The carbon cost of materials discovery: Can machine learning really accelerate the discovery of new photovoltaics? Supplementary Information

Matthew Walker* and Keith T. Butler†

Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, United Kingdom

I. PREDICTING THE SLMES OF EXTERNAL TEST SETS

Figure 1 shows how Model I, which predicts SLMEs directly, has considerably larger errors when predicting on test sets calculated with different theoretical approaches to the GGA + HSE scissor correction approach used to calculate the training data. The Δ-sol approach [1] is not necessarily more accurate than GGA + HSE, so these errors are not too concerning, but the large errors on the highly accurate GW calculations from Yu & Zunger [12] stress the need for high-fidelity training data.

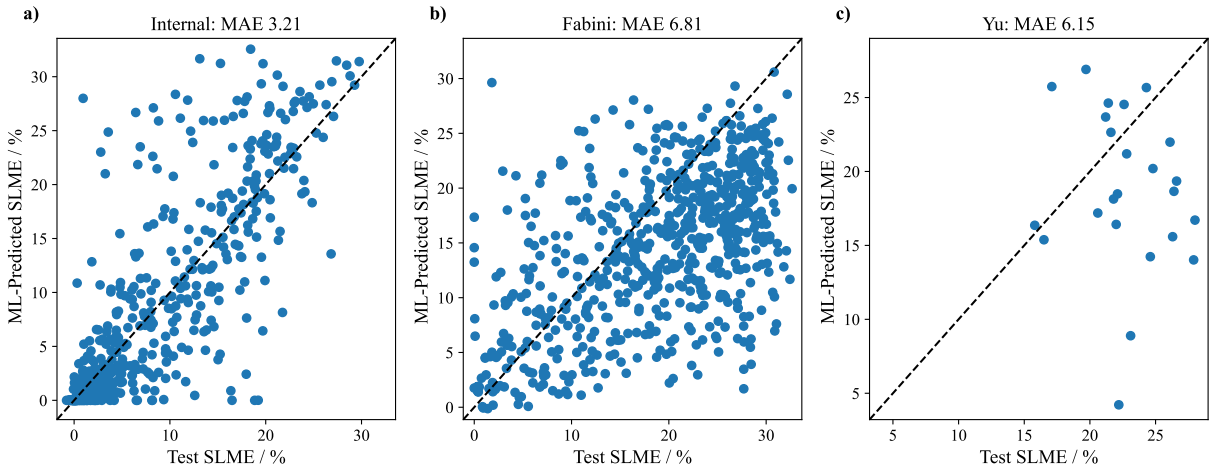


FIG. 1. ML predictions of SLMEs applied to a) the left-out test set from the dataset used for training, b) the Δ-sol dataset from Fabini et al. [4], and c) the GW dataset from Yu and Zunger [12].

* matthew.walker.21@ucl.ac.uk
† k.t.butler@ucl.ac.uk

II. SPECTRAL PREDICTION IMPROVEMENT WITH TRAINING DATASET SIZE

Figure 2 shows the improvement in the predicted spectrum of GaAs by the ALIGNN [2] model used in this work as the size of the training dataset increases. Each value is scaled separately, leading to initially noisy spectra, but by 10^3 training data points, the point-to-point correlation has been learned and the spectra become smooth.

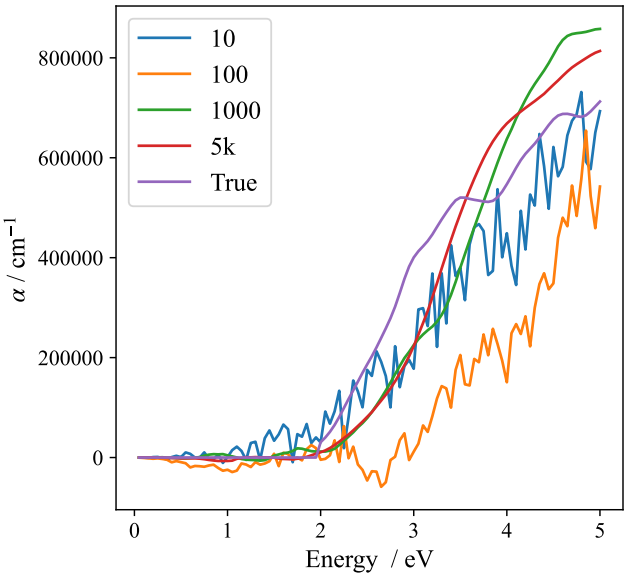


FIG. 2. ML predictions of absorption spectrum of GaAs (in ground state) as training dataset size increases. Note the curve becoming smooth when increasing from 100 to 1000 data points.

III. EVEN GOOD SPECTRAL PREDICTION CAN GIVE BAD SLMES

Figure 3 compares the predictive performance of ALIGNN from Choudhary et al. [2] and the equivariant GNNOpt from Hung et al. [6] based on e3nn [5], when both are trained on the training set used in this work. Note that here the offset used is the true value, so the errors are lower than Method II in the main text. Out-of-the-box, ALIGNN gives slightly better R^2 scores on the $\bar{\alpha}$ prediction on the test set, while GNNOpt gives SLMEs with an MAE nearly 2 percentage points lower. The R^2 score for the SLMEs for GNNOpt is significantly lower than reported by Hung et al. (0.50 vs 0.81), likely due to the more diverse training data with more atoms per unit cell. This appears to be more detrimental to performance than the improvement

gained from training on 4x as much training data as they used. These graphs also cast doubt on $\bar{\alpha}(E)$ as a metric to quantify spectrum reproduction quality, especially when the desired application is calculating SLMEs.

IV. CARBON COSTS

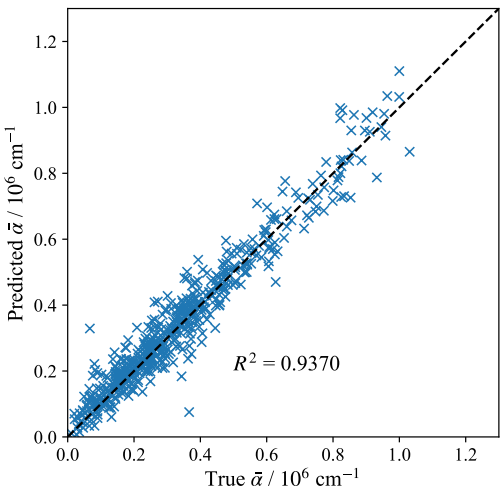
Tables I, II, and III show the energy and carbon costs of a set of DFT calculations in VASP [8–10], used to estimate the typical cost of such calculations (averages in figure III), and an ML inference using ALIGNN [2]. These were used for the Pareto fronts and to generate the area plot of the method costs (Figure 1 in the main text), for which the area is calculated as:

$$A_{\text{method}} = \ln \frac{C_{\text{method}}}{C_{\text{min}}} + 1,$$

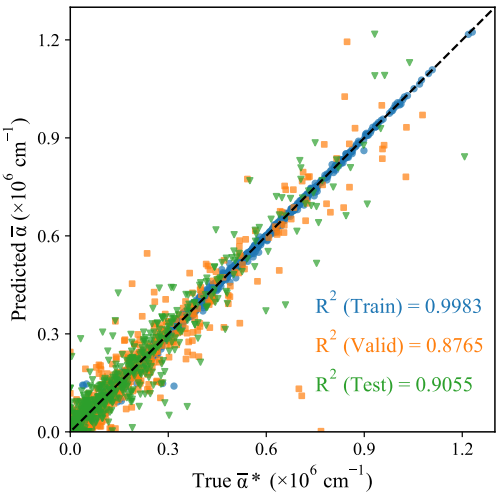
where the addition of 1 gives the smallest cost an area of 1. Then these areas are scaled to a reasonable size in the plot.

	Static calculation			**Bands**			**Optics**		
	VASP time/ s	Energy / kWh	kg CO_2	VASP time/ s	Energy / kWh	kg CO_2	VASP time/ s	Energy / kWh	kg CO_2
GaAs	21.555	0.0022	0.0005	98.181	0.0101	0.0024	16.543	0.0017	0.0004
InP	12.566	0.0013	0.0003	5.735	0.0006	0.0001	7.056	0.0007	0.0002
CdTe	13.908	0.0014	0.0003	6.577	0.0007	0.0002	7.789	0.0008	0.0002
$MAPbI_3$	121.952	0.0125	0.003	130.802	0.0135	0.0032	198.756	0.0204	0.0049
$GaCuSe_2$	29.675	0.0031	0.0007	20.594	0.0021	0.0005	35.82	0.0037	0.0009
$InCuSe_2$	33.098	0.0034	0.0008	26.208	0.0027	0.0006	41.523	0.0043	0.001
$ZnCu_2SnS_4$	32.064	0.0033	0.0008	24.935	0.0026	0.0006	50.326	0.0052	0.0012
$CuSbS_2$	49.224	0.0051	0.0012	43.89	0.0045	0.0011	105.457	0.0108	0.0026
Sb_2S_3	78.297	0.0081	0.0019	42.645	0.0044	0.001	85.909	0.0088	0.0021
Cu_2O	13.429	0.0014	0.0003	7.06	0.0007	0.0002	13.238	0.0014	0.0003
SnS	36.981	0.0038	0.0009	27.329	0.0028	0.0007	59.427	0.0061	0.0015
Sb_2Se_3	86.269	0.0089	0.0021	44.308	0.0046	0.0011	89.532	0.0092	0.0022
Si	8.323	0.0009	0.0002	4.938	0.0005	0.0001	5.556	0.0006	0.0001

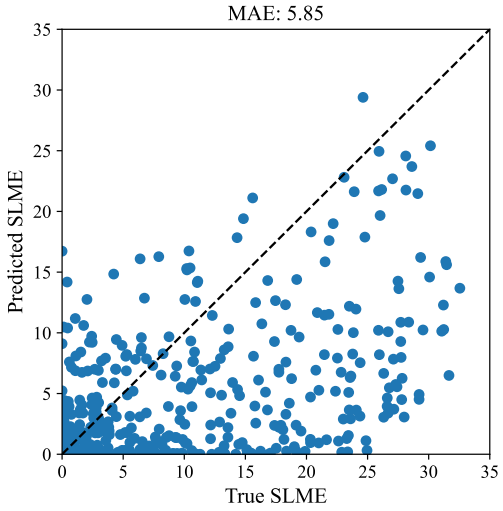
TABLE I. Table of the GGA calculations used to calculate the SLME of an inorganic material with VASP [8–10] CPU time, energy consumption, and carbon equivalent reported for a range of high-performance PVs, with a range of atom types and system sizes. Energy and carbon cost calculated with CodeCarbon [3].



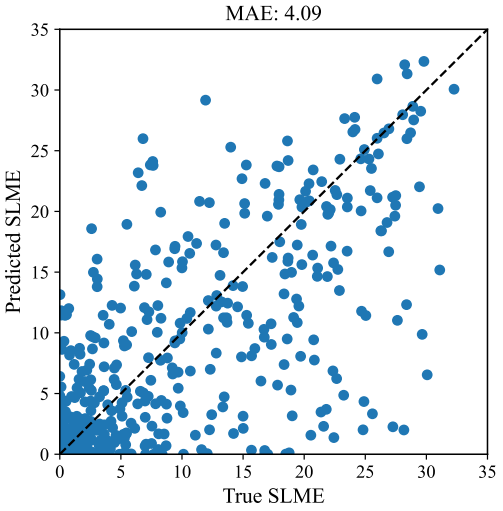
(a) High R^2 score for ALIGNN model predicting absorption spectra, when performance is measured by $\bar{\alpha}(E)$.



(b) Similarly strong performance by Hung et al.'s model [6], graph produced by Juypter Notebook from Hung et al.



(c) Considerable error in SLMEs calculated from absorption spectra predicted using ALIGNN.



(d) Still considerable error when Hung et al.'s predicted spectra [6] are used to calculate SLMEs.

FIG. 3. Comparing our absorption spectrum-predicting model using ALIGNN [2] to that of Hung et al. [6] when both are trained on the dataset we used, with spectra from Wood-Robinson et al. [11] and scissor corrections from Kim et al. [7].

	Static calculation			Bands			Optics		
	VASP time/ s	Energy / kWh	CO_2e / kg	VASP time/ s	Energy / kWh	CO_2e / kg	VASP time/ s	Energy / kWh	CO_2e / kg
GaAs	545.494	0.0561	0.0133	643.675	0.0662	0.0157	702.546	0.0722	0.0172
InP	231.889	0.0238	0.0057	237.624	0.0244	0.0058	213.787	0.022	0.0052
CdTe	252.441	0.026	0.0062	259.018	0.0266	0.0063	190.582	0.0196	0.0047
$MAPbI_3$	596.506	0.0613	0.0146	727.308	0.0748	0.0178	505.214	0.052	0.0123
$GaCuSe_2$	749.109	0.077	0.0183	769.703	0.0791	0.0188	635.948	0.0654	0.0155
$InCuSe_2$	786.352	0.0809	0.0192	812.56	0.0836	0.0199	624.712	0.0642	0.0153
$ZnCu_2SnS_4$	4583.033	0.4713	0.112	4607.968	0.4738	0.1126	3878.417	0.3988	0.0948
$CuSbS_2$	5219.081	0.5367	0.1275	5262.971	0.5412	0.1286	2842.61	0.2923	0.0694
Sb_2S_3	14583.39	1.4996	0.3563	14626.035	1.504	0.3573	12362.016	1.2712	0.302
Cu_2O	378.456	0.0389	0.0092	385.516	0.0396	0.0094	212.456	0.0218	0.0052
SnS	1935.761	0.1991	0.0473	1963.09	0.2019	0.048	1439.235	0.148	0.0352
Sb_2Se_3	3051.384	0.3138	0.0745	3095.692	0.3183	0.0756	2637.996	0.2713	0.0644
Si	189.814	0.0195	0.0046	194.752	0.02	0.0048	173.962	0.0179	0.0043

TABLE II. Table of the HSE calculations used to calculate the SLME of an inorganic material with VASP [8–10] CPU time, energy consumption, and carbon equivalent reported for a range of high-performance PVs, with a range of atom types and system sizes. Energy and carbon cost calculated with CodeCarbon [3].

Calculation	**Energy / kWh**	**CO_2e / kg**
GGA static	0.004250	0.001010
GGA bands	0.003822	0.0009081
GGA optics	0.005671	0.001347
HSE static	0.2618	0.06221
HSE bands	0.2657	0.06312
HSE optics	0.2090	0.04965
ML inference	1.911×10^{-6}	4.5396×10^{-7}

TABLE III. Average energy and carbon cost of the VASP [8–10] processes from Tables I and II compared to a single ML inference of an ALIGNN [2] model trained to predict SLMEs. All data calculated using CodeCarbon [3]

[1] Chan, M. K. Y. and Ceder, G., Physical Review Letters **105**, 196403 (2010).

[2] Choudhary, K. and DeCost, B., npj Computational Materials **7**, 1 (2021).

[3] Courty, B., Schmidt, V., Luccioni, S., Goyal-Kamal,, MarionCoutarel,, Feld, B., Lecourt, J., LiamConnell,, Saboni, A., Inimaz,, supatomic,, Léval, M., Blanche, L., Cruveiller, A., ouminasara,, Zhao, F., Joshi, A., Bogroff, A., Lavoreille, H. d., Laskaris, N., Abati, E., Blank, D., Wang, Z., Catovic, A., Alencon, M., Michał Stęchły,, Bauer, C., Araújo, L. O. N. d., JPW,, and MinervaBooks,, "mlco2/codecarbon: v2.4.1," (2024), version Number: v2.4.1.

[4] Fabini, D. H., Koerner, M., and Seshadri, R., Chemistry of Materials **31**, 1561 (2019).

[5] Geiger, M. and Smidt, T., "e3nn: Euclidean Neural Networks," (2022).

[6] Hung, N. T., Okabe, R., Chotrattanapituk, A., and Li, M., Advanced Materials **36**, 2409175 (2024).

[7] Kim, S., Lee, M., Hong, C., Yoon, Y., An, H., Lee, D., Jeong, W., Yoo, D., Kang, Y., Youn, Y., and Han, S., Scientific Data **7**, 387 (2020).

[8] Kresse, G. and Furthmüller, J., Computational Materials Science **6**, 15 (1996).

[9] Kresse, G. and Furthmüller, J., Physical Review B **54**, 11169 (1996).

[10] Kresse, G. and Hafner, J., Physical Review B **47**, 558 (1993).

[11] Woods-Robinson, R., Xiong, Y., Shen, J.-X., Winner, N., Horton, M. K., Asta, M., Ganose, A. M., Hautier, G., and Persson, K. A., Matter **6**, 3021 (2023).

[12] Yu, L. and Zunger, A., Physical Review Letters **108**, 068701 (2012).