



RadonPy

プロジェクト概要

データ駆動型高分子材料研究の諸問題

吉田 亮^{1,2,3,4}

代表機関：統計数理研究所

協力機関：東京工業大学

連携機関：東京大学，東京薬科大学

三菱ケミカル株式会社，JSR株式会社，出光興産株式会社，株式会社ダイセル，横浜ゴム株式会社，三井化学株式会社，株式会社日本触媒，東洋紡株式会社，株式会社アスミス，東ソー株式会社，旭化成株式会社，日東電工株式会社，住友ベークライト株式会社，株式会社デンソー

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

⁴ 物質・材料研究機構 統合型材料開発・情報基盤部門

データ駆動型高分子材料研究におけるデータ資源の不足

- APIによるデータの自動抽出ができない（ウェブ閲覧用）
- 複数の物性が同時に測定されたデータがほとんどない。
- バックグラウンド（プロセス条件など）の情報が欠けている。

データベース	概要
PoLyInfo (polymer.nims.go.jp)	学術文献から抽出したデータをまとめたデータベース。18,015種類のモノマーから重合されたポリマー群の約100物性のデータを収録している。
Polymer Genome – Khazana (khazana.gatech.edu)	24の出版物から抽出した実験データと第一原理計算で算出した物性値を提供しているプラットフォーム。データベースには、1,412種類の結晶性ポリマーと2,657種類の無機材料の特性データが収録されている。
Polymer Property Predictor and Database (pppdb.uchicago.edu)	文献から抽出した263件のFlory-Huggins χ パラメータと212件のガラス転移温度のデータを含む。
NanoMine (materialsmine.org)	ポリマーコンポジットの微細組織構造の組成、プロセス、電子顕微鏡データ、物性を含む。
CROW (polymerdatabase.com)	ポリマーの熱物性データを含むデータベース。文献から抽出した実験データや定量的構造活性相関解析から算出した計算物性のデータを含む。
Polymers: A Property Database (poly.chemnetbase.com)	Wiley出版社の書籍“Polymers: A Property Database”の付録として提供されている高分子物性データ

無機結晶・低分子化合物の計算物性ビッグデータ

第一原理計算データベースの誕生がデータ駆動型研究の発展・普及を大きく牽引

■ Materials Project (DFT)

~130,000 virtual and real compounds

Accessible via API

<https://materialsproject.org/>

Jain et al. APL Materials, 1, 011002 (2013).

■ AFLOW (DFT)

~3,000,000 virtual and real compounds

Accessible via API

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

Curtarolo et al. Comput Mater Sci, 58, 218–226 (2012). <http://quantum-machine.org/gdml/#datasets>

■ OQMD (DFT)

~600,000 virtual and real compounds

Accessible via API

<http://oqmd.org/>

Kirklin et al. npj Comput. Mater, 1, 1-15 (2015).

■ QM9

~134k stable small organic molecules

Batch download available

<http://quantum-machine.org/datasets/>

Ramakrishnan et al. Sci Data, 1, 1–7 (2014).

■ MD17

150k to nearly 1M conformational geometries

MD trajectories of 10 molecules

Batch download available

<http://quantum-machine.org/gdml/#datasets>

Chmiela et al. Sci Adv, 3(5), e1603015 (2017).

■ Open Catalyst Project

130 millions of metal catalyst surfaces and molecules

Accessible via API

Facebook AI Research and CMU

<https://opencatalystproject.org/>

短中期的には大学のラボや1企業で生産可能なデータが標準に

データ駆動型高分子材料研究が発展・普及する上で大きな阻害要因に

コスト

実験・シミュレーション・試料作製・物性評価等のコストが高い。
高分子物性計算のハイスループット化の技術障壁

ニーズ 多様性

設計変数（材料種，プロセスなど）や高次構造等の多様性
コモンデータを創出しようという動きが起きにくい。

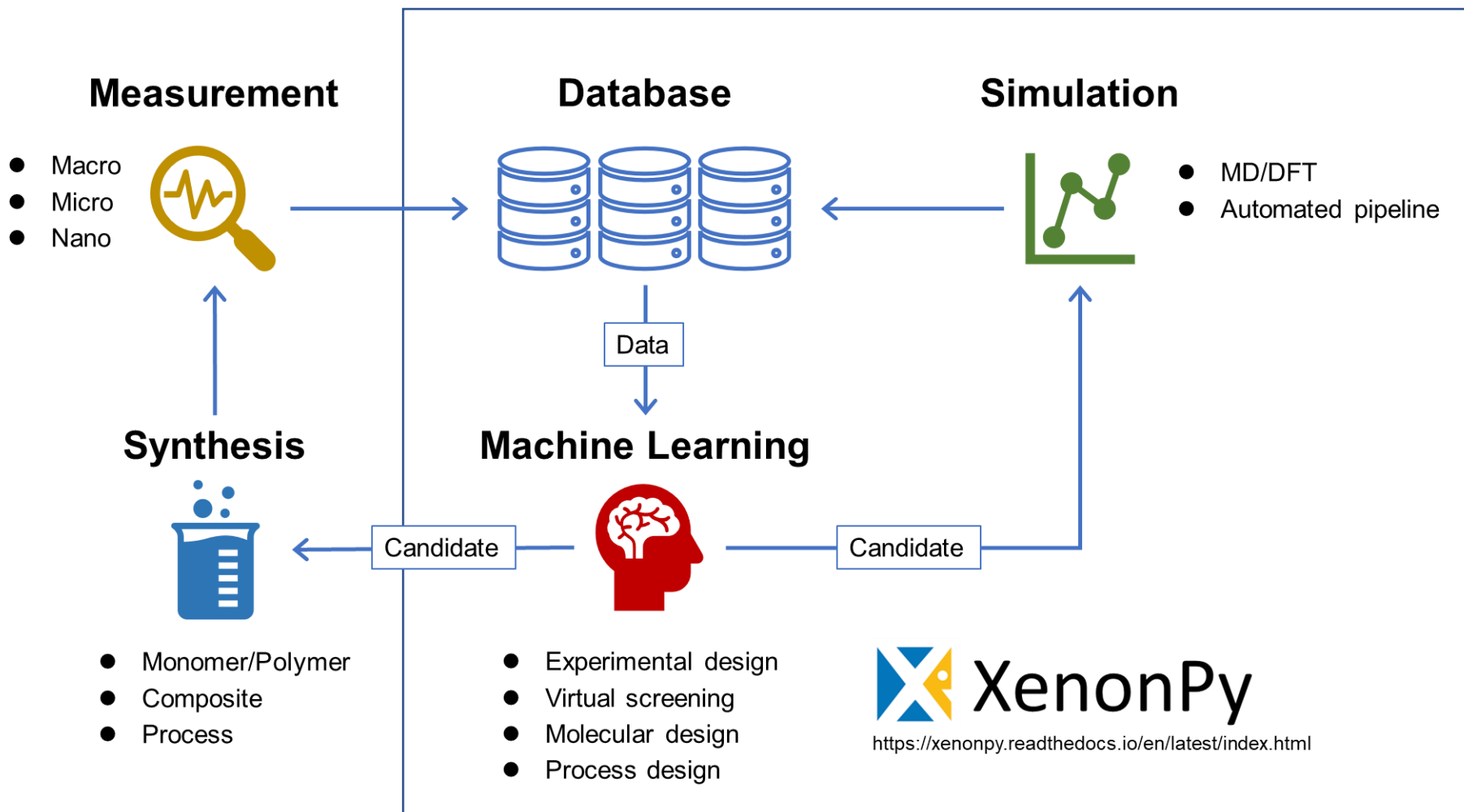
インセン ティブ欠如

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働きにくい。

高分子熱物性のデータ駆動研究プラットフォーム

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





MDシミュレーション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. arXiv:2203.14090 (2022).

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

Latest: 15 properties implemented

(2022/03/31)

- ◆ 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)
- ◆ Refractive index

Polymer systems

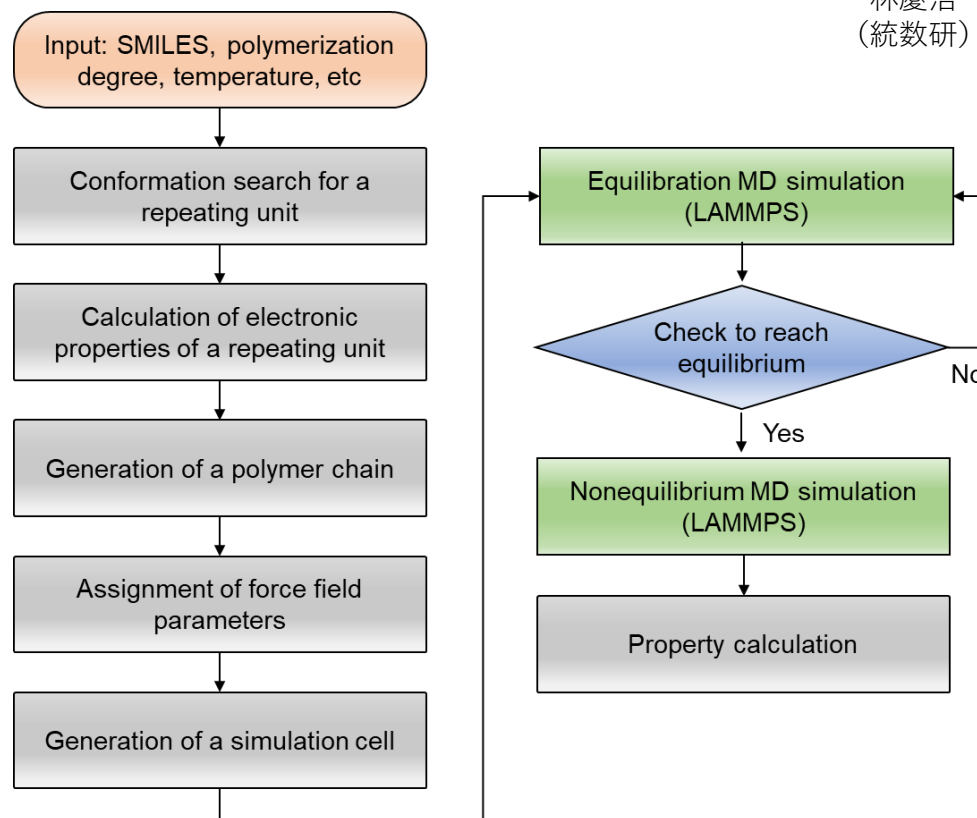
- ◆ Amorphous/crystalline states
- ◆ Oriented structures
- ◆ Mixtures

RadonPy (OSS)

Python library for the automation of all-atom MD simulations using LAMMPS



林慶浩
(統数研)

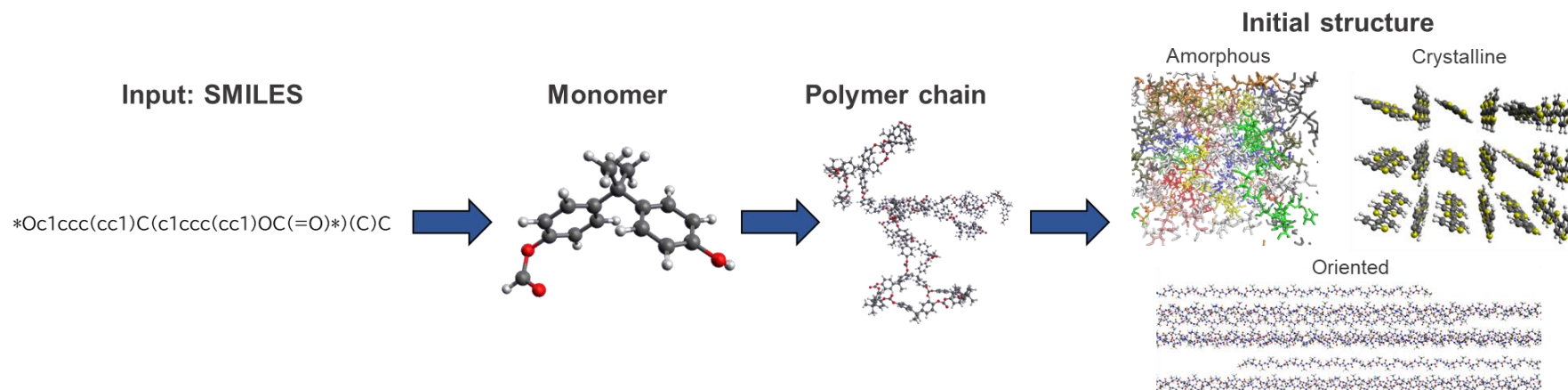


高分子物性自動計算の難しさ

高分子物性計算の自動化：煩雑なプロトコル設定

パイプラインを構築するためのツール群が必要

- 分子力場パラメータの自動割り当て
- 原子電荷の第一原理計算
- シミュレーション条件（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、物性値毎に検討が必要）
- 系の平衡化の終了判定（各種物理量の収束状況から判定）
- 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）



実験データとの系統的比較に基づく計算条件の決定

- 2021年度: PoLyInfo 1,070 ポリマー（5回の独立な試行）15物性の計算
- 実験値との比較検証を実施し，計算条件を決定

Hayashi et al. RadonPy: Automated Physical Property Calculation using All-atom Classical Molecular Dynamics Simulations for Polymer Informatics.
arXiv:2203.14090 (2022).

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・19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」

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）

1国研・3大学・19企業（R3年度）によるデータの共同生産

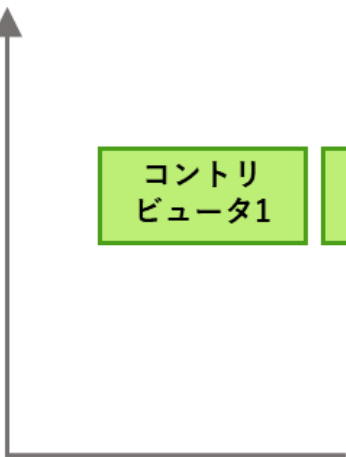
コントリビュータ1

コントリビュータ2

コントリビュータ3




RadonPy



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

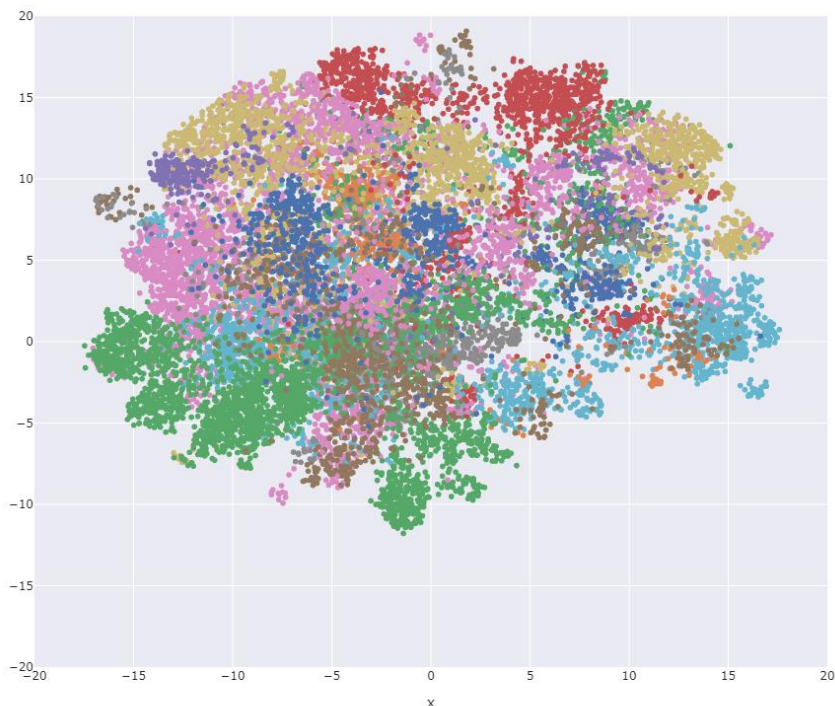
高分子物性の大地図を作成

10^5 - 10^7 個の分子骨格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を観測し、パレートフロンティアの位置や構造的特徴を明らかに
- 特異な特性を持つポリマーを予測・合成
- 機械学習（サロゲートモデリングや転移学習など）のデータ資源としての利用

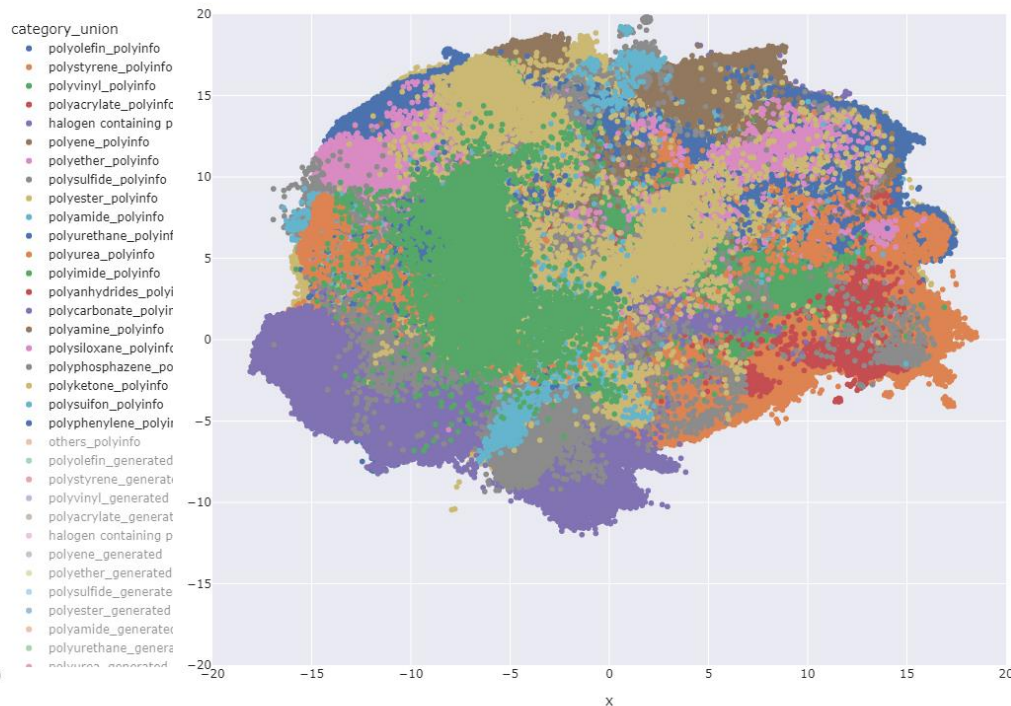
これまでに合成されたポリマーの分布（UMAP）

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

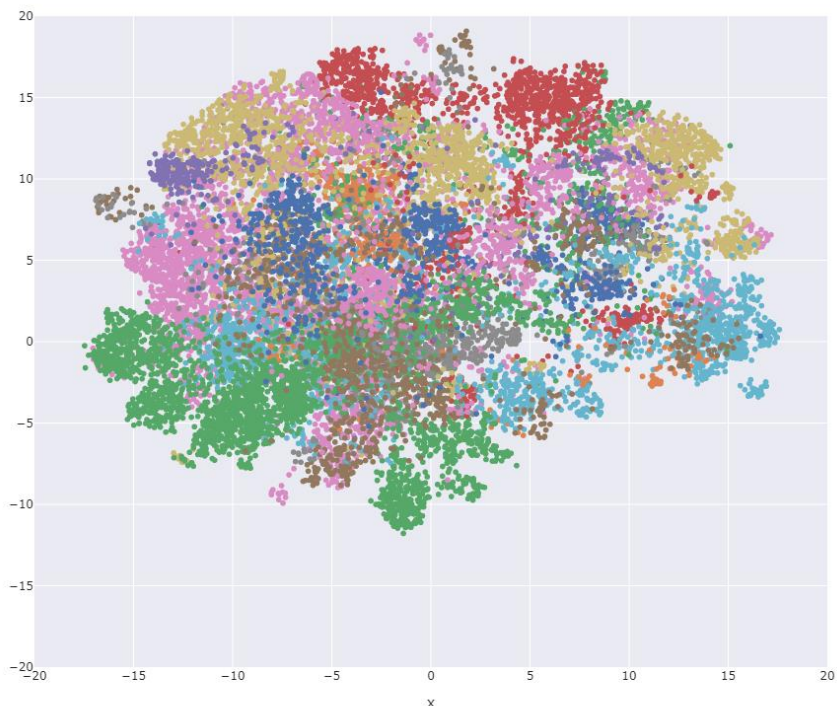
Ikebata et al. J Compt Aided Mol Des. 31:379-391 (2017)



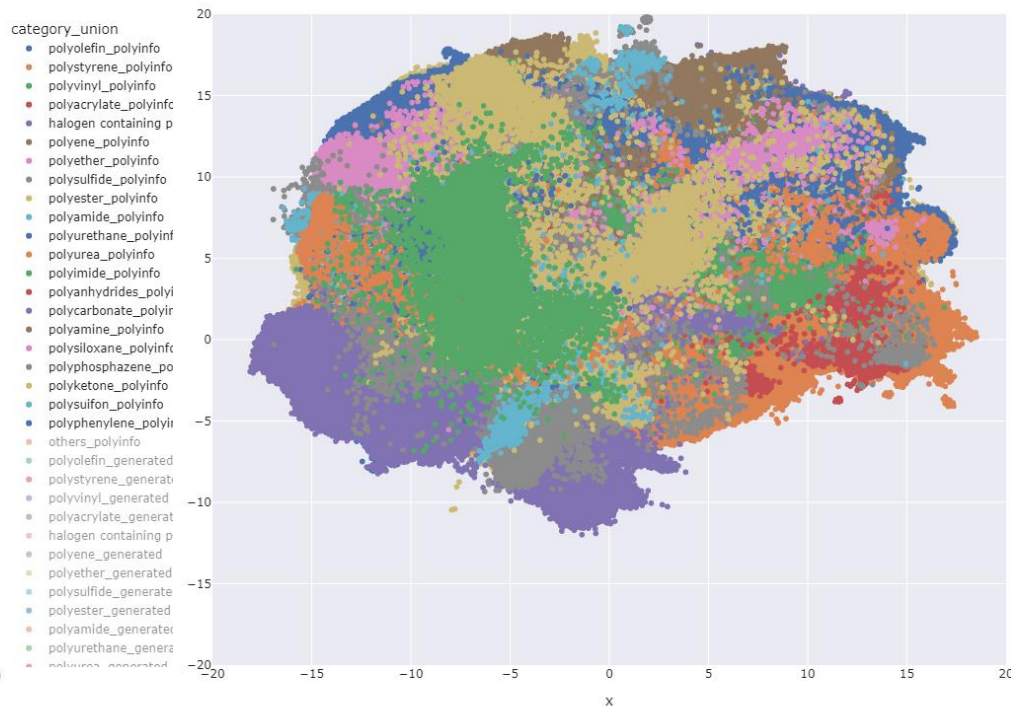
第1世代 仮想ライブラリ：ケミカルスペースを埋め尽くす

P01	PHYC	Polyolefins (hydrocarbons)	1,708	P11	PURT	Polyurethanes	146,887
P02	PSTR	Polystyrenes	20,233	P12	PURA	Polyureas	55,968
P03	PVNL	Polyvinyls	23,824	P13	PIMD	Polyimides	176,614
P04	PACR	Polyacrylics	95,855	P14	PANH	Polyanhydrides	61,776
P05	PHAL	Polyhalo-olefins	1,614	P15	PCBN	Polycarbonates	38,891
P06	PDIE	Polydienes	778	P16	PIMN	Polyimines	34,658
P07	POXI	Polyethers	91,035	P18	PPHS	Polyphosphazenes	13,762
P08	PSUL	Polysulfides	30,631	P19	PKTN	Polyketones	81,027
P09	PEST	Polyesters	458,961	P20	PSFO	Polysulfones	49,147
P10	PAMD	Polyamides	370,460	P21	PPNL	Polyphenylenes	24,210

PoLyInfo の 21クラス分類



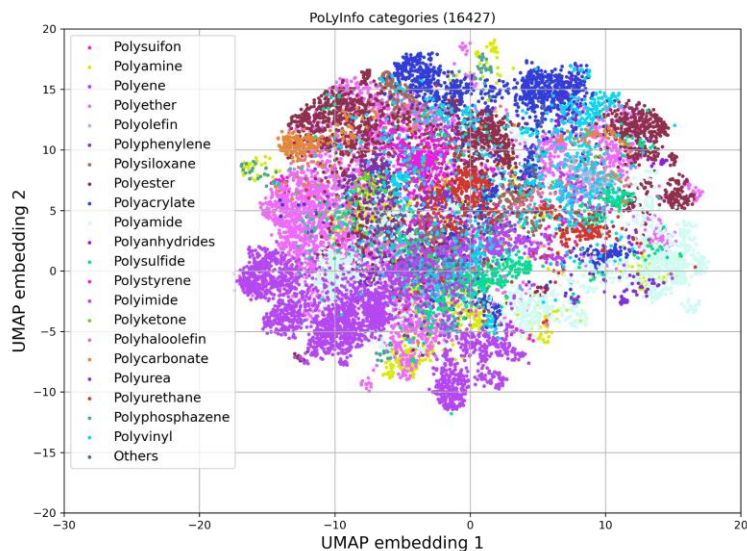
機械学習で構築した仮想ポリマーの分布



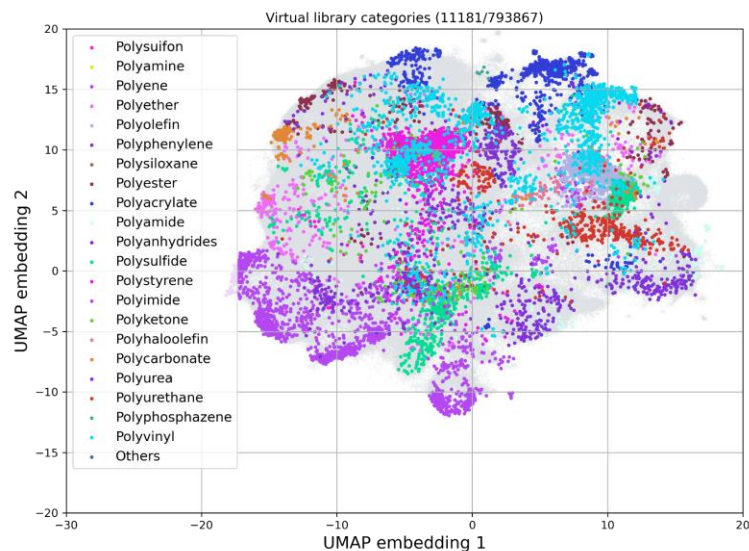
プロジェクトの現在地点

- RadonPyの富岳への移植完了
- 2022/02/28時点：11,181ポリマーの計算を完了（所要期間：約2か月, 1080,000 NH）

PoLyInfo の 21クラスポリマーの分布
16,427ポリマー



仮想ポリマー：1,778,039
計算完了ポリマー：11,181

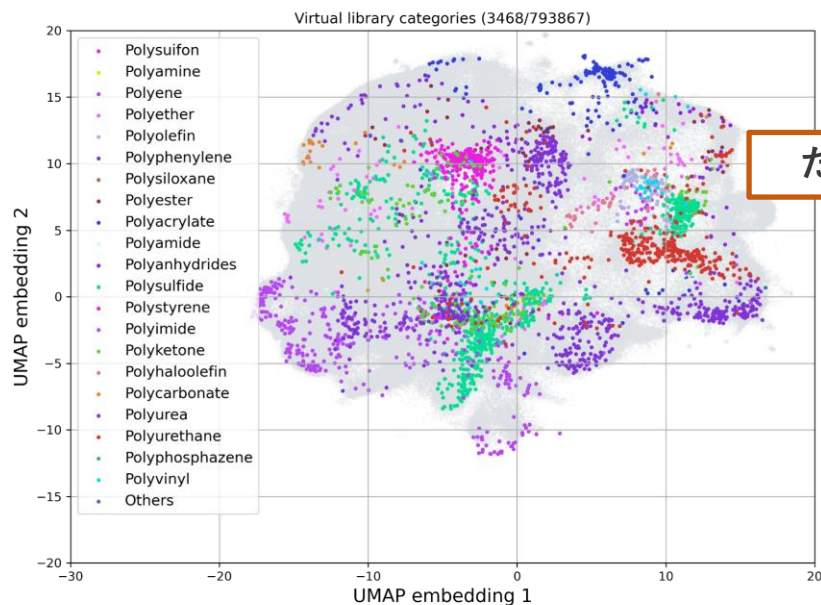


P01	PHYC	Polyolefins (hydrocarbons)	1,708	P11	PURT	Polyurethanes	146,887
P02	PSTR	Polystyrenes	20,233	P12	PURA	Polyureas	55,968
P03	PVNL	Polyvinyls	23,824	P13	PIMD	Polyimides	176,614
P04	PACR	Polyacrylics	95,855	P14	PANH	Polyanhydrides	61,776
P05	PHAL	Polyhalo-olefins	1,614	P15	PCBN	Polycarbonates	38,891
P06	PDIE	Polydienes	778	P16	PIMN	Polyimines	34,658
P07	POXI	Polyethers	91,035	P18	PPHS	Polyphosphazenes	13,762
P08	PSUL	Polysulfides	30,631	P19	PKTN	Polyketones	81,027
P09	PEST	Polyesters	458,961	P20	PSFO	Polysulfones	49,147
P10	PAMD	Polyamides	370,460	P21	PPNL	Polyphenylenes	24,210

富岳の超高火力

- 2022年1月後半にようやく富岳でのパイプライン構築・チューニングが完了
- 参画者が一丸となりデータ生産を本格始動

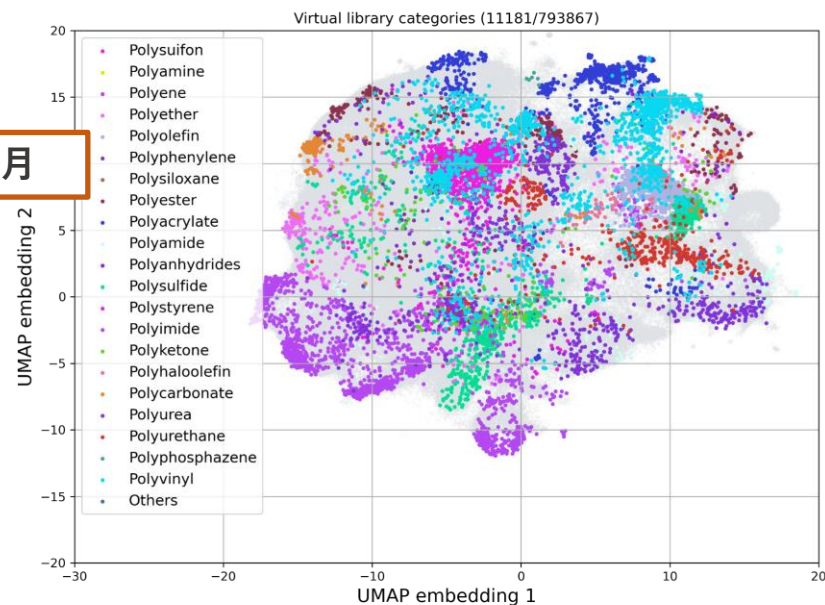
2022/02/05
計算完了ポリマー：3,468



たった一カ月



2022/02/28
計算完了ポリマー：11,181

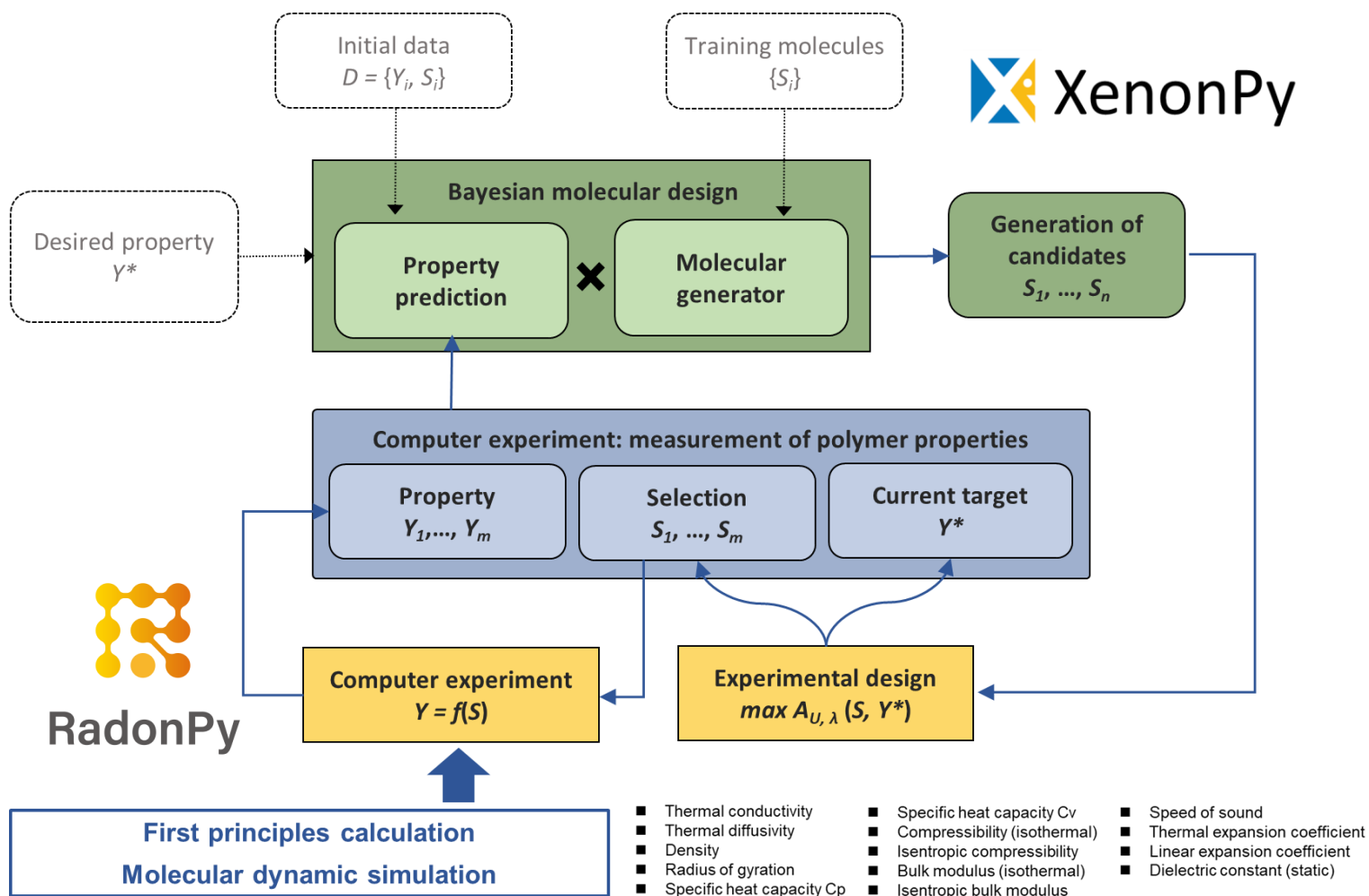


高分子物性自動計算と統計的機械学習の融合

SAPCIER: Virtual Automated Laboratory

Integration of Machine Learning Polymer Design and Automated Computer Experiments

- 適応的実験計画法に基づきデータが存在しない未踏領域を埋めていく
- 高分子物性自動計算 + 機械学習による自動分子設計 は未踏課題

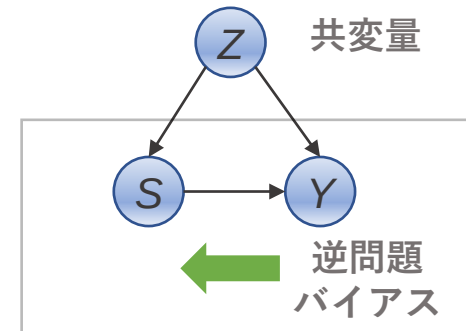


データ駆動型研究に資するデータベースとは

なぜデータ科学の国研が材料データベースを開発するのか？

データ科学への適用に求められるデータベース

- 設計や制御を目的とする場合、モデルは因果効果を適切に表現する必要がある。
- 因果的効果をバイアスなく推定するためのデータ取得系の設計やバイアスを補正に利用できる共変量の情報が必要に



データベースには様々な「選択バイアス」や「交絡バイアス」

- 平衡化に成功しやすいポリマーに偏る
- 作製しやすいサンプルに偏る
- 特定の物性水準やケミカルスペースを重点的に探索したデータ（例：ベイズ最適化）
- 文献情報データベースの場合：出版バイアスや未観測のプロセス条件など

RadonPy データベース共同開発事業コンソーシアム

産学共創のモデルケースを作り出したい

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



RadonPy

RadonPyデータベース共同開発事業コンソーシアム

吉田 亮^{1,2,3,4}

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

⁴ 物質・材料研究機構 統合型材料開発・情報基盤部門

活動内容：運営会議・テクニカルミーティングなど

■ キックオフミーティング（これまでに交流があった研究者・組織に呼びかけ）

- 2021年4月27日

■ テクニカルミーティング（2か月に1回 原則 第2火曜日 15:00-17:00）

- 6月8日
- 8月10日
- 10月5日
- 12月7日
- 2月8日

■ 運営会議（2か月に1回 原則 第2火曜日 15:00-16:00）

- 7月6日
- 9月21日
- 11月9日
- 1月11日
- 3月8日

■ Slack等での情報交換

活動内容：参画者 $\cong$ コントリビュータ

アルゴリズム・コード開発

- ガラス転移温度（企業コントリビュータ）
- コイル・グロビュール転移と溶媒和自由エネルギー（統数研・東京薬科大 山田寛尚）
- レオロジー物性（企業コントリビュータ）
- 量子化学計算（企業コントリビュータ）
- 熱硬化性樹脂（企業コントリビュータ）
- ポリマー重合反応モデル（企業コントリビュータ）

データ生産の分担

- RadonPy及び富岳を初めとする各種計算資源を利用

コンソーシアム会員募集の予定

コンソーシアム規約を作成中

受入体制が整っておらず，一般募集をかけても対応できないかもしれないという不安

アンケートを取り，関連分野のニーズを調査したい．

あるいは個別の面談も受け付けます（yoshidar@ism.ac.jp）