



Materials property prediction from limited and multi-fidelity datasets



Gian-Marco Rignanese^{1,2} 1 Université catholique de Louvain, Louvain-la-Neuve (Belgium) 2 Northwestern Polytechnical University, Xi'an (China)

4th IKZ-FAIRmat Winterschool "Machine Learning in Materials Science and Crystal Growth" Berlin (Germany), 23-25 January 2023

Many materials DB have become available online



Each of these databases has

its own user base and specific API



The queries can be very different for asking for the same thing...



http://www.crystallography.net/cod/result.php?formula=O2%20Si



http://www.materialsproject.org/rest/v2/materials/SiO2/vasp/structure? API_KEY=YOUR_API_KEY



http://aflowlib.duke.edu/search/API/?species(Si,O),nspecies(2)

... and the responses have very different formats



JSON Raw Data He	aders
Save Copy Collapse All	
0:	
file:	"1010921"
a:	"7.16"
siga:	null
b:	"7.16"
sigb:	null
c:	"7.16"
sigc:	null
alpha:	"90"
sigalpha:	null
beta:	"90"
sigbeta:	null
gamma:	"90"
siggamma:	null
vol:	"367.1"
sigvol:	null
celltemp:	null
sigcelltemp:	null
diffrtemp:	null
sigdiffrtemp:	null
cellpressure:	null
sigcellpressure:	null
diffrpressure:	null
sigdiffrpressure:	null
thermalhist:	null
pressurehist:	null
compoundsource:	null
nel:	"2"
sg:	"P 21 3"



JSON Raw Data	Headers
JSON Copy Collapse A	u .
<pre>response:</pre>	
- 0:	
<pre>material_id:</pre>	"mp-600033"
<pre>▼ structure:</pre>	
@module:	"pymatgen.core.structure"
@class:	"Structure"
charge:	null
<pre>valattice:</pre>	
<pre>> matrix:</pre>	[]
a:	9.01708962
b:	9.01708962
c:	9.01708962
alpha:	90
beta:	90
gamma:	90
volume:	733.160668139128
▼ sites:	
₹0:	
<pre>species</pre>	
₹0:	
el	ement: "Si"
00	:cu: 1
▼abc:	
0:	0.25
1:	0.5
2:	0
▼ xyz:	
0:	2.254272405
	4 50054401



Headers

JSON

Raw Data

Save Copy Col	lapse All Expand All
- 0:	
compound:	"016Si8"
auid:	"aflow:590a543e005fcdd0"
▼aurl:	"aflowlib.duke.edu:AFLOWDATA/ICSD_WEB/0
species:	"0,Si"
nspecies:	"2"
v 1:	
compound:	"08Si4"
auid:	"aflow:fe6cb4a748ca8f04"
▼aurl:	"aflowlib.duke.edu:AFLOWDATA/ICSD_WEB/0
species:	"0,Si"
nspecies:	"2"
₹2:	
compound:	"048Si24"
auid:	"aflow:a461b6af4b750e1c"
▼aurl:	"aflowlib.duke.edu:AFLOWDATA/ICSD_WEB/0
species:	"0,Si"
nspecies:	"2" ·
▼3:	
compound:	"024Si12"
auid:	"aflow:3dd0d3cf29cc4b04"
⇒aurl:	"aflowlib.duke.edu:AFLOWDATA/ICSD_WEB/0
species:	"0,Si"
nspecies:	"2"
₹4:	
compound:	"016Si8"
auid:	"aflow:390c258fcaa1a88b"
▼aurl:	"aflowlib.duke.edu:AFLOWDATA/ICSD_WEB/0
species:	"0,Si"

A common API has been defined



- The initial release was developed by the participants of the workshops "Open Databases Integration for Materials Design" held at:
 - the Lorentz Center (October 2016)







Check for updates

scientific data

OPEN OPTIMADE, an API for exchanging ARTICLE materials data

Casper W. Andersen et al.#

The Open Databases Integration for Materials Design (OPTIMADE) consortium has designed a universal application programming interface (API) to make materials databases accessible and interoperable. We outline the first stable release of the specification, v1.0, which is already supported by many leading databases and several software packages. We illustrate the advantages of the OPTIMADE API through worked examples on each of the public materials databases that support the full API specification.

*A full list of authors and their affiliations appears at the end of the paper.

Thanks to OPTIMADE, it is possible to search many materials DBs with the same query...



Predicting different properties requires very different computing time

4.7 million properties; 57 million CPU hours; 730,000 calculations...



How can we predict the phase stability of polymorphs at different temperatures?

• At T=0K: for exemple, the Cu-O system



• At T>0K, the vibrational entropy needs to be taken into account. This can be done by DFPT but it is very demanding.

An automatic workflow was developed



The vibrational properties were calculated for 1521 semiconductors

- The dataset includes:
 - phonon band structure
 - LO-TO splitting
 - phonon DOS
 - Born effective charges
 - dielectric tensor
 - derived quantities:
 - ΔF , $\Delta E_{\rm ph}$, C_v and S
- The dataset is openly available!



G. Petretto, S. Dwaraknath, H.P.C. Miranda, D. Winston, M. Giantomassi, M.J. van Setten, X. Gonze, K.A. Persson, G. Hautier, and G.-M. Rignanese, Sci. Data **5**, 180065 (2018).



The vibrational properties are available on the Materials Project website



... but only for those 1521 semiconductors

This is where the power of machine learning becomes very handy



One of the most crucial steps is to choose the features representing the atomic structures

• The aim is to transform the structure into a descriptive vector



• Many methods have been proposed:

Matminer [Ward *et al.*, Comput. Mater. Sci. **152**, 60 (2018)]
 (<u>https://hackingmaterials.lbl.gov/matminer</u>)

Megnet [Chen *et al.*, Chem. Mater. **31**, 3564 (2019)]
 (<u>https://github.com/materialsvirtuallab/megnet</u>)

Matminer

Composition

- Element fractions
- Mean atomic mass
- Mean atomic row
- Stoechiometry

. . .

• Electronegativity

Structure

- Space group number
- Crystal system
- Radial distribution function
- Bond fractions
- Coulomb matrix
- ...

Site

- Radial environment
- Motif matching
- ...







Megnet

• Graph networks allow for the representation the attributes of atoms





• It was originally trained for the formation energy and the band gap



The other one is the model whose predictive power depends on the amount of data available



Amount of data available

Computationally demanding material properties are precisely those with little available data



In terms of quantity, is it really Big Data?

The sorites paradox

If a heap is reduced by a single grain at a time, the question is: at what exact point does it cease to be considered a heap?

Material Optimal Descriptor Network (MODNet)

- <u>Concept:</u> feedforward neural network with an optimal set of descriptors.
- <u>Idea:</u> Feature selection by relevance-redundancy algorithm
 - Prior physical knowledge and constraints are taken into account by adopting physically-meaningful features.
 - This reduces the optimization space without relying on a massive amount of data.
- Bonus: Novel architecture that learns on multiple properties

To be relevant, the selected features should present some kind of interrelation with the target property



Pearson correlation coefficient is a measure of the interrelation between two variables



Pearson correlation coefficient

presents, however, a series of limitations



R=0

In MODNet, feature selection is based on the Normalized Mutual Information (NMI)

- The mutual information (MI) of two random variables is a measure of the mutual dependence between the two variables.
 - It quantifies the "amount of information" (entropy) obtained about one random variable through observing the other random variable.



[P.-P. De Breuck et al., npj Computational Materials 7, 83(2021)]

The feature *f* having the highest NMI with the target variable *y* will be chosen the first one

• This provides some understanding of the underlying physics.

Indeed, it pinpoints the most important and complementary variables.

1	Mean average bond length (AverageBondLength)
2	Mean row in the periodic table (ElementProperty)
3	GRDF centered at 3.0 Å and width of $1.0 \text{ Å}(\text{GeneralizedRDF})$
4	HOMO element (AtomicOrbitals)
5	Crystal system (GlobalSymmetryFeatures)
6	Mean minimum vornoi volume (VoronoiFingerprint)
7	Mean square co-planar CN 4 (OPSiteFingerprint)
8	Mean A:2 (i.e angular) environment (ChemEnvSiteFingerprint)
9	Mean spherical harmonic l=7 (BondOrientationParameter)
10	Standard deviation of the Voronoi volume(VoronoiFingerprint)

The feature *f* having the highest NMI with the target variable *y* will be chosen the first one

- This provides some understanding of the underlying physics. Indeed, it pinpoints the most important and complementary variables.
- For instance, the vibrational entropy is found to be strongly related to
 - the inter-atomic bond length
 - the valence range of the constituent elements (ionicity of the bond).



[P.-P. De Breuck et al., npj Computational Materials 7, 83(2021)]

The feature *f* having the highest NMI with the target variable *y* will be chosen the first one

- This provides some understanding of the underlying physics.
 Indeed, it pinpoints the most important and complementary variables.
- For instance, the refractive index is found to be strongly related to
 - an estimation of the bandgap
 - the density.



[P.-P. De Breuck et al., npj Computational Materials 7, 83(2021)]

For the next chosen features,

redundancy should also be avoided

• To this end, we define a relevance and redundancy *RR* score: given

 \blacklozenge a subset of selected features \mathcal{F}_s extracted from the set \mathcal{F}

• another feature f

$$RR(f) = \frac{\text{NMI}(f, y)}{\left[\max_{f_s \in \mathscr{F}_s} \left(\text{NMI}(f, f_s)\right)\right]^p + c}$$

where *p* and *c* are determine the relevance/redundancy balance.

- In practice, varying *p* and *c* dynamically seems to work better, as redundancy is a bigger issue with a small amount of features.
- The selection proceeds until the number of features reaches a threshold (fixed arbitrarily or, better, optimized to minimize the model error).

MODNet introduces the possibility of learning on multiple properties simultaneously



[P.-P. De Breuck et al., npj Computational Materials 7, 83(2021)]

How can we predict the phase stability of polymorphs at different temperatures?

• At T=0K: for exemple, the Cu-O system



• At T>0K, the vibrational entropy needs to be taken into account. This can be done by DFPT but it is very demanding.

Early attempts with ML were based on RF using only chemical composition features



[F. Legrain et al., Chem. Mater. 29, 6220 (2017)]

Including structural features

clearly improves the predicting power



<u>NB:</u> Performing feature selection on the input space has no effect on the results as a RF intrinsically selects optimal features while learning.

Neural-network models perform better than RF approaches whatever the size of the data set



[P.-P. De Breuck et al., npj Computational Materials 7, 83(2021)]

Feature selection is really important especially for small training size



[P.-P. De Breuck et al., npj Computational Materials 7, 83(2021)]

The joint-learning approach (m-MODNet) shows on average a slight improve in accuracy



and provides a single model for multiple properties

In particular, it is possible to obtain curves of the thermodynamic properties vs. temperature



Thanks to this approach, we can build temperature dependent stability graphs



MODNet performs very well

on the curated MatBench test suite



[A. Dunn et al., npj Comput. Mater 6, 138 (2020); https://github.com/hackingmaterials/matbench]

A probabilistic MODNet has also been developed



A probabilistic MODNet has also been developed

• Example: refractive index [F. Naccarato et al., Phys. Rev. Mater. 3, 044602 (2019)]



[Uncertainty-toolbox: K. Tran et al., Mach. Learn. Sci. Technol. 1, 025006 (2020)]

What about the quality of the data?





Behind the Paper

Boosting machine learning for materials properties by denoising multi-fidelity data

SPRINGER NATURE Materials Community

Let us consider a student preparing for the theoretical driving license exam



Books and internet



Materials DBs contain orders of magnitude less high-accuracy results than low-accuracy ones.

- Obtaining computational results faster usually requires to resort to more important approximations and hence leads, as a general rule, to a lower accuracy (cost vs. accuracy trade-off).
- Obtaining experimental results generally necessitates even more time.



Materials DBs contain orders of magnitude less high-accuracy results than low-accuracy ones.

- Obtaining computational results faster usually requires to resort to more important approximations and hence leads, as a general rule, to a lower accuracy (cost vs. accuracy trade-off).
- Obtaining experimental results generally necessitates even more time.
- Example: the electronic band gap





The multi-fidelity data are not necessarily available for the same materials

 $\begin{array}{l} \textbf{E} \rightarrow experimental \\ \textbf{P} \rightarrow DFT \ with \ \textbf{PBE} \\ \textbf{H} \rightarrow DFT \ with \ \textbf{HSE} \\ \textbf{S} \rightarrow DFT \ with \ \textbf{SCAN} \\ \textbf{G} \rightarrow DFT \ with \ \textbf{GLLB} \end{array}$







For the band gap, the distribution is different between the multi-fidelity dataset Н G PNHNSNGNE All Only σ σ σ - 1.44 1.67 - 1.47 1.67 **—** 1.71 1.71 - 2.61 2.10 - 1.92 1.77 - 3.87 3.72 **—** 1.99 1.89 **—** 4.15 2.72 - 3.88 2.67 - 3.24 3.09 - 1.20 1.67 - 1.61 1.68 - 3.05 2.69 Band Gap (eV) Splitting in 3 categories: metals, as well as small-gap ($E_g < 2$) and wide-gap ($E_g \ge 2$) semiconductors Ε Η 472 52348 6030 G 2703 2290 $E_g = 0$ $E_a = 0$ $E_{a}=0$ 2775 19903 $E_q < 2$ 1384 (46%) 531 (38%) (51%) $E_g = 0$ (23%) $E_q < 2$ *E*_{*g*} < 2 16 $E_g < 2$ 291 15218 (3%)1019 (62%) $E_q \ge 2$ *E*_a<2 $E_g \ge 2$ (29%) $E_g \ge 2$ (17%) $E_g \ge 2$ $E_g \ge 2$ 1759 762 557 17227 165 2236 (77%) (21%) (28%) (33%)(35%) (37%)

The DFT results present systematic deviations



By analyzing these systematic deviations, the data can be scaled to reduce the errors



A machine learning model built only on E data has a rather limited the accuracy



[X. Liu et al., npj Computational Materials 8, 233 (2022)]

A multi-fidelity approach was recently proposed



Fidelity-to-state embedding



[C. Chen et al., Nature Computational Science 1, 46 (2022)]

The accuracy clearly improves compared to ML on E data only



[X. Liu et al., npj Computational Materials 8, 233 (2022)]

• Transfer learning



• Joint learning





• Deep-Stacking Ensemble Learning



• Comparison of the different approaches:

Learning technique	Training sets	Samples	Test set (5-fold)	MAE	
Single-Fidelity	EXP	2480	EXP	0.382	0%
Single-Fidelity ^c	EXP ^c	4604	EXP	0.366	-4%
Transfer Learning	$PBE \cup HSE \cup EXP \to HSE \cup EXP \to EXP$	2480	EXP	0.397	+4%
Joint learning	PBE ∩ EXP	2480	EXP	0.368	-4%
Stacking Ensemble Learning	$PBE \parallel HSE \parallel EXP \rightarrow EXP$	2480	EXP	0.367	-4%
Deep-Stacking Ensemble Learning	$PBE \parallel HSE \parallel EXP \rightarrow EXP$	2480	EXP	0.370	-3%
PBE as a feature	PBE ∩ EXP	2480	EXP	0.371	-3%
Correction Learning (PBE)	$PBE \cap EXP$	2480	EXP	0.318	-17%
Single-Fidelity		325	$HSE \cap EXP$	0.582	0%
Correction Learning (PBE)	$PBE \cap HSE \cap EXP$	325	$HSE \cap EXP$	0.442	-24%
Correction Learning (HSE)	$PBE \cap HSE \cap EXP$	325	$HSE \cap EXP$	0.402	-31%
Single-Fidelity	EXP ^c	4604	$HSE \cap EXP$	0.438	0%
Correction Learning (PBE)	$PBE \cap EXP$	2480	$HSE \cap EXP$	0.356	-19%
Correction Learning (HSE)	$HSE \cap EXP$	325	$HSE \cap EXP$	0.402	-8%

[P.-P. De Breuck et al., J. Mater. Inf. 2, 10 (2022)]

• Comparison of the different approaches:

Learning technique	Training sets	Samples	Test set (5-fold)	MAE	
Single-Fidelity	EXP	2480	EXP	0.382	0%
Single-Fidelity ^c	EXP ^c	4604	EXP	0.366	-4%
Transfer Learning	$PBE \cup HSE \cup EXP \to HSE \cup EXP \to EXP$	2480	EXP	0.397	+4%
Joint learning	$PBE \cap EXP$	2480	EXP	0.368	-4%
Stacking Ensemble Learning	$PBE \parallel HSE \parallel EXP \rightarrow EXP$	2480	EXP	0.367	-4%
Deep-Stacking Ensemble Learning	$PBE \parallel HSE \parallel EXP \rightarrow EXP$	2480	EXP	0.370	-3%
PBE as a feature	$PBE \cap EXP$	2480	EXP	0.371	-3%
Correction Learning (PBE)	$PBE \cap EXP$	2480	EXP	0.318	-17%
Single-Fidelity	PBE ∩ HSE ∩ EXP	325	HSE ∩ EXP	0.582	0%
Correction Learning (PBE)	$PBE \cap HSE \cap EXP$	325	$HSE \cap EXP$	0.442	-24%
Correction Learning (HSE)	PBE ∩ HSE ∩ EXP	325	$HSE \cap EXP$	0.402	-31%
Single-Fidelity	EXP ^c	4604	$HSE \cap EXP$	0.438	0%
Correction Learning (PBE)	$PBE \cap EXP$	2480	$HSE \cap EXP$	0.356	-19%
Correction Learning (HSE)	HSE ∩ EXP	325	$HSE \cap EXP$	0.402	-8%

[P.-P. De Breuck et al., J. Mater. Inf. 2, 10 (2022)]

Recently, we tested various approaches

• Comparison of the different approaches:

Learning technique Training sets		Samples Test set MAE (5-fold)			
Single-Fidelity	EXP	2480	EXP	0.382	0%
Single-Fidelity ^c	EXP ^c	4604	EXP	0.366	-4%
Transfer Learning	$PBE \cup HSE \cup EXP \to HSE \cup EXP \to EXP$	2480	EXP	0.397	+4%
Joint learning	$PBE \cap EXP$	2480	EXP	0.368	-4%
Stacking Ensemble Learning	$PBE \parallel HSE \parallel EXP \rightarrow EXP$	2480	EXP	0.367	-4%
Deep-Stacking Ensemble Learning	$PBE \parallel HSE \parallel EXP \rightarrow EXP$	2480	EXP	0.370	-3%
PBE as a feature	$PBE \cap EXP$	2480	EXP	0.371	-3%
Correction Learning (PBE)	PBE ∩ EXP	2480	EXP	0.318	-17%
Single-Fidelity Correction Learning (PBE) Correction Learning (HSE)	PBE IN HSE IN EXP PBE IN HSE IN EXP PBE IN HSE IN EXP	325 325 325	$\begin{array}{l} HSE \cap EXP \\ HSE \cap EXP \\ HSE \cap EXP \end{array}$	0.582 0.442 0.402	0% -24% -31%
Single-Fidelity Correction Learning (PBE)	EXP ^c PBE ∩ EXP	4604 2480	HSE ∩ EXP HSE ∩ EXP	0.438 0.356	0% -19%
Correction Learning (HSE)	HSE O EXP	325	$HSE \cap EXP$	0.402	-8%

[P.-P. De Breuck et al., J. Mater. Inf. 2, 10 (2022)]

• We compared two different methods of transfer/curriculum learning:



[X. Liu et al., npj Computational Materials 8, 233 (2022)]

• All possible path were considered for the *one-by-one* training approach



[X. Liu et al., npj Computational Materials 8, 233 (2022)]

• All possible path were considered for the *one-by-one* training approach



[X. Liu et al., npj Computational Materials 8, 233 (2022)]

• All possible path were considered for the *onion* training approach



[X. Liu et al., npj Computational Materials 8, 233 (2022)]

• All possible path were considered for the *onion* training approach



[X. Liu et al., npj Computational Materials 8, 233 (2022)]

• The *onion* training approach is the best on average and it is especially good when the sequence is in increasing fidelity of the data.

		MAE			
Approach	Sequence	Global	$E_g = 0$	<i>E</i> _g < 2	<i>Eg</i> ≥ 2
only-E	E	0.680	0.490	0.534	1.131
all-together	PHSGE	0.501	0.150	0.567	1.091
one-by-one	$S \rightarrow G \rightarrow P \rightarrow E \rightarrow H$ (best)	0.484	0.228	0.571	0.884
	$H \rightarrow P \rightarrow S \rightarrow E \rightarrow G$ (worst)	0.993	1.039	0.733	1.097
	$H \rightarrow P \rightarrow G \rightarrow S \rightarrow E \text{ (worst}^*)$	0.599	0.434	0.510	0.964
onion	$PHSGE \rightarrow PHSE \rightarrow HSE \rightarrow HE \rightarrow E$ (best)	0.438	0.239	0.515	0.743
	$PHSGE \to PHSG \to PSG \to SG \to G \text{ (worst)}$	0.916	0.937	0.634	1.083
	$PHSGE \to PHSG \to PSG \to PG \to G \ (2^{nd}\text{-worst})$	0.889	0.883	0.665	1.064
	$PHSGE \to PSGE \to SGE \to GE \to E \text{ (worst*)}$	0.495	0.338	0.485	0.790

[X. Liu et al., npj Computational Materials 8, 233 (2022)]

• The denoising procedure can thus be applied...



[X. Liu et al., npj Computational Materials 8, 233 (2022)]

The effect of the denoising is quite remarkable...



The accuracy is further improved compared to the previous multi-fidelity approach



[X. Liu et al., npj Computational Materials 8, 233 (2022)]

To assess the generality of the procedure, we further apply it using MODNet

MAE (eV)

Approach	Denoiser	MODNet	MEGNet	MFGNet
only-E		0.446	0.701	
all-together		0.477	0.467	0.423
all-together (3-fi)		0.533	0.487	0.416
onion		0.403	0.397	0.395
onion (3-fi)		0.429	0.388	0.400
onion	onion	0.396	0.373	

[X. Liu et al., npj Computational Materials 8, 233 (2022)]

Take-home message

- MODNet (Material Optimal Descriptor Network) has been proposed to deal with small datasets.
- A method to take full advantage of the availability of multi-fidelity data has also been presented. It is based on:
 - an appropriate combination of all the data into a training sequence
 - a simple denoising procedure.
- Both approaches provide a sensible way to improve the results that can be achieved when limited and multi-fidelity data are available (which is particularly the case in materials science).