Machine Learning Framework for Impurity Level Prediction in Semiconductors

Arun Mannodi-Kanakkithodi[+], Michael Toriyama[+], Fatih G. Sen[+], Michael J. Davis[✵], Maciej P. Polak[○], Ryan Jacobs[○], Dane Morgan[○], Xiaofeng Xiang[❐], Laura Jacoby[❐], Robert Biegaj[❐] and Maria K.Y. Chan[+]

[+]Center for Nanoscale Materials, Argonne National Laboratory
[✵]Chemical Sciences and Engineering, Argonne National Laboratory
[○]Department of Materials Science and Engineering, University of Wisconsin-Madison
[❐]Direct Capstone Program, University of Washington

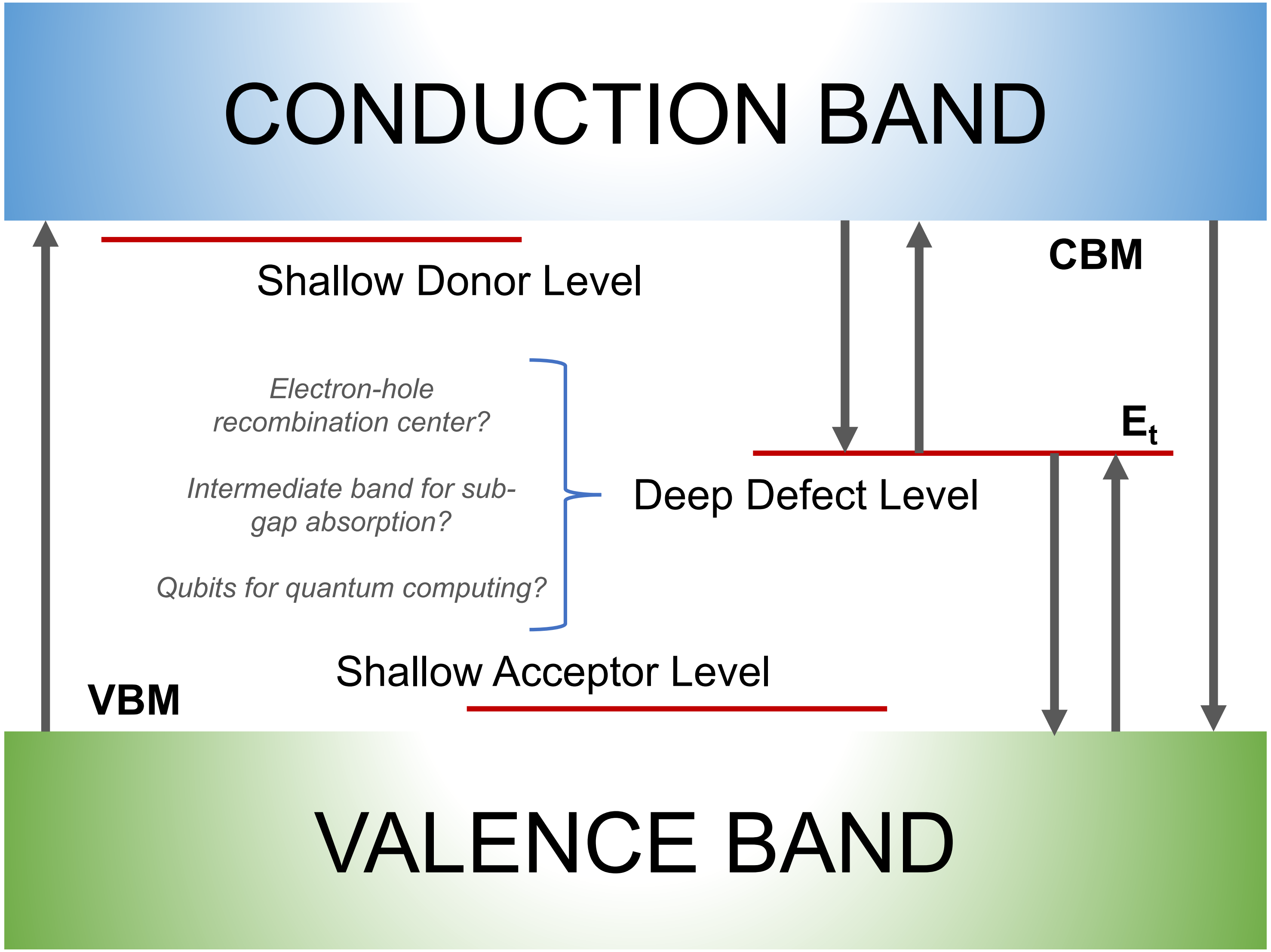
U.S. DEPARTMENT OF ENERGY
Office of Science

Argonne NATIONAL LABORATORY

vith

found in the file.

Impurity Levels in Semiconductors


CONDUCTION BAND
CBM
Shallow Donor Level
Electron-hole recombination center?
Intermediate band for sub-gap absorption?
Qubits for quantum computing?
Deep Defect Level
E_t
Shallow Acceptor Level
VBM
VALENCE BAND

CHALLENGES

Experimental

- Sample preparation difficult with DLTS or CL.
- Difficult to assign observed levels to particular defect.

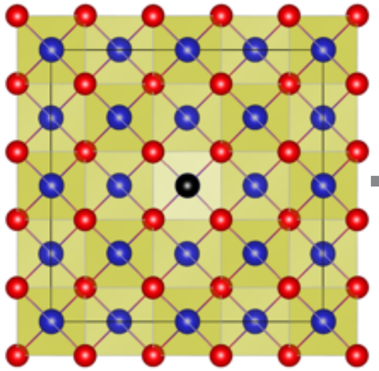
Density Functional Theory (DFT)

- Large supercells, charge states → expensive.
- Prior knowledge not utilized for new defect levels.

Predicting Impurity Behavior in Semiconductors

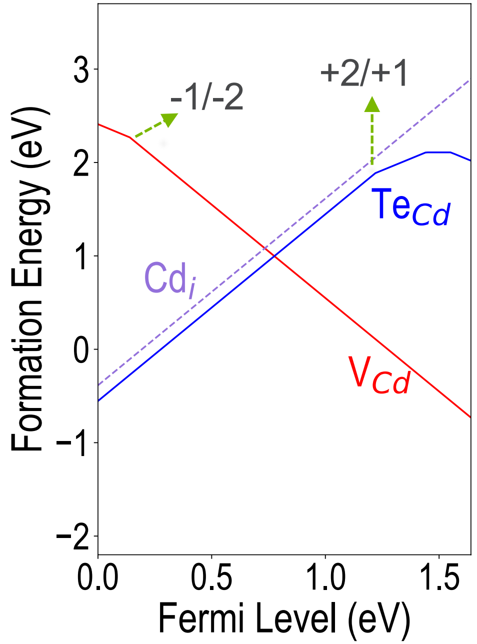
- “Computational Study of Pb Substitution in MAPbBr3”, ***Chem. Mater.*** (2019).
- "Machine-learned impurity level prediction for Cd chalcogenides”, ***npj Comput. Mater.*** (2020).
- “Universal ML Framework for Impurity Level Prediction in Group IV, III-V & II-VI Semiconductors”, *in prep*.
- “Accelerated Screening of Functional Atomic Impurities in Halide Perovskites using High-Throughput Computations and Machine Learning”, *in prep*.

Semiconductor + impurity



Density Functional Theory

- $E^f(q) = E(D^q) – E(bulk) + \sum n_i\mu_i + q(E_F+E_{vbm}) + E_{corr}$
- Impurity levels: $\varepsilon(q_1/q_2) = [\ E^f(q_1) - E^f(q_2)\ ] / (q_2-q_1)$

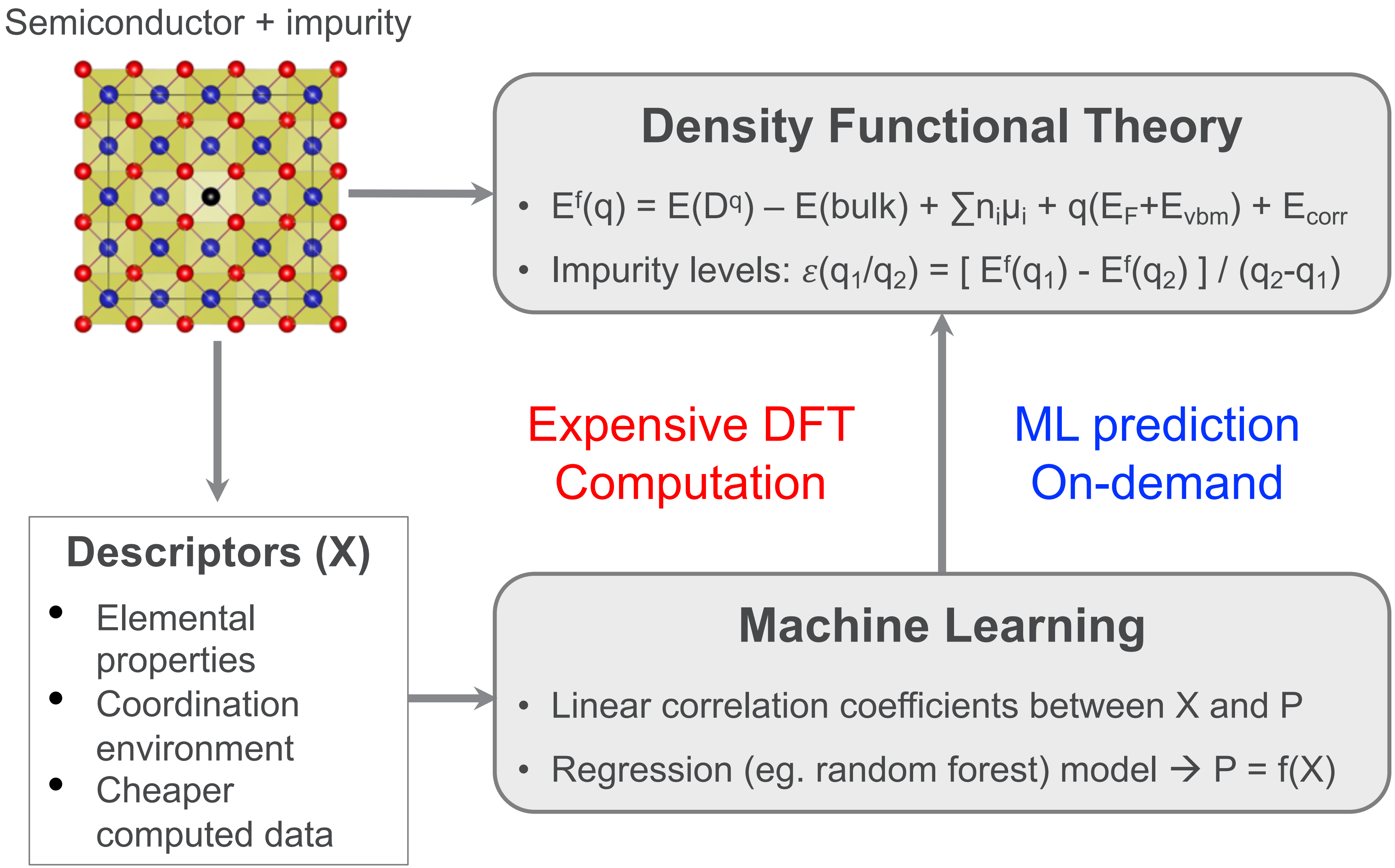


Impurity levels $\varepsilon(q_1/q_2)$:

Fermi energies (E_F) where defect transitions from one stable charge state (q_1) to another (q_2)

Predicting Impurity Behavior in Semiconductors

- “Computational Study of Pb Substitution in MAPbBr$_3$”, ***Chem. Mater.*** (2019).
- "Machine-learned impurity level prediction for Cd chalcogenides”, ***npj Comput. Mater.*** (2020).
- “Universal ML Framework for Impurity Level Prediction in Group IV, III-V & II-VI Semiconductors”, *in prep*.
- “Accelerated Screening of Functional Atomic Impurities in Halide Perovskites using High-Throughput Computations and Machine Learning”, *in prep*.



Semiconductor + impurity

Density Functional Theory

- $E^f(q) = E(D^q) – E(bulk) + \sum n_i\mu_i + q(E_F+E_{vbm}) + E_{corr}$
- Impurity levels: $\varepsilon(q_1/q_2) = [\ E^f(q_1) - E^f(q_2)\ ] / (q_2-q_1)$

Expensive DFT Computation

ML prediction On-demand

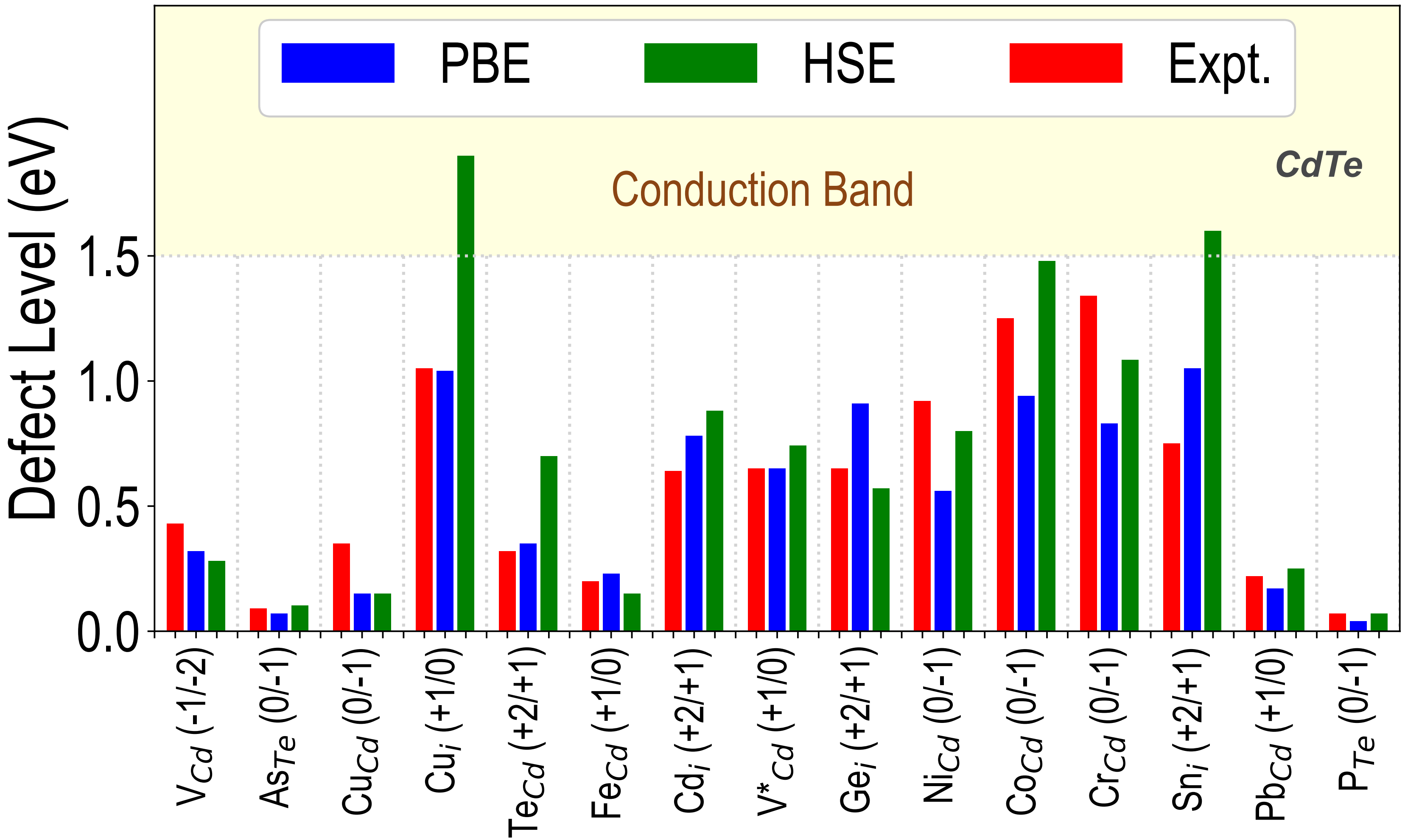
Descriptors (X)

- Elemental properties
- Coordination environment
- Cheaper computed data

Machine Learning

- Linear correlation coefficients between X and P
- Regression (eg. random forest) model → P = f(X)

$\varepsilon(q_1/q_2)$: DFT vs Experiments



Root mean square error (RMSE) between DFT (PBE) and experiments = 0.22 eV

ML models with similar accuracy can replace DFT computations and lead to accelerated screening.

"Machine-learned impurity level prediction for Cd chalcogenides", *npj Comput. Mater.* (2020).

Steps Involved in Training a Material → Property Regression Model

1. READ DATA: Labels, computed properties, descriptors.
2. SELECT ML TECHNIQUE: Random Forest / Kernel Ridge / LASSO / etc.
3. TRAIN DEFAULT MODEL: With a (for eg.) 70-30 training-test split, train a model using chosen technique.
4. HYPERPARAMETER OPTIMIZATION: Improve model by tuning every hyperparameter to minimize test prediction error.
5. CROSS-VALIDATION: Divide training data into n folds, tune hyperparameters to minimize cross-validation test error.
6. LEARNING CURVES: Model prediction performance vs training set size.
7. DEPLOY BEST MODEL: Make new predictions and discovery.

☒ The image part with relationship ID rId36 was not found in the file.

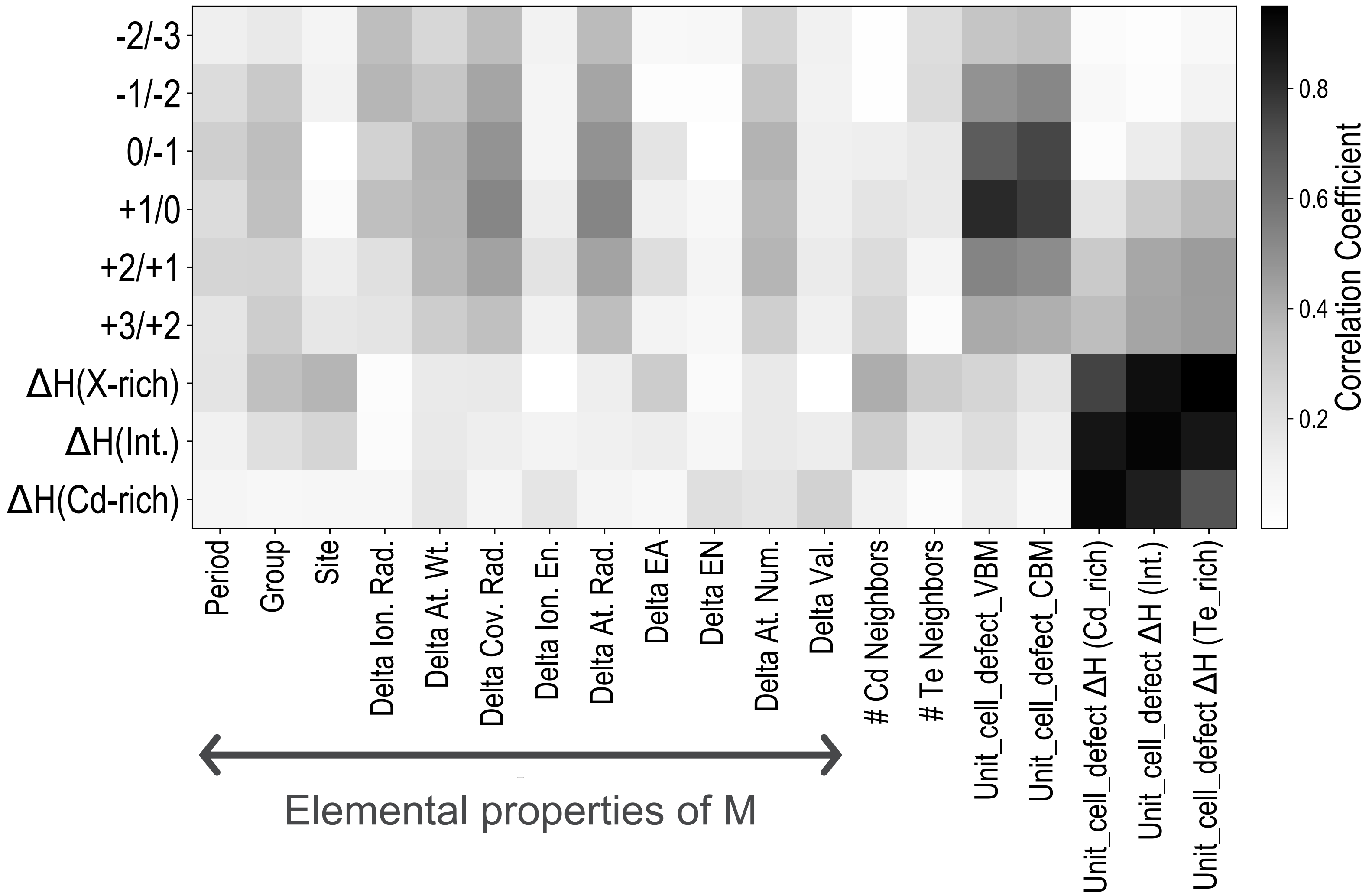
DFT Dataset for Machine Learning

CdX	Doping Site	M	∆H(Cd-rich)	∆H(Mod)	∆H(Te-rich)	(+3/+2)	(+2/+1)	(+1/0)	(0/-1)	(-1/-2)	(-2/-3)
CdTe	M_Te	N	2.09	2.48	2.88	-0.97	-0.67	-0.40	-0.08	1.44	1.90
CdTe	M_Te	O	1.17	1.17	1.43	-0.91	-0.65	-0.33	0.95	1.37	2.11
CdTe	M_i_Te_site	Rh	3.46	3.93	4.88	-0.52	0.05	0.56	1.12	1.83	2.14
CdTe	M_Te	Re	6.75	7.22	7.70	0.68	-0.49	-1.05	0.61	1.72	2.18
CdSe	M_Se	Si	2.59	3.27	5.21	-0.64	-0.33	0.30	0.70	1.10	2.21
CdTe	M_Cd	Be	0.58	0.46	0.34	-0.92	-0.65	-0.32	1.40	1.84	2.22
CdTe	M_i_Cd_site	F	1.72	1.48	1.24	-0.85	-0.56	-0.29	0.03	1.77	2.23
CdTe	M_i_Te_site	F	2.61	2.38	2.14	-0.84	-0.51	-0.22	0.09	1.80	2.26
CdSe	M_Cd	Cu	2.49	1.23	1.05	-0.88	-0.51	-0.14	0.36	1.79	2.33
CdSe	M_Se	Os	5.88	6.53	7.17	-0.55	-0.55	0.05	1.44	1.90	2.36
CdSe	M_i_Cd_site	F	1.95	1.63	1.31	-0.83	-0.51	-0.21	0.12	1.82	2.38
CdTe	M_i_Cd_site	Hg	1.78	1.47	1.76	-0.65	-0.15	0.29	1.66	2.06	2.42
CdTe	M_i_old	Cu	2.50	1.94	2.22	-0.74	-0.40	1.26	1.72	2.08	2.43
CdSe	M_i_Se_site	Cl	3.68	3.36	3.03	-0.65	-0.32	0.05	0.38	1.94	2.51
CdTe	M_Cd	Sr	1.06	1.06	1.06	-0.64	-0.41	-0.10	1.66	2.11	2.52
CdS	M_i_S_site	S	5.23	4.57	3.91	-0.06	-0.05	0.78	1.13	1.59	2.69
CdS	M_i_Cd_site	S	4.81	4.15	3.49	-0.56	-0.27	0.56	1.02	1.50	2.71
CdS	M_Cd	O	7.52	6.22	5.56	-0.87	-0.30	0.10	0.47	2.11	2.75
CdS	M_i_old	S	4.65	3.99	3.33	-0.65	0.83	0.83	1.48	1.95	2.75
CdSe	M_i_Se_site	Pd	2.22	1.58	2.02	-0.62	-0.19	0.27	1.56	2.09	2.89
CdS	M_Cd	S	6.04	4.73	3.42	-0.80	0.46	0.83	1.32	2.32	2.94
CdSe	M_Cd	Pb	0.76	0.73	0.74	-0.68	-0.16	0.34	1.67	2.19	2.97
CdS	M_i_old	Pt	3.03	2.38	3.25	-0.43	0.23	0.72	1.98	2.57	3.38
CdS	M_S	Se	0.19	0.20	0.21	-0.77	-0.42	-0.12	1.95	2.53	3.39

Dataset 1: Formation Enthalpies (∆H) → 945 impurities → ***945 points*** (each)

Dataset 2: Charge Transition Levels ($\varepsilon(q_1/q_2)$) → 381 impurities → ***2286 points*** (combined)

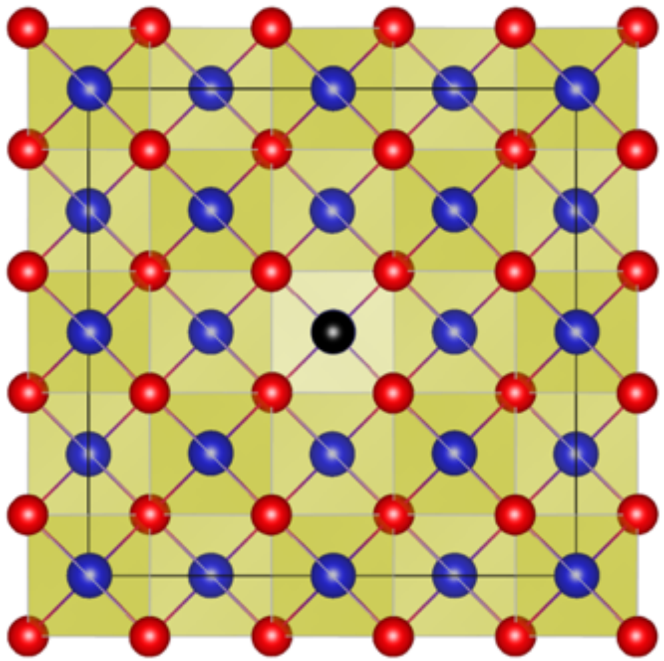
Correlation between descriptors & properties



↔ Elemental properties of M

↔ Unit cell defect calculation

DATASET: 381 impurities in CdTe, CdSe and CdS.



magnitude cheaper calculation

→ *unit cell defect calculation descriptors*

The image part with relationship ID rId36 was not found in the file.

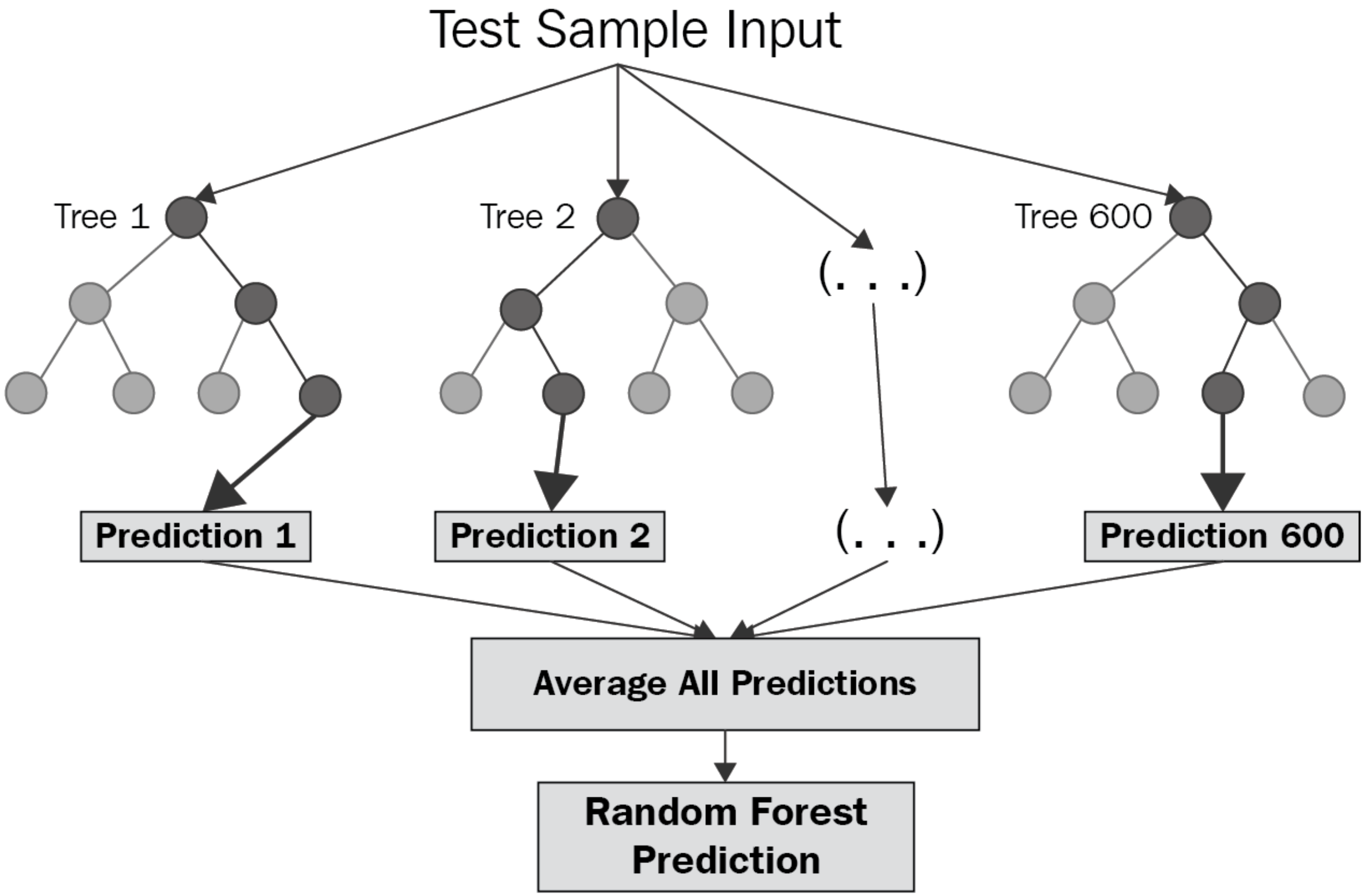
Descriptors for Machine Learning

CdX	Doping Site	M	Delta Ion. En.	Delta At. Rad.	Delta EN	Delta Val.	…	Unit_cell_def CBM	Unit_cell_def ∆H(Mod)	Unit_cell_def ∆H(X-rich)
CdTe	M_Te	N	-347.5	0.061	-0.71	-1		1.156	1.048	0.812
CdTe	M_Te	O	-371.9	0.411	-0.76	-1		1.316	1.173	1.083
CdTe	M_i_Te_site	Rh	-448.9	0.861	-0.87	-1		1.647	1.942	1.965
CdTe	M_Te	Re	-464.7	0.991	-0.87	2		2.140	2.326	2.481
CdSe	M_Se	Si	-492	1.181	-0.9	-1		2.000	3.000	3.000
CdTe	M_Cd	Be	31.7	-0.369	-0.12	0		2.718	0.770	0.651
CdTe	M_i_Cd_site	F	-130	0.111	-0.38	0		2.907	0.113	0.114
CdTe	M_i_Te_site	F	-277.9	0.481	-0.69	0		2.995	0.864	0.865
CdSe	M_Cd	Cu	-318.2	0.661	-0.74	0		2.995	1.276	1.277
CdSe	M_Se	Os	-364.8	0.731	-0.8	0		2.897	1.807	1.808
CdSe	M_i_Cd_site	F	-67.1	-0.509	0.35	1		3.300	3.320	2.846
CdTe	M_i_Cd_site	Hg	-290.1	-0.059	-0.08	1		3.547	0.954	1.214
CdTe	M_i_old	Cu	-288.9	-0.079	0.12	1		3.032	0.757	0.856
CdSe	M_i_Se_site	Cl	-309.4	0.171	0.09	1		3.327	0.723	0.920
CdTe	M_Cd	Sr	-278.4	0.221	0.35	1		2.854	0.999	0.796
CdS	M_i_S_site	S	218.7	-0.579	0.86	2		3.506	5.499	5.025
CdS	M_i_Cd_site	S	-81.3	-0.169	0.21	2		3.408	1.755	1.281
CdS	M_Cd	O	-105.6	-0.119	0.32	2		3.541	0.897	0.737
CdS	M_i_old	S	-159.1	0.131	0.27	2		3.491	1.076	1.077
CdSe	M_i_Se_site	Pd	-152.2	0.261	0.64	2		3.457	0.980	0.981
CdS	M_Cd	S	534.6	-0.569	1.35	3		2.140	4.523	3.970
CdSe	M_Cd	Pb	144	-0.209	0.5	3		2.945	2.574	2.045
CdS	M_i_old	Pt	79.3	-0.099	0.47	3		3.116	1.953	1.502
CdS	M_S	Se	-34	0.101	0.36	3		3.375	4.104	4.342

Descriptor Set 1: Elemental Properties (14 features)
Descriptor Set 2: Unit Cell Defect Properties (5 features)
Descriptor Set 3: Elemental + Unit Cell (19 features)

☒ The image part with relationship ID rId36 was not found in the file.

Random Forest Regression



Test Sample Input

Tree 1

Tree 2

(. . .)

Tree 600

Prediction 1

Prediction 2

(. . .)

Prediction 600

Average All Predictions

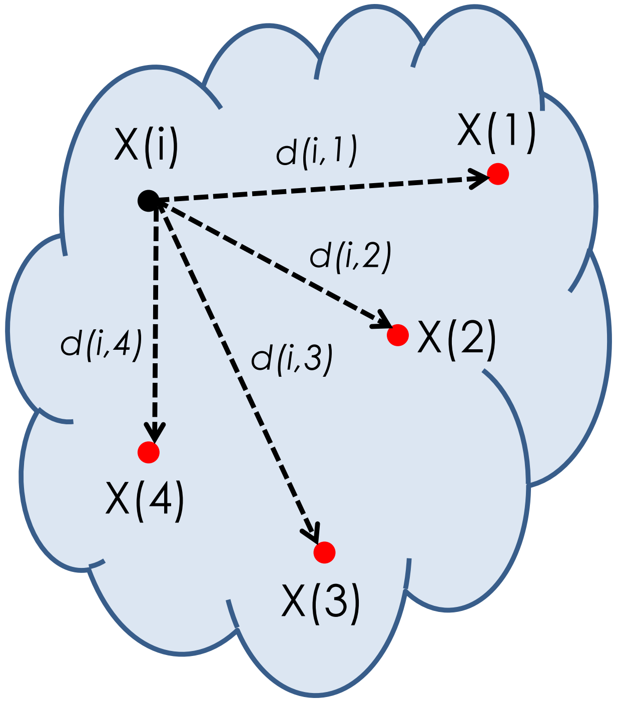
Random Forest Prediction

https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f

The image part with relationship ID rId36 was not found in the file.

Kernel Ridge Regression

Chemical Space



$X(i) = \{x_1, x_2, x_3 \ldots x_m\}$

Similarity-based regression

Measure of Similarity: Euclidean Distance

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + \ldots + (x_{im} - x_{jm})^2}$$

Property = Weighted sum of Gaussians

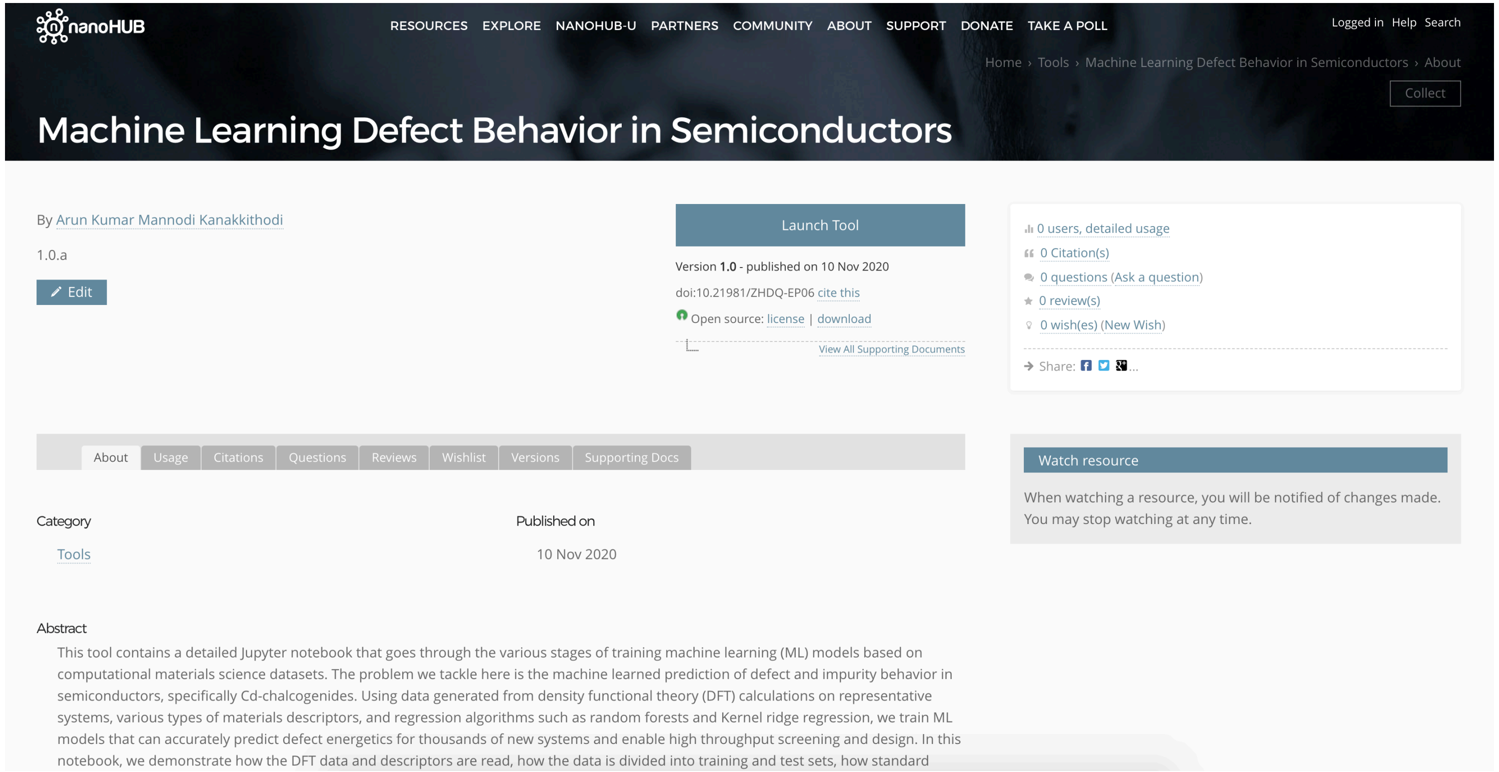
$$f(i) = \sum_{k=1}^{N} a_k \, . \, \exp\left(-\frac{1}{2\sigma^2} \, . \, [d(i, i_k)]^2\right)$$

A. Mannodi-Kanakkithodi et al., Sci. Rep. ***2016****.*
T. D. Huan et al., Phys. Rev. B. ***2015****.*

The image part with relationship ID rId36 was not found in the file.

Launching the Jupyter tool on Nanohub

- Login to your account on Nanohub.
- Go to the following link: https://nanohub.org/resources/mldefect/
- Click on “Launch Tool”; it may take a minute or so to load.
- You should be able to see the following Jupyter notebook.



Machine Learning Defect Behavior in Semiconductors

By Arun Kumar Mannodi Kanakkithodi

1.0.a

Version **1.0** - published on 10 Nov 2020

doi:10.21981/ZHDQ-EP06 cite this

Open source: license | download

View All Supporting Documents

0 users, detailed usage

0 Citation(s)

0 questions (Ask a question)

0 review(s)

0 wish(es) (New Wish)

When watching a resource, you will be notified of changes made. You may stop watching at any time.

Category

Tools

Published on

10 Nov 2020

Abstract

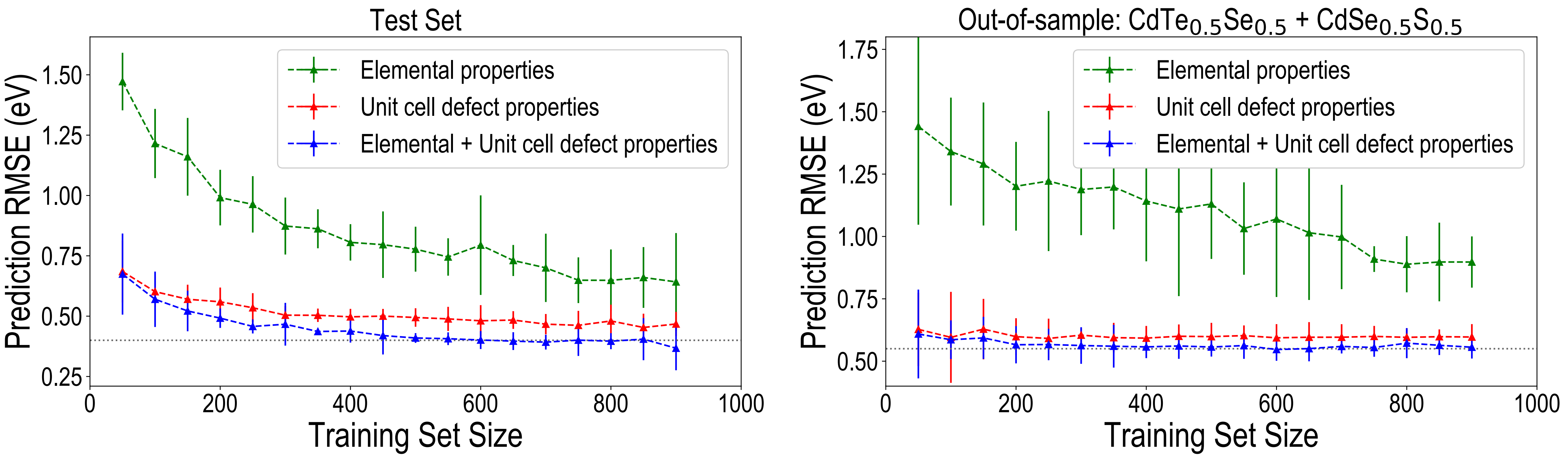
This tool contains a detailed Jupyter notebook that goes through the various stages of training machine learning (ML) models based on computational materials science datasets. The problem we tackle here is the machine learned prediction of defect and impurity behavior in semiconductors, specifically Cd-chalcogenides. Using data generated from density functional theory (DFT) calculations on representative systems, various types of materials descriptors, and regression algorithms such as random forests and Kernel ridge regression, we train ML models that can accurately predict defect energetics for thousands of new systems and enable high throughput screening and design. In this notebook, we demonstrate how the DFT data and descriptors are read, how the data is divided into training and test sets, how standard

Arun Kumar Mannodi Kanakkithodi (2020), "Machine Learning Defect Behavior in Semiconductors," https://nanohub.org/resources/mldefect. (DOI: 10.21981/ZHDQ-EP06).

The image part with relationship ID rId36 was not found in the file.

Random Forest: Formation Enthalpy

Models trained for ΔH (Cd-rich) on CdTe+CdSe+CdS data

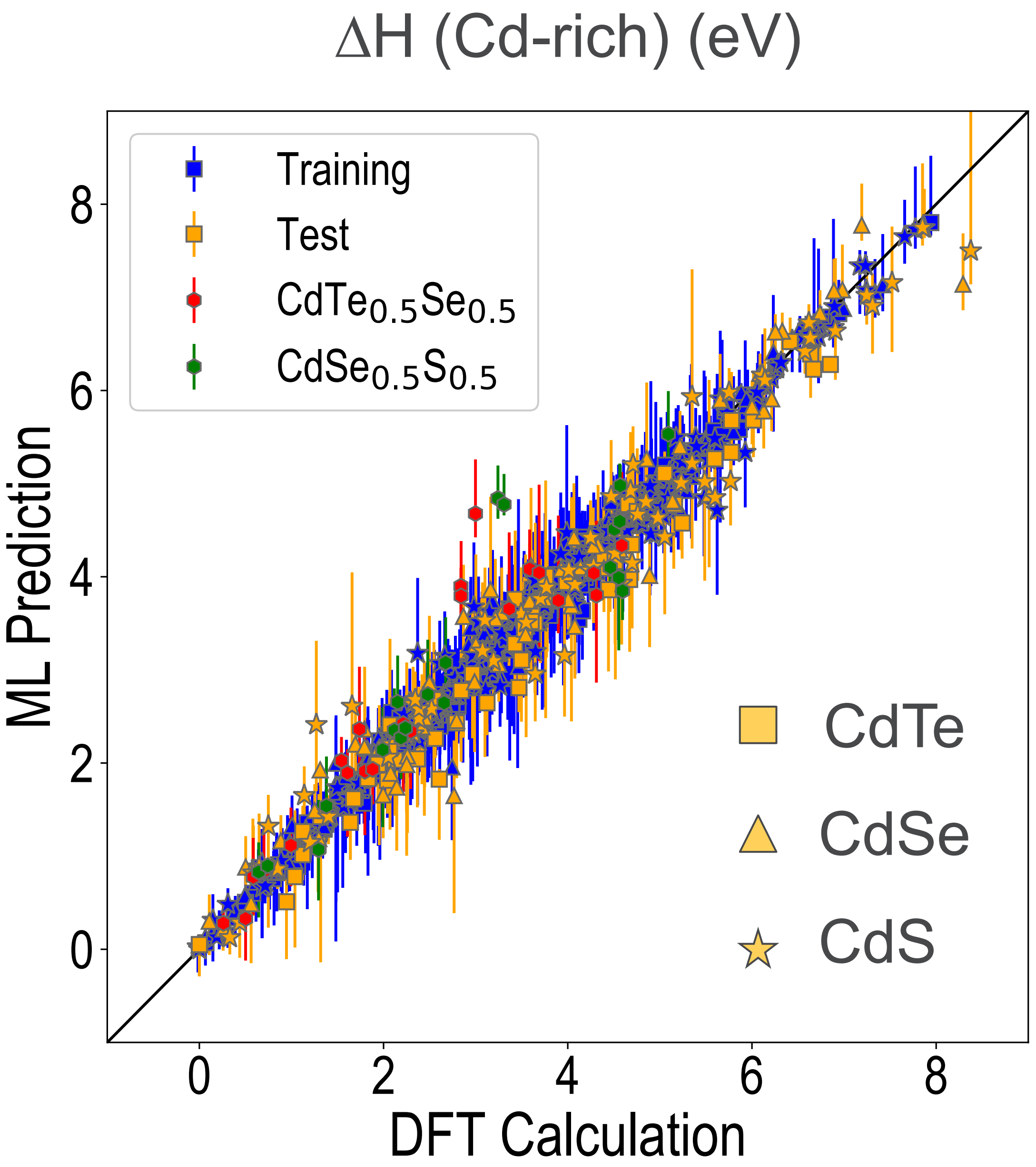


Elemental properties + unit cell defect properties lead to best models.

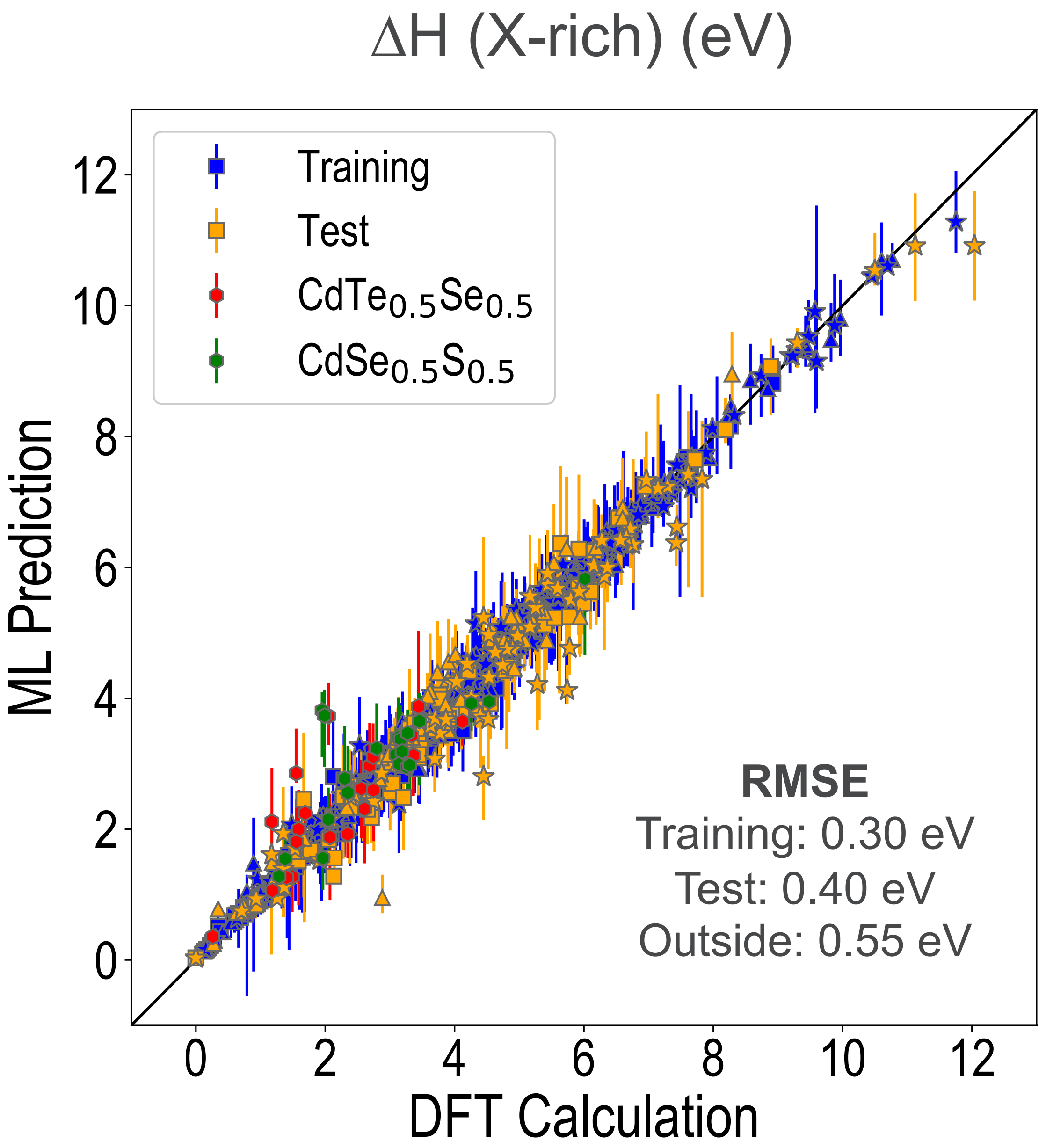
"Machine-learned impurity level prediction for Cd chalcogenides", *npj Comput. Mater.* (2020)

The image part with relationship ID rId36 was not found in the file.

Random Forest: Formation Energy



ΔH (Cd-rich) (eV)



ΔH (X-rich) (eV)

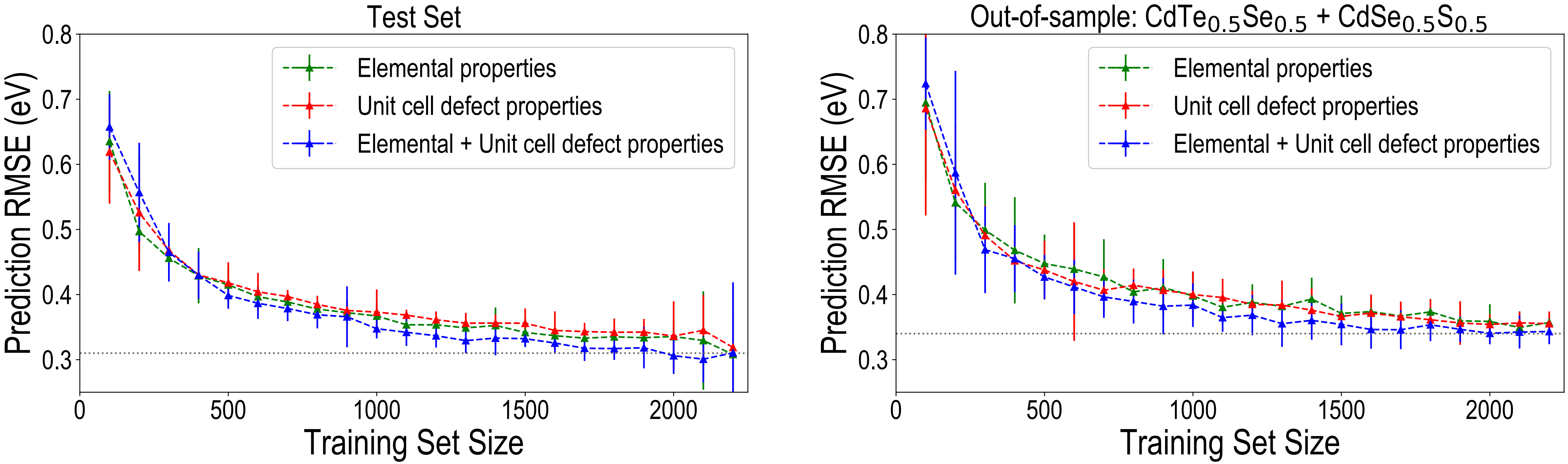
RMSE
Training: 0.30 eV
Test: 0.40 eV
Outside: 0.55 eV

th

found in the file.

Random Forest: Impurity Levels

Models trained for $\varepsilon(q_1/q_2)$ on CdTe+CdSe+CdS data



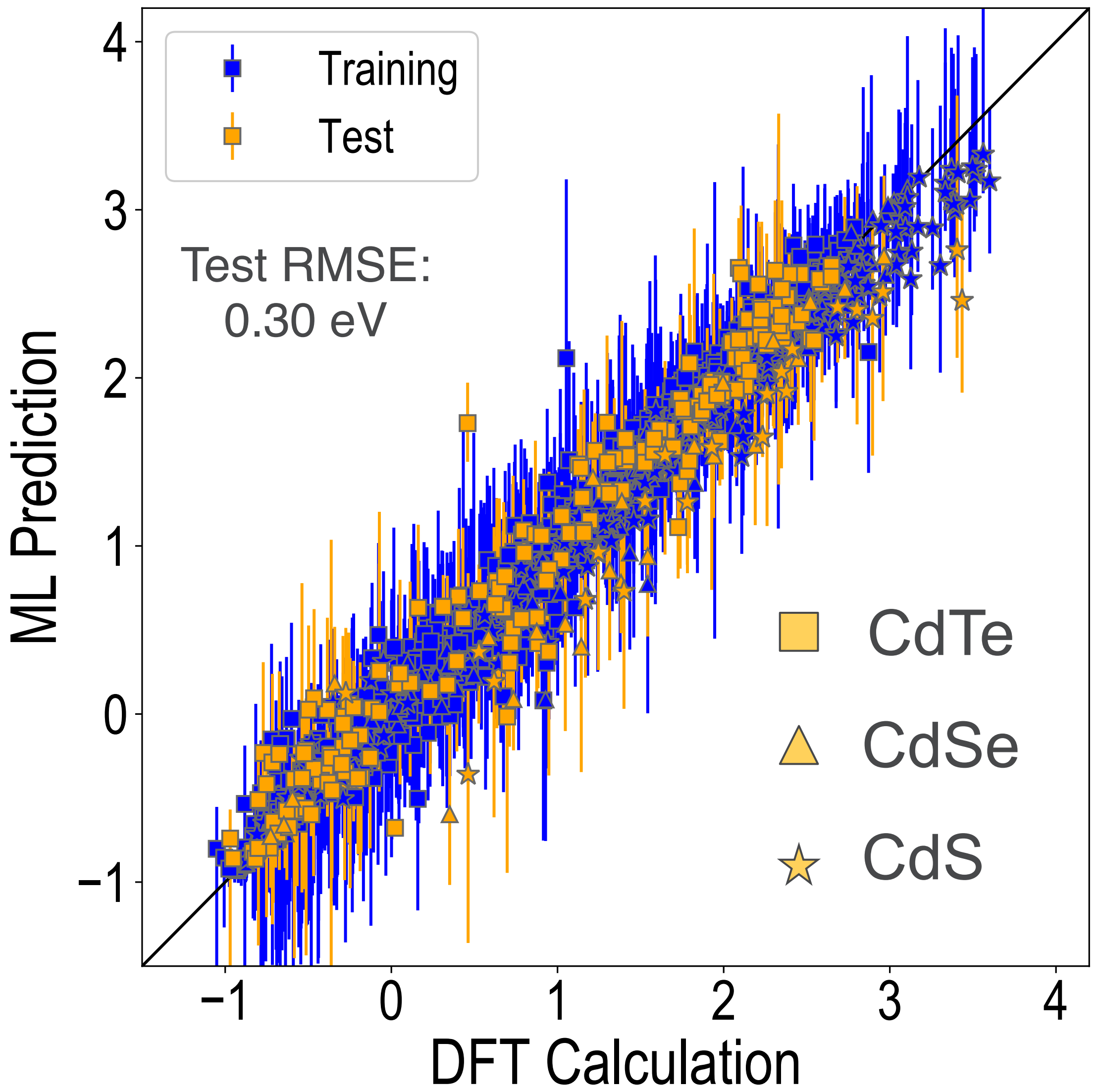
Elemental properties + unit cell defect properties lead to best models.

"Machine-learned impurity level prediction for Cd chalcogenides", *npj Comput. Mater.* (2020)

The image part with relationship ID rId36 was not found in the file.

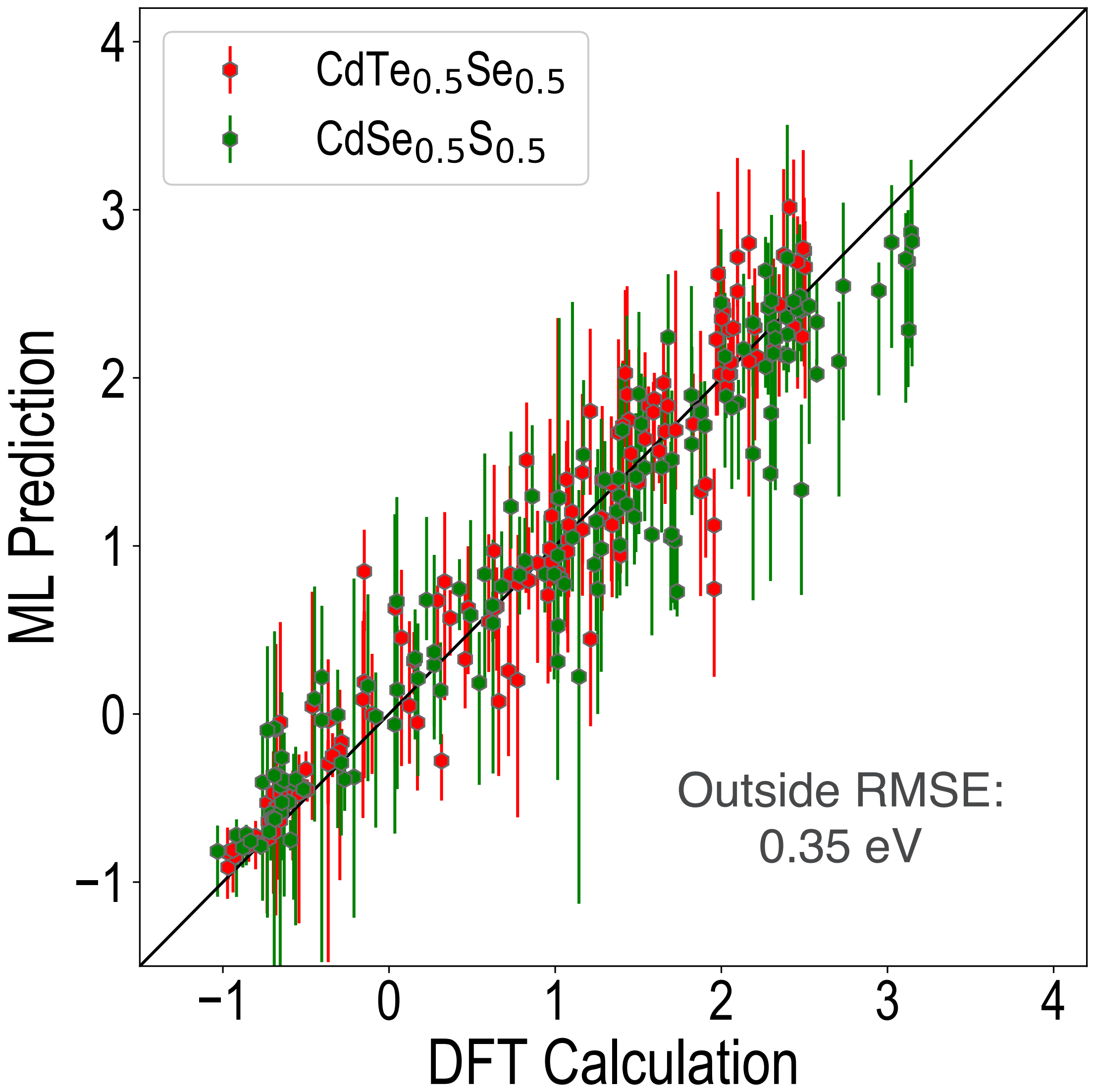
Random Forest: Transition Levels

RF Model to predict $\varepsilon(q_1/q_2)$ (eV)



Test RMSE: 0.30 eV

Out-of-sample prediction

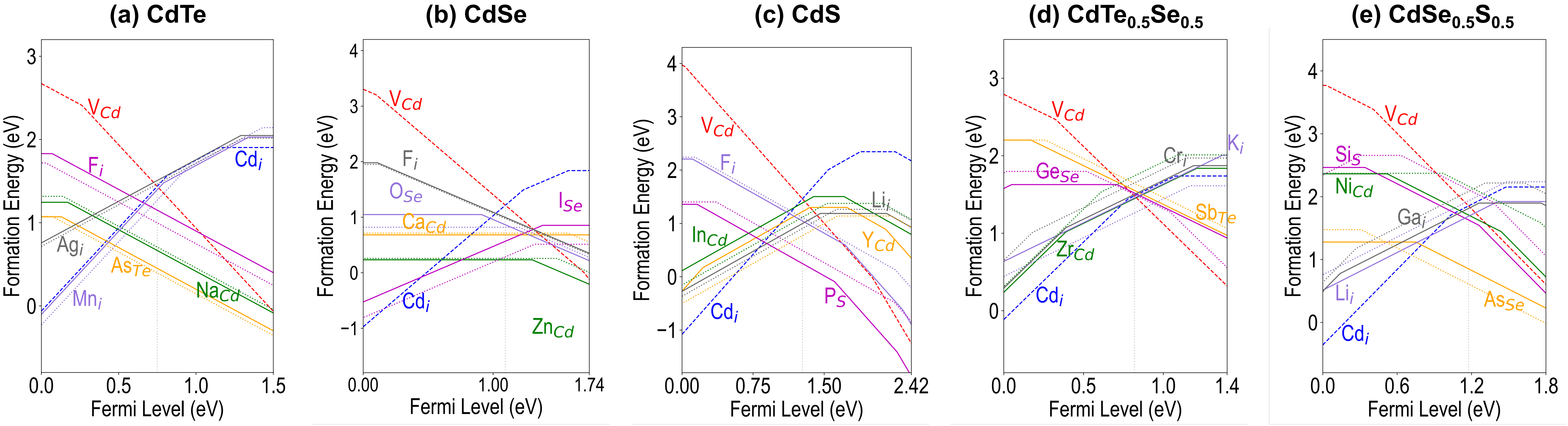


Outside RMSE: 0.35 eV

h

found in the file.

Impurity Formation Energies: DFT vs ML



ML models trained on dataset of 381 impurities in CdTe, CdSe & CdS are used to predict complete formation energies for 1827 impurities in 5 compounds; in theory applicable to any impurity in any Cd-Te-Se-S compound.

The image part with relationship ID rId36 was not found in the file.

Extensions of current work

New semiconductors, impurity atoms, structures (Wurtzite vs ZB)

II-VI

A	B
Cd	O
Zn	S
	Se
	Te

8 candidates

III-V

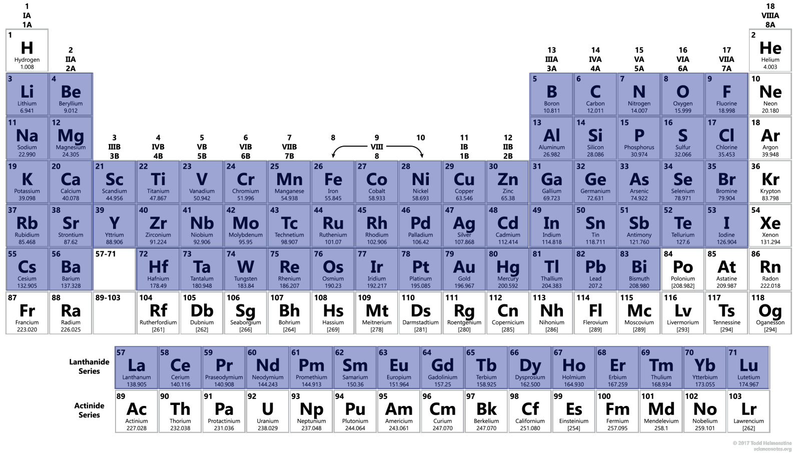
A	B
B	N
Al	P
Ga	As
In	Sb

16 candidates

IV-IV

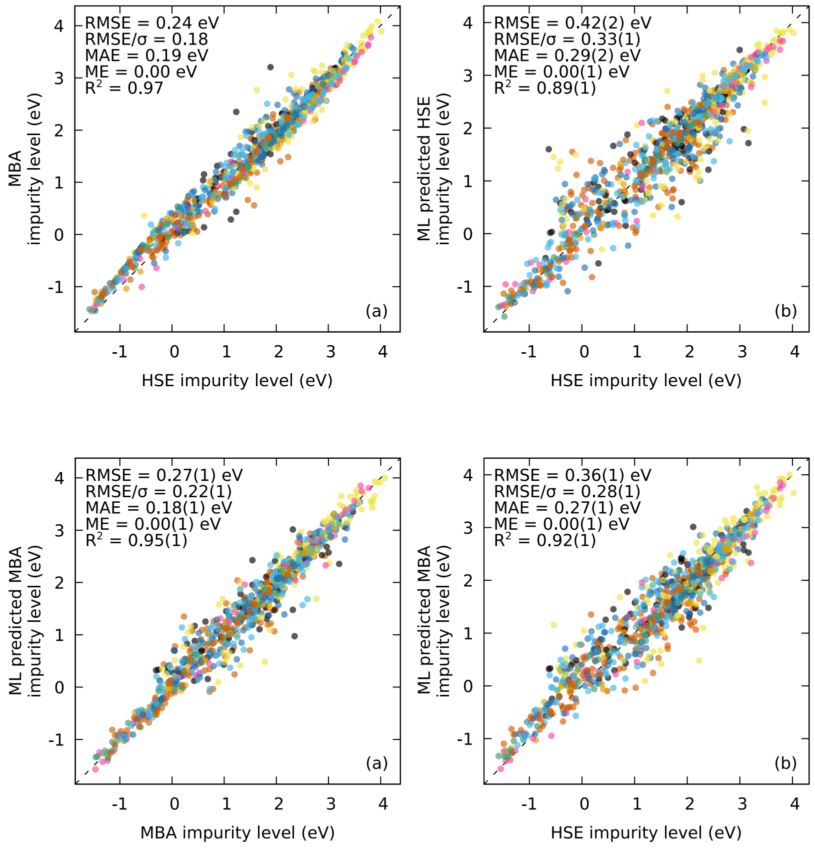
A	B
C	C
Si	Si
Ge	Ge
Sn	Sn

10 candidates


H He
Li Be B C N O F Ne
Na Mg Al Si P S Cl Ar
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te I Xe
Cs Ba 57-71 Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi Po At Rn
Fr Ra 89-103 Rf Db Sg Bh Hs Mt Ds Rg Cn Nh Fl Mc Lv Ts Og
Lanthanide Series La Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb Lu
Actinide Series Ac Th Pa U Np Pu Am Cm Bk Cf Es Fm Md No Lr

A.M.K et al., *in prep.*

Advanced theory: Modified band alignment (MBA) & HSE



MBA vs HSE

ML HSE vs HSE

ML MBA vs MBA

ML MBA vs HSE

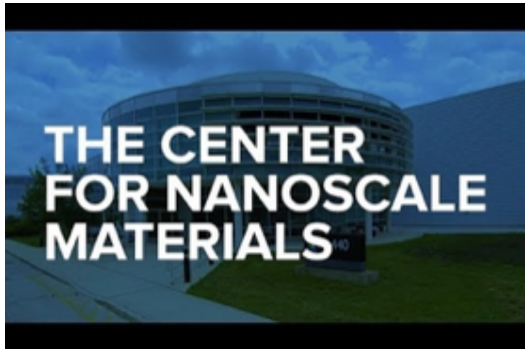
M.P. Polak et al., *under review.*

The image part with relationship ID rId36 was not found in the file.

Acknowledgements

Argonne LDRD: Office of Science #DE-AC02-06CH11357

EERE PVRD: SunShot program #DOE DEEE005956.



NERSC

National Energy Research Scientific Computing Center

U.S. DEPARTMENT OF ENERGY

Office of Science

Argonne NATIONAL LABORATORY

CONTACT: mannodiarun@anl.gov

The image part with relationship ID rId36 was not found in the file.

THANK YOU