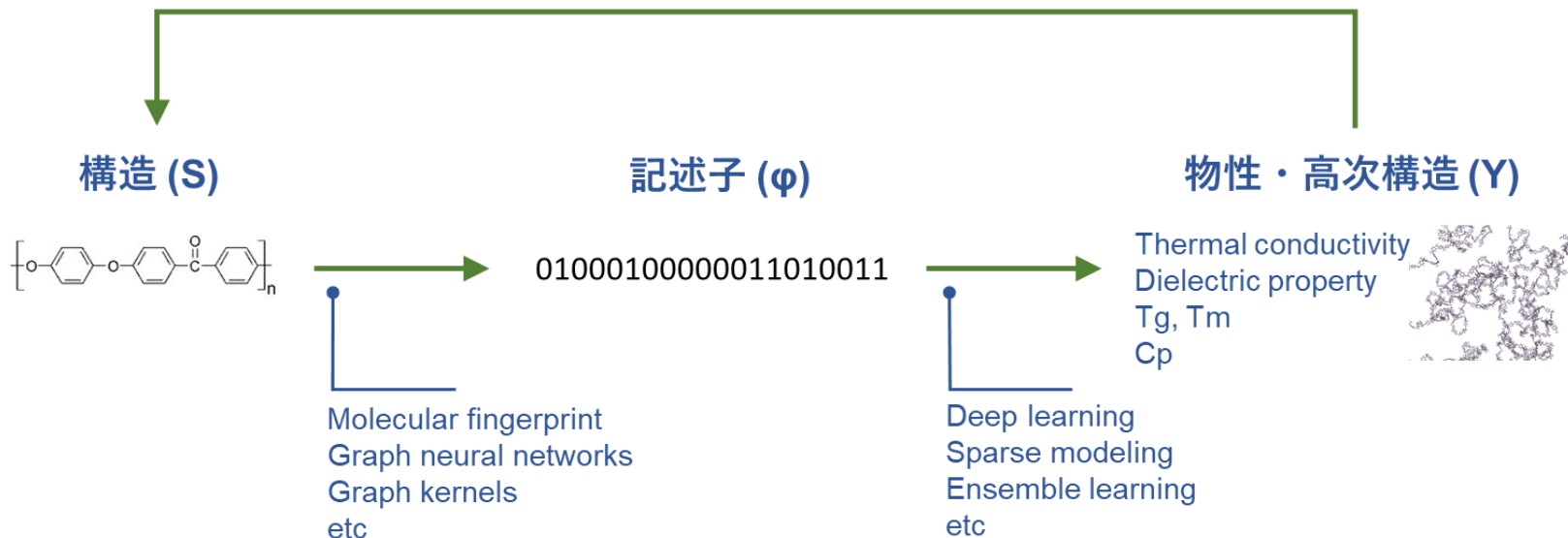
高分子物性データベースの共創

広大なマテリアルスペースの地図を作成する

所望の特性を有する高分子材料の予測・発見

高分子材料インフォマティクスの実践・実証

逆問題：所望の特性 Y^* を持つ構造 S を予測 $S = f^{-1}(Y^*)$



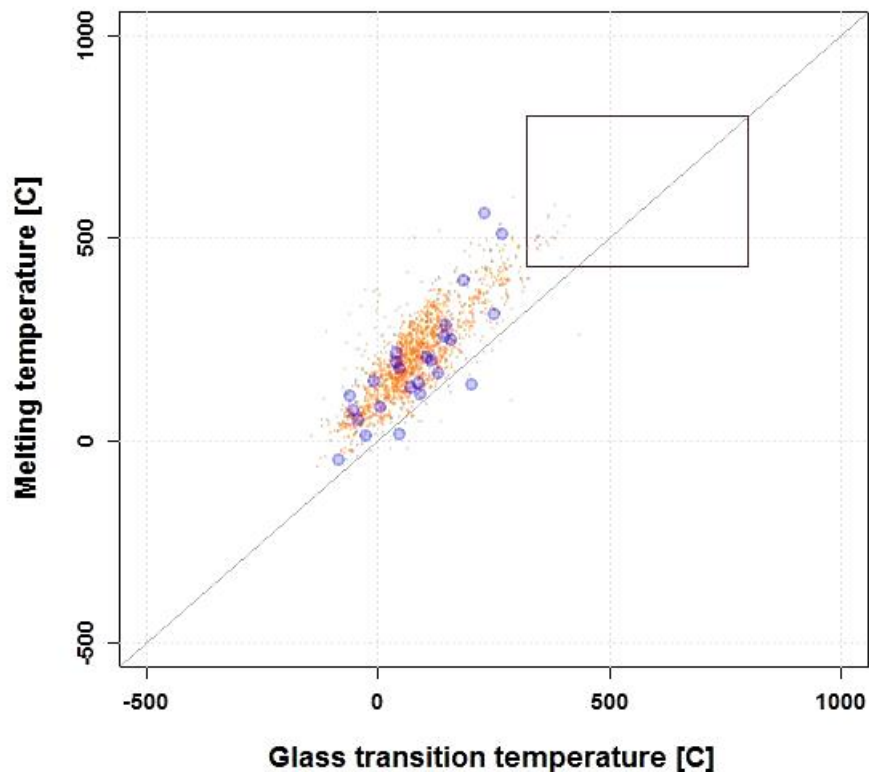
順問題：(Y, S)のデータを用いて $Y = f(S)$ を訓練

2017-高分子材料インフォマティクスに関する発表論文

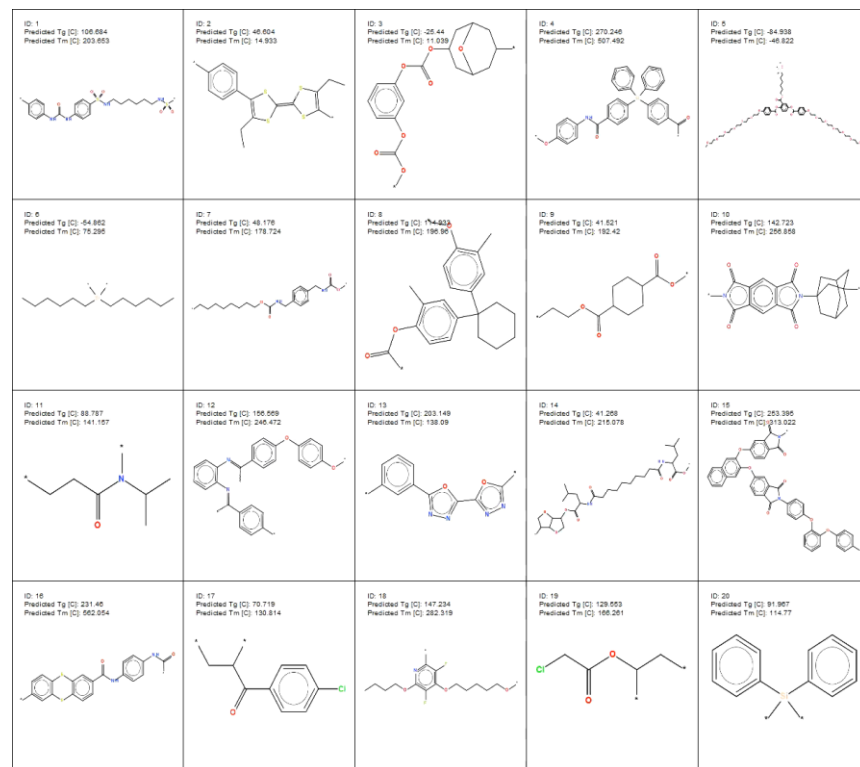
- Guo et al. Bayesian algorithm for retrosynthesis. J Chem In. Model. 60:4474–4486 (2020)
- Wu et al. iQSPR in XenonPy: a Bayesian inverse molecular design algorithm. Mol Inform. 39:1900107 (2019)
- Yamada et al. Predicting materials properties with little data using shotgun transfer learning. ACS Cent Sci. 5:1717-1730 (2019)
- Wu et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. npj Comput Mater. 5:66 (2019)
- Ikebata et al. Bayesian molecular design with a chemical language model. J Comput Aided Mol Des. 31:379-391 (2017)

逆問題：機械学習で所望の特性を持つ高分子を設計する

目標物性：ガラス転移温度 ↗ ・ 融点 ↗



モノマーの化学構造を設計
ケミカルスペースを自由自在に走査



- Ikebata et al. Bayesian molecular design with a chemical language model. *J Comput Aided Mol Des.* 31:379-391 (2017)
- Wu et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput Mater.* 5:66 (2019)
- Wu et al. iQSPR in XenonPy: a Bayesian inverse molecular design algorithm. *Mol Inform.* 39:1-2 (2020)

データベースの不在が続く高分子材料インフォマティクス

無機材料では第一原理計算データベースがMIの発展と普及を牽引

高分子物性データベース

■ PoLyInfo (literature survey)

~100 properties for 18,015 polymers

NO API

<https://polymer.nims.go.jp/>

■ Polymer Genome (DFT)

~10 properties for ~800 polymers

NO API

<https://www.polymergenome.org/?m=home>

■ CROW (literature survey)

NO API

<http://www.polymerdatabase.com/>

■ CRIPT (not yet open)

MIT + Citrine Informatics

<https://cript.mit.edu/>

無機化合物のデータベース

■ Materials Project (DFT)

~130,000 virtual and real compounds

Accessible via API

<https://materialsproject.org/>

■ AFLOW (DFT)

~3,000,000 virtual and real compounds

Accessible via API

<http://www.aflowlib.org/>

■ OQMD (DFT)

~600,000 virtual and real compounds

Accessible via API

<http://oqmd.org/>

■ ICSD (experiment)

~240,000 real compounds

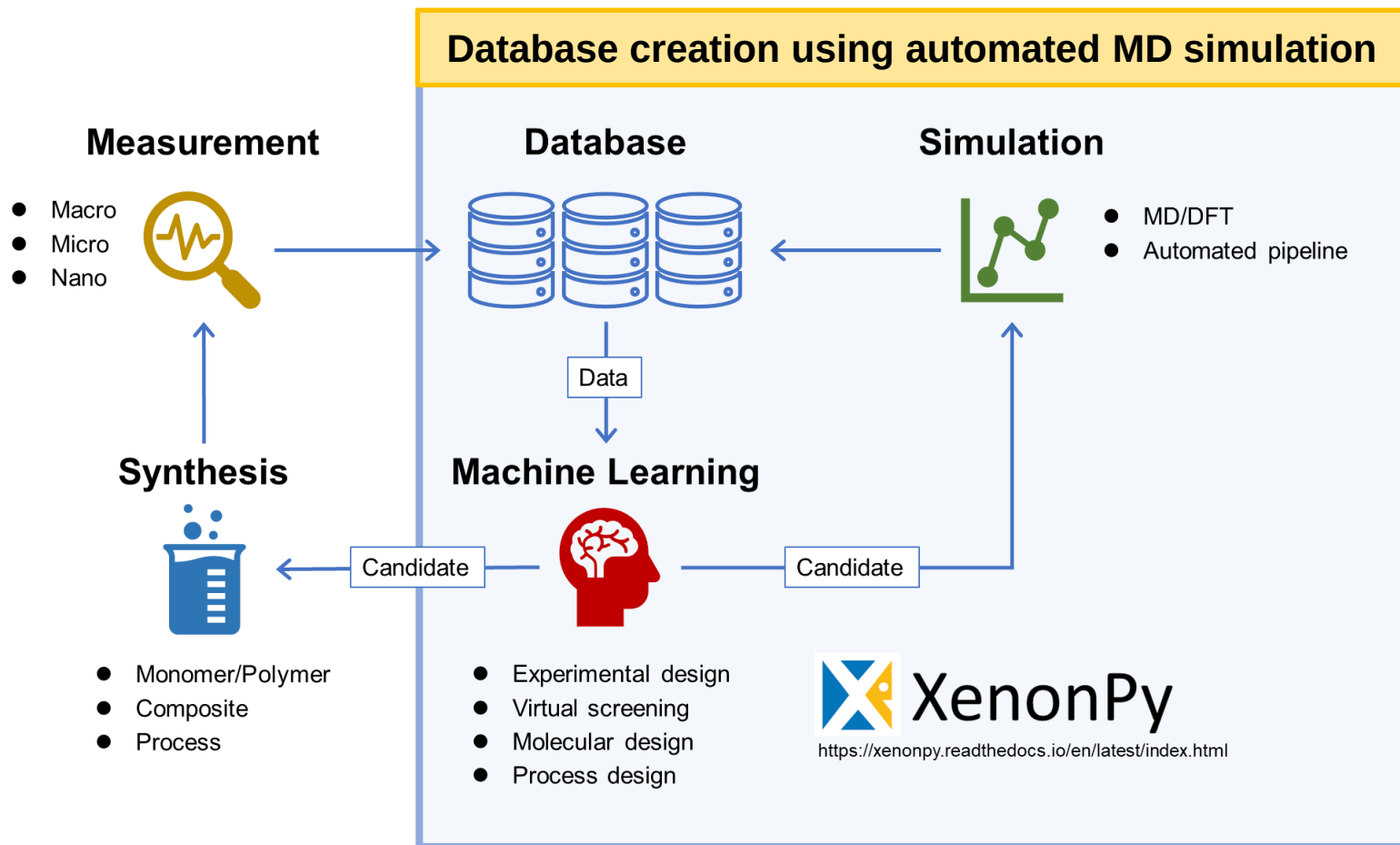
Accessible via API

<https://icsd.products.fiz-karlsruhe.de/>

高分子材料インフォマティクスの学術基盤の構築

JST-CREST 熱制御「高分子の熱物性マテリアルズインフォマティクス」(2019.10-) 代表：森川淳子（東工大）

機械学習・計算科学グループ：吉田亮・林慶浩・Wu Stephen・野口瑤・Chang Liu・塩見淳一郎・川内進





全原子分子動力学シミュレーションによる高分子物性自動計算

RadonPy

最新版：14物性の自動計算に対応

(2021/09/16)

- ◆ Thermal conductivity
- ◆ Thermal diffusivity
- ◆ Density
- ◆ Radius of gyration
- ◆ Specific heat capacity C_p
- ◆ Specific heat capacity C_v
- ◆ Compressibility (isothermal)
- ◆ Isentropic compressibility
- ◆ Bulk modulus (isothermal)
- ◆ Isentropic bulk modulus
- ◆ Self-diffusion coefficient
- ◆ Thermal expansion coefficient
- ◆ Linear expansion coefficient
- ◆ Dielectric constant (static)

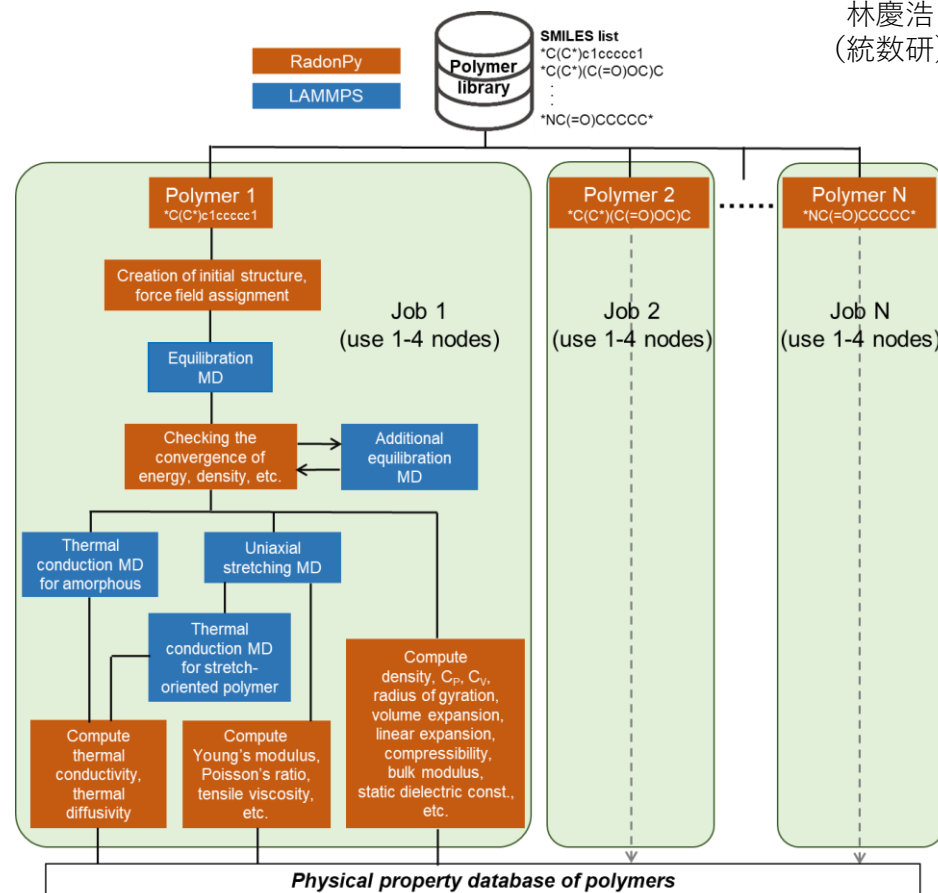
- 力場の割り当て、初期構造生成、エラー処理、平衡・非平衡MDによる物性評価までの全工程を完全に自動化
- 高分子集合系、高分子液晶、架橋高分子
- 最新版：アモルファスポリマーや延伸配向したポリマーの自動計算に対応

RadonPy (OSS)

古典分子動力学計算ソフトウェアLAMMPSによる高分子物性計算を自動化するPythonライブラリ



林慶浩
(統数研)



包括的データベースの創出：計算資源の確保が鍵

小規模な1研究グループには難しい

2020年度までは自然科学研究機構 分子科学研究所などの計算資源を活用
延伸配向した2,000ポリマーの計算 = ABCIで換算すると2,500万円分

	自然科学研究機構 計算科学研究センター NEC LX 2U-Twin2 サーバ406Rh-2	産業総合研究所 ABCI	東京大学物性研究所 Ohtaka
ノード数 / ポリマー	1	1	1
CPU / ノード	12 core	4 core	128 core
GPU / ノード	1 GPU	1 GPU	-
メモリ / ノード	192 GB	60 GB	256 GB
実行時間 / ポリマー (アモルファス構造)	30-50 時間	32-50 時間	36-44 時間
実行時間 / ポリマー (延伸配向構造)	+ 約100 時間	+ 約100 時間	+ 約100 時間

産学連合体による高分子物性オープンデータベースの共創

2021- 統計数理研究所 ものづくりデータ科学研究センター

10⁵-10⁷ポリマーの包含する世界初の体系的な高分子物性データベースの創出

- 統数研・3大学・21企業が組織の垣根を越えてデータを共同生産・共有
- 「富岳」成果創出加速プログラム「データ駆動型高分子材料研究を変革するデータ基盤創出」（吉田）

計算候補ポリマーライブラリ

- ◆ 市販ポリマー・公共データベースのポリマー
- ◆ 機械学習で生成した仮想ポリマー



機械学習（実験計画法）



コントリ
ビュータ1

コントリ
ビュータ2

コントリ
ビュータ3

1国研・3大学・21企業（R3年
度）によるデータの共同生産



包括的な高分子物性データベース

- ◆ 10万種類以上の高分子骨格
- ◆ 高分子集合系，高分子液晶，架橋高分子
- ◆ 熱物性・力学的特性など



RadonPy

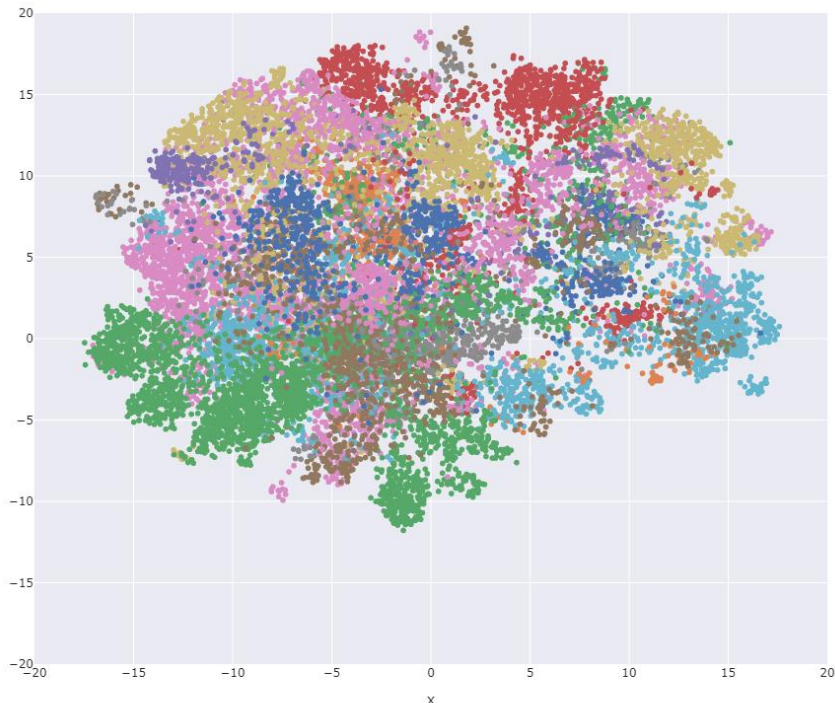
高分子物性の大地図を作成

10⁵-10⁷個の高分子が張るケミカルスペースの全体像を明らかにする

- 複数物性の同時分布
- パレートフロンティアの位置・構造的特徴
- 特異な特性を持つ高分子を発見

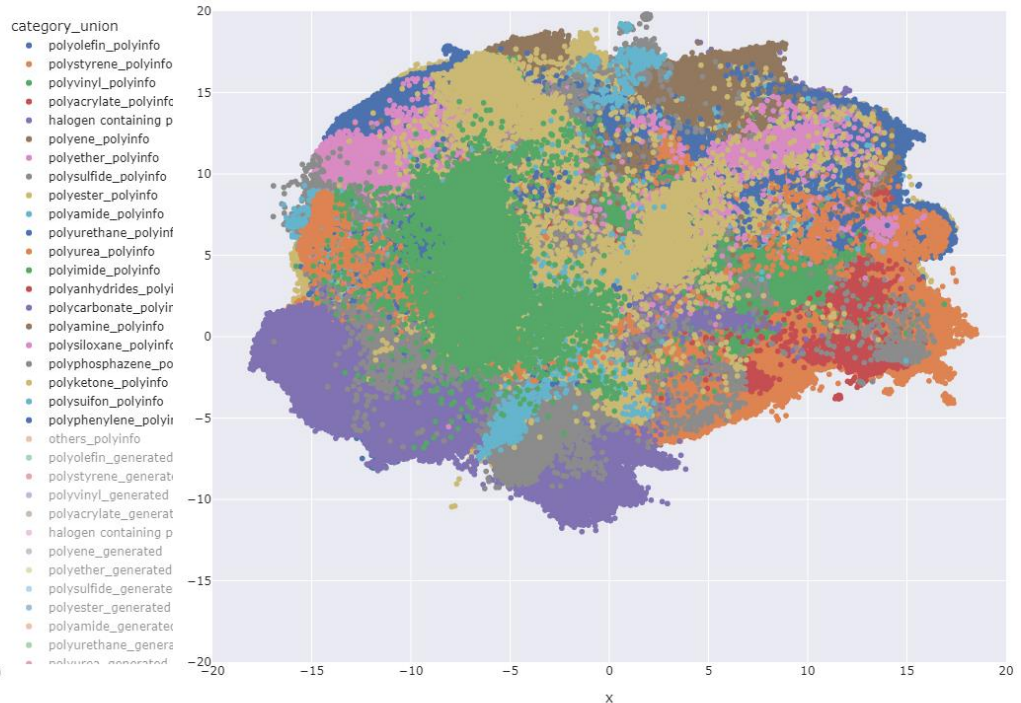
これまでに合成されたポリマーの分布 (UMAP)

PoLyInfo の 21クラス分類



機械学習※で構築した仮想ポリマーの分布

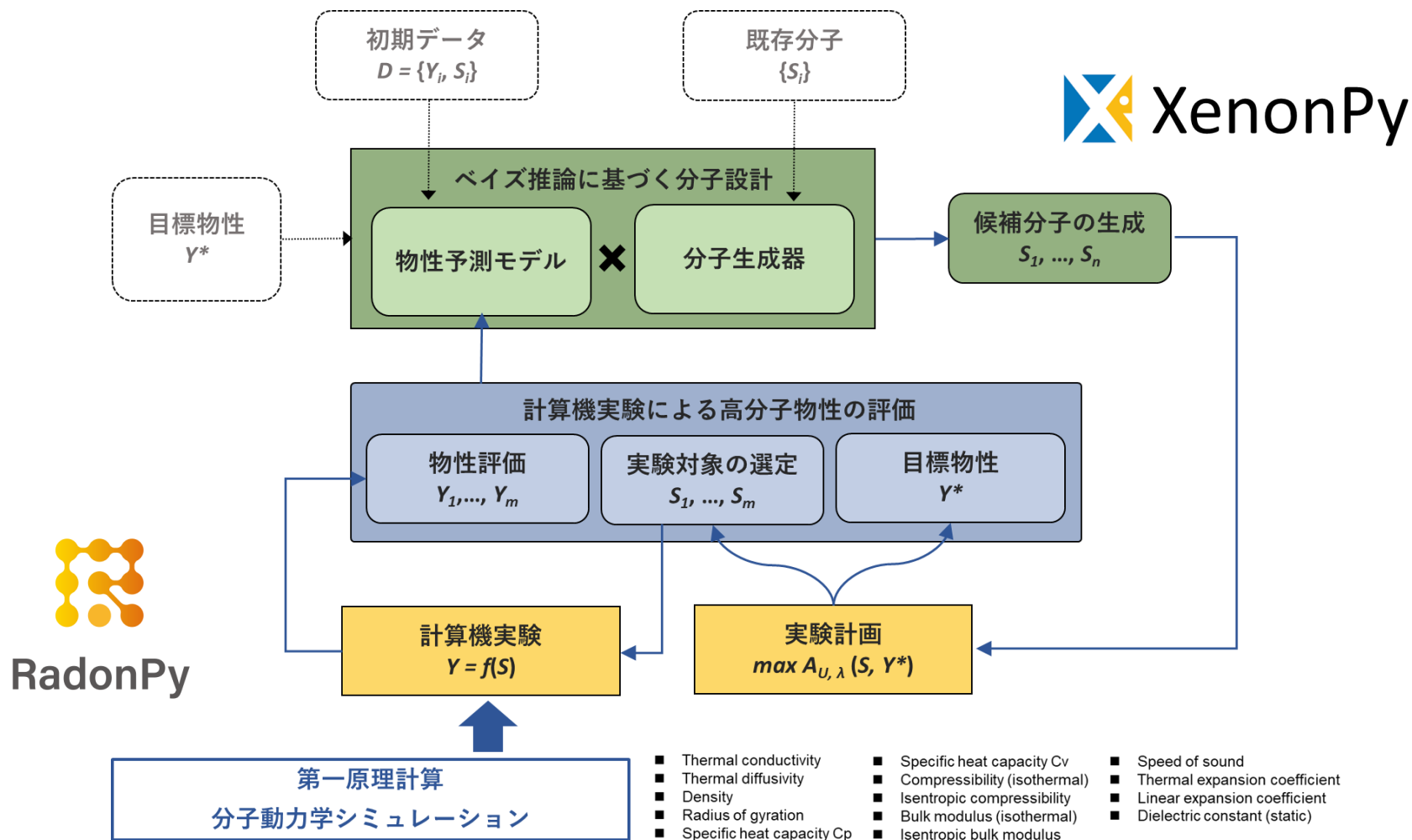
Ikebata et al. J Compt Aided Mol Des. 31:379-391 (2017)



仮想ラボの構築

機械学習による高分子設計と高分子物性自動計算システムの統合

仮想空間内に実験計画法に基づき設計・合成・物性計測を自動化したパイプラインを構築



産学連携の新しいモデルケースを創出

様々なコントリビュータ

① データ生産



ノルマ = X samples per years

② コード開発



自動計算アルゴリズムの開発
機械学習等の手法やコード

③ 研究資金



データを活用した共同研究
データ生産の代行

参画機関にとってのメリット

データやコードのフルアクセス

DB system



RadonPy

例：自社独自のDB開発

研究開発への自由利用



技術・人的交流
人材育成に活用



毎月テクニカルミーティング
Slack等での議論・情報交換

産学連携の在り方：バーチャルラボを作る

2017年のセンター設立時に構想

- サイエンスの複合化が進むと、1研究グループで到達可能な領域が限られてくる。
- 従来の「委託・受託」的な産学連携ではなく、創造的なアイデアを生み出し、実践するための「パートナーシップ」を作ることに注力
- 将来的には「データ」を中心に据えた産学連携・学融合の形態を探求・実践していきたい。

