Active learning for optimum experimental design: insight into perovskite oxides

Lourenco, Maicon Pierre; Tchagang, Alain; Shankar, Karthik; Thangadurai, Venkataraman; Salahub, Dennis R.

Publisher's version / Version de l'éditeur:
https://doi.org/10.1139/cjc-2022-0198
Canadian Journal of Chemistry, 101, 9, pp. 734-744, 2023-04-06

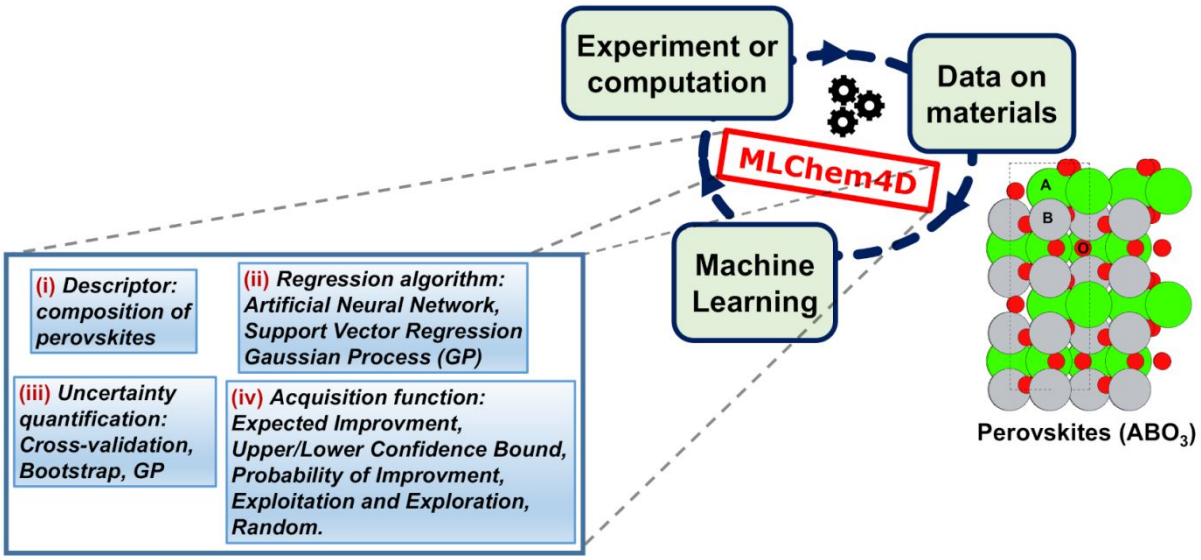
Active Learning for Optimum Experimental Design – Insight into Perovskite Oxides

Maicon Pierre Lourenço[1*], Alain Tchagang[2], Karthik Shankar[3], Venkataraman Thangadurai[4] and Dennis R. Salahub[5]

1- Departamento de Química e Física – Centro de Ciências Exatas, Naturais e da Saúde – CCENS – Universidade Federal do Espírito Santo, 29500-000, Alegre, Espírito Santo, Brasil.
2- Digital Technologies Research Centre, National Research Council of Canada, 1200 Montréal Road, Ottawa, ON, K1A 0R6 Canada.
3- Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, T6G 1H9, Canada.
4- Department of Chemistry, University of Calgary, 2500 University Drive NW, Calgary, AB T2N 1N4, Canada.
5- Department of Chemistry, Department of Physics and Astronomy, CMS Centre for Molecular Simulation, IQST Institute for Quantum Science and Technology, Quantum Alberta, University of Calgary, 2500 University Drive NW, Calgary, AB, T2N 1N4, Canada.

** Address correspondence to:* maiconpl01@gmail.com(MPL).

Table of Contents



Abstract

Finding the optimum material with improved properties for a given application is challenging because data acquisition in materials science and chemistry is time consuming and expensive. Therefore, dealing with small datasets is a reality in chemistry, whether the data is obtained from synthesis or computational experiments. In this work, we propose a new artificial intelligence method based on active learning (AL) to guide new experiments with as little data as possible, for optimum experimental design. The AL method is applied to ABO_3 perovskites where a descriptor based on atomic properties was developed. Several regressor algorithms were employed: artificial neural network, Gaussian process and support vector regressor. The developed AL method was applied in the experimental design of two important materials: non-stoichiometric perovskites ($Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$) due to substituting ionic sites with different concentrations and elements (A = Ca, Sr, Cd; B = Zr, Sn, Hf), aiming at the maximization of the energy storage density; stoichiometric ABO_3 perovskites where different elements are changed in the A and B sites for the minimization of the formation energy. AL for experimental design is implemented in the *machine learning agent for chemistry and design* (MLChem4D) software; which has the potential to be applied in inorganic and organic synthesis (e.g.: search for the optimum concentrations, catalysts, reactants, temperatures and pH to improve the yield) and materials science (e.g.: search the periodic table for the proper elements and their concentrations to improve the materials properties). The latter marks the first MLChem4D application for the design of perovskites.

Keywords

Materials, Perovskites, Machine Learning, Active Learning, Design of Experiment, Transition Metal Oxides.

1. Introduction

Gaining insight about the next experimental conditions to find new materials with improved properties (such as band-gap, mechanical properties, formation energy, convex hull energy and energy storage density) by walking through a huge chemical space[1,2] is challenging because data acquisition is time consuming, error prone and expensive: whether the data is obtained from synthesis and characterization in the laboratory or from modeling or simulations ("computational" experiments). With complex multi-element oxides, the intuition gained by studying related binary oxides is not always applicable due to the emergence of unexpected properties and behavior[3-6], and an alternative method is needed to provide the necessary guidance regarding promising compositions for the specific end-application. Therefore, dealing with small and unbalanced datasets is a reality in chemistry and materials science and methods that assist with optimal experimental design, guided by data, are fundamental.

In this work, we propose a new artificial intelligence method based on active learning (AL) or Bayesian optimization[7] to guide new experiments, with as little data as possible, for optimum experimental design. AL is based on supervised machine learning (ML) where the uncertainty is obtained from ML regression models analytically, as in Gaussian process (GP) methods[8], or from their resampling by K-fold cross-validation[9] (CV) or non-parametric bootstrap[10] (BS), as in artificial neural network (ANN) and support vector regressor (SVR) techniques, for instance. The uncertainty is used to evaluate acquisition functions for decision-making[11], i.e., the experimental conditions in the unobserved space to be chosen for subsequent experiments. More data is acquired during the AL cycle, and the ML regression is improved, increasing the probability of finding the optimum property with the fewest new experiments.

AL methods have been used with first-principles calculations for accelerating high-throughput searches for new alloys[12]. For example, intelligent methods have been used for accelerating materials discovery[13] by an agent-based approach used to decide which experiments to carry out using previous knowledge from the surrogate model together with different exploration-exploitation strategies for the discovery of stable Fe-X binary compounds. Also, AL has been applied for the search of electrocatalysts for CO_2 reduction

and H_2 evolution[14]. There is a rapidly growing literature on "ML for design" of new materials[15-18]; MLChem4D embodies the unique characteristics outlined below.

The novelty of this work is the AL as implemented in the MLChem4D software – developed to provide optimum experimental design for several areas of chemistry: materials design, physical chemistry as well as experimental variable optimization in organic and inorganic synthesis. Here, we present the first application of the software for perovskite oxide design with the following functionalities available (boxes in Fig. 1): (i) we have developed a descriptor for non-modified and modified perovskites by considering just their atomic properties and their elemental concentrations; (ii) several machine learning regression algorithms (i.e. SVR, ANN and GP); (iii) different uncertainty quantification methods: CV and bootstrap (BS) resampling when using SVR or ANN and by Bayesian statistics when using the GP method; (iv) different acquisition functions are available in just one code: pure exploitation – max(μ) or min(μ) – and pure exploration (max(σ)), tradeoff between exploitation and exploration by the lower or upper confidence bound[11] (LCB or UCB), probability of improvement (PI)[11], expected improvement[7] (EI), and random indication of next experiments, for performance efficiency. The MLChem4D software is available and ready to aid experimentalists for new design on reasonable request to the authors. Last but not least, MLChem4D has the potential to be applied for the design of experiments in several areas of chemistry: inorganic and organic synthetic conditions (e.g.: the nature and concentration of catalysts and reactants, temperature, pH, among others); physical chemistry (e.g.: theoretical chemistry modeling, electrochemistry and kinetics experiments) and materials science design. The latter marks the first MLChem4D application for the design of perovskites.

The AL method proposed here is applied to ABO_3 perovskite-type structure oxides where a set of descriptors was developed based on their atomic properties. Several regressor algorithms: ANN, GP and SVR, were employed in the AL for material design. Two systems were investigated by AL. First, non-stoichiometric perovskites ($Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$)[19,20] formed by substituting ionic sites with different elements (A = Ca, Sr, Cd; B = Zr, Sn, Hf) and concentrations, aiming at the maximization of the energy storage density[19]. Second, ABO_3 perovskites where different elements substitute the A and B sites, resulting in materials whose formation energies were used as the objective function.

When using NN and SVR in AL, the uncertainty for AL exploration was obtained by CV[9] or BS[21]. The uncertainty in the GP regressor is obtained analytically[8]. The stoichiometric ABO_3 perovskites were searched when different ionic sites are changed for the minimization of the formation energy, the AL search was performed by using the ANN regressor algorithm with the uncertainty obtained by non-parametric bootstrap. The decision-making used in this work, to guide new experiments, was the expected improvement[7] (EI) (for all systems explored in this work) and the lower confidence bound[22] (LCB) (for the ABO_3 formation energy minimization).

The data set used in the AL in this work for the energy storage density of $Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$ (A = Ca, Sr, Cd; B = Zr, Sn, Hf) was obtained from Refs[19,20]. The database used in the AL for the ABO_3 stoichiometric perovskites for formation energy was obtained from the Materials Project[23].

The AL for experimental design is implemented in the ML agent for chemistry and design (MLChem4D) software, joining the QMLMaterial software developed by the authors using AL for the automatic structural determination[24] based on quantum methods for non-stoichiometric materials[25,26], nanoparticles[27,28] and adsorbate-substrate[29] materials.

2. Methods

2.1 Active learning

AL uses supervised ML algorithms (regressions in this work) and their uncertainties to make decisions for the next conditions for further experiments[11]. The ML regression predictions are improved in each cycle, increasing the probability of finding the optimum material or synthetic condition with the property of interest as response.

The sequence of the AI method based on AL developed in the present work is shown on Fig. 1:

(A) *Data on materials*: a database with N observed materials (synthesized/computed or from legacy data) is set up. At this moment, the descriptor for composition of ABO_3

perovskites is provided (i) for the observed materials (N) and for the unexplored materials ($N^{k}_{virtual}$). Details about the descriptors for ABO_3 perovskites developed in this work will be provided below.

(B) Machine learning or *decision-making (the agent)*: a ML regression model (ii) is obtained for the N observed materials and the uncertainty is computed (iii). From that, the agent – based on decision-making (iv) like the EI (Eq. 1) and the LCB (Eq. 2) – informs *next experimental condition* or new material (in the unexplored space: $N^{k}_{virtual}$), $N^{k+1}_{selected}$, to be measured in the laboratory.

(C) *New experiments or computations*: $N^{k+1}_{selected}$ *new experiments* are realized and the property of the materials evaluated, increasing the database: ($N = N + N^{k+1}_{selected}$). If the new material with its property is not yet satisfactory, the algorithm returns to step (A) with the new measurement added to the initial database: $N = N + N^{k+1}_{selected}$. New ML regression models and uncertainties are obtained (step B). The cycles continues (k) until the property of interest is optimized or the budget has been exhausted.

The blue boxes from (i) to (iv) in Fig. 1 highlight the novelty and features available in the MLChem4D software for optimum experimental design. Moreover, the MLChem4D software has the option of defining in the input different initial data size (N) and indications or selections ($N^{k+1}_{selected}$) of the next materials in $N^{k}_{virtual}$ to be observed or measured in the laboratory, for instance. This features allow the use of previous legacy knowledge – from the researchers or even from literature data – of the material to be designed to be incorporated in the initial database, besides giving flexibility to explore different options for the experimental design.

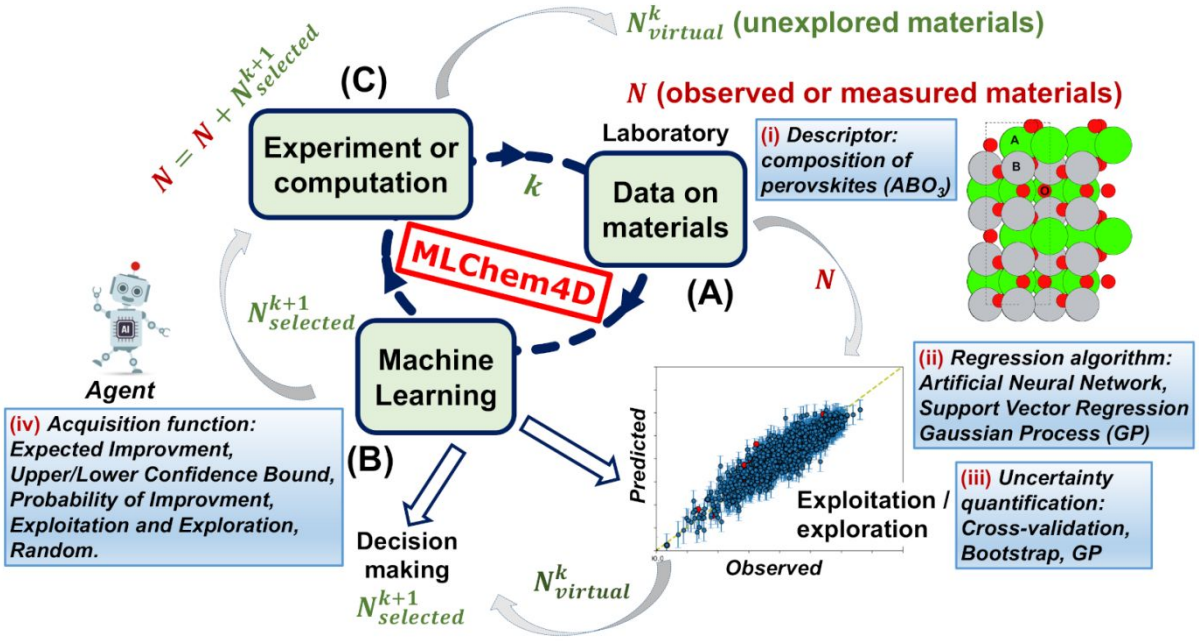


Figure 1- The artificial intelligence (AI) workflow based on active learning (AL) for optimum experimental design as implemented in MLChem4D.

2.2 Decision-making

The acquisition function expected improvement[7,27-29] (EI) for minimization is expressed as:

$$E[I(x^{(j)})] = (f_{min} - \mu(x^{(j)}))\Phi\left(\frac{f_{min} - \mu(x^{(j)})}{\sigma(x^{(j)})}\right) + \sigma(x^{(j)})\phi\left(\frac{f_{min} - \mu(x^{(j)})}{\sigma(x^{(j)})}\right) \quad \textbf{(1)}$$

and for maximization[25,30] is:

$$E[I(x^{(j)})] = (\mu(x^{(j)}) - f_{max})\Phi\left(\frac{\mu(x^{(j)}) - f_{max}}{\sigma(x^{(j)})}\right) + \sigma(x^{(j)})\phi\left(\frac{\mu(x^{(j)}) - f_{max}}{\sigma(x^{(j)})}\right) \quad \textbf{(2)}$$

where *j* means the *j-th* material in the virtual space ($N_{virtual}^{k}$): j = 1, ..., $N_{virtual}^{k}$. The mean and standard deviation are $\mu(x^{(j)})$ and $\sigma(x^{(j)})$, respectively. They can be obtained analytically from the GP method and from CV or BS for NN and SVR.

The lowest (i.e. formation energy) and the highest (i.e. energy storage density) target property observed so far is f_{min} and f_{max}, respectively. $\Phi(\cdot)$ is the cumulative density function and $\phi(\cdot)$ the probability density function. The EI definition can be found

in the literature as early as 1978 in the work of Mockus[31] and has been extensively used in global searches, including in the field of materials design[17,18]. A contour plot of the EI numerical ranges is shown in Fig. 2.

Given the problem of minimization (Eq. 1) as an example, the trade-off between pure exploitation, min(μ), and exploration, max(σ), is contemplated in the EI. In Fig. 3, when T (the target prediction) and σ(X) (the uncertainty) are higher, the EI will be higher as well: $EI = T\Phi(\cdot) + \sigma(X)\phi(\cdot)$, Eq. 1. Then, the search takes place farther from the already observed search surface. On the other hand, by having T higher and σ(X) lower, EI ~ T, which means we are exploiting the already known region of the surface. This is known as exploitation and is more predominant for local minimum searches. Another scenario is when we have T ~ 0 (i.e.: when $f_{min} \sim \mu$) and high σ(X), thus EI ~ σ(X), which is known as exploration and it allows jumping to more distant parts of the search landscape. The same analysis is valid for the problem of maximization: Eq. 2.

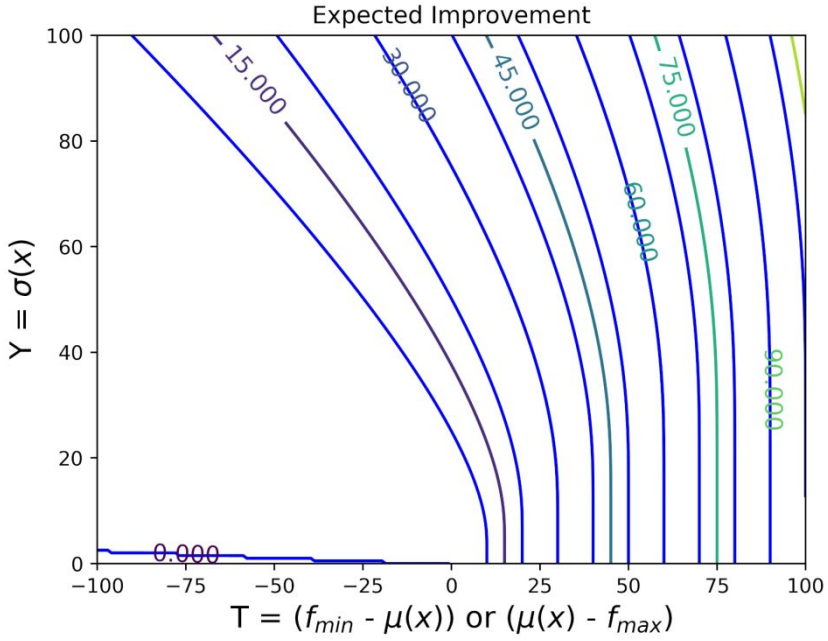


Figure 2- Contour plot of the expected improvement (EI) as a function of the target prediction, T = (f_{min} - μ(X)) or (μ(X) - f_{max}), and the uncertainty, Y = σ(X), Eq. 1. The minimum and maximum property observed so far is f_{min} and f_{max}; respectively; μ and σ are, respectively, the mean and the standard deviation of the prediction obtained by GP. For the NN and SVR predictions (μ), σ is obtained by K-fold cross-validation

(CV) or non-parametric bootstrap (BS). The materials' configurations are represented by the descriptor X.

The lower confidence bound (LCB) for minimization[11,32] (i.e. formation energy) is expressed as:

$$LCB\left[I(x^{(j)})\right] = \mu(x^{(j)}) - C\sigma(x^{(j)}), \textbf{(3)}$$

where C is an empirical parameter that will weigh the amount of uncertainty considered in the decision-making. As C becomes larger, regions farther from the known search surface are considered (exploration). In Fig. 3, a contour plot of the LCB (Eq. 2) numerical ranges, when C=3 (exploration factor), is shown.

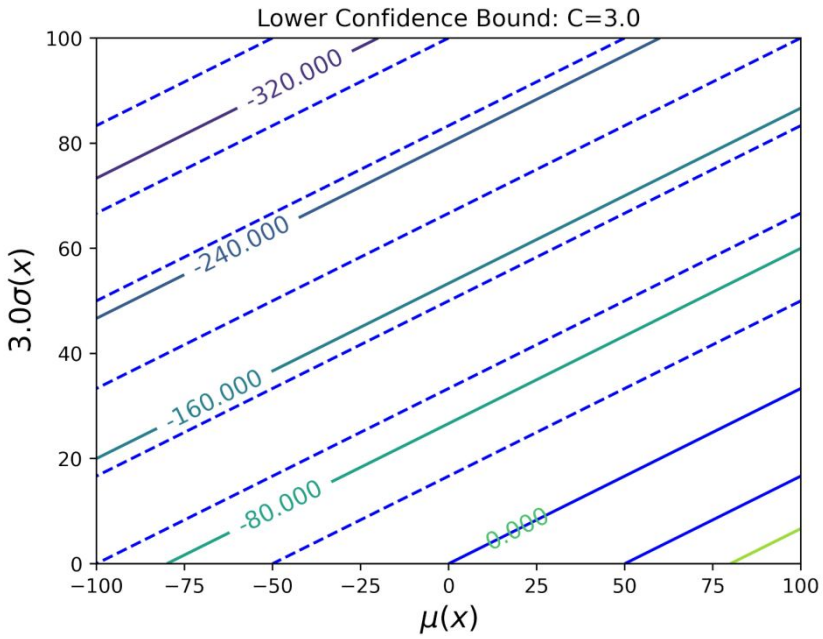


Figure 3- Contour plot of the lower confidence bound (LCB): $\mu(X) - C\sigma(X)$, with C = 3. Where, μ and σ are, respectively, the mean and the standard deviation of the prediction obtained from NN with K-fold cross-validation (CV) or non-parametric bootstrap (BS). X represents the descriptor for different materials' configurations. The continuous and dashed lines in the contour represent the positive and negative regions, respectively.

2.3 The perovskite descriptor

The set of descriptors for non-stoichiometric perovskites – as for $Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$ for energy storage density search – is composed of the weighted atomic properties ($\overline{P}_J$): $\overline{P}_J = \sum_J^{N_{ions}} x_J P_J$; where x_J is the fraction of ions that compose the sites A or B of the ABO_3 perovskite and P_J is the property of the Jth ion at the A or B site. N_{ions} is the number of ions in the A or B sites. Also, the following tolerance factors[33] are used as descriptor: $t_f = (\bar{r}_A + r_O)/[\sqrt{2}(\bar{r}_B + r_O)]$ and $t_t^{new} = \left(\frac{\bar{r}_O}{\bar{r}_B}\right) - Q_A\left[[Q_A - \left(\frac{\bar{r}_A}{\bar{r}_B}\right)/ln\left(\frac{\bar{r}_A}{\bar{r}_B}\right)\right]$; where Q_A is the oxidation state of site A and $\bar{r}_A$, and $\bar{r}_B$ are the average atomic radius of the A and B sites, respectively, obtained by $\overline{P}_J$; r_O, is the ionic ratio of oxygen.

The set of descriptors developed in this work is composed of a list with the following elements:

(I) The tolerance factors t_f and t_t^{new}.

(II) The following properties of the A ($\overline{P}_A$) and B ($\overline{P}_A$) sites of ABO_3: (1) Shannon ionic radius[34,35]; (2) ideal bond length of A-O and B-O; (3) electronegativity; (4) van der Waals radius; (5) first ionization energy; (6) molar volume; (7) atomic number; (8) atomic mass. As they are for the both A and B sites, 2×8 = 16 features result.

(III) The properties related to the A and B sites, as described in II, divided ($\overline{P}_A/\overline{P}_B$) and multiplied ($\overline{P}_A.\overline{P}_B$); resulting in another 16 features. Therefore, the set of descriptors developed is composed of a list with 34 properties. This set of descriptors for the 242 $Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$ perovskites in the database with their respective energy storage density (mJ/cm^3) can be found in the Supporting Information (SI).

Finally, when we have stoichiometric perovskites, as in the case of formation energy investigation of ABO_3, the set of descriptors is obtained by considering $N_{ions} = 1$ and x_J=1 for both A and B ABO_3 perovskite sites. This set of descriptors also is a list with 34 properties. The set of descriptors for the 86 ABO_3 perovskites in the database with their formation energy (eV/atom) can be found in the SI.

The aforementioned atomic properties used to create sets of descriptors for the perovskites were obtained from the Python Materials Genomics[36] (pymatgen) library.

2.4 Machine learning set up

The regression mean (μ) and its uncertainty (σ) were obtained from a non-parametric Bayesian model by using a GP that uses a prior and a covariant function. For the ANN and SVR algorithms, the μ and σ were obtained by CV or BS with 20-fold resampling (K = 20): CV20 and BS20, respectively. From those, the acquisition functions EI or LCB were evaluated to indicate the next perovskites to be investigated in the laboratory. ANN, SVR and GP supervised learning regression algorithms are available in the current version of MLChem4D, which was developed in Python3.x[37] and uses the scikit-learn[38,39] library. In the design loop, the ML regression fit is made on 95 % of the observed data (the training set) and tested on 5 % of it (the test set).

Table 1. The statistics of the hyperparameters used in this work, as defined in the scikit-learn library for the perovskites systems. 95 % of the data were randomly chosen for the training set and 5 % for testing for all systems. **The cost function used to evaluate the quality of the regression was the mean absolute error (MAE) and it is presented for different data sizes obtained by AL. "Hyper." means: hyperparameters used in the ML regression. The MAE for $Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$ and ABO_3 is mJ/cm^3 and eV/atom, respectively.**

System I	Hyper.	MAE train	MAE test	MAE train	MAE test	MAE train	MAE test
		73 data		~100 data		~150 data	
$Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$	ANN10, α = 10.0	4.71	8.33	5.77	8.36	5.94	6.99
	Kernel = ConstantKernel + RBF, const. = 1.0, l=10.0, α = 10.0	1.86	7.56	1.84	3.2	1.87	5.51
	SVR; C = [20; 200; 30]; γ = [0.03; 0.07; 0; 0.10]	5.58	4.53	2.53	4.61	3.30	5.35

System II	Hyper.	MAE train	MAE test	MAE train	MAE test	MAE train	MAE test
		5 data		~30 data		~50 data	
ABO_3	ANN10, α = 1.0	0.10	0.06	0.16	0.28	0.13	0.18

The regression (ANN, GP and SVR) hyperparameters that resulted in the smallest mean absolute error (MAE) for the training and the testing set are shown in Tab. 1. For the ANN regressor, a network with one hidden layer with 10 neurons (NN10) for both perovskite systems was used. The α (L2 regularization) was equal to 10.0 and 1.0 for the $Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$ and ABO_3 systems, respectively. Also, the maximum number of iterations was set to 2×10^5, the LBFGS solver and the rectified linear Unit (ReLU) activation function were used. For GP regression for the $Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$ perovskites, the "const." parameter in the ConstantKernel kernel defines the covariance. *l* defines the length scale of the radial-basis function (RBF) kernel. The α parameters in the GP regressor are added to the diagonal of the kernel matrix during fitting. The combination of two kernels is represented as ConstantKernel + RBF. The SVR regression algorithm used the C = [20; 200; 30] and γ = [0.03; 0.07; 0.10]. The three values in brackets (list) are the optimum C and γ hyperparameters found for the dataset (for $Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$) of size: 73, 100 and 150, respectively. Further information can be found in the documentation of the scikit-learn library[38].

The graphs and statistical analysis were performed with the Python3.x libraries Matplotlib[40], Numpy[41] and Scipy[42].

3. Results and discussion

3.1 Application to energy storage density of $Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$ (A = Ca, Sr, Cd; B = Zr, Sn, Hf) perovskites

In order to evaluate the efficiency of the AL search – as implemented in MLChem4D – in finding the modified perovksite with the maximum energy storage density (GMax) with few data, we emulated a laboratory condition by starting with 73 $Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$ perovskites from an experimental database of 242 perovskites (as presented in SI)

obtained from Ref[20]. The optimum material (the GMax) in the database is $Ba_{0.88}Ca_{0.12}Ti_{0.78}Zr_{0.22}O_3$, which has an energy storage density of 71.7 mJ/cm^3. The 73 initial materials (N = 73) were generated randomly and with different distributions for each independent experimental run. Moreover, the 73 perovskites that compose the initial dataset are selected randomly, considering their properties far from the GMax – below 65 mJ/cm^3, to emulate a challenging optimization scenario. We performed 30 independent experimental runs to test the efficiency of the AL for material discovery. In each AL cycle or iteration (k), $N_{selected}^{k+1}$ indicates the number of selections by the acquisition function or agent in the unexplored space ($N_{virtual}^{k}$) for new $Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$ perovskites observations which is set at: $N_{selected}^{k+1}$ = 1. The one-shot learning (OSL) study was done by one AL cycle (k = *0*), N = 73, $N_{selected}^{1}$ = 30 (EI-30) and $N_{selected}^{1}$ = 60 (EI-60) with the GP model and EI for decision making (GP-EI-30-OSL and GP-EI-60-OSL).

Fig. 4 presents the cumulative success calculated for 30 independent experimental runs as a function of the number of new experiments for different ML regressors: ANN, GP and SVR. The uncertainty for the ANN10 and SVR were obtained with CV20 while for GP it was obtained analytically from the method itself. The EI was used as the acquisition function.

Based on the results in Fig. 4, after 40 new experiments, the AL with the regression algorithms ANN, GP and SVR yielded 70 % cumulative success to maximize the energy storage density and the random search (RS) ~30 %. When AL iterations continue, 100 % cumulative success is attained with 70 new experiments. The RS remains at just ~40 %.

AL with GP yielded the best performance, finding the GMax in 21.9 ± 11.6 new experiments, averaged over 30 runs; AL with ANN and SVR found the GMax in 34.1 ± 10.0 and 33.5 ± 18.5 new experiments, respectively. AL with ANN performed better than SVR. The ANN found the GMax in ~60 new experiments while the SVR found it in 70. In the one-shot-learning study – GP-EI-30-OSL and GP-EI-60-OSL –the agent indicated from the EI 30 and 60 new perovskites to be observed in just one AL cycle (k = *1*). The cumulative success for OSL with EI-30 and EI-60 for 30 independent runs are 53 % and 83 %, respectively (Fig. 4). OSL performed better than the random search. AL considering the

learning in steps ($N_{selected}^{k+1}$ = 1) – ANN10-EI, GP-EI and SVR-EI – outperform the random search as well as OSL, as shown in Fig. 4.

On the whole, the new AL method applied for the global maximization of the energy storage density of doped ABO_3 perovskites outperformed the RS for different AL regression algorithms, as the results of the cumulative success indicate. In general, the AL developed and presented in this work successfully achieved the optimum material ($Ba_{0.88}Ca_{0.12}Ti_{0.78}Zr_{0.22}O_3$) with 55 % of observed data (73 initial and 60 new perovskites guided by AL), considering a search space composed of 242 $Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$ perovskites.

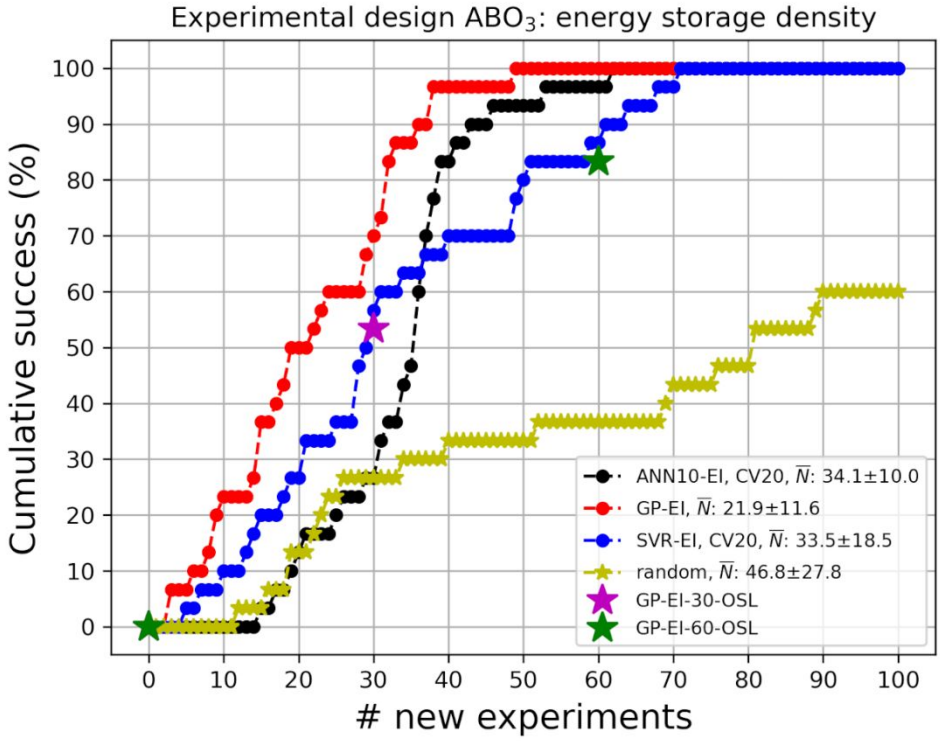


Figure 4- Cumulative success obtained by 30 independent experimental design runs, for the maximization of the energy storage density of doped ABO_3 perovskites, as a function of the number of new experiments. ANN10: Artificial Neural Network with one hidden layer and ten hidden neurons. SVR: Support Vector Regressor. GP: Gaussian process. CV-20: K-fold cross-validation with 20 splits. EI: Expected Improvement. RS: random search. GP-EI-X-OSL (X=30, 60): one-shot learning (OSL) with GP where k = 1 and $N_{selected}^{1}$ = 30 and 60.

Fig. 5 highlights the average energy storage density obtained for 30 independent experimental executions as a function of the number of new experiments indicated by the AL method. In ~50 new experiments the GP-EI has average energy equal to the GMax ($Ba_{0.88}Ca_{0.12}Ti_{0.78}Zr_{0.22}O_3$: 71.7 mJ/cm^3), indicating 100 % cumulative success. The ANN10-EI and SVR-EI achieved the average energy equal to GMax in ~60 and ~70 new experiments, respectively. The RS method presented an average energy storage density smaller than the AL method by ANN10-EI, GP-EI and SVR-EI.

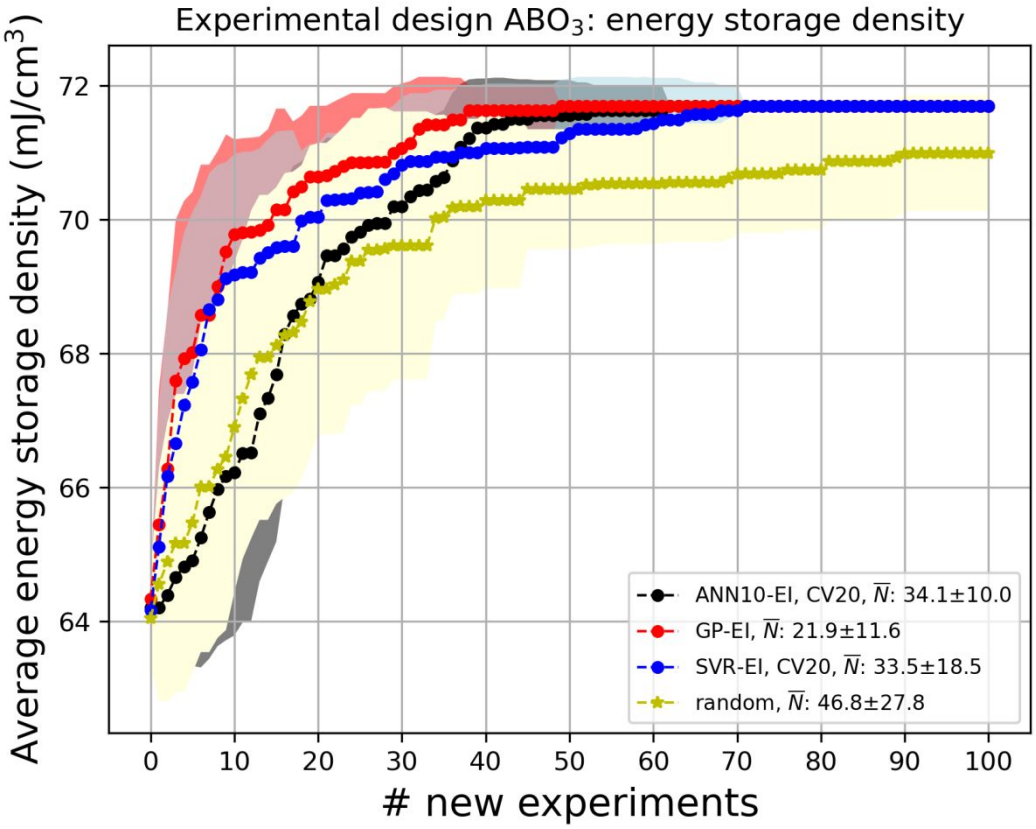


Figure 5- Average energy storage density of doped ABO_3 perovskites obtained by 30 independent runs as a function of the number of new experiments. ANN10: Artificial Neural Network with one hidden layer and ten hidden neurons. SVR: Support Vector Regressor. GP: Gaussian process. CV-20: K-fold cross-validation with 20 splits. EI: Expected Improvement. RS: random search. The shadows around the lines indicate the standard deviations computed for 30 independent runs.

The results of the ANN10, GP and SVR regression algorithms by using the proposed set of descriptors can be found in Fig. 6. The observed and predicted axes have the energy

storage density in mJ/cm^3. The diagonal dashed line is the ideal correlation, when the predicted data by ML is equal to the observed one. The ML graphs were obtained for one independent run of MLChem4D. Also, the data and the ML regressions shown are from one iteration after the GMax, $Ba_{0.88}Ca_{0.12}Ti_{0.78}Zr_{0.22}O_3$, was indicated by the AL for new experiment. The MAE of training set (blue) and test set (red) for ANN10 is 5.356 and 9.355 mJ/cm^3, respectively. The GP has the MAE equal to 2.032 mJ/cm^3 and 6.180 mJ/cm^3 for the training and test set, respectively. The SVR regression presents the MAE for the training and test set equal to 2.838 mJ/cm^3 and 3.018 mJ/cm^3, respectively. The GMax (71.7 mJ/cm^3) found by using the AL is indicated by an arrow in all cases.

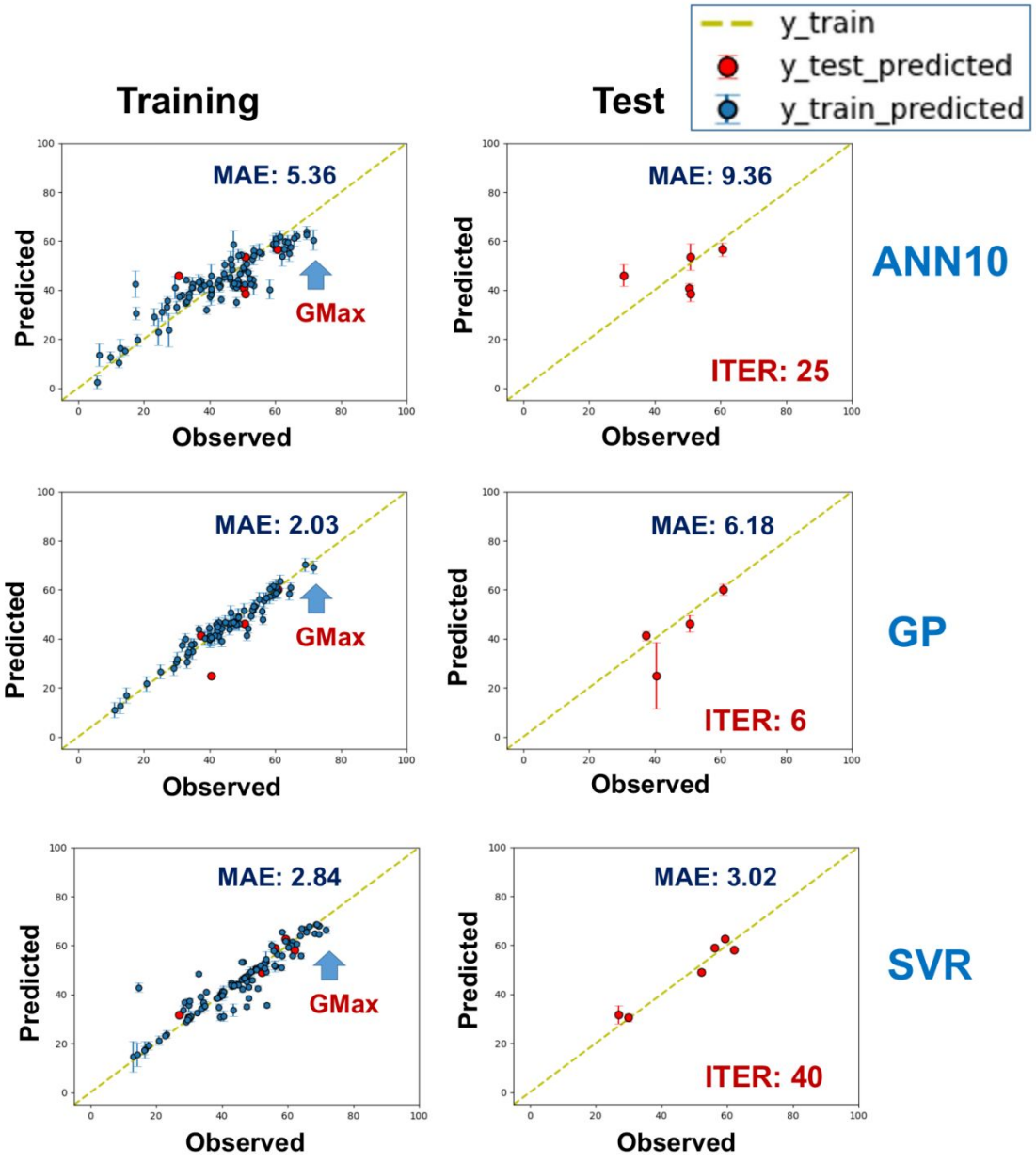


Figure 6- Average predicted and observed energy storage density (mJ/cm³) by ANN10, GP and SVR within the AL cycle. Error bars represent the standard deviation of the average prediction of energy storage density. They were obtained from cross-validation (CV20) in the case of ANN10 and SVR and analytically for GP. Blue points are from the training set (95 %) and red points from the test set (5 %). The above statistical labels mean: the Mean Absolute Error (MAE) is in mJ/cm³. ITER is the next iteration after the indication from decision-making, by AL, of the optimum perovskite, $\mathbf{Ba_{0.88}Ca_{0.12}Ti_{0.78}Zr_{0.22}O_3}$***, to be observed.***

3.2 Application to formation energy of ABO_3 perovskites

The AL, as implemented in MLChem4D, was applied to the minimization of the formation energy of ABO_3 perovskites. The database was composed of 86 perovskites obtained from the Materials Project[23] from computations. The database with the set of descriptors and the formation energies can be found in the SI. The global minimum (GMin) in the database is the perovskite $SrHfO_3$, which has a formation energy of -3.8160 eV/atom. The AL method applied for this problem used as starting point for the AL design just 5 initial data (N = 5), which were obtained randomly and making sure that their formation energies were always above -3.0 eV/atom. This was enough to make the minimum energy in the initial database far from the GMin, to provide a challenging search situation. The AL method was evaluated by considering 30 independent experimental design runs, where each one presented an initial database of different ABO_3 perovskites. In each AL cycles (k) the number of new ABO_3 perovskites in the unexplored space ($N^k_{virtual}$) selected by the acquisition function or the agent ($N^{k+1}_{selected}$) to be observed is: $N^{k+1}_{selected}$ = 1. The one-shot learning (OSL) study for the formation energy optimization was done by considering: one AL cycle ($k = 0$), N = 5, $N^1_{selected}$ = 15 (EI-15) and $N^1_{selected}$ = 30 (EI-30) with the ANN10 regression model, BS20 and EI for decision making (ANN10-EI-15-OSL, ANN10-EI-30-OSL).

The convex hull is defined in Ref[16] as "a surface of the formation energy as a function of the chemical composition that passes through all lowest energy phases that are 'thermodynamically stable': that do not decompose into other phases." As our AL search involves ABO_3 perovskites with defined composition but different A and B ions, the formation energy is an equivalent metric as the maximum decomposition energy (i.e. the

energy above the hull) for material's synthesis stability[13,16]. On the other hand, when performing AL campaigns for the search for stable Mn_xS_y materials, where the x and y are integers whose ratios (compositions) change, the formation energies as a function of the compositions (the convex hull energy) are used as objective function, as shown in Ref [13].

Fig. 7 highlights the cumulative success for 30 independent runs – aimed at finding the perovskite with the minimum formation energy – as a function of the number of new experiments. Here, the AL used an ANN with BS20 for uncertainty quantification. Also, the EI and LCB acquisition functions were used for decision-making: to choose the next materials for further evaluations.

In Fig. 7, the AL with ANN10-EI found the GMin perovskite with 25 new experiments while the AL with ANN10-LCB found it with 28. Until 20 new experiments, the AL with NN10-LCB achieved a better performance than the AL with ANN10-EI. In general, the EI and LCB had similar performance. On the contrary, the RS results show that trying to find the GMin by chance results in very poor performance. For instance, in 30 new experiments it is observed that the AL with ANN-EI and ANN-LCB presented 100 % cumulative success while the RS presented less than 40 %. The one-shot-learning ANN10-EI15-OSL and ANN10-EI30-OSL means that the agent indicated from the EI (decision making) with 15 and 30 new perovskites ($N_{selected}^{k+1}$ = 15 and 30) to be observed in just one AL cycle (*k = 1*), Fig. 7. The cumulative success for OSL with EI-15 and EI-30 obtained for 30 independent runs are 27 % and 50 %, respectively. The OSL performed just only slightly better than the random search. The AL considering the learning in steps ($N_{selected}^{k+1}$ = 1) – ANN10-EI and ANN10-LCB – outperform the random search and the OSL.

Overall, the AL achieved the optimum material ($SrHfO_3$) in 80 % of the independent experiment design runs with just 29 % of observed data (5 initial and 20 new perovskites guided by AL), considering a search space of 86 stoichiometric ABO_3 perovskites. This is an indication of the efficiency of the proposed AL method to deal with small and imbalanced datasets for material discovery.

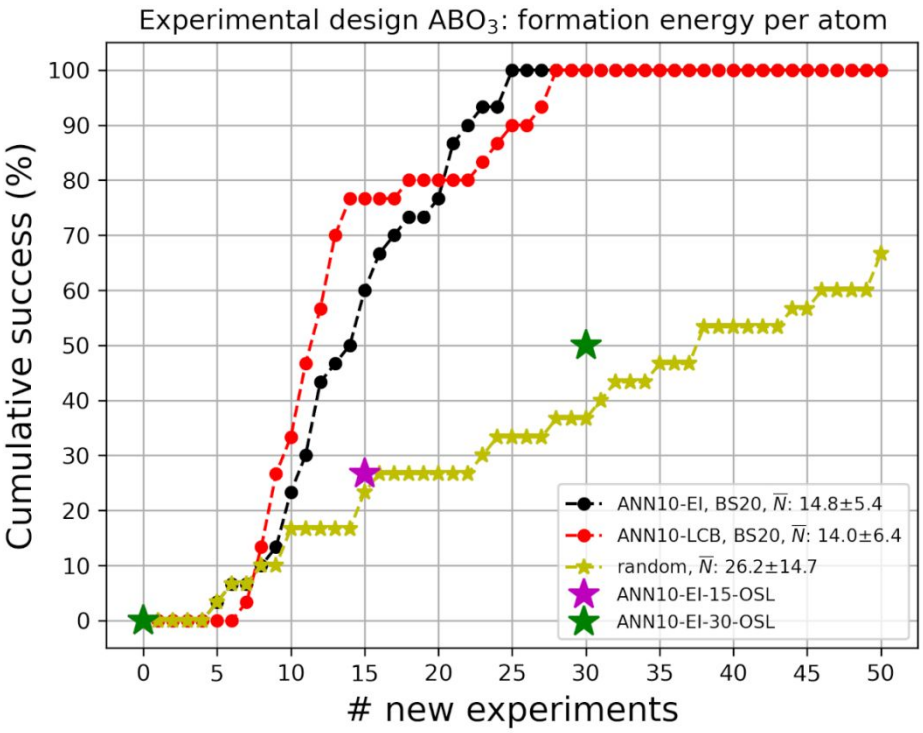


Figure 7 - Cumulative success obtained by 30 independent experimental design runs, for the minimization of formation energy of ABO_3 perovskites as a function of the number of new experiments. ANN10: Artificial Neural Network with one hidden layer and ten hidden neurons. EI: Expected Improvement; LCB: Lower Confidence Bound with C=3. BS20: non-parametric bootstrap with 20 resampling. RS: random search. ANN-10-EI-X-OSL (X=15, 30): one-shot learning (OSL) with ANN-10 where k = 1 and $N^1_{selected}$ = 15 and 30.

The average formation energy per atom (eV/atom) as a function of the number of new experiments, for 30 independent AL runs, is shown in Fig. 8. The RS presented an average formation energy above the one found by ANN10-EI e ANN10-LCB. After 25 new experiments or measurements indicated by the AL, as implemented in MLChem4D, the average formation energy was -3.0 eV/atom, which corresponds to finding the $SrHfO_3$ in all independent runs.

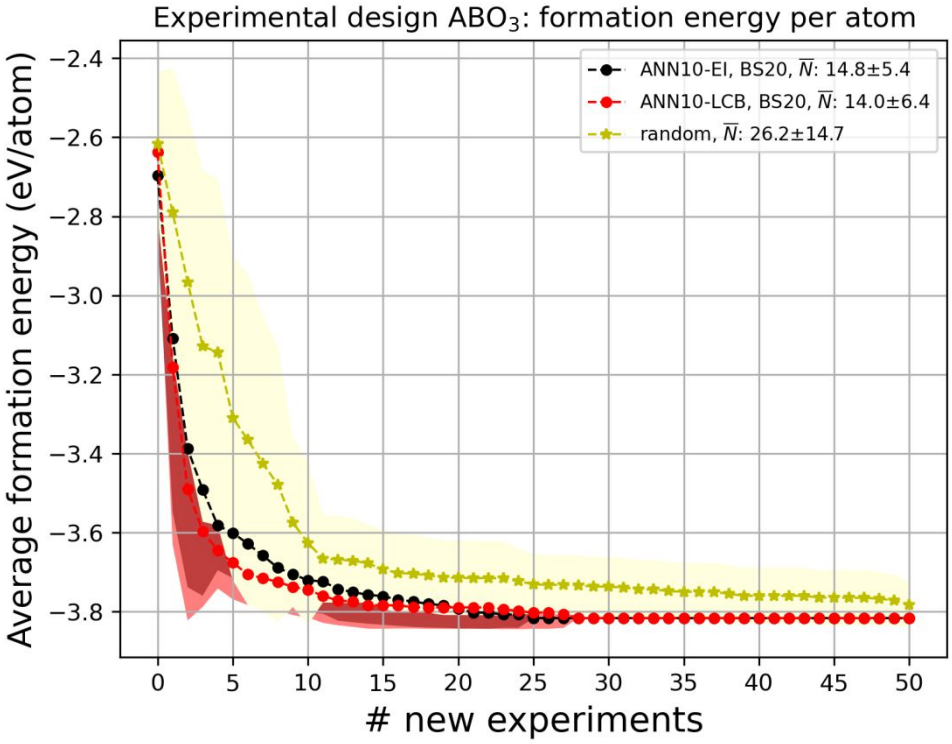


Figure 8- Average formation energy of ABO_3 perovskites obtained by 30 independent runs as a function of the number of new experiments. ANN10: Artificial Neural Network with one hidden layer and ten hidden neurons. EI: Expected Improvement; LCB: Lower Confidence Bound with C=3. BS20: non-parametric bootstrap with 20 resampling. RS: random search. The shadows around the lines indicate the standard deviations computed for 30 independent runs.

In Fig. 9, the results of the ANN10-EI and ANN10-LCB regression algorithms using the proposed set of descriptors are shown. The abscissa (observed) and ordinate (predicted) axes are in eV/atom. The ML graphs were obtained for one independent run of the AL as implemented in MLChem4D. The yellow dashed line in the diagonal corresponds to an ideal correlation: when the predicted property by ML is equal to the observed one. The ML regressions plot are from one iteration after the optimum structure $SrHfO_3$ (GMin: -3.8160 eV/atom) was indicated by the AL. The MAE of training set (blue) and test set (red) for ANN10-EI is 0.189 and 0.226 eV/atom, respectively. The regression plot for ANN10-LCB shows the MAE for the training and test set equal to 0.157 eV/atom and 0.181 eV/atom, respectively. The GMin (-3.8160 eV/atom) found by AL is depicted by an arrow.

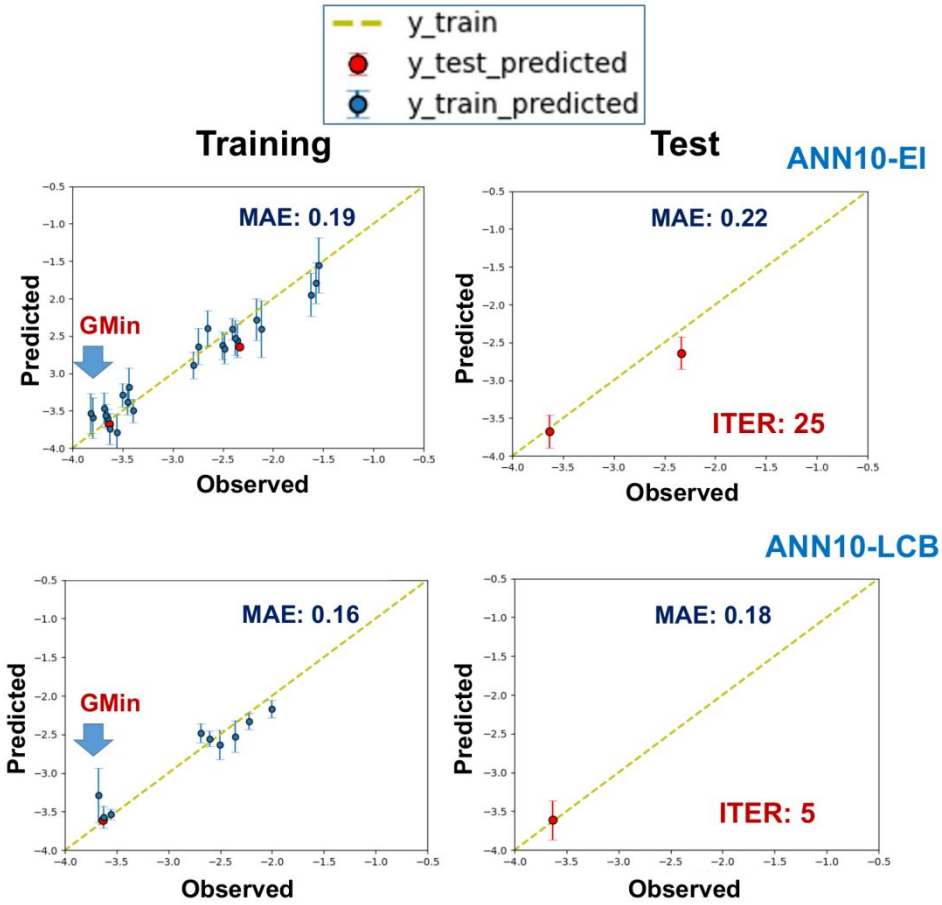


***Figure 9- Average predicted and observed formation energy (eV/atom) by ANN10 within the AL cycle. Error bars represent the standard deviation of the average prediction of formation energy from 20 ML models, obtained from non-parametric bootstrap (BS20). Blue points are from the training set (95 %) and red points from the test set (5 %). The Mean Absolute Error (MAE) is in eV/atom. ITER is the next iteration after the indication from decision-making, by AL, of the* optimum perovskite, $SrHfO_3$, to be observed.**

Finally, the regressions quality reported in Fig. 6 for energy storage density and Fig. 9 for formation energy are in line with other works involving AL for materials design, such as adsorption site search of C_{60}@TiO_2(anatase)[29], structural elucidation of Al^{3+} doping Fe^{3+} sites in goethite (FeOOH)[25] and band gap engineering of apatites[43].

4 Concluding remarks

Artificial intelligence (AI) methods based on active learning (AL) have been shown to be efficient, in particular, for decision-making in situations where small datasets of

materials are available in chemistry and materials science: whether it is from synthesis or from computational modeling or simulation. The new AL developed in the current work is implemented in the MLChem4D software, which was used to obtain the results in this work.

Two materials with their properties were investigated by AL: first, the maximization of energy storage density of $Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$ perovskites whose database was obtained from synthetic (laboratory) data[19,20]. Second, the formation energy of stoichiometric ABO_3 perovskites obtained from a computational database (the Materials Project)[23]. The database used in the study for the two systems was composed of 242 and 86 perovskites, respectively. An atomic descriptor was developed for both problems. Different regression algorithms – ANN, GP and SVR – were employed in the AL design for the maximization of the energy storage density of $Ba_{(1-x)}A_xTi_{(1-y)}B_yO_3$ perovskites, where the EI was used for decision-making. For the formation energy of stoichiometric ABO_3 perovskites, the NN with EI and LCB for decision-making was used, where the uncertainty was obtained by BS. The efficiency of the AL method was compared to RS.

On the whole, the results of the AL method for both aforementioned problems proved to be very efficient, allowing us to walk in regions of chemical space[2] for new discoveries, even in small data scenarios. Thus, as a perspective, an on-the-fly computation of the ABO_3 perovskites formation energy or convex hull energy will be provided from DFT calculations[16] (e.g.: using Quantum ESPRESSO[44] code) by interfacing with MLChem4D. Moreover, the software is ready to be used in the design of new perovskite synthesis in the laboratory, including double-perovskites. Finally, a graphical user interface (GUI) using the Python Tkinter library is under development to aid experimentalists and theoreticians to propose new experimental design in an iterative (AL) and interactive (GUI) manner – not just for inorganic materials design, but for organic synthesis as well.

Finally, the MLChem4D, written in Python3.x, is currently under development with the aim of providing a general AI framework for experimental design to aid chemists in finding synthetic conditions or the materials with the proper elements or composition to be synthesized or to have their properties evaluated, aiming to improve the desired

properties with as few experiments as possible (from the laboratory or from computational modeling or simulations).

Corresponding Author

*Maicon Pierre Lourenço: maiconpl01@gmail.com.

Supporting Information

The database and the set of descriptors of the ABO_3 perovksites used in this work.

Statements & Declarations

Competing Interests

The authors have no financial interests to disclose.

Availability of data and material

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Code availability

The MLChem4D software is available from the corresponding author on reasonable request.

Author contributions

Not applicable.

Acknowledgements

The support of the Brazilian agencies: Fundação de Amparo à Pesquisa do Espírito Santo (FAPES), Conselho Nacional para o Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES) are gratefully acknowledged. Work supported by the National Research Council of Canada, Artificial

Intelligence for Design program and by the Natural Sciences and Engineering Research Council of Canada, Discovery Grant (RGPIN-2019-03976).

Bibliography

(1) Kirkpatrick, P.; Ellis, C. *Nature* **2004**, *432*, 823.
(2) Le, T. C.; Winkler, D. A. *Chem Rev* **2016**, *116*.
(3) Lu, J.; Luo, W.; Feng, J.; Xiang, H. *Nano Letters* **2018**, *18*, 595.
(4) Banerjee, R.; Chatterjee, S.; Ranjan, M.; Bhattacharya, T.; Mukherjee, S.; Jana, S. S.; Dwivedi, A.; Maiti, T. *ACS Sustainable Chemistry & Engineering* **2020**, *8*, 17022.
(5) Kumar, P.; Mulmi, S.; Laishram, D.; Alam, K. M.; Thakur, U. K.; Thangadurai, V.; Shankar, K. *Nanotechnology* **2021**, *32*, 485407.
(6) Ye, X.; Wang, X.; Liu, Z.; Zhou, B.; Zhou, L.; Deng, H.; Long, Y. *Dalton Transactions* **2022**, *51*, 1745.
(7) Jones, D. R.; Schonlau, M.; Welch, W. J. *Journal of Global Optimization* **1998**, *13*, 455.
(8) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. *Chemical Reviews* **2021**, *121*, 10073.
(9) Kohavi, R. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*; Morgan Kaufmann Publishers Inc.: Montreal, Quebec, Canada, 1995, p 1137.
(10) Efron, B. *Biometrika* **1981**, *68*, 589.
(11) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; Freitas, N. d. *Proceedings of the IEEE* **2016**, *104*, 148.
(12) Gubaev, K.; Podryabinkin, E. V.; Hart, G. L. W.; Shapeev, A. V. *Computational Materials Science* **2019**, *156*, 148.
(13) Montoya, J. H.; Winther, K. T.; Flores, R. A.; Bligaard, T.; Hummelshøj, J. S.; Aykol, M. *Chemical Science* **2020**, *11*, 8517.
(14) Tran, K.; Ulissi, Z. W. *Nature Catalysis* **2018**, *1*, 696.
(15) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. *npj Computational Materials* **2019**, *5*.
(16) Schmidt, J.; Shi, J.; Borlido, P.; Chen, L.; Botti, S.; Marques, M. A. L. *Chem Mater* **2017**, *29*.
(17) Zhang, Y.; Apley, D. W.; Chen, W. *Scientific Reports* **2020**, *10*, 4924.
(18) Bassman, L.; Rajak, P.; Kalia, R. K.; Nakano, A.; Sha, F.; Sun, J.; Singh, D. J.; Aykol, M.; Huck, P.; Persson, K.; Vashishta, P. *npj Computational Materials* **2018**, *4*, 74.
(19) Yuan, R.; Tian, Y.; Xue, D.; Xue, D.; Zhou, Y.; Ding, X.; Sun, J.; Lookman, T. *Advanced Science* **2019**, *6*, 1901395.
(20) Tian, Y.; Yuan, R.; Xue, D.; Zhou, Y.; Ding, X.; Sun, J.; Lookman, T. *Journal of Applied Physics* **2020**, *128*, 014103.
(21) Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans*.
(22) Auer, P. *J. Mach. Learn. Res.* **2003**, *3*, 397.
(23) 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. *APL Materials* **2013**, *1*, 011002.
(24) Lourenço, M. P.; Hostaš, J.; Herrera, L. B.; Calaminici, P.; Köster, A. M.; Tchagang, A.; Salahub, D. R. *Journal of Computational Chemistry*, *n/a*.

(25) Lourenço, M. P.; dos Santos Anastácio, A.; Rosa, A. L.; Frauenheim, T.; da Silva, M. C. *Journal of Molecular Modeling* **2020**, *26*, 187.

(26) Lourenço, M. P.; Herrera, L. B.; Hostaš, J.; Calaminici, P.; Köster, A. M.; Tchagang, A.; Salahub, D. R. *Physical Chemistry Chemical Physics* **2022**.

(27) Lourenço, M. P.; Galvão, B. R. L.; Barrios Herrera, L.; Hostaš, J.; Tchagang, A.; Silva, M. X.; Salahub, D. R. *Theoretical Chemistry Accounts* **2021**, *140*, 62.

(28) Lourenço, M. P.; Herrera, L. B.; Hostaš, J.; Calaminici, P.; Köster, A. M.; Tchagang, A.; Salahub, D. R. *Theoretical Chemistry Accounts* **2021**, *140*, 116.

(29) Lourenço, M. P.; Herrera, L. B.; Hostaš, J.; Calaminici, P.; Köster, A. M.; Tchagang, A.; Salahub, D. R. *Journal of Molecular Modeling* **2022**, *28*, 178.

(30) Lookman, T.; Balachandran, P. V.; Xue, D.; Yuan, R. *npj Computational Materials* **2019**, *5*.

(31) Mockus, J.; Tiesis, V.; Zilinskas, A. 1978; Vol. 2, p 117.

(32) Jørgensen, M. S.; Larsen, U. F.; Jacobsen, K. W.; Hammer, B. *J Phys Chem A* **2018**, *122*.

(33) Bartel, C. J.; Sutton, C.; Goldsmith, B. R.; Ouyang, R.; Musgrave, C. B.; Ghiringhelli, L. M.; Scheffler, M. *Science Advances* **2019**, *5*, eaav0693.

(34) Shannon, R. D.; Prewitt, C. T. *Acta Crystallographica Section B* **1969**, *25*, 925.

(35) Shannon, R. *Acta Crystallographica Section A* **1976**, *32*, 751.

(36) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. *Computational Materials Science* **2013**, *68*, 314.

(37) Rossum, G. V.; Drake, F. L. *Python 3 Reference Manual*; CreateSpace, 2009.

(38) Pedregosa, F.; Ga; #235; Varoquaux, l.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; #201; Duchesnay, d. *J. Mach. Learn. Res.* **2011**, *12*, 2825.

(39) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E.; Louppe, G. *Journal of Machine Learning Research* **2012**, *12*.

(40) Hunter, J. D. *Computing in Science & Engineering* **2007**, *9*, 90.

(41) Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. *Nature* **2020**, *585*, 357.

(42) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; Vijaykumar, A.; Bardelli, A. P.; Rothberg, A.; Hilboll, A.; Kloeckner, A.; Scopatz, A.; Lee, A.; Rokem, A.; Woods, C. N.; Fulton, C.; Masson, C.; Häggström, C.; Fitzgerald, C.; Nicholson, D. A.; Hagen, D. R.; Pasechnik, D. V.; Olivetti, E.; Martin, E.; Wieser, E.; Silva, F.; Lenders, F.; Wilhelm, F.; Young, G.; Price, G. A.; Ingold, G.-L.; Allen, G. E.; Lee, G. R.; Audren, H.; Probst, I.; Dietrich, J. P.; Silterra, J.; Webber, J. T.; Slavič, J.; Nothman, J.; Buchner, J.; Kulick, J.; Schönberger, J. L.; de Miranda Cardoso, J. V.; Reimer, J.; Harrington, J.; Rodríguez, J. L. C.; Nunez-Iglesias, J.; Kuczynski, J.; Tritz, K.; Thoma, M.; Newville, M.; Kümmerer, M.; Bolingbroke, M.; Tartre, M.; Pak, M.; Smith, N. J.; Nowaczyk,

N.; Shebanov, N.; Pavlyk, O.; Brodtkorb, P. A.; Lee, P.; McGibbon, R. T.; Feldbauer, R.; Lewis, S.; Tygier, S.; Sievert, S.; Vigna, S.; Peterson, S.; More, S.; Pudlik, T.; Oshima, T. *Nature Methods* **2020**, *17*, 261.

(43) Balachandran, P.; Xue, D.; Theiler, J.; Hogden, J.; Gubernatis, J.; Lookman, T. 2018, p 59.

(44) Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; Dal Corso, A.; de Gironcoli, S.; Fabris, S.; Fratesi, G.; Gebauer, R.; Gerstmann, U.; Gougoussis, C.; Kokalj, A.; Lazzeri, M.; Martin-Samos, L.; Marzari, N.; Mauri, F.; Mazzarello, R.; Paolini, S.; Pasquarello, A.; Paulatto, L.; Sbraccia, C.; Scandolo, S.; Sclauzero, G.; Seitsonen, A. P.; Smogunov, A.; Umari, P.; Wentzcovitch, R. M. *Journal of Physics: Condensed Matter* **2009**, *21*, 395502.