ケモ/マテリアルズインフォマティクスの現状と将来における展望

東京大学大学院工学系研究科化学システム工学専攻
（兼務）奈良先端科学技術大学院大学
データ駆動型サイエンス創造センター・研究ディレクター
船津　公人

2018年10月24日
国際高等研究所
第26回エジソンの会

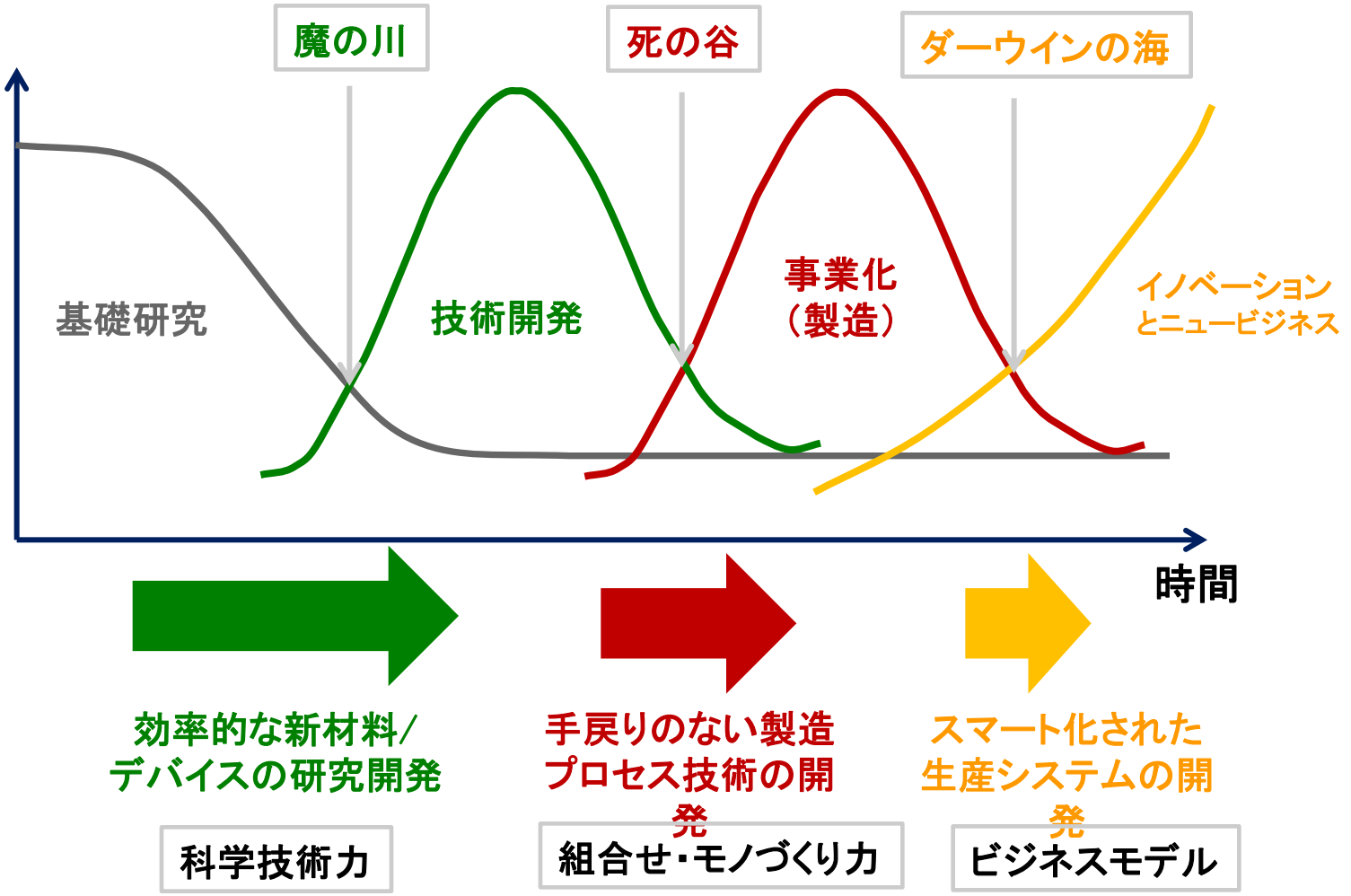
講演内容

- 材料・デバイス系ものづくりにおけるイノベーション
- ケモインフォマティクスとは
- データのモデル化法について
- モデルの逆解析について
- 分子設計の事例紹介
  - 医薬品分子設計
    - ・化学空間の可視化による構造探索
- ポリマーアロイを題材とした有機材料の設計
  - ・輸液用液体パックの設計
  - ・輝度向上性フィルムの成分組成とプロセス条件の最適化
- ソフトセンサー
  - ・化学プラントの監視のためのソフトセンサー
- プロセス・インフォマティクス
  - ・材料設計・プロセス設計・品質管理と制御の連動
- 最後にもうちょっと

材料・デバイス系ものづくりにおけるイノベーション

材料・デバイス系ものづくりにおけるイノベーション

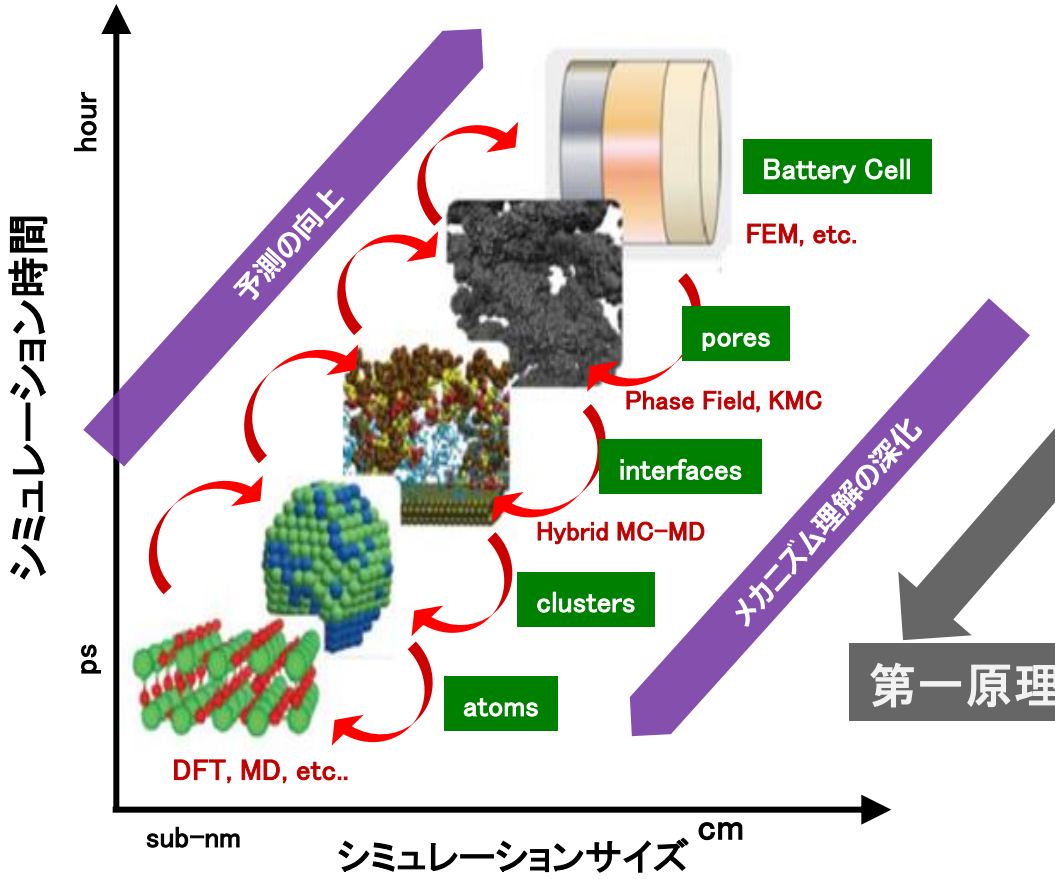
■基礎研究からイノベーションへは、いくつもの障壁を乗り越えなければならない。



材料からデバイス・プロセスへ

■ 材料からデバイスへの効率的な開発にはマルチスケールシミュレーションが有用である。

エネルギーデバイスにおけるマルチスケールシミュレーション



実験(計測データも含む
（ハイスループット）

経験的モデル ＝ 精度を高めるには、多量の実験データが必要

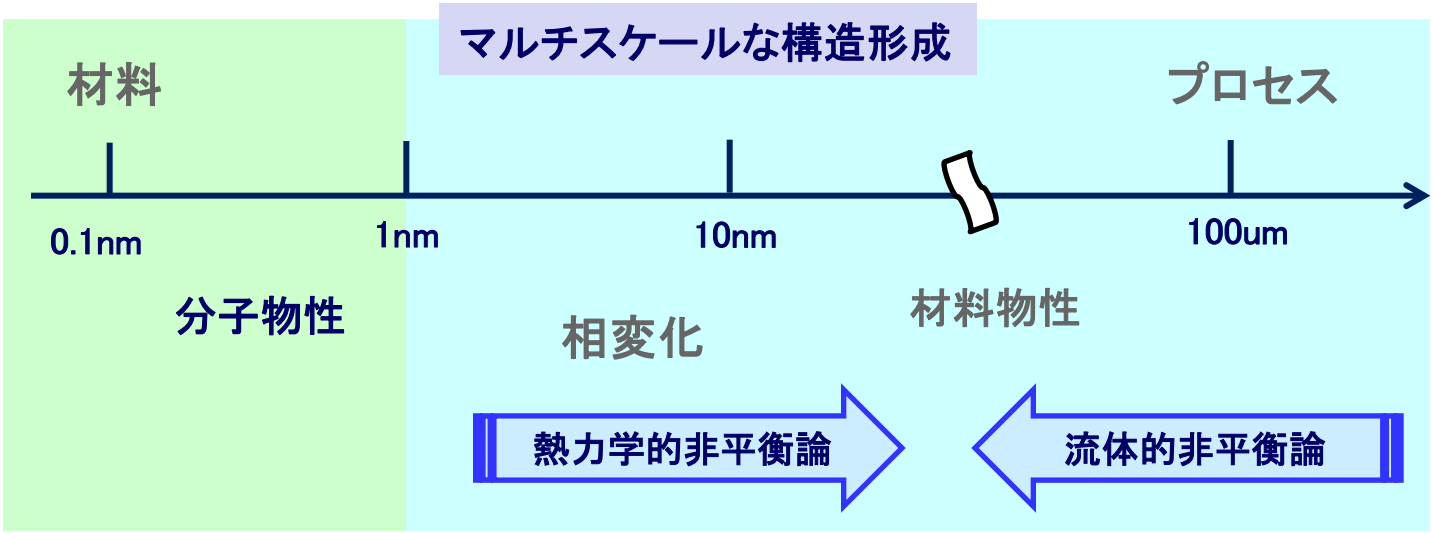
データ科学 x 計算科学

第一原理 ⇔ 理論

図は、A.A.Franco, RSC Advance, 3(2013)p.13027より引用/改変

材料開発からプロセス開発へ

■マルチスケールな構造形成を理解する効率的な課題解決ツールが求められている。



【製品の機能発現プロセスにおける課題】
- 材料と個々のプロセス装置の相互関係が強い 。
- (複合化しているため)材料物性の評価が難しい。
- ものづくりの現場主導 ←解析が不十分で感性と経験に依存

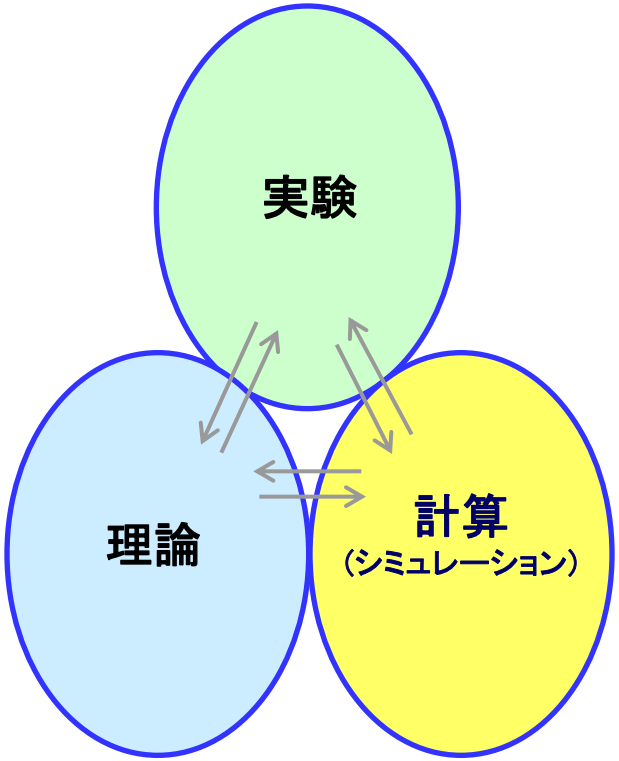
➡ この過程で膨大な実験データが蓄積される。

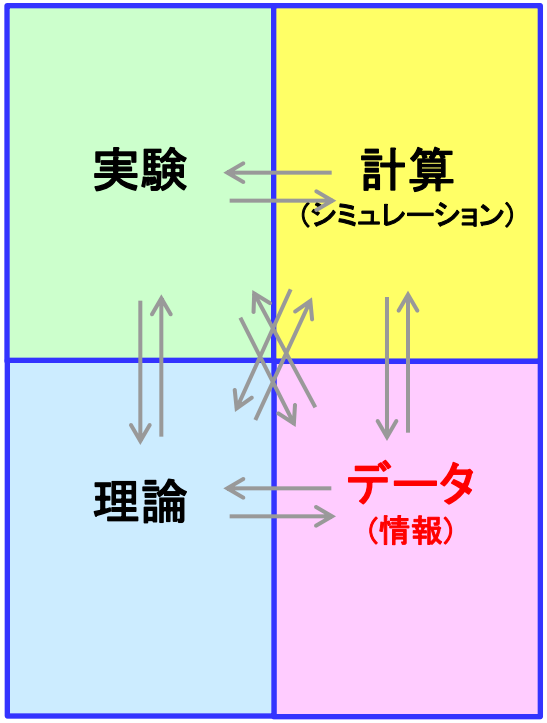
➡ 膨大な実験データ x 大規模シミュレーション ➡ 課題解決の有力なツール

データ科学

効率的な材料・デバイスの研究開発にむけて

■従来の手法に加えて、データ科学を融合し、材料・デバイス開発を加速する。





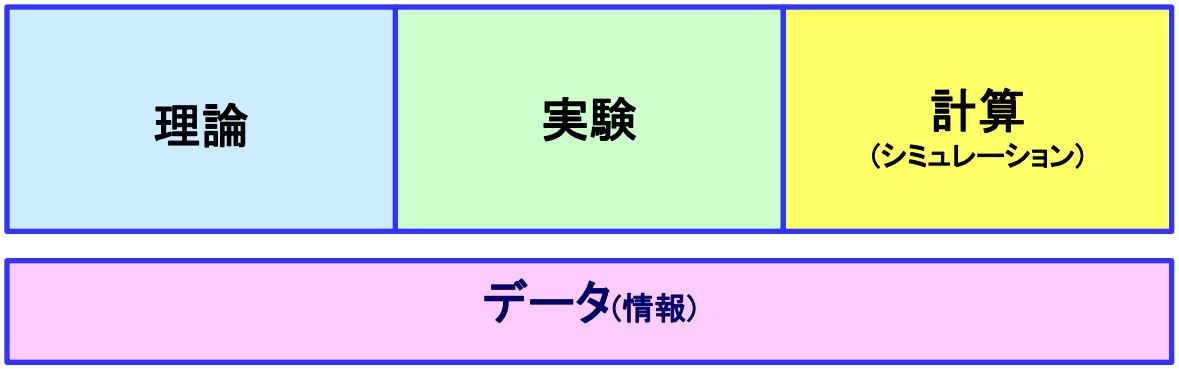
次世代の材料・デバイス研究開発インフラ

理論・実験・計算・データ科学の連携・融合した共通基盤プラットフォームの構築

次世代材料・デバイス開発 基盤プラットフォーム

■材料デバイス開発プラットフォームは、効率的なプロセス開発に必要な礎にもつながる。

次世代材料・デバイス開発 共通基盤プラットフォーム



・所望の新材料の発見、所望の機能出現、それにもとずく新デバイスの原理の発明
・システマティックな材料・デバイス設計の本質的道筋の構築

材料の本質の理解につながるので、
・ミクロ～メソ～マクロにわたるマルチスケールシミュレーション技術
・材料・デバイス製造プロセス技術
に対して、しっかりとした土台が構築しやすくなる。

- 材料開発でボトルネックとなっている要素はなにか？
  （コスト、時間、知識、技術、データ、人材）
- MIの現状をどうとらえているか？
- MIの現状を打破する要素はなにか？
  （コンソーシアム、ソフトウェア、データ蓄積、
  データマニング・機械学習・人工知能）

有機材料・無機材料のMI比較

質問内容：①　材料データは十分にあるか？　あるもの　不足しているもの
②　解析ツールは十分に用意されているか？　あるもの　不足しているもの
③　今後の課題

有機材料系

材料データ	解析ツール	今後の課題
ある：添加剤、界面活性剤など主に低分子材料に関して、典型的なものはハントブック等に集約されている。 ない：半導体、発光素子など先端的なもの、あるいはインフォマティクスを進めていく上で必要なネガティブデータなどは企業の中にあり、公開されていないものは多い。 また、高分子や複合材料など複数の素材からなる材料に関しては、構造や特性が組み合わせ、プロセスに大きく依存するので、入手できるデータは極めて限られている。	ある：低分子材料で量子化学計算が適用できるような系に関しては、ツールが利用されており、材料特性を直接アウトプットしないまでも、特性と相関づけられるある程度定量的なデータが得られている。 ない：高分子、複合材料では機能と大きく関連する、高次構造（結晶、相分離、分散etc.）を予測する時点ですでに計算化学の適用は限界があり、多くの場合定性的な結果を与えるに留まる。	有機材料は一部の添加剤などを除いて、分子単体で機能を発現する用途は少なく、集合体の高次構造が機能を左右する。まずは、それら高次構造を予測できる計算科学の技術、および計測技術の進歩が必要だと考えられる。 さらに、多くの有機材料は複合材料、混ぜ物として使われるので、構成素材、組成、プロセスなど材料データとして非常に複雑なものになり、構造、特性データを含めた複雑なデータを整理するデータベースの仕様を作っていくというのも今後の課題になるのではないかと思われる。

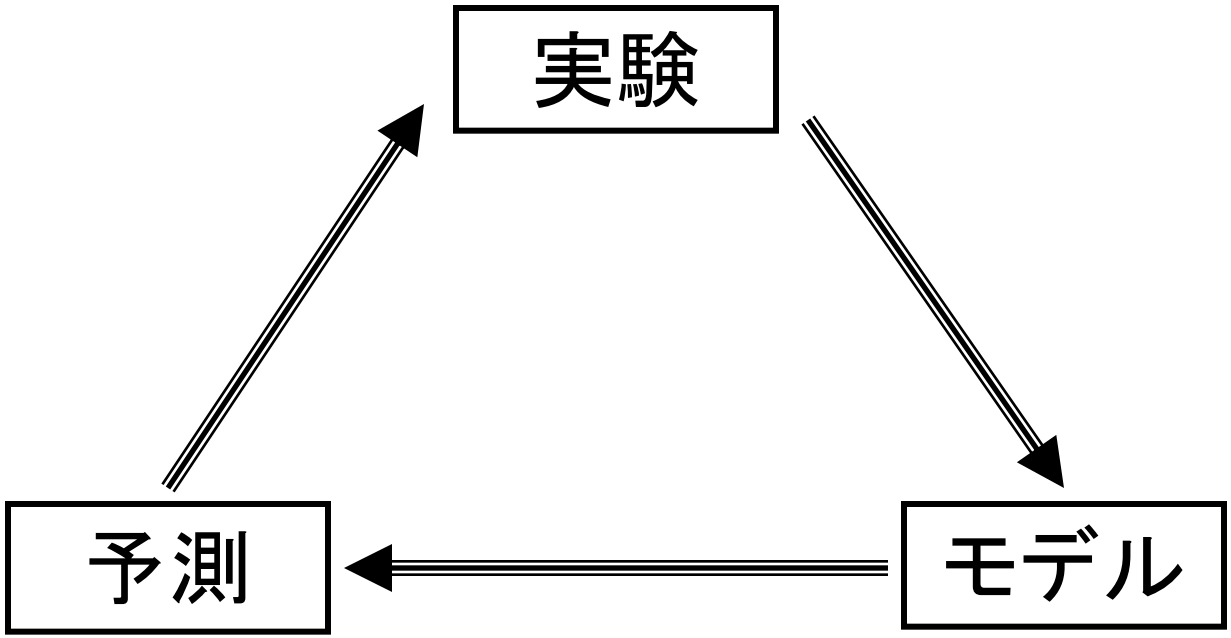
有機材料・無機材料のMI比較

質問内容：①　材料データは十分にあるか？　あるもの　不足しているもの
②　解析ツールは十分に用意されているか？　あるもの　不足しているもの
③　今後の課題

無機材料系

材料データ	解析ツール	今後の課題
ある：超伝導材料、金属化合物の相図、酸化物の一部などは、金属便覧や各種相図などはデータブックとして公開されている。 　また、NIMSのMatnaviなどで材料データが公開されている。アメリカではNISTに材料データが蓄積されている。 ない：半導体、耐熱材料など先端的なもの、組成を連続的に変化させたときの状態などのデータは不足。 概して、バイオ、有機材料データと比較して実測値に基づくデータは不足している。	ある：X線回折、表面分析、光学計測など多くの最先端計測機器が開発されている。 ない：高速に計測できる装置が少ない。計測では大量にデータが発生するが、それを解析するツールは現在、開発中。また、組成やプロセス条件を系統的にあつめたデータはない。	バイオ、有機などはかなりの実測データを保有しているが、無機材料系のデータは少ない。また、材料は多面的で一つの材料が複数の特性を示すために単純ではない（TiO2など） また、一般に材料合成に時間がかかるために、データを蓄積に時間がかかっている。

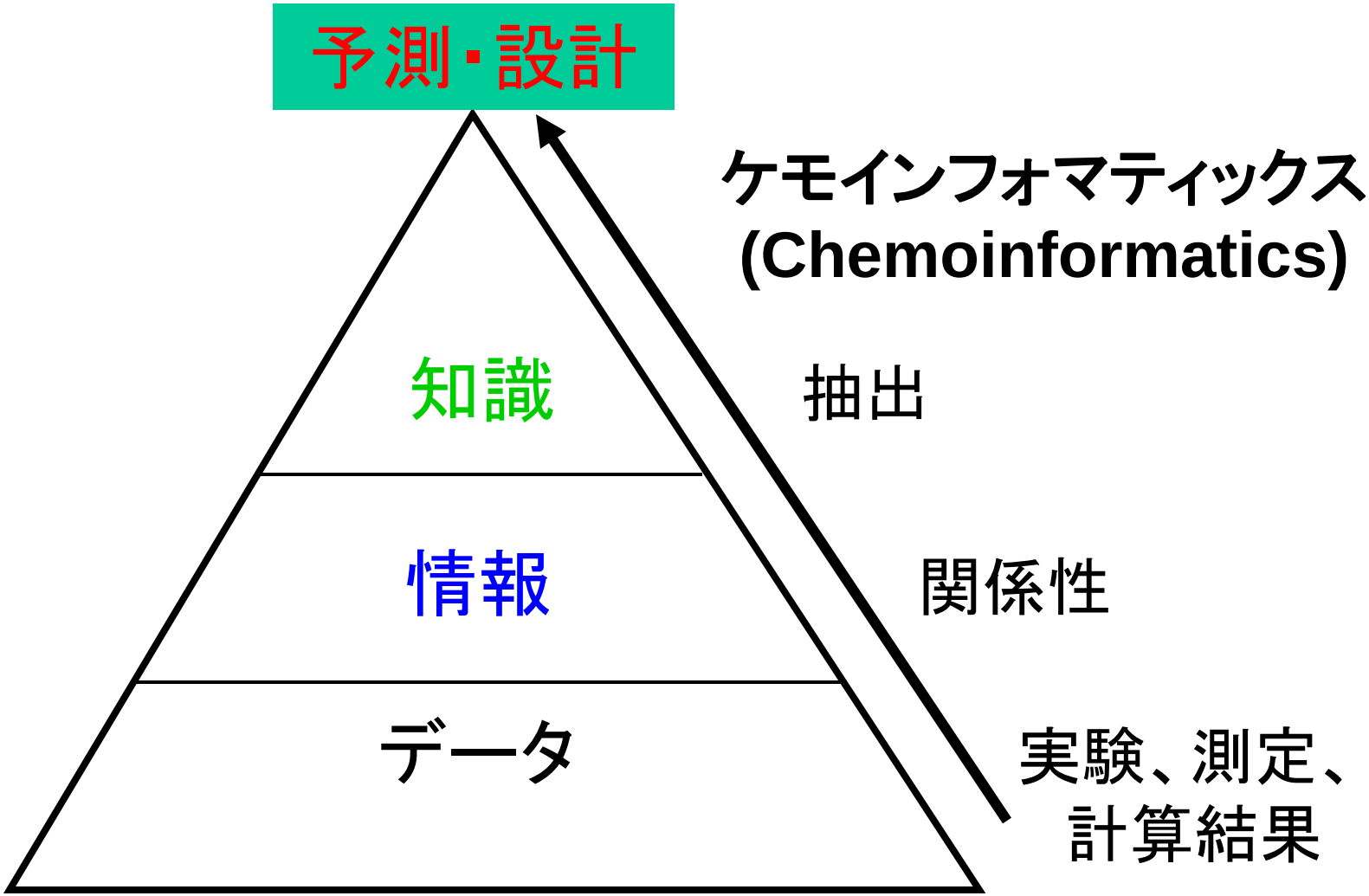
帰納的学習



・数式モデル

・データ・情報の相関関係

・直感モデル



データ、情報、知識の階層図

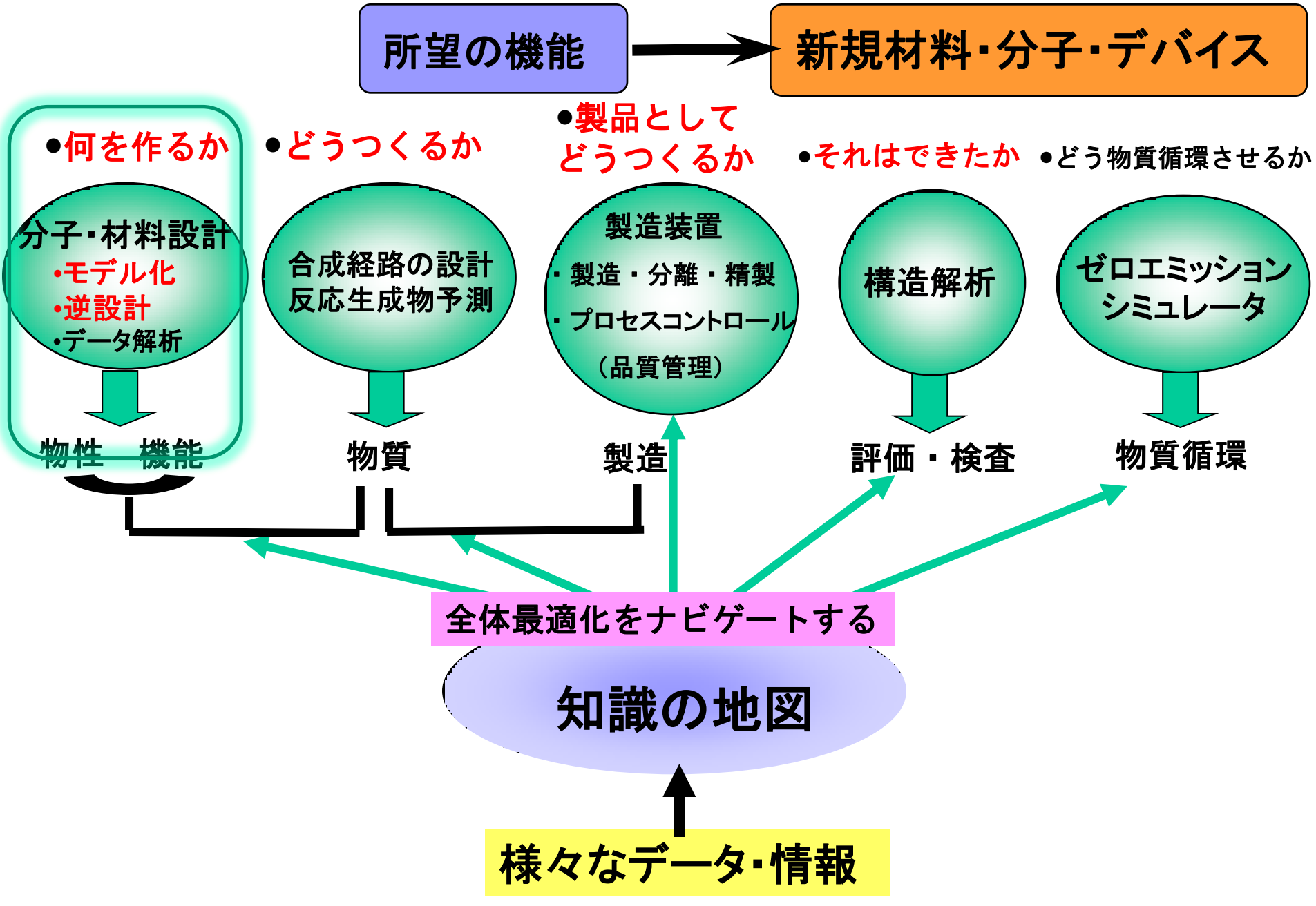
ケモインフォマティックスが目指すこと

データ・情報から知識へ　知識から設計へ

単なる解析や説明だけではなく、具体的な予測や設計つながる新たな知識をコンピュータの特徴（ex.網羅性）を利用して作り出すことであり、まさに化学の姿勢そのもの

- 新しい規則の発見、それによる新規分子構造、材料候補などの創出への期待
- 先入観による見落としのカバー

データ駆動型化学の守備範囲



データ・情報の知識化と予測・設計への活用

・データ・モデリングによる知識化とその利用

・構造―物性（活性）相関モデルの構築

・モデルの逆解析による目的物性を満足する材料・分子構造候補（プロセス条件を含む）の創出

・構造発生をドライビングフォースとし、モデルを制約条件とする候補構造の創出

薬物分子設計

高機能樹脂設計(モノマー設計を含む)

触媒設計

・量子化学計算

与えた構造、材料の評価はできても、逆に具体的な構造、材料候補の提案は基本的に不可能。

順方向予測や解析は可能。逆方向予測は不可能。

まさにこの逆解析がデータ集約型化学に求められる重要項目の一つであると理解され始めてきた。

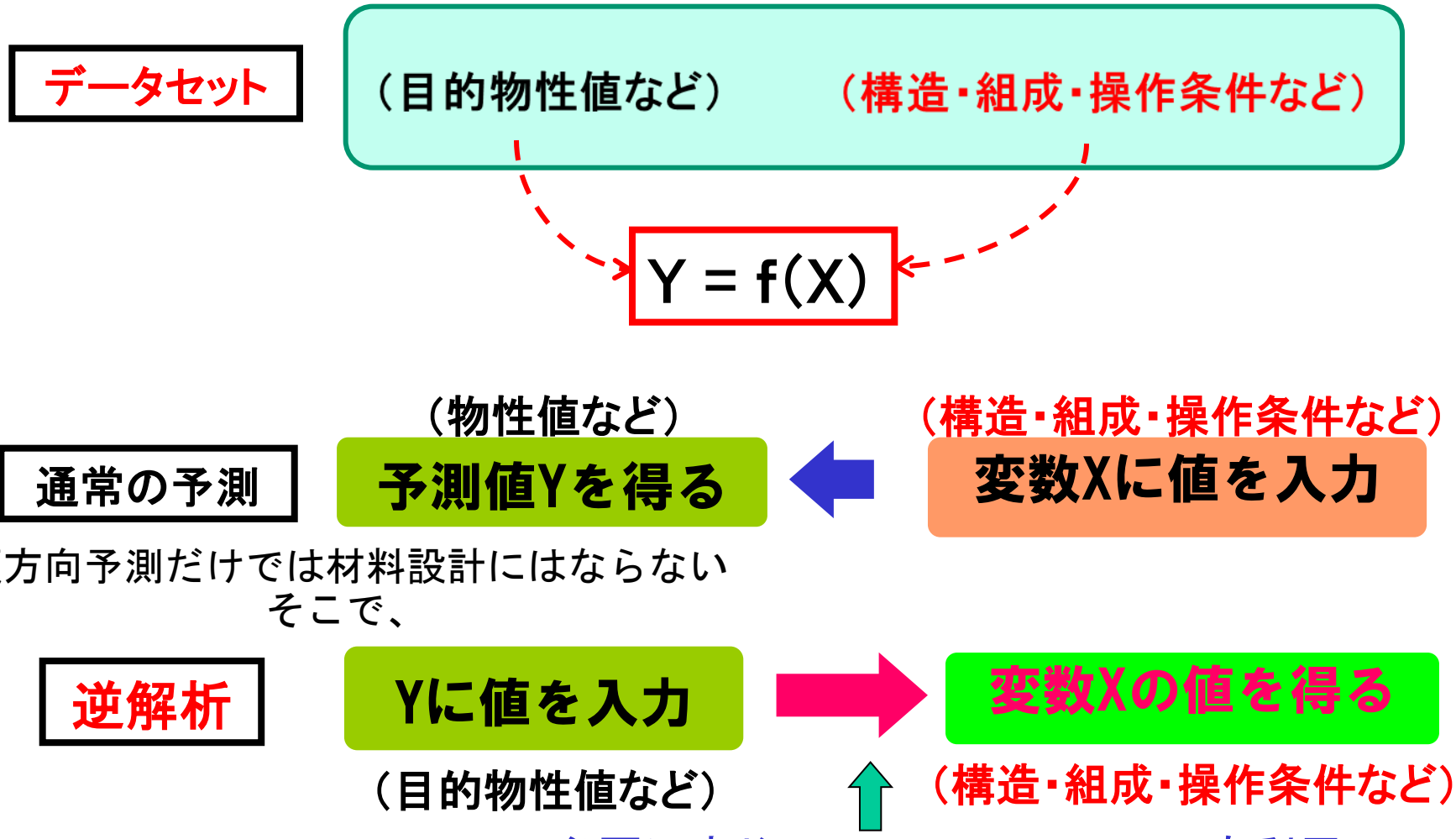
・データ駆動型化学＆量子化学計算

役割が異なるために車の両輪としての利用が必要

つまり設計と評価

これこそが目指すマテリアルズ・インフォマティクスの姿

実験データを用いた予測モデル作成



•T.Miyao, M. Arakawa, K. Funatsu, Exhaustive Structure Generation for Inverse-QSPR/QSAR, *Molecular Informatics*, 29, 111-125, (2010).

•T. Miyao, H. Kaneko, K. Funatsu, Inverse QSPR/QSAR Analysis for Chemical Structure Generation (from y to x), *Journal of Chemical Information and Modeling*, 56(2), 286-299, 2016.

データ・情報の知識化と予測・設計への活用

・データ・モデリングによる知識化とその利用

・構造―物性（活性）相関モデルの構築

・モデルの逆解析による目的物性を満足する材料・分子構造候補（プロセス条件を含む）の創出

・構造発生をドライビングフォースとし、モデルを制約条件とする候補構造の創出

薬物分子設計

高機能樹脂設計(モノマー設計を含む)

触媒設計

薬物設計・材料設計など

１）薬物設計

- Inverse QSPR/QSARを利用した薬物分子構造設計
- 薬物の代謝酵素予測および代謝部位予測に関する研究
- 活性を考慮した化学空間およびタンパク質空間の可視化
- 薬物分子構造重ね合わせによるPharmacophoreモデルの自動創出と新規リード化合物の探索手法の開発
- 可視化による化学空間探索型構造生成

２）材料設計

- ポリマーアロイ、エンジニアリングプラスティック、共重合ポリマーなどの組成設計、操作条件の最適化
- 共重合に用いる新規モノマー設計

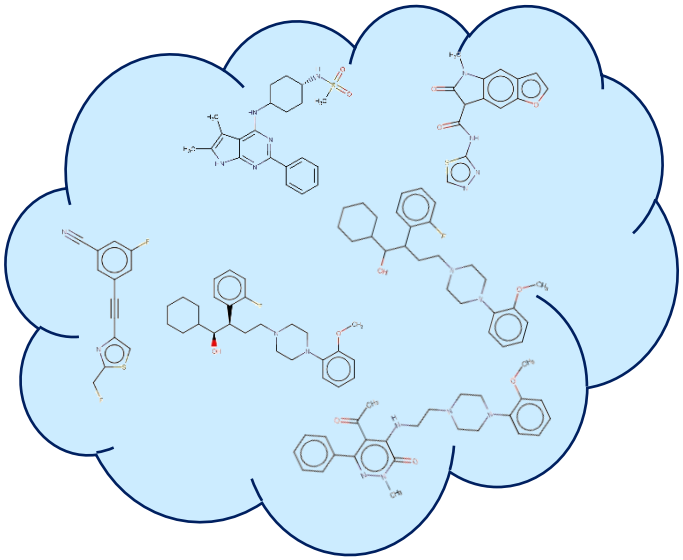
３）触媒設計

- 固体触媒設計
- 自動車用排ガス触媒の構造設計

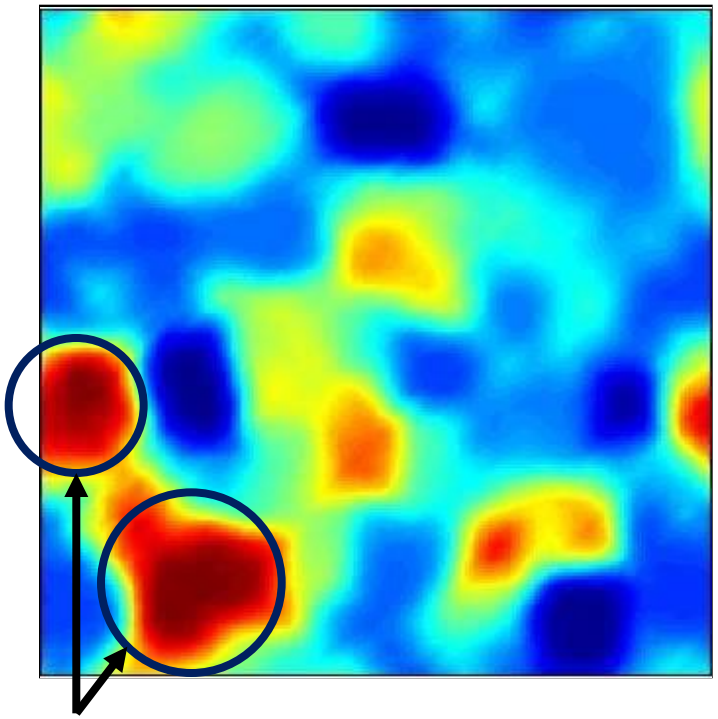
可視化による化学構造探索

- 化学空間の可視化
  - 化学空間内での活性分布を把握
  - 活性分布を元に構造探索のターゲットを決定　⇒　探索の効率化

化学空間
（化学構造の記述子空間）



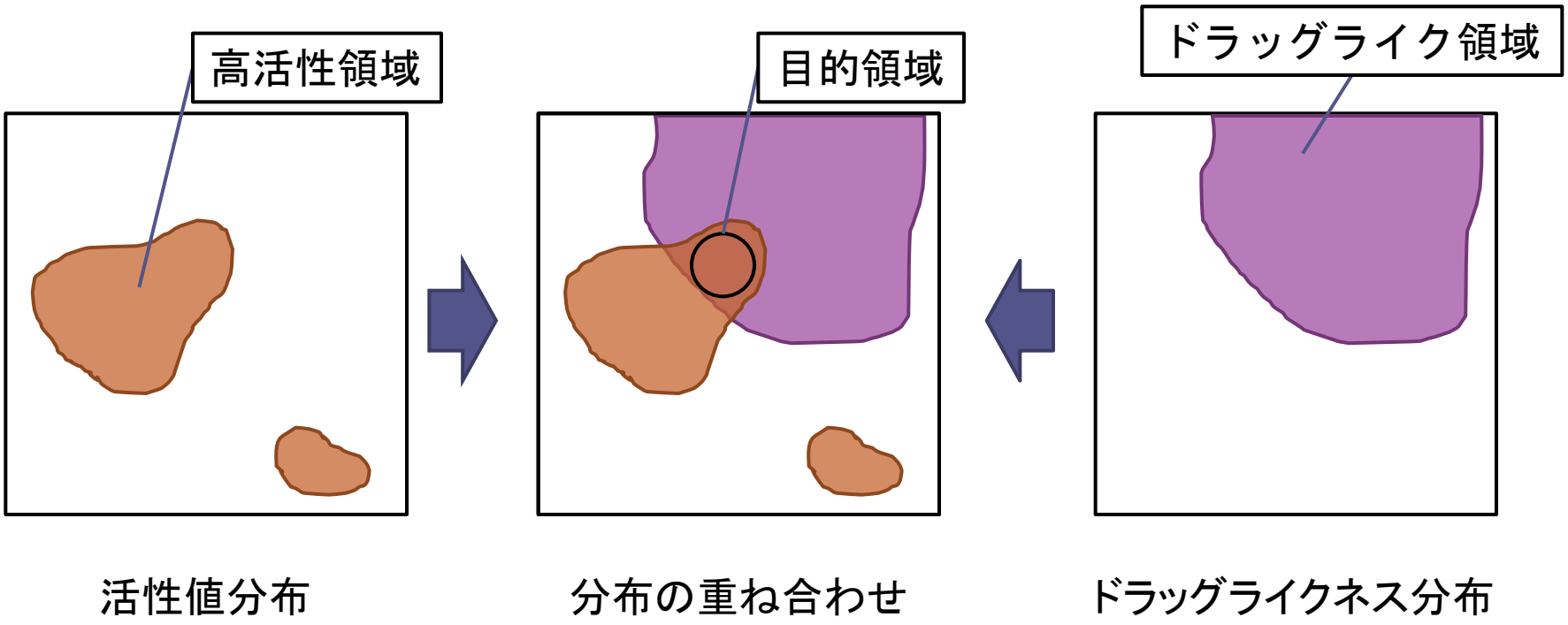
可視化



ドラッグライクネス分布

同じマップ上で活性値とドラッグライクネスを
それぞれ予測

両方の分布を考慮して領域を選択



ケーススタディ

目標タンパク質：ヒスタミンH1受容体
元構造データ：ChEMBL[1]より取得した522構造と活性値pKi
記述子：PubChem[2]のフィンガープリントより142個

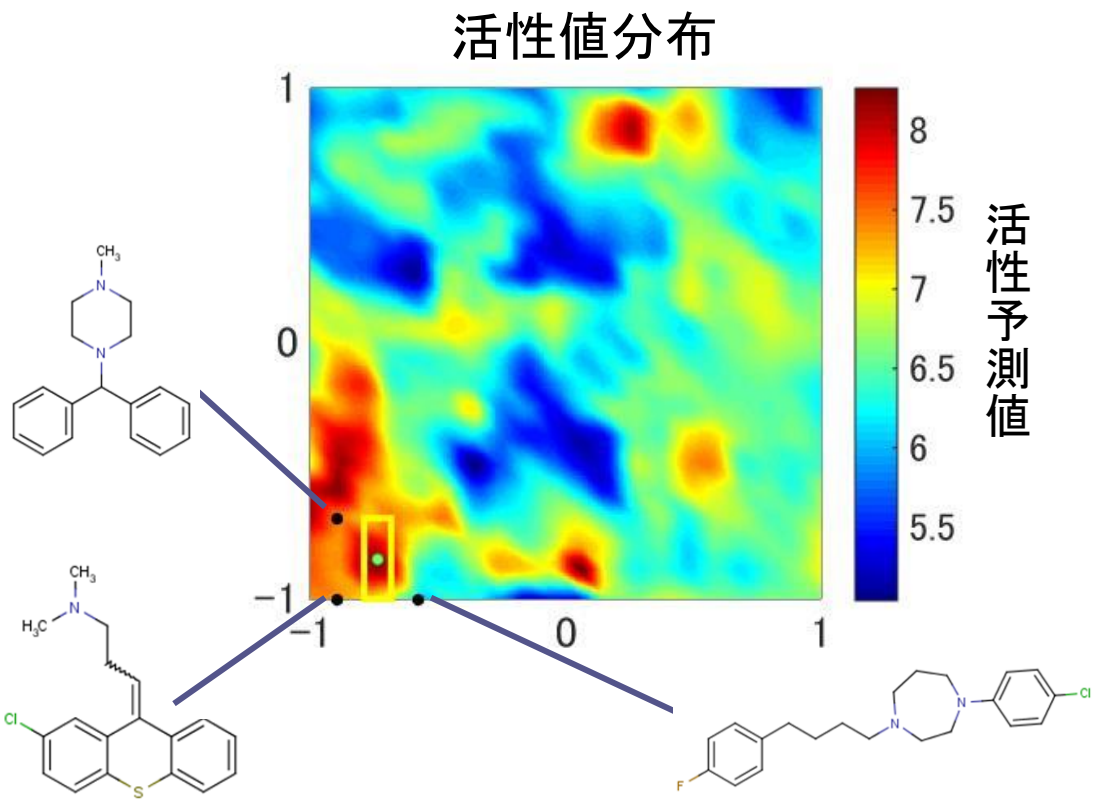
ドラッグライクネス判定モデル構築用データ:
CMC[3]より1000構造　ACD[3]より1000構造　計2000構造
判別モデルはSVM(Support Vector Machine)で構築

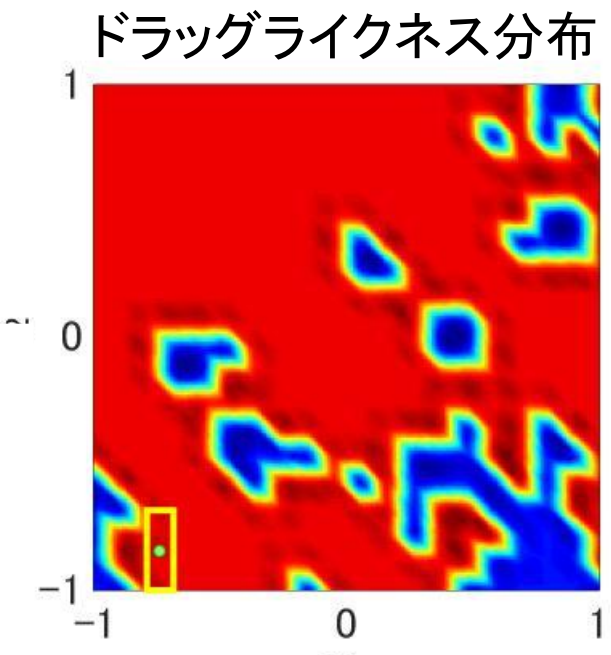
[1] https://www.ebi.ac.uk/chembl/
[2] http://pubchem.ncbi.nlm.nih.gov/
[3] http://accelrys.co.jp/products/databases/

分布を考慮した構造生成

ドラッグライクネスが低い・高い領域で構造を生成

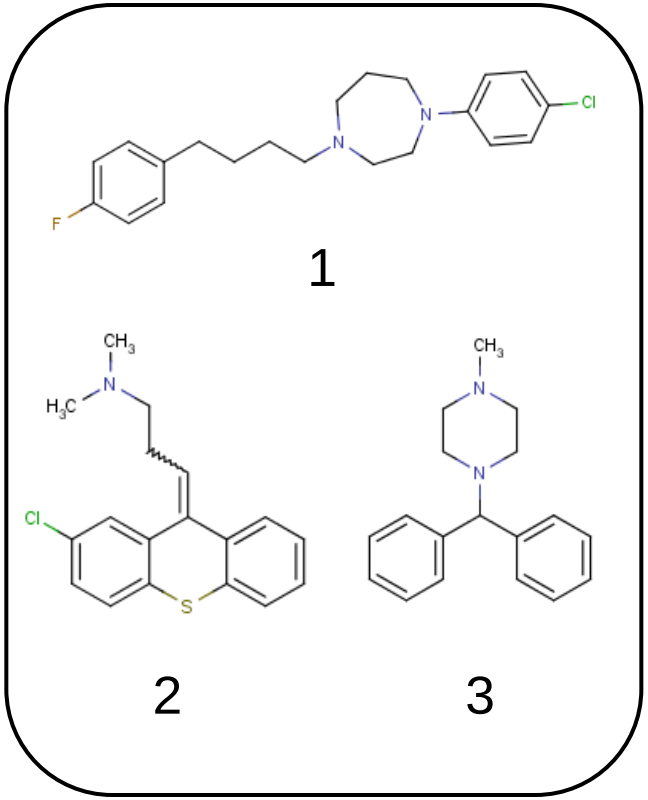
② 高い領域



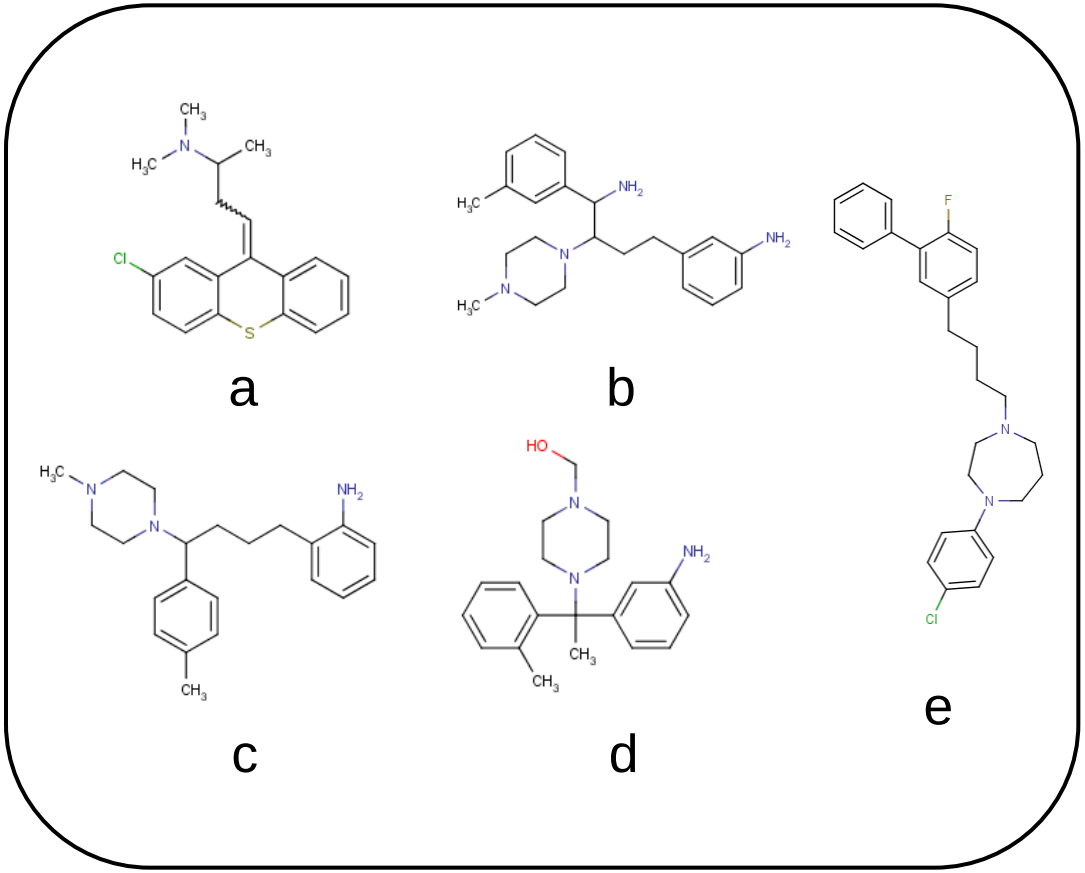


赤：ドラッグライク
青：非ドラッグライク

生成構造



初期シード構造

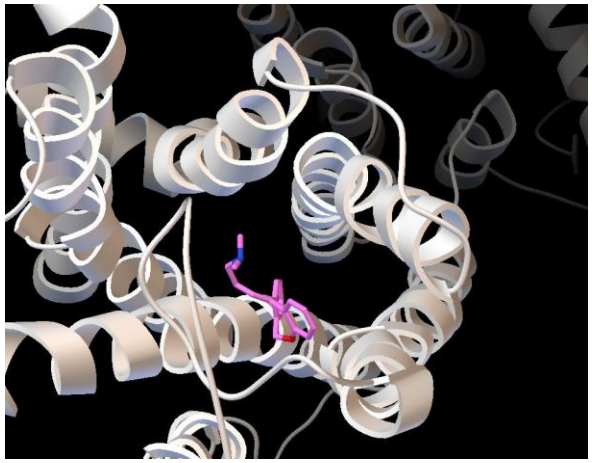


部分構造あり
生成構造の一部

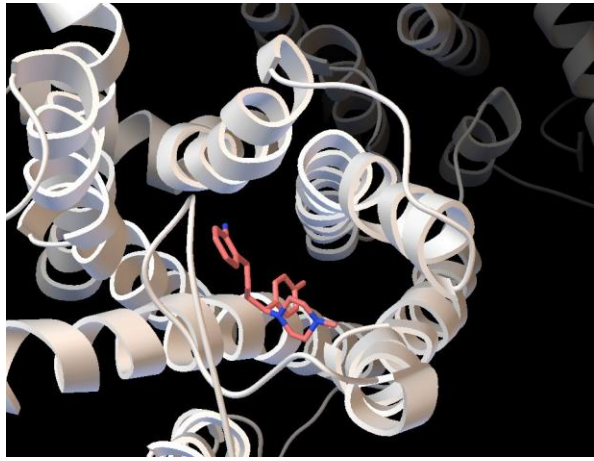
ドッキングシミュレーション

AutoDock Vina[8]を用いてドッキングシミュレーションを実行
ターゲット：HistamineH1 receptor (PDBID : 3RZE)

	結合エネルギー kcal/mol
1	-8.0
2	-7.3
3	-8.4
a	-8.7
b	-8.2
c	-9.5
d	-8.4
e	-9.2



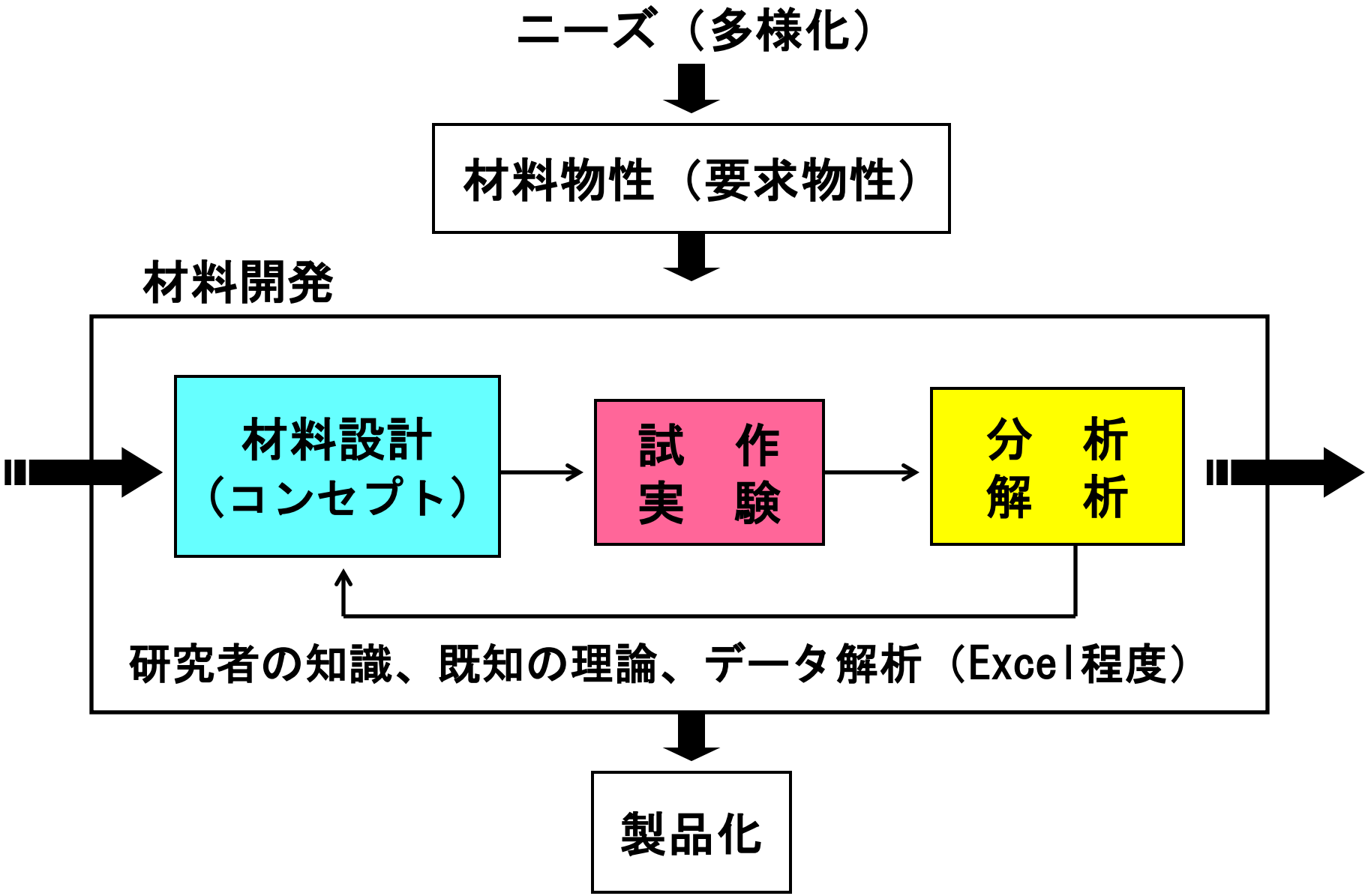
参考構造　Doxepine
結合エネルギー：-9.4 kcal/mol



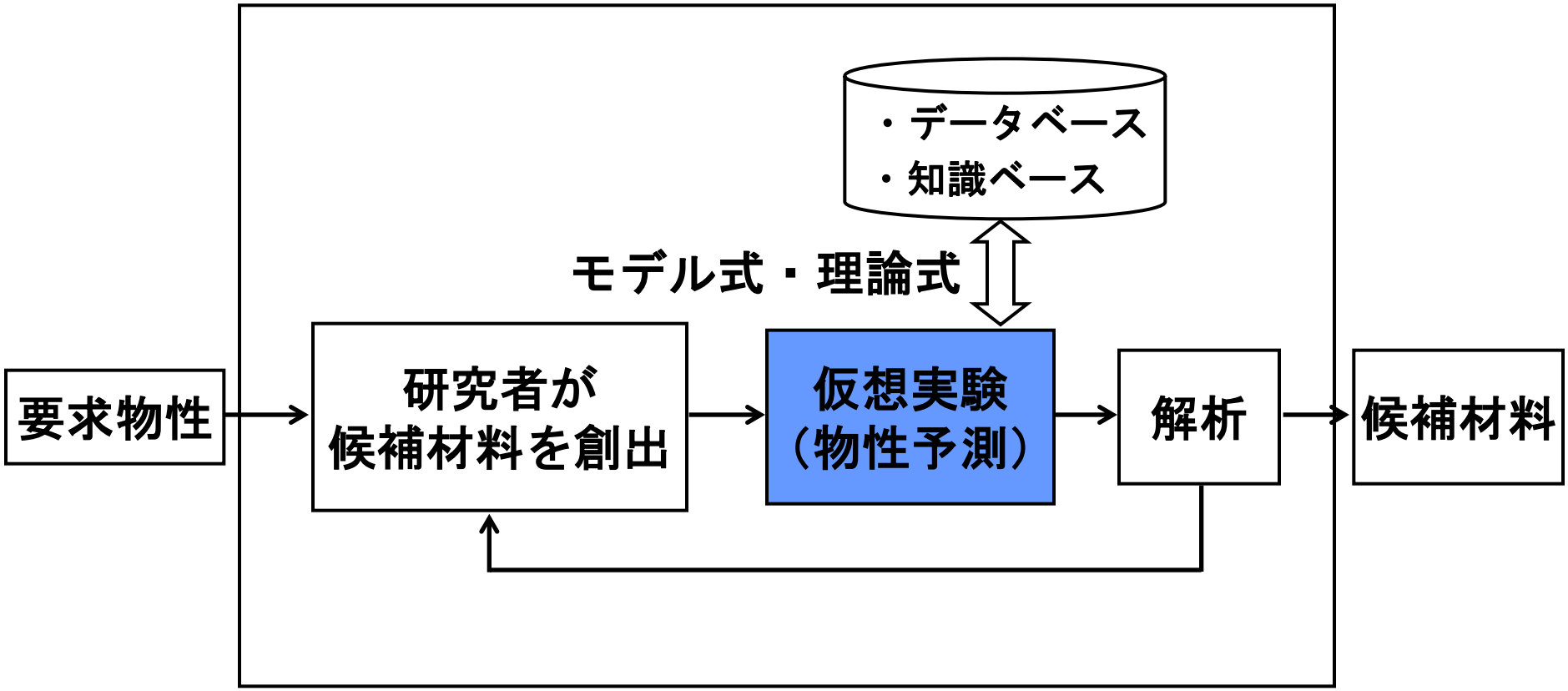
c

[1] O. Trotto, A. J. Olson, J. Comput. Chem., 31.2, 455-461, 2010.

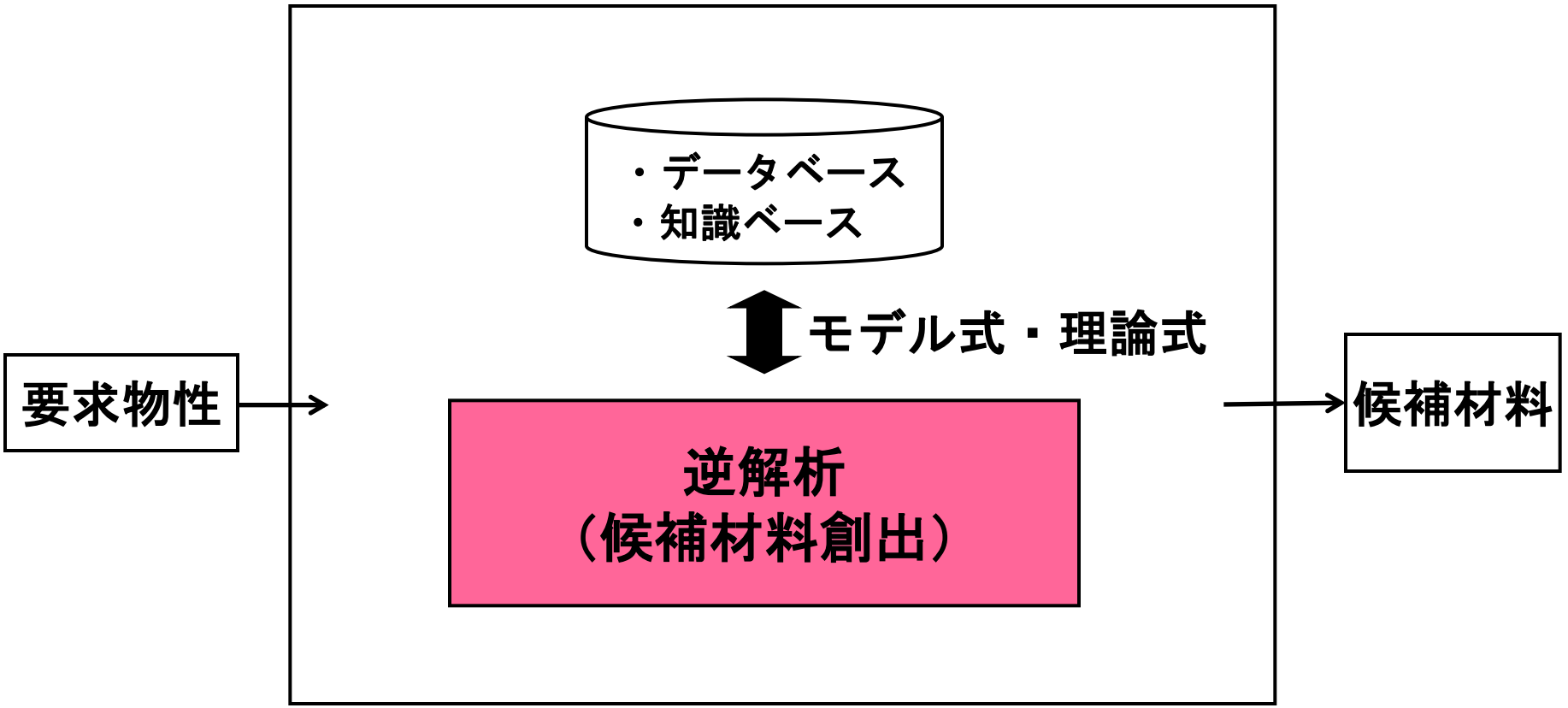
材料開発の流れ


ニーズ（多様化）
材料物性（要求物性）
材料開発
材料設計
（コンセプト）
試　作
実　験
分　析
解　析
研究者の知識、既知の理論、データ解析（Excel程度）
製品化

材料設計（コンセプト）

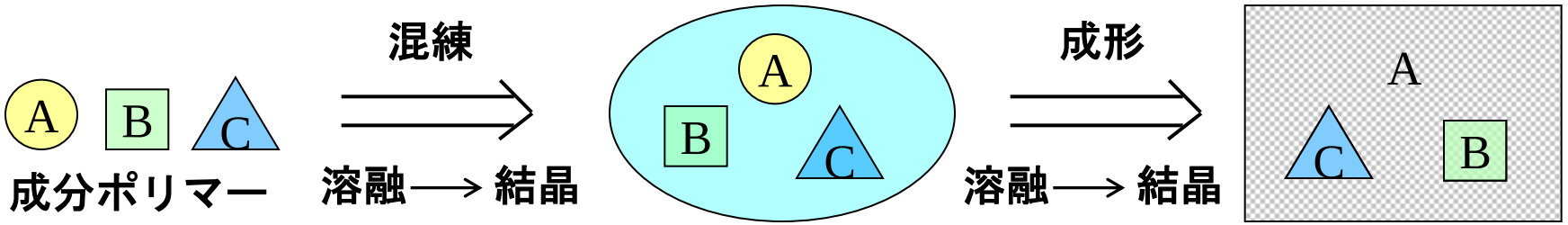


逆解析



事例

ポリマーアロイ



データ項目

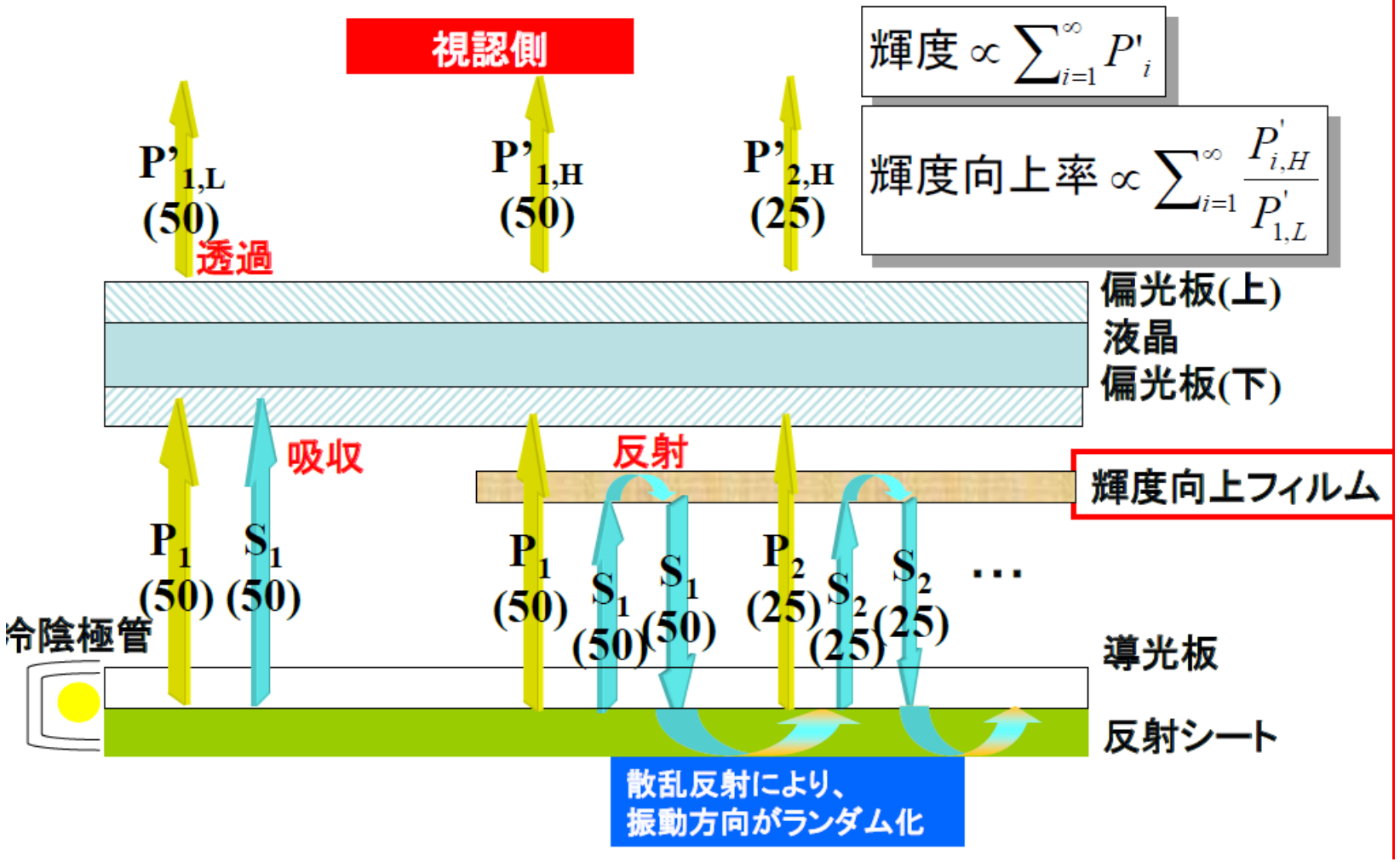
- 各成分ポリマーの組成・物性値
- 混練条件
- 成形条件

（プロセス条件）

- アロイ物性値（目的物性）

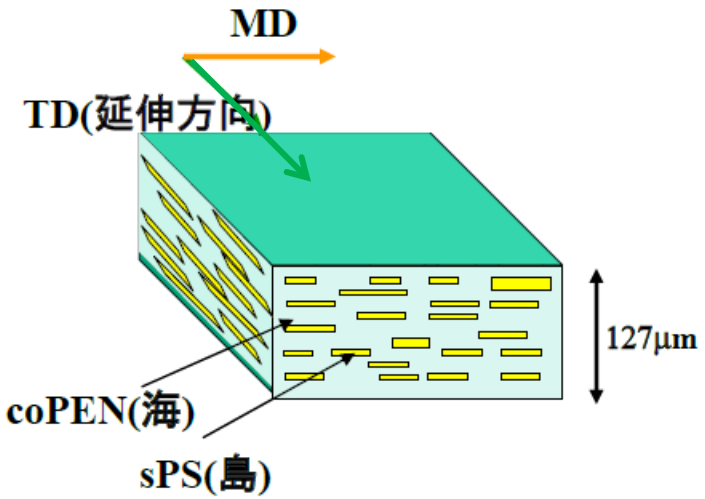
100～200種

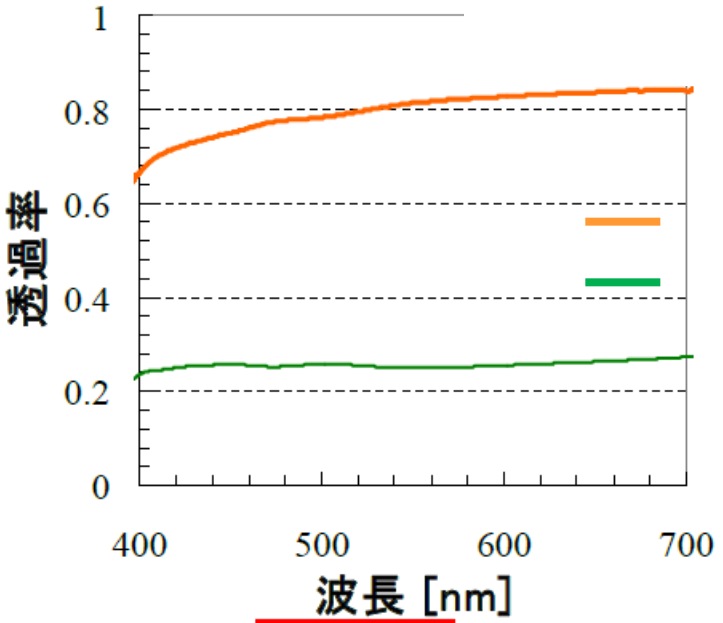
輝度向上フィルムの輝度向上原理

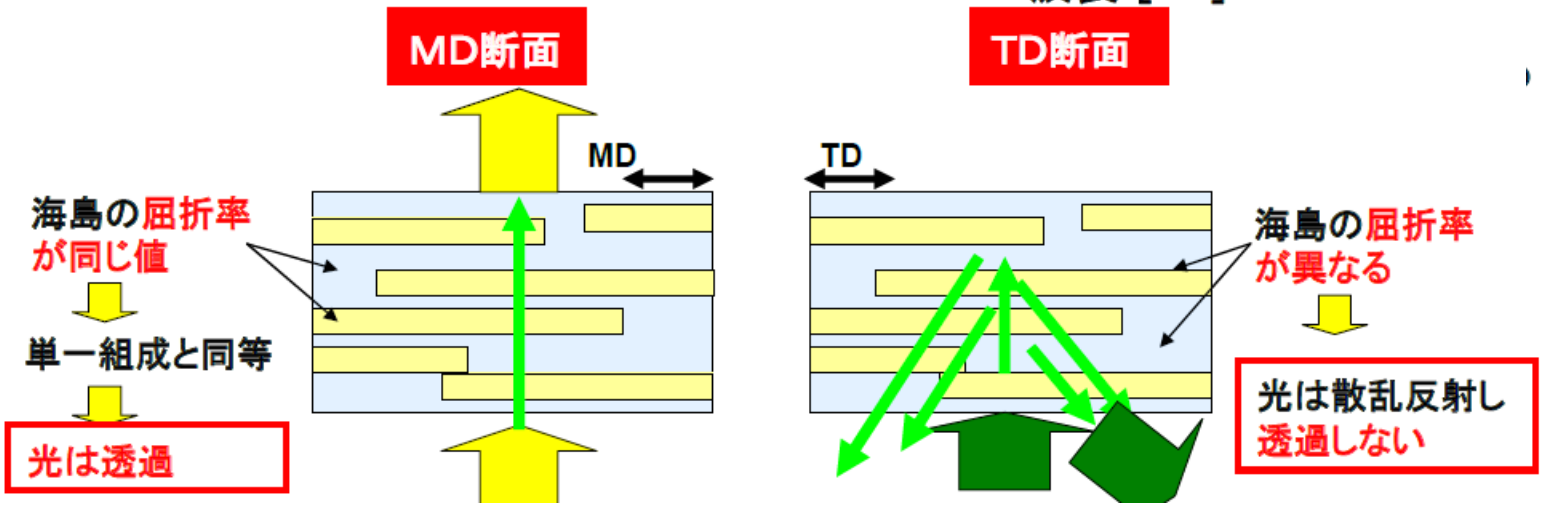


従来吸収されていた偏向(Si)のみを反射して再利用ー＞高輝度化

偏光透過/反射のメカニズム







材料組成だけではなくプロセス条件により特性が変化

- 開発目標

  - 輝度向上フィルムの特性に関する定量モデルを構築し、より性能の高いフィルムの設計を行う
  - その際に、押出条件などの性能に影響するパラメータ(プロセス条件)の最適化を目指したい

目標物性

- 輝度5400以上、MD透過率82%以上、TD透過率20%以下

・解析　１

– 目的変数(Y)：輝度[cd/m2]
– 説明変数(X)：
  組成（PEN [%], PET [%] , sPS [%]）（相溶化剤1, 2, 3 [%]）
  延伸温度[℃], 押出機機種（1 or 2）, 延伸倍率, 厚み[um]

・解析　２

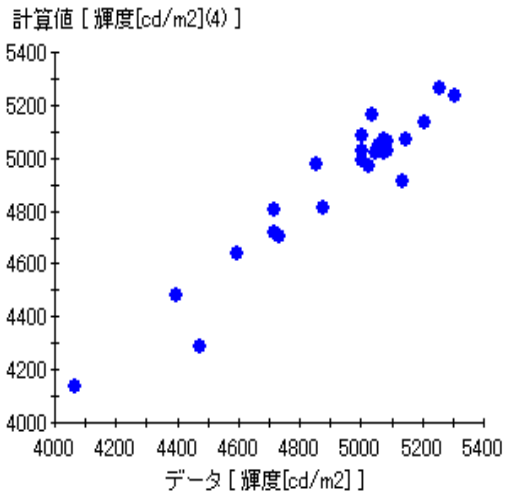
– 目的変数(Y)：MD透過率[%] or TD透過率[%]
– 説明変数(X)：解析１と同じ

サンプル数：２６

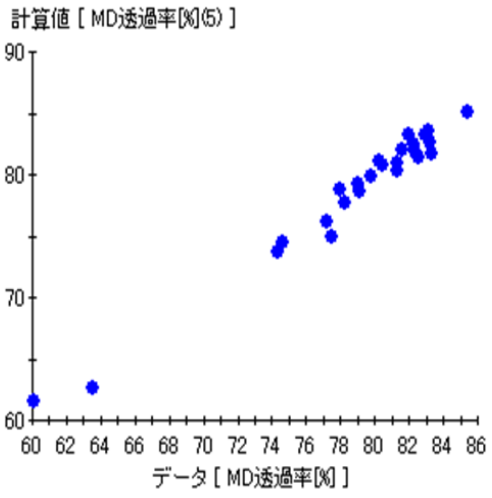
説明変数の数：１０

PLS（線形モデル化手法）による解析結果

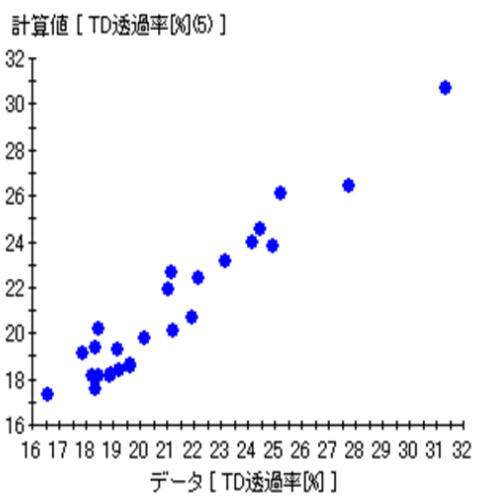
- 目的変数：輝度
- 結果
  - $R^2 = 0.916$
  - $Q^2 = 0.682$



- 目的変数：MD透過率
- 結果
  - $R^2 = 0.977$
  - $Q^2 = 0.920$



- 目的変数：TD透過率
- 結果
  - $R^2 = 0.930$
  - $Q^2 = 0.746$



・上の3つのモデルの同時に逆解析

・トレードオフを考慮した多目的最適化
（パレート最適解の探索）

多目的最適化

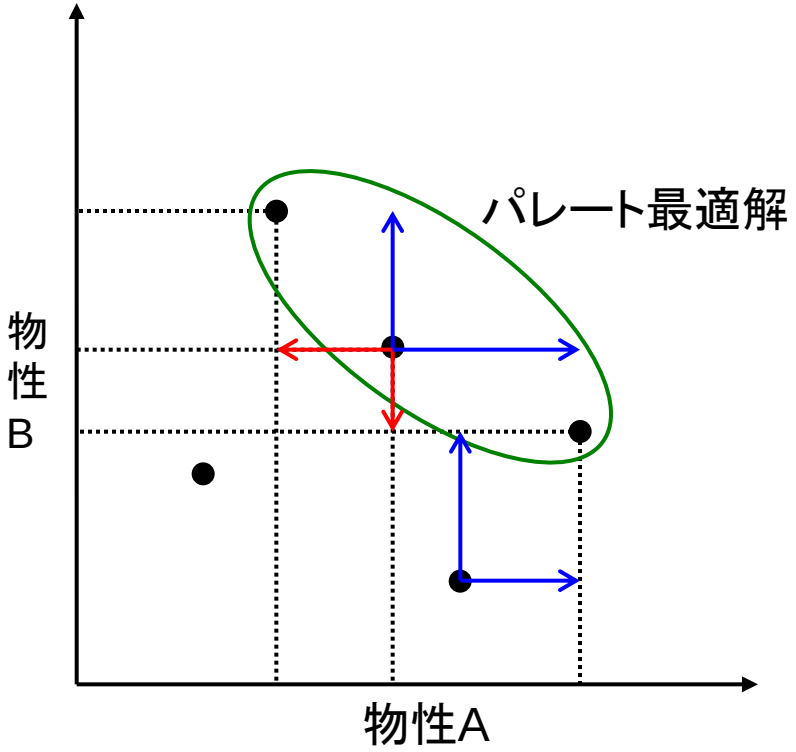
一般に物性間にはトレードオフの関係が存在する場合が多い

目的とする物性を同時に最適化することはできない

パレート最適の概念を用いた多目的最適化を行なう

✧パレート最適とは

他の目的変数の値を減らすことなしには、どの目的変数の値も増加させることができないような状態



応用事例

- 共重合ポリエステル樹脂の組成設計および新規モノマー設計（成功事例あり）
- 自動車用各種プラスティック部品の組成設計および操作条件最適化

  （実際に成分ポリマー・カタログとアクセスをし、コスト面でも良好なポリマーアロイを短時間で設計する仕組みを開発し日常的に利用）
- 導電性有機材料開発（成功事例あり）
- プロピレンオキシド製造用触媒

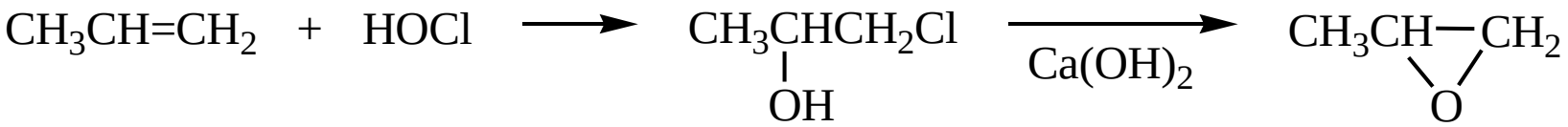
など

・棚田東作、荒川正幹、西村竜一、船津公人、ケモメトリックス手法を用いた材料開発支援プログラムの開発、J. Comput.-Aided Chem., 1, 35-46, (2000).
・後藤俊、荒川正幹、船津公人、ポリマー設計のための物性推算法と逆解析手法の開発、J. Comput.-Aided Chem., 10, 30-37 (2009).

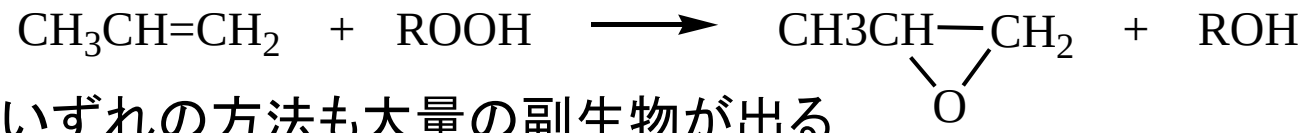
プロピレンオキシド製造用触媒

プロピレンオキシドの製造法

- クロロヒドリン法



- 有機過酸化物法



いずれの方法も大量の副生物が出る

→ 副生物の出ない酸素酸化のための触媒開発が必要
Agを主成分とする多成分系酸化物触媒の研究が主流

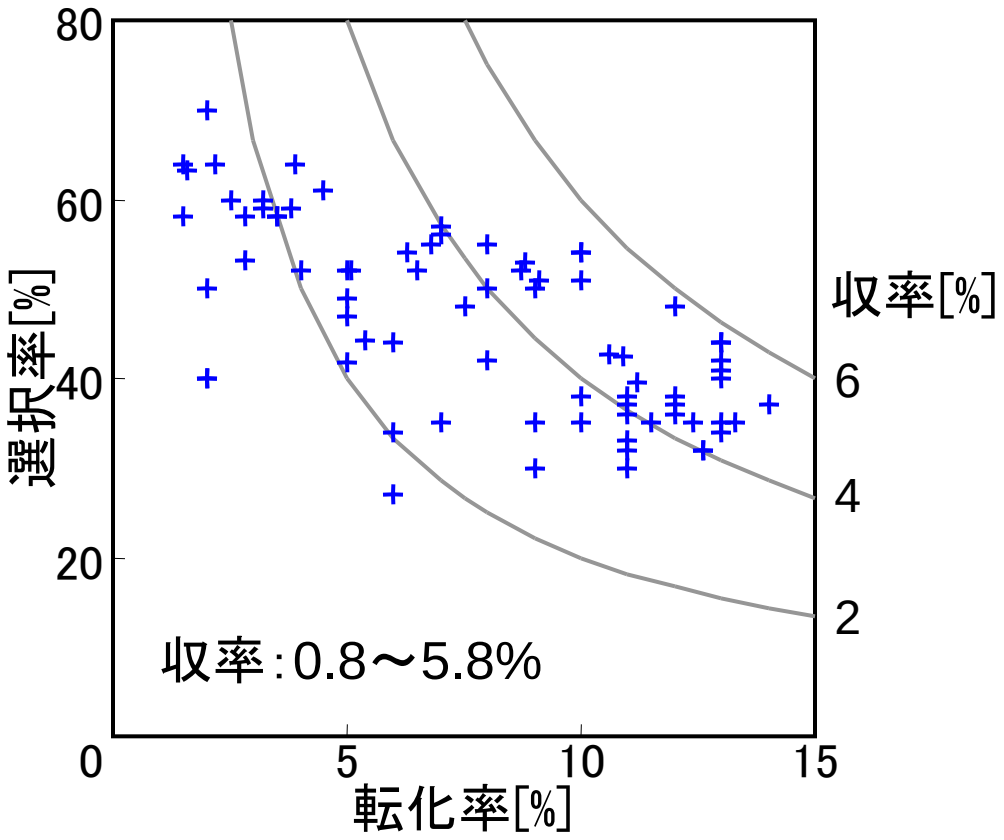
- 回帰分析を行い、回帰モデルの予測性を検証する
- 触媒候補を探索して、組成の検討を行なう

データ収集

特許19件

- 反応物：プロピレン
- 生成物：プロピレンオキシド
- 酸化剤：酸素
- 気固触媒反応
- 反応温度：200℃以上
- 担体：$CaCO_3$

データ数：75



説明変数と目的変数

説明変数X

触媒組成

Ag [wt%]　**K** [wt%]　**Mo** [wt%]
W [wt%]　**Re** [wt%]　**Au** [wt%]
Cl [wt%]　**Mg** [wt%]　**Fe** [wt%]

反応条件

温度 [℃]　**圧力** [psig]
空間速度 [hr^{-1}]
原料気体組成
O_2 [mol%]　**プロピレン** [mol%]
NO_X [ppm]　**EtCl** [ppm]
CO_2 [mol%]

触媒前処理

有り 1、無し 0

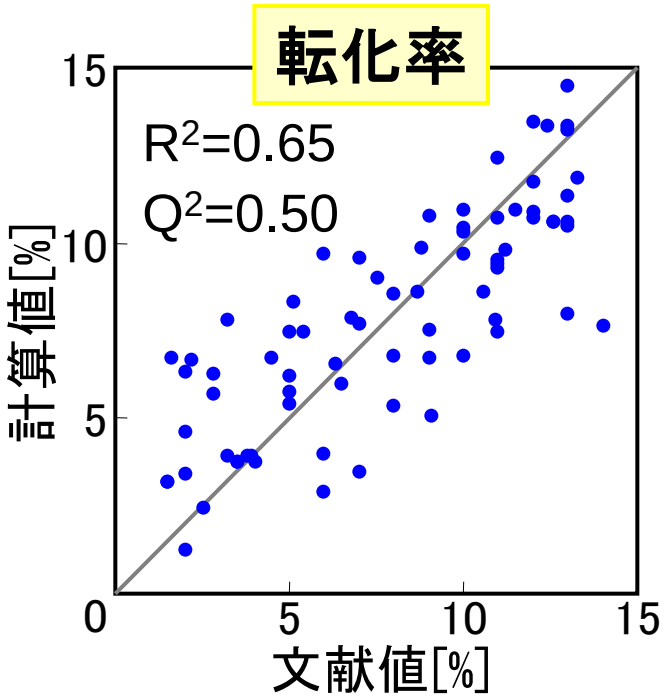
NO_X　**EtCl**　**CO_2**

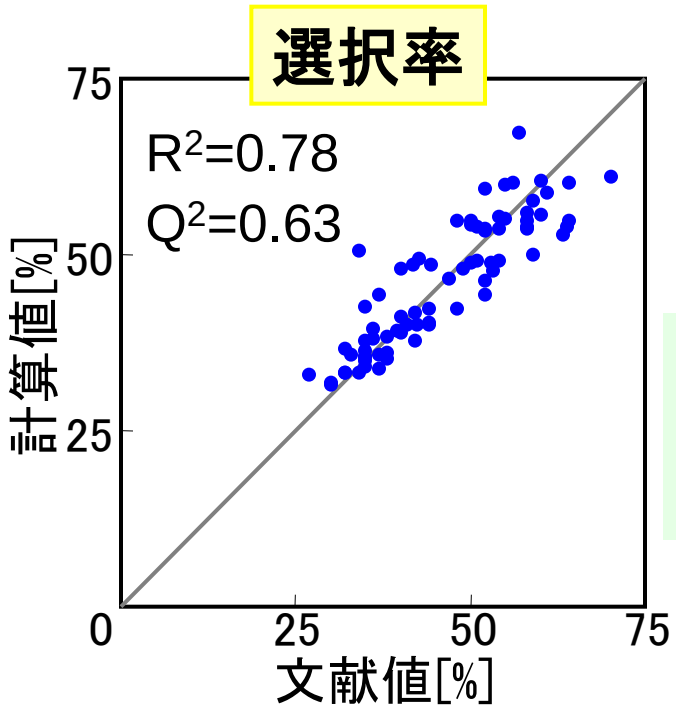
目的変数Y

⟷ **転化率** [%]

⟷ **選択率** [%]

PLSの結果





R^2：モデルの精度
Q^2：モデルの予測性
1に近いほどよい

PLSで線形回帰を行なったがモデルの予測性が不十分

データに非線形性があるため

説明変数の2乗項、交差項を追加 → 変数230個 → **変数選択**

GAPLSによる変数選択

- ✓説明変数の数が多いほどモデルの精度は高くなる
- ✓意味のない変数が含まれていたり、互いに相関の高い変数があるような場合にはモデルの予測性が低くなる

目的変数の予測に必要な説明変数のみを選び出す

↓

モデルの予測性が向上する

GAPLS法

K. Hasegawa et al. , J. Chem. Inf. Comput. Sci. 37 (1997) 306-310.

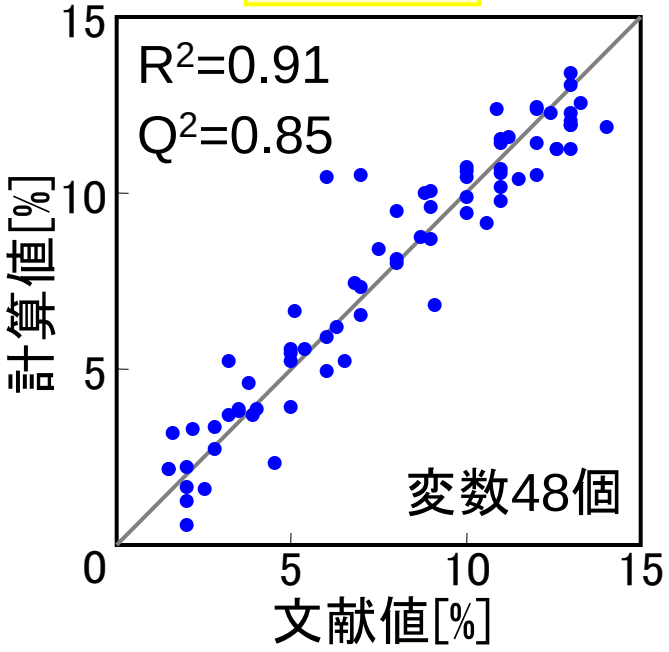
遺伝的アルゴリズムを用いてPLSのQ^2値が最大になるような
変数の組み合わせを見つけ出す手法

遺伝的アルゴリズム(GA:Genetic Algorithm)

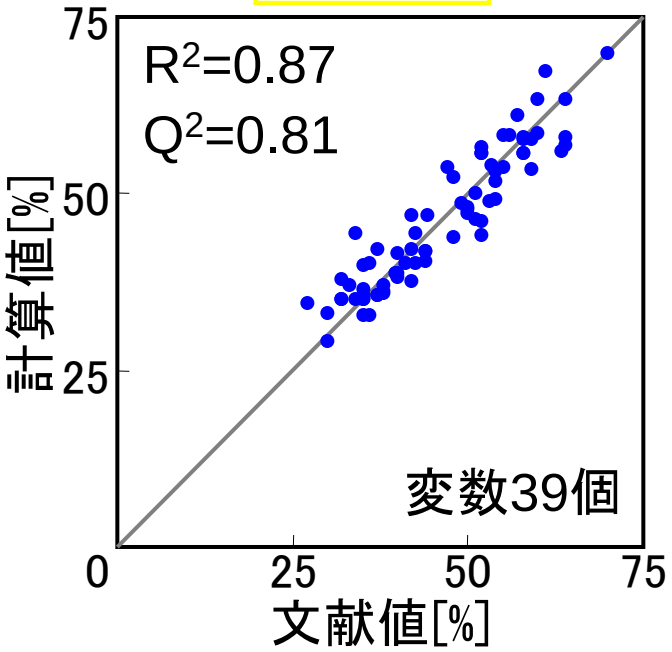
生物の遺伝の仕組みを模倣した汎用の最適化アルゴリズム

GAPLSの結果

転化率



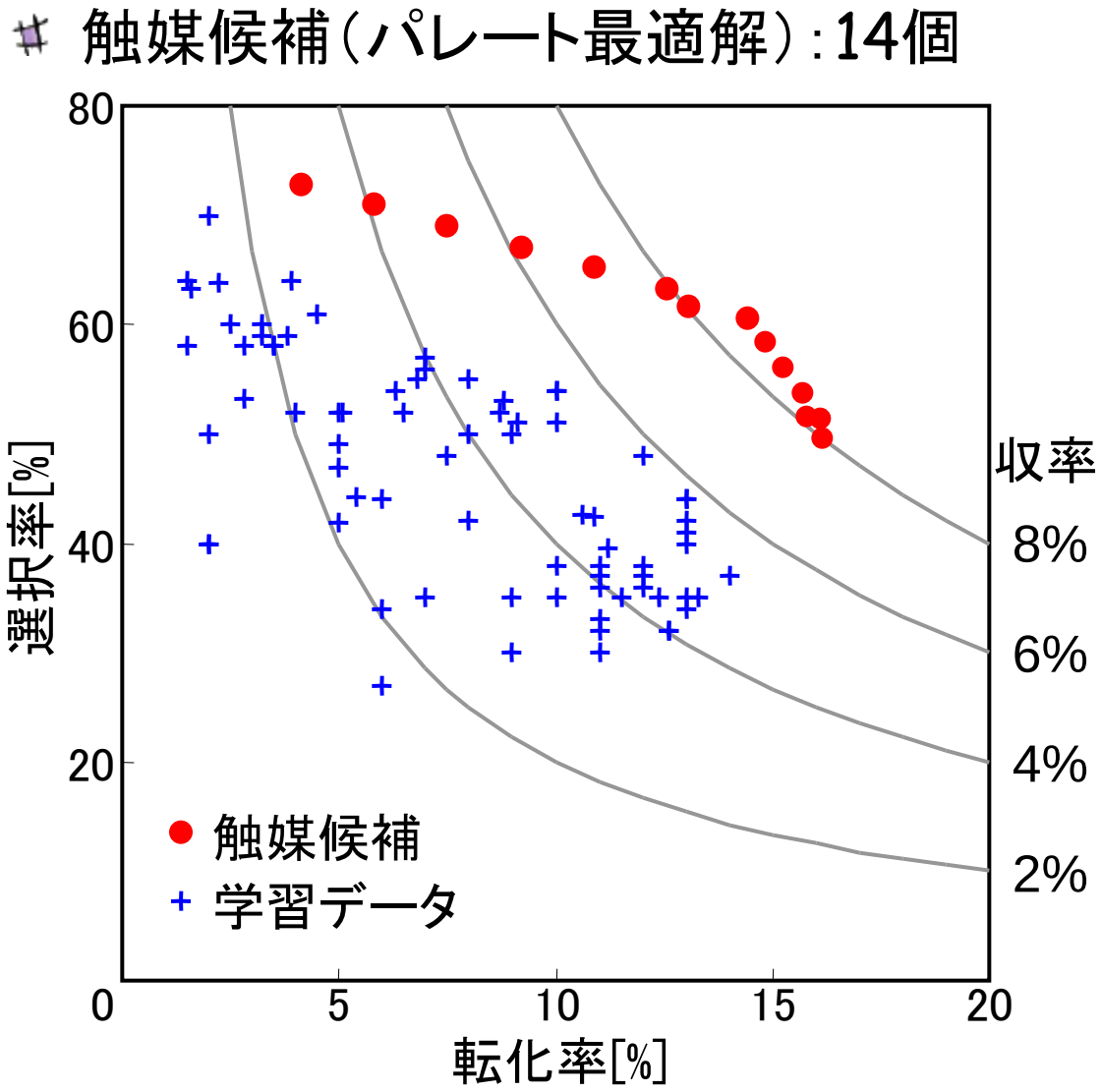
選択率

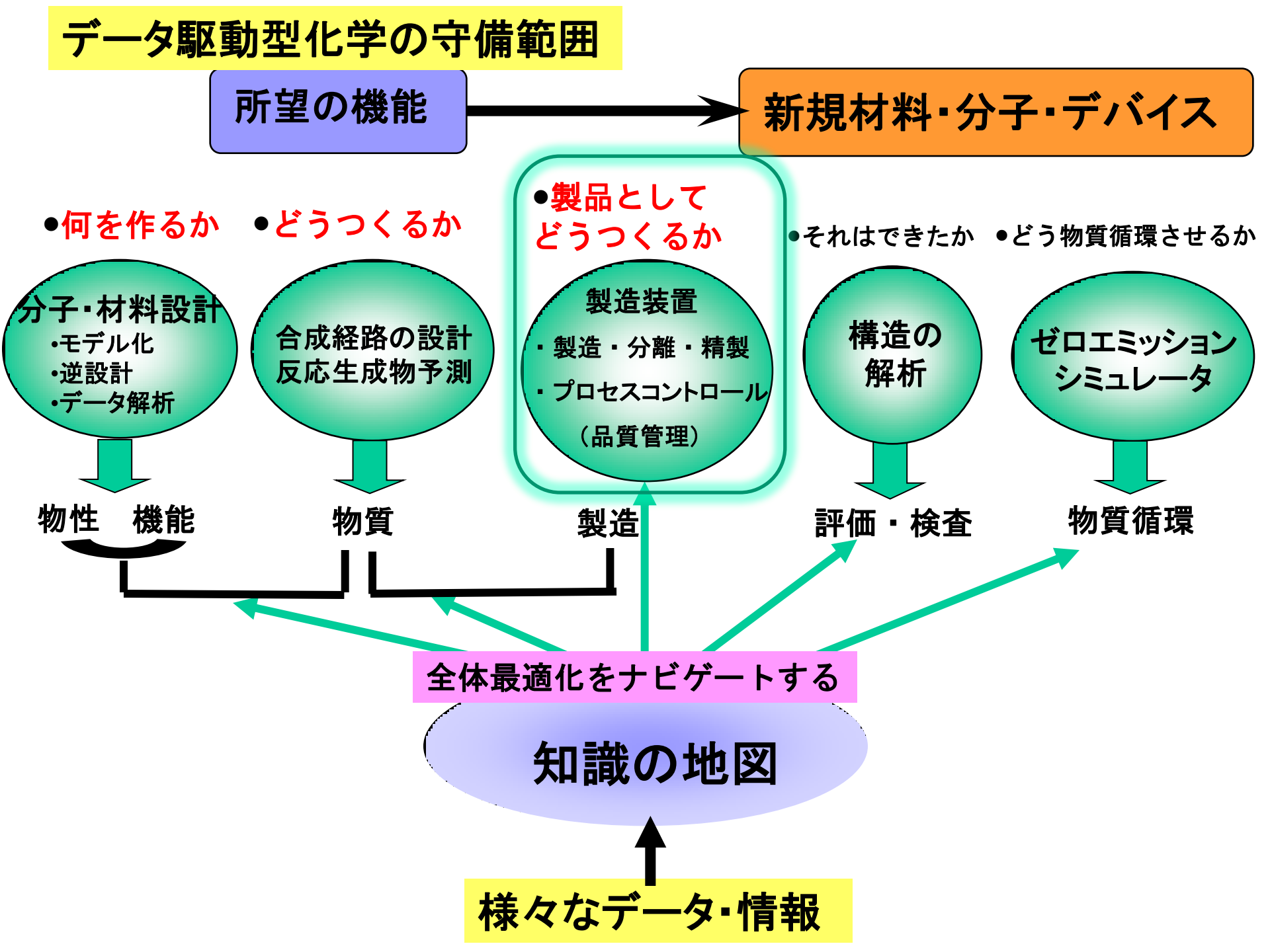


転化率、選択率ともにR^2、Q^2が高い
文献値と計算値のグラフも点が対角線付近に集まった

予測性の高いモデルを構築できた

パレート最適解の分布


触媒候補（パレート最適解）：14個
80
60
40
20
0
5
10
15
20
選択率[%]
転化率[%]
収率
8%
6%
4%
2%
触媒候補
学習データ


データ駆動型化学の守備範囲
所望の機能
新規材料・分子・デバイス
●何を作るか
●どうつくるか
●製品として
どうつくるか
●それはできたか
●どう物質循環させるか
分子・材料設計
•モデル化
•逆設計
•データ解析
合成経路の設計
反応生成物予測
製造装置
・製造・分離・精製
・プロセスコントロール
（品質管理）
構造の
解析
ゼロエミッション
シミュレータ
物性　機能
物質
製造
評価・検査
物質循環
全体最適化をナビゲートする
知識の地図
様々なデータ・情報

ソフトセンサー手法の開発と
プロセス管理への応用

観測が容易な温度・圧力などのデータから、オンライン観測が困難な生産品の濃度・特性など品質をリアルタイムに予測しながらプラントを監視・制御する。

生産の効率化および安全確保の観点から確かなソフトセンサー手法が求められている。生産支援技術として注目

対象：　重合リアクター、薄膜製造、蒸留塔など

船津公人、ソフトセンサー入門、コロナ社、2015年

ソフトセンサーとは何か

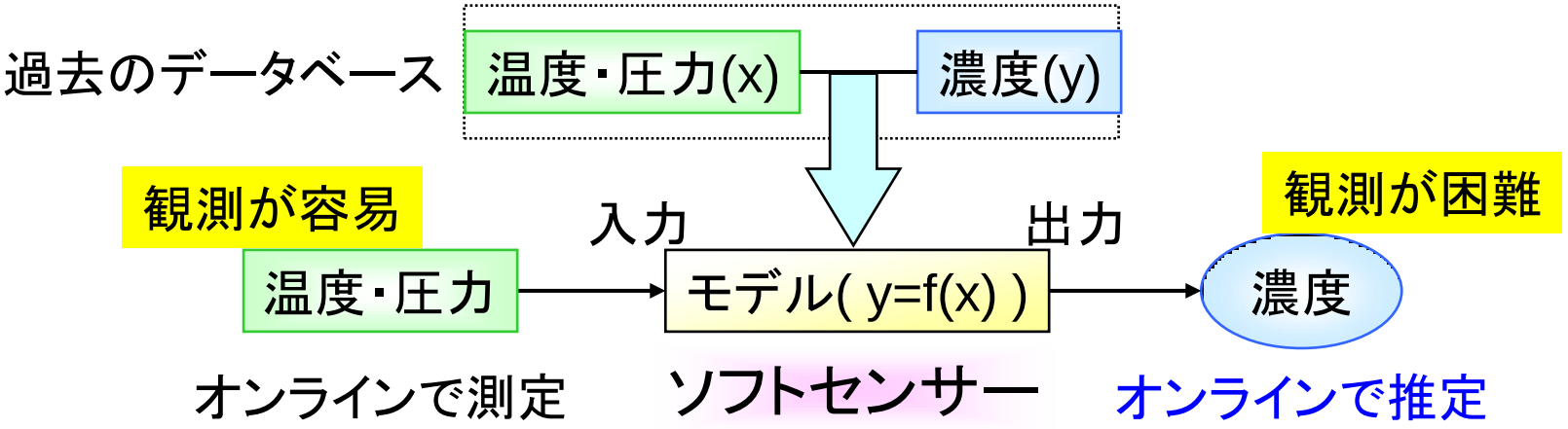
化学プラントを正常に運転するためには、
プロセスを監視し、適切に制御する必要がある

温度 圧力 濃度 など

そのためには、プロセス変数を正確に低コストで測定することが重要である

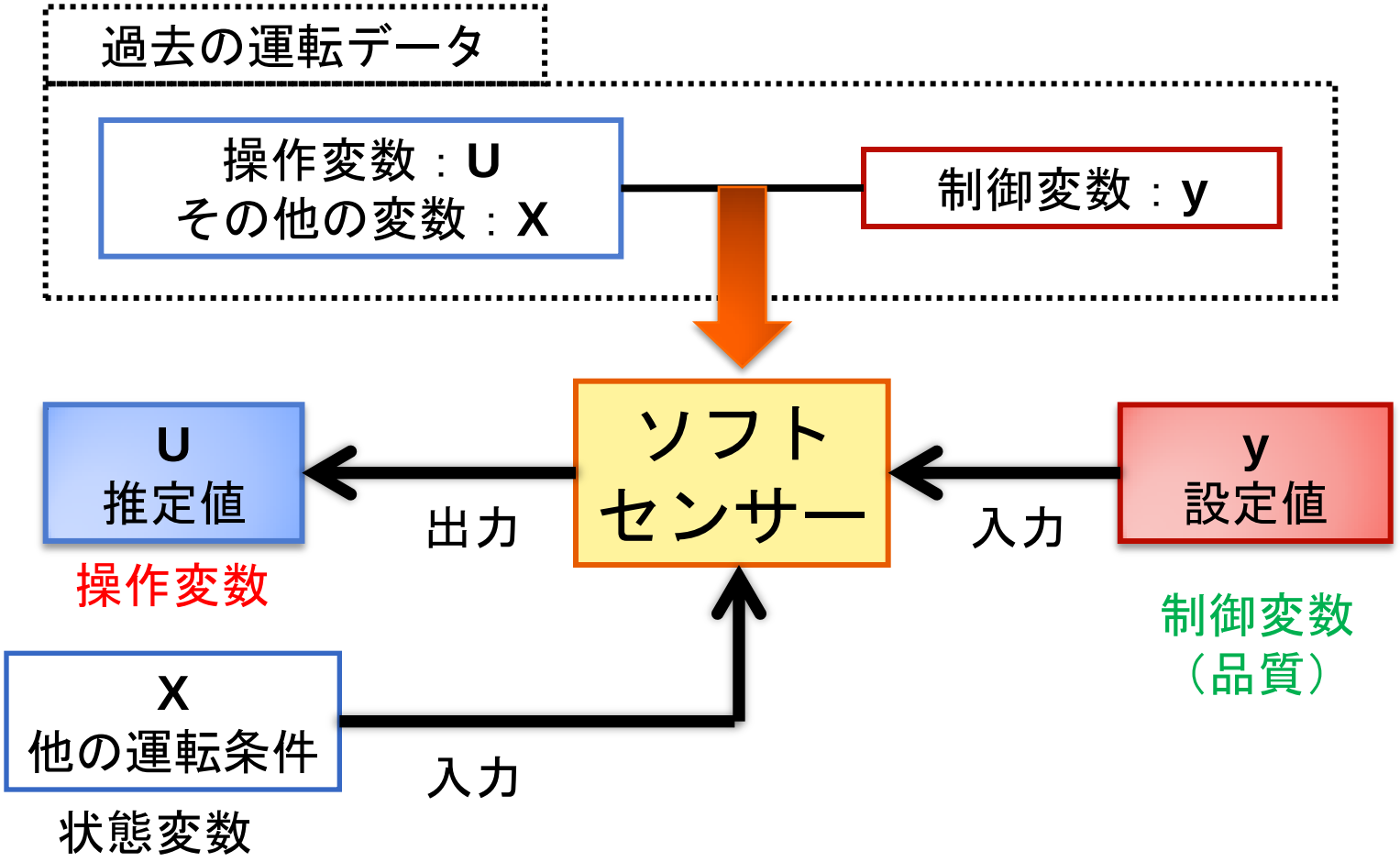
しかし、全ての変数を簡単に測定できるとは限らない

✓技術的に困難　　✓分析に時間がかかる

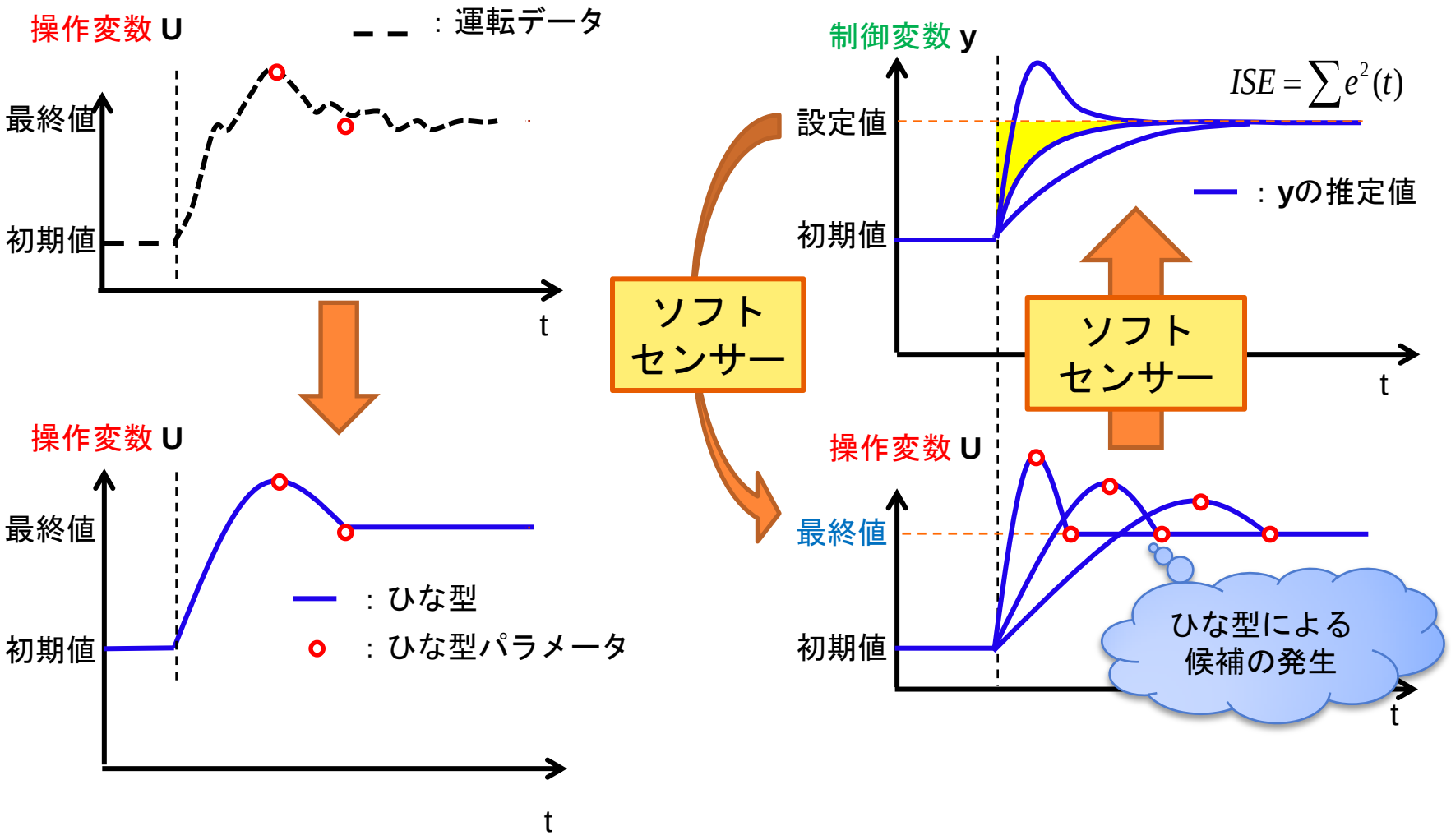


逆向きに使うと制御に利用できる

ソフトセンサーを用いたプロセス制御

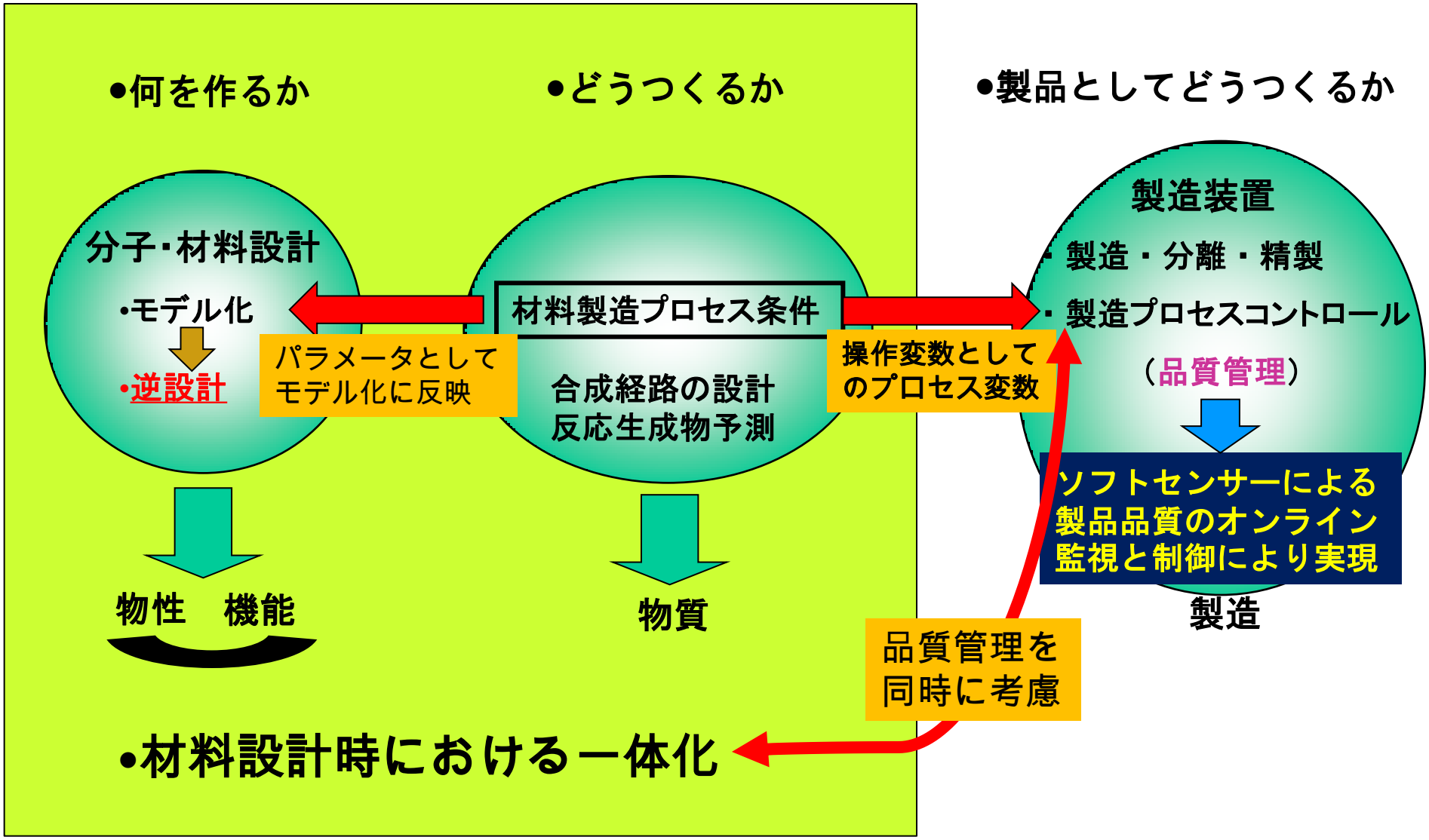


ソフトセンサーの逆解析による操作量決定の流れ



•木村 一平, 金子 弘昌, 船津 公人, ソフトセンサーとその逆解析を利用した新規 フィードフォワード制御手法の開発, 化学工学論文集, 41(1), 29-37. 2015

プロセス・インフォマティクスを考える



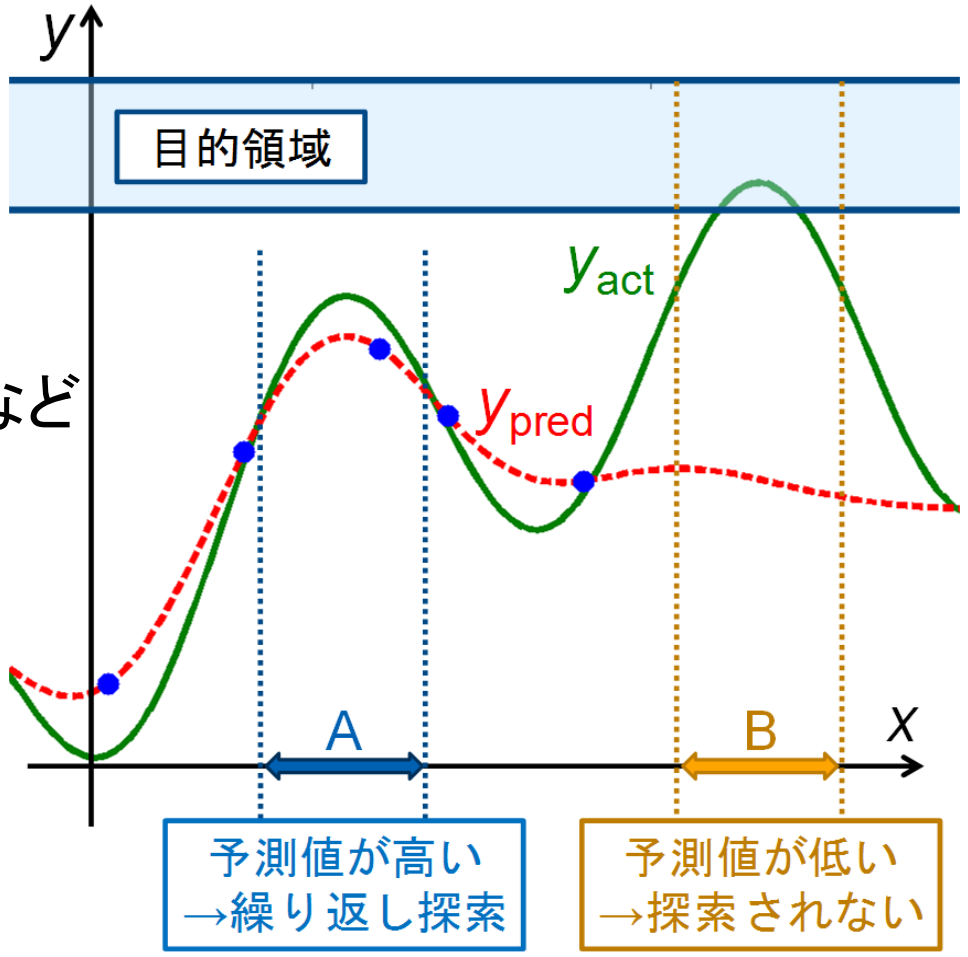
材料設計・プロセス条件検討・品質管理までを一気通貫で扱える

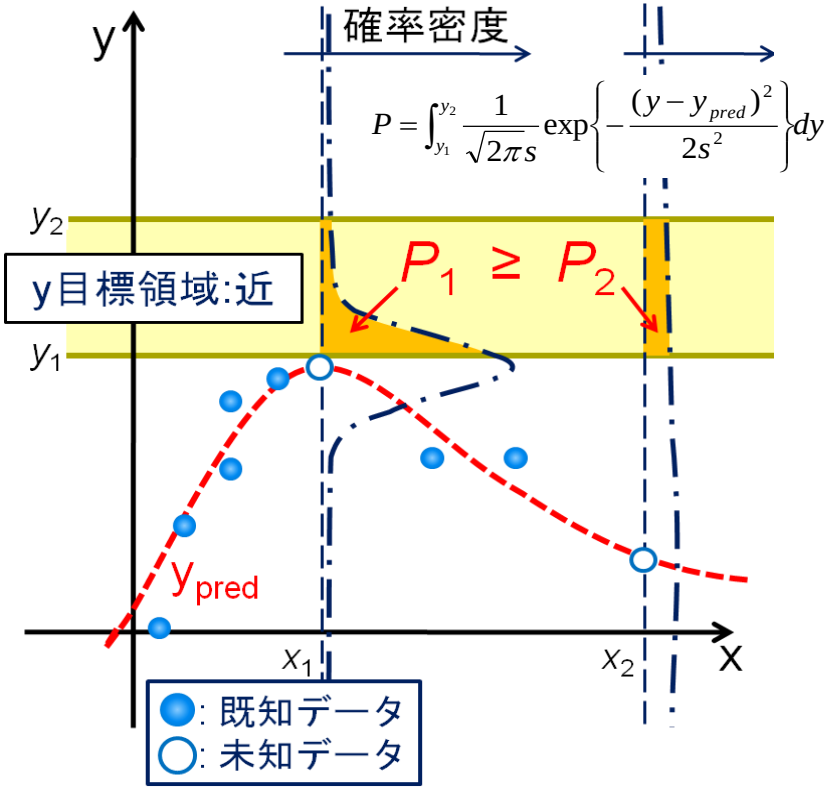
最後にもうちょっと

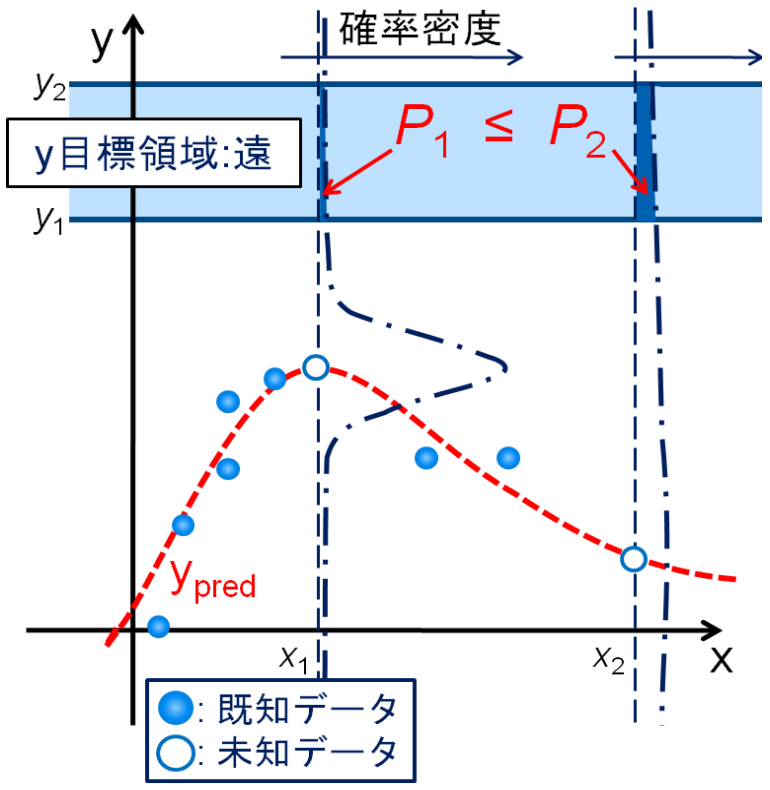
少ない実験データから出発して
目標物性を達成するには

y:　目標物性

x:　組成、構造、プロセス条件など







(a) 目標領域と予測値が近いケース　　(b) 目標領域と予測値が遠いケース

GP法によって計算された予測値とその分散を用いて、物性yが目標値を満たす確率*P*が計算できる

学習を通しながら迅速に目標に到達

・*X*は材料のパラメータだけではなく、運転条件、装置構造に関するパラメータも扱える。

・*X*を利用した時間のかかる様々なシミュレーションも効率よく行え、目的特性を持つ材料/装置構造候補に迅速に到達できる。

・T. Kishio, H. Kaneko, K. Funatsu, Strategic Parameter Search Method Based on Prediction Errors and Data Density for Efficient Product Design, Chemometrics and Intelligent Laboratory Systems, 127(1), 70-79, 2013

・A. Nakao, H. Kaneko, K. Funatsu, Development of an Adaptive Experimental Design Method Based on Probability of Achieving a Target Range through Parallel Experiments, Industrial & Engineering Chemistry Research, 55(19), 5726-5735.

ビッグデータ・サイエンスとは

第1のパラダイム

科学歴史において、アリストテレスの時代からはじまった経験・実験的手法

第2のパラダイム

ライプニッツ、ケプラー、ニュートン、マクスウェル（電磁気学創始者）など。
観測データ群を分析し、その背後にある論理・法則を見いだしていく方法

第3のパラダイム

解析解が得られない非線形方程式をノイマン型計算機、すなわちコンピュータの演算能力によって数値解の形で解いていく方法。
計算機シミュレーションにより解決することを目指した方法論

ビッグデータ・サイエンスとは

第4のパラダイム

大量のデータをもとに統計的な推論モデルにより科学を展開することを目指して提案された。

点から線へ、線から面へ、面から多次元へ。
そのために必要もしくは欠落している情報・データを実験、シミュレーションによって意図的に得る。

(1)データを集約（量だけではなくDiversityが重要）し、
(2)データにもとづいたモデリングを行い、
(3)分野を越えて融合研究を進める科学研究スタイル。

分野融合の発想で研究が進められている。
学際的な融合から新しい研究の方向性を創出する。

マテリアルズ・インフォマティクスを根付かせるには

- 日常的なデータの取り扱いを議論・検討する、部門を超えた若手の集まりを組織すること
- 積極的にデータ・情報活用を推進し、たとえわずかでも利用可能なものは積極的に活用していこうとする意識改革が必要

往々にして人はこのようなデータ解析結果や予測システムに対しては完璧性を求めがちで、完璧でないものは使えない、使わないという極端な発想を持つことが多い。しかし、どんなものにもscopeとlimitationがあるように、データ活用による予測・設計にもデータの偏りや測定誤差などに起因する予測精度の不十分さが付きまとうことがある。

- 対象に応じてどこまで予測できれば良いのか、どこまで使えるのか、何が欠けているのかを正しく見極められる人材育成が必要

そして

- 材料開発、それに基づくデバイス開発、そして生産などで得られるデータ・情報は次の新たなものを生み出すための土壌のようなものである。
- 土壌やその耕し方を他に依存するのは、切り花を買ってくるようなものでしかない。そのような花はすぐ枯れる。
- 世界的にもはやデータ駆動型化学の意識が後戻りすることはない。我々が意識すべきは、土壌の日常的な充実とその土壌を耕す技術の蓄積である。
- まさにこのための継続的な人材育成が求められる。