Data-driven Discovery of Photoactive Quaternary Oxides using First-principles Machine Learning

Daniel W. Davies,[†] Keith T. Butler,[‡] and Aron Walsh[*,†,¶]

†*Department of Materials, Imperial College London, Exhibition Road, London SW7 2AZ, UK*

‡*SciML, Scientific Computing Division, Rutherford Appleton Laboratory, Harwell Oxford, Didcot, Oxfordshire OX11 0QX, UK*

¶*Global E*^{3} *Institute and Department of Materials Science and Engineering, Yonsei University, Seoul 120-749, Korea*

E-mail: a.walsh@imperial.ac.uk

Abstract

We present a low-cost, virtual high-throughput materials design workflow and use it to identify earth-abundant materials for solar energy applications from the quaternary oxide chemical space. A statistical model that predicts bandgap from chemical composition is built using supervised machine learning. The trained model forms the first in a hierarchy of screening steps. An ionic substitution algorithm is used to assign crystal structures, and an oxidation state probability model is used to discard unlikely chemistries. We demonstrate the utility of this process for screening over 1 million oxide compositions. We find that, despite the difficulties inherent to identifying stable multi-component inorganic materials, several compounds produced by our workflow are calculated to be thermodynamically stable or metastable and have desirable optoelectronic properties according to first-principles calculations. The predicted oxides

are Li_2MnSiO_5, $MnAg(SeO_3)_2$ and two polymorphs of $MnCdGe_2O_6$, all four of which are found to have direct electronic bandgaps in the visible range of the solar spectrum.

Introduction

The past decade has seen the construction of extensive databases for computed materials properties from quantum mechanical calculations.[1–6] These databases have enabled the virtual screening of thousands of compounds for new target properties in the fields of photovoltaics,[7–9] solar fuels,[10–14] thermoelectrics,[15–17] and others.[18,19] They are also facilitating the move towards predictive materials design using data-mining and machine learning (ML). A growing infrastructure of ML tools has enabled its application to complex problems across many areas of molecular and materials science.[20] This includes building models that relate readily-available descriptors to desirable properties including bandgap,[21–24] thermodynamic stability,[25–27] thermal transport properties[28,29] and the probability for crystal structure types to form.[30,31] These approaches constitute computationally affordable ways to explore the vast chemical space that is otherwise intractable to high-throughput first-principles computation.[32]

While the development of more advanced statistical techniques for chemical and materials science continues,[33] it is already possible to add ML models to the list of tools that can be used in materials design workflows. In this paper, we present a virtual high-throughput screening process in which ML joins the ranks of other data-driven models and density functional theory (DFT) calculations to constitute a hierarchy of filtering stages. The overall workflow is capable of translating from a compositional search space of over 1 million quaternary oxides ($A_wB_xC_yO_z$) to compounds predicted to have target optoelectronic properties by explicit quantum-mechanics calculations.

Our workflow consists of five steps. In the first, which deals with the largest number of configurations, an ML model is used to screen for compositions predicted to have a bandgap

within a window for potential applications for solar energy conversion. The next stage of filtering, illustrated in Figure 1, combines multiple low-cost data-driven approaches to further reduce the search space. We make use of the Herfidahl Hirschman Index of Resource Availability (HHI_R)[34] to focus on the most sustainable element compositions. Two established models are used to assign high-ranking compositions to likely crystal structures,[35] then assess the feasibility of these new compounds in terms of oxidation states.[36] Finally, automated electronic structure calculations are carried out in order to accurately predict the thermodynamic stability and bandgap of candidate materials. We demonstrate the overall process by screening 1.1 million quaternary oxide compositions to identify four new compounds with suitable bandgaps for solar energy applications comprising of earth-abundant elements. These data-driven approaches are used to drastically reduce the required computational resources compared to a brute-force first-principles investigation.

Step 1: Machine learning model of oxide bandgaps

Supervised ML can be used to build statistical models that relate input values (features) to target values (labels) for a set of training samples. These models can then be used predictively given new data. There exists a wide variety of supervised ML approaches, many of which are being applied to numerous problems relating to first-principles materials modelling.[20] We now provide a brief outline of the key concepts and training procedure needed to build a gradient boosting regression (GBR) model, which is employed in this work to predict bandgaps from chemical compositions. The GBR model is trained and subsequently applied using the `scikit-learn` Python library.[39]

Model structure

In GBR, an ensemble of individual weak learners (usually decision trees) is used. By *weak* learners, we mean that each individual learner has poor predictive power if applied in iso-

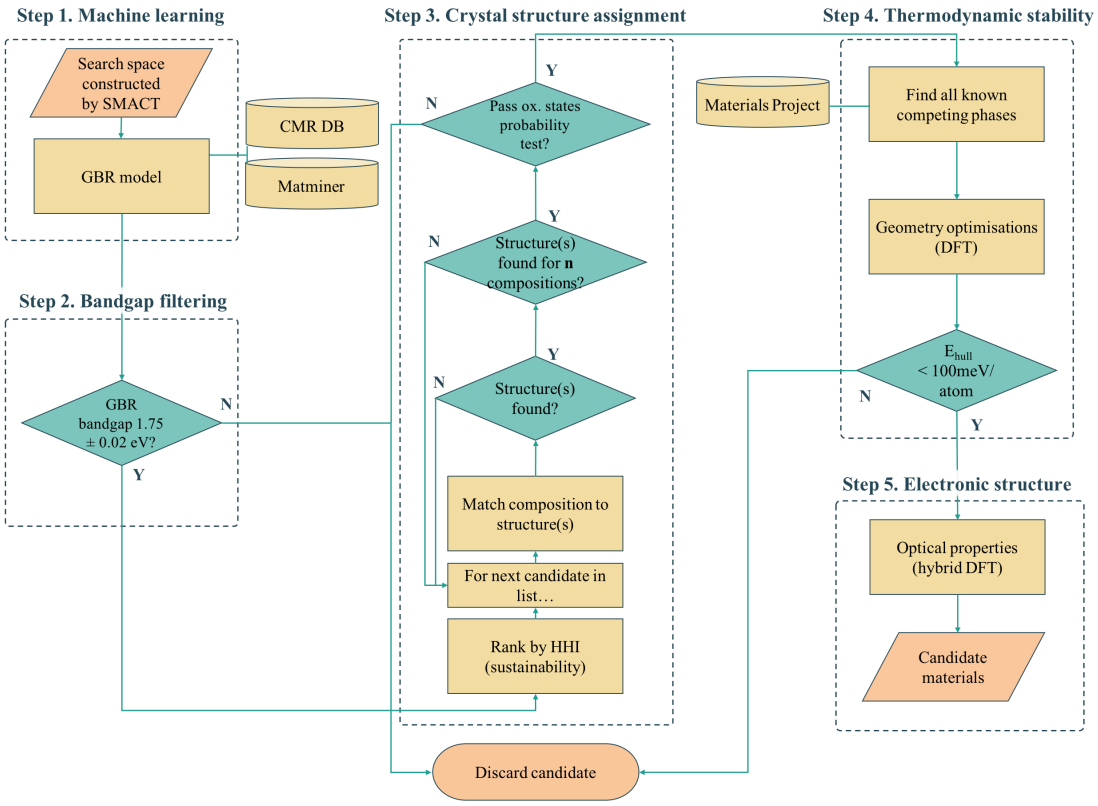

Figure 1: Computer-aided design workflow: Data from the Computational Materials Repository (CMR) database is used in conjunction with `Matminer`[37] to construct a gradient boosting regression (GBR) model (step 1), which is then used as a bandgap filter (step 2). Compositions are ranked using the Herfindahl Hirschman Index ($\mathrm{HHI_R}$),[34] appropriate structures generated with a structure substitution algorithm,[38] and a probabilistic oxidation state model filters out unlikely species combinations (step 3). Thermodynamic stability (step 4) and bandgaps (step 5) are calculated from first-principles using semi-local density functional theory (DFT) and non-local hybrid DFT.

lation. When building decision trees, the goal is to predict the value of sample labels by learning simple decision rules from the sample features. Individual trees are constructed using the classification and regression trees (CART) algorithm.[40] In brief, for a given node of a decision tree (Figure 2a), the sample space is split into two parts that are as homogeneous as possible according to their labels. A decision rule involving one of the sample features is selected to best achieve this goal, i.e. to minimise the impurity of the node. This process is carried out recursively until some stopping criteria is met. For regression, the mean value of the ground truth labels at a given leaf node is taken as the prediction of the model for

samples at that node.

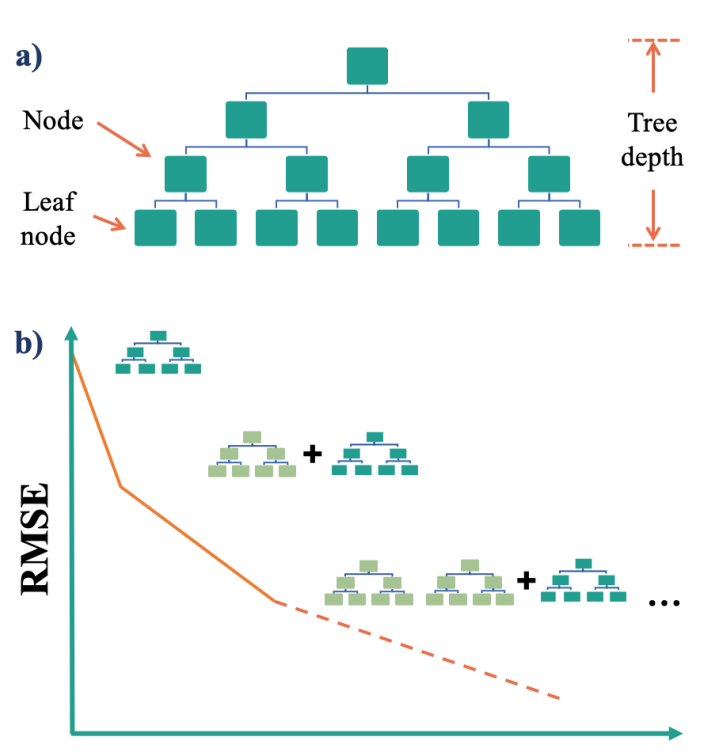

Figure 2: Schematic of the gradient boosting regression (GBR) model. a) A decision tree splits the sample space recursively at nodes based on feature values, grouping samples into leaf nodes. b) Multiple decision trees are constructed during the GBR process with each consecutive tree trained on the residuals of the existing model, minimising the root mean squared error (RMSE).

One problem with decision trees – and indeed the reason that they fall into the category of *weak* learners – is that by splitting the sample space on the basis of one feature at each node, they fail to include predictive power from multiple, overlapping regions of feature space. As such, decision trees of a small depth tend to ignore valuable information from unused features, while those of a large depth are likely to be fit to random noise in the dataset. This shortcoming is countered within GBR by constructing multiple decision trees sequentially. As depicted in Figure 2b, the overall model is built by adding trees in a forward, stagewise fashion with each consecutive tree trained not on the sample labels, but on the residuals of the current model. The result is that each consecutive tree can consider the whole sample space and serves to improve the overall performance of the model by minimizing a chosen

loss function, in this case the root-mean-squared-error (RMSE).

Data representation

The target property that we wish to predict is the bandgap calculated using the GLLB-sc functional,[41] which has been shown to give more reliable estimations of bandgap than semi-local DFT functionals that operate within the generalised-gradient approximation (GGA).[11] The bandgap values produced by Castelli *et al.* are used as a training set,[13] and are available from the Computational Materials Repository (CMR) database.[3] This set is comprised of 2,289 inorganic materials, 799 of which are oxides (i.e. contain oxygen and at least one other element), which is used as training data.

The compositions of the materials are represented using the element properties from the `Magpie` package.[42] The features used are the minimum, maximum, range, mean, mode and mean absolute deviation (MAD) of atomic number, Mendeleev number, atomic mass, melting temperature, electronegativity, among others (see Supplementary Information of Ref. 42 for the full list). In addition, we use the number of valence electrons, elemental frontier orbital energies calculated from neutral atoms with DFT, and the bandgap center position calculated using the geometric mean of electronegativities as demonstrated by Nethercot.[43] All of the 149 features are generated using the `Matminer` Python library.[37]

Model training

While the model parameters are set automatically during the learning process as described above, several key *hyperparameters* must be chosen at the start. For GBR, as well as the self-explanatory tree-specific hyperparameters, there are three key boosting parameters (Table 1). The fraction of compounds to fit each tree dictates the maximum number of samples in the training set that any individual tree can use, introducing some level of diversity into the ensemble, which helps to mitigate against overfitting. The number of decision trees and learning rate refer to the number of boosting stages used in the final ensemble and the factor

by which the contribution of each new tree is multiplied, respectively. The overall model is given by

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{1}$$

where M is the total number of decision trees, $h_m(x)$ are the individual trees and γ_m is the learning rate.

The total error in ML approaches comes from a combination of bias, variance, and irreducible errors. Gradient boosting reduces bias of individual trees, but runs the risk of increasing the variance (error from sensitivity to noise in the training data). Upon changing a given hyperparameter, it is crucial to check how the model performs on unseen data, even if the fit to the training data appears to be improving (see Figure 3). Each time a model is built using a trial set of hyperparameters, 10-fold cross validation (CV) is performed whereby the model is trained on 90% of the data, then tested on the remaining 10%. This process is repeated such that every 10% chunk of data is used for testing, then the mean RMSE is calculated.

Optimal hyperparameter values for this GBR model were found by Bayesian optimisation and are listed in Table 1. This was achieved using the `scikit-optimize` Python library,[44] and involves approximating the model using Gaussian processes. The next set of hyperparameters to trial is chosen by an acquisition function over the Gaussian prior, which is cheaper to evaluate than the model itself. A more detailed explanation of this approach can be found in Ref. 45. Using these parameters, as well as removing oxide gases such as CO_2 and SO_2, and complex anions containing uncommon oxidation states such as phosphites and perphosphates, yields a final model with an RMSE of 0.95 eV.

Finally, it might be assumed that the correlation between bandgap calculated using GGA exchange-correlation functionals, which tend to be consistently underestimated, and that calculated using GLLB-sc could be high enough to use predicatively. If this were the case, a ML model could be trained using a larger database, such as the Materials Project (MP), which contains ∼86,000 inorganic structures with bandgaps calculated using the PBE functional.[46]

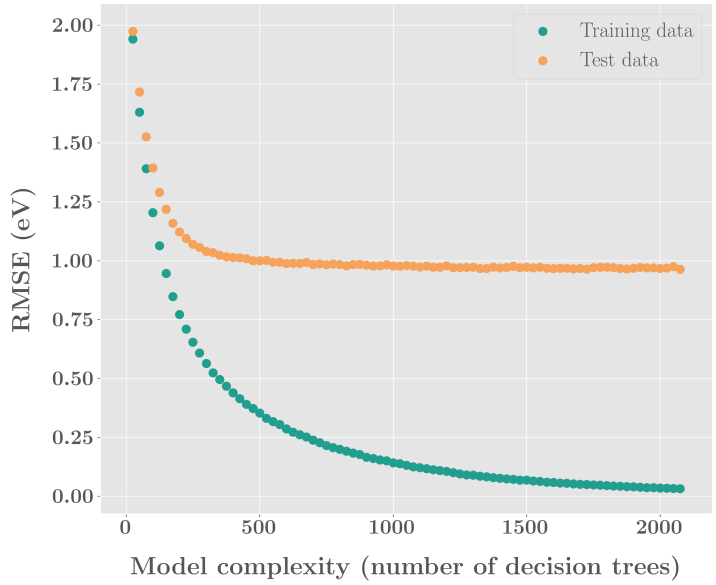

Figure 3: Effect of the number of decision trees (boosting stages) on model performance for bandgap prediction. At a certain threshold (∼1000), increasing the number of trees in the model ceases to improve its performance on unseen test data, even though it appears to better fit the training data.

We find that while the expected linear relationship is observed between bandgaps calculated using PBE and GLLB-sc, there is significant deviation from the relationship, and that this is larger in general for oxides (see Figure S1 in the Supplementary Information). The standard deviation is 0.85 eV, thus for a two-step approach to be advantageous, the RMSE of the model trained on the large dataset of PBE bandgaps would have to be unreasonably low ($<$ 0.1 eV).

Model performance and limitations

Features representing the crystal structures of inorganic compounds to ML algorithms are the subject of much recent development.[47–50] The use of such features has been shown to improve the predicted properties of inorganic solids beyond compositional representations alone. As such, the accuracy of our model is limited because atomic connectivity is not accounted for. This effect is particularly prevalent for oxides, as their structural diversity results in a wide

Table 1: Hyperparameter values used in final GBR model for oxide bandgap prediction.

Tree-specific parameters	
Min. compounds needed to split nodes	65
Max. depth of tree	20
Min. compounds required at leaf nodes	1
Max. features considered per tree	86
Boosting parameters	
Fraction of compounds to fit each tree	0.9
Learning rate	0.01
Number of decision trees	1000

variety of local bonding arrangements. We have quantified this phenomenon by showing that the unscreened Madelung site potential of the oxide anion – a quantity that reflects the electrostatic potential of an ion in a crystal by approximating ions as point charges – varies across all binary metal oxides with a striking range of 16 V.[51] The distribution of the maximum (PBE) bandgap difference between polymorphs for all oxide compositions in the MP database is shown in Figure 4. While for a large number of oxides, polymorphism results in a bandgap difference of < 0.5 eV, the difference can be as large as 4.18 eV (e.g. $LiFePO_4$) and the mean difference is 0.57 eV. This highlights the extent to which crystal structure plays a role in dictating bandgap, and a model that considers chemical composition alone can only be used as a pre-screening filter. In this context, a composition-only model with an RMSE of 0.95 eV is reasonable.

It is also instructive to compare this approach to existing heuristic methods. For example, the solid state energy (SSE) scale,[52,53] derived from the relationship between electron affinity (EA) and ionization potential (IP) and bandgap for a selection of binary closed-shell inorganic semiconductors and insulators, can be used to estimate bandgaps for new compounds.[54] The SSE has knowledge only of the EA and IP values of the constituent cations and anions, respectively. The range and standard deviation of IP values for the 56 binary oxides used in the construction of the SSE model are 4.9 eV and 1.44 eV, respectively, giving O the largest associated uncertainty of all the anions featured. For this reason, there is no correlation between the bandgap predicted using the SSE scale and the GLLB-sc bandgap of the 799

oxides in the training dataset (see Figure S2). By taking into account more information about the constituent elements, the GBR model we developed is able to predict bandgaps to a higher level of accuracy.

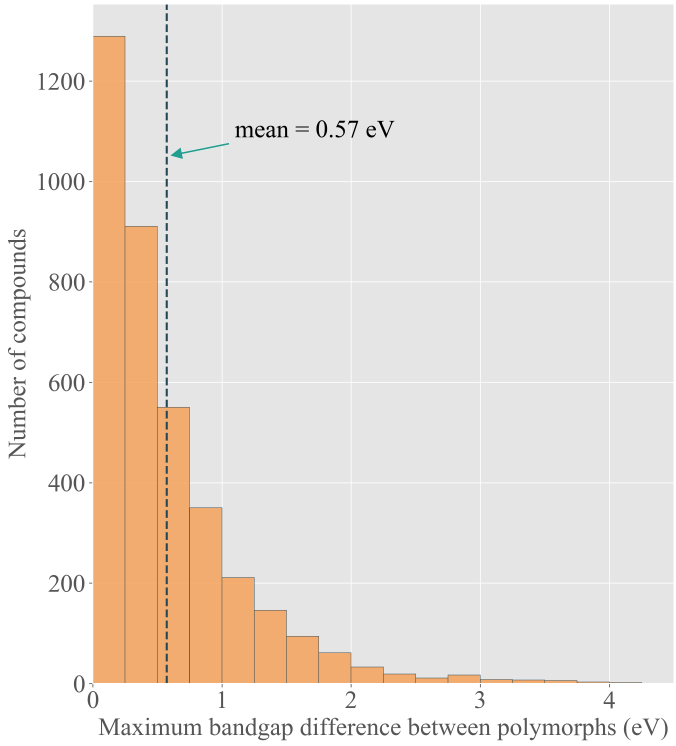

Figure 4: Distribution of maximum bandgap difference between polymorphs for oxides in the Materials Project database that exhibit polymorphism. Only compounds with an energy above the convex hull of < 0.1 eV and a maximum bandgap difference of > 0.05 eV are included. Bandgaps are calculated in the Materials Project using the GGA-PBE functional.

Finally, we can inspect which features are most important in the final GBR model using the decrease impurity method.[40] Figure 5 shows the mean absolute deviation (MAD) of covalent radius is the most important feature. The mean value for volume per atom and MAD of melting temperature are also relatively important. The extent to which this can be interpreted as meaningful depends on how highly correlated the features are. For example, covalent radius and volume per atom are strongly correlated, which makes it harder to

decouple their contributions to the overall model. In general, a number of features contribute significantly to the final model. Investigation into the effect of systematically removing correlated features, and retraining the model, is an avenue for further study and a means of extracting physically-intuitive relationships.

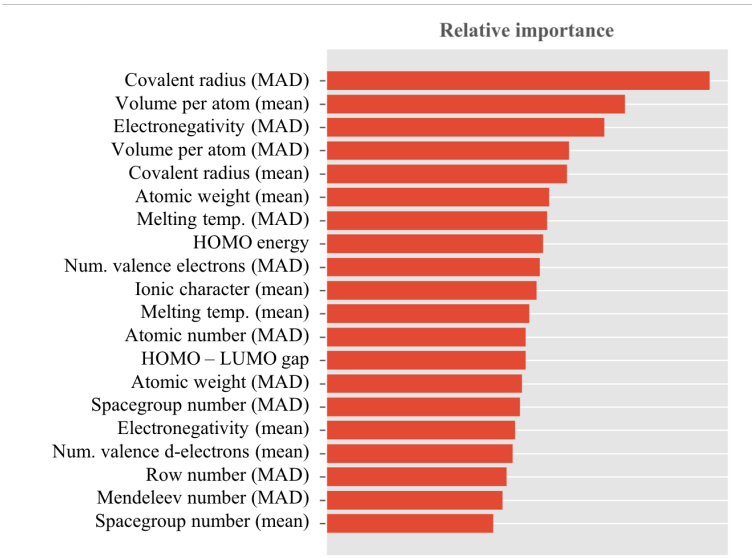

Figure 5: Relative importance of the 20 most important features in the final gradient boosting regression model. HOMO and LUMO energies refer highest occupied and lowest unoccupied molecular orbital energy as calculated within DFT, respectively, (taken directly from the `Matminer` Python library). Ionic character refers to Pauling's empirical ionic character between pairs of atoms calculated using electronegativities.[55]

Step 2: Bandgap filter

We now use the trained GBR model to search for promising candidates from a large search space. A pool of 1.1 million hypothetical quaternary oxide compositions was generated using the `SMACT` Python library, implementing the heuristic chemical rules employed in that code.[32] The target bandgap window of 1.0–2.5 eV will capture a wide range of photoactive materials. Smaller gaps may be more suitable for single-junction photovoltaic applications, while wider gaps could be used in tandem systems or solar fuel processes.[56,57]

The distribution of errors obtained using the GBR model is shown in Figure 6a. Materials

predicted to have a bandgap at the centre of the target window (1.75 eV) have a 60% probability of having a GLLB-sc bandgap within the window. In contrast, Figure 6b shows the distribution of bandgaps of all oxides in the CMR dataset and the probability of choosing one at random with a bandgap in the target window is just 8%. We filter the 1.1 million candidates that do not have a predicted bandgap of 1.75 ± 0.02 eV, leaving 17,833 viable compositions. This approach does not aim to capture all the hypothetical compositions that fall between within the target bandgap window. Rather, those compositions that are most likely to have useful bandgaps according to the GBR model are targeted. This screening step corresponds to a greater than 60-fold reduction of the search space.

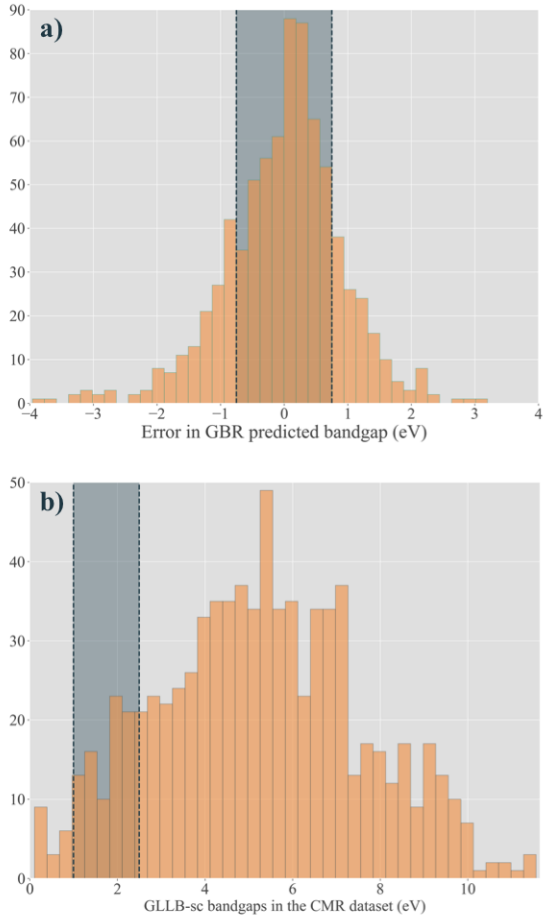

Figure 6: a) Distribution of error in predicted bandgap by the trained GBR model. The shaded region corresponds to an error of ± 0.75 eV and encloses 60% of all predictions. b) Distribution of GLLB-sc bandgaps for oxides in the CMR training dataset. The shaded region corresponds to a bandgap of 1.75 ± 0.75 eV.

Step 3: Crystal structure assignment

The surviving 17,833 compositions are ranked by sustainability using the HHI_R scale.[34] Starting with the most sustainable composition, chemically plausible quaternary oxide crystal structures are constructed using the structure substitution algorithm developed by Hautier and coworkers.[38] For each predicted structure, an oxidation state probability model[54] is applied as an additional filter, to check that the combination of ions in that structure is chemically plausible. A low probability threshold of 0.005 is used so only very unlikely species combinations are eliminated. We also choose to eliminate Ti^{3+} compounds due to the d^1 electronic configuration being linked to fast electron-hole recombination for solar applications. This procedure was repeated until 235 candidate materials were generated, corresponding to 61 unique chemical compositions. This pool is small enough to allow for explicit first-principles calculations, and we take these candidate materials forward to calculate their thermodynamic stability.

Step 4: Thermodynamic stability

Competing phases are identified using the chemical potentials from the MP database. Then full geometry optimization is carried out on candidate compounds and all competing phases using DFT at the GGA (PBEsol) level, with equivalent computational setup. This is done in high-throughput using the `Atomate`[58] and `Fireworks`[59] Python libraries. The DFT total energies are used to determine thermodynamic stability *via* the distance from the 3D convex hull of the quaternary phase diagram.

Of the 235 compounds, 27 are calculated to be within the predefined metastability window of 100 meV/atom of the convex hull. Four of the 27 compounds were found to be structurally identical to one other compound in the set, leaving 23 unique compounds, corresponding to 8 distinct compositions. The presence of identical structures can occur when different parent structures are found for one composition using the structure substitution algorithm, which

then ultimately yield the same crystal structure following geometry optimisation.

The relatively small proportion of stable and metastable compounds is unsurprising given the existence of a large number of stable binary and ternary oxides that act as competing phases. The energies above the convex hull for all 23 compounds are given in Table S1. Only one compound, Li_2MnSiO_5, has been previously reported in the MP database, but has not been synthesised experimentally to the authors' knowledge. Shown in Figure 7a, the compound $ZrMnSi_2O_7$ is the only one predicted to be thermodynamically stable, while a second polymorph of $ZrMnSi_2O_7$ along with a Li_2TiMnO_4 structure are predicted to be $<$ 10 meV/atom above the convex hull, as shown in Figure 7b and Figure 7c, respectively.

While three polymorphs of Li_2TiMnO_4 are in the MP database, including one that has been investigated as a possible active material for Li-ion battery applications,[60] none of the crystal structures adopted by the candidate compounds have previously been reported. The new phase of Li_2TiMnO_4 differs from the three previously reported polymorphs as the metals are in tetrahedral environments as opposed to octahedral. It also has a wide electronic bandgap of 4.21 eV, as calculated using a hybrid DFT functional in the following section, whereas the previously reported compounds all have PBE-calculated bandgaps of less than 0.4 eV. To the best of our knowledge, no compounds have previously been reported for any of the other 7 compositions.

Table 2: Summary of compounds found to have (predicted HSE06/DFT) bandgaps that fall within the target window of 1.0–2.5 eV.

Formula	Spacegroup	E_{hull} (meV/atom)	Bandgap (eV)
Li_2MnSiO_5	P4/nmm	86	2.24
$MnCdGe_2O_6$	$P2_1/c$	99	2.47
$MnCdGe_2O_6$	C2/c	99	1.76
$MnAg(SeO_3)_2$	$Pna2_1$	36	2.31

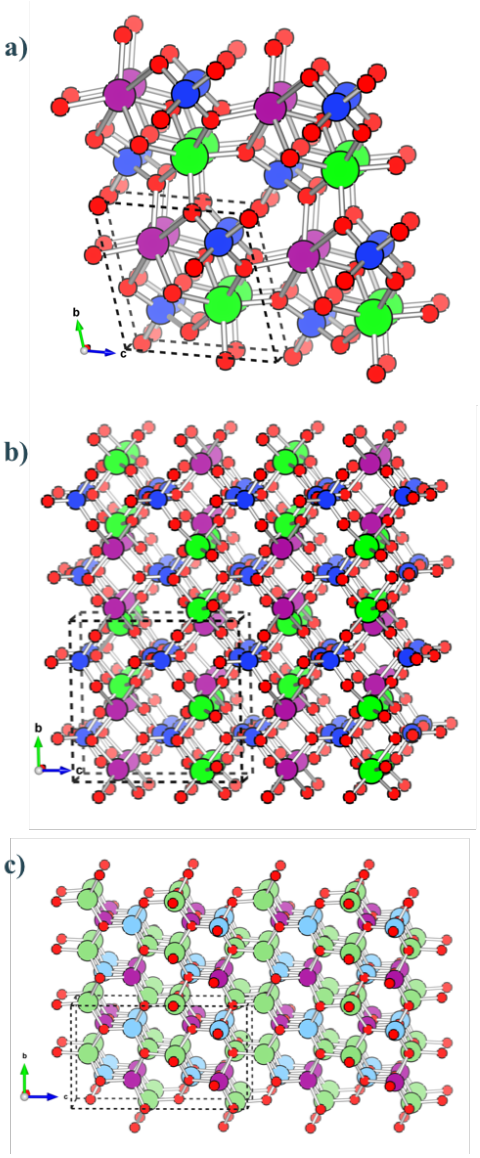

Figure 7: Three most stable compounds identified by the workflow. a) and b) are different polymorphs of $ZrMnSi_2O_7$ in which Si, Zr and Mn atoms are depicted as blue, green and purple circles, respectively. c) A Li_2TiMnO_4 structure in which Li, Ti and Mn atoms are depicted as green, blue and purple circles, respectively. O atoms are red circles in all three structures.

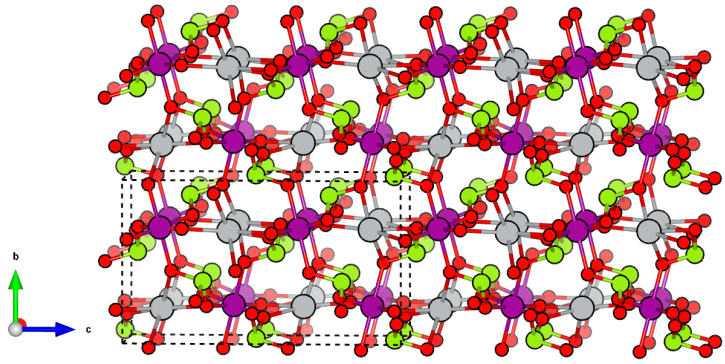

Figure 8: The most stable compound identified by the high-throughput workflow with a bandgap within the target window, $MnAg(SeO_3)_2$. Mn, Ag, Se and O atoms are depicted as purple, silver, green and red circles, respectively.

Step 5: Electronic structure

The electronic structure of the 23 remaining candidate compounds were calculated with the HSE06 hybrid functional.[61,62] While being more computationally demanding, this approach yield more accurate electronic structure information that what is available in current materials databases. The majority of compounds have a calculated bandgap of > 4 eV, which is well outside the target bandgap window (see Table S1). Four of the compounds are calculated to have bandgaps within the target window and are listed in Table 2. The most thermodynamically stable compound with a bandgap within the target window is $MnAg(SeO_3)_2$ and is shown in Figure 8.

Encouragingly, the four compounds with useful bandgaps include three different compositions. Since the original GBR model is trained on composition alone, this indicates a coarse 37.5% success rate. While the success rate is not as high as the original 60% as indicated by the 10-fold CV results, the latter should be considered a maximum achievable success rate when using this model predictively. Cross validation can give some indication of model performance, but there are limited options to glean further insight before applying the model predictively where existing data is scarce.

Crucially, this study represents a small sample size making it impossible to draw strong conclusions. Qualitatively, it promising that we have identified four candidate compounds

using only 235 first-principles calculations, given the "needle in a haystack" nature of the problem. Without using the data-driven screening stages, computationally prohibitive structure optimization calculations would have been required for each of the top compositions suggested by the ML model. The overall virtual high-throughput screening process constitutes a multi-objective optimization, in which bandgap, sustainability and stability are all targeted sequentially (see Figure 1). The latter of these, stability, is likely to be a significant bottleneck for any screening of quaternary materials as compared with binary or ternary phases, given the expected lower stability window due to an increase in possible decomposition pathways.

Finally, the model was trained on bandgaps calculated using the GLLB-sc functional, whereas the bandgaps of the new compounds are calculated using the HSE06 functional. In the original work by Castelli *et al.* in which they calculate the bandgaps used here for training data, they show that bandgaps calculated using HSE06 and GLLB-sc are generally in good agreement.[13] However, they also show that for smaller bandgaps such as those considered here, the GLLB-sc functional has a tendency to underestimated as compared with the HSE06 functional. This could be another reason for getting a lower success rate and would also explain why no compounds had bandgaps calculated using HSE06 smaller than the target window (< 1.0 eV). The availability of a comprehensive database with electronic and thermodynamic properties of materials at a consistent high-level of theory would greatly benefit the training of ML models and future data-driven studies.

Conclusion

We outlined a multi-stage computational procedure to reduce a chemical space of over 1 million compositions to 4 target compounds using a combination of techniques and chemical filters. The majority of the study has been performed on a single-processor workstation. A GBR model was trained to predict bandgaps for quaternary oxide compositions. This model

is shown to outperform established chemical heuristics for ability to predict bandgap and allows for a 60-fold reduction of the initial search space, with an order of magnitude better chance of identifying suitable compounds compared to random filtering. Additional screening based on sustainability, oxidation state combinations, and thermodynamic stability was used, before performing high-quality electronic structure calculations on a pool of 23 candidate materials. Finally, we identified four new quaternary oxides not previously reported or explored for solar energy applications. The workflow that we present here can be a blueprint for using a combination of machine learning and first-principles calculations to allow efficient, targeted screening of the vast chemical structure-composition hyperspace.

Computational Methods

Full information on the workflow is available in Supporting Information. It makes use of the Python libraries `SMACT`,[32] `Pymatgen`,[63] `Matminer`,[37] `Scikit-learn`,[39] `Atomate`,[58] and `Fireworks`.[59]

Electronic structure caluclations

First-principles calculations are carried out using Kohn-Sham DFT with a projector-augmented plane wave basis[64] as implemented in the Vienna Ab-initio Simulation Package (VASP).[65,66] We use the PBEsol exchange-correlation functional[67] and a k-point grid is generated for each calculation with a density of 120 Å^3 in the reciprocal lattice. The kinetic-energy cut-off is set at 600 eV and the forces on each atom minimised to below 0.005 eVÅ^{-1}.

Semi-local exchange-correlation treatments such as the PBEsol functional provide an accurate description of crystal structures but tend to underestimate the electronic bandgaps of semiconductors. To overcome this issue, more accurate electronic structure calculations are performed using the hybrid non-local functional HSE06,[62] which includes 25% screened Hartree-Fock exact exchange. Γ-centred homogeneous k-point grids are used with a density

of 64 Å^3 in the reciprocal lattice and the kinetic energy cutoff is set at 520 eV.

Acknowledgement

Via our membership of the UK's HEC Materials Chemistry Consortium, which is funded by EPSRC (EP/L000202), this work used the ARCHER UK National Supercomputing Service (http://www.archer.ac.uk) for all DFT calculations. DWD is supported by the EPSRC via the Doctoral Prize Fellowship and AW is supported by a Royal Society University Research Fellowship. This research was also supported by the Creative Materials Discovery Program through the National Research Foundation of Korea (NRF) funded by Ministry of Science and ICT (2018M3D1A1058536).

Supporting Information Available

Code showing all steps of the screening process is available in Jupyter notebook form at `https://doi.org/10.5281/zenodo.2609120`, whilst the `SMACT` code is available from `https://doi.org/10.5281/zenodo.2609134`. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Setyawan, W.; Curtarolo, S. High-throughput electronic band structure calculations: Challenges and tools. *Computational Materials Science* **2010**, *49*, 299312.

(2) Setyawan, W.; Gaume, R. M.; Lam, S.; Feigelson, R. S.; Curtarolo, S. High-Throughput Combinatorial Database of Electronic Band Structures for Inorganic Scintillator Materials. *ACS Combinatorial Science* **2011**, *13*, 382–390.

(3) Landis, D. D.; Hummelshøj, J. S.; Nestorov, S.; Greeley, J.; Dułłak, M.; Bligaard, T.;

Nørskov, J. K.; Jacobsen, K. W. The computational materials repository. *Computing in Science and Engineering* **2012**, *14*, 51–57.

(4) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **2013**, *1*, 011002.

(5) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *Jom* **2013**, *65*, 1501–1509.

(6) The NoMaD Repository. `http://nomad-repository.eu/-[Accessed:02-09-2017]`.

(7) Yu, L.; Kokenyesi, R. S.; Keszler, D. A.; Zunger, A. Inverse Design of High Absorption Thin-Film Photovoltaic Materials. *Advanced Energy Materials* **2012**, *3*, 43–38.

(8) Krishnamoorthy, T.; Ding, H.; Yan, C.; Leong, W. L.; Baikie, T.; Zhang, Z.; Sherburne, M.; Li, S.; Asta, M.; Mathews, N.; Mhaisalkar, S. G. Lead-free germanium iodide perovskite materials for photovoltaic applications. *J. Mater. Chem. A* **2015**, *3*, 23829–23832.

(9) Hinuma, Y.; Hatakeyama, T.; Kumagai, Y.; Burton, L. A.; Sato, H.; Muraba, Y.; Iimura, S.; Hiramatsu, H.; Tanaka, I.; Hosono, H.; Oba, F. Discovery of earth-abundant nitride semiconductors by computational screening and high-pressure synthesis. *Nature Communications* **2016**, *7*, 11962.

(10) Castelli, I. E.; Landis, D. D.; Thygesen, K. S.; Dahl, S.; Chorkendorff, I.; Jaramillo, T. F.; Jacobsen, K. W. New cubic perovskites for one- and two-photon water splitting using the computational materials repository. *Energy & Environmental Science* **2012**, *5*, 9034–9043.

(11) Castelli, I. E.; Olsen, T.; Datta, S.; Landis, D. D.; Dahl, S.; Thygesen, K. S.; Jacobsen, K. W. Computational screening of perovskite metal oxides for optimal solar light capture. *Energy Environ. Sci.* **2012**, *5*, 5814–5819.

(12) Wu, Y.; Lazic, P.; Hautier, G.; Persson, K.; Ceder, G. First principles high throughput screening of oxynitrides for water-splitting photocatalysts. *Energy Environ. Sci.* **2013**, *6*, 157–168.

(13) Castelli, I. E.; Hüser, F.; Pandey, M.; Li, H.; Thygesen, K. S.; Seger, B.; Jain, A.; Persson, K. A.; Ceder, G.; Jacobsen, K. W. New light-harvesting materials using accurate and efficient bandgap calculations. *Advanced Energy Materials* **2015**, *5*, 1400915.

(14) Pandey, M.; Vojvodic, A.; Thygesen, K. S.; Jacobsen, K. W. Two-Dimensional Metal Dichalcogenides and Oxides for Hydrogen Evolution: A Computational Screening Approach. *The Journal of Physical Chemistry Letters* **2015**, *6*, 1577–1585.

(15) Toher, C.; Plata, J. J.; Levy, O.; de Jong, M.; Asta, M.; Nardelli, M. B.; Curtarolo, S. High-throughput computational screening of thermal conductivity, Debye temperature, and Grüneisen parameter using a quasiharmonic Debye model. *Physical Review B* **2014**, *90*, 174107.

(16) Sparks, T. D.; Gaultois, M. W.; Oliynyk, A.; Brgoch, J.; Meredig, B. Data mining our way to the next generation of thermoelectrics. *Scripta Materialia* **2016**, *111*, 10–15.

(17) Faghaninia, A.; Yu, G.; Aydemir, U.; Wood, M.; Chen, W.; Rignanese, G.-M.; Jeffrey, S.; Hautier, G.; Jain, A. A computational assessment of the electronic, thermoelectric, and defect properties of bournonite (CuPbSbS3) and related substitutions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 6743–6756.

(18) de Jong, M.; Chen, W.; Geerlings, H.; Asta, M.; Persson, K. A. A database to enable discovery and design of piezoelectric materials. *Scientific Data* **2015**, *2*, 150053.

(19) Miller, S. A.; Gorai, P.; Aydemir, U.; Mason, T. O.; Stevanović, V.; Toberer, E. S.; Snyder, G. J. SnO as a potential oxide thermoelectric candidate. *J. Mater. Chem. C* **2017**, *5*, 8854–8861.

(20) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.

(21) Zhuo, Y.; Mansouri Tehrani, A.; Brgoch, J. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *The Journal of Physical Chemistry Letters* **2018**, *9*, 1668–1673.

(22) Zhu, Z.; Dong, B.; Yang, T.; Zhang, Z.-D. Fundamental Band Gap and Alignment of Two-Dimensional Semiconductors Explored by Machine Learning. **2017**,

(23) Weston, L.; Stampfl, C. Machine learning the band gap properties of kesterite I$_2$-II-IV-V$_4$ quaternary compounds for photovoltaics applications. **2017**,

(24) Lee, J.; Seko, A.; Shitara, K.; Tanaka, I. Prediction model of band-gap for AX binary compounds by combination of density functional theory calculations and machine learning techniques. **2015**,

(25) Legrain, F.; Carrete, J.; van Roekeghem, A.; Madsen, G. K. H.; Mingo, N. Materials Screening for the Discovery of New Half-Heuslers: Machine Learning versus Ab Initio Methods. *The Journal of Physical Chemistry B* **2018**, *122*, 625632.

(26) Seko, A.; Hayashi, H.; Tanaka, I. Compositional descriptor-based recommender system accelerating the materials discovery. *Arxiv* **2017**, 1711.06387.

(27) Faber, F. A.; Lindmaa, A.; Von Lilienfeld, O. A.; Armiento, R. Machine Learning Energies of 2 Million Elpasolite (ABC2D6) Crystals. *Physical Review Letters* **2016**, *117*, 135502.

(28) Ju, S.; Shiga, T.; Feng, L.; Hou, Z.; Tsuda, K.; Shiomi, J. Designing Nanostructures for Phonon Transport via Bayesian Optimization. *Physical Review X* **2017**, *7*, 021024.

(29) Legrain, F.; Carrete, J.; van Roekeghem, A.; Curtarolo, S.; Mingo, N. How Chemical Composition Alone Can Predict Vibrational Free Energies and Entropies of Solids. *Chemistry of Materials* **2017**, *29*, 6220–6227.

(30) Ryan, K.; Lengyel, J.; Shatruk, M. Crystal Structure Prediction via Deep Learning. *Journal of the American Chemical Society* **2018**, jacs.8b03913.

(31) Liu, Y.-H.; van Nieuwenburg, E. P. Discriminative Cooperative Networks for Detecting Phase Transitions. *Physical Review Letters* **2018**, *120*, 176401.

(32) Davies, D. W.; Butler, K. T.; Jackson, A. J.; Morris, A.; Frost, J. M.; Skelton, J. M.; Walsh, A. Computational Screening of All Stoichiometric Inorganic Materials. *Chem* **2016**, *1*, 617–627.

(33) Gubernatis, J. E.; Lookman, T. Machine learning in materials design and discovery: Examples from the present and suggestions for the future. *Physical Review Materials* **2018**, *2*, 120301.

(34) Gaultois, M. W.; Sparks, T. D.; Borg, C. K. H.; Seshadri, R.; Bonificio, W. D.; Clarke, D. R. Data-Driven Review of Thermoelectric Materials: Performance and Resource Considerations. *Chemistry of Materials* **2013**, *25*, 2911–2920.

(35) Hautier, G.; Ong, S. P.; Jain, A.; Moore, C. J.; Ceder, G. Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability. *Physical Review B* **2012**, *85*, 155208.

(36) Davies, D. W.; Butler, K. T.; Isayev, O.; Walsh, A. Materials discovery by chemical analogy: role of oxidation states in structure prediction. *Faraday Discussions* **2018**, *211*, 553–568.

(37) Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **2018**, *152*, 60–69.

(38) Hautier, G.; Fischer, C.; Ehrlacher, V.; Jain, A.; Ceder, G. Data mined ionic substitutions for the discovery of new compounds. *Inorganic Chemistry* **2011**, *50*, 656–663.

(39) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(40) Breiman, F. J. O. R., L.; Stone, C. *Classification and Regression Trees*; Taylor & Francis Group: Boca Raton, 1984; Vol. 1.

(41) Kuisma, M.; Ojanen, J.; Enkovaara, J.; Rantala, T. T. Kohn-Sham potential with discontinuity for band gap materials. *Physical Review B* **2010**, *82*, 115106.

(42) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2016**, *2*, 16028.

(43) Nethercot, A. H. Prediction of Fermi energies and photoelectric thresholds based on electronegativity concepts. *Physical Review Letters* **1974**, *33*, 1088–1091.

(44) Head, T. Scikit-optimize. `https://scikit-optimize.github.io/-[Accessed: 05-09-18]`.

(45) Rasmussen, C. E.; Williams, C. K. I. *MIT Press*, 1st ed.; MIT Press: Cambridge Massachusetts, 2006.

(46) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters* **1996**, *77*, 3865–3868.

(47) Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Physical Review B* **2014**, *89*, 205118.

(48) Ward, L.; Liu, R.; Krishna, A.; Hegde, V. I.; Agrawal, A.; Choudhary, A.; Wolverton, C. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Physical Review B* **2017**, *96*, 024104.

(49) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal Fragment Descriptors for Predicting Electronic Properties of Inorganic Crystals. *Nature Communications* **2017**, *8*, 15679.

(50) Jain, A.; Bligaard, T. Atomic-position independent descriptor for machine learning of material properties. *Physical Review B* **2018**, *98*, 214112.

(51) Walsh, A.; Butler, K. T. Prediction of electron energies in metal oxides. *Accounts of Chemical Research* **2014**, *47*, 364–372.

(52) Pelatt, B. D.; Ravichandran, R.; Wager, J. F.; Keszler, D. a. Atomic solid state energy scale. *Journal of the American Chemical Society* **2011**, *133*, 16852–16860.

(53) Pelatt, B. D.; Kokenyesi, R. S.; Ravichandran, R.; Pereira, C. B.; Wager, J. F.; Keszler, D. A. Atomic solid state energy scale: Universality and periodic trends in oxidation state. *Journal of Solid State Chemistry* **2015**, *231*, 138–144.

(54) Davies, D. W.; Butler, K. T.; Skelton, J. M.; Xie, C.; Oganov, A. R.; Walsh, A. Computer-aided design of metal chalcohalide semiconductors: from chemical composition to crystal structure. *Chemical Science* **2018**, *9*, 1022–1030.

(55) Pauling, L. *The Nature of the Chemical Bond*, 3rd ed.; Cornell University Press: Ithaca, 1960.

(56) Bak, T.; Nowotny, J.; Rekas, M.; Sorrell, C. Photo-electrochemical hydrogen generation from water using solar energy. Materials-related aspects. *International Journal of Hydrogen Energy* **2002**, *27*, 991–1022.

(57) Pinaud, B. A.; Benck, J. D.; Seitz, L. C.; Forman, A. J.; Chen, Z.; Deutsch, T. G.; James, B. D.; Baum, K. N.; Baum, G. N.; Ardo, S.; Wang, H.; Miller, E.; Jaramillo, T. F.; Turner, J. A.; Dinh, H. N. Technical and economic feasibility of centralized facilities for solar hydrogen production via photocatalysis and photoelectrochemistry. *Energy & Environmental Science* **2013**, *6*, 1983–2002.

(58) Mathew, K. et al. Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows. *Computational Materials Science* **2017**, *139*, 140–152.

(59) Jain, A.; Ong, S. P.; Chen, W.; Medasani, B.; Qu, X.; Kocher, M.; Brafman, M.; Petretto, G.; Rignanese, G. M.; Hautier, G.; Gunter, D.; Persson, K. A. FireWorks: A dynamic workflow system designed for high-throughput applications. *Concurrency Computation* **2015**, *27*, 5037–5059.

(60) Küzma, M.; Dominko, R.; Meden, A.; Makovec, D.; Bele, M.; Jamnik, J.; Gaberšček, M. Electrochemical activity of Li2FeTiO4 and Li2MnTiO4 as potential active materials for Li ion batteries: A comparison with Li2NiTiO4. *Journal of Power Sources* **2009**, *189*, 81–88.

(61) Heyd, J.; Scuseria, G.; Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *The Journal of chemical physics* **2003**, *118*, 8207–8215.

(62) Krukau, A. V.; Vydrov, O. A.; Izmaylov, A. F.; Scuseria, G. E. Influence of the exchange screening parameter on the performance of screened hybrid functionals. *Journal of Chemical Physics* **2006**, *125*, 224106.

(63) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **2013**, *68*, 314–319.

(64) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B* **1999**, *59*, 1758–1775.

(65) Kresse, G.; Furthmüller, J. Efficiency of Ab-initio Total Energy Calculations for Metals and Semiconductors Using a Plane-wave Basis Set. *Computational Materials Science* **1996**, *6*, 1550.

(66) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-energy Calculations Using a Plane-wave Basis Set. *Physical Review B* **1996**, *54*, 11169–11186.

(67) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L. A.; Zhou, X.; Burke, K. Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces. *Physical Review Letters* **2008**, *100*, 136406.

Supporting information for: Data-driven Discovery of Photoactive Quaternary Oxides using First-principles Machine Learning

Daniel W. Davies,[†] Keith T. Butler,[‡] and Aron Walsh[*,†,¶]

†*Department of Materials, Imperial College London, Exhibition Road, London SW7 2AZ, UK*

‡*SciML, Scientific Computing Division, Rutherford Appleton Laboratory, Harwell Oxford, Didcot, Oxfordshire OX11 0QX, UK*

¶*Global E^3 Institute and Department of Materials Science and Engineering, Yonsei University, Seoul 120-749, Korea*

E-mail: a.walsh@imperial.ac.uk

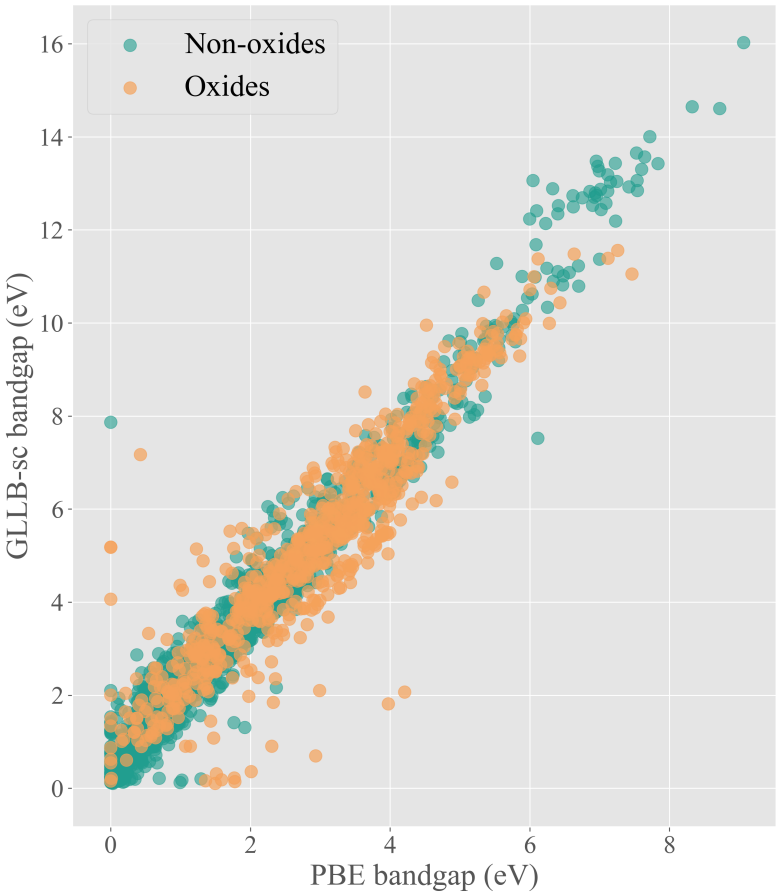

Figure S1: PBE calculated bandgap vs GLLB-sc calculated bandgap for the materials in the database used to train the GBR model. The GLLB-sc values are those taken from the Computational Materials Respository (CMR) while the PBE bandgaps are taken from the Materials Project database.

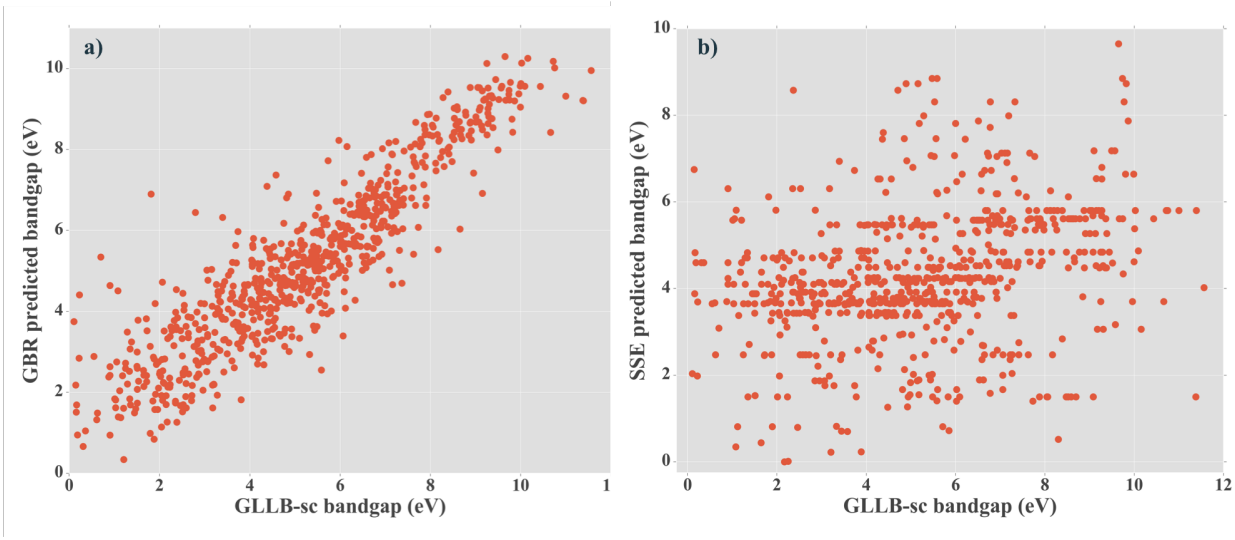

Figure S2: Predictive power of the GBR model (a) and the SSE approach (b) for predicting GLLB-sc bandgaps of the materials in the dataset used to train the GBR model. The values shown in a) relate to estimates from 10-fold cross validation, whereby the predicted bandgaps shown are for each compound when in the 10% of data not used to train the model. The values in b) relate to the bandgap from the limiting SSEs, i.e., the difference between the lowest cation SSE and highest anion SSE.

Table S1: Summary of most stable compounds found after high-throughput DFT calculations. Bandgaps calculated with hybrid DFT that fall within the target window of 1.0 – 2.5 eV are shown in bold.

Number	Formula	spacegroup symbol	E_{hull} (meV/atom)	Bandgap (eV)
1	$MgFe(SO_4)_2$	$P2_1/m$	99	4.07
2	$MgFe(SO_4)_2$	C2/m	11	4.15
3	Li_2MnSiO_5	P4/nmm	86	**2.24**
4	$MnCd(GeO_3)_2$	$P2_1/c$	99	**2.47**
5	$MnCd(GeO_3)_2$	C2/c	99	**1.76**
6	$ZrMnSi_2O_7$	C2	0	4.64
7	$ZrMnSi_2O_7$	P-1	40	4.32
8	$ZrMnSi_2O_7$	P-1	72	3.95
9	$ZrMnSi_2O_7$	$P2_1/m$	3	4.33
10	$ZrMnSi_2O_7$	$P2_1/c$	39	4.40
11	$ZrMnSi_2O_7$	$P2_1/c$	36	5.12
12	Na_2YFeO_4	Pc	79	4.27
13	Na_2YFeO_4	$Pmn2_1$	90	4.33
14	$MnAg(SeO_3)_2$	$Pna2_1$	36	**2.31**
15	Li_2TiMnO_4	$P2_1/c$	38	4.10
16	Li_2TiMnO_4	I-42m	96	4.05
17	Li_2TiMnO_4	$Pna2_1$	40	4.19
18	Li_2TiMnO_4	$Pmn2_1$	11	4.23
19	Li_2TiMnO_4	Pnma	4	4.21
20	Li_2TiMnO_4	$P2_1/c$	31	4.58
21	Li_2TiMnO_4	Pnma	60	4.05
22	$NaCaFeO_3$	$Pna2_1$	61	3.73
23	$NaCaFeO_3$	$P2_1/c$	60	2.87