Supplementary Information (SI) for Journal of Materials Chemistry C.

Supporting Information for

Defect Formation in $CsSnI_3$ from Density Functional Theory and Machine Learning

Chadawan Khamdang[1] and Mengen Wang[1, 2*]

[1]*Department of Electrical and Computer Engineering, State University of New York at Binghamton, Binghamton, New York 13902, United States*

[2] *Materials Science and Engineering Program, State University of New York at Binghamton, Binghamton, New York 13902, United States*

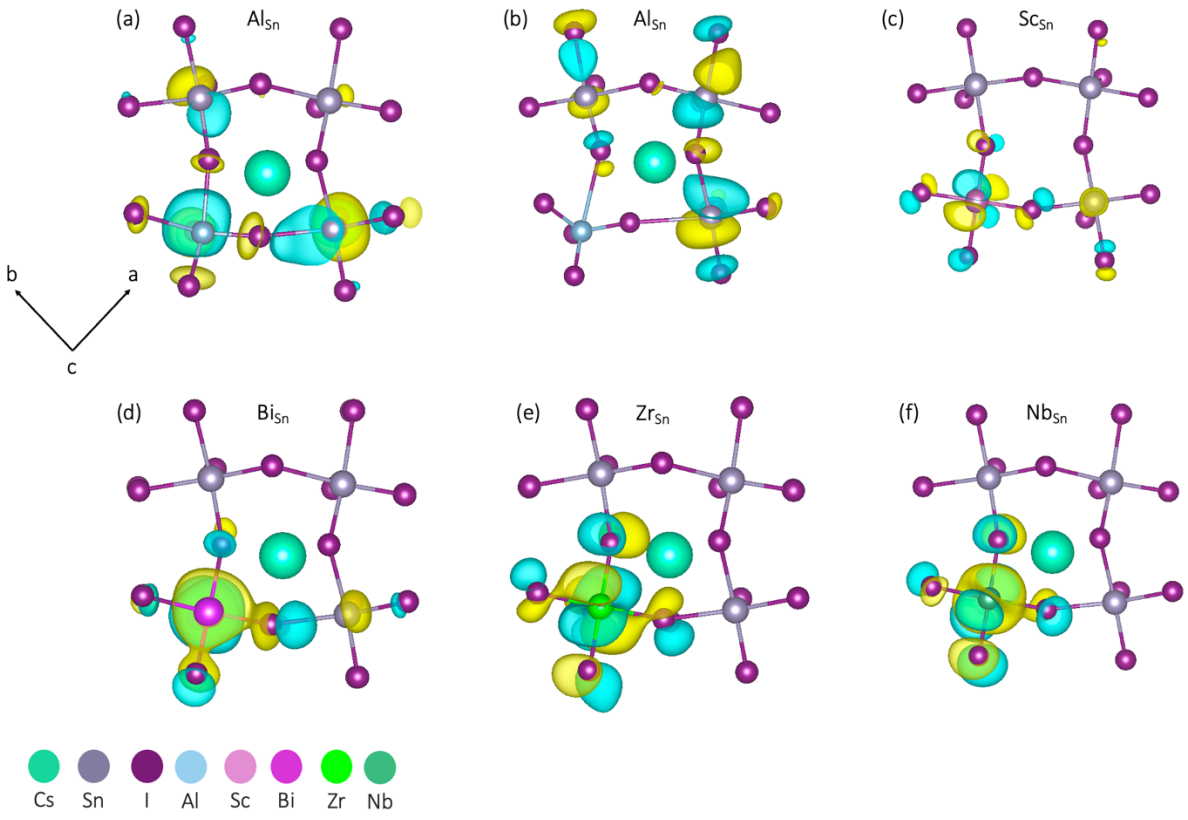

Figure S1. Electron localization in X_{Sn} in the neutral charge state including (a) the ground state and (b) the metastable structure of Al_{Sn}, followed by (c)–(f) Sc_{Sn}, Bi_{Sn}, Zr_{Sn}, and Nb_{Sn}. The isosurfaces in cyan and yellow represent the wave function of the localized electron in the defect.

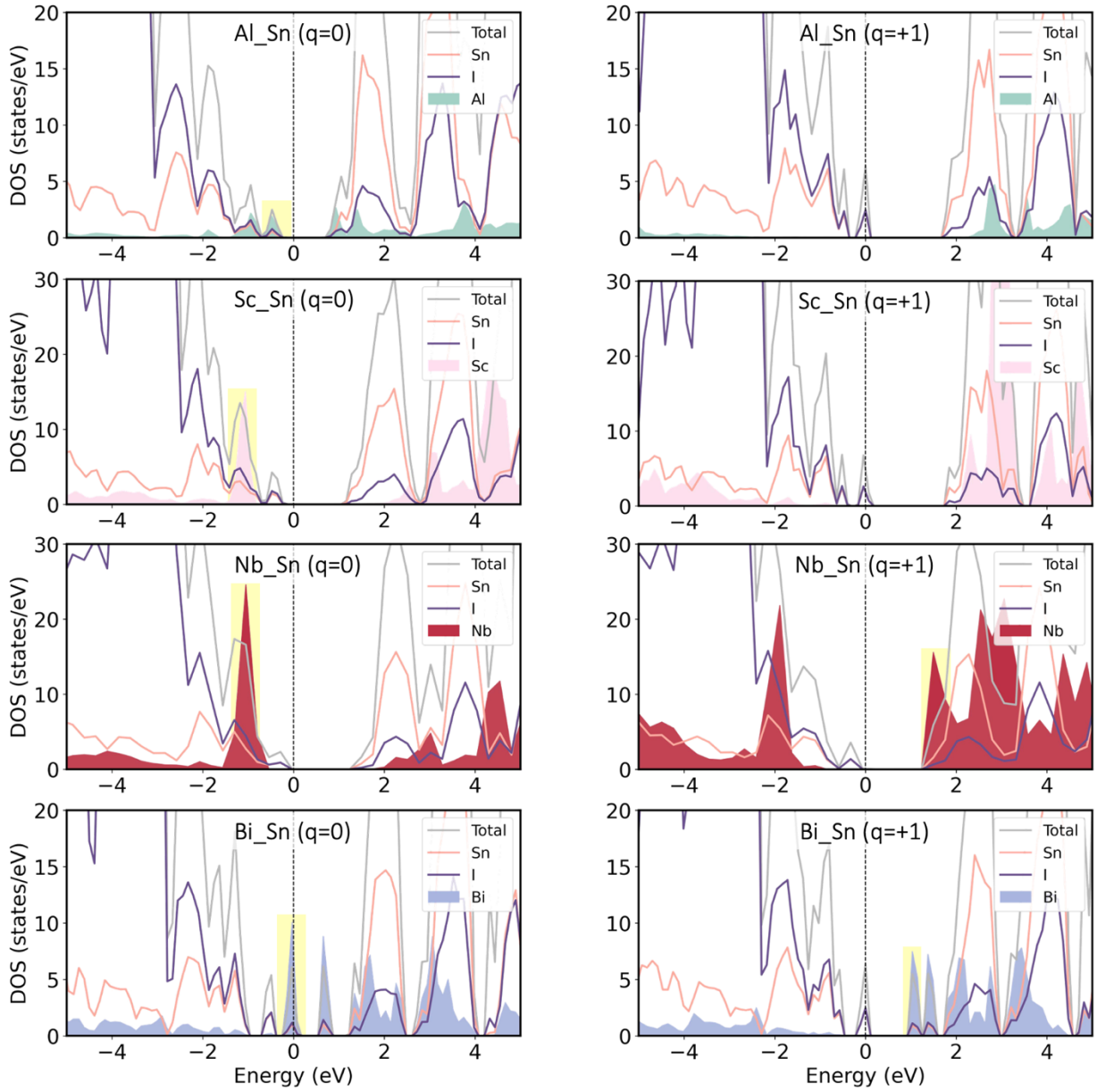

Figure S2. The projected density of states (PDOS) for Al_{Sn}, Sc_{Sn}, Nb_{Sn}, and Bi_{Sn} in the *q*=0 and *q*=+1 charge states, with the Fermi level set to zero (dashed line). The yellow shading highlights the localized defect states. The PDOS of the dopants was scaled by a factor of five.

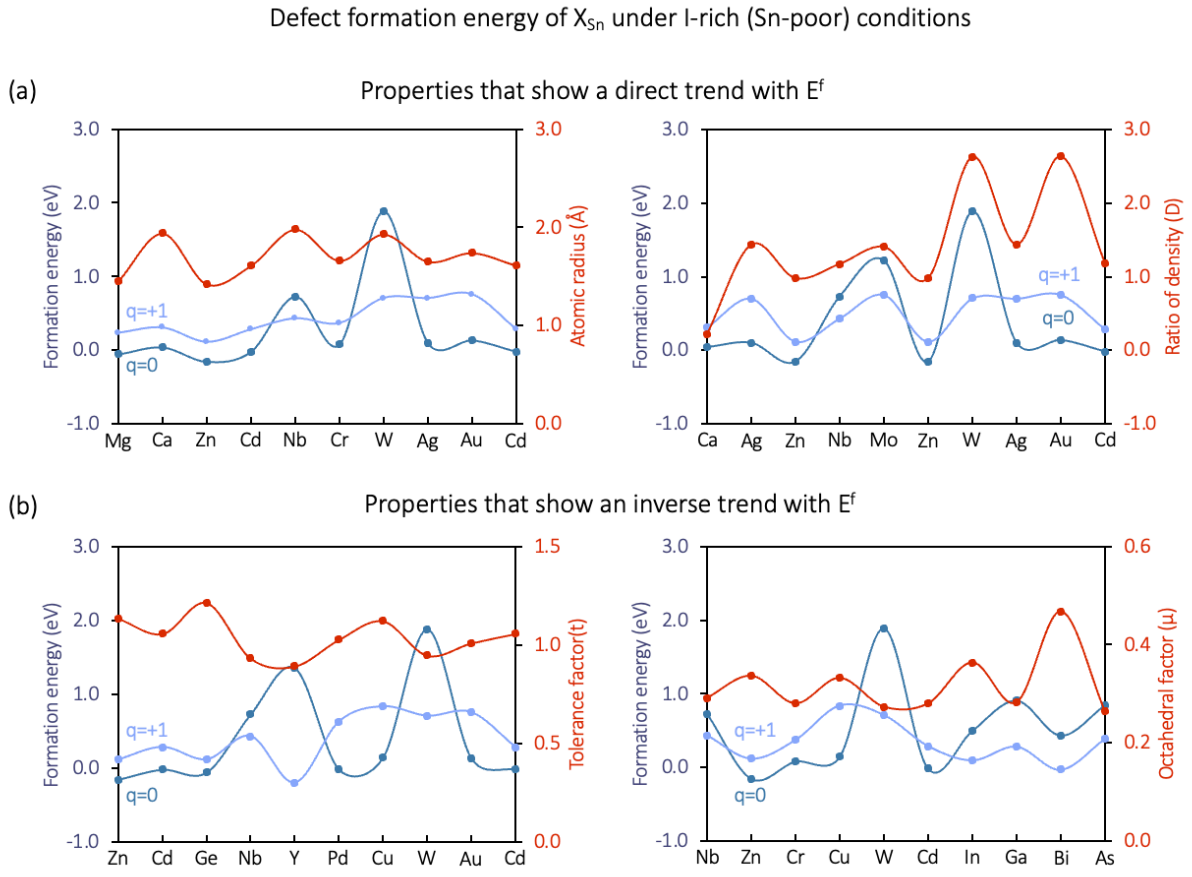

Figure S3. Trends between elemental properties and formation energies under the I-rich condition. (a) Elemental properties that show a direct trend with formation energies, and (b) Elemental properties that show an inverse trend with formation energies.

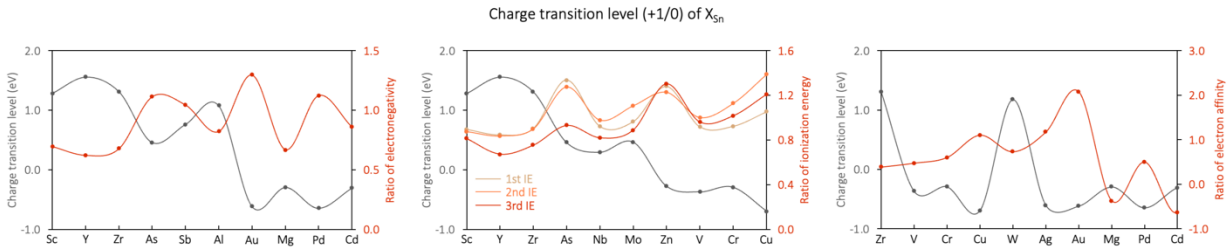

Figure S4. Trends between elemental properties and the charge transition level [CT(+1/0)].

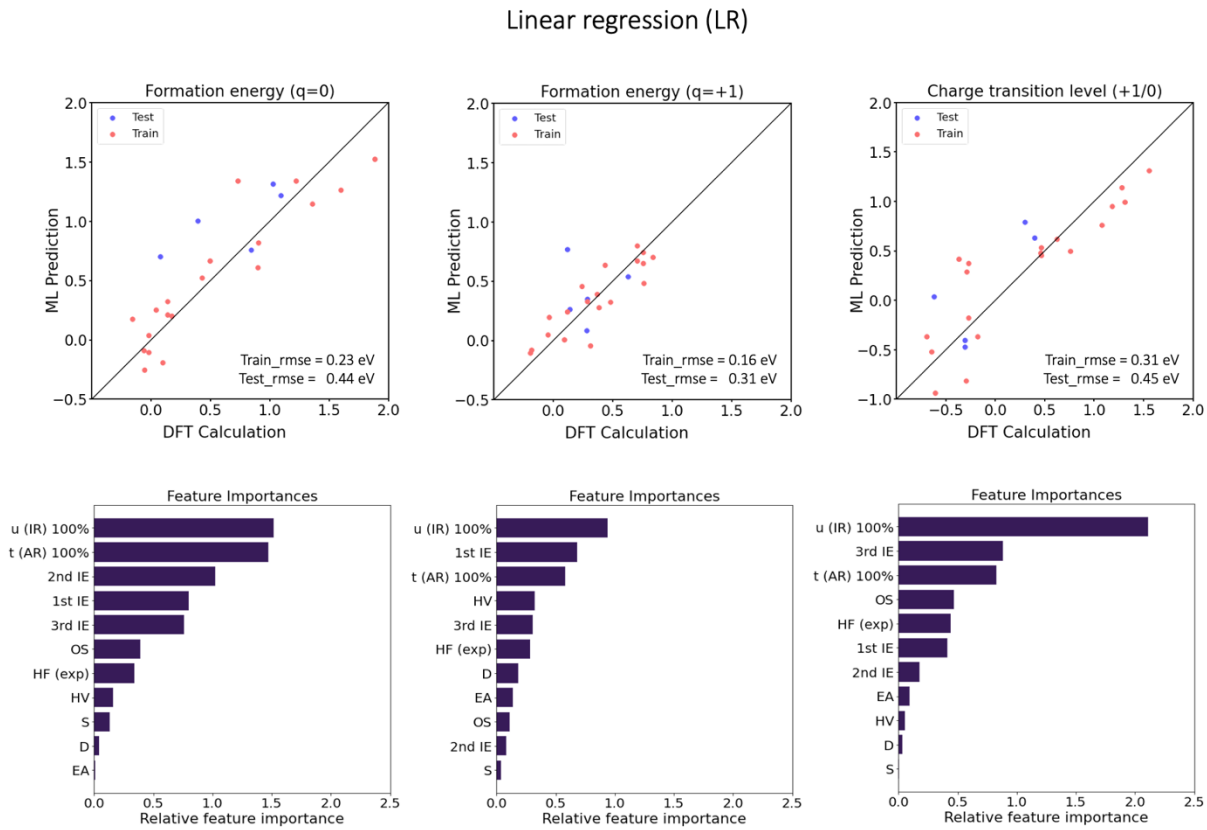

Figure S5. The parity plot and feature importance from the linear regression model training for the formation energy and charge transition levels of X_{Sn}.

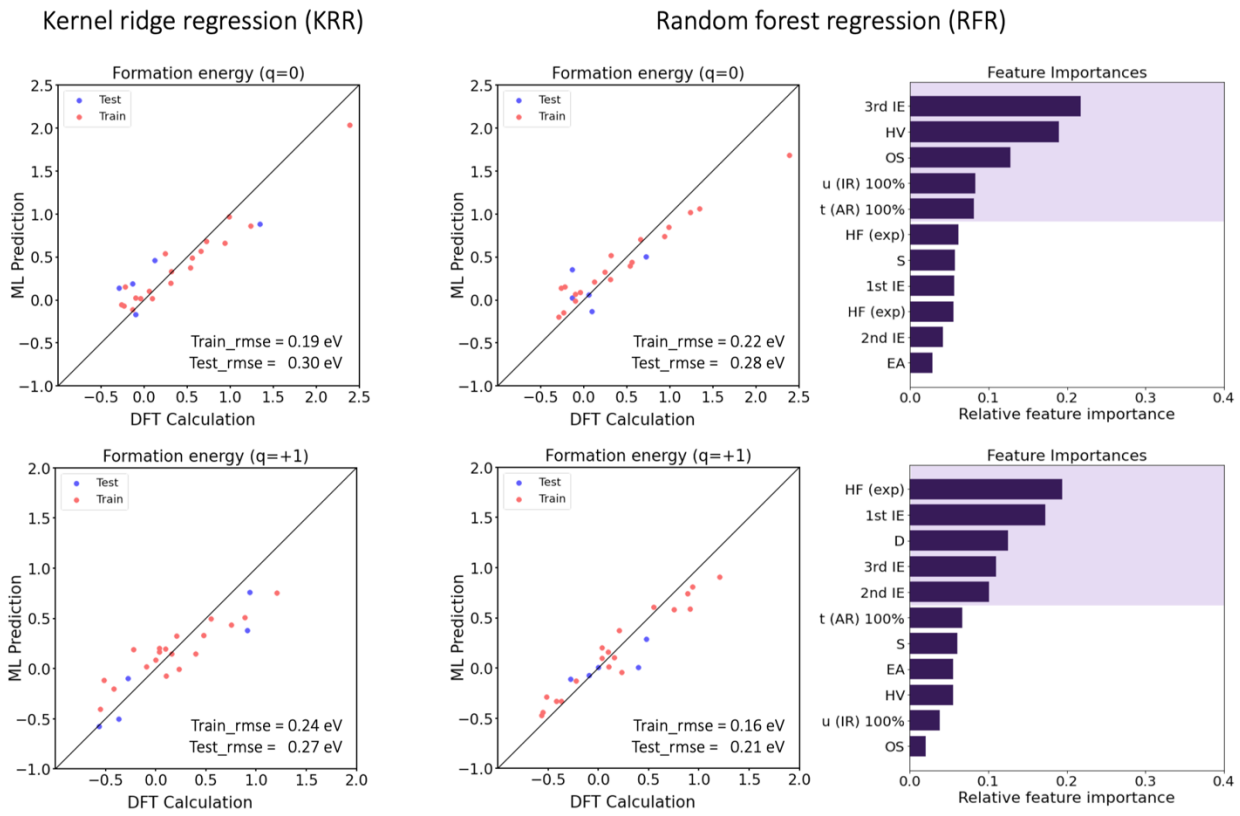

Figure S6. Parity plots from kernel ridge regression and random forest regression, along with the feature importances for formation energies under the I-poor condition of defect X_{Sn}. The top five most important features are highlighted.

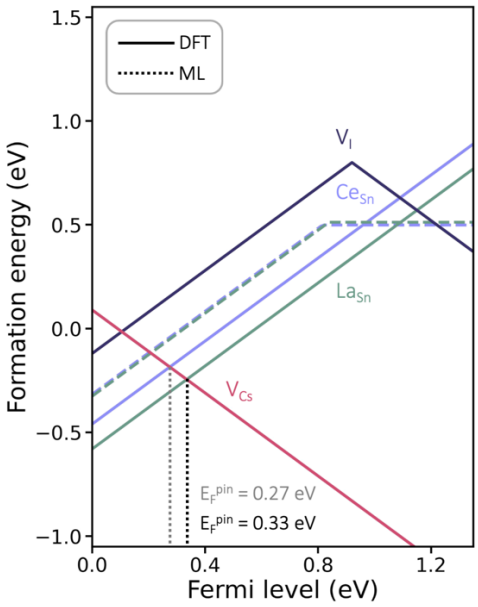

Figure S7. The defect formation energy diagrams for the calculated (solid lines) and ML-predicted (dashed lines) Ce_{Sn} and La_{Sn}.

Table S1. Optimized hyperparameters in ML regression models

Regression	Properties (train/test)	Best Hyperparameters
LR	E^f *q=0* (0.23/0.44)	fit_intercept=False, n_jobs=2
	E^f *q=+1* (0.16/0.31)	fit_intercept=False, n_jobs=2
	CT (0.31/0.45)	fit_intercept=False, n_jobs=2
GPR	E^f *q=0* (0.21/0.32)	alpha=0.1, kernel=Matern(length_scale=2.15e+03, nu=2), n_restarts_optimizer=250
	E^f *q=+1* (0.18/0.23)	alpha=1.0, kernel=DotProduct(sigma_0=1), n_restarts_optimizer=100
	CT (0.16/0.31)	alpha=0.1, kernel=Matern(length_scale=0.01, nu=2), n_restarts_optimizer=200
KRR	E^f *q=0* (0.15/0.25)	alpha=0.1, kernel=RationalQuadratic(alpha=4.64, length_scale=1.67)
	E^f *q=+1* (0.16/0.19)	alpha=0.1, kernel=ExpSineSquared(length_scale=0.599, periodicity=35.9)
	CT (0.15/0.27)	alpha=0.1, kernel=RationalQuadratic(alpha=21.5, length_scale=1.67)
RFR	E^f *q=0* (0.17/0.20)	max_depth=20, max_features=7
	E^f *q=+1* (0.11/0.15)	max_depth=40, max_features=3, n_estimators=50
	CT (0.17/0.22)	max_depth=50, max_features=7

Table S2. Prediction of formation energy for out-of-sample dopants on the Sn site under the I-poor condition.

	I-poor conditions	
	q=0 (eV)	*q=+1* (eV)
La	0.5122	-0.3258
Ce	0.4985	-0.3167
Pr	0.5019	-0.3167
Sr	0.1333	-0.1225
Ba	0.1588	-0.1030
Hf	0.4372	-0.0588
Ta	0.6371	0.3994
Re	0.6436	0.7057
Fe	0.0749	0.2767
Ru	0.1814	0.5201
Os	-0.0938	0.6638
Co	0.0879	0.2965
Rh	-0.0132	0.4624
Ir	0.4915	0.6952
Pt	0.5493	0.8346
B	0.3560	0.3623
Tl	0.2223	0.2469
N	0.7218	0.3661
P	0.9116	0.3577
O	0.4781	0.5142
S	0.8348	0.4263
Se	1.0073	0.4797
Te	0.7500	0.5037