大学共同利用機関法人 情報・システム研究機構
統計数理研究所
SOKENDAI

RadonPy XenonPy

データ駆動型材料研究における実験・シミュレーション・機械学習の融合

吉田 亮 [1,2,3]

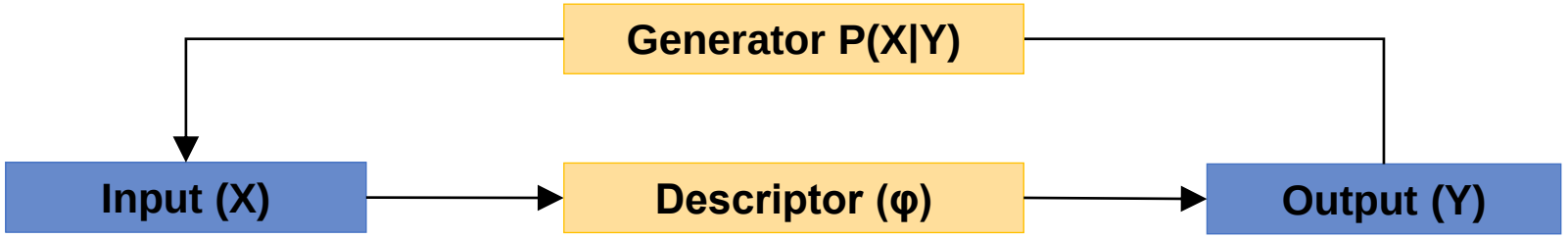
[1] 情報・システム研究機構 統計数理研究所 データ科学研究系
[2] 情報・システム研究機構 統計数理研究所 ものづくりデータ科学研究センター
[3] 総合研究大学院大学 複合科学研究科 統計科学専攻

XenonPy

データ駆動型材料研究の順問題・逆問題

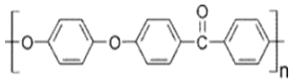
Ikebata et al. Bayesian molecular design with a chemical language model. *J Comput Aided Mol Des* **31**, 379-391 (2017).

Backward prediction S ~ p(Y|X)

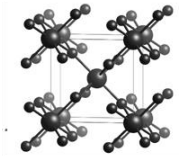


Forward prediction p(Y|X) = p(Y|φ(X))

Chemical structure



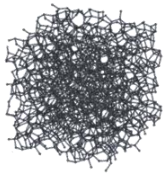
Crystal



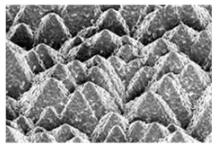
Composition

$Sr_2Ti_2O_6$

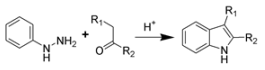
Amorphous



Microstructure

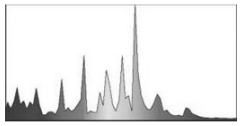


Chemical reaction



Θ, T, P

Spectrum



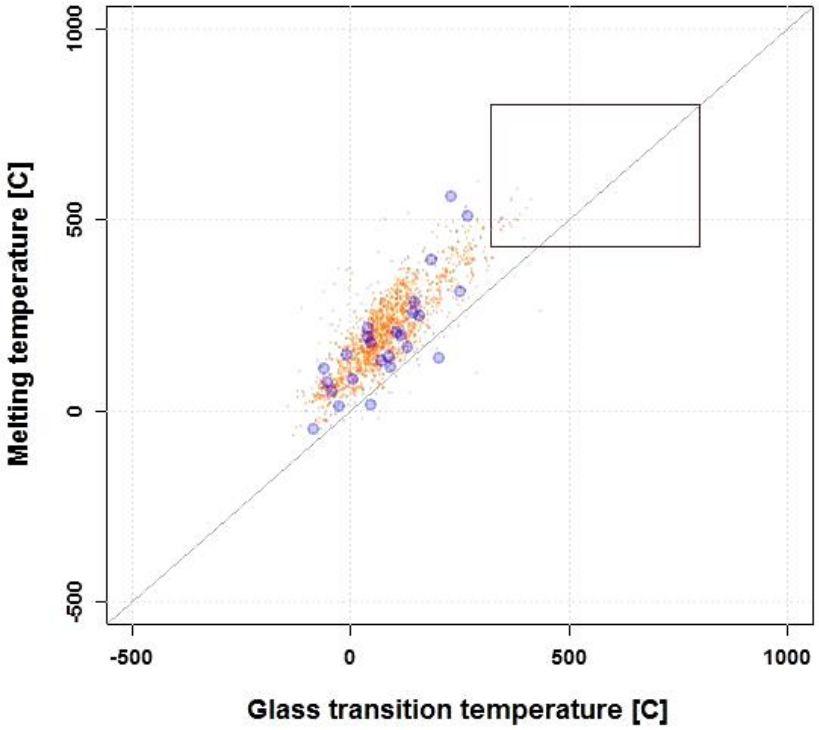
XenonPy

分子生成言語モデルを用いた高熱伝導非晶質高分子の発見

Wu et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput Mater* **5**, 66 (2019).

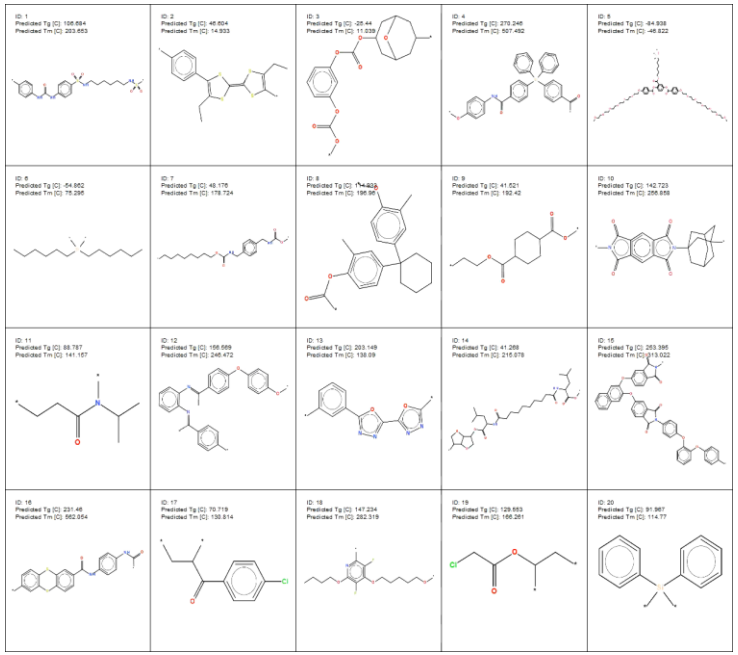
Refinement of polymer properties

Tg ↗ and Tm ↗



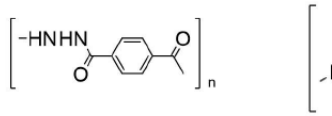
Designed molecules (monomers)

Generative models of chemical structures

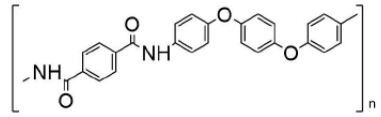


Discovery of new polymers

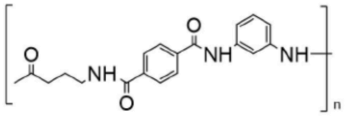
Three thermally conductive polymers



4

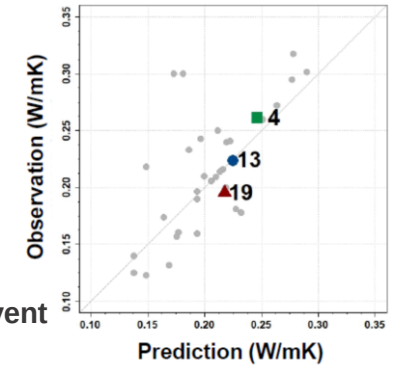


13



19

- **Filmed polymers**
- **Crystallinity**
- **Glassy states**
- **Soluble in organic solvent**
- **TC = ~0.408 W/mK**



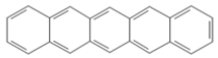
2017年～ものづくりデータ科学研究センター ～ 基盤技術・学術資源

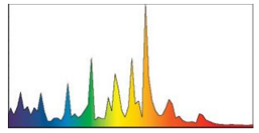
Process & Composition **Microstructure**

$Cr_x Al_y O_z N$
Θ, T, P

Iwayama et al. J. Chem. Inf. Model. 62(20):4837–4851 (2022)

Chemical structure **Spectrum**

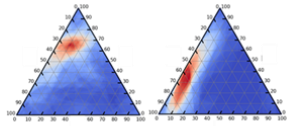




Iwayama et al. J. Chem. Inf. Model. 62(20):4837–4851 (2022)

Composition **Phase diagram**

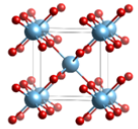
$Al_xTM_yTM_z$



Liu et al. Adv. Mater. 33(36):e2102507 (2021)
Liu et al. Phys Rev Mater (2023) (under review)

Composition **Crystal structure**

TiO_2

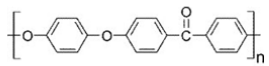


Kusaba et al. Comput. Mater. Sci. 211:111496 (2022)
Liu et al. arXiv (2023)

Reactants **Synthetic product**

Guo et al. J. Chem. Inf. Model. 60(10):4474–4486 (2020)
Zhang et al. Sci. Technol. Adv. Mater. Methods. (2022)
Ohno et al. ChemRxiv (2023)

Crystal or molecule **Property**



Tg, Tm, Cp

Yamada et al. ACS Cent. Sci. 5(10):1717-1730 (2019)
Ju et al. Phys. Rev. Mater. 5:053801 (2021)
Minami et al. AAAI 35(10):8992-8999 (2021)
Ikebata et al. J. Compt. 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)
Torres et al. J Phys: Condens Matter. 34(13):135702 (2022)
Hayashi et al. npj Comput Mater 8, 222 (2022)
Aoki et al. Macromolecules (2023) under review

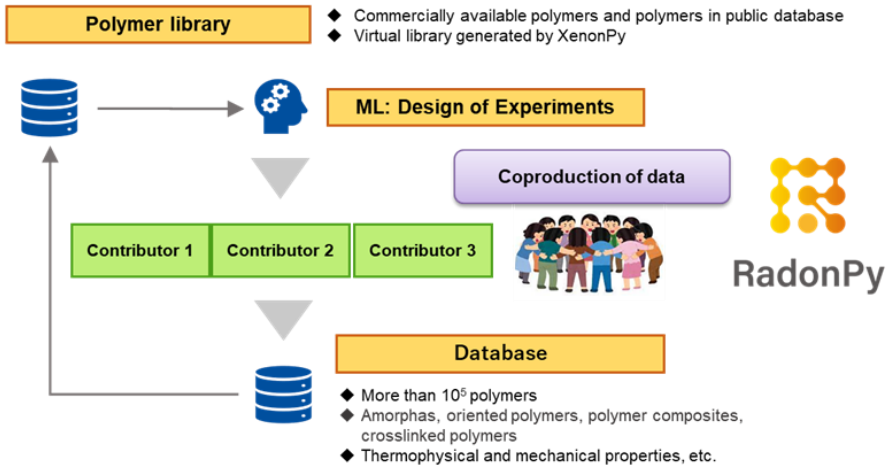
Software

RadonPy Python library for polymer properties calculation using fully automated MD simulation
https://github.com/RadonPy/RadonPy

XenonPy Python library for inverse materials design
https://xenonpy.readthedocs.io/en/latest/

Database

Co-creation of the world's largest polymer properties database (10^5-10^7 polymers) with Industry-Academia Consortium



データ駆動型材料研究におけるデータ資源の不足

オープンデータの不足：技術的・文化的問題 ～ 短中期的解決は困難

- **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/

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

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

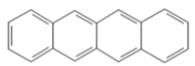
競合相手に対する情報秘匿の意識

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．

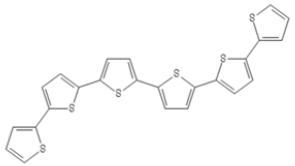
革新的な材料の周辺にはデータが存在しない

データ科学の内挿的予測の限界

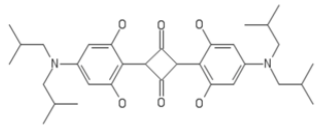
tetracene (2005)

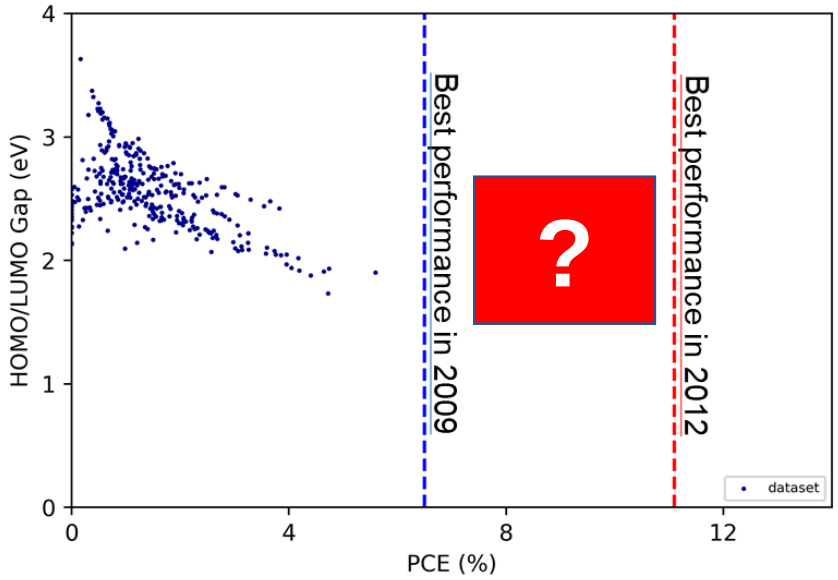


α-sexithiophene (2008)



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





データ駆動型材料研究の在り方とセンターのミッション定義

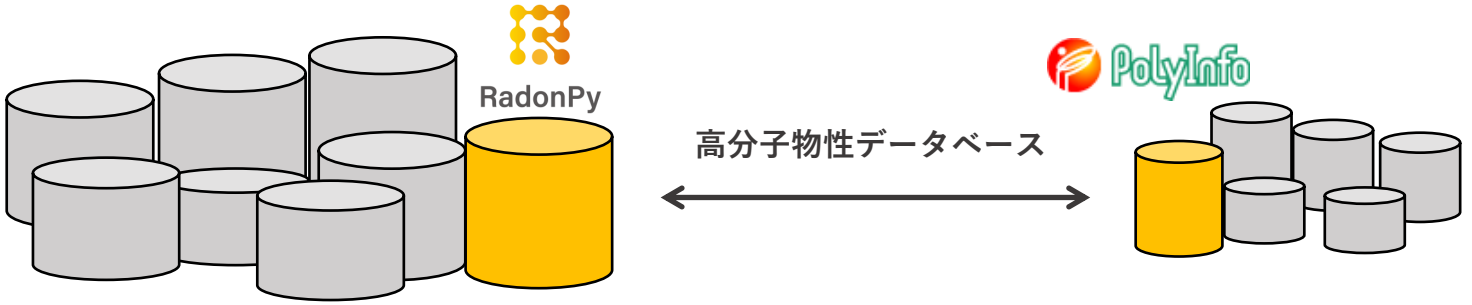
オープン領域

シミュレーション基礎物性・構造DB + 学習済み基盤モデル

第一原理電子状態計算・分子動力学計算・流体計算

実験 DB（大学・大型実験施設等） + 学習済み基盤モデル

文献・特許・自動実験：オープン化にやや壁



異種データ統合解析のためのデータ科学・API群
（転移学習・キャリブレーションなど）

クローズ領域

素材企業　大学の研究室　中規模国プロ　コンソーシアム　AIスタートアップ

林慶浩
（統数研）

RadonPy: 全原子分子動力学法による高分子物性計算の自動化

RadonPy

Hayashi et al. RadonPy: Automated Physical Property Calculation using All-atom Classical Molecular Dynamics Simulations for Polymer Informatics. npj Comput Mater npj Comput Mater 8, 222 (2022).

GitHub: https://github.com/RadonPy/RadonPy

Latest: 17 properties implemented

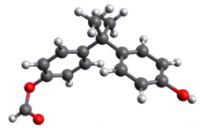
(2023/03/13)

- Thermal conductivity
- Thermal diffusivity
- Density
- Radius of gyration
- Specific heat capacity Cp
- Specific heat capacity Cv
- Compressibility (isothermal)
- Isentropic compressibility
- Bulk modulus (isothermal)
- Isentropic bulk modulus
- Self-diffusion coefficient
- Thermal expansion coefficient
- Linear expansion coefficient
- Dielectric constant (static)
- Refractive index
- Glass transition temperature
- Abbe's number

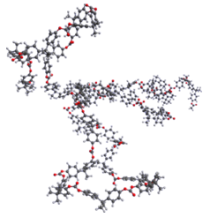
Polymer systems

- Amorphous/crystalline states
- Oriented structures
- Mixtures

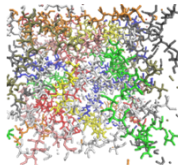
Oc1ccc(cc1)C(c1ccc(cc1)OC(=O))(C)C
SMILES



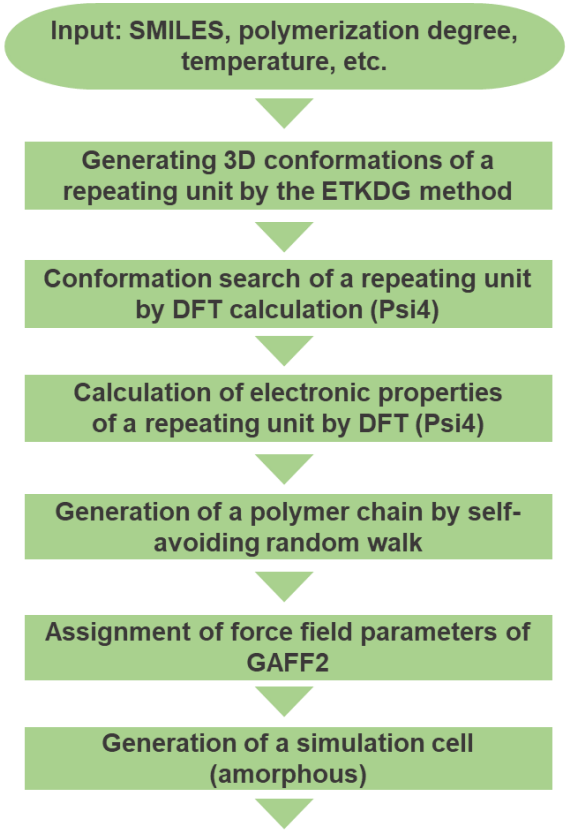
Repeating unit

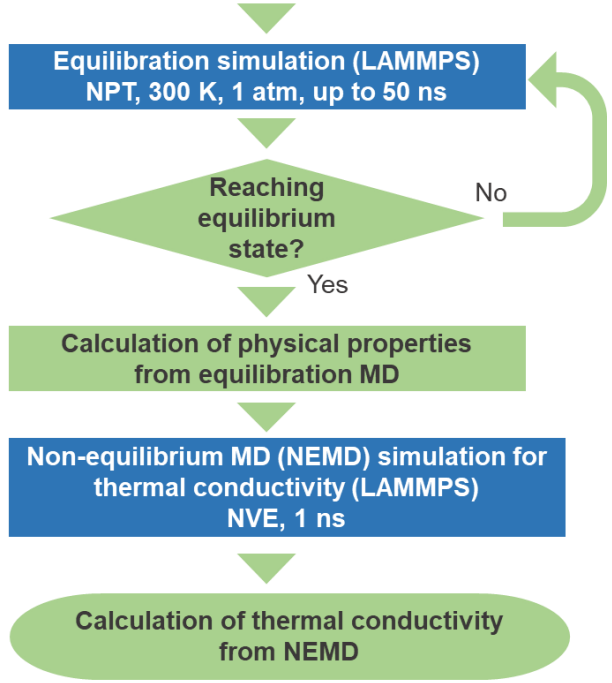


Polymer chain
(about 1,000 atoms)



Amorphous cell
(10 polymer chains)





高分子MD計算の全自動化の難しさ：ノウハウとコスト

計算条件・パラメータ設定に関する「ノウハウ」に大きく依存（属人的）

- 計算条件：アニーリング条件・重合度・時間刻み幅・長距離相互作用など，物性・系ごとに検討が必要
- MD計算の妥当性の評価（温度勾配・配向度など，物性値毎に検討が必要）

膨大な計算コスト

- 1ポリマーの計算に150時間以上（延伸配向したポリマーの熱伝導率@分子研スパコン）
- 延伸配向した2,000ポリマーの計算 = スパコン利用料金で換算すると2,500万円分

RadonPy

- MD専門家による計算条件の標準化（プリセット）
- MD計算の民主化：実験研究者・データ科学研究者など，誰でもカジュアルにMD計算

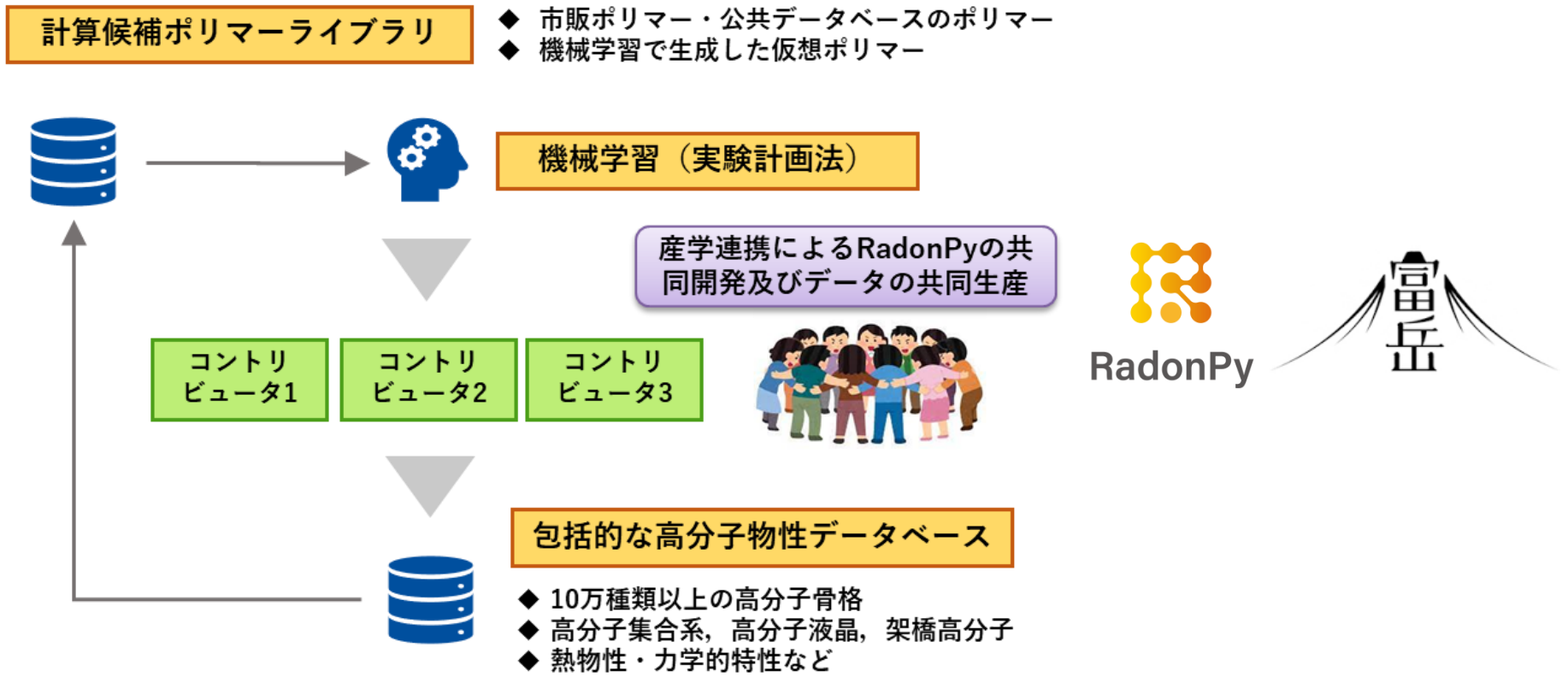
産学連携による RadonPy 及び 高分子物性オープンデータベースの共創

2021- RadonPy共同開発コンソーシアム@統計数理研究所ものづくりデータ科学研究センター

10^5-10^7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

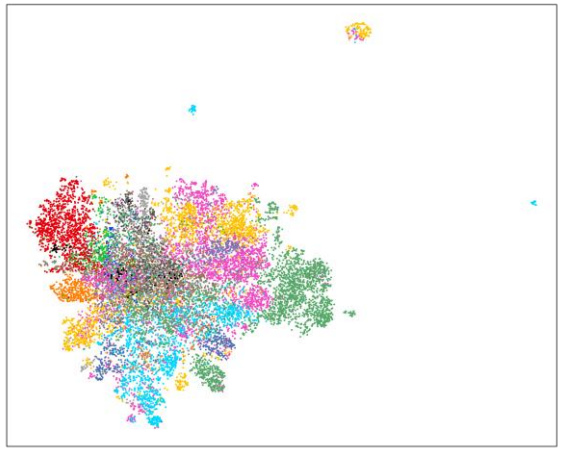
■ 統数研・3大学・24企業（参画者数：約138名）が組織の垣根を越えてデータを共同生産・共有

■ 「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」（代表：吉田亮）

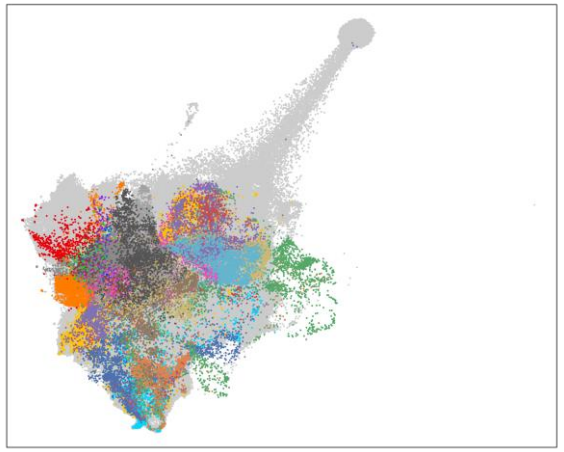


高分子物性の世界地図：広大な物質空間の全貌を明らかに

これまでに合成された高分子の分布
（PoLyInfo: 16,427）

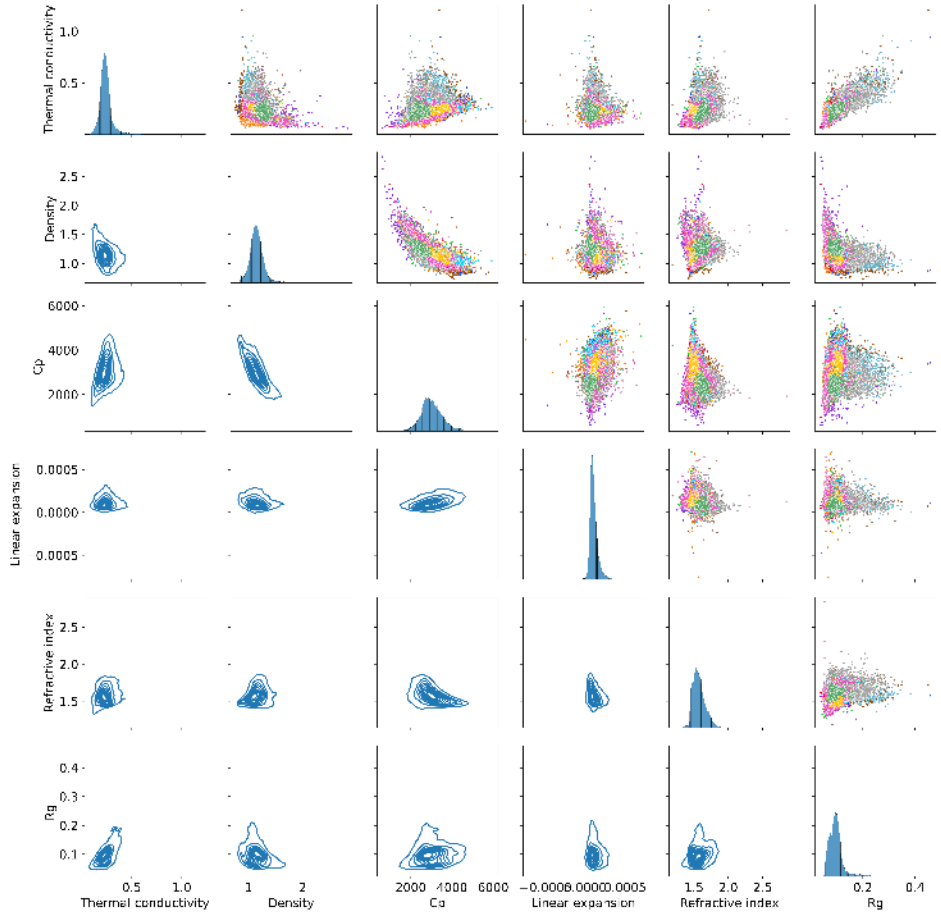


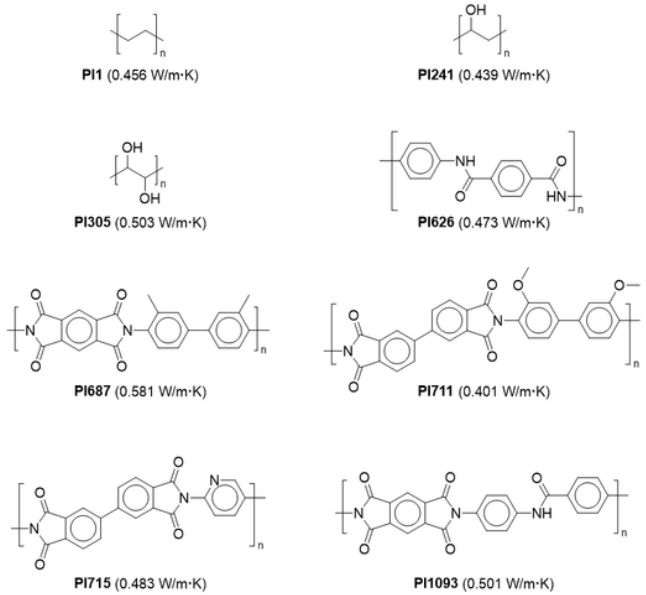
計算が完了した47,500ポリマー
(● 未計算仮想ポリマー)



- 複数物性の同時分布の観測：パレートフロンティアの位置や構造的特徴が明らかに
- 特異な物性を発現する高分子の候補を同定
- 実験だけではこのような網羅的観測は不可能

Hayashi et al. npj Comput Mater 8, 222 (2022).





熱伝導率が0.5 W / (m・K)を超えるアモルファスポリマーが存在するという予想

RadonPyの活用例：Sim2Real 転移学習

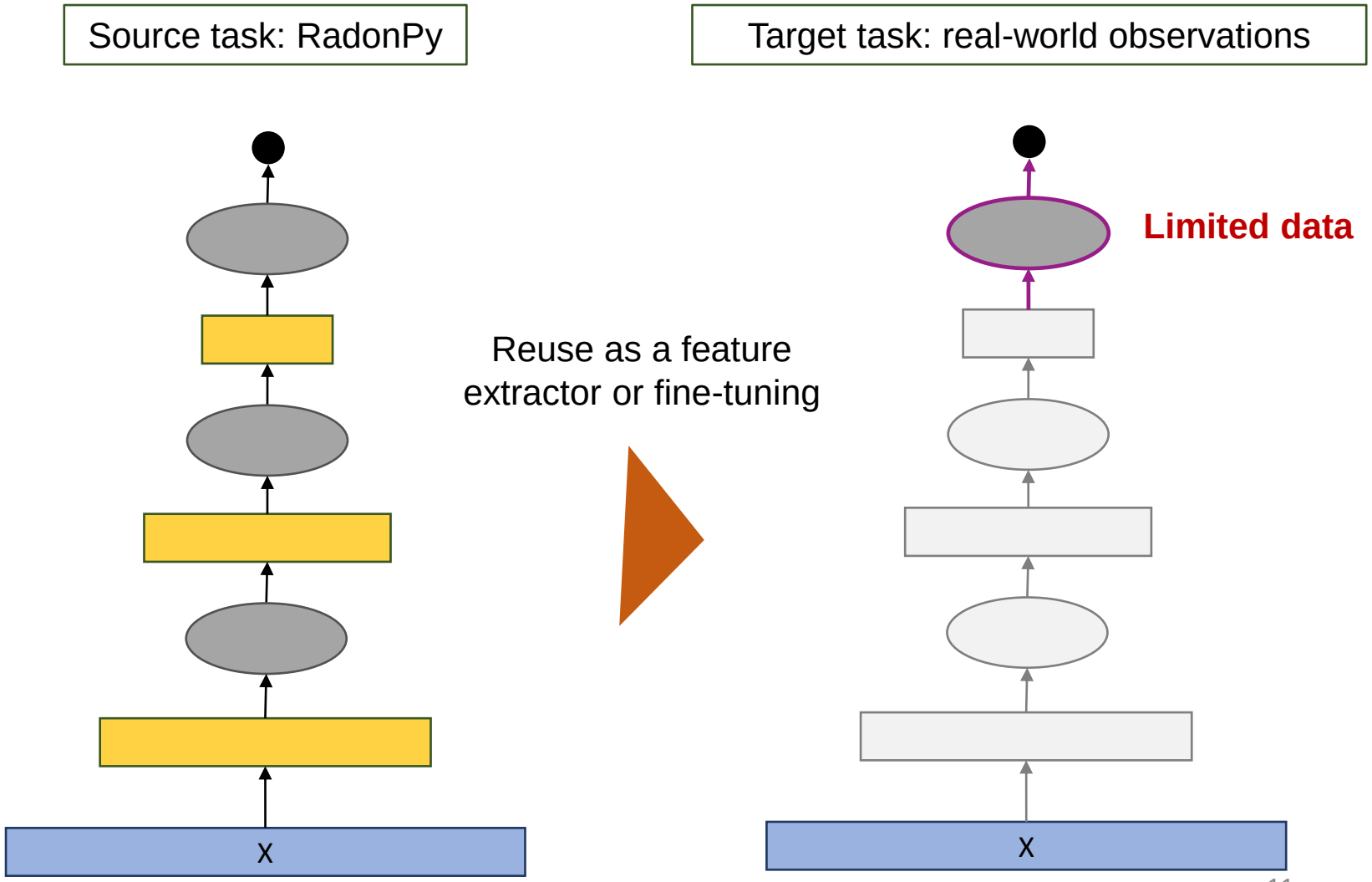
大量のシミュレーションデータと少数の実験データの統合解析（転移学習・ドメイン適応・マルチフィディリティ学習など）

Latest: 17 properties implemented
(2023/03/13)

- Thermal conductivity
- Thermal diffusivity
- Density
- Radius of gyration
- Specific heat capacity Cp
- Specific heat capacity Cv
- Compressibility (isothermal)
- Isentropic compressibility
- Bulk modulus (isothermal)
- Isentropic bulk modulus
- Self-diffusion coefficient
- Thermal expansion coefficient
- Linear expansion coefficient
- Dielectric constant (static)
- Refractive index
- Glass transition temperature
- Abbe's number

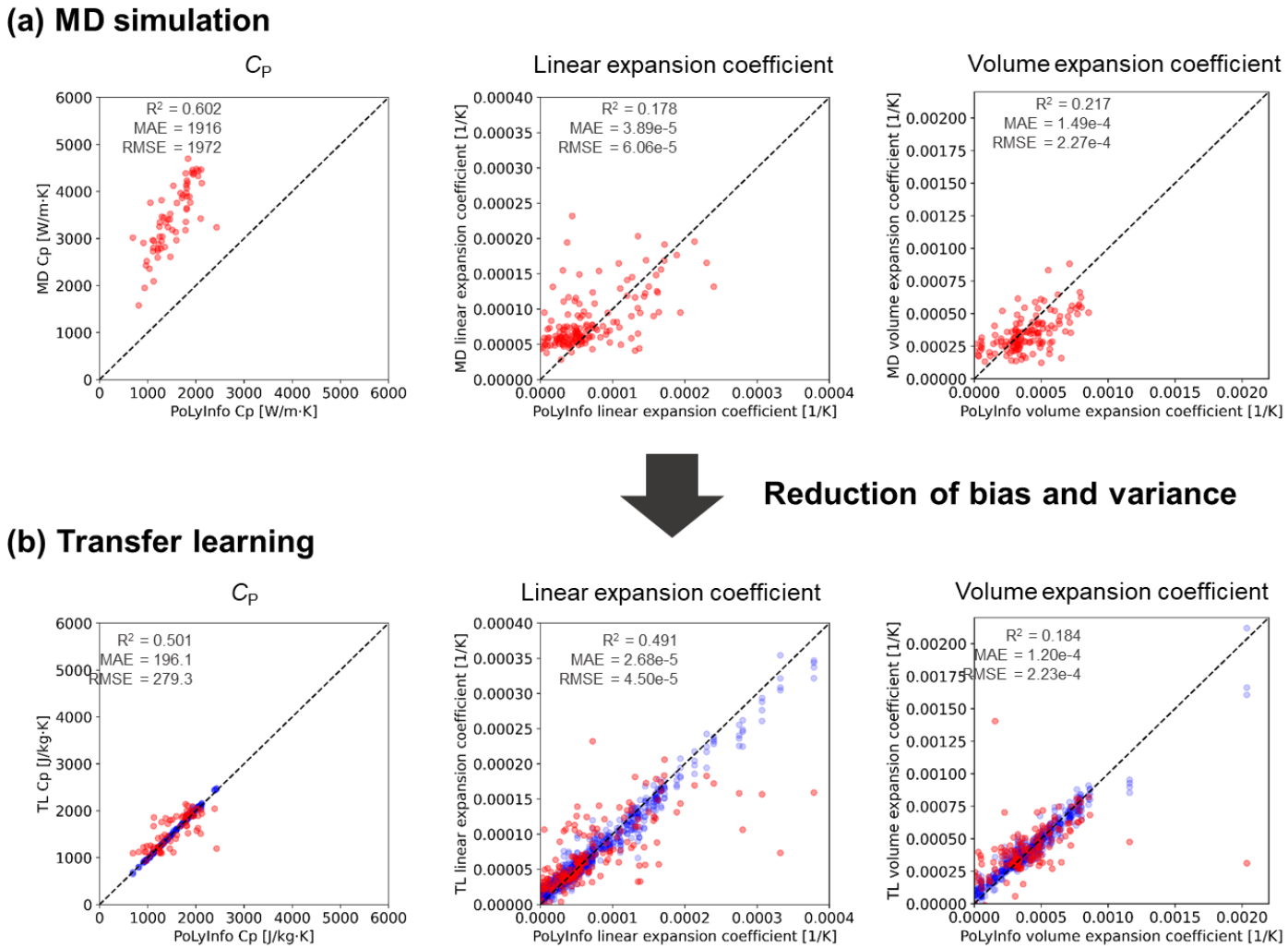
Polymer systems

- Amorphous/crystalline states
- Oriented structures
- Mixtures



不完全な計算と不確かな現実系の乖離を機械学習で埋める

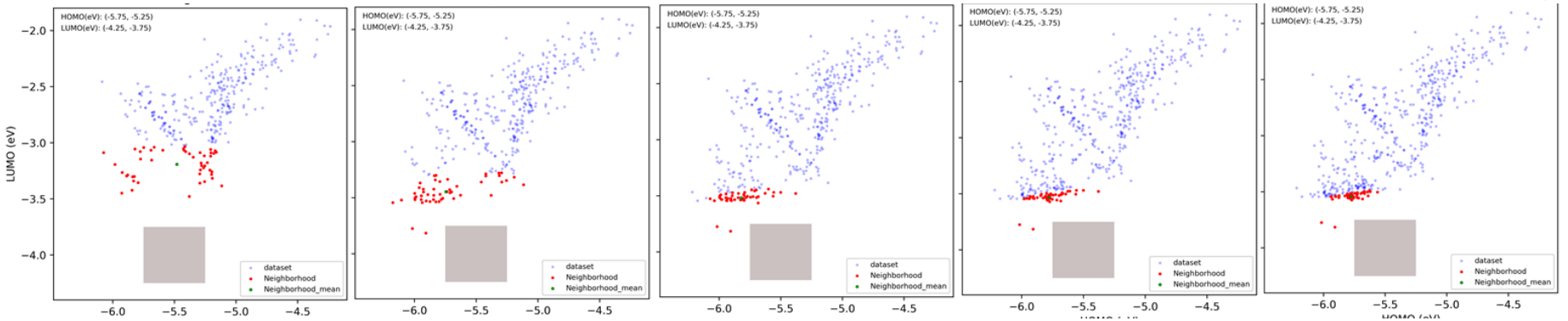
Hayashi et al. npj Comput Mater 8, 222 (2022).



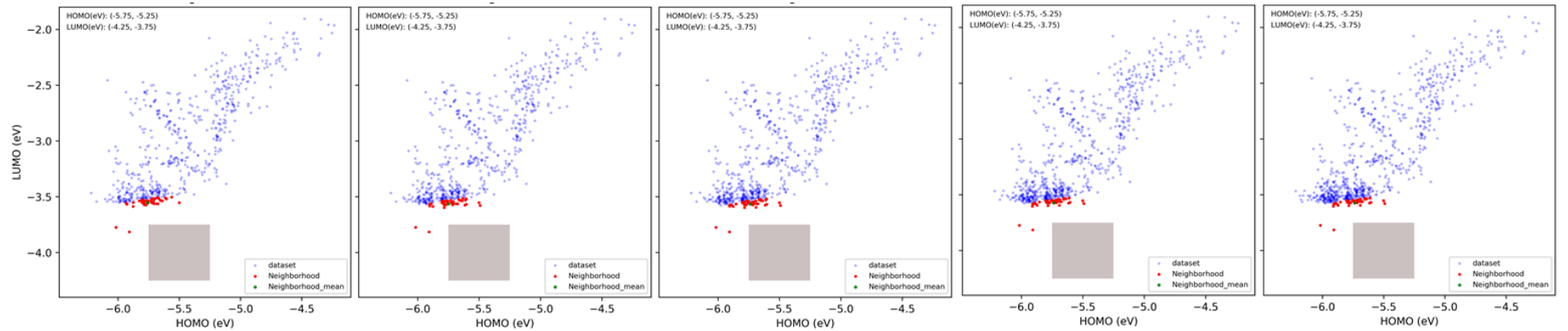
未踏領域に存在する新物質を同定する

受動的に得られたデータを単純に解析するだけでは到達できない

t = 1



t = 6



Scharber's formula for PCBM (Adv. Mater. 18, 789 (2006))

PCE & HOMO-LUMO gap ➤ HOMO & LUMO (DFT-computable)

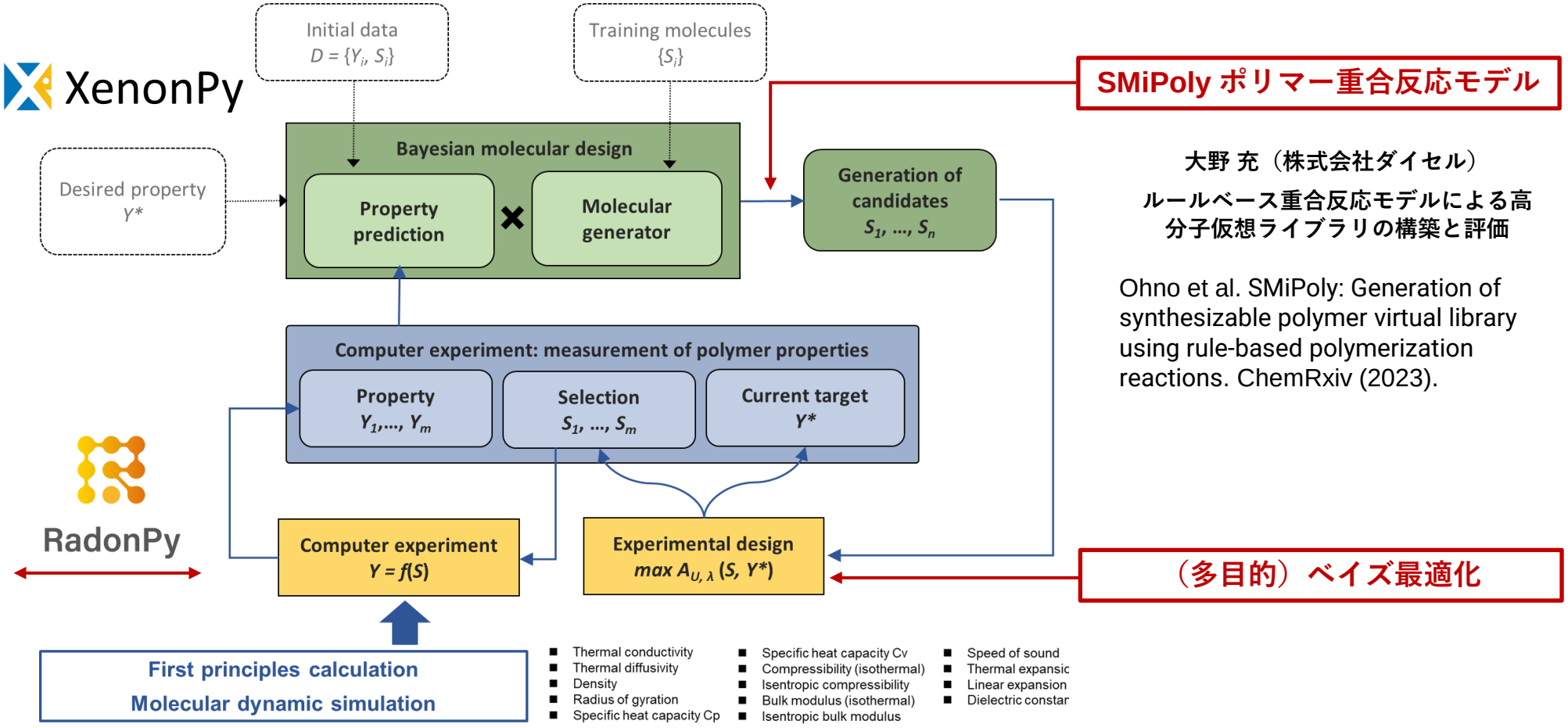
Y-axis: LUMO
X-axis: HOMO
Blue: generated molecules
Red: molecules selected for DFT
Blue zone: final target
Yellow zone temporary target

SPACIER: RadonPy自動実験×キャリブレーション×生成AI×適応的実験計画法

世界初のポリマー物性自動計算×自動設計システム

南條舜
（総研大）

Arifin (JSR)



SMiPoly ポリマー重合反応モデル

大野 充（株式会社ダイセル）

ルールベース重合反応モデルによる高分子仮想ライブラリの構築と評価

Ohno et al. SMiPoly: Generation of synthesizable polymer virtual library using rule-based polymerization reactions. ChemRxiv (2023).

実験データ
キャリブレーション

南 俊匠& 福水健次（ISM）

アフィン型転移学習

（多目的）ベイズ最適化

Minami & Fukumizu et al. Transfer learning with affine model transformation. arXiv (2022) https://doi.org/10.48550/arXiv.2210.09745.

設計空間を自由自在に走査できる統計的生成モデル

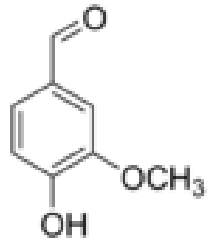
PoLyInfoのPIDの接頭番号と分子骨格の21クラス

P01〜	炭化水素系 (hydrocarbons)
P02〜	ポリスチレン系 (polystyrenes)
P03〜	ポリビニル、ポリビニリデン (polyvinyl, polyvinylidene)
P04〜	アクリル樹脂 (polyacrylate)
P05〜	ハロゲン含有ポリマー (halogen containing polymers)
P06〜	ポリエン (polyene)
P07〜	ポリエーテル (polyether)
P08〜	ポリチオエーテル (polythioether)
P09〜	ポリエステル (polyester)
P10〜	ポリアミド (polyamide)
P11〜	ポリウレタン (polyurethane)
P12〜	ポリウレア (polyurea)
P13〜	ポリイミド (polyimide)
P14〜	ポリ無水物 (polyanhydrides)
P15〜	ポリカーボネート (polycarbonate)
P16〜	ポリアミン？ポリアミド以外の窒素含有ポリマー？ (polyamine)
P17〜	ポリシラン、ポリシロキサン (polysilane, polysiloxane)
P18〜	ポリホスファゼン (polyphosphazene)
P19〜	不明
P20〜	ポリスルホン (polysulfon)
P21〜	ポリフェニレン (polyphenylene)

確率的言語モデルを用いて，各クラスのポリマーを模倣した仮想ライブラリを作製

$$p(S) = p(s_1) \prod_{i=2}^{p} p(s_i | s_{1:i-1})$$

- **SMILES文字列集合からモデルを訓練**
- **頻出部分構造や化学結合，重合点の位置，トポロジーを学習**



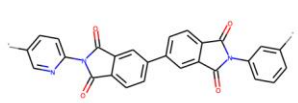
vanillin $C_8H_8O_3$

S = O=Cc**1**ccc**(**O**)**c**(**OC**)**c**1**

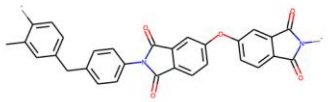
- Ikebata et al. Bayesian molecular design with a chemical language model. *J Compt Aided Mol Des*. 31:379-391 (2017)
- Wu et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput Mater*, 5:66 (2019)
- Wu et al. iQSPR in XenonPy: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

Virtual library example

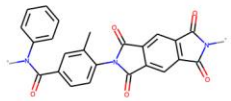
Machine-learned polyimide



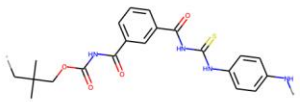
P13-443.0



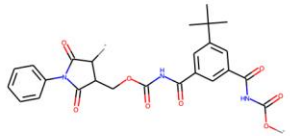
P13-486.0



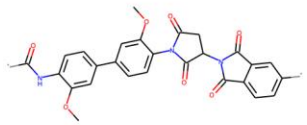
P13-423.0



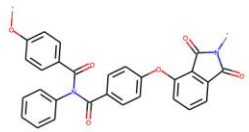
P43-426.0



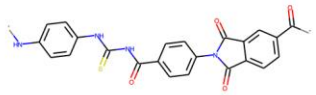
P43-493.0



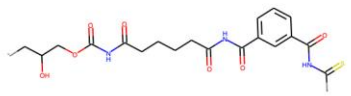
P43-497.0



PIMD-476.0

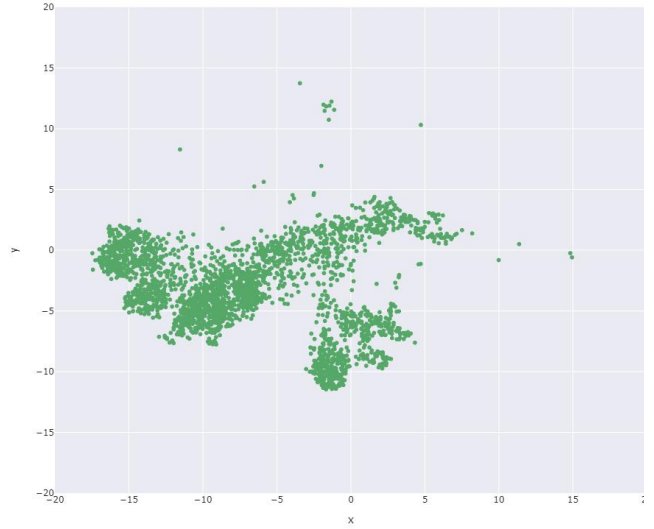


PIMD-442.0

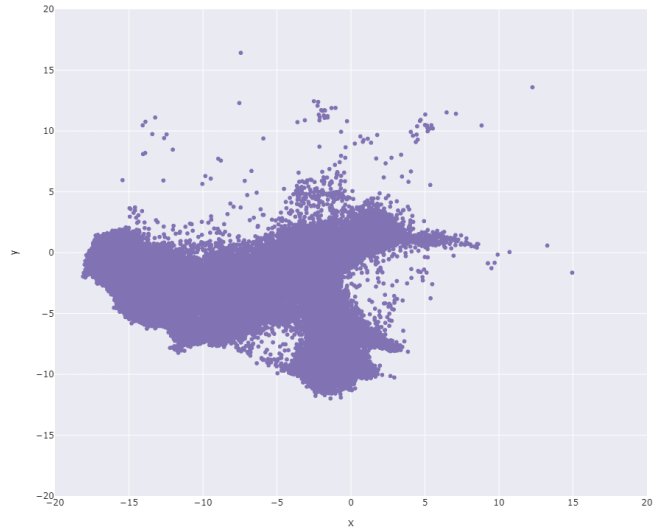


PIMD-435.0

PoLyInfo



XenonPy



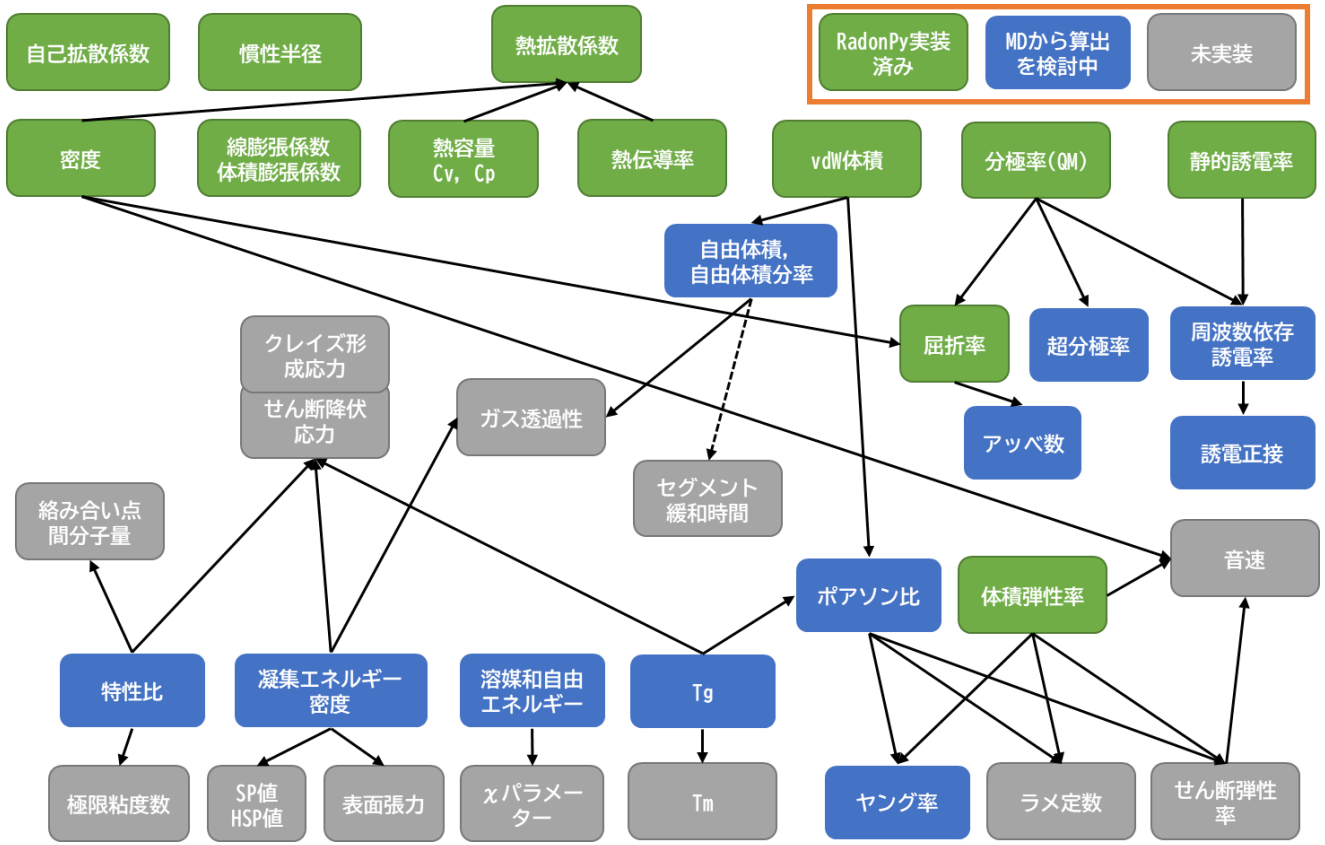
RadonPyの多機能化・高度化：コンソーシアム型オープンソース開発

産学連携コンソーシアムによるRadonPyオープンソース開発

- 多彩な高度専門人材による開発の加速
- 世界標準のソフトウェアに

■ ガラス転移温度（杉澤 宏樹　三菱ケミカル株式会社）
■ 配向構造の物性（古屋 秀峰　東京工業大学）
■ 溶媒和自由エネルギー（山田 寛尚　東京薬科大）
■ レオロジー物性（企業コントリビュータ）
■ 量子化学計算（林 慶浩　統数研）
■ 架橋ポリマー（企業コントリビュータ）
■ 誘電特性（古屋 秀峰　東京工業大学）
■ 共重合体（南條 舜　総研大）
■ 高分子溶液
■ 物性の温度依存性・組成依存性
■ 低分子化合物の材料物性（企業コントリビュータ）
■ 生分解性ポリマー（篠田恵子　統数研）

計算対象物性の拡張計画

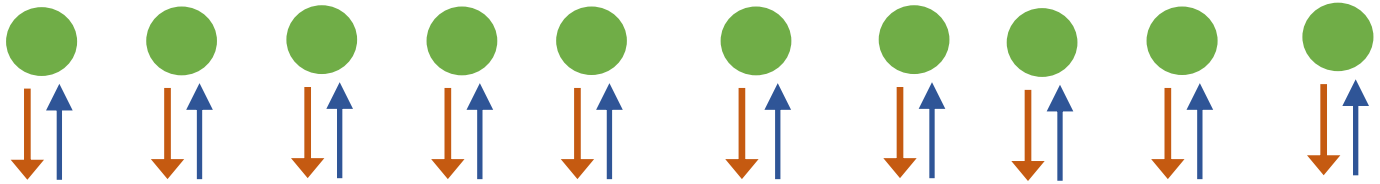


バーチャルなラボを作る

- 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．
- 「パートナーシップ」を築き，ラボの活動を外に開放：100-200名規模のバーチャルラボ（単独のグループではできない研究）

バーチャルラボ：共同研究・研究員派遣・セミナー・勉強会・日常的な交流・議論の場

高分子　磁性　超伝導　準結晶　半導体　バイオポリマー　電池　合成　計算化学　自動実験ロボット



大学共同利用機関法人 情報・システム研究機構
統計数理研究所

MI 研究者

教授1名
准教授1名
助教1名
特任研究員10名

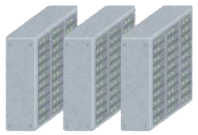
マテリアルズインフォマティクス

結晶構造予測
分子・合成経路設計
物性予測モデル
転移学習
第一原理計算
分子動力学計算

オープンソースソフトウェア

XenonPy
RadonPy

高分子物性データベース



共同研究部門

MI Frontier

三菱ケミカル株式会社
ISM-MCC フロンティア材料設計拠点

JSR株式会社
JSR-ISM スマートケミストリーラボ

スピンアウト

国プロ・大型競争的資金

コンソーシアム

セミナー・勉強会

複数機関での共同開発

RadonPyプロジェクトの成り立ち

わずか2年

2020年中頃	RadonPyプロトタイプが概ね完成
	共同研究者に構想を投げかけ
	当初はコンソ限定のデータベース
2021年2月	共同研究者（大学・企業）と意見交換会
2021年4月27日	コンソ非公式発足: 3大学・19企業
	当初は計算資源を分担
	計算資源が足りないことに気付く
2021年6月9日	「富岳」成果創出加速P 応募・採択
	採択条件：データベースをオープンに
	コンソ内で公開に反対する意見は一切なく，世界最高峰の高分子物性オープンデータベースの創出を目指すことに
2021年8月	「富岳」成果創出加速P 始動
2022年8月	コンソーシアム公式発足（規約施行）

2021年2月 意見交換会アンケート抜粋

- 日本国内での材料DB作成に関する日本的で閉鎖的な議論が近年続いているものと認識しています．本取り組みが呼び水となり、様々なDB創出のための産業界の協調の機運が起これば、と思いました．
- 本事業ができれば取組そのものが世界初として認知されると思っています．
- 自社で囲い込む研究開発が難しくなってきている昨今，ある程度シェアできるところはオープンにした方が産業界全体に良いように思います．本事業がそのような取り組みのロールモデルになれば，大きなインパクトがあるとおもいます．