核融合科学学際連携センター　第四回 フュージョン・インフォマティクス勉強会

材料科学分野のベンチマークの実例に学ぶ：核融合関連データのベンチマーク化にむけて

草場 穫

核融合科学研究所　学際連携センター　特任助教

今日の勉強会の動機

・現在、NIFS学際連携センターでLHD実験データを使った学際連携の推進を支援する活動を行っている。

→詳細:「学際連携センターにおけるLHDデータ資源を活用した学際連携推進の支援案(別資料)」参照

・まず初めに、LHD実験データの中でも早く実現できそうなデータ領域・問題設定でベンチマークセットを作成・発表を目指す。

→向井清史、草場穫、剱持尚輝、庄司主、横山雅之、鈴木優也（学生）からなる主に放射崩壊の異常検知をテーマとしたベンチマークセット作成グループを立ち上げた。

→ 他の分野のベンチマークセットについて学び、ベンチマーク作成の指針や論文化の際の参考にする。

今日の勉強会の動機

→発表者が専門とするマテリアルズインフォマティクス（MI) 、特に無機のMI分野でのベンチマークセットに関する論文を紹介する。

- MatBench: 無機MIのベンチマークとして著名 (引用数>300)
- CSPBench: 著名では無いが発表者の専門分野 (結晶構造予測)

→材料科学の細かい部分は今回の議題の中心では無いため、省いて説明する。

→論文の説明に入る前に機械学習分野におけるベンチマークセットについて説明する。

機械学習分野におけるベンチマークとは?

ベンチマークセット: 機械学習モデルの性能を客観的かつ再現可能に比較・評価するために用いられる、標準的なデータセットを指す。

→例えば、新しいモデルを提案した際に、その提案者が自分に都合の良い設定で学習、モデル評価できないようにする。
→公正にモデルを評価するため、共通ベンチマーク＋オープンデータ＋オープンソースが機械学習分野では非常に重要視されている。

一般的に、ベンチマークを発表する際は、評価指標 (MAE, Accuracy など)とベースラインモデルの予測精度を共に報告する。

→現状の予測精度を概ね把握することができ、新しいモデルの提案者がベースラインモデルの実装・実験に割く時間を節約できる。

機械学習分野におけるベンチマークとは?

・著名な例として、MNISTと imagenet がある。どちらも画像を入力としてその画像を正しく分類できるかを評価するためのものである(入力:画像, 出力: 画像のラベル)。

🔁 MNISTとILSVRCの比較まとめ (ChatGPTによるまとめ)

観点	MNIST	ILSVRC
登場時期	1998年	2010年（2012年に転機）
画像サイズ	28×28（グレースケール）	約256×256（カラー）
クラス数	10	1,000
データ数	70,000枚	約1,400,000枚
対象技術	SVM, MLP, 初期CNN	深層CNN、GPU学習、転移学習
影響	ML研究の標準入門	ディープラーニングのブレイクスルーと大衆化

*imagenetは厳密に言うと2009年に発表されたデータベースで、それを使用してILSVRCという国際コンペが2010~2017年に開催された。2012年度にこのコンペ用に作成されたベンチマークは2017年まで使用され続けた。よって、一般に imagenetと言う場合、ILSVRC 2012を指す。MNISTはNISTによって作成されたベンチマーク。

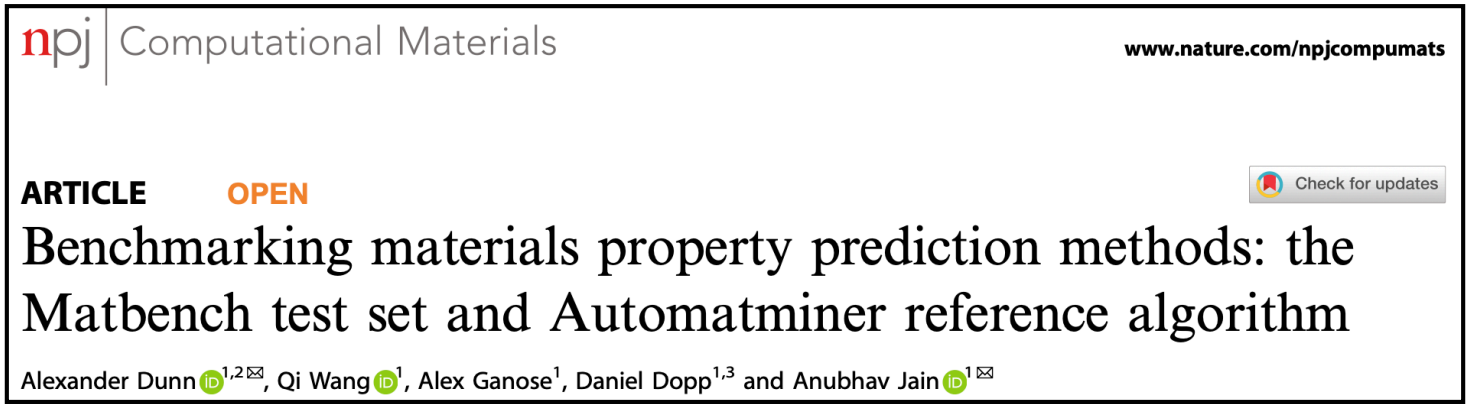
機械学習分野におけるベンチマークとは?

・MNIST: tensorflowやpytorch（機械学習用標準ライブラリ）などからダウンロードできる（→グーグルコラボ参照）。ライセンスはCC BY-SA 3。訓練セットとテストセットを提供 (双方ラベル付き、検証セットが必要な手法は訓練セットから自分で作る)。評価指標はaccuracyが用いられる。

・Imagenet: 公式ページに申請して、通れば使用できる。独自ライセンスと理解している。訓練セット、検証セット、テストセットを提供。なお、テストセットの正解ラベルはコンペが終わった現在でも未公開である（そのため、検証セットをテストセットとして使用することが行われている）。評価指標はTop-1 AccuracyとTop-5 Accuracy (モデルが出力した上位5クラスの中に正解ラベルが含まれるかどうかを判定; 1000クラスもあるため)。

→機械学習分野の発展にこれらベンチマークは極めて重要な役割を果たしており、核融合分野でもこのようなものができれば良いと考えている。

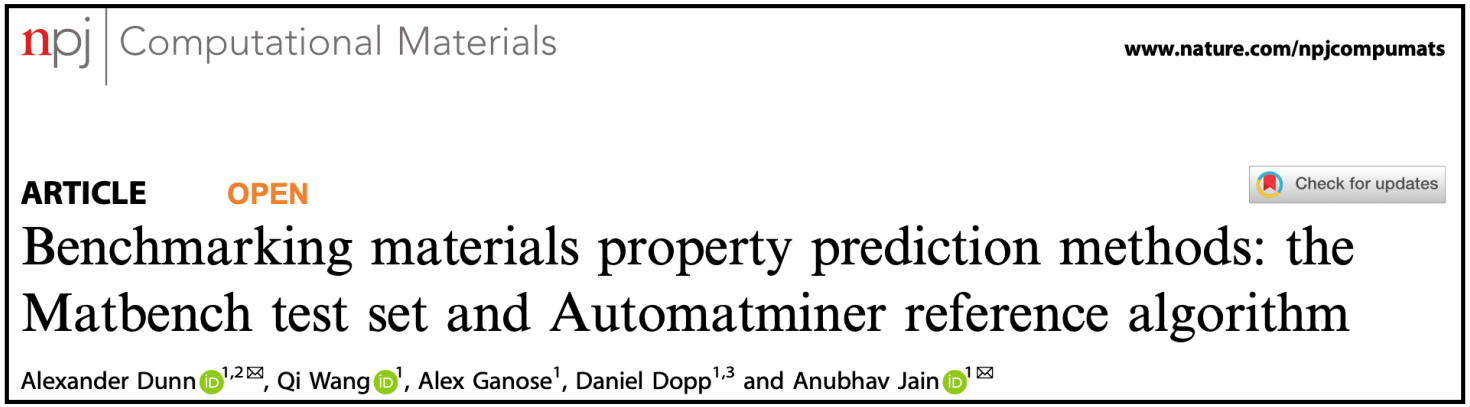
Matbenchの論文を読む



論文概要: 機械学習を用いた無機材料物性予測において、モデル性能を公正かつ標準的に評価するためのベンチマーク「Matbench」と、全自動MLパイプライン「Automatminer」を提案し、その性能を評価、報告した。

→ 既存のデータベースであるMaterials ProjectとMatminer（無機MIの標準的ライブラリ）に登録されているデータセットからベンチマークを作った（つまり、加工編集のみでデータを生成したわけではない）。
また、automatminerは基本的に既存の機械学習ライブラリを組み合わせて作ったもの（新しいモデルの開発や実装を行なったわけではない; ほぼsklearnとmatminer）。

Matbenchの論文を読む



・Matbenchは、pipからライブラリmatbenchがインストールでき、python上でベンチマークをダウンロードできる(→グーグルコラボ参照)。また、Automatminer はこれとは別にpipからautomatminerがインストールでき、利用できる (*2020年7月に更新が止まっている)。コードはMITライセンス、データはCC by。コードはgithubで公開されており、独立したホームページもある。

→ 論文上でのautomatminerの説明は長いが（さまざまなモデルを使っているため）、本論文の業績の主題はベンチマークの作成である。

Matbenchの論文を読む (ベンチマークセット)

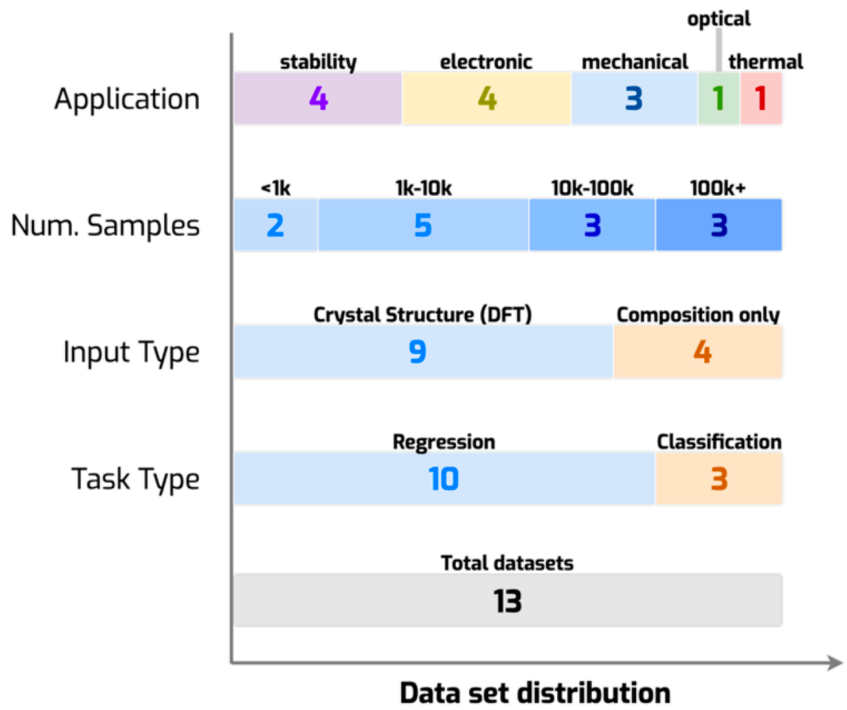


Fig. 1 Categorical dataset distribution of the 13 machine learning tasks in the Matbench test suite v0.1. Methods of categorization are listed on the left: Application describes the ML target property of the task as it relates to materials, Num. samples describes the number of samples in each task, Input Type describes the materials primitives that serve as input for each task, and Task Type designates the supervised ML task type. Numbers in the bars represent the number of tasks fitting the descriptor above it (e.g., there are 10 regression tasks).

・さまざまなデータソースから、実験・計算両方を含む13種類のタスクに対するベンチマークスイートを作成した。

Matbenchの論文を読む (ベンチマークセット)

ベンチマークスイートの詳細

Table 1. The dataset test suite.

Target property (unit)	Task type	Data source	Samples	Structure available	Method
Bulk modulus (GPa)	Regression	Materials Project[43–45]	10,987	Yes	DFT-GGA
Shear modulus (GPa)	Regression	Materials Project[43–45]	10,987	Yes	DFT-GGA
Band gap (eV)	Regression	Materials Project[43,44]	106,113	Yes	DFT-GGA
Metallicity (binary)	Classification	Materials Project[43,44]	106,113	Yes	DFT-GGA
Band gap (eV)	Regression	Zhuo et al.[46]	4604	No	Experiment
Metallicity (binary)	Classification	Zhuo et al.[46]	4921	No	Experiment
Bulk metallic glass formation (binary)	Classification	Landolt-Bornstein Handbook[28,47]	5680	No	Experiment
Refractive index (no unit)	Regression	Materials Project[43,44,48]	4764	Yes	DFPT-GGA
Formation energy (eV/atom)	Regression	Materials Project[43,44]	132,752	Yes	DFT-GGA
Formation energy of Perovskite cell (eV)	Regression	Castelli et al.[9]	18,928	Yes	DFT-GGA
Freq. at last phonon PhDOS peak (cm^{-1})	Regression	Materials Project[43,44,49]	1296	Yes	DFPT-GGA
Exfoliation energy (meV/atom)	Regression	JARVIS DFT 2D[50]	636	Yes	DFT-vDW-DF
Steel yield strength (MPa)	Regression	Citrine Informatics[51]	312	No	Experiment

The test suite contains 13 separate ML tasks spread across 10 datasets. The test suite's datasets are diversified across multiple metrics, including target property, number of samples (representing several orders of magnitude), and method for determining the target property.

Matbenchの論文を読む (ベンチマークセット)

このベンチマークセットは事前にクリーニング済みであり、その手続きをまとめている。

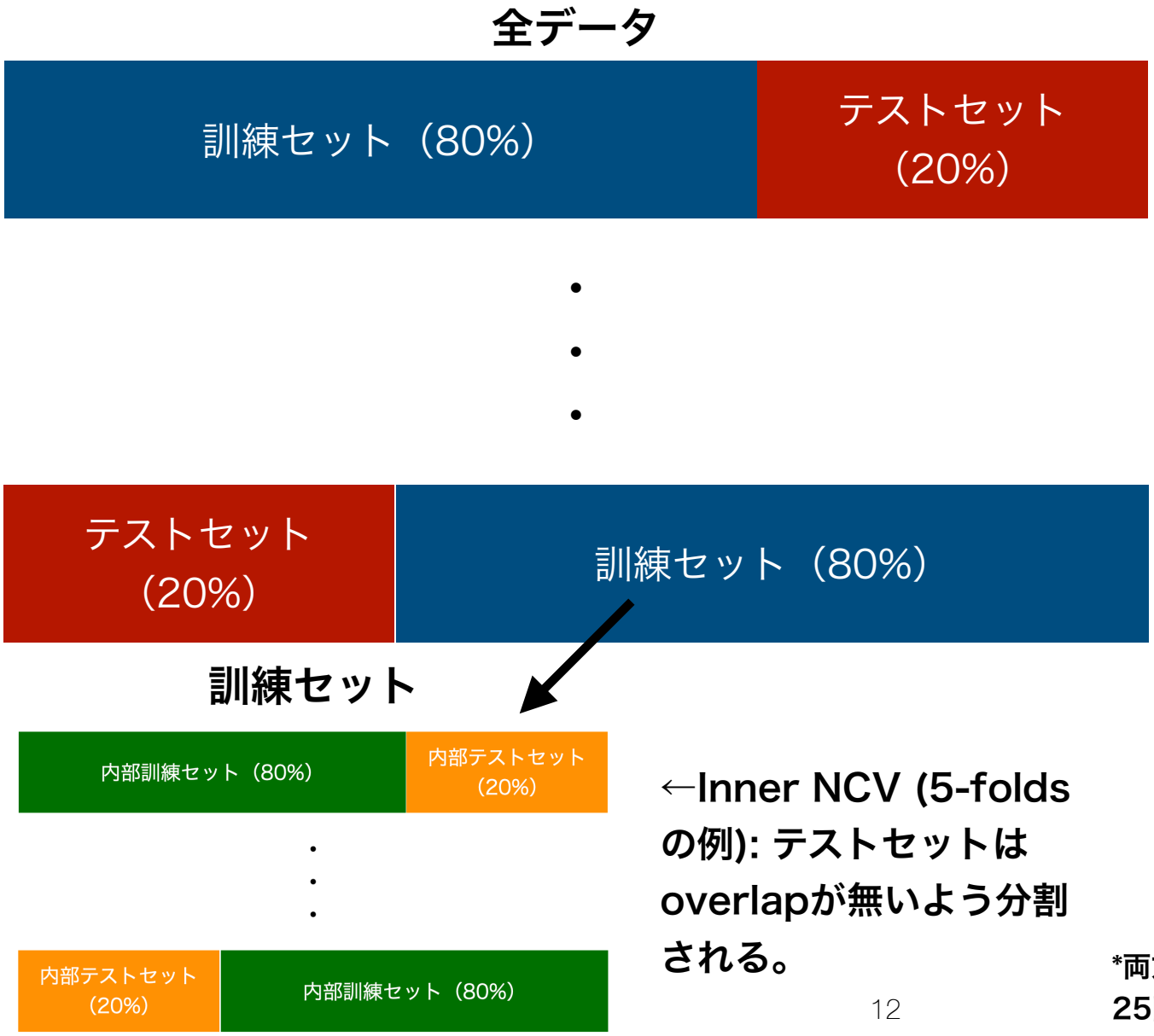
Table 2. Procedures and sources for creating datasets in Matbench v0.1.

Task name	Target property (unit)	Original source	Matminer source dataset	Additional modifications
log_kvrh	Bulk modulus (GPa)	Materials Project[43–45]	None[a]	1,2,3,6,7
log_gvrh	Shear modulus (GPa)	Materials Project[43–45]	None[a]	1,2,3,6,7
mp_gap	Band gap (eV)	Materials Project[43,44]	None[a]	1,6,7
mp_is_metal	Metallicity (binary)	Materials Project[43,44]	None[a]	1,6,7
expt_gap	Band gap (eV)	Zhuo et al.[46]	expt_gap	8,9,10
expt_is_metal	Metallicity (binary)	Zhuo et al.[46]	expt_gap	8,10,11
glass	Bulk metallic glass formation (binary)	Landolt-Bornstein Handbook[28,47]	glass_ternary_landolt	8,12
dielectric	Refractive index (no unit)	Materials Project[43,44,48]	None[a]	1,4,6,7
mp_e_form	Formation energy (eV/atom)	Materials Project[43,44]	None[a]	5,6,7
perovskites	Formation energy per Perovskite cell (eV)	Castelli et al.[9]	castelli_perovskites	7
phonons	Freq. at last phonon PhDOS peak (cm^{-1})	Materials Project[43,44,49]	phonon_dielectric_mp	1,7
jdft2d	Exfoliation energy (meV/atom)	JARVIS DFT 2D[50]	jarvis_dft_2d	7
steels	Steel yield strength (MPa)	Citrine Informatics[51]	steel_strength	8

Original Source denotes the original work that produced the raw data, which needs not be in tabular form. Matminer source datasets are tabular versions of this raw data, which can be retrieved with Matminer and may apply additional postprocessing or filtering to the original source data. More information on these datasets can be found on Matminer's dataset summary page and in the Matminer source code. Additional modifications are enumerated.
[a]Generated using the Materials Project API[44] on 4/12/2019.
(1) Remove entries having a formation energy or energy above the convex hull more than 150 meV.
(2) Remove entries having G_{Voigt}, G_{Reuss}, G_{VRH}, K_{Voigt}, K_{Reuss}, or K_{VRH} less than or equal to zero.
(3) Remove entries failing $G_{Reuss} \leq G_{VRH} \leq G_{Voigt}$ or $K_{Reuss} \leq K_{VRH} \leq K_{Voigt}$
(4) Remove entry with refractive index less than 1.
(5) Remove entries having formation energies greater than 3.0 eV. This operation removes ~1500 1-dimensional crystal structures likely resulting from mis-converged DFT structure optimizations of Half-Heuslers present in the Materials Project database as of the generation date.
(6) Remove entries containing noble gases.
(7) Remove all columns except structure and the target variable.
(8) Remove all columns except composition and the target variable.
(9) Filter according to unique compositions by ensuring no composition has conflicting metallicity.
(10) Correct erroneous $GaAs_{0.1}P_{0.9}$ composition from Zhou et al.[46] originally aggregated from Kiselyova et al.[52].
(11) Filter according to unique compositions by removing compositions with a range of reported band gap values of more than 0.1 eV. For each remaining composition, select the value closest to the mean of that composition's reported values.
(12) Filter according to unique compositions, removing compositions with any conflicting bulk metallic glass formation classifications.

Matbenchの論文を読む (ベンチマークセット)

Matbenchはデータ分割にNested Cross-Validation (NCV)を採用している。



← Outer NCV (5-foldsの例): テストセットはoverlapが無いよう分割される。

←Inner NCV (5-foldsの例): テストセットはoverlapが無いよう分割される。

Matbenchでは、Outer NCVのみ提供している。→ 訓練セット分割はユーザーの自由

*両方5-foldsなら、一つのモデルを評価するのに、25回モデルが訓練される。

Matbenchの論文を読む (ベンチマークセット)

まとめ

データ: 様々なソースからなる13タスクに対するデータセット集合

→入力は化学組成と結晶構造をそのまま提供している。データの標準化等は行わないが、明らかにおかしいデータを除去する作業は行なっている。

評価指標:　回帰→MAE, R2, RMSE(一部), MAE/MAD(一部, 後述)

分類→ROC-AUC, F1 Score(一部)

データ分割:　NCV→ Outer NCVを提供

精度比較用のベースラインモデル：

CGCNN, Megnet (2020年当時では最先端の結晶構造を入力とするNNモデル), ランダムフォレスト(一般的なML), Dummy (平均値), Automatminer

Matbenchの論文を読む (Automatminer)

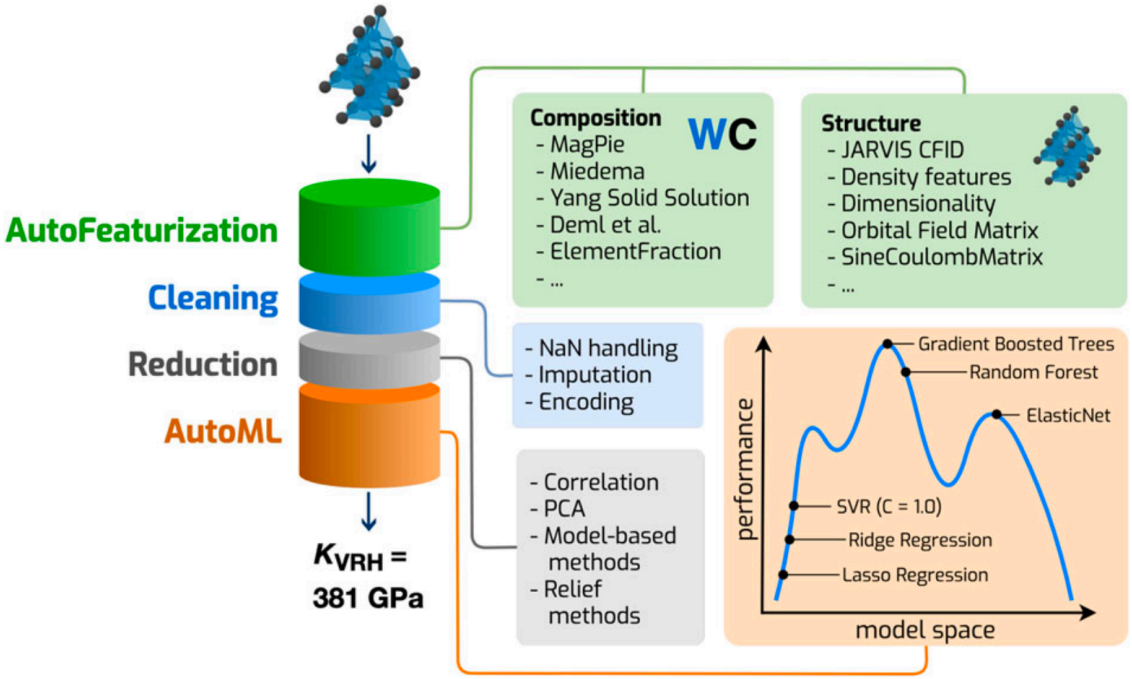


Fig. 2 The AutoML + Matminer (Automatminer) pipeline. The pipeline can be applied to composition-only datasets, structure datasets, and datasets containing electronic bandstructure information. Once fit, the pipeline accepts one or more materials primitives and returns a prediction of a materials property. During autofeaturization, the input dataset is populated with potentially relevant features using the Matminer library. Next, data cleaning and feature reduction stages prepare the feature matrices for input to an AutoML search algorithm. During training, the final stage searches ML pipelines for optimal configurations; during prediction, the best ML pipeline (according to internal validation score) is used to make predictions.

Automatminerとは、化学組成と結晶構造の記述子生成 (固定長ベクトルなどに変換すること)、記述子のクリーニング（欠損値の補完など）、次元削減、最適なモデルの選択とその訓練を全て自動で行うシステムである。

記述子の候補集合としてMatminerに登録された記述子生成手法をほぼ網羅、モデルの候補集合はsklearnにある様々モデルを網羅している。ニューラルネットワークや非線形SVMなどは無い(→基本的に伝統的なモデル)。

Matbenchの論文を読む (Automatminer)

・Automatminerは、記述子やモデル候補集合をあらかじめ用意しておき、Tree-based Pipeline Optimization Tool (TPOT) により自動で最適化する（タスクごとに最も良い記述子、記述子の前処理、モデル、そのモデルのハイパーパラメータを選択すること）。TPOTは木構造の進化的アルゴリズムである。記述子→モデル→そのモデルのハイパラのように木構造があるのでTPOTが相応しい。

・記述子やモデル候補集合+学習にかける時間の大きさを3つに分けてpresetとして提供している。以下にまとめられる。

プリセット名	特徴	使用する記述子	AutoML時間	精度	計算コスト
`debug`	最速でテスト向け	最小限（Magpieのみ）	数分	低め	非常に軽い
`express`	実用性重視の標準	中程度（高速かつ信頼性高い）	24時間以内	高	中程度
`heavy`	精度重視（研究用）	多数の高コスト記述子	48h〜	わずかに高	高

Matbenchの論文を読む (補足)

MAE/MADの説明

🔷 **MAE（Mean Absolute Error）：**

- モデルの予測誤差の平均（単位つき）
- 式：

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

🔷 **MAD（Mean Absolute Deviation from Mean）：**

- **データそのものの散らばり具合**（平均からのズレの平均）を表す
- 式：

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \bar{y}|$$

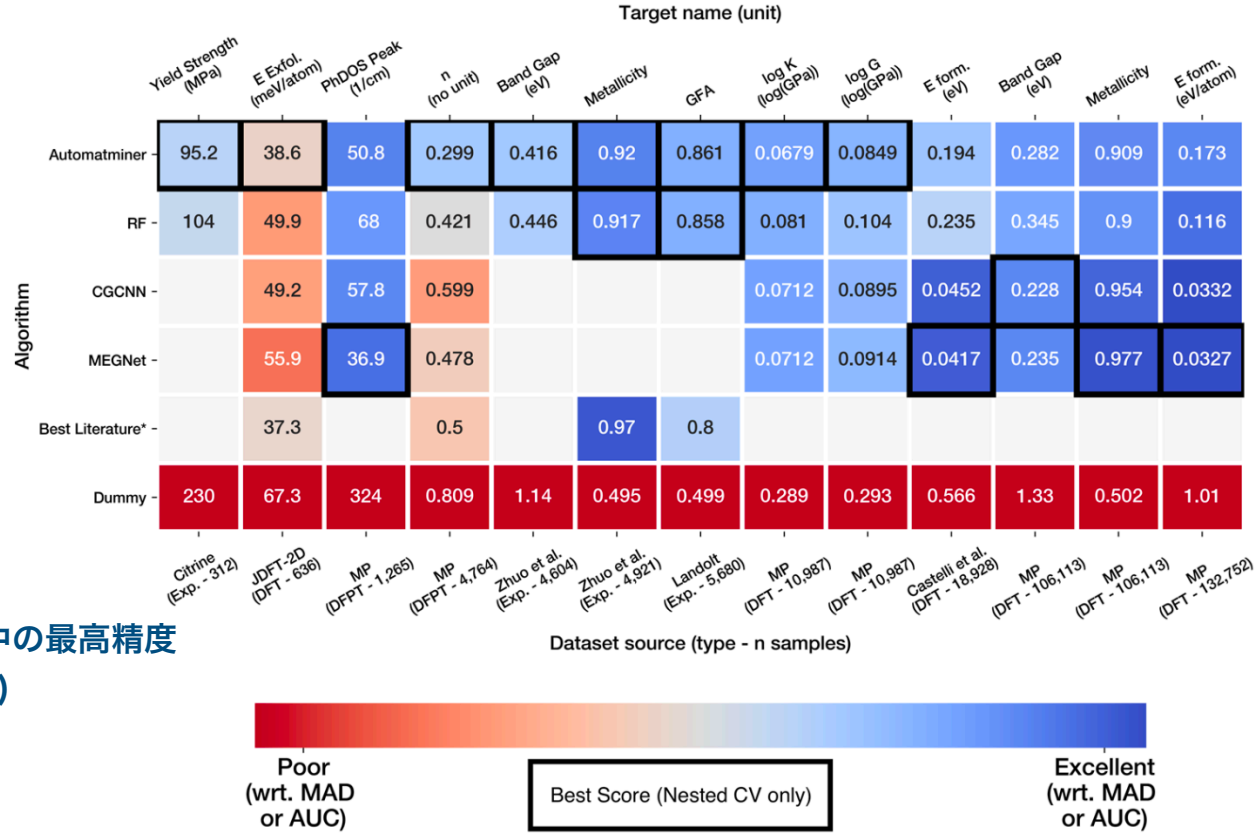
ここで $\bar{y}$ はデータの平均値

🔢 **MAE / MAD の意味**

- **値が1.0の場合**：モデルは「常に平均を予測する」のと同じ性能
- **値が1未満**：モデルは「平均よりも良い予測」をしている
- **値が0に近い**：モデルは「理想的に正確な予測」をしている
- **値が1を超える**：モデルは「平均予測よりも悪い」

Dummy (平均値予測)では必ずこの値が1になる。

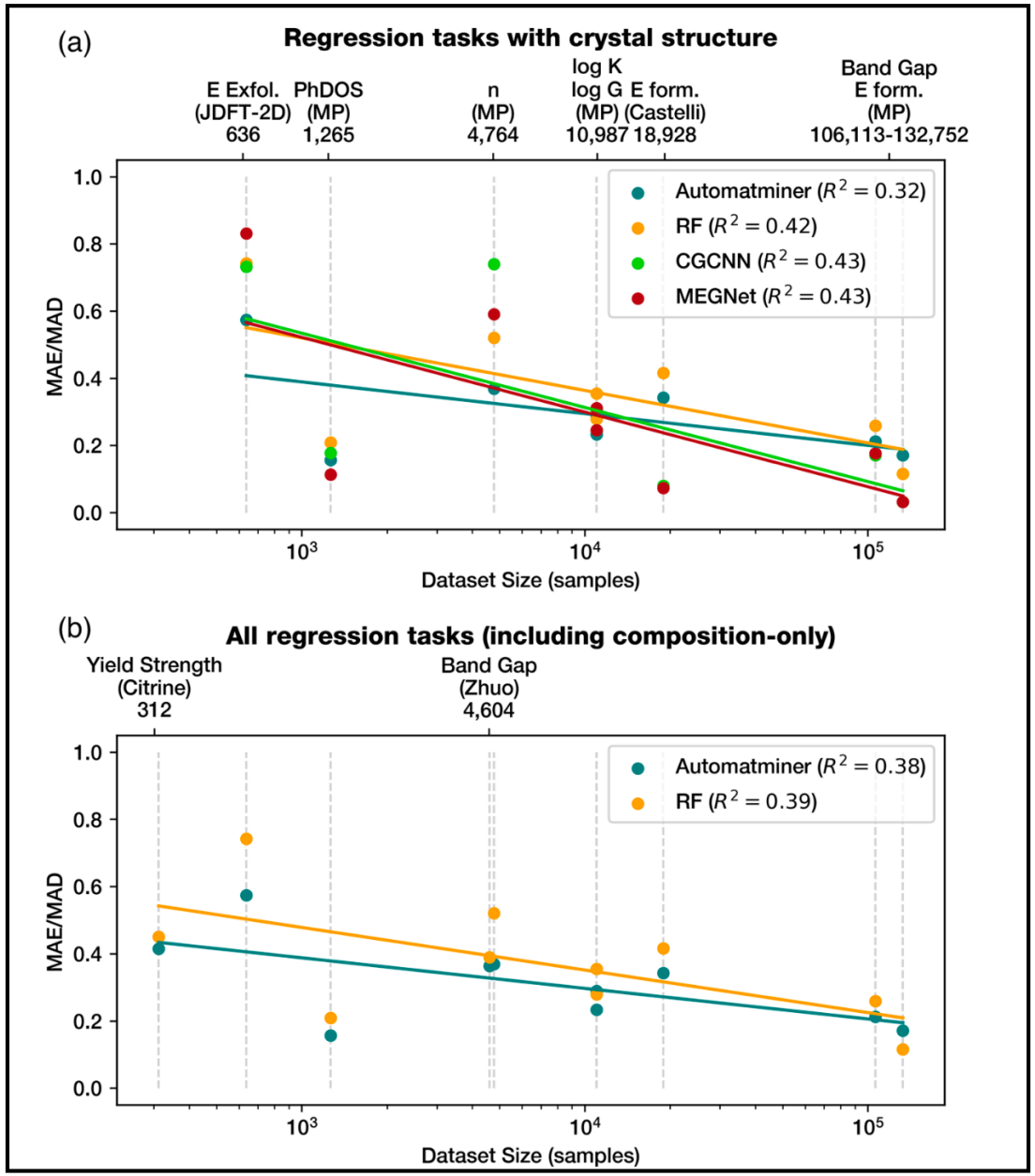
Matbenchの論文を読む (予測結果)



* Best literatureは文献中の最高精度
(データが異なるので参考用)

Fig. 3 Comparison of machine learning algorithm accuracies on the Matbench v0.1 test suite. See Table 1 for more details of the test sets. Numbers on each square represent either the mean average error (regression) or mean ROC-AUC (classification) of a five-fold nested cross validation (NCV), except for Best Literature scores. Best Literature scores were taken from published literature models[33,46,53] evaluated on similar tasks or datasets, often subsets of those in Matbench, and do not use NCV. Colors represent prediction quality (analogous to relative error) with respect to either the dataset target mean average deviation (MAD) for regression or the high/low limits of ROC-AUC (0.5 is equivalent to random, 1.0 is best) for classification; blue and red represent high and low prediction qualities, respectively, with respect to these baselines. Accordingly, red-hued columns indicate more difficult ML tasks where no algorithm has high predictive accuracy when compared to predicting the mean. Red-hued rows therefore indicate poorly performing algorithms across multiple tasks. The best score for each task is outlined with a black box (The Best Literature scores are excluded because they do not use the same testing protocol). To account for variance from choice of NCV split, multiple scores may be outlined if within 1% of the true best score. A comparison with a pure Random Forest (RF) model using Magpie[28] and SineCoulombMatrix[29] features is provided for reference. Dummy predictor results are also shown for each task. All Automatminer, CGCNN, MEGNet, and RF results were generated using the same NCV test procedure on identical train/test folds; all featurizer (descriptor) fitting, hyperparameter optimization, internal validation, and model selection were done on the training set only. A full breakdown of all error estimation procedures can be found in Methods.

Matbenchの論文を読む (予測結果)



・タスクが異なっても、データが増えるごとに予測精度が向上する傾向にあることが示された。

・異なるタスク間で比較が可能なようにMAE/MADを指標としている（前述）。

・Automatminerはデータが少ない時に優秀だが、多くなるとNNベースの手法に負ける傾向がある。

CSPbenchの論文を読む

CSPBENCH: A BENCHMARK AND CRITICAL EVALUATION OF CRYSTAL STRUCTURE PREDICTION

論文概要: 結晶構造予測（CSP）は新材料探索に不可欠だが、これまで統一されたベンチマークデータセットや性能評価指標が不足していた。この論文では、180種類の結晶構造（CSP180）を用いたベンチマークセットと、複数の定量評価指標を用いて、13種類の最先端CSPアルゴリズムを比較した。

→ 既存のデータベースであるMaterials Project から、空間群や結晶系、化学組成や予測難易度が多様になるように180個の結晶構造が選択された (手動)。この論文では、テストセットのみが提供されている。結晶構造予測の評価指標はこの分野で統一的なものがないので、様々な評価指標のセットを一括して使っている（一括して算出するコードは同じ研究グループが先行研究で発表済み）。CSPBench は GitHub 上で CSV としてデータを公開している。

→ちなみに、結晶構造予測とは与えられた化学組成の熱力学的に安定な構造を予測する問題である。

CSPbenchの論文を読む (ベンチマーク)

✅ データセットの選定方法

論文より：

> *"We selected a total of 180 crystal structures, named CSP180, from the Materials Project database... ensuring a diverse and representative sample."*

- 出典：Materials Project
- 対象：**binary（2元）・ternary（3元）・quaternary（4元）** の化合物、各60個ずつ
- 手動選定時に考慮された要素：
  - 原子数・元素数の多様性
  - 結晶系の分布（立方、六方、斜方など）
  - 空間群（space group）のバリエーション（225, 139, 216, 221, 194など）
  - 難しさの多様性（予測困難な構造も含む）
  - 既存データベースにおける類似構造の有無

つまり、**自動ランダム抽出ではなく、予測性能の評価に適したバランスの取れた代表構造群を著者が慎重に設計・選定した**ことが明言されています。

CSPbenchの論文を読む (ベンチマーク)

🧩 難易度の決定基準

論文より：

> *"The criteria for difficulty classification include factors such as space group classification, template-based categorization, and the prototype ratios defining the crystal structures."*

難易度分類の主な要素：

分類要素	説明
空間群	複雑な対称性（例：低対称の群）は予測が難しい傾向がある
テンプレートベース分類	類似構造が存在するかどうか（存在しない＝難しい）
原子数・格子構造の比	prototype 比（例：ThB_{12}など極端な組成比は難易度が高い）

- 例：`binary_easy` では DyCu や GaCo のような単純で高対称な立方構造
- `binary_hard` では $BePd_2$ や Sn_8Pd_2 のように、対称性が低く予測困難な構造

これらの基準で、**easy / medium / hard** に難易度ラベルが付けられています（特にbinary系では明確に分類表あり）。

CSPbenchの論文を読む (ベンチマーク)

選択された構造の例:

Table 2: **Details of the binary_easy, binary_medium and binary_hard data for benchmark crystals used in this work. See Table S1 for the whole set.**

Material id	Pretty formula	Space group	Crystal system	Category
mp-2334	DyCu	221	Cubic	binary_easy
mp-2226	DyPd	221	Cubic	binary_easy
mp-1121	GaCo	221	Cubic	binary_easy
mp-2735	PaO	225	Cubic	binary_easy
mp-1169	ScCu	221	Cubic	binary_easy
mp-30746	YIr	221	Cubic	binary_easy
mp-24658	SmH_2	225	Cubic	binary_easy
mp-20225	$CePb_3$	221	Cubic	binary_easy
mp-788	Co_2Te_2	194	Hexagonal	binary_easy
mp-20176	$DyPb_3$	221	Cubic	binary_easy
mp-1231	$Cr6Ga_2$	223	Cubic	binary_easy
mp-12570	ThB_{12}	225	Cubic	binary_easy
mp-20132	InHg	166	Trigonal	binary_medium
mp-2209	$CeGa_2$	191	Hexagonal	binary_medium
mp-30497	$TbCd_2$	191	Hexagonal	binary_medium
mp-30725	YHg_2	191	Hexagonal	binary_medium
mp-2731	$TiGa_3$	139	Tetragonal	binary_medium
mp-2510	ZrHg	123	Tetragonal	binary_medium
mp-2740	$ErCo_5$	191	Hexagonal	binary_medium
mp-570875	Ga_4Os_2	70	Orthorhombic	binary_medium
mp-861	$Hf4Ni_2$	140	Tetragonal	binary_medium
mp-1566	$SmFe_5$	191	Hexagonal	binary_medium
mp-2387	Th_4Zn_2	140	Tetragonal	binary_medium
mp-1607	$YbCu_5$	191	Hexagonal	binary_medium
mp-13452	$BePd_2$	139	Tetragonal	binary_hard
mp-11359	Ga_2Cu	123	Tetragonal	binary_hard
mp-1995	PrC_2	139	Tetragonal	binary_hard
mp-30501	Ti_2Cd	139	Tetragonal	binary_hard

CSPbenchの論文を読む (比較手法)

比較用の結晶構造予測手法一覧

Table 1: A summary of the main CSP softwares. MLP: machine learning potentials; MOGA: multi-objective genetic algorithm;

Algorithm	Year	Category	Open-source	URL link	Program Lang
USPEX [4]	2006	De novo (DFT)	No	link	Matlab
CALYPSO [19]	2010	De novo (DFT)	No	link	Python
ParetoCSP [28]	2024	MOGA+MLP	Yes	link	Python
GNOA [9]	2022	BO/PSO + MLP	Yes	link	Python
TCSP [17]	2022	Template	Yes	link	Python
CSPML [29]	2022	Template	Yes	link	Python
GATor [30]	2018	GA + FHI potential	Yes	link	Python
AIRSS [31, 32]	2011	Random + DFT or pair Potential	Yes	link	Fortran
GOFEE [33]	2020	ActiveLearning + Gaussian Pot.	Yes	link	Python
AGOX [13]	2022	Search + Gaussian Potential	Yes	link	Python
GASP [34]	2007	GA + DFT	Yes	link	Java
M3GNet [35]	2022	Relax with MLP	Yes	link	Python
ASLA [36]	2020	NN + RL	No	link	N/A
CrySPY [37]	2023	GA/BO + DFT	Yes	link	Python
XtalOpt [38]	2011	GA + DFT	Yes	link	C++
AlphaCrystal [39, 40]	2023	GA + DL	Yes	link	Python

CSPbenchの論文を読む (評価指標)

結晶構造予測の評価指標一覧

2.4 Evaluation metrics

Evaluation metrics are essential in materials science research as they quantitatively assess the performance and effectiveness of different materials. Currently, numerous evaluation metrics exist in molecular research, such as RDKit [62] and MOSES [63]. However, in the field of materials informatics, there is no unified standard for evaluating new structures. Recently, we introduced a set of distance metrics for CSP performance comparisons in benchmark studies [64], including M3GNet energy distance, minimal rmse distance, minimal mae distance, rms distance, rms anonymous distance, Sinkhorn distance, Chamfer distance, Hausdorff distance, superpose rmsd distance, edit graph distance, Fingerprint distance, to standardize the training and comparison of material structure generation models. For test structures in the polymorph category, we employ a detailed evaluation approach. We compare the predicted structures with multiple ground truth structures, each representing different polymorphs. As each sample corresponds to multiple ground truth polymorphs, this results in several evaluation metrics for each sample. To identify the most accurate predictions, we select the evaluation metrics associated with the ground truth structure that has the minimum M3GNet energy distance. This method ensures that the selected metrics reflect the closest match to the predicted structure, providing a reliable measure of prediction accuracy. The distance metrics are shown below. Table 3 shows selected distance scores for various test samples generated by the AGOX-pt algorithm.

- Wyckoff position fraction coordinate RMSE distance
- Wyckoff Minimal MAE distance
- M3GNet Energy distance
- Pymatgen RMS distance
- Sinkhorn distance
- Chamfer distance
- Hausdorff distance
- Superpose RMS distance
- CrystalNN Fingerprint distance
- Edit Graph distance
- XRD distance
- OFM distance

***与えられた化学組成の予測構造と真の構造の誤差によって、予測精度を比較するが、主に結晶構造の数値表現に統一的な手法が存在しない（どれも一長一短）ことに起因して、様々な評価指標がある。**

＋Pymatgen (無機MIの標準ライブラリ, matminerよりコアな部分を担当) のStructureMatcherと空間群が合うかどうか

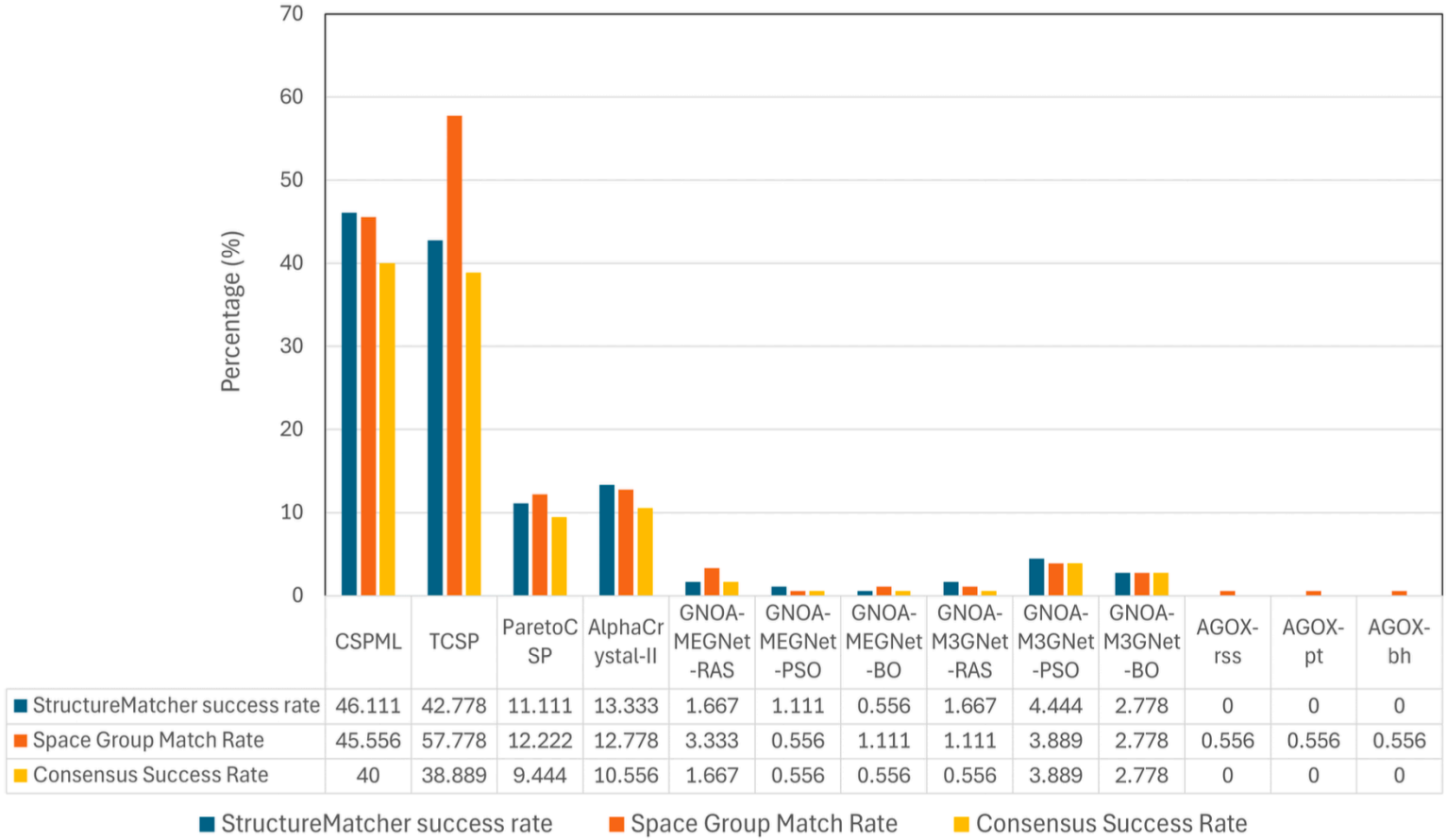
CSPbenchの論文を読む (評価指標)

結晶構造予測の評価指標に基づいたランキングスコアを算出

Ranking scores of algorithms: To evaluate how different performance metrics reflect the actual closeness of the predicted structures to the ground truth structure, we employed quantitative distance matrices of CSP [64] to assess the quality of all structures generated by the algorithms. We adopted a ranking scheme to evaluate candidate CSP algorithms based on the quality of their predicted structure against the ground truth structure. For each test structure, all algorithms are first ranked based on the quality of their predicted structures, i.e., their distances to the ground truth structure. Ranking scores on a 0-100 scale are assigned to the algorithms using a standardized scoring method to ensure fairness in ranking. The ranking scheme is illustrated as follows: for example, if there are five algorithms for comparison, five evenly distributed scores ranging from 100 to 0 are assigned to the five algorithms sorted by their performance from the highest to the lowest. Specifically, the algorithm in the first place receives a score of 100 (reflecting the smallest distance), and the second-placed algorithm earns a score of 75, followed by 50 for the third place, 25 for the fourth place, and 0 for the fifth place. In cases where multiple algorithms produced structures with identical quality/distances, they were assigned the same rank, and scores were averaged according to their rankings. For instance, if the first and second place algorithms tie in the quality of their predicted structures, their scores are set as the average of 100 and 75 [(100 + 75)/2]. Similarly, if all five algorithms have the same performance, their scores are set as the average of the five scores [(100 + 75 + 50 + 25 + 0)/5]. Figure 5 shows the ranking scores based on the overall average distances for each algorithm.

XX distanceの平均で評価すると、ものすごく外れた結晶構造予測結果数例の影響を受けすぎてしまう。これを防ぐためにランキングシステムが導入されている。
基本的にこれか、予測の成功率 (StructureMatcherの結果と空間群の正解率)で評価。

CSPbenchの論文を読む (予測結果の例)



* consensus SRは他二つを同時に満たす確率

Figure 2: Comparison of structure prediction performance in terms of the success rate (%) measured as the percentage of the predicted structures (out of 180 test samples) that match the ground truth structures according to Pymatgen StructureMatcher and have identical space groups.

→ちなみにCSPMLは私が開発した手法

CSPbenchの論文を読む (予測結果の例)

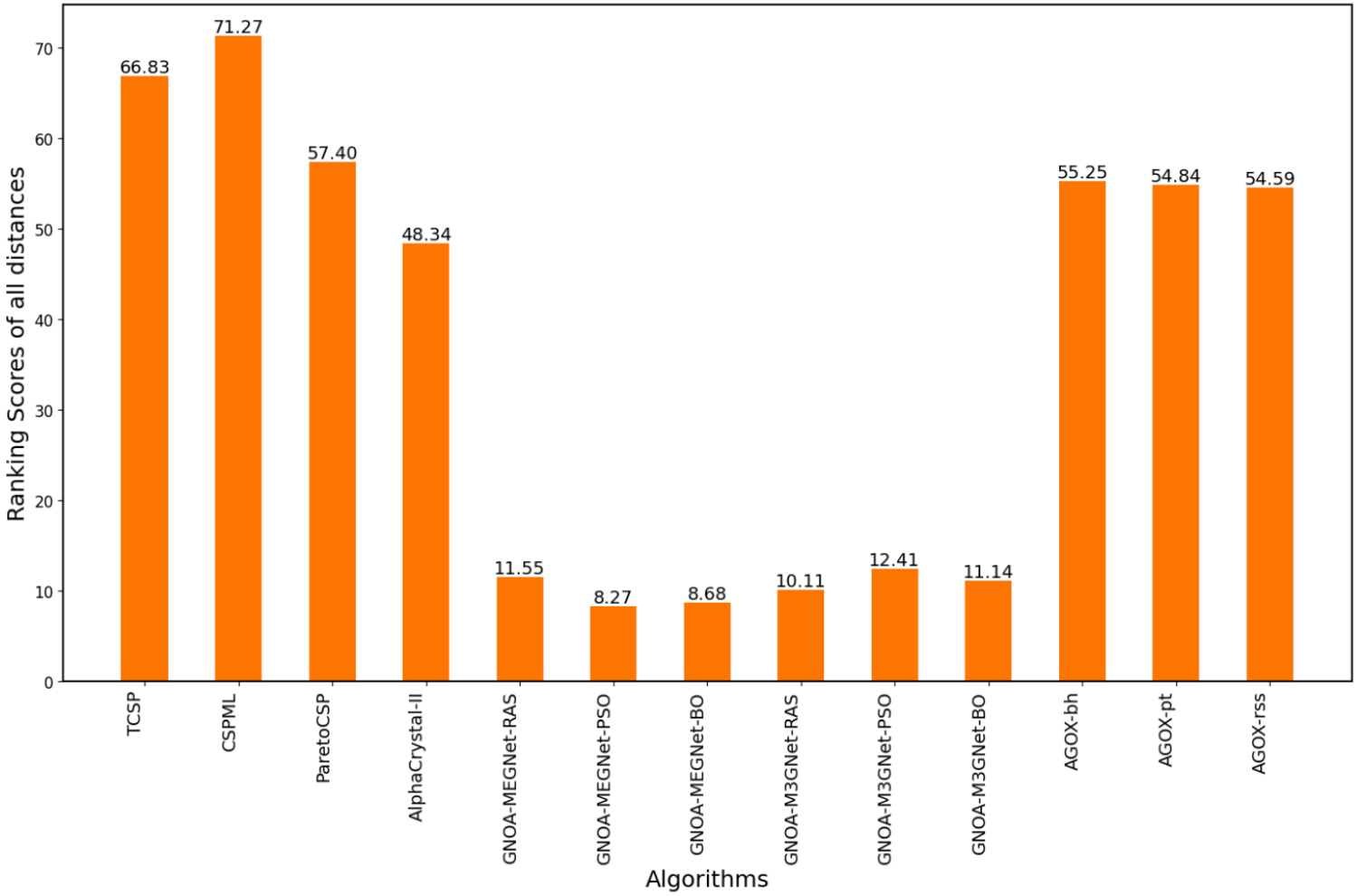


Figure 5: Comparison of CSP algorithms by their ranking scores based on the average of all 12 distance metrics of the predicted structures against the ground truth structures.

放射崩壊のベンチマークセット作成に向けて学んだことメモ

・データは標準化・正規化せず生のままで出すべきだが、物理的におかしなデータは除去するべき。記述子生成などは別のシステムに任せる(matbenchとautomatminerの分離)。

・データは訓練とテストに分割して提供する (CVにするか、単純分割か)。

・評価指標と比較用のベースラインモデルを選定しておく。

・ライセンスはコード部分はMIT, データ部分はCC by が望ましい。

・ベンチマークセットは、pythonライブラリなどで提供するか、CSV等をGithub, figshareなどに置いておく。多様なものならば前者の方が良い。

・論文化する場合は、ベースラインモデルの実装と予測結果の簡単な解析をするべき (+コードの公開)。