A01: 新しい概念に基づいたアルゴリズム・最適化の問題創出とその効率的求解方法の研究

帰納バイアスと分子の組合せ的表現・幾何的表現

たきがわ　いちがく
瀧川 一学

https://itakigawa.github.io

理化学研究所 革新知能統合研究センター
iPS細胞連携医学的リスク回避チーム

北海道大学 化学反応創成研究拠点
(WPI-ICReDD)

AFSA
https://afsa.jp

自己紹介：瀧川 一学 (たきがわ いちがく)

専門：機械学習と機械発見

特に離散構造を伴う機械学習 + データ中心的な自然科学研究

現在の主業務：**幹細胞生物学(理研) + 化学(北大)**

10年 北大 **工学研究科** システム情報工学専攻

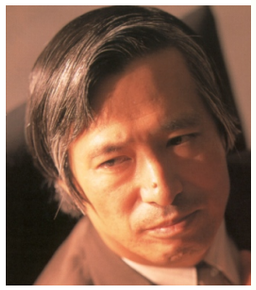
(1995～2004) "劣決定信号源分離のL1ノルム最小解の理論分析"

7年 京大 **化学研究所 バイオインフォマティクスセンター**

(2005～2011) **薬学研究科 医薬創成情報科学専攻**

7年 北大 **情報科学研究科** 情報理工学専攻

(2012～2018) **JSTさきがけ 材料インフォマティクス領域**



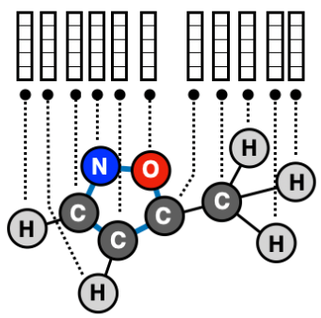
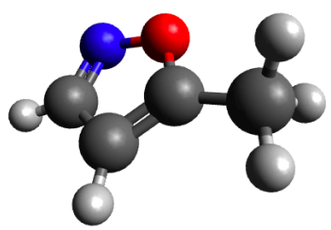
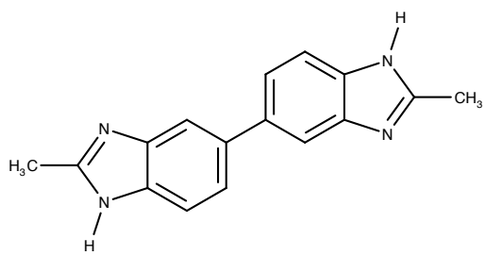
？年 理研(京都) **AIPセンター iPS細胞連携医学的リスク回避チーム**

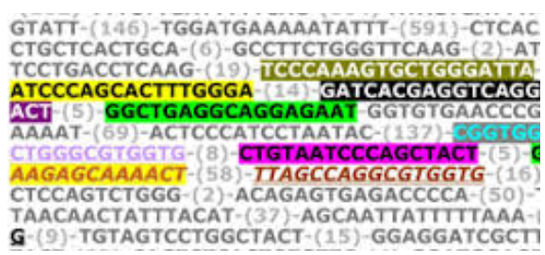
(2019～) **北大 化学反応創成研究拠点 (クロスアポイント)**

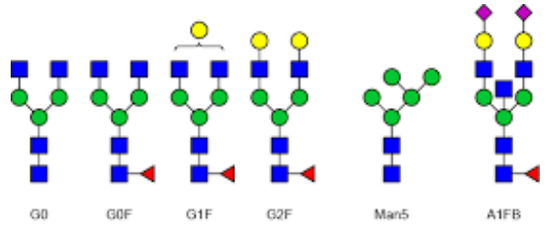
離散構造を伴う機械学習

集合、系列/文字列、組合せ、置換、木構造、グラフ/ネットワーク構造、…

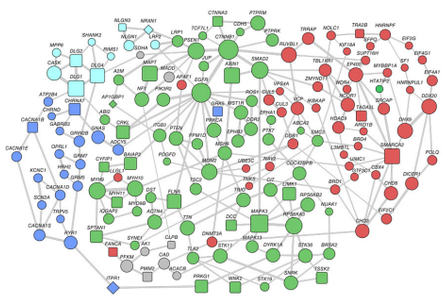
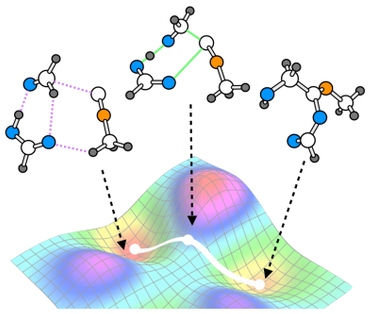
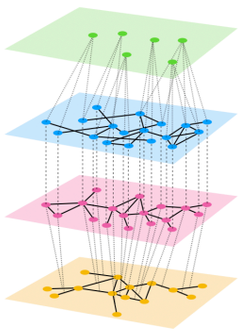
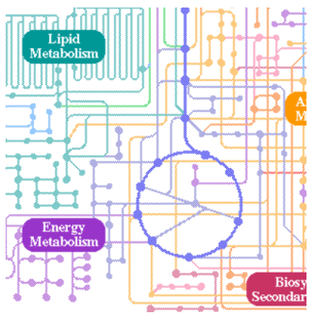
- **対象が「離散構造」を持つ**

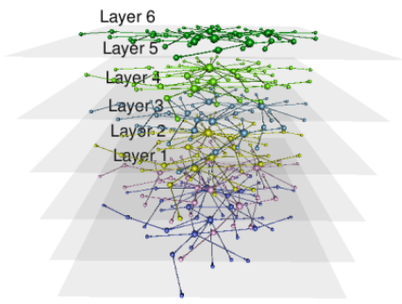




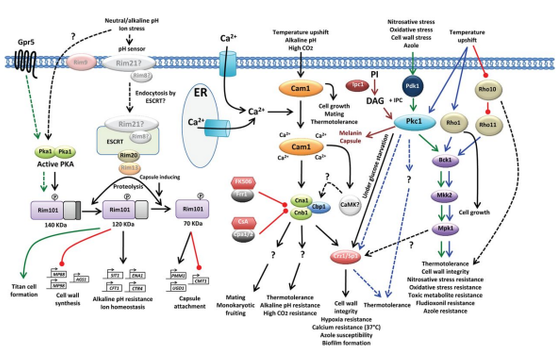




- **対象の関係が「離散構造」を持つ**

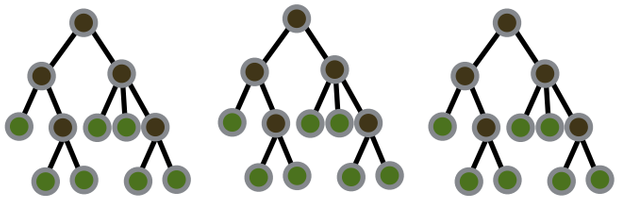







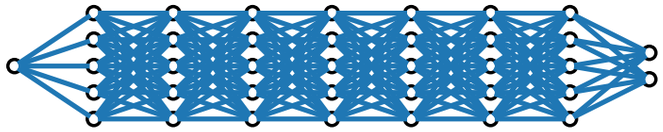


- **モデルが「離散構造」を持つ**

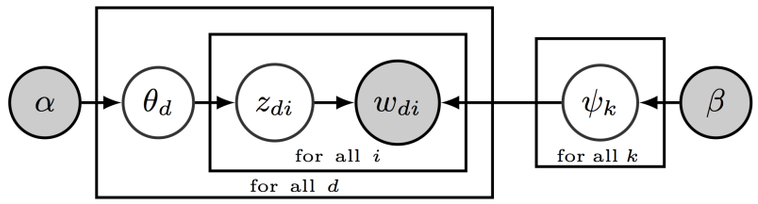
決定木・決定DAG



ニューラルネットワーク



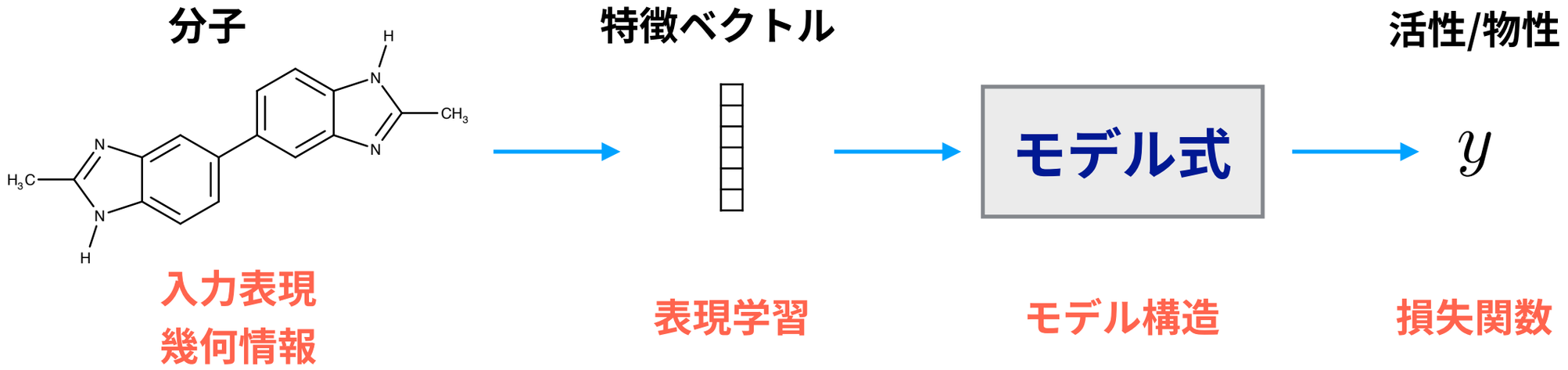
確率的プログラミング



帰納バイアスと分子の組合せ的表現・幾何的表現

お題の背景にあるテーマ：機械学習から機械発見へ

主業務を踏まえた主に次のような機械学習の問題がいつも頭にあります。



- **Underspecification**の問題：自由度が高いため、モデルの近似が有限サンプルの擬似相関にトラップされた表面的でmisleadingなものに帰することが多い
- **グレイボックス最適化**：機械学習を科学研究を支える「良い道具立て」に格上げするには**この辺**に科学的洞察に応じた**「意味のある制約 (帰納バイアス)」**を専門家と協働してデザインする必要がある (データにできる情報は常に部分的)
- 機械発見に向けて機械学習が使われる場面に適合した**新しい定式化**も必要

宇野班での私の関心

「A01: 新しい概念に基づいたアルゴリズム・最適化の問題創出とその効率的求解方法の研究」

研究はとても面白かったけど**全然使わなかった(使われなかった)技術たち**を整理し**主業務の良い道具立て**となる「新しい問題創出・定式化」を得る

→ なぜ使われなかった？どうすればもっとうまくやれた？の理解

→ 主業務であるリアルな化学・生命科学研究のための**機械学習・機械発見**

核となる関心

1. **有限の連続値ランダムデータを要素に含む離散最適化**
   Rashomon効果と解釈多様性
2. **グレイボックス最適化**
   融合研究と帰納バイアスのデザイン
3. **人間を要素として含む最適化**
   誰のためのデザイン？

本日のケーススタディ：巨人の肩と反省の上に立つ

① Subclass coverの研究

ユースケースを真剣に考える

「すごく計算時間のかかるその厳密最適化は本当に必要ですか？」

② Graph miningとそれに基づく機械学習の研究

データマイニングは「発見」じゃない！？

「ビール買うとき実際にオムツを見かけたことありますか？」

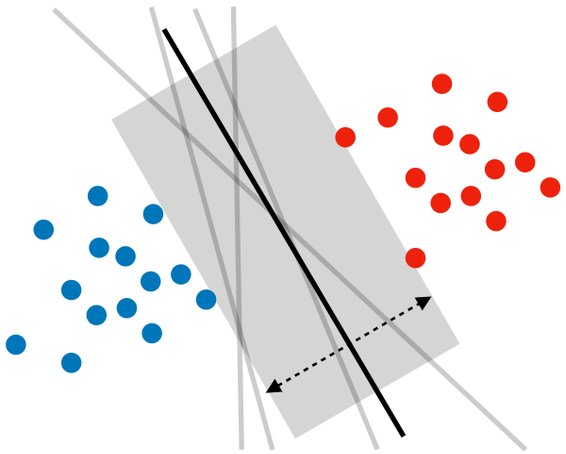
③ SIG-FPAIと発見科学の研究

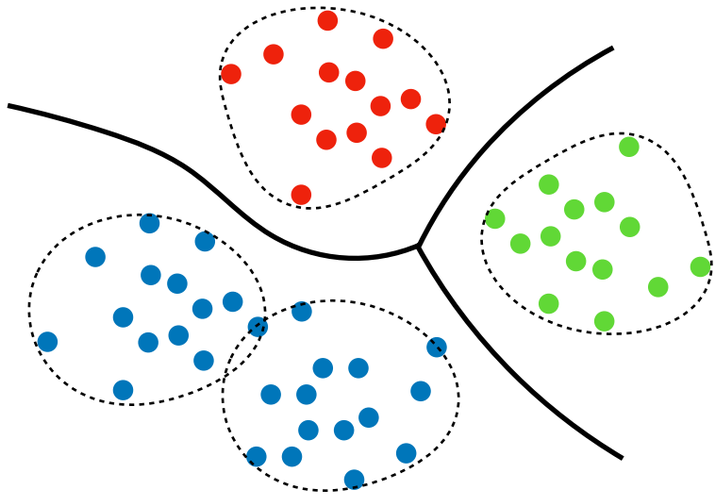
機械学習から機械発見へ：「発見」は合理化できるのか？

- 有川先生の科研費特定領域「巨大学術社会情報からの知識発見に関する基礎研究 (略称:発見科学)」(1998 ~ 2001)の計画・方法・成果を再考する
- 領域出版物「発見科学とデータマイニング」(2001, 森下 真一・宮野 悟 編)

① Subclass coverの研究

背景とアイデア：凸集合による部分クラス被覆





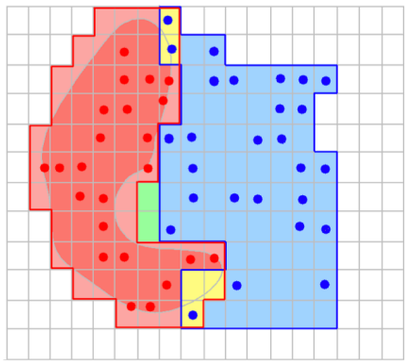
機械学習のマージン理論
線形分離可能な場合、マージン最大化(or ベイズポイント)が汎化に良い

線形分離可能じゃない場合は？

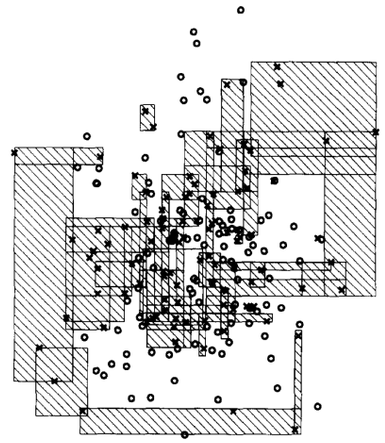
多クラスの場合は？
(どことどこのマージンを最大化？)

部分クラスに分解して局所的な分離として解けると自然
(classificationは部分クラスへのk-最小射影で解く)

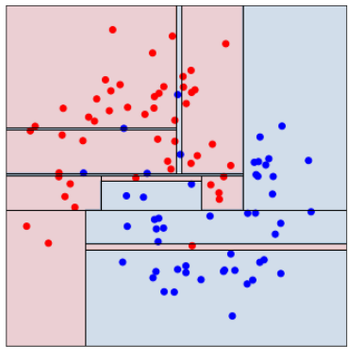
→「点集合Xから一意に定まる凸集合S(X)が他のクラスの点を含まないような極大な部分集合Xの列挙」として解く



Logical analysis of data (LAD)
by Peter L. Hammer



Subclass method
by Mineichi Kudo



Decision Trees / CART
by BFOS

① Subclass coverの研究

やったこと: 最小包含球族による被覆 + 凸包による被覆への拡張

Takigawa I, Kudo M, Nakamura A, Convex sets as prototypes for classifying patterns. (EAAI 2009)
Takigawa I, Kudo M, Nakamura A, The convex subclass method: combinatorial classifier based on a family of convex sets (MLDM 2005)

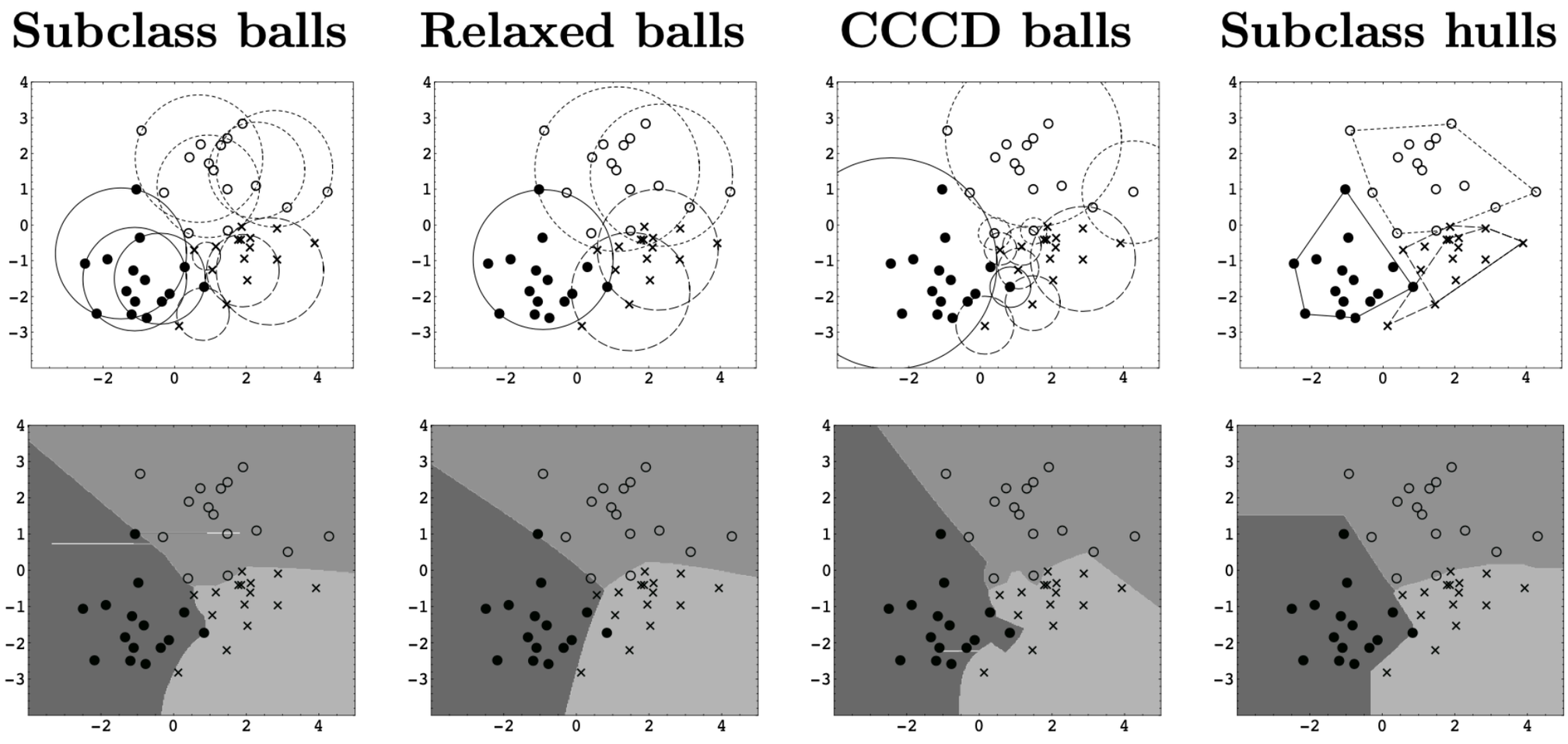


Fig. 4. Subclass covers and decision boundaries. A relaxed subclass cover with balls of $\xi = 1$, and a class cover with Class Cover Catch Digraph Method (Priebe *et al.*, 2003) are also shown.

① Subclass coverの研究

この問題は考えだすととても面白いが…

- **極大集合(台集合がSperner族になるように)の列挙**なのでmaximal freqent setsやhypergraph traversals (or 単調な論理関数の双対化)と本質的に同じ構造を含み、当時、宇野先生や牧野先生の解説や論文で勉強 (c.f. Heikki Mannilaのborders of theories)
- 一方で、本当に全ての凸包部分クラス被覆を厳密列挙すると**「本来の目的としては好ましくないもの」が色々含まれる**ことも判明(e.g. “spiky” hulls)。最小包含「楕円体」versionも同様
- 計算時間かかるわりにCV精度的にはSVMやRF等の主流の方法と比べて有意に良いわけでもなかった。

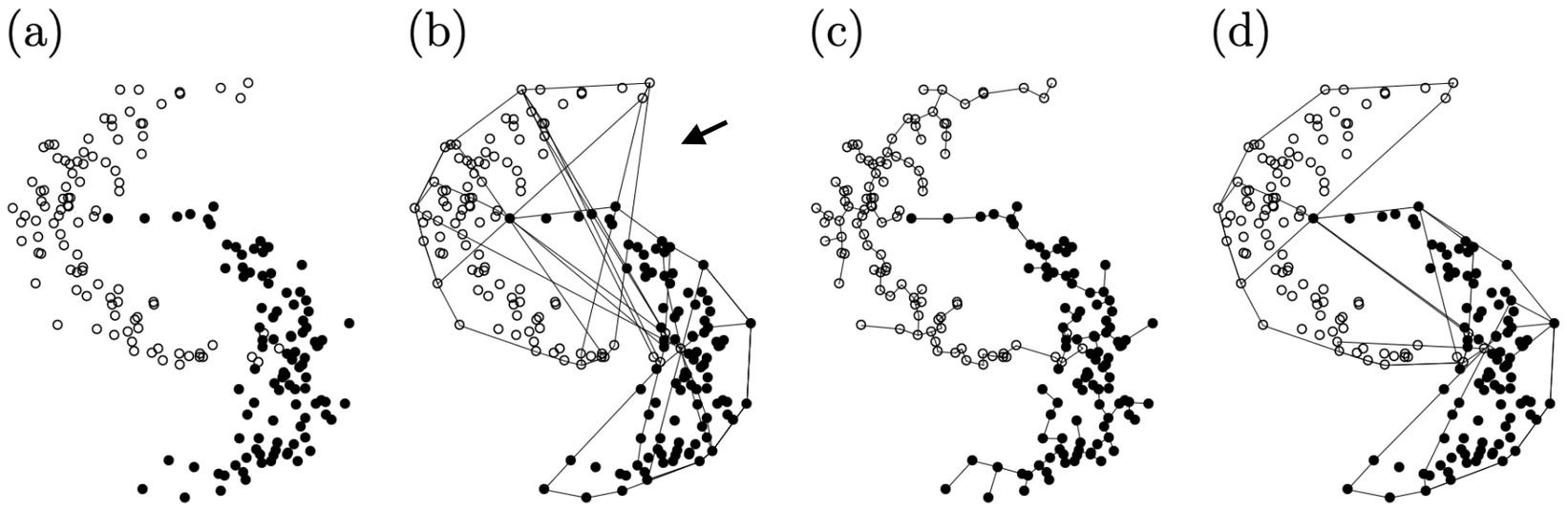
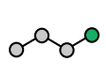
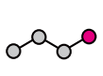
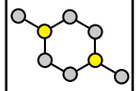
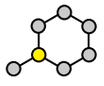
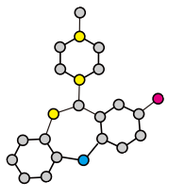
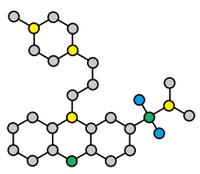


Fig. 3. (a) An example dataset consisting of 2 classes, (b) the subclass covers of each class by Algorithm 2 including the “spiky” convex hulls, (c) the minimum spanning trees of each class for the complete proximity graphs on the positive samples, and (d) the subclass cover of each class by Algorithm 3 using the trees (c).

② Graph miningとそれに基づく機械学習の研究

背景：グラフマイニングとグラフ分類

部分構造

y	入力構造							…
0.1		0	0	1	1	1	0	…
0.7		1	0	0	0	0	1	…
0.9		1	1	0	1	1	0	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱
1.2		1	0	1	1	1	0	…

どういう部分構造を用意すると良い？

Table 19.2. SAR performance of different descriptors.

Datasets	fp	ECFP	MK	FS	GF
NCI1	0.30	0.32	0.29	0.27	**0.33**
NCI109	0.27	**0.32**	0.24	0.26	**0.32**
NCI123	0.25	**0.27**	0.24	0.23	**0.27**
NCI145	0.30	0.35	0.28	0.30	**0.37**
NCI167	0.06	0.06	0.04	0.06	**0.07**
NCI220	**0.33**	0.28	0.26	0.21	0.29
NCI33	0.26	0.31	0.26	0.25	**0.33**
NCI330	0.34	**0.36**	0.31	0.24	**0.36**
NCI41	0.25	**0.36**	0.28	0.30	**0.36**
NCI47	0.26	**0.31**	0.26	0.24	**0.31**
NCI81	0.27	**0.28**	0.25	0.24	**0.28**
NCI83	0.26	**0.31**	0.26	0.25	**0.31**

fp: あるサイズまでの**すべての**サイクルとパス

ECFP: 出現する**すべての**r-近傍部分グラフ

MK: 人間が専門的見地から決め打ちで選んだもの

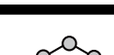
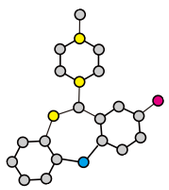
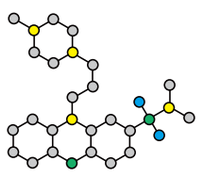
FS: 頻度σ以上の**すべての**部分グラフ (頻出部分グラフ)

GF: あるサイズまでの出現する**すべての**部分グラフ

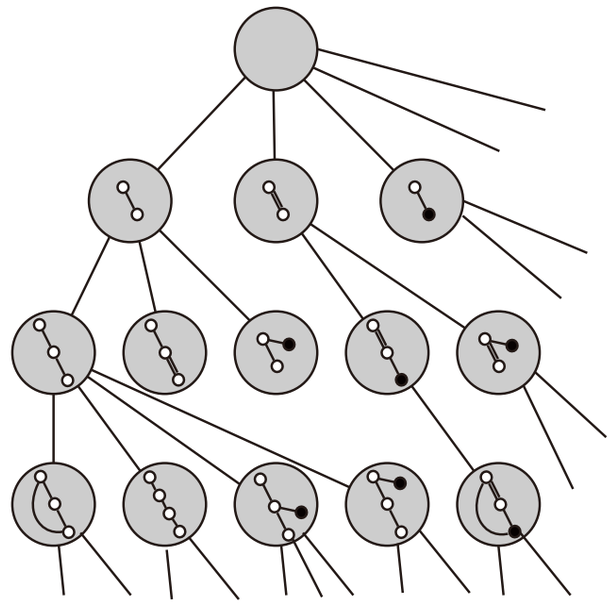
② Graph miningとそれに基づく機械学習の研究

背景：グラフマイニングとグラフ分類

部分構造

y	入力構造							…
0.1		0	0	1	1	1	0	…
0.7		1	0	0	0	0	1	…
0.9		1	1	0	1	1	0	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱
1.2		1	0	1	1	1	0	…

前提となる知見：出現部分構造は探索木を辿って逐次重複なく漏れなく調べられる



gSpanの木型の探索空間 (列挙木)

↓

適切に打ち切れば下記が得られる

FS: 頻度σ以上の**すべての**部分グラフ **(**頻出部分グラフ)
GF: あるサイズまでの出現する**すべての**部分グラフ

② Graph miningとそれに基づく機械学習の研究

やったこと: すべての出現部分構造から「予測に貢献するのだけ」を選ぶ

① 線形モデルに対して「loss + L1罰則 + L2罰則」を座標勾配降下で解く！

Takigawa I, Mamitsuka H, Generalized sparse learning of linear models over the complete subgraph feature set (TPAMI 2017)
Takigawa I, Mamitsuka H, Efficiently mining δ-tolerance closed frequent subgraphs (Machine Learning 2011)

方針：Branch & Boundで更新後の係数が非ゼロのものだけ計算

Brute-force (現実的には動かない)

until 解が収束
  for x in すべての出現部分グラフ:
    $b_x \leftarrow b_x - \eta \partial L / \partial b_x$
$pred_y = b_0 + \Sigma_x b_x$

※ 実際は対角Hessianも使うBCGD法に適用

提案アルゴリズム (スパース条件下では解が求まる)

until 解が収束
  for x in DFS of 列挙木(すべての出現部分グラフ):
    if xの子孫zでは更新後のb_zがゼロ:
      prune
    $b_x \leftarrow b_x - \eta \partial L / \partial b_x$
$pred_y = b_0 + \Sigma_x b_x$

② 決定木学習のsplitter searchについて同様にして非線形モデルを得る！

Shirakawa R, Yokoyama Y, Okazaki F, Takigawa I. Jointly learning relevant subgraph patterns and nonlinear models of their indicators. (MLG 2018)

「いま手元にある事例Sを最もよく分ける部分構造xの探索」を再帰的に行うので、boundを求めておけば直接branch & boundが適用できる (津田先生のgBoostやgLARS的に)

→回帰木を得られれば(正則化)勾配ブースティングなどで木アンサンブルへも拡張できる

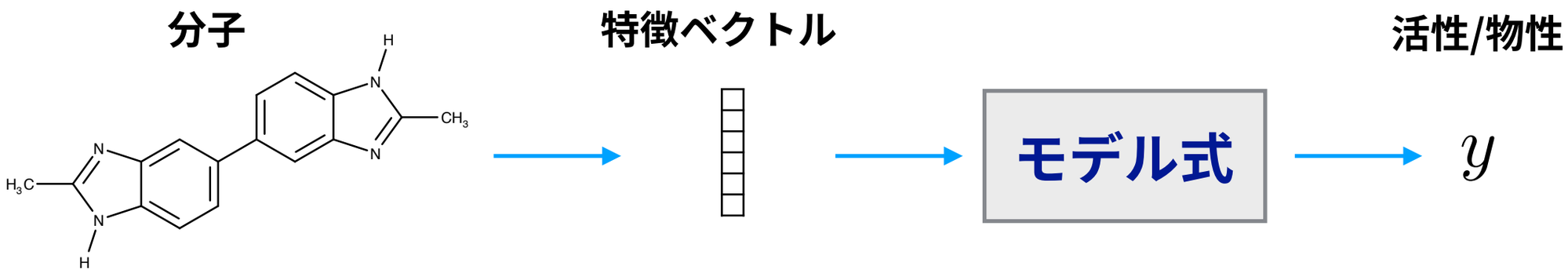
② Graph miningとそれに基づく機械学習の研究

- 「離散ラベルのついたグラフ」の分類なら**「部分構造の有無(or カウント)」は最も汎用性の高い特徴量**で、これを厳密に全部調べてその中から予測に貢献するものだけを選ぶ、という方針でかなり広いクラスの線形学習・非線形学習が実現できることを示した。
- **非常に残念なことに**「分子グラフの活性・物性予測」のタスクについてはフェアに比べると**ヒューリスティックな~~もっといいかげんな~~方法と比べて精度向上は有意とは言えない**
- GFは予測精度を出そうとすると特徴数が膨大になり結果機械学習にも**計算時間がかかる**。私たちの方法も同様に部分グラフパターンの厳密列挙に基づいており、計算時間が必要
- 一方、ECFPは非常に優秀で「グラフデータ全部なんか見ずに」**各入力グラフ一つだけ**から超速アルゴリズム(Morgan Algorithm)で素早く計算できる。WL + “Hashed Fingerprint”
- 精度が同程度ならあえて提案法やGFを使うメリットはすくない (他の選択肢を優越しない)
- 特に分子データは数が多いため、部分グラフ列挙に依拠して手法を構築すると**スケールしない**。そのためGFや提案法だけでなく、gBoostやgLARSも含めて実用的に使われる例をほとんどみないし我々自身も使っていない… (訓練データの全対比較が必要なグラフカーネルも似たような立場？)。一方、ECFPと共によく使われるGNNはこの制約をうけない。

  参考) RECAP, BRICSなど分子構造データから系統的にフラグメント集合を生成する手法も広く使われる。

③ SIG-FPAIと発見科学の研究

そもそもなぜ↓を考えるかというと真の目的は「予測」ではなく「発見」…



- **「機械学習」と「機械発見」のギャップ**に悩んでいたころ、たまたまAI学会FPAI研究会の主査をやっており、AI学会から30周年で研究会特集をやるから一種研究会のFPAIは何か書くようにと依頼がくる。
- 私はそもそもFPAI界隈で仕事をしてきたわけではないので、何か書けと言われても30年の歴史なんてよく知らない。そこで、文献調査を行いSIG-FPAI自体について「研究」することにした。
- FPAI(正確には改称前のFAI)とAI学会は同時にできたもので、初代主査の有川先生の率いていた科研費特定課題がまさに**データマイニングによる「発見」**を体系的に目指していたことを知る

③ SIG-FPAIと発見科学の研究

KAKEN　研究課題をさがす　研究者をさがす　KAKENの使い方　日本語

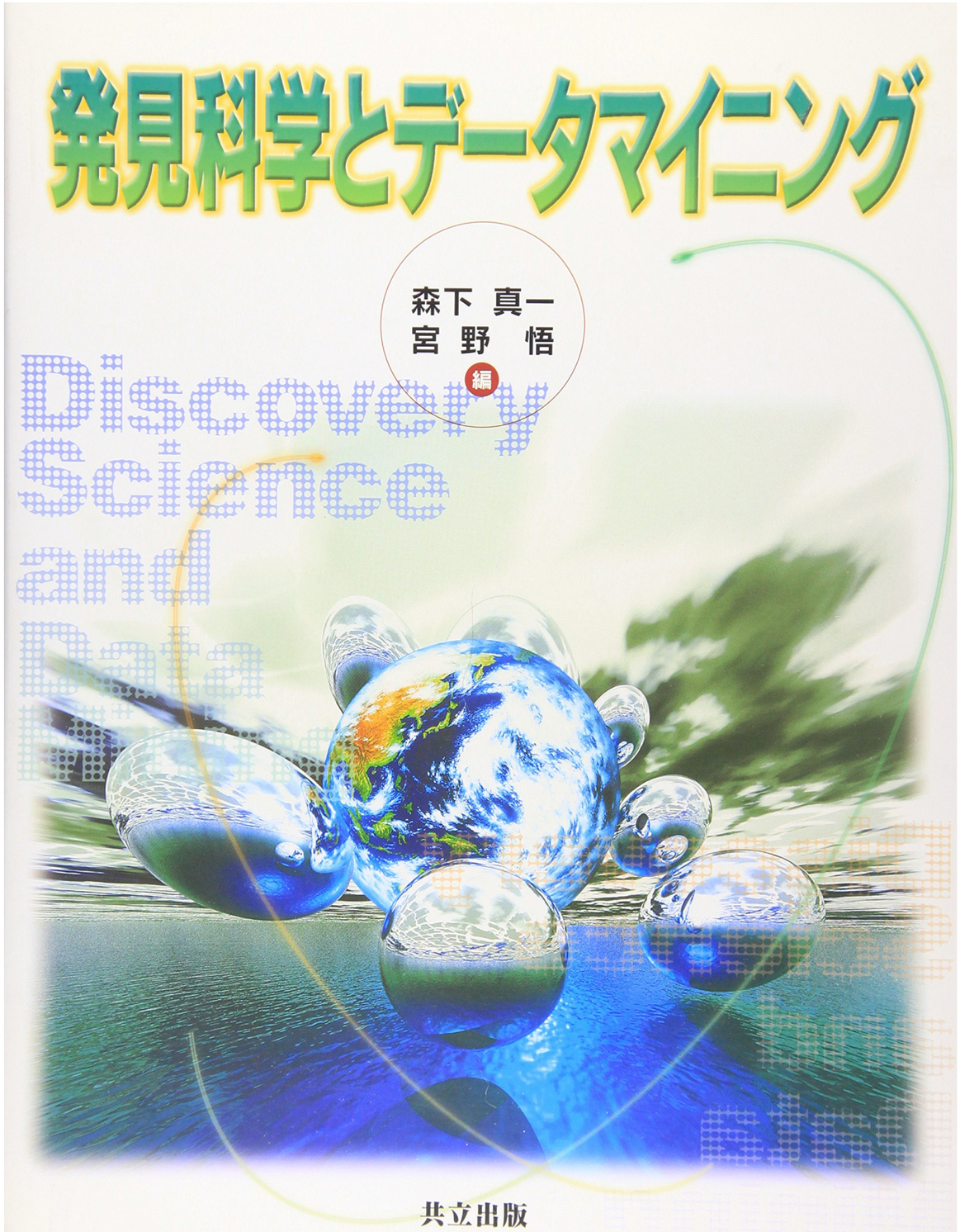
巨大学術社会情報からの知識発見に関する基礎研究　研究課題

サマリー

研究課題/領域番号	10143106
研究種目	特定領域研究(A)
配分区分	補助金
研究機関	九州大学
研究代表者	**有川 節夫**　九州大学, 大学院・システム情報科学研究院, 教授 (40037221)
研究分担者	丸岡 章　東北大学, 大学院・情報科学研究科, 教授 (50005427) 佐藤 泰介　東京工業大学, 大学院・情報理工学研究科, 教授 (90272690) 佐藤 雅彦　京都大学, 大学院・情報学研究科, 教授 (20027387) 金田 康正　東京大学, 情報基盤センター, 教授 (90115551) 宮野 悟　東京大学, 医科学研究所, 教授 (50128104)
研究期間 (年度)	**1998 – 2000**
研究課題ステータス	完了 (2001年度)
配分額 *注記	**66,400千円 (直接経費: 66,400千円)** 2001年度: 3,000千円 (直接経費: 3,000千円) 2000年度: 15,300千円 (直接経費: 15,300千円) 1999年度: 19,100千円 (直接経費: 19,100千円) 1998年度: 29,000千円 (直接経費: 29,000千円)
キーワード	発見科学 / 知識科学 / データマイニング / データベース / 科学的発見の論理 / アブダクション / 機械学習 / ネットワークエージェント / 知識発見

→ **発見科学、科学的発見の論理、知識発見、機械学習、アブダクション**

③ SIG-FPAIと発見科学の研究



ノップ

機械学習から機械発見へ

Our Studies on Machine Learning and Machine Discovery

有川 節夫*
Setsuo Arikawa

* 九州大学大学院システム情報科学研究科情報理学専攻
Dept. of Informatics, Kyushu University.

1996 年 8 月 28 日受理

Keywords : machine learning, machine discovery, algorithmic learning theory, discovery science.

1. は じ め に
—創造工学を機械化できないか—

1968 年から 1972 年の頃であったと思う．北川敏男先生や国沢清典先生，森口繁一先生達の企画で日本科学技術連盟主催のセミナー(講習会)が定期的に開催されていた．その一つに「創造工学」があった．KJ 法や等価変換論といった代表的な創造工学の手法の提唱者自身による講演があり，私も，北川先生の好意で出席させてもらった．その当時計算理論の研究をしていたので，このような講演が非常に新鮮に感じられ，また，そうした創造工学の手法が，非常に主観的で精神論的なものに感じられた．

もう少し客観的に機械的にそうした手法を実現できないものか，そのような研究はきっと非常に重要になり盛んになるはずだ，というような生意気なことを北川先生に話したように記憶している．

また，1970 年代には日米計算機会議というのが，2〜3 年おきに開催され，その最初の会議で，アメリカの A.W.Biermann が，有限オートマトンを対象にした，文法推論に関する非常に興味深い研究を発表し，それを記述するプログラムがデータのサイズそのものとほとんど変わらないとき，ランダムであるという．したがって，データにアルゴリズム的な規則性がなければ，ランダムということになり，文法推論可能性や学習可能性と対極をなす概念と考えられる．文法推論可能であれば，データ圧縮が可能であるという観点から，簡単な報告を書き，この方面の研究を本格的に展開するつもりでいた．

2. 情報検索は人工知能の基礎である

しかし，その頃スタートしたデータベースと情報検索システムに関連した特定研究で，北川先生が責任者の一人であったため，情報検索関連の研究をすることになり，この計画は無期延期になった．情報検索の研究では，学術情報の生産者でありかつユーザである研究者をシステムのなかに積極的に位置づけ，研究者の主観や偏見を検索に反映できる知的なシステムを構築した．これは，研究者が使い込んでいくとどんどん賢くなり，自分自身の知識が検索に生かされるようになるもので，当時としては非常に新鮮なシステムとして，JICST などでも評価していただいた．情報検索

③ SIG-FPAIと発見科学の研究

発見科学とデータマイニング

この書籍は現在お取り扱いできません。

森下 真一・宮野 悟 編

ISBN	978-4-320-12018-1
判型	A4変型
ページ数	318ページ
発行年月	2001年06月
価格	4,840円（税込）

Pat LangleyとHeikki Mannilaの
重要論文の和訳もついている

③ SIG-FPAIと発見科学の研究

- 大前提として**「私たちはまだ機械発見にはたどりついていない」**
- ということは今までの試みは何か失敗している・何か問題があった
- 2021年現在**「機械による発見」「科学的発見のコンピュータ支援」**は技術やハードの進歩に加えて、「実験自動化」やAI・DXなどのコモディティ化により様々な自然科学分野に渡って、も**のすごく再燃しているトピック**
- 書籍「発見科学とデータマイニング」の内容には、2021年現在世界中で行われている色々な試みがデジャブに見えるくらいテーマの類似が見られる。
  つまり、**方法は変わったが「関心自体はほぼ変わっていない」**

↓

第二次AIブームの頃の研究プロジェクトなので**「(explicitな)知識表現」を暗黙のうちに仮定している**ことが実問題との主たる齟齬かと思われるが(この本で「機械学習」と呼ばれるものは今「機械学習」と呼ばれているものとだいぶ違う)、同時に**今ちょうど見直すべき様々なアイデアや議論**があり、ひきつづき鋭意研究中…

参考：研究会紹介特集で浮きまくった謎エッセイ

学会誌「人工知能」Vol. 34, No 5, 2019年9月 (オープンアクセス)

Permalink : http://id.nii.ac.jp/1004/00010296/

人工知能基本問題研究会（FPAI）

Special Interest Group on Fundamental Problems in Artificial Intelligence

瀧川　一学　理化学研究所革新知能統合研究センター
Ichigaku Takigawa　RIKEN Center for Advanced Intelligence Project.
ichigaku.takigawa@riken.jp, https://itakigawa.github.io/

"論理，学習，知識の表現と獲得，並列計算モデル，知的プログラミング，自然言語理解，パターン理解などに関する人工知能としての基礎的研究 (有川, 1990)"

ところで、今月の人工知能学会誌の瀧川さんのSIG-FPAI記事、
ただの研究会の紹介記事のハズのなのに、えらい見識になっていて、
他の研究会の記事との落差に困惑しました。
みなさまも是非。

kashi_pong先生

考察：新しい最適化の問題創出を目指して

本日のケーススタディ

① Subclass coverの研究
ユースケースを真剣に考える

② Graph miningとそれに基づく機械学習の研究
データマイニングは「発見」じゃない！？

③ SIG-FPAIと発見科学の研究
機械学習から機械発見へ：「発見」は合理化できるのか？

宇野班での私の関心課題

1. **有限の連続値ランダムデータを要素に含む離散最適化**
   Rashomon効果と解釈多様性
2. **グレイボックス最適化**
   融合研究と帰納バイアスのデザイン
3. **人間を要素として含む最適化** ⟷
   誰のためのデザイン？

ORの問題と違い自然科学の対象は自然法則なので**人間を考えなくて良い**と思い違いをしていた

誰のための「発見」？

科学は**人間が**世界を理解するための手段。
発見や理解は**人間の営み。**

↓

“科学研究もまた人間の営みである”
データを人間が集める以上、私たちが抱える認知的バイアス・社会的バイアスから不可避

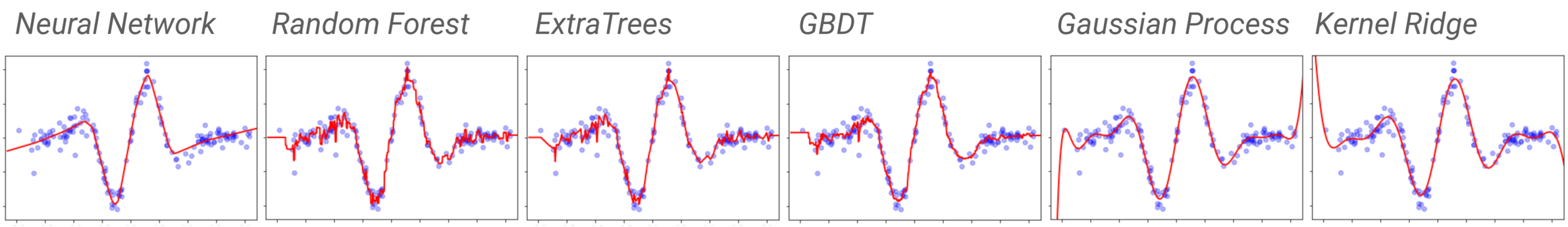
BreimanのRashomon効果と解釈多様性

Rashomon効果 = 予測精度の高い機械学習モデルの多重性(非一意性)

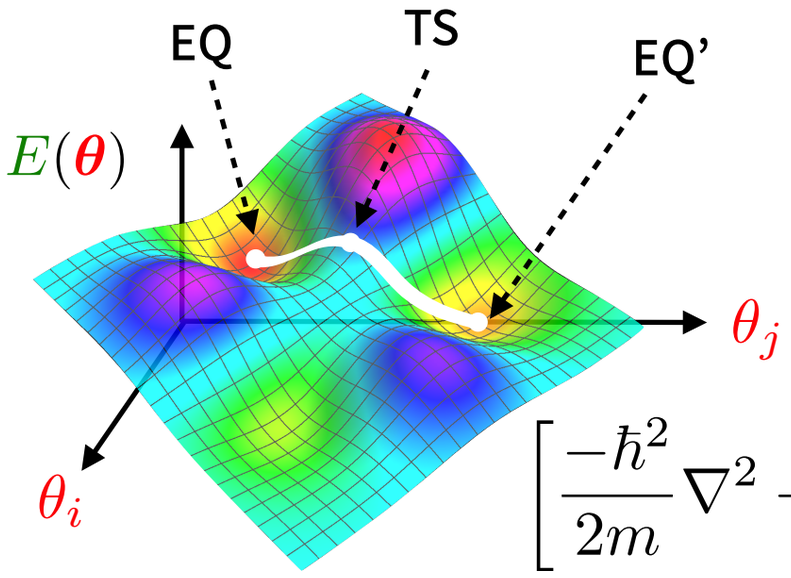
一般のかなり広いタスクにおいて「ほとんど同じくらい予測精度が高い(が、全くやり方は異なる)機械学習モデル」は複数(無数に)存在する

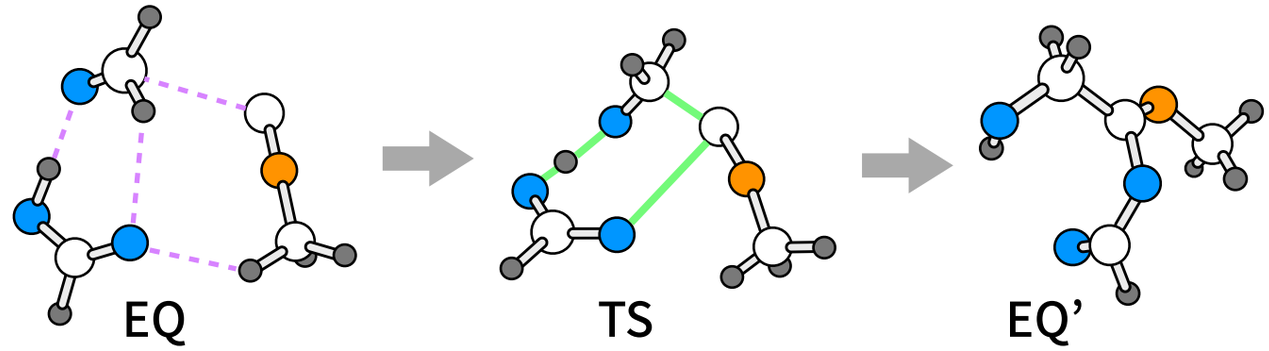
⟶ 例: MLコンペの上位解法は実用的な意味で予測精度は同じだが方法は多様

- 訓練データで定義される複雑な目的関数をがんばって厳密最小化する方法を考えても実用予測精度がヒューリスティックな方法に優越しない主原因
- 機械学習モデルの「解釈」を考えると、解釈の方法の数だけ異なる説明がありえる上に、同精度のモデルも多数あるという実用上の難しさを示唆する。



化学反応のデザインと発見 (北大WPI課題)





$$\left[\frac{-\hbar^2}{2m}\nabla^2 + V(\boldsymbol{\theta})\right]\Psi = E(\boldsymbol{\theta})\,\Psi$$

CO2の資源利用：人工光合成

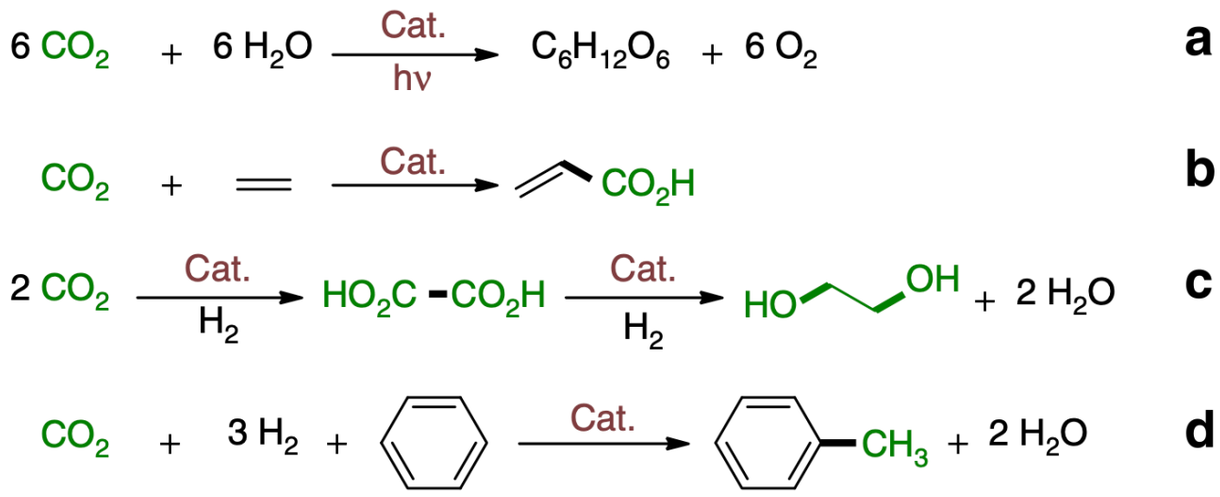
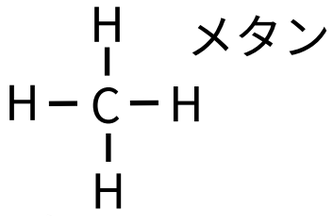


Figure 8 | Dream reactions towards artificial photosynthesis. (**a**) Photosynthesis reaction process. (**b**) Direct synthesis of acrylic acid from ethylene and CO_2. (**c**) Ethylene glycol synthesis via reductive coupling of CO_2. (**d**) Reductive methylation of benzene using H_2 and CO_2.

C1化学

原料はCが一つ (天然ガス・オイルシェール・石炭・バイオマスなど**石油以外**から得られる)

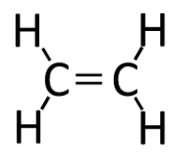
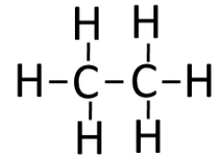


合成ガス

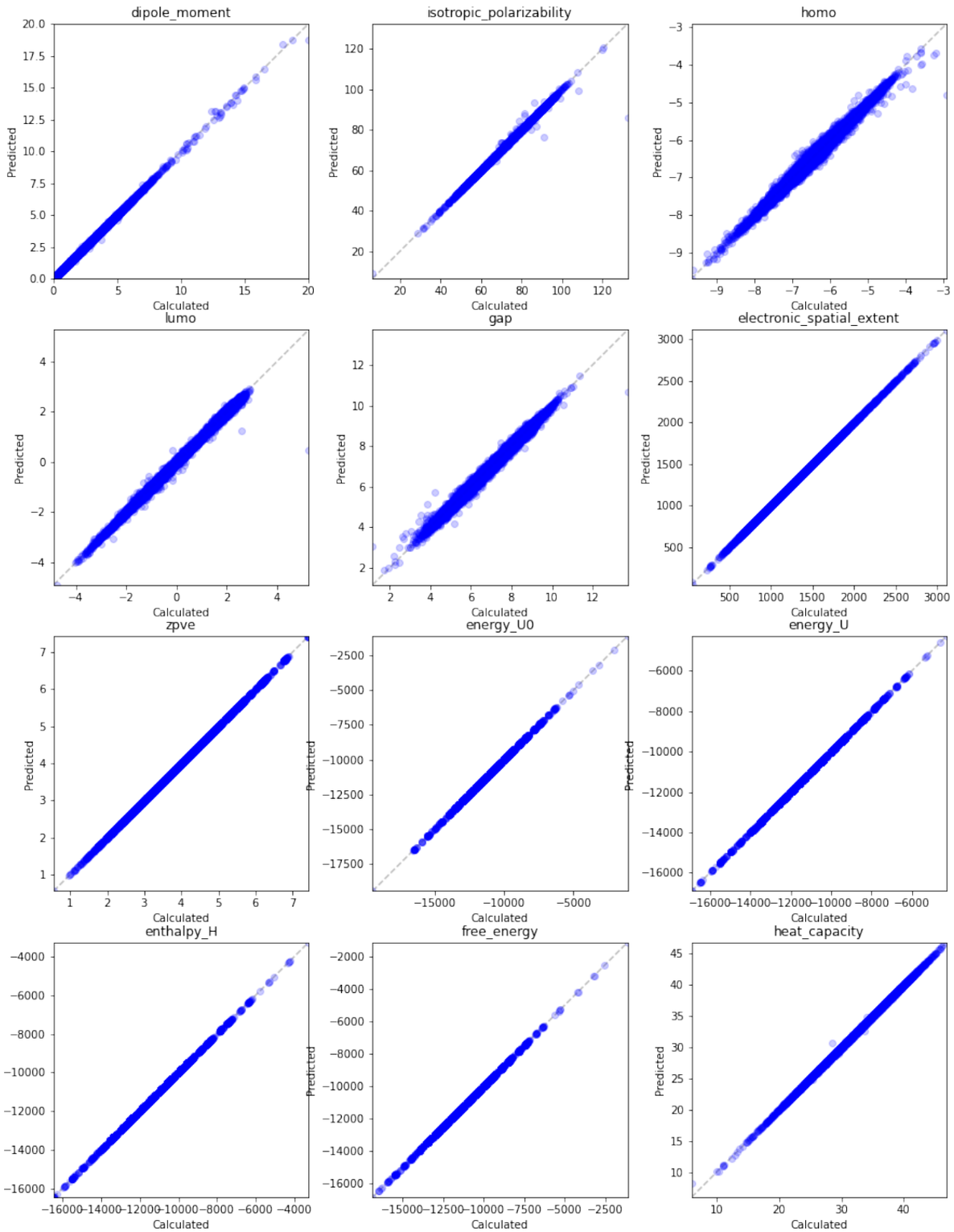
$CO + H_2$

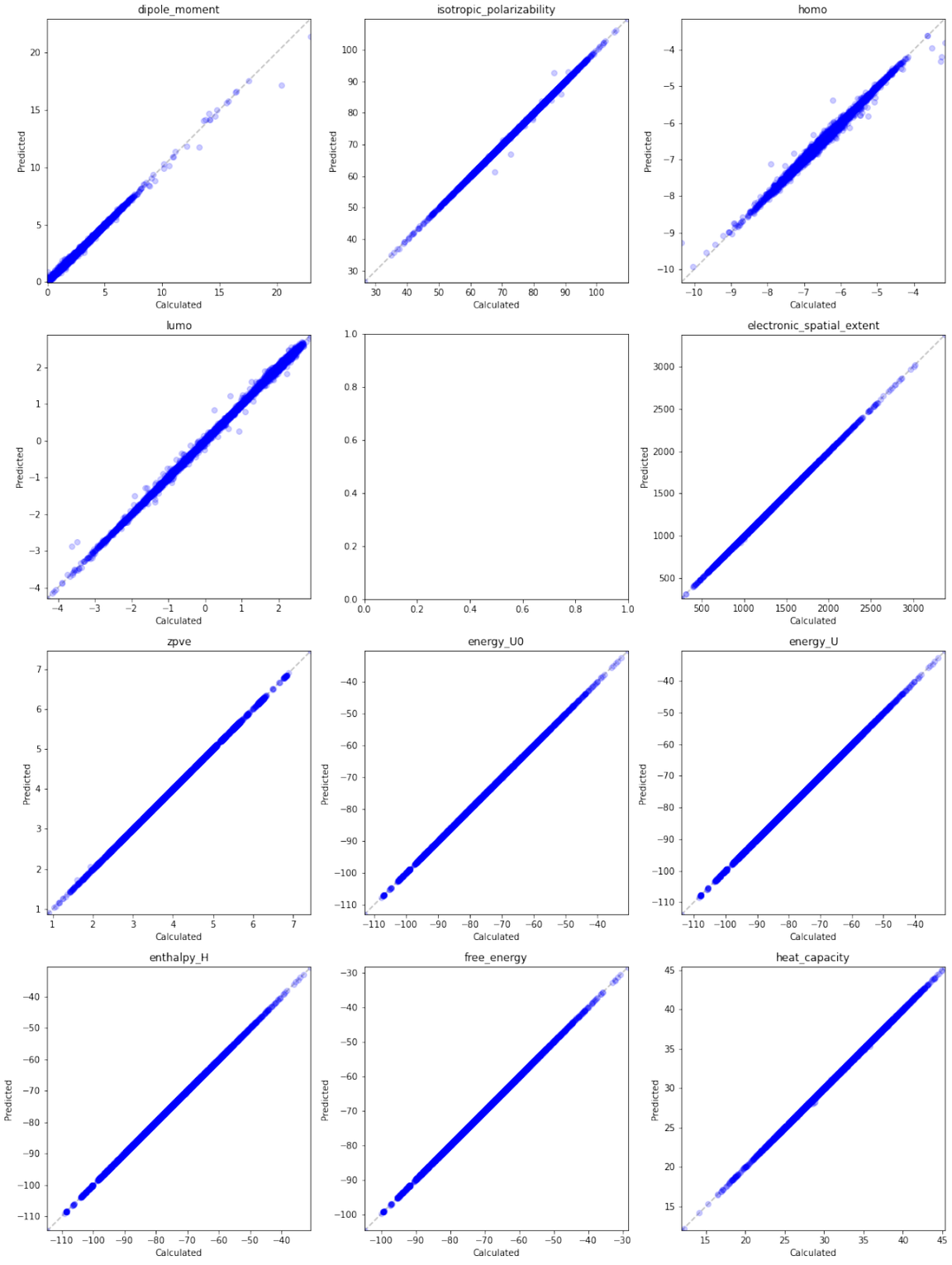
酸化カップリング

C2炭化水素類の合成

エチレン：C_2H_4	エタン：C_2H_6
	

機械学習を量子化学計算や知識ベースと融合させる?

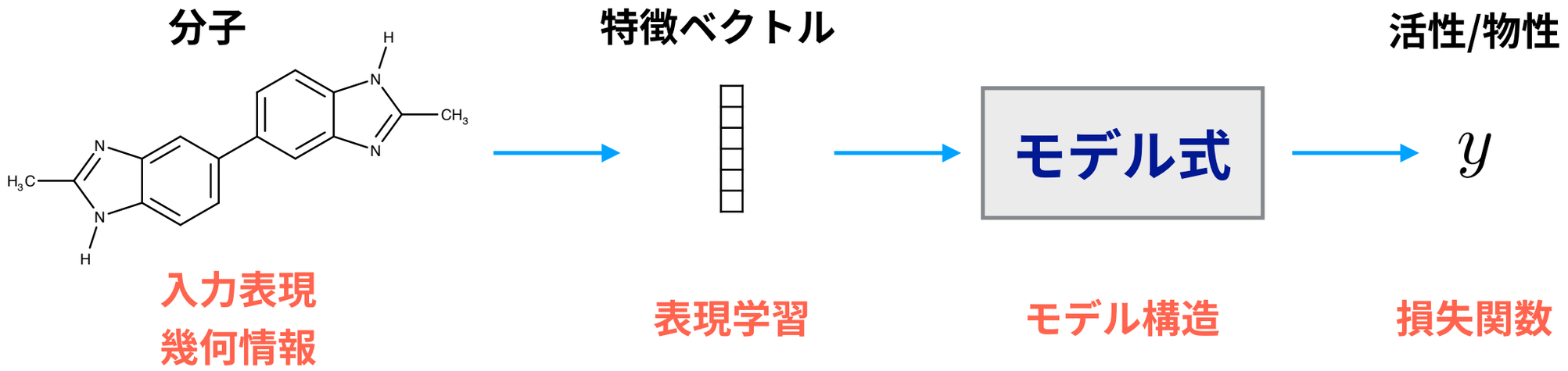




帰納バイアスと分子の組合せ的表現・幾何的表現

お題の背景にあるテーマ：機械学習から機械発見へ

主業務を踏まえた主に次のような機械学習の問題がいつも頭にあります。



- **Underspecification**の問題：自由度が高いため、モデルの近似が有限サンプルの擬似相関にトラップされた表面的でmisleadingなものに帰することが多い
- 機械学習を科学研究を支える「良い道具立て」に格上げするには**この辺**に「科学的洞察」に応じた**「意味のある制約 (帰納バイアス)」**を専門家と協働してデザインする必要がある (データとして観測できる情報は常にごく一部)
- 機械発見に向けて機械学習が使われる場面に適合した**新しい定式化**も必要

まとめ：新しい最適化の問題創出を目指して

本日のケーススタディ

① Subclass coverの研究
ユースケースを真剣に考える

② Graph miningとそれに基づく機械学習の研究
データマイニングは「発見」じゃない！？

③ SIG-FPAIと発見科学の研究
機械学習から機械発見へ：「発見」は合理化できるのか？

宇野班での私の関心課題

1. **有限の連続値ランダムデータを要素に含む離散最適化**
   Rashomon効果と解釈多様性
2. **グレイボックス最適化**
   融合研究と帰納バイアスのデザイン
3. **人間を要素として含む最適化**
   誰のためのデザイン？