データ駆動型材料研究の諸問題：現状と展望

統計数理研究所 ものづくりデータ科学研究センター

吉田 亮[1,2] yoshidar@ism.ac.jp

XenonPy
https://xenonpy.readthedocs.io/en/latest/

[1] 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
[2] 総合研究大学院大学 複合科学研究科 統計科学専攻

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
The Institute of Statistical Mathematics

ものづくりデータ科学研究センター

2017年7月設立：産学の共創による“ものづくり” の実践

重点分野：マテリアルズインフォマティクス

産学連携を推進力：2020年度 15社 参画者100名以上

共同研究部門

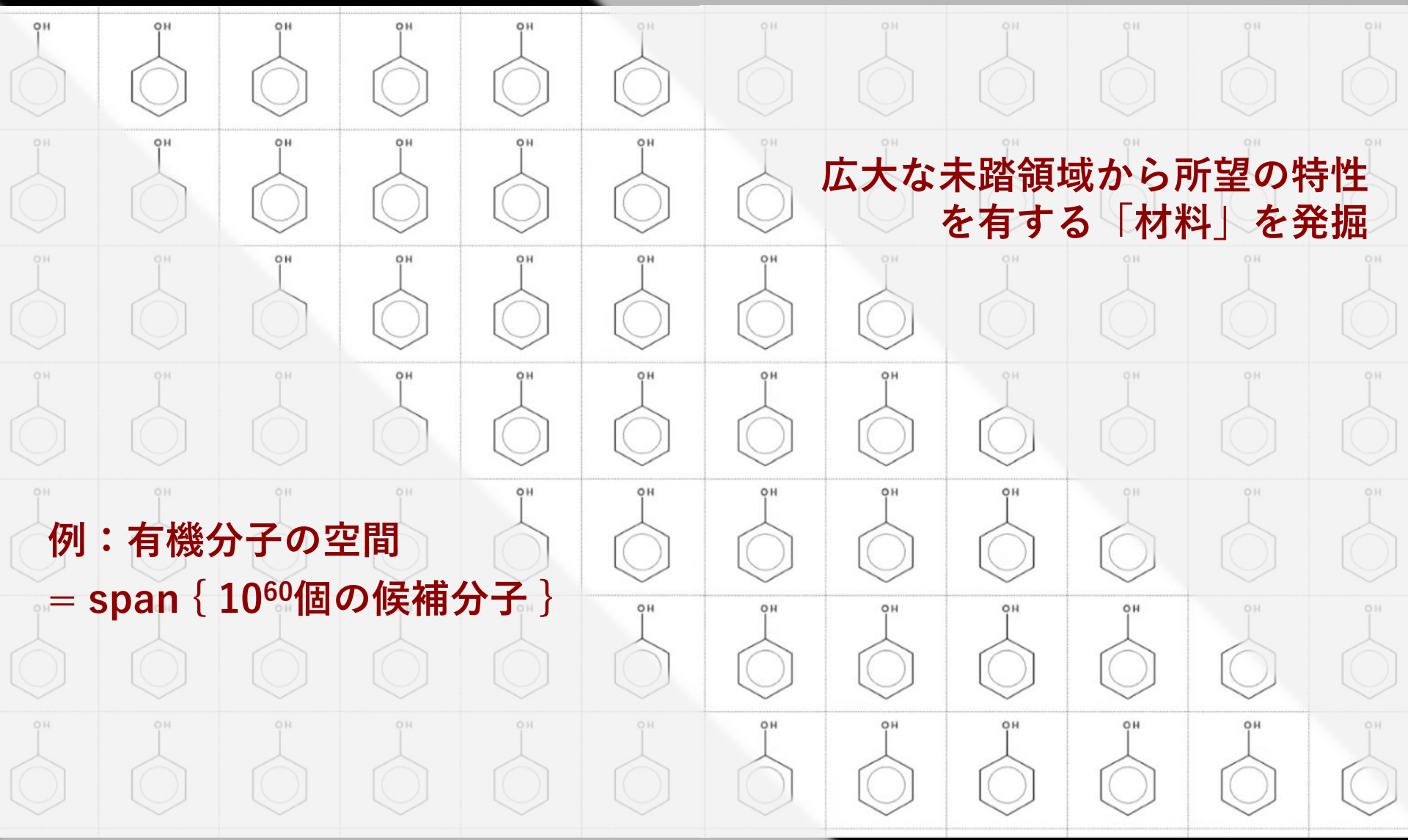
MI Frontier

三菱ケミカル株式会社

ISM-MCCフロンティア材料設計拠点 (2019.10-)

JSR 株式会社

JSR-ISM スマートケミストリーラボ (2020.10-)



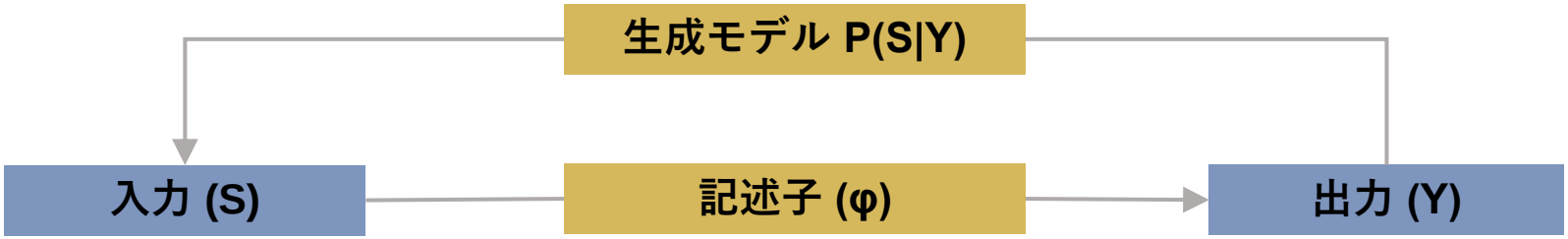
広大な未踏領域から所望の特性
を有する「材料」を発掘

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

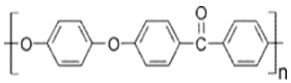
ベクトル表現の方法が非自明な変数（物質・材料）の「表現」と「生成」

$$S = f^{-1}(Y)$$

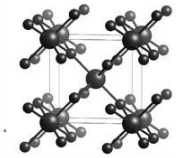


$$Y = f(S) = g^{\theta}\varphi(S)$$

Chemical structure



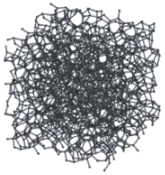
Crystal



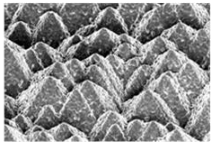
Composition

$Sr_2Ti_2O_6$

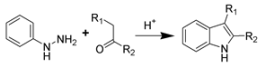
Amorphous



Microstructure

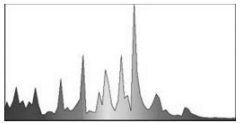


Chemical reaction



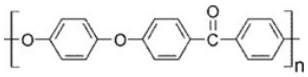
Θ, T, P

Spectrum



順問題：高分子の繰り返しユニットから材料特性を予測

構造 (S)



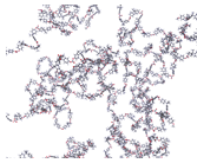
記述子 (φ)

01000100000011010011

Molecular fingerprint
Graph neural networks
Graph kernels
etc

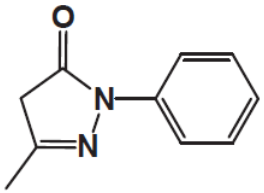
物性・高次構造 (Y)

Thermal conductivity
Dielectric property
Tg, Tm
Cp



Deep learning
Sparse modeling
Ensemble learning
etc

e.g. **molecular fingerprint**

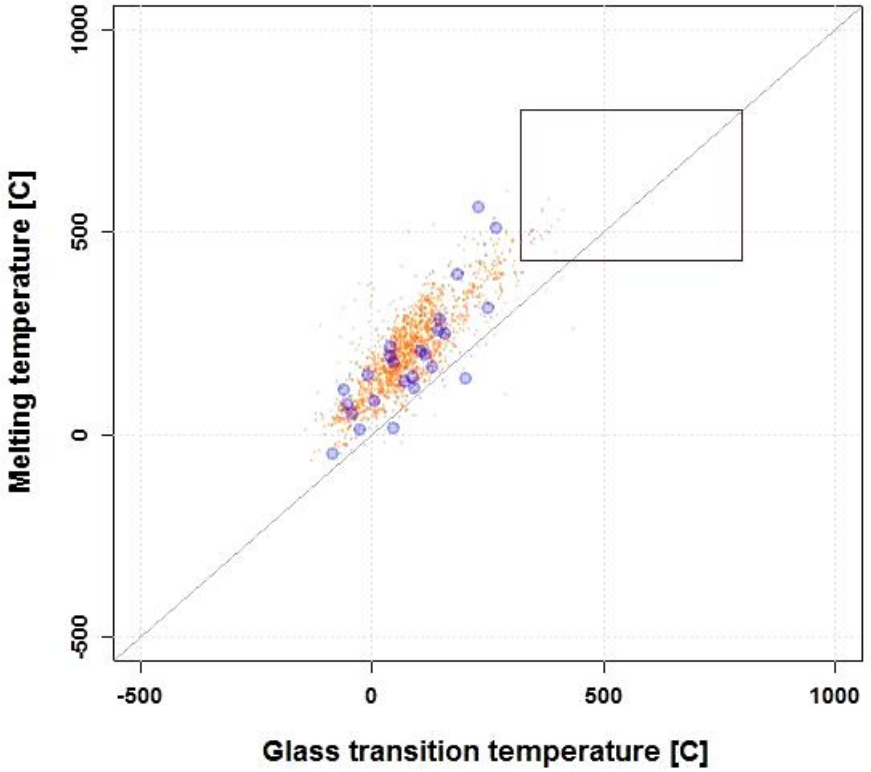


$$f \in \boldsymbol{F}$$     .....

$$\phi(S) = (\ 1,\quad 0,\quad 1,\quad 1,\quad 1, \cdots\cdots\ )$$

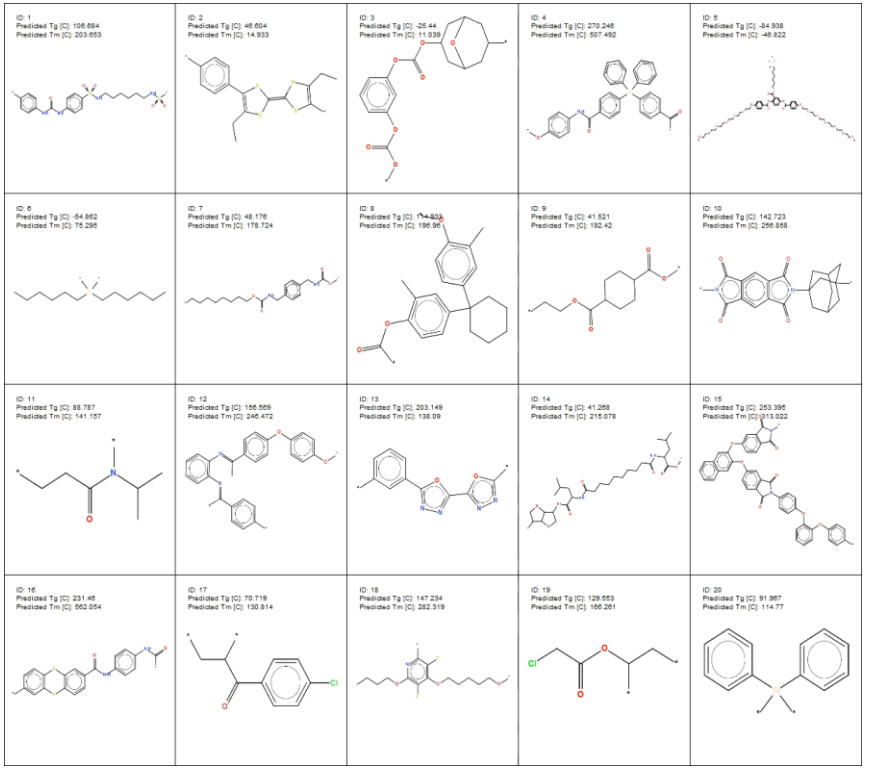
逆問題：所望の特性を持つ物質を設計

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

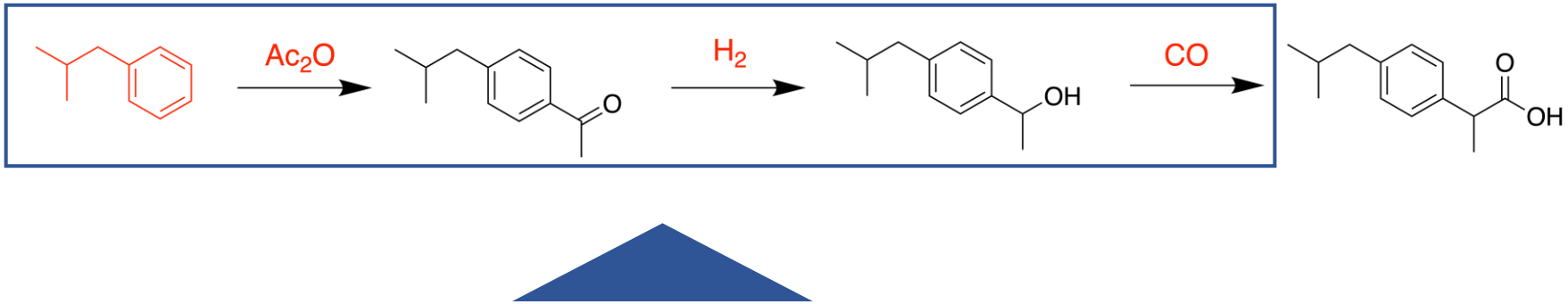


- Ikebata et al. Bayesian molecular design with a chemical language model. *J Compt Aided Mol Des*. 31:379-391 (2017)
- Wu et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput Mater*, 5:66 (2019)
- Wu et al. iQSPR in XenonPy: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

機械学習で目標分子の合成経路を策定する

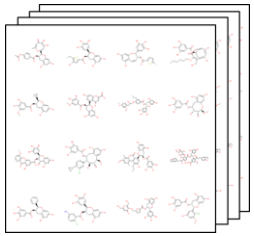
商用化合物のリストから組み合わせを選択

目標分子



商用化合物のカタログ (10^6-10^7)

- 出発化合物
- 反応物
- 触媒・溶媒
- 温度・圧力など



順問題: 合成反応の生成物の予測

機械翻訳の問題に定式化し，深層学習で生成物を予測する

Input

Reactant 1	Reactant 2	Reagent
Nc1nnc(C2CCC2)s1	. O=C(Cl)Cl	> A_CCOC(C)=O

→

Output

Product
O=C=Nc1nnc(C2CCC2)s1

[SMILES] Seq2Seq (https://github.com/google/seq2seq)
[SMILES] Transformer (https://github.com/pschwllr/MolecularTransformer)
[Graph] Rexgen (https://github.com/connorcoley/rexgen_direct)

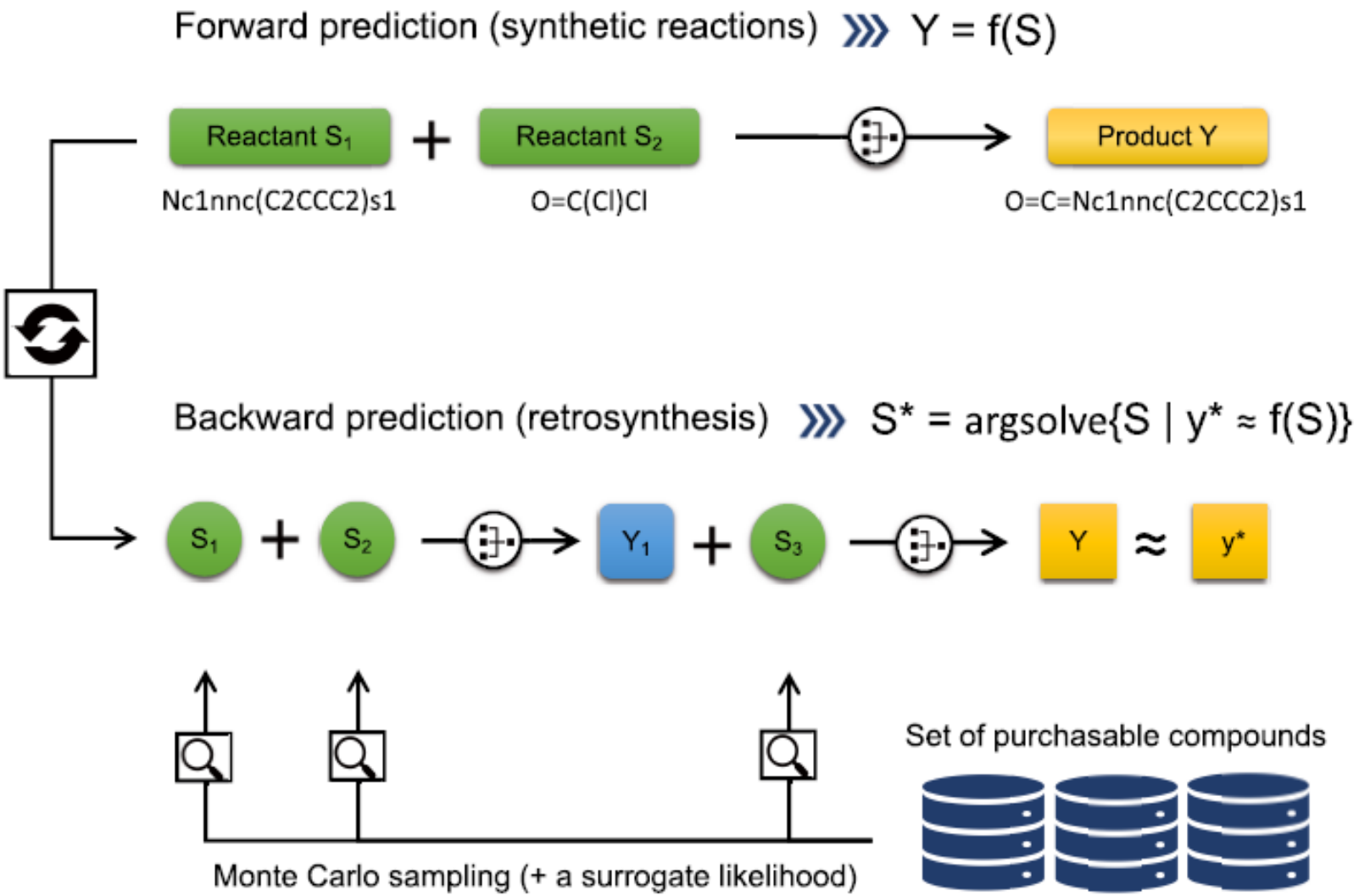
Table 1: Performance of recently appeared deep neural networks on the prediction of synthetic reactions in forward and backward manners. Top-1, top-3, top-5 and top-10 accuracies in [%] are shown for each.

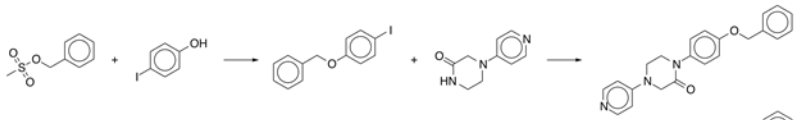
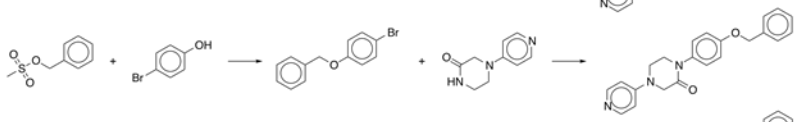
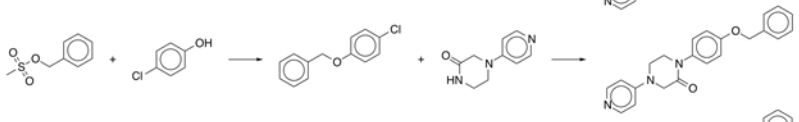
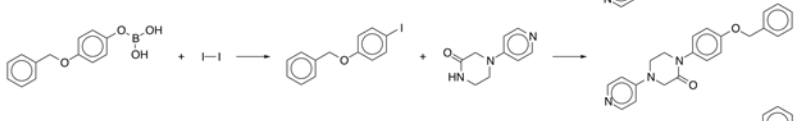
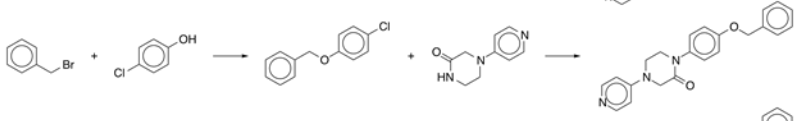
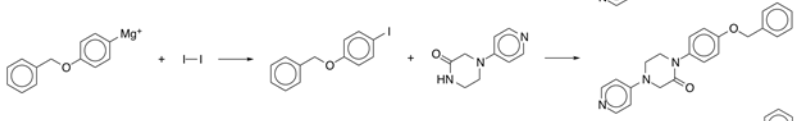
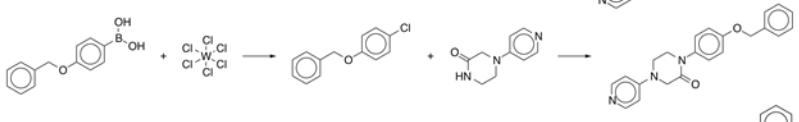
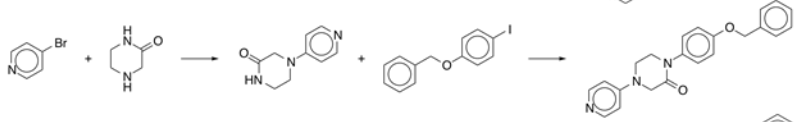
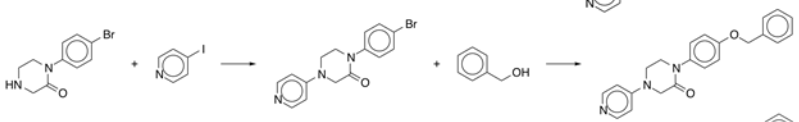
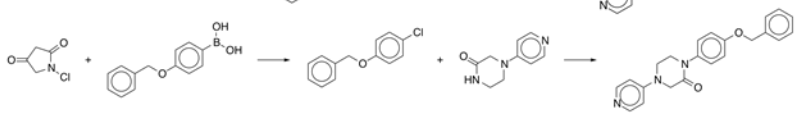
Task	Model	top-1	top-3	top-5	top-10
Backward	Similarity (Coley et al. 2017)[6]	37.3	54.7	63.3	74.1
	SCROP (Zheng et al. 2019)[8]	43.7	60.0	65.2	68.7
	Lin et al. 2019[9]	43.1	64.6	71.8	78.7
Forward	Template-based (Coley et al. 2017)[13]	71.8	86.7	90.8	94.6
	WLDN (Jin et al. 2017)[14]	79.6	87.7	89.2	-
	Molecular Transformer (Schwaller et al. 2019)[10]	90.4	94.6	95.3	-

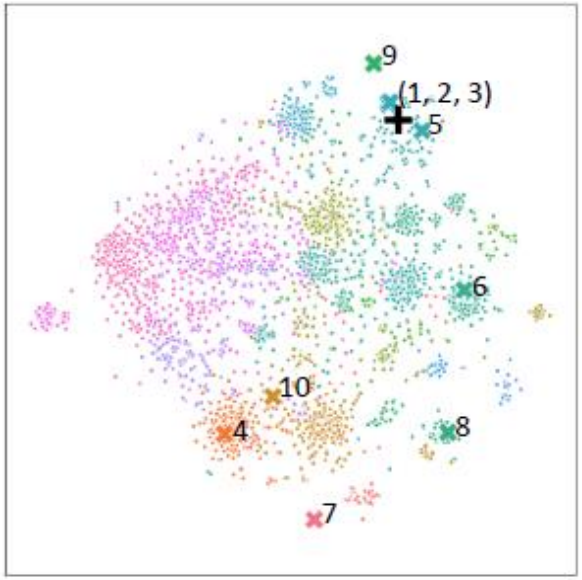
Table 1 in Guo et al. Bayesian algorithm for retrosynthesis. JCIM 60(10):4474–4486 (2020)

逆問題: 目標化合物に達する合成経路を予測

Guo et al. Bayesian algorithm for retrosynthesis. JCIM 60(10):4474–4486 (2020)



Candidates of synthetic route	Target	Step 1	Step 2
		1	1
		1	1
		1	1
		3	1
		1	1
		1	1
		3	1
		1	1
		1	1
		3	1



t-SNE projections of identified reactant pairs which formed nearly 98 clusters

- Identification of nearly 100 candidate routes based on 10^6 purchasable compounds
- 35-60% would be chemically valid according to judgments made by expert chemists

Fig 4 in Guo et al. Bayesian algorithm for retrosynthesis. JCIM 60(10):4474–4486 (2020)

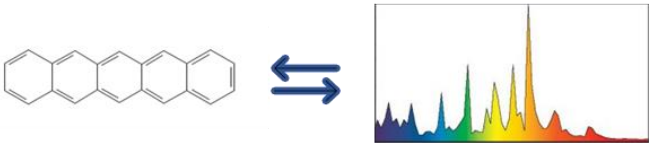
データ科学のユニークな視点から「問題設定」を発掘

組成・プロセス ⇆ **微細組織**

$Cr_x Al_y O_z N$
Θ, T, P

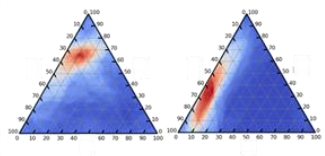
Banko et al. Commun Mater 1:15 (2020)
Yang et al. J Mech Des 140(10): 111416 (2018)
Li et al. Sci Rep. 8:13461 (2018)

化学構造 ⇆ **スペクトル**



化学組成 ⇆ **準結晶**

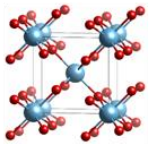
$Al_xTM_yTM_z$



Liu et al. Adv. Mater. (2021) in press

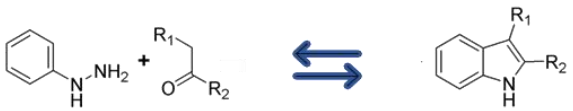
化学組成 ⇆ **結晶構造・対称性**

TiO_2



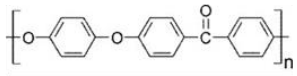
Yamashita et al. Phys Rev Mater 2:013803 (2018)
Yamada et al. ACS Cent. Sci. 5(10):1717-1730 (2019)

反応物 ⇆ **生成物**



Guo et al. J. Chem. Inf. Model. 60(10):4474–4486 (2020)
Segler et al. Nature. 555:604–610 (2018)
Mikulak-Klucznik et al. Nature 588, 83–88 (2020)

結晶構造・高分子 ⇆ **物性**

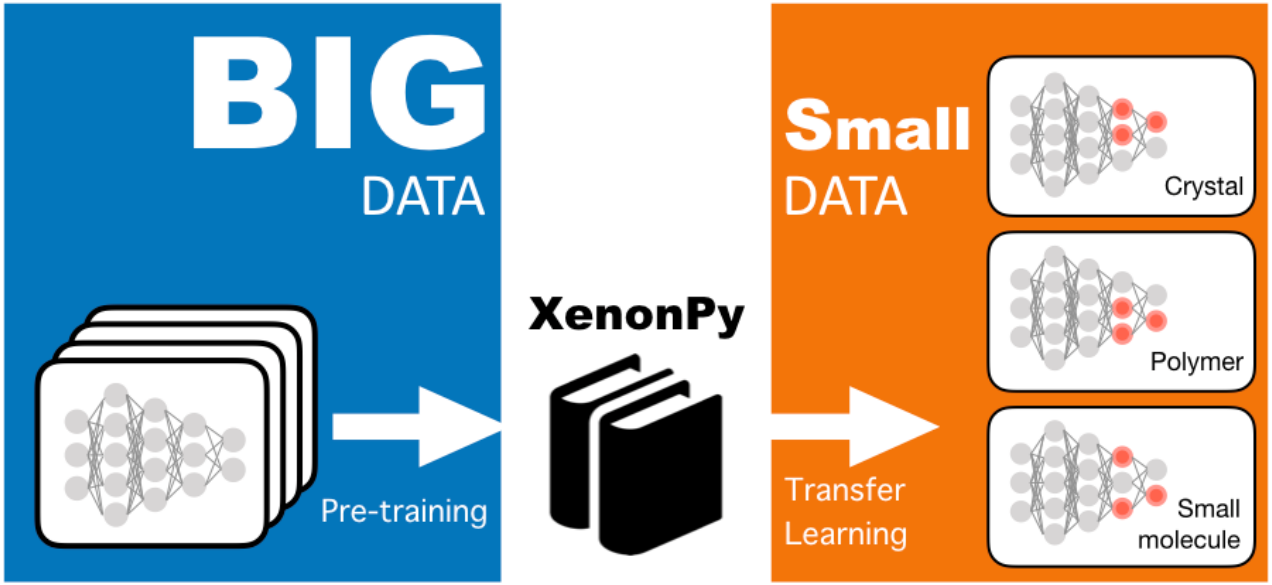


Tg, Tm, Cp

Yamada et al. ACS Cent. Sci. 5(10):1717-1730 (2019)
Ju et al. Phys. Rev. Mater. 5:053801 (2021)
Minami et al. AAAI (2021) in press
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)

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多くの領域



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.

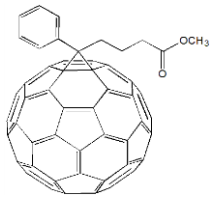
そもそも革新的な材料の周辺にはデータが存在しない

例：有機薄膜太陽電池のドナー分子

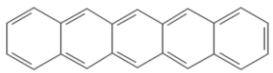
Acceptor = PCBM

Donners = design targets

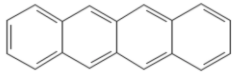
PCBM



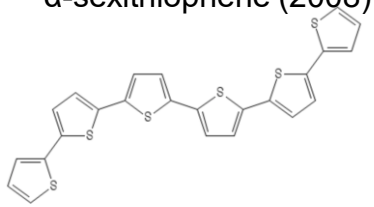
pentacene (2004)



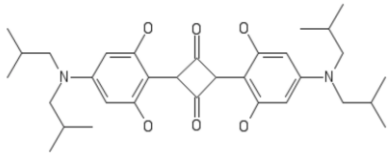
tetracene (2005)



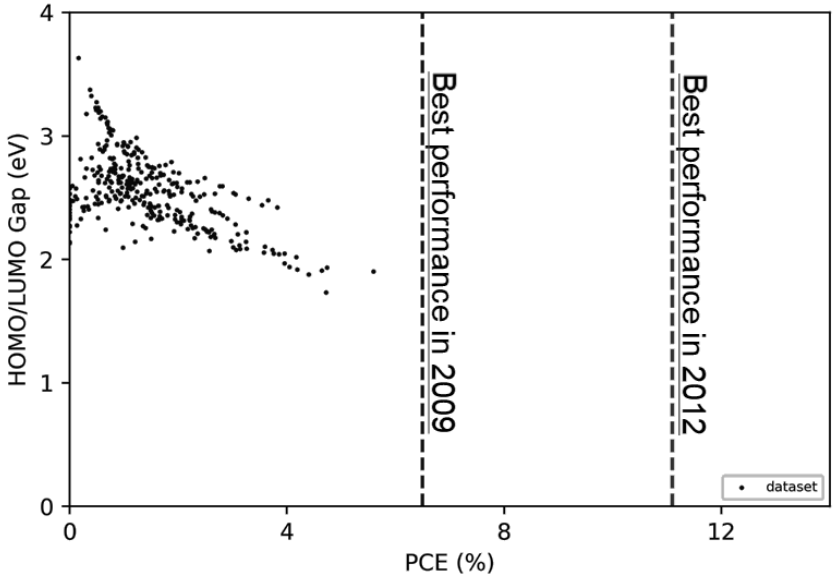
α-sexithiophene (2008)



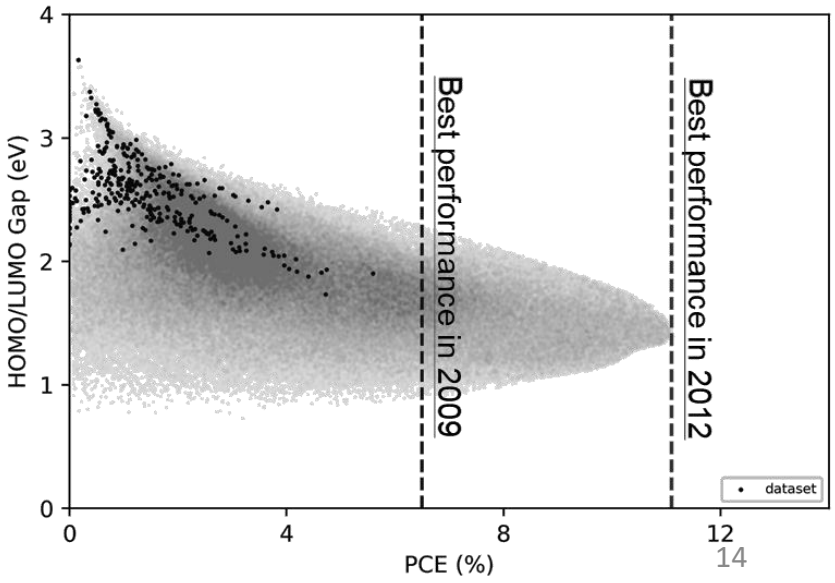
2,4-Bis[4-(*N*,*N*-diisobutylamino)-2,6-dihydroxyphenyl] squaraine (2009)



2009年の性能分布



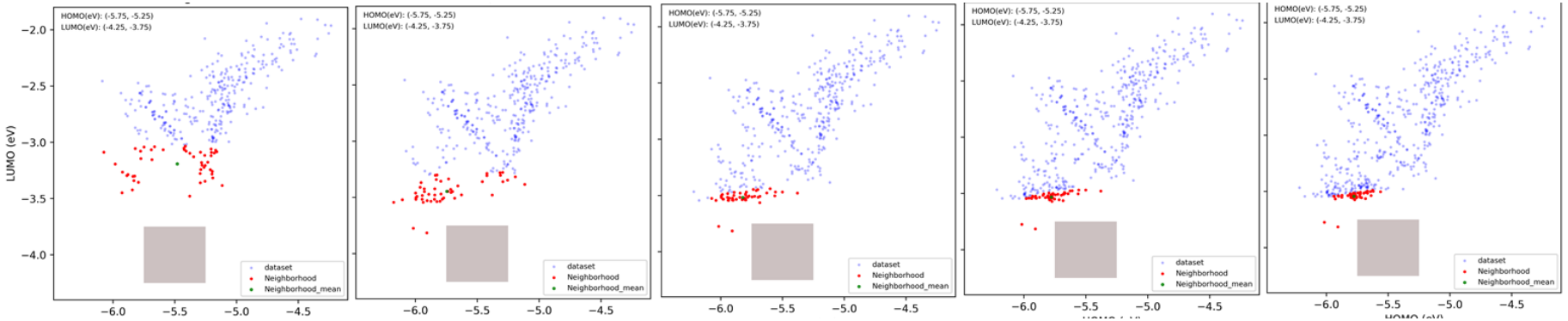
2012の性能分布



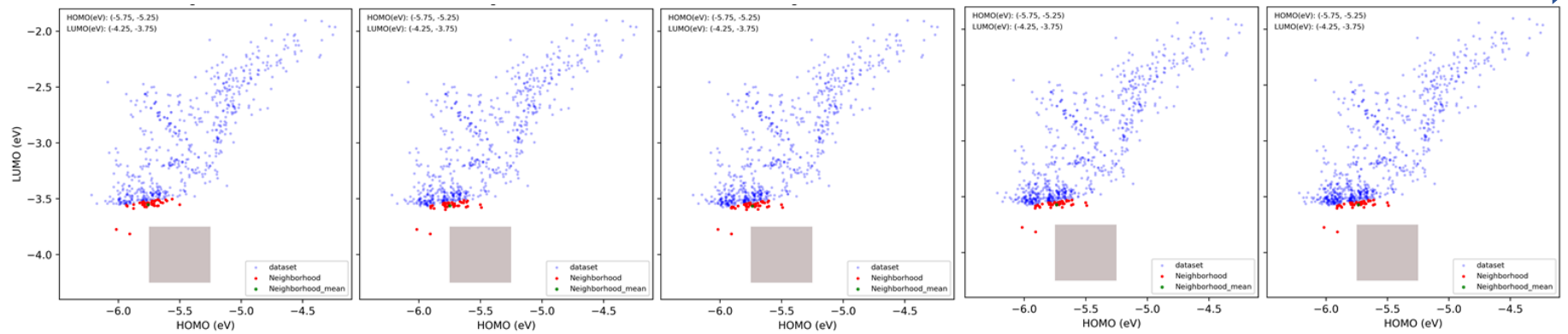
データ科学の内挿的予測の限界

「入力が近ければ，出力も近い」という予測をやっているに過ぎない

t = 1



t = 6



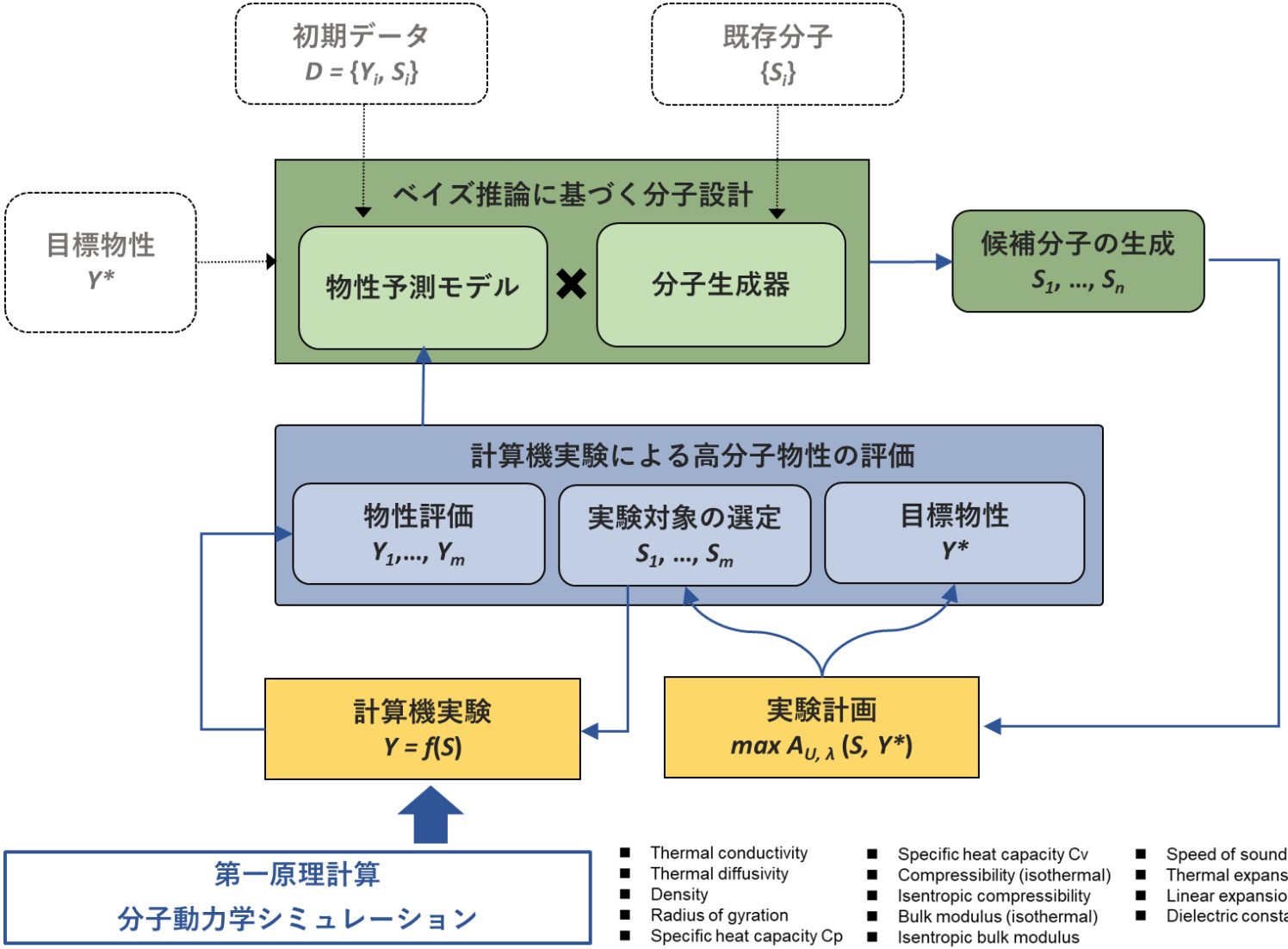
Scharber's formula for PCBM (Adv. Mater. 18, 789 (2006))

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

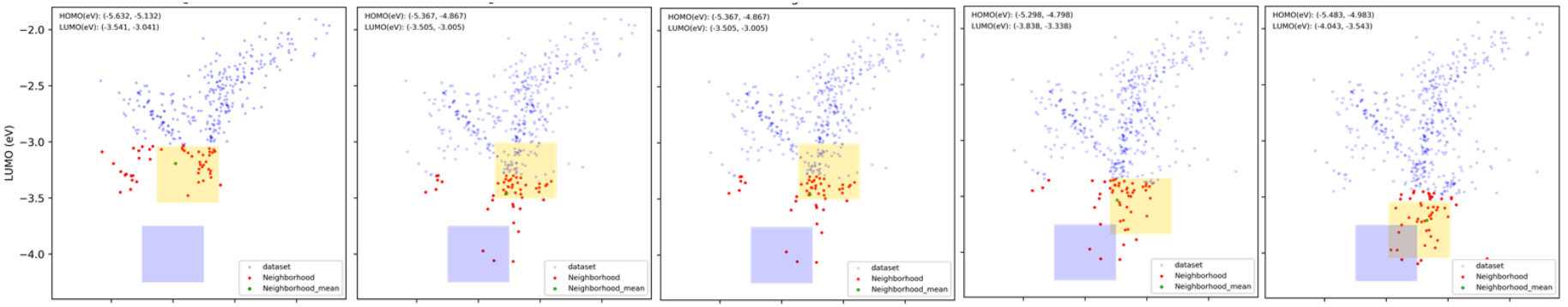
Y-axis: LUMO
X-axis: HOMO
Blue: generated molecules
Red: molecules selected for DFT
Gray zone: final target

SPACIER Go Beyond Interpolative Prediction

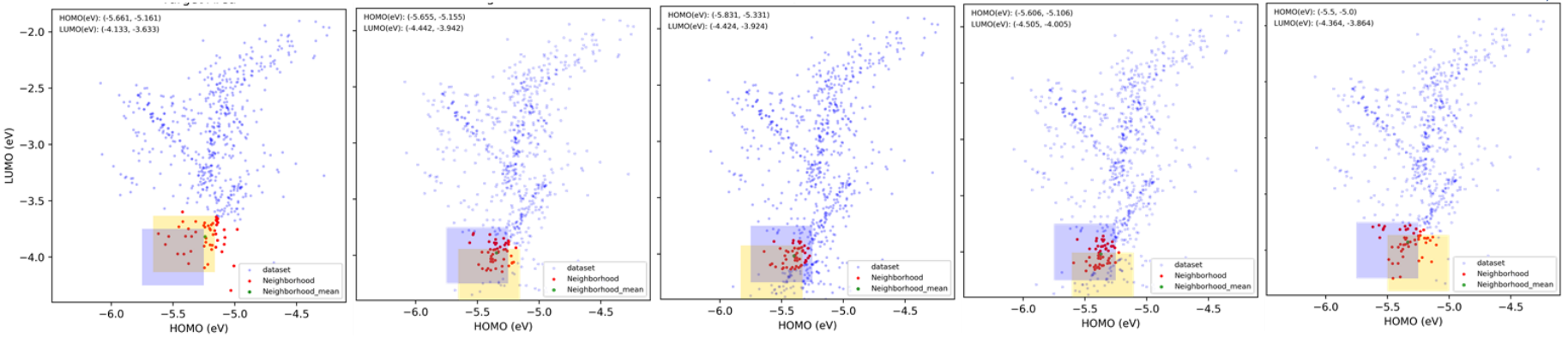
実験計画法に基づいて機械学習アルゴリズムを自動実験システムに組み込むことが主流

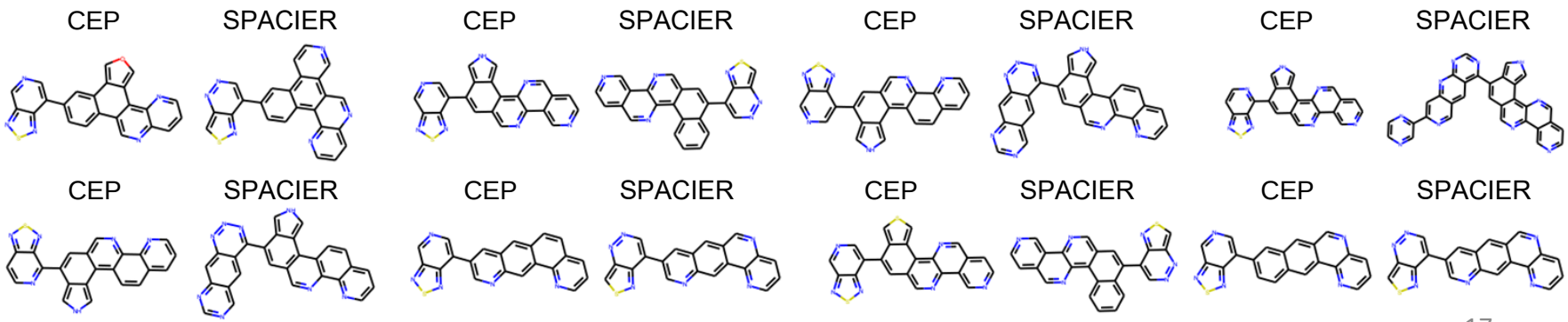


t = 1


HOMO(eV): (-5.632, -5.132)
LUMO(eV): (-3.541, -3.041)
HOMO(eV): (-5.367, -4.867)
LUMO(eV): (-3.505, -3.005)
HOMO(eV): (-5.367, -4.867)
LUMO(eV): (-3.505, -3.005)
HOMO(eV): (-5.298, -4.798)
LUMO(eV): (-3.838, -3.338)
HOMO(eV): (-5.483, -4.983)
LUMO(eV): (-4.043, -3.543)
LUMO (eV)
dataset
Neighborhood
Neighborhood_mean

t = 6


HOMO(eV): (-5.661, -5.161)
LUMO(eV): (-4.133, -3.633)
HOMO(eV): (-5.655, -5.155)
LUMO(eV): (-4.442, -3.942)
HOMO(eV): (-5.831, -5.331)
LUMO(eV): (-4.424, -3.924)
HOMO(eV): (-5.606, -5.106)
LUMO(eV): (-4.505, -4.005)
HOMO(eV): (-5.5, -5.0)
LUMO(eV): (-4.364, -3.864)
LUMO (eV)
HOMO (eV)
dataset
Neighborhood
Neighborhood_mean


CEP
SPACIER
CEP
SPACIER
CEP
SPACIER
CEP
SPACIER
CEP
SPACIER
CEP
SPACIER
CEP
SPACIER
CEP
SPACIER

ラボオートメーション：三つの問い

どこまでハイスループット化が進行するか？

- 実験装置のハイスループット化
- 計算機実験（シミュレーション）

コモディティ化へのシナリオ

- 現状では，資本力のあるラボ・企業しか導入できない
- 大型共用施設・研究者によるラボ内DIY

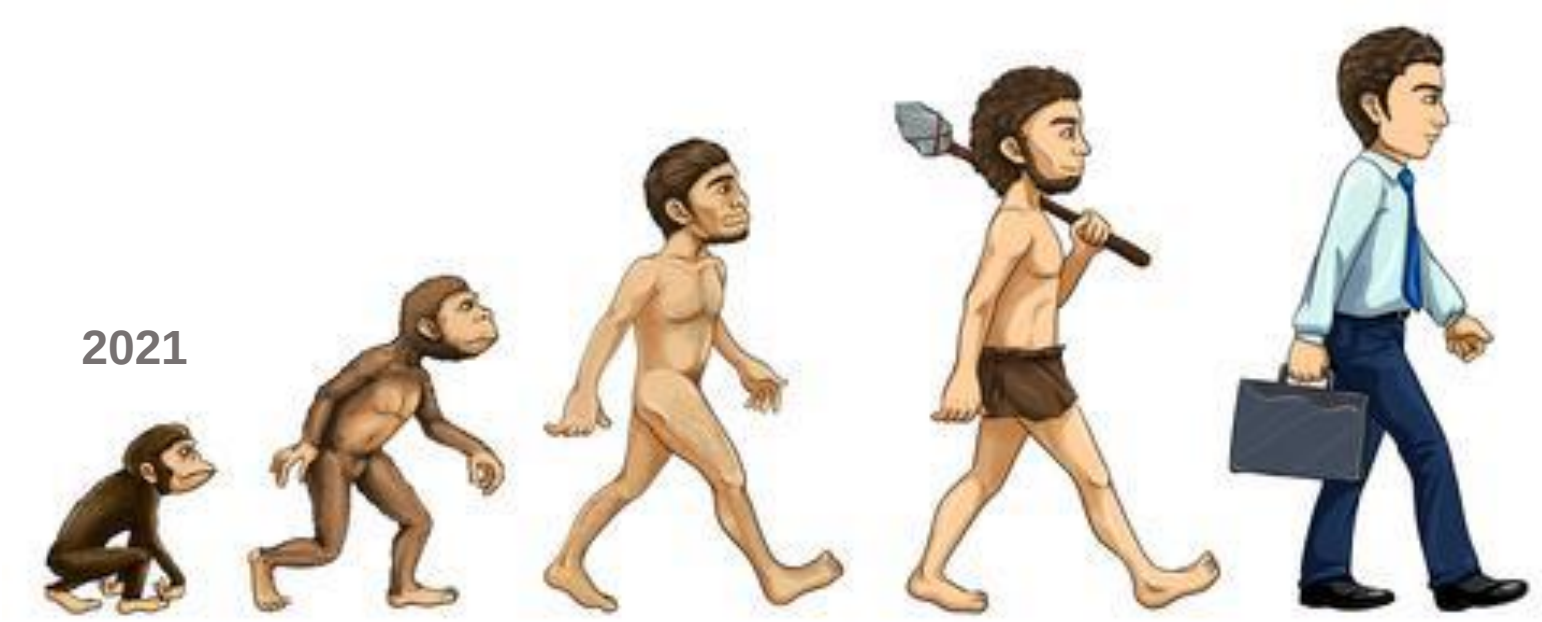
メリット・デメリット

- 労働集約的ラボ運営，施設稼働率改善，低再現性・研究不正防止，危険な実験
- 物質特許のすり抜け，資本の格差

データの量と多様性は単調に増加する

Data is the “infinitely increasable” oil

- **データ駆動型科学の究極：データを持つものが勝つという資本のゲーム**
- **化学や材料科学には「共有」の文化がない** (Mat Todd, The University of Sydney, Nature Digest 11. 2014)
- **格差緩和と技術向上のに，コミュニティが一体となってコモンデータを共創しようという機運 ... 全くと言っていいほど高まっていない**



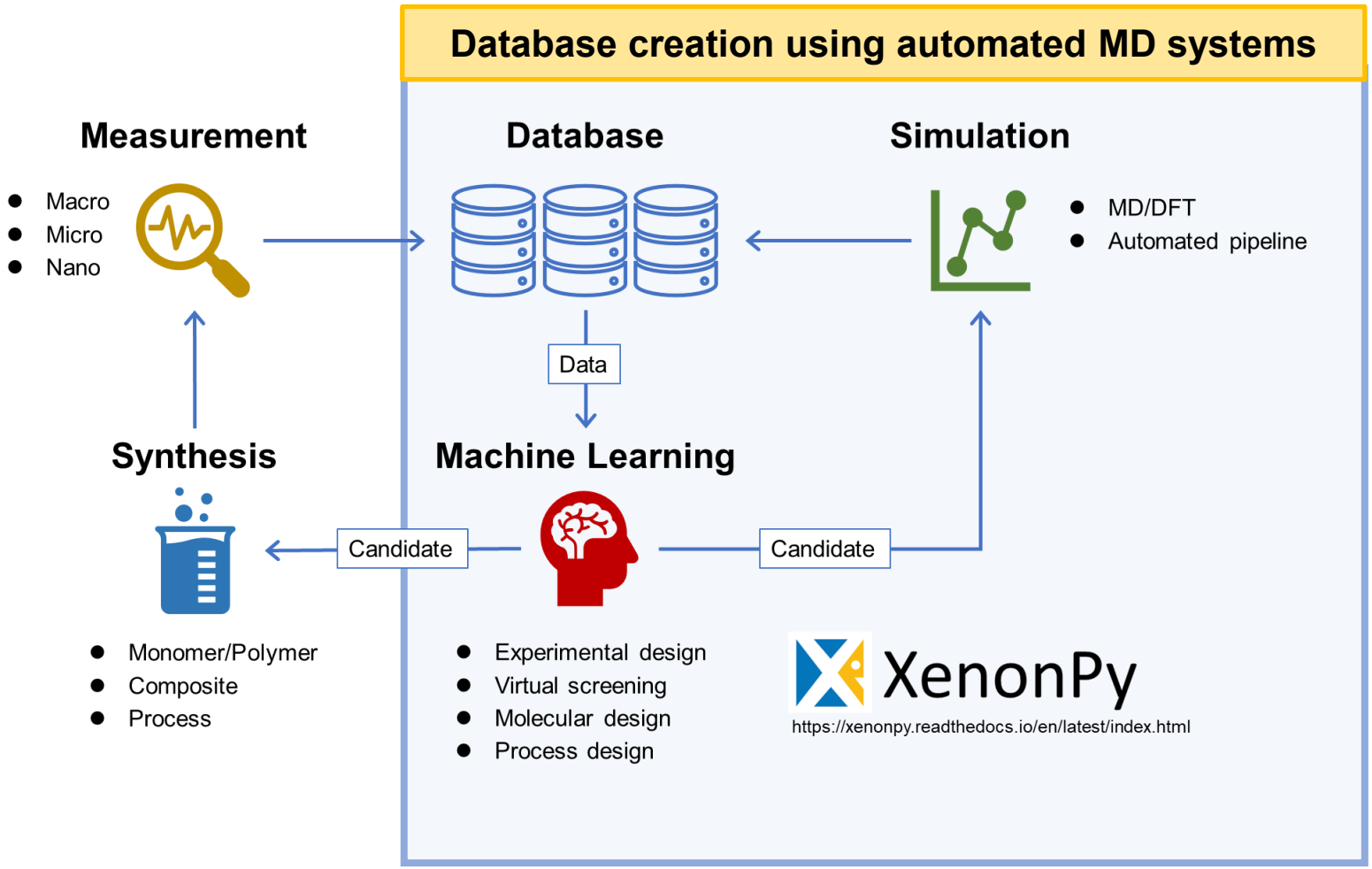
データ量と多様性は単調に増大 ➔ 格差の増大

データ駆動型研究に資する高分子物性データベースが不足

NO DATA, NO DATA SCIENCE

データベース	概要
PoLyInfo (polymer.nims.go.jp)	国立研究開発法人物質・材料研究機構が提供している学術文献から抽出したデータをまとめた高分子物性データベース（18,044件の文献データ）．18,015種類のモノマーから重合されたポリマー群の物性データ367,711点を収録している．
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種類のポリマー/有機材料と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)	CHiMaDが提供しているデータベース．文献から抽出した263件のFlory-Huggins $¥chi$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，物性を含むデータベースならびにデータ共有のためのプラットフォーム．
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``¥textit{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．
CAMPUS (campusplastics.com)	ポリマー9,236種を含む市販の材料特性データベース．

高分子インフォマティクスの学術基盤の構築



JST-CREST “Materials Informatics for Thermophysical Properties of Polymers” (2019.10-) Morikawa, J. (Tokyo Tech Inst)
ML & CS group：Yoshida, R., Hayashi, Y., Wu, S., Noguchi, Y., Shiomi, J., Kawauchi S.

分子動力学に基づく高分子物性測定の全自動化

林慶浩
(ISM)

最新版：14物性の自動計算に対応

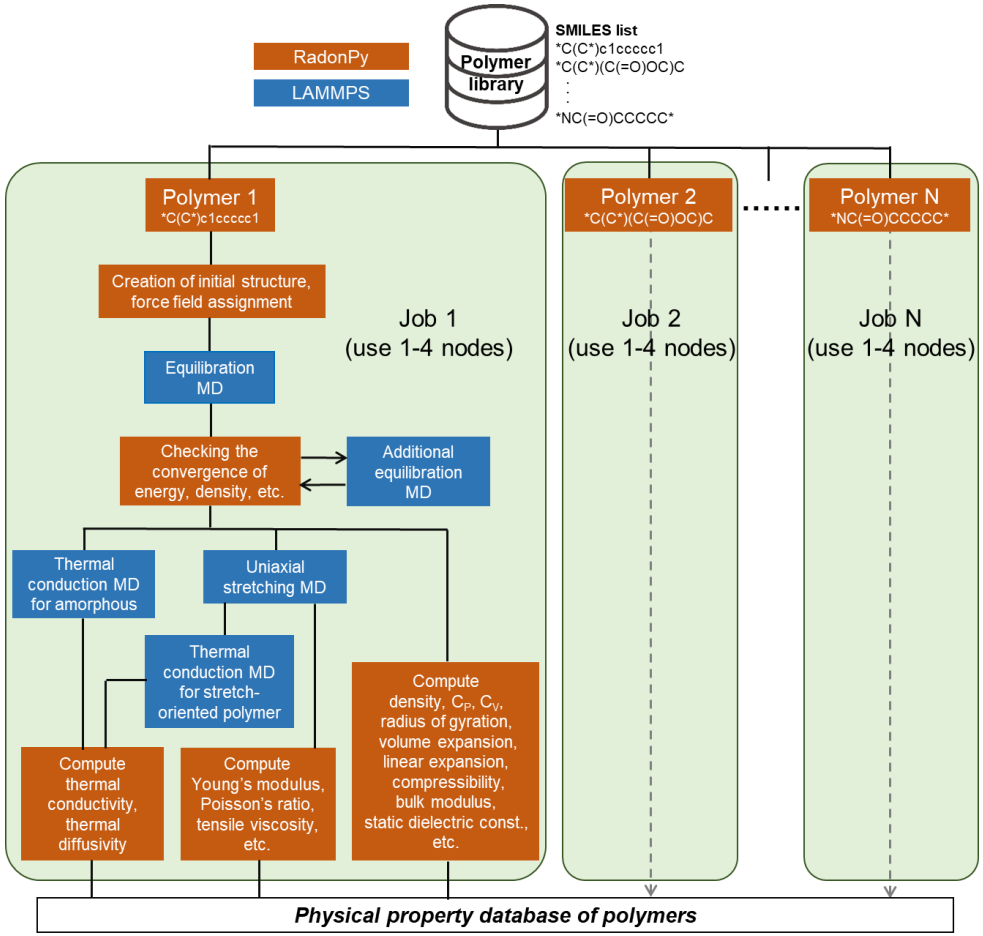
- Thermal conductivity
- Thermal diffusivity
- Density
- Radius of gyration
- Specific heat capacity Cp
- Specific heat capacity Cv
- Compressibility (isothermal)
- Isentropic compressibility
- Bulk modulus (isothermal)
- Isentropic bulk modulus
- Speed of sound
- Thermal expansion coefficient
- Linear expansion coefficient
- Dielectric constant (static)

ポリマーの繰り返し単位の化学構造を入力し，力場の割り当て，初期構造の生成，エラー処理，平衡・非平衡MD計算による物性評価までの全工程を完全に自動化

最新版では，アモルファスポリマーや配向したポリマーに対し，14種類の物性を自動計算できる．

RadonPy (open-source software)

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の自動化を実行するPythonライブラリ



計算資源の確保が鍵

富岳，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

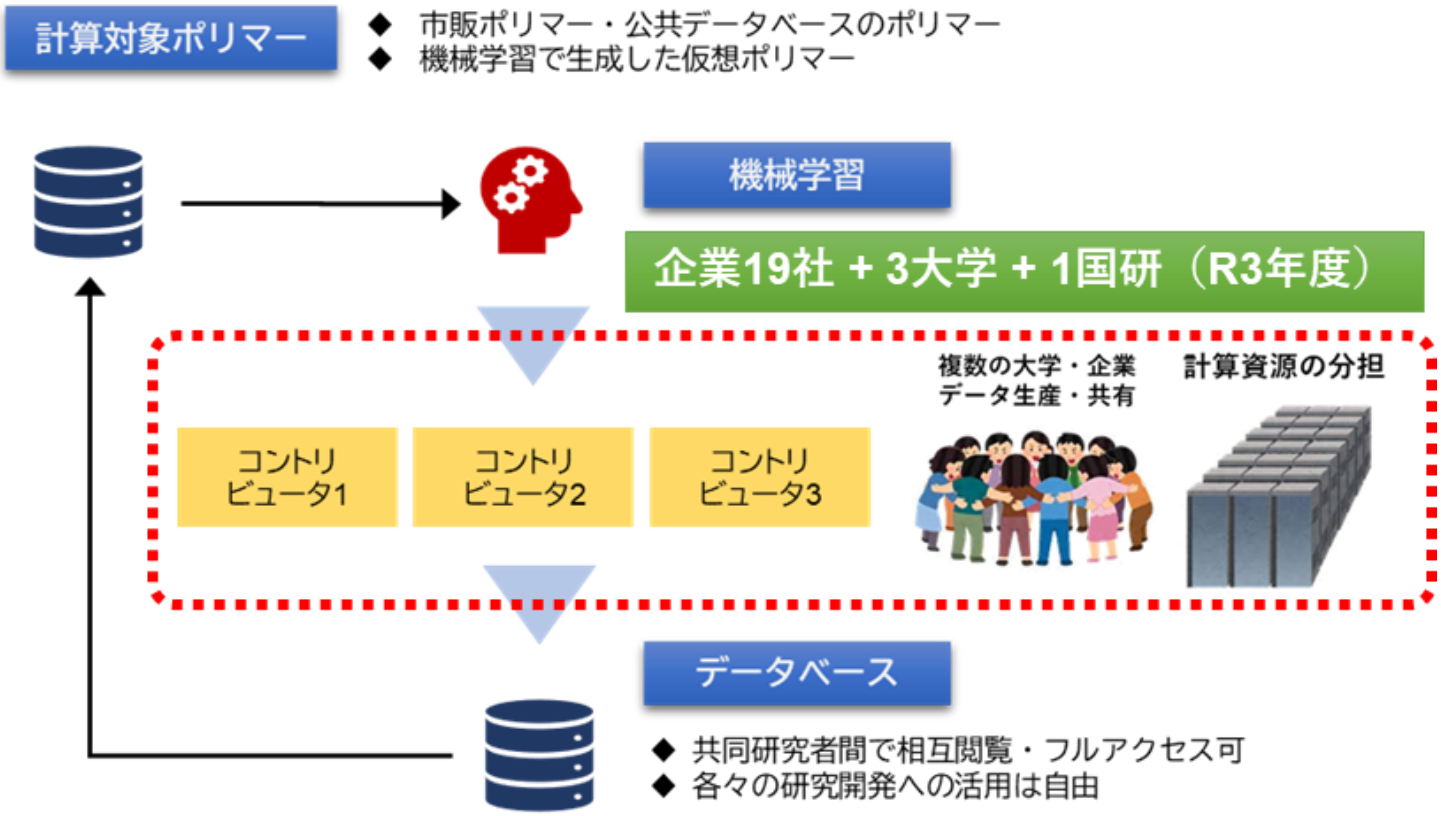
	自然科学研究機構 計算科学研究センター NEC LX 2U-Twin2 サーバ406Rh-2	産業総合研究所 ABCI	東京大学物性研究所 ohtaka
ノード数 / ポリマー	1	1	1
CPU / ノード	12 core	4 core	128 core
GPU / ノード	1 GPU	1 GPU	-
メモリ/ ノード	192 GB	60 GB	256 GB
実行時間 / ポリマー (アモルファス構造)	30-50 時間	32-50 時間	36-44 時間
実行時間 / ポリマー (延伸配向構造)	+ 約100 時間	+ 約100 時間	+ 約100 時間

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

目標データ数：10^5-10^7ポリマー / 5年のデータ取得

- 大学・企業・国研が組織の垣根を越えてデータを共同生産・共有
- 計算資源の分担（スパコン・大規模クラスタシステム）



コントリビュータ種別

① データ生成
（計算資源）

ノルマ = X samples per year

② コード開発

ノルマなし

③ 研究資金

ISMがデータ生成を代行

参画メリット

データやコード
フルアクセス可

API

RadonPy

研究開発への
利用は自由

技術・人的交流
テクニカルミーティング

データの共有・共創を目的とする連合体形成

10^5-10^7個の高分子骨格が張るケミカルスペースを観測

- 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
- 特異な高分子を発見

民主的なデータベース開発：モデルケースの発信

- 組織の垣根を超えた協調の機運を高める布石に
- ボトムアップ的にデータの共創を目的とする小中規模の連合体が生まれることを期待