Citrine Informatics

The data analytics platform for the physical world

The Latest from Citrine

Summit on Data and Analytics for Materials Research

31 October 2016

Our Mission is Simple

Add as much value
to your work as possible,
immediately,
using data

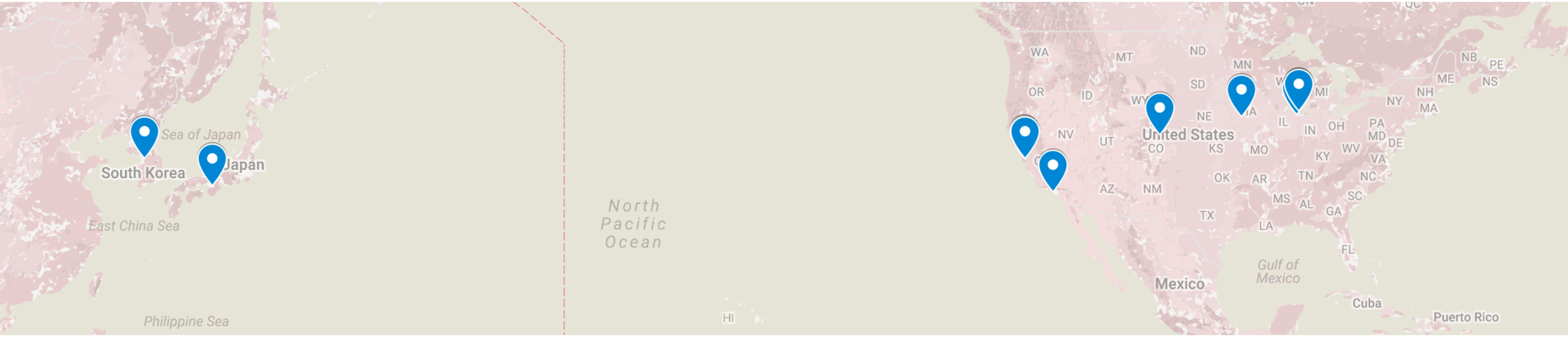
Keys to Industrial Relevance

UBIQUITY

EASE OF USE

OBVIOUS ROI

Citrine Platform: Worldwide Deployments



Several Fortune 500 companies

Citrine is the community cloud for materials data, predictive models, & post-processing

+ **All relevant data** in one place, unified from databases, research groups, papers

+ **Predictive AI**, physics-based simulations, and post-processing tools seamlessly integrated with the data

+ **Vibrant ecosystem** of researchers and developers

All Relevant Data

17m+ free data records as pif's on citrination.com (& API)

ASM and MMPDS are now official data partners, providing premium data to the platform; 6 free NIST SRD's & much more

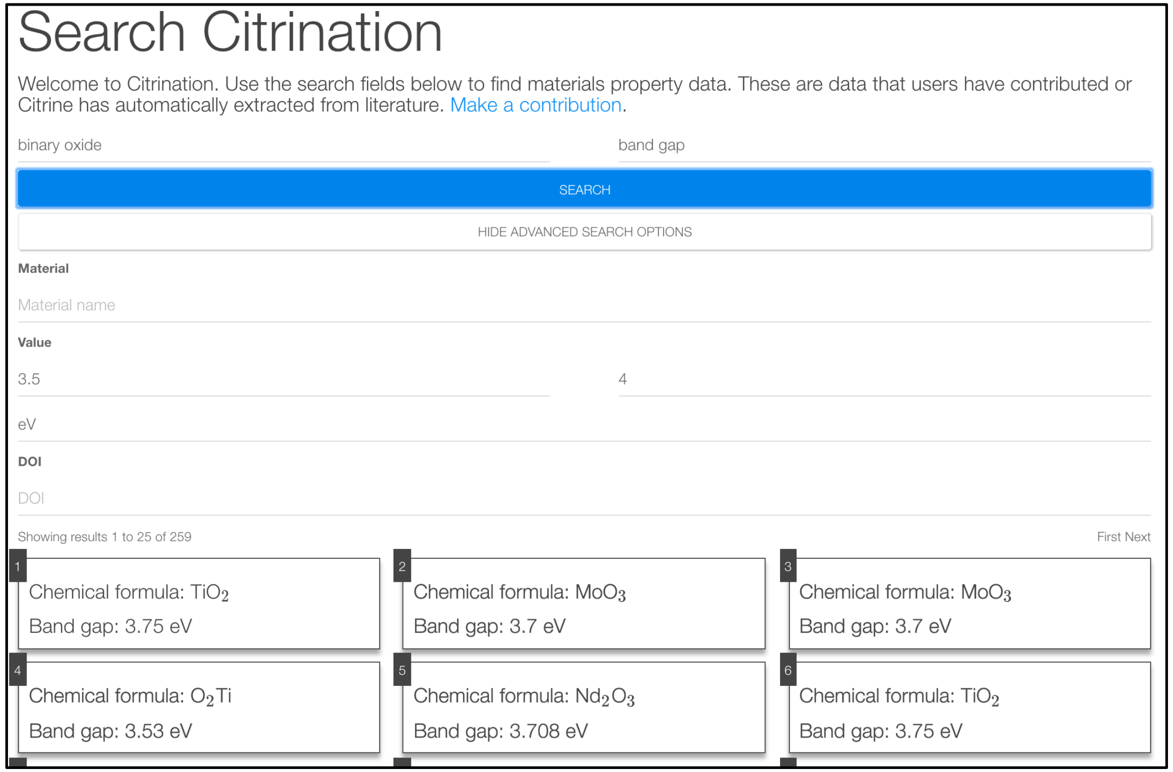
General Plots

SHOW FILTERS | SAVE COLUMN LAYOUT | RESTORE COLUMN LAYOUT | HIDE EMPTY COLUMNS | SHOW ALL COLUMNS | EXPORT TO CSV

	Alloy	Properties				
Record	Name	Specification	Product form	Design tensile ultimate stress	Design tensile yield stress	Design ultimate bearing str
View	AISI 1025	AMS-T-5066	Tubing	55 ksi	36 ksi	90 ksi
View	AISI 1025	AMS 5046	Sheet, strip, and plate	55 ksi	36 ksi	90 ksi
View	AISI 1025	AMS 5075	Tubing (Seamless)	55 ksi	36 ksi	90 ksi
View	AISI 1025	ASTM A 108	Bar	55 ksi	36 ksi	90 ksi
View	AISI 4130	AMS 6345	Sheet, strip, and plate	90 ksi	70 ksi	190 ksi
View	AISI 4130	AMS 6360, AMS 6373	Tubing	90 ksi	70 ksi	190 ksi
View	AISI 4130	AMS 6374	Tubing	95 ksi	75 ksi	200 ksi
View	AISI 4135	AMS 6365	Tubing	95 ksi	80 ksi	180 ksi
View	AISI 4130	AMS 6361	Tubing	125 ksi	100 ksi	251 ksi
View	AISI 4130	AMS 6362	Tubing	150 ksi	135 ksi	285 ksi

1 to 10 of 303 First Previous Page 1 of 31 Next

Graphical & API (Semantic) Search



"Show me binary oxides with band gap between 3.5 and 4 eV"

Open Data Matters

JOM
August 2016, Volume 68, Issue 8, pp 2116–2125

Semi-Supervised Approach to Phase Identification from Combinatorial Sample Diffraction Patterns

Authors | Authors and affiliations

Jonathan Kenneth Bunn, Jianjun Hu, Jason R. Hattrick-Simpers

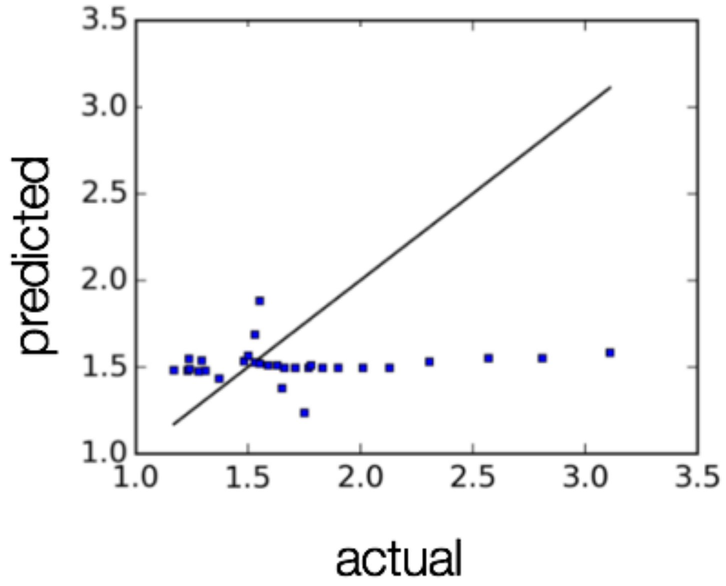
"In the current implementation, SS-AutoPhase (semi-supervised AutoPhase) was used to phase map 278 diffractograms from a FeGaPd "open-data" combinatorial thin-film library.**[citation for Citrination]**

...

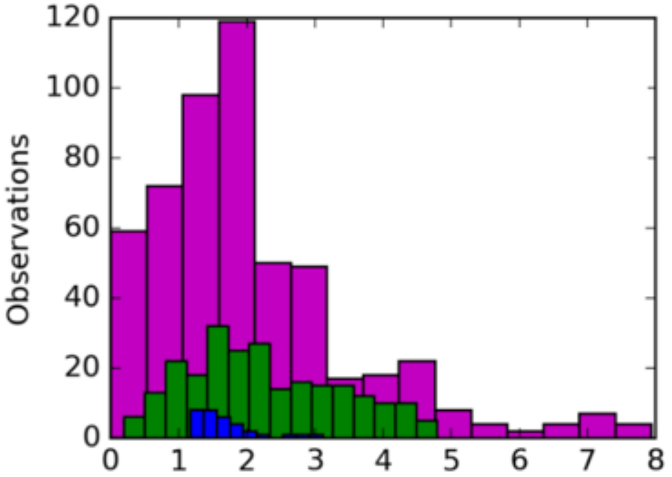
In this study, the open FeGaPd structural data not only allowed for the validation of SS-AutoPhase, but also it enabled a **new materials discovery from data produced >10 years ago**. By making these data open, the value of the data to the materials community was increased."

Value of Data Scale in Practice

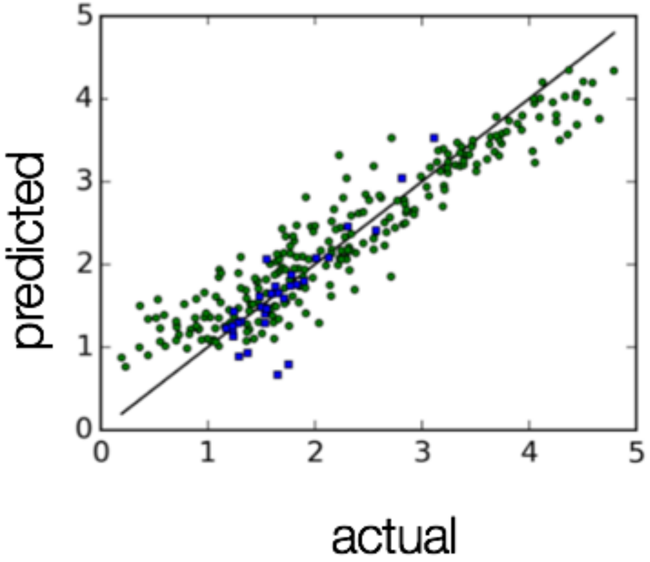
Initial dataset too small for signal



→Larger training set via Citrine platform



Predictive model drove real-world discovery



The Citrine Predictive Approach

Start with known physical and chemical relationships

(priors = DFT ground states, CALPHAD simulations, design rules…)

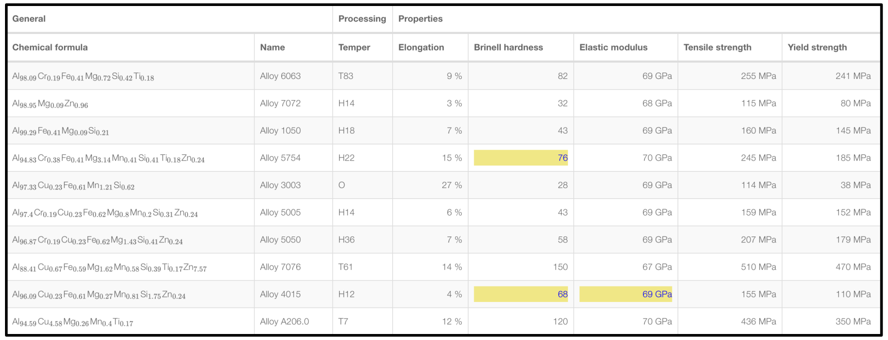
then

fit remaining variance to reality (huge quantities of relevant measurements) with machine learning

Platform Machine Learning Capabilities

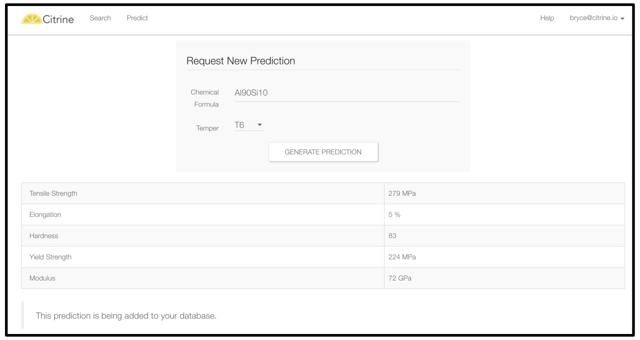
Citrine's platform exposes machine learning in 3 ways

Filling in Data Gaps

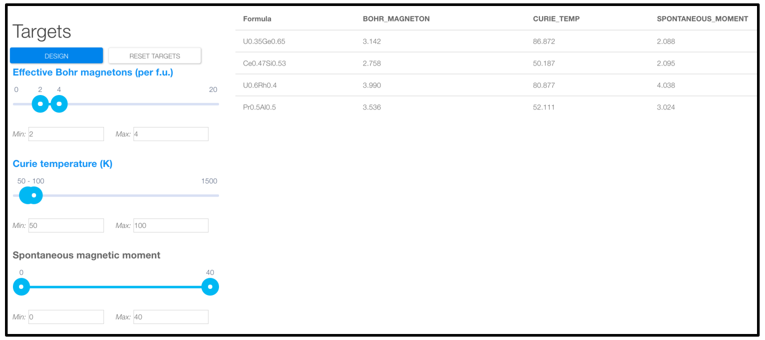


General		Processing	Properties				
Chemical formula	Name	Temper	Elongation	Brinell hardness	Elastic modulus	Tensile strength	Yield strength
$Al_{98.09}Cr_{0.19}Fe_{0.41}Mg_{0.72}Si_{0.42}Ti_{0.18}$	Alloy 6063	T83	9 %	82	69 GPa	255 MPa	241 MPa
$Al_{98.95}Mg_{0.09}Zn_{0.96}$	Alloy 7072	H14	3 %	32	68 GPa	115 MPa	80 MPa
$Al_{99.29}Fe_{0.41}Mg_{0.09}Si_{0.21}$	Alloy 1050	H18	7 %	43	69 GPa	160 MPa	145 MPa
$Al_{94.83}Cr_{0.38}Fe_{0.41}Mg_{3.14}Mn_{0.41}Si_{0.41}Ti_{0.18}Zn_{0.24}$	Alloy 5754	H22	15 %	76	70 GPa	245 MPa	185 MPa
$Al_{97.33}Cu_{0.23}Fe_{0.61}Mn_{1.21}Si_{0.62}$	Alloy 3003	O	27 %	28	69 GPa	114 MPa	38 MPa
$Al_{97.4}Cr_{0.19}Cu_{0.23}Fe_{0.62}Mg_{0.8}Mn_{0.2}Si_{0.31}Zn_{0.24}$	Alloy 5005	H14	6 %	43	69 GPa	159 MPa	152 MPa
$Al_{96.87}Cr_{0.19}Cu_{0.23}Fe_{0.62}Mg_{1.43}Si_{0.41}Zn_{0.24}$	Alloy 5050	H36	7 %	58	69 GPa	207 MPa	179 MPa
$Al_{88.41}Cu_{0.67}Fe_{0.59}Mg_{1.62}Mn_{0.58}Si_{0.39}Ti_{0.17}Zn_{7.57}$	Alloy 7076	T61	14 %	150	67 GPa	510 MPa	470 MPa
$Al_{96.09}Cu_{0.23}Fe_{0.61}Mg_{0.27}Mn_{0.81}Si_{1.75}Zn_{0.24}$	Alloy 4015	H12	4 %	68	69 GPa	155 MPa	110 MPa
$Al_{94.59}Cu_{4.58}Mg_{0.26}Mn_{0.4}Ti_{0.17}$	Alloy A206.0	T7	12 %	120	70 GPa	436 MPa	350 MPa

Predict Interface



Inverse Design

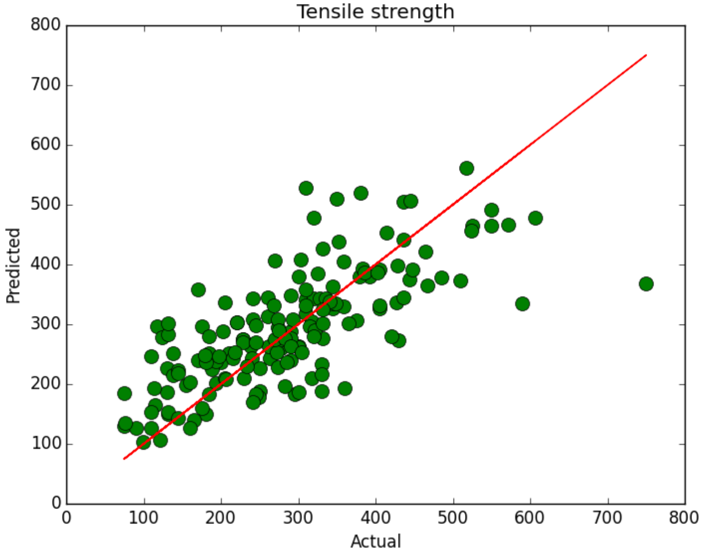


Predictive Artificial Intelligence for Materials

Collaboration with Computherm to demonstrate benefits of CALPHAD data in training AI to predict Al alloy mech properties

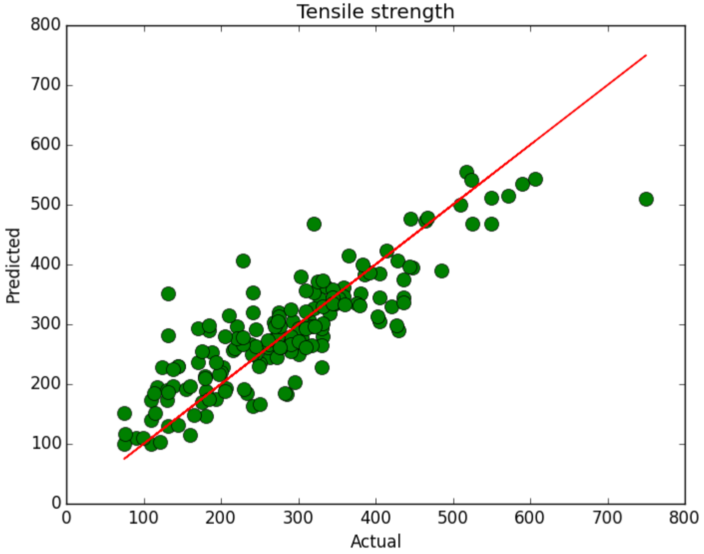
AI without CALPHAD

RMSE = 82 MPa



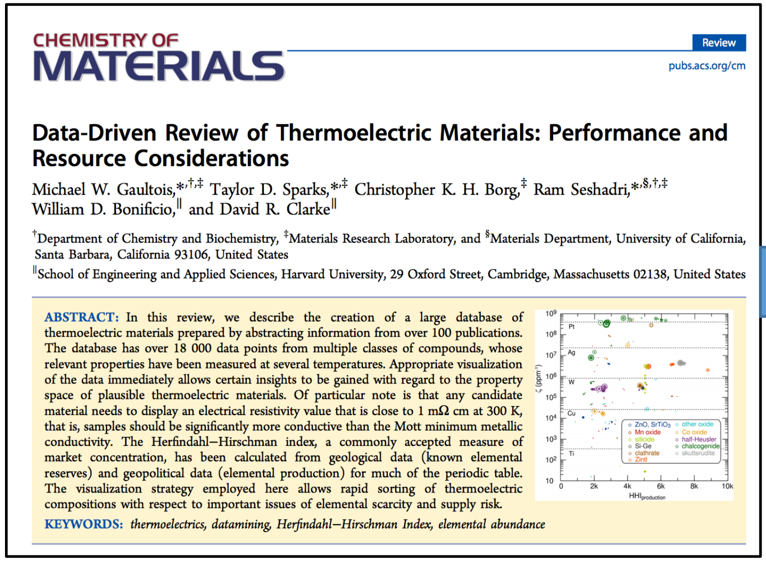
AI with CALPHAD

RMSE = 61 MPa



Machine Learning on Demand

Paper with valuable data


CHEMISTRY OF MATERIALS

Review

pubs.acs.org/cm

Data-Driven Review of Thermoelectric Materials: Performance and Resource Considerations

Michael W. Gaultois,*[†,‡] Taylor D. Sparks,*[‡] Christopher K. H. Borg,[‡] Ram Seshadri,*[§,†,‡] William D. Bonificio,[∥] and David R. Clarke[∥]

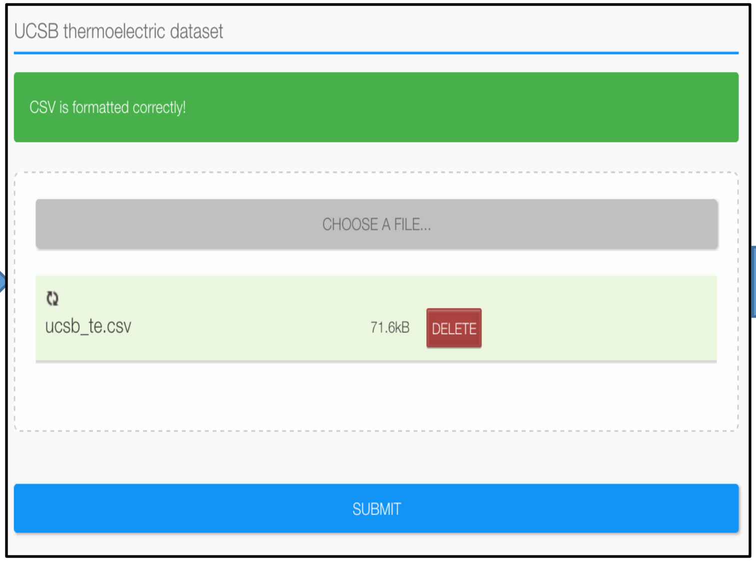
[†]Department of Chemistry and Biochemistry, [‡]Materials Research Laboratory, and [§]Materials Department, University of California, Santa Barbara, California 93106, United States

[∥]School of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, Massachusetts 02138, United States

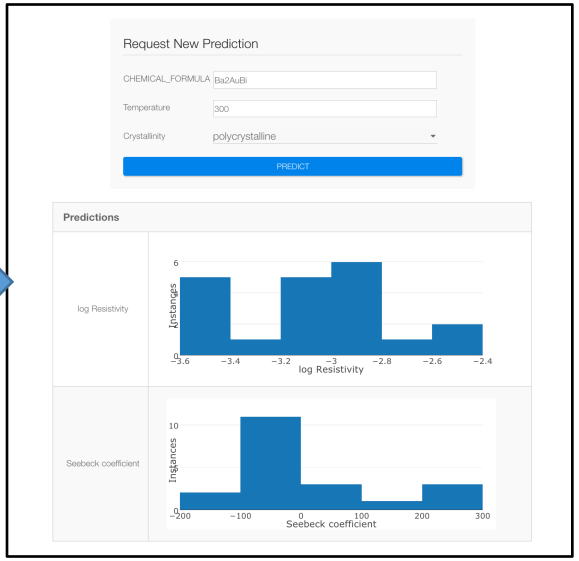
ABSTRACT: In this review, we describe the creation of a large database of thermoelectric materials prepared by abstracting information from over 100 publications. The database has over 18 000 data points from multiple classes of compounds, whose relevant properties have been measured at several temperatures. Appropriate visualization of the data immediately allows certain insights to be gained with regard to the property space of plausible thermoelectric materials. Of particular note is that any candidate material needs to display an electrical resistivity value that is close to 1 mΩ cm at 300 K, that is, samples should be significantly more conductive than the Mott minimum metallic conductivity. The Herfindahl−Hirschman index, a commonly accepted measure of market concentration, has been calculated from geological data (known elemental reserves) and geopolitical data (elemental production) for much of the periodic table. The visualization strategy employed here allows rapid sorting of thermoelectric compositions with respect to important issues of elemental scarcity and supply risk.

KEYWORDS: *thermoelectrics, datamining, Herfindahl−Hirschman Index, elemental abundance*

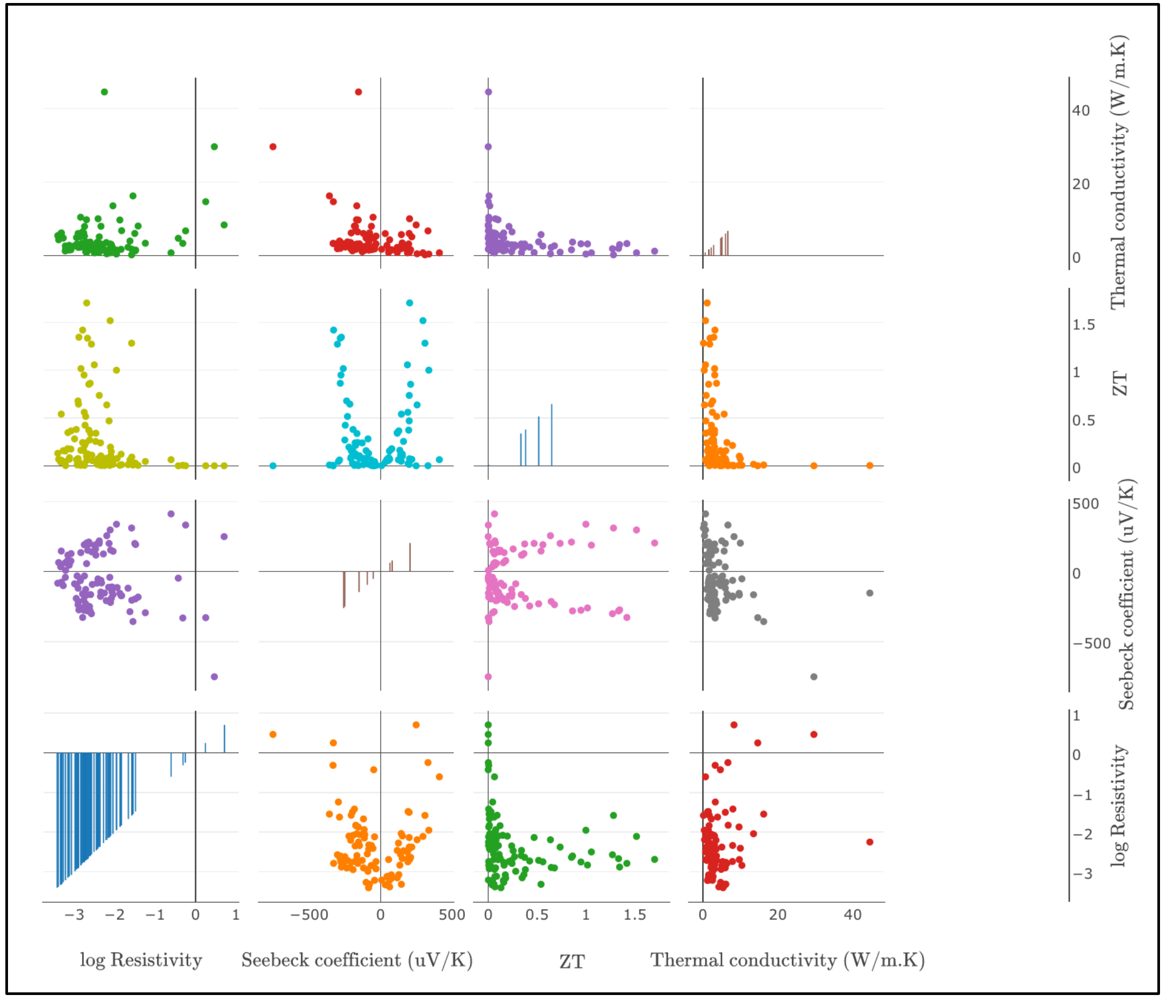
Drag and drop .csv



Interactive models



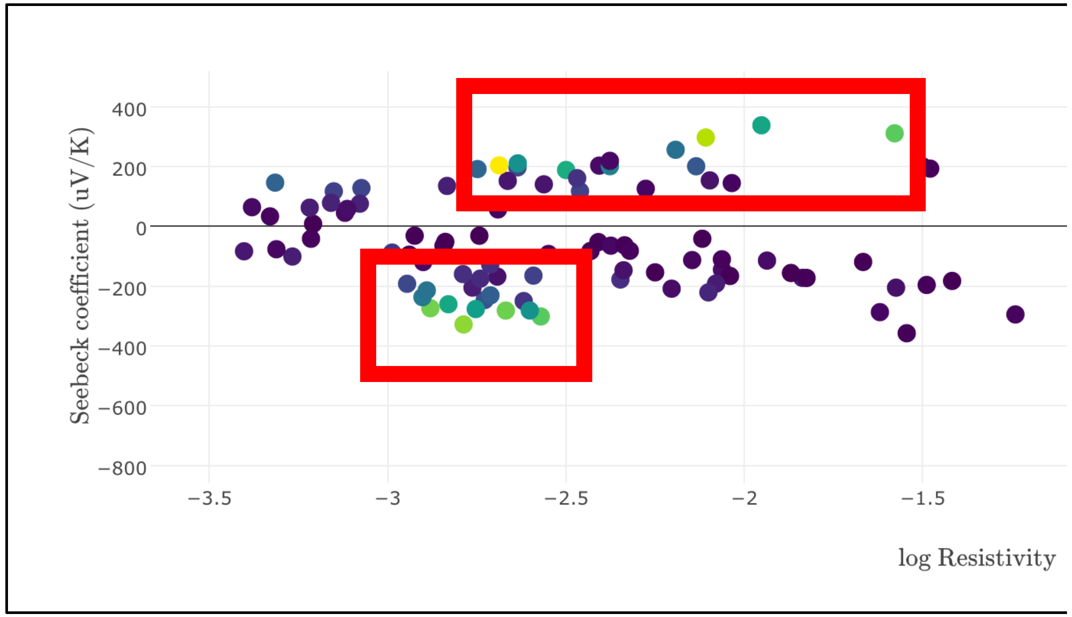
Dataset Visualization



Scatterplot of UCSB thermoelectrics dataset

Gaultois *et al.*, *Chem Mater* **25** (2013)

Dataset Visualization



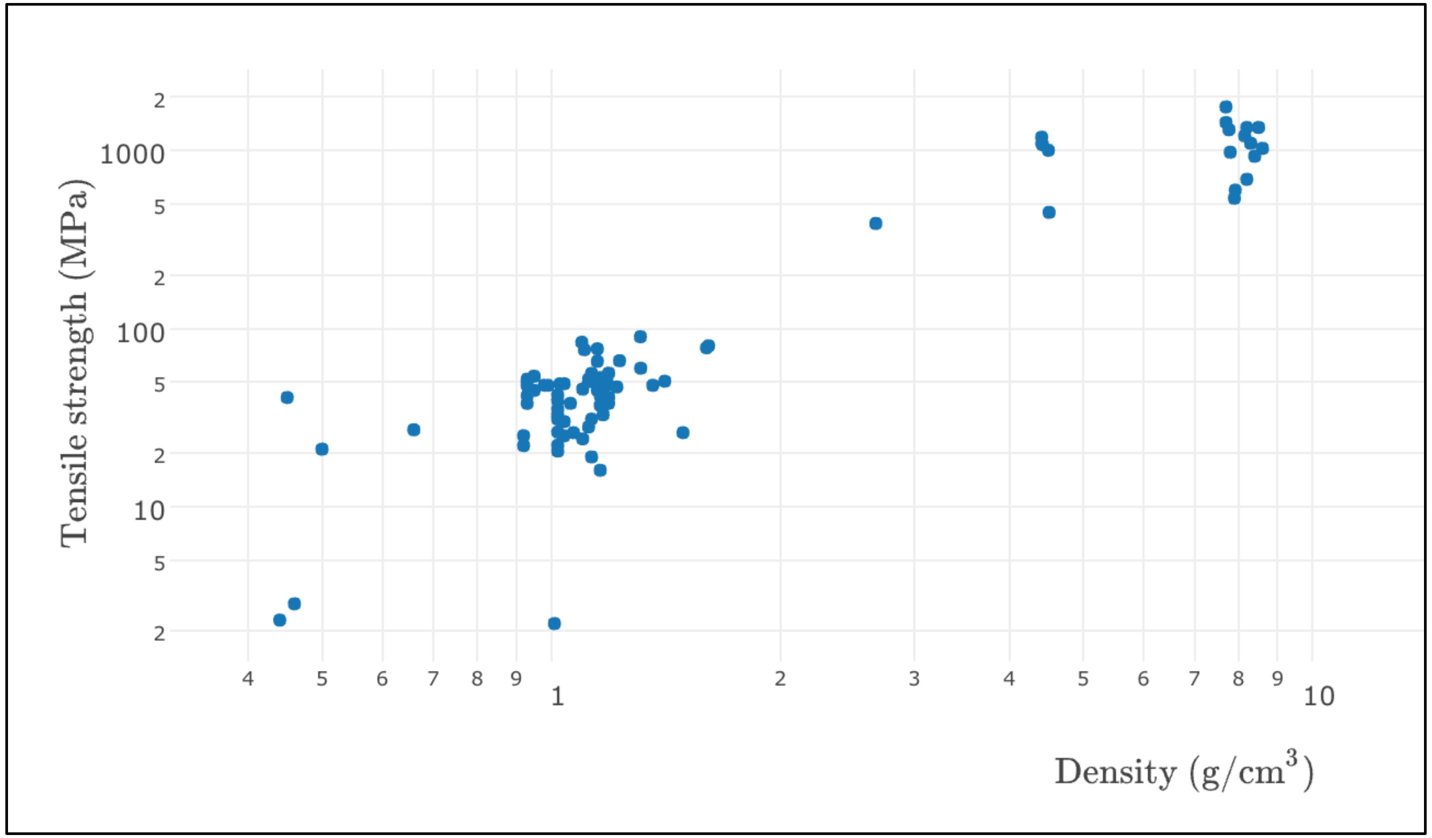
Material families: Mn oxides, Co oxides, ZnO and $SrTiO_3$, other oxides, chalcogenides, clathrates, skutterudites, half-Heuslers, Zintls, Si and Ge, Silicide

Citrine platform recreates visuals from the paper interactively

Gaultois *et al.*, *Chem Mater* **25** (2013)

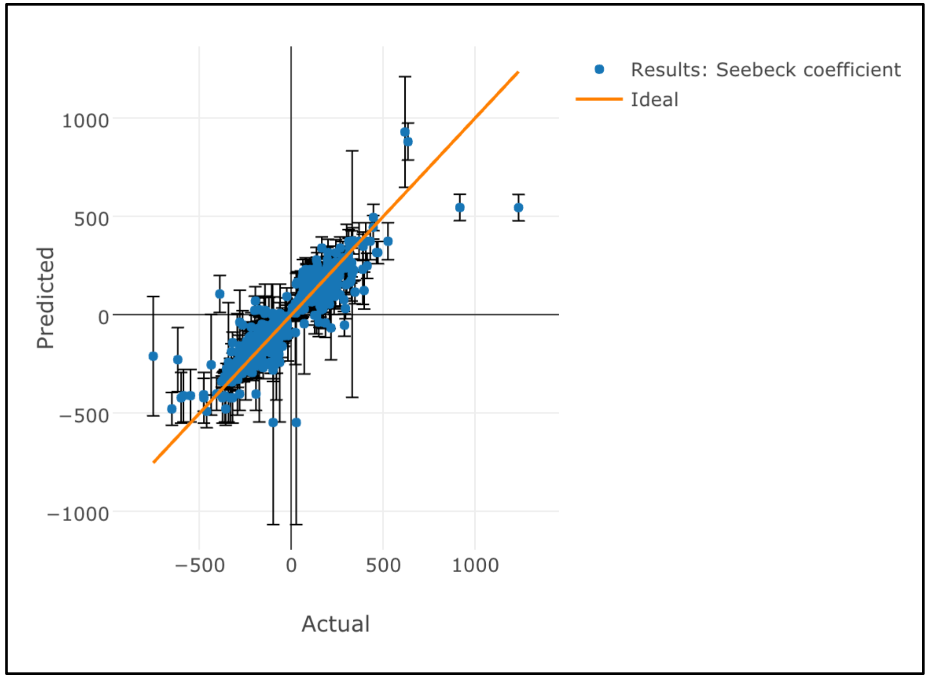
Dataset Visualization

Dynamic Ashby plot of commercial 3D printing materials



Uncertainty Quantification

All Models Have Error Bars



Predictions are Distributions

Feature Selection & Importance

Magpie feature set
bitbucket.org/wolverton/magpie
doi:10.1038/npjcompumats.2016.28

We are working with the informatics community to build a comprehensive library of *all published features*

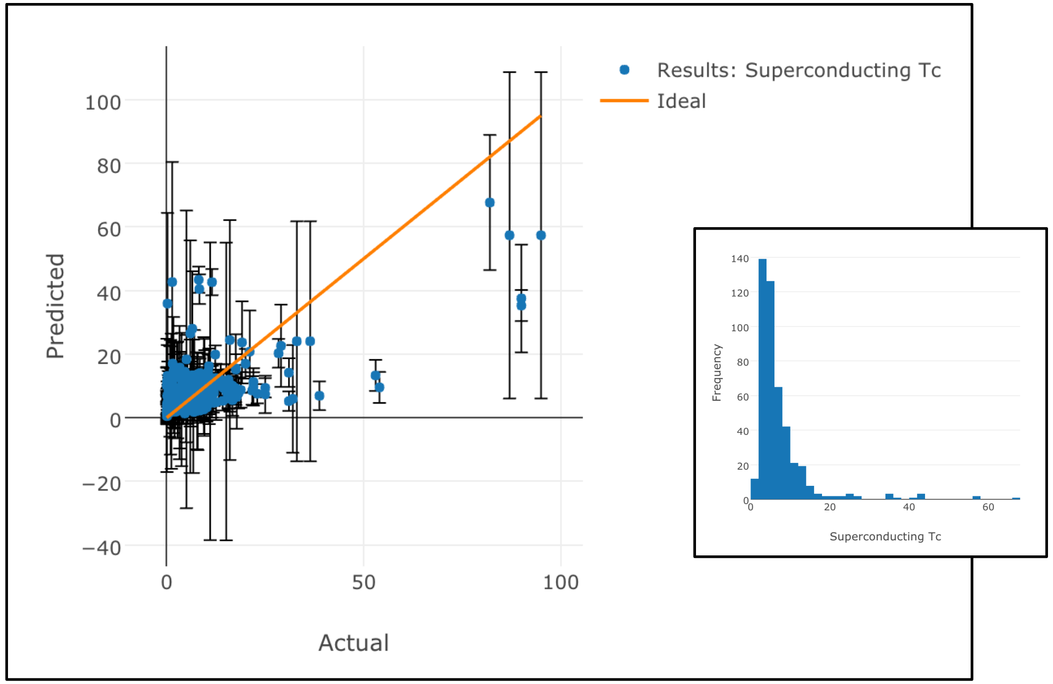
Important Features

Seebeck coefficient	
CHEMICAL_FORMULA_ElectronAffinity_l1	0.11953138134990215
CHEMICAL_FORMULA_NsUnfilled_l1	0.10335721226261824
CHEMICAL_FORMULA_NUnfilled_l1	0.09780109721022519
CHEMICAL_FORMULA_NsValence_l1	0.08118081419616913
CHEMICAL_FORMULA_GSestFCClatcnt_l1	0.07888443644268245
CHEMICAL_FORMULA_ICSDVolume_l1	0.07696848738961315
CHEMICAL_FORMULA_Row_l1	0.07500187458125034
CHEMICAL_FORMULA_MiracleRadius_l1	0.06839587008787573
CHEMICAL_FORMULA_GSestBCClatcnt_l1	0.06776567820725884
CHEMICAL_FORMULA_BoilingT_l1	0.06771500457780066
CHEMICAL_FORMULA_GSvolume_pa_l1	0.06425279861122248
CHEMICAL_FORMULA_ShearModulus_l1	0.06199032157983999
Temperature	0.03715502350354161

Model Anything!

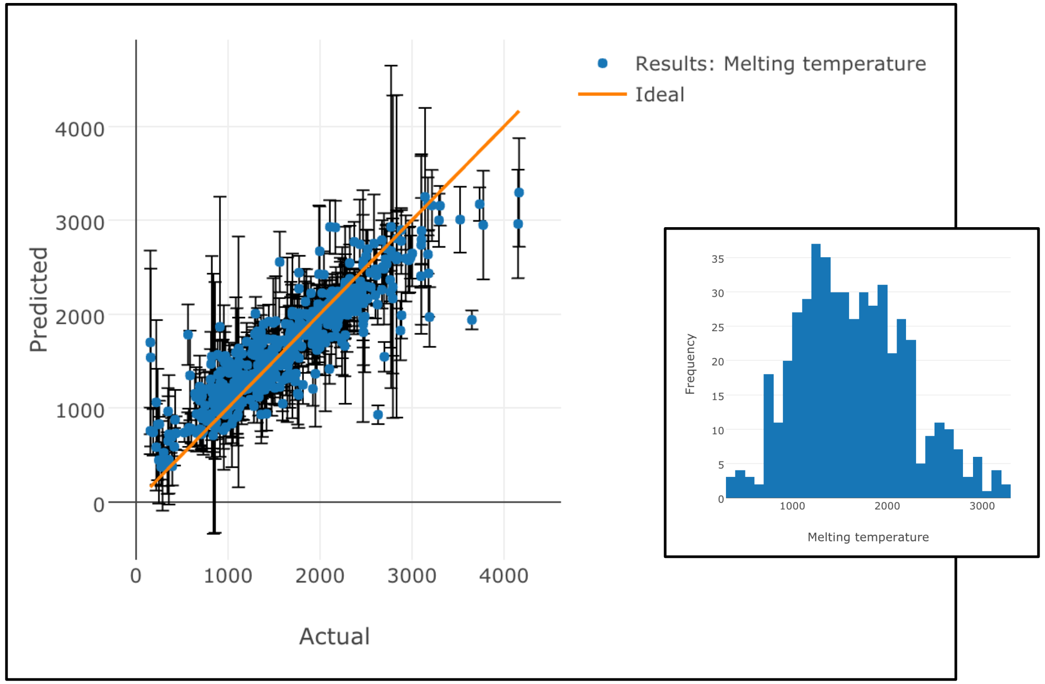
NIMS Superconductor Dataset

(turns out, superconductors = not easy)



NIMS Melting Point Dataset

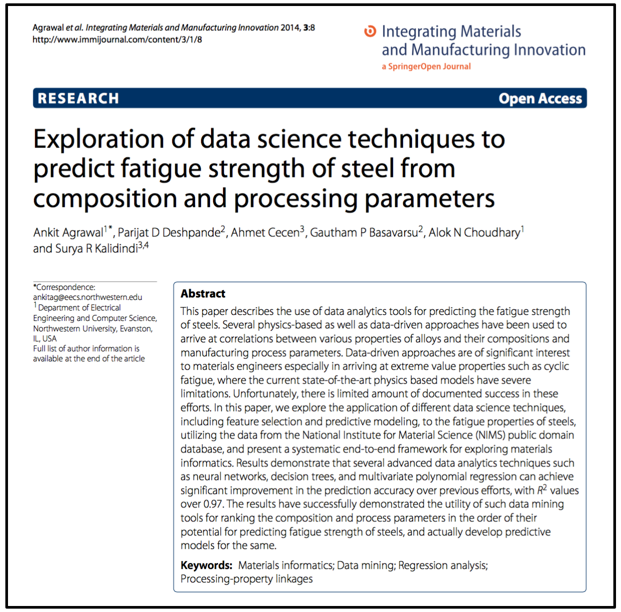
(melting point = much easier)



Model Anything!

Citrine platform creates steel fatigue model from published dataset

Agrawal *et al.*, *IMMI* **3** (2014)


Agrawal *et al.* *Integrating Materials and Manufacturing Innovation* 2014, **3**:8
http://www.immijournal.com/content/3/1/8

Integrating Materials and Manufacturing Innovation
a SpringerOpen Journal

RESEARCH **Open Access**

Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters

Ankit Agrawal[1*], Parijat D Deshpande[2], Ahmet Cecen[3], Gautham P Basavarsu[2], Alok N Choudhary[1] and Surya R Kalidindi[3,4]

*Correspondence: ankitag@eecs.northwestern.edu
[1] Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA
Full list of author information is available at the end of the article

Abstract

This paper describes the use of data analytics tools for predicting the fatigue strength of steels. Several physics-based as well as data-driven approaches have been used to arrive at correlations between various properties of alloys and their compositions and manufacturing process parameters. Data-driven approaches are of significant interest to materials engineers especially in arriving at extreme value properties such as cyclic fatigue, where the current state-of-the-art physics based models have severe limitations. Unfortunately, there is limited amount of documented success in these efforts. In this paper, we explore the application of different data science techniques, including feature selection and predictive modeling, to the fatigue properties of steels, utilizing the data from the National Institute for Material Science (NIMS) public domain database, and present a systematic end-to-end framework for exploring materials informatics. Results demonstrate that several advanced data analytics techniques such as neural networks, decision trees, and multivariate polynomial regression can achieve significant improvement in the prediction accuracy over previous efforts, with R^2 values over 0.97. The results have successfully demonstrated the utility of such data mining tools for ranking the composition and process parameters in the order of their potential for predicting fatigue strength of steels, and actually develop predictive models for the same.

Keywords: Materials informatics; Data mining; Regression analysis; Processing-property linkages

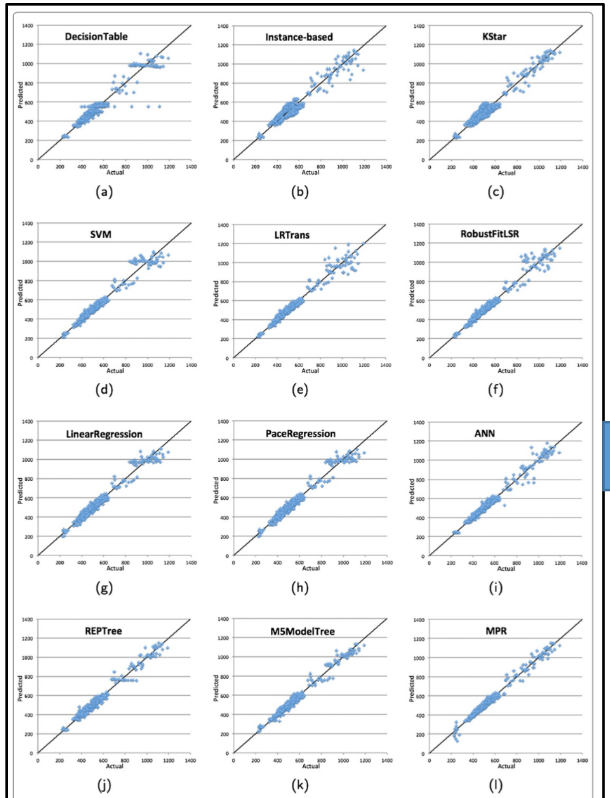
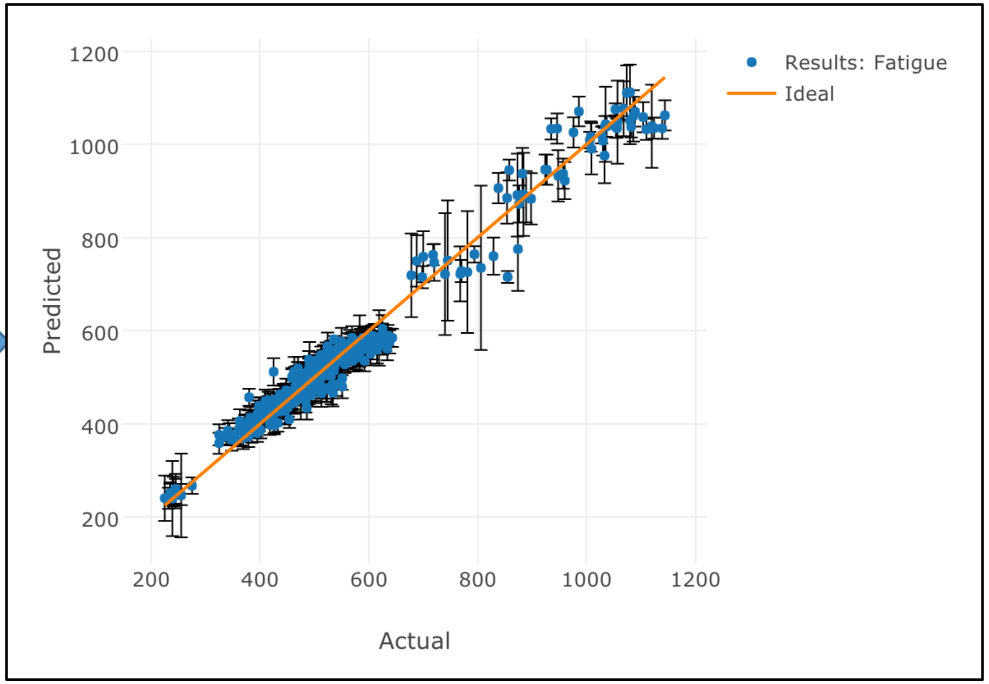


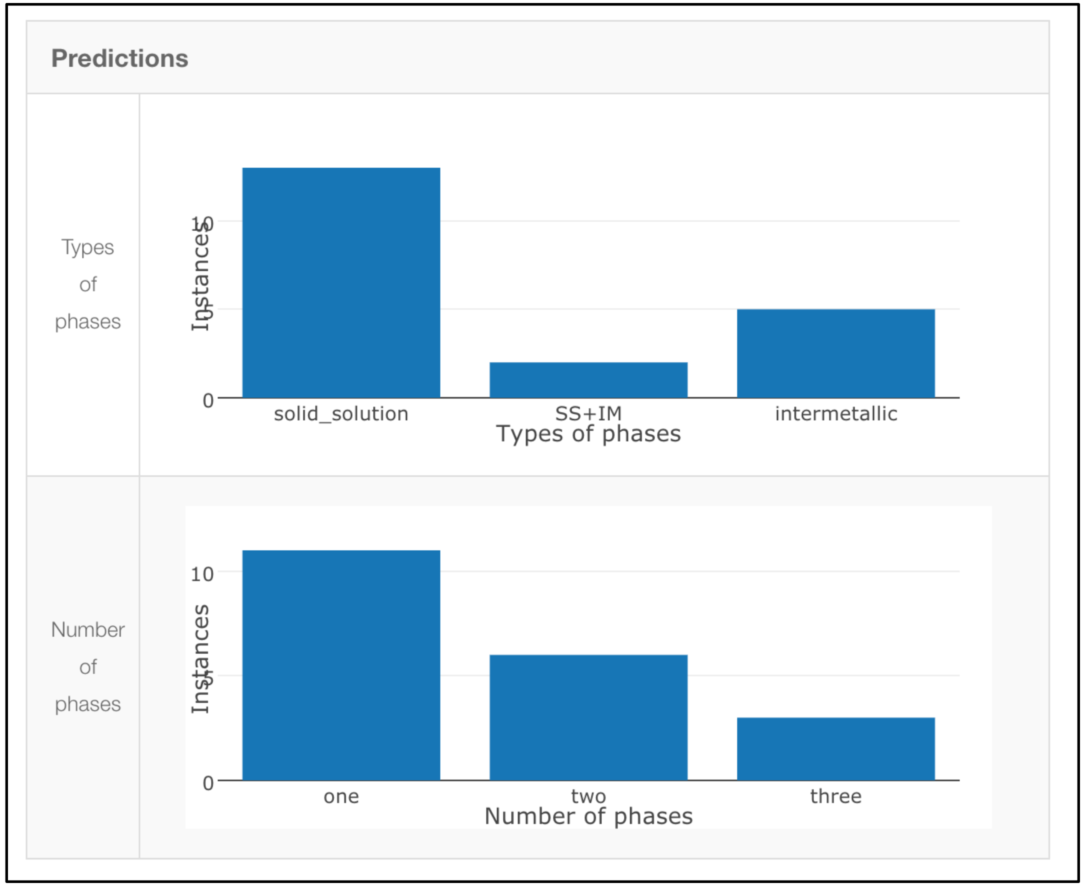
Figure 6 Scatter plots. Scatter plots for the 12 modeling techniques. X-axis and Y-axis denote the actual and predicted fatigue strength (in MPa) respectively. **a)** Decision Table; **b)** Instance-based; **c)** KStar; **d)** Support Vector Machines; **e)** Regression with Transformed Terms; **f)** RobustFit Regression; **g)** Linear Regression; **h)** Pace Regression; **i)** Artificial Neural Networks; **j)** Reduced Error Pruning Trees; **k)** M5 Model Trees; **l)** Multivariate Polynomial Regression.



Model Anything!

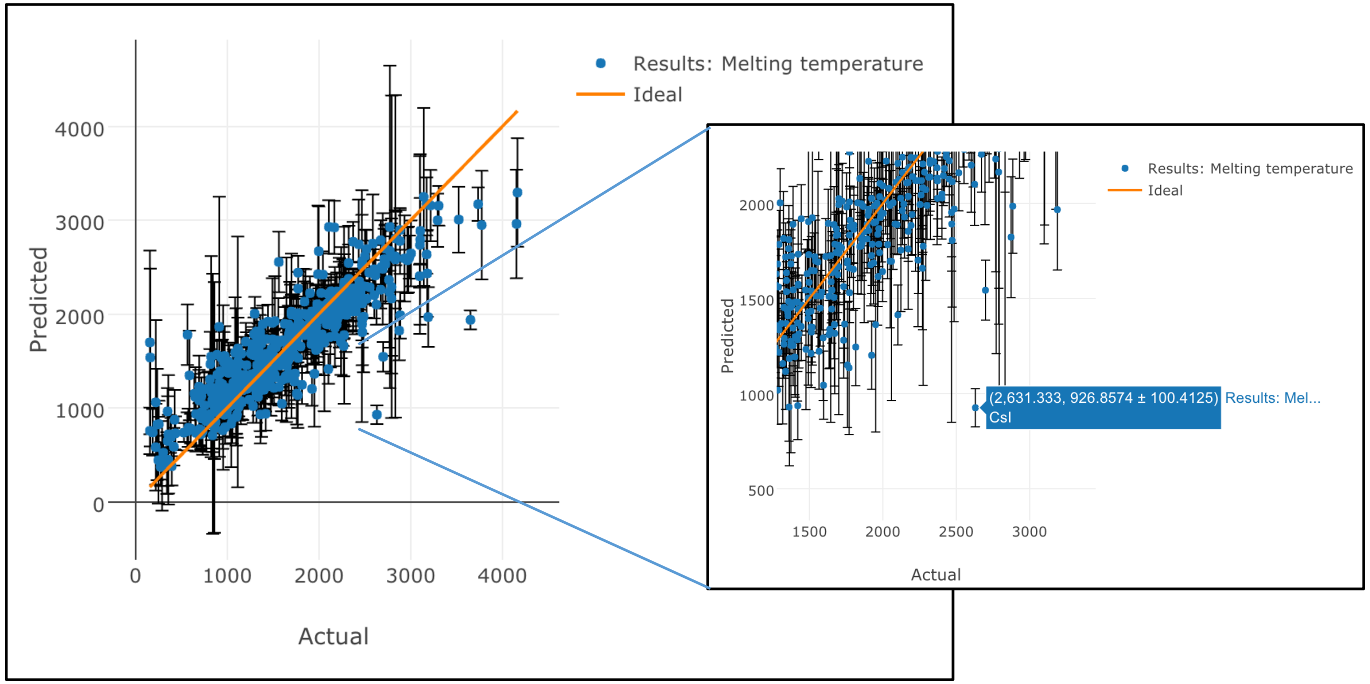
Citrine platform trained on
HEA phase stability database
D Miracle & O Senkov, *Acta Mater* 2016

Ex: MoRhRu correctly predicted to be single-phase SS



Machine Learning-Assisted Data Curation

NIMS Melting Point Dataset



Properties

Show details of ...

No.	Chemical formula	Data type	Property	Property value	Source references
1	CsI	Property	congruent melting temperature, T_{fus}	2660.(76) K	Phys. Rev. B: Condens. Matter,1996,53,,556-563,Boehler R., Ross M., Boercker D.B.
			congruent melting temperature, T_{fus}	3466.(134) K	
			congruent melting temperature, T_{fus}	2500 K	
			congruent melting temperature, T_{fus}	3500 K	
2	CsI	Property	congruent melting temperature, T_{fus}	894 K	J. Alloys Compd.,1993,201,,217-221,Qiao Z., Zhuang W., Wu S., Wang S., Zhao X.
3	CsI	Property	congruent melting temperature, T_{fus}	899 K	J. Less-Common Met.,1989,149,,95-99,Chen X.Z., Wang S.H., Jiang S.B.
4	CsI	Property	congruent melting temperature, T_{fus}	3500 K	Phys. Rev. B: Condens. Matter,1985,31,,1457-1462,Radousky H.B., Ross M., Mitchell A.C., Nellis W.J.

CsI

Predicted: 927 K

Training: 2631.333 K

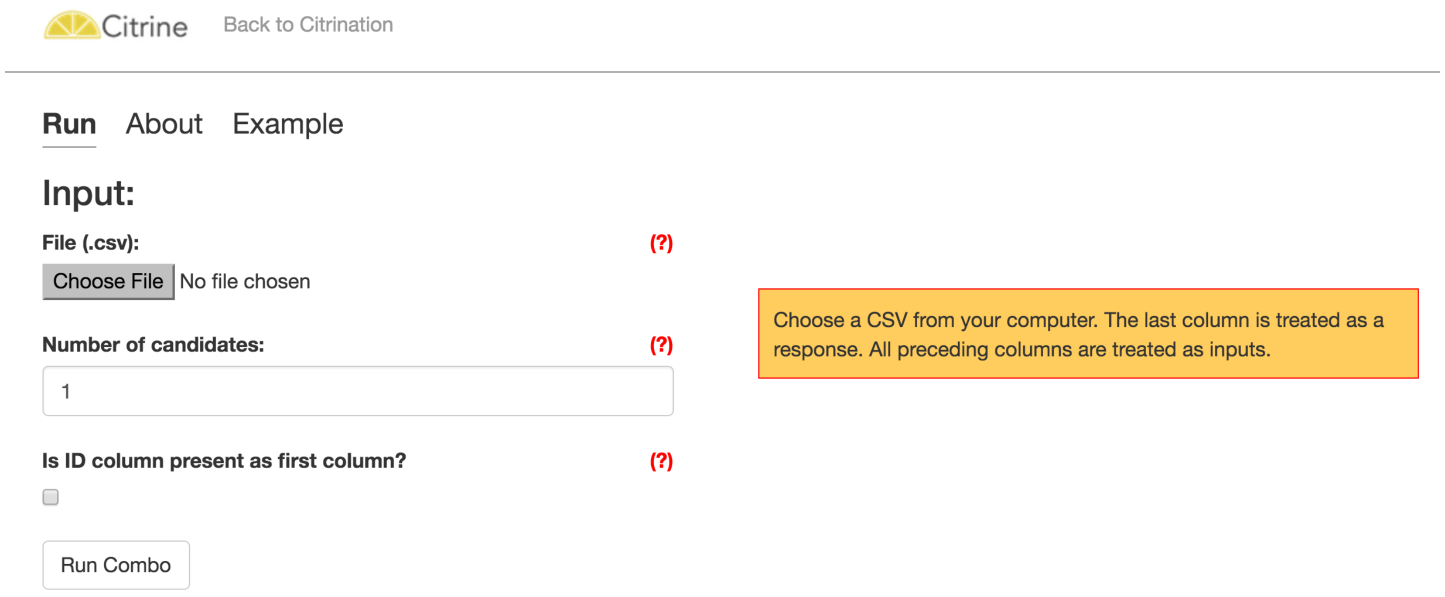
1 atm value: 831 K

Vibrant Ecosystem

Citrine has a new developers' program to enable researchers to publish code that integrates on Citrination

COMBO Bayesian Optimization Package

K Tsuda, Univ Tokyo / NIMS



“Powered by Citrine” Launch

Anchor set of university labs deploying Citrine lab-wide

We are training these users on our API, dataset templates, machine learning templates, PIF data format, and pdf->dataset extraction tools

Data-Driven Materials Community

Data-Driven Materials Science & Chemistry Newsletter (citrine.io/ddms-newsletter) has >200 weekly readers

"Your new research highlights are great. There's nothing else out there like this for materials informatics … Particularly when there's a ton of stuff to do in a day, the 1-2 paragraphs plus a figure is a perfect length to start off the day with a hit of research." –a reader

Citrine Business Model

Free platform (data & apps) available to everyone

Users of the free platform allow Citrine's algorithms to learn from their data (*Gmail model=monetizing data, not users*)

Industrial users pay for data privacy, while tapping the insights of the free platform

Some premium platform content (e.g., commercial databases)

Sustainability

Citrine's team of 15+ spends $mm/year to create a scalable, secure, extensible, supported materials data infrastructure for thousands of users—this is not fast, easy, cheap, or temporary

Things we build, track, or have:

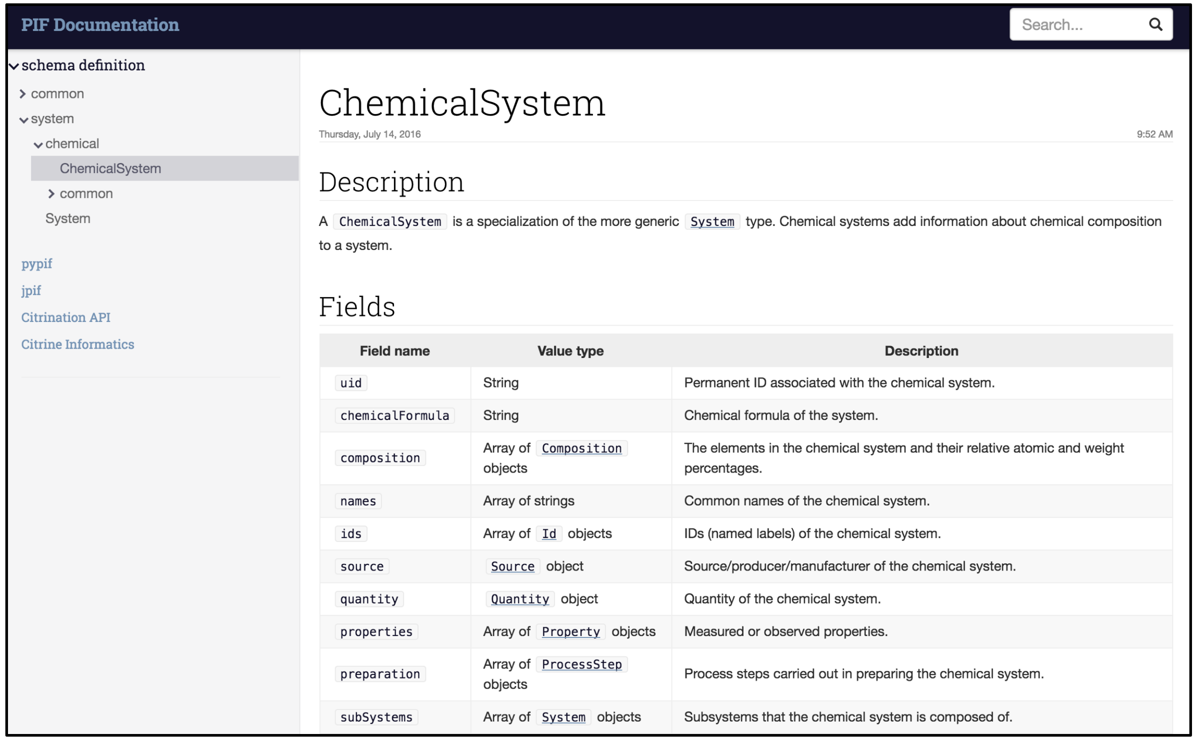
- Uptime
- Performance
- Feature velocity
- Security
- Support
- Quality assurance
- Decades of enterprise s/w engineering experience

Citrine Informatics

Citrine Does Not Lock Users In

Our data structure (pif) is completely open-source JSON: you can export all of your data out of Citrine and back it up elsewhere

We want users using us because they love our platform, not because their data are trapped



PIF Documentation

Search...

schema definition
- common
- system
  - chemical
    - ChemicalSystem
    - common
  - System

pypif

jpif

Citrination API

Citrine Informatics

ChemicalSystem

Thursday, July 14, 2016 9:52 AM

Description

A `ChemicalSystem` is a specialization of the more generic `System` type. Chemical systems add information about chemical composition to a system.

Fields

Field name	Value type	Description
`uid`	String	Permanent ID associated with the chemical system.
`chemicalFormula`	String	Chemical formula of the system.
`composition`	Array of `Composition` objects	The elements in the chemical system and their relative atomic and weight percentages.
`names`	Array of strings	Common names of the chemical system.
`ids`	Array of `Id` objects	IDs (named labels) of the chemical system.
`source`	`Source` object	Source/producer/manufacturer of the chemical system.
`quantity`	`Quantity` object	Quantity of the chemical system.
`properties`	Array of `Property` objects	Measured or observed properties.
`preparation`	Array of `ProcessStep` objects	Process steps carried out in preparing the chemical system.
`subSystems`	Array of `System` objects	Subsystems that the chemical system is composed of.

citrine.io/pif

(also see *MRS Bull* article on pif)

Let's Create Community Infrastructure

Lots of groups working on roughly the same core web platform features and data plumbing

How can Citrine make it easier for you to build on top of or integrate with our core platform capabilities?

"Let Citrine handle the IT so you can focus on science"