Machine Learning Accelerated Genetic Algorithms for Computational Materials Search

Paul C. Jennings,[1,2] Steen Lysgaard,[3] Jens Strabo Hummelshøj,[4] Tejs Vegge,[3,a)] and Thomas Bligaard[1,2,b)]
[1)] *SUNCAT Center for Interface Science and Catalysis, Department of Chemical Engineering, Stanford University, Stanford, CA 94305, USA*
[2)] *SLAC National Accelerator Laboratory, 2575 Sand Hill Road, CA 94025, USA*
[3)] *Department of Energy Conversion and Storage, Technical University of Denmark, Lyngby, Denmark*
[4)] *Toyota Research Institue, Los Altos, CA 94022, USA*

(Dated: 3 December 2018)

A machine learning (ML) model is trained on-the-fly as a computationally inexpensive energy predictor before analyzing how to augment convergence in Genetic Algorithm (GA)-based approaches by using the ML model as a surrogate. This leads to a machine learning accelerated genetic algorithm (MLaGA) combining robust qualities of the GA with rapid learning of the ML. The MLaGA is used to search for stable, compositionally variant nanoparticle alloys to illustrate its capability for accelerated materials discovery, *e.g.*, nanoalloy catalysts. The MLaGA, in this case, yields a 50-fold reduction in the number of required energy calculations compared to a traditional "brute force" GA. This makes searching through the space of all compositions of a binary alloy particle in a given structure feasible, using density functional theory calculations.

I. INTRODUCTION

The current rate of discovery of clean energy materials remains a key bottleneck in the transition to renewable energy, and computational tools enabling accelerated prediction of the chemical ordering and structure of such materials, *e.g.*, nanoparticle alloys and catalysts, are in high demand.

Genetic algorithms (GAs) are metaheuristic optimization algorithms inspired by Darwinian evolution. Performing crossover, mutation and selection operations, the algorithm progresses a population of evolving candidate solutions. Selecting well designed operators and optimal parameters, GAs have exhibited a high degree of robustness in terms of finding ideal solutions to difficult optimization problems[1,2]. The robustness results from the evolutionary process being able to advance solutions that would have been very difficult to predict *a priori*. Though, GAs often require a large number of function evaluations, resulting from typical offspring not being very "fit" solutions. Modern machine learning (ML) methods have the capacity to fit complex functions in high-dimensional feature spaces while controlling overfitting[3,4]. However, the high-dimensional feature space means that finding an optimum in an ML model is not a simple task. The robustness of the GA is analyzed while accelerating its convergence through integration with an on-the-fly established Gaussian process (GP) regression model of the feature space.

For materials applications, GAs have typically employed (semi-) empirical potentials (EP)[5–11] to describe the potential energy surface (PES).[12–15] The utilization of more accurate methods to describe the PES, such as density functional theory (DFT) has been limited, due to computational cost. To account for the increased computational cost of searching the PES directly with DFT, studies have often been limited in size,[16] though these methods have successfully been used in a number of investigations.[17–25] This study focuses on utilizing the GA to gain an understanding of chemical ordering within larger particles. Searching across a range of compositions is particularly important in the field of materials discovery, where composition can have a profound effect on the desired property e.g. catalytic activity.[26,27] Further, the optimal composition may vary with the size of the nanoparticle. Therefore, the accurate description of chemical ordering is important; where, for certain motifs, the ordering is very complex.[28] Focus is placed on expediting a fast unbiased homotop search by reducing the number of energy evaluations needed to explore the PES and locate the putative global minimum (GM) for a given template structure.

II. RESULTS AND DISCUSSION

The chemical ordering of atoms is optimized for a 147-atom Mackay icosahedral template structure.[29] Searches elucidate the full convex hull of possible Pt_xAu_{147-x} for $x \in [1, 146]$ compositions. A small number of PtAu compositions will preferentially distort to form rosette-icosahedral instead of the Mackay icosahedral structures.[30] The GA locates these structures in a number of cases, though as structure optimization is not the focus of these benchmarks, when the rosette distortion occurs, the calculations are prevented from entering the population preserving the template structure.

The excess energy is used to assign fitness within the GA, as in Equation 1.

[a)] Electronic mail: teve@dtu.dk
[b)] Electronic mail: bligaard@stanford.edu

$$E_{excess} = E_{AB} - \frac{E_A \cdot n_A}{N} - \frac{E_B \cdot n_B}{N} \quad (1)$$

For particles containing a total of N atoms, n_A and n_B are the number of atom type A and B, respectively. E_{AB} is the total energy of the mixed particle, while E_A and E_B are the energies of the pure particles. The number of homotops for each particle rises combinatorially toward the 1:1 composition. The number of possible homotops is given by Equation 2.

$$H_N = \frac{N!}{N_A!N_B!} \quad (2)$$

There are a total of 1.78×10^{44} homotops for all 146 compositions. The total number homotops for each composition is shown in Fig. 1 as well as an example of a randomly ordered icosahedral structure under consideration in this study.

We first run a traditional GA (described in details in the Methods section) to baseline our benchmark and then describe the ML extensions and their results. When using the traditional GA, it is possible to locate the hull of local minima with ∼16,000 energy evaluations. This is significantly lower than the total number of homotops that are present, and thus the number of energy calculations required if a brute-force method was used (1.78×10^{44}). However, this is still typically above the number of energy calculations one would wish to perform if a more expensive energy calculator were being employed. To overcome inefficiencies in this method, the underlying search algorithm is optimized and coupled with machine learning selection. A GP regression model is used to predict excess energies of nanoparticles before employing electronic structure calculations. The squared exponential kernel was utilized for the mapping function, as in Equation 3.

$$k(\mathbf{x}, \mathbf{x}') = exp\left(-\frac{1}{2w^2}\|\mathbf{x} - \mathbf{x}'\|^2\right) \quad (3)$$

The kernel is applied to determine relationships between the fingerprint vectors ($\mathbf{x}$) of two candidates, where $\|x\|$ is the Euclidean L^2-norm and w denotes the kernel width.

The training dataset is comprised of unique numerical fingerprint vectors, with features representing distinct chemical ordering within a particle, based on a simple measure, the averaged number of nearest neighbors, as in Equation 4.

$$f_d = \left[\frac{\#A\text{-}A}{N}; \frac{\#A\text{-}B}{N}; \frac{\#B\text{-}A}{N}; \frac{\#B\text{-}B}{N}; M\right] \quad (4)$$

Where, e.g. $\#A$-A accounts for the number of homoatomic bonds between atom type A. The summed mass (M) is appended to account for compositional changes. The ML model is trained on relaxed nanoparticles, though predictions are based on features generated for the unrelaxed structure. The set of descriptors generated in the fingerprint vector are invariant to small changes to the coordinate system, such as a small expansion or contraction of the lattice resulting from the geometry relaxation. A similar Δ-learning method, has been discussed by von Lilienfeld *et al.*[31]

Within the ML accelerated GA (MLaGA) implementation, exists two tiers of energy evaluation, one by the ML functions giving a predicted fitness and the other by the energy calculator providing the actual fitness. A nested GA has been implemented to search the surrogate model representation, generated by the ML. This acts as a high throughput screening function based solely on predicted fitness, running in the "master" GA. The nested surrogate GA takes the current population and is able to progress through additional search iterations, where evaluation and selection are based only on the current model of the PES. The final population from the nested GA returns unrelaxed candidates to the master GA.

This is well suited to making large steps on the PES without performing expensive energy evaluations. A difficulty when searching with the MLaGA is that convergence criteria typically used in these studies is no longer suitable. The MLaGA methodology is specifically implemented to limit the number of energy evaluations that are performed. Therefore, every candidate in the generation typically progresses the population. This progression within the population continues until the ML routine is unable to find new candidates that are predicted to be better, essentially stalling the search. For this reason, convergence is considered to have been achieved by the point at which the ML routine prevents new candidates from being evaluated. The general MLaGA methodology is shown in Figure 2.

The GA can be run with a pool or generational population. When running the MLaGA with a generational population, a ML model is trained and utilized to search for a full generation of e.g. 150 candidates. When combining the MLaGA with the generational nested GA, a greater number of candidates are generated in total, compared with the traditional GA. However, the majority of candidates generated in the nested GA routine are discarded prior to the expensive energy evaluation step. Therefore, the MLaGA with a nested search, is able to locate the full convex hull of minima in an average of 1200 candidates. It is possible to reduce the total number of energy calculations through the employ of different acceptance criteria. Tournament acceptance was particularly efficient at reducing the number of required energy calculations, reducing to fewer than 600 for the search.

Tournament acceptance is able to improve search efficiency by restricting the number of candidates passed from the nested, to the master GA. To exploit this further, the MLaGA can also be run with a pool based population where the surrogate model is trained for each new data point resulting from an electronic structure calcu-

lation. In this case, the search must progress in serial. Despite the potential for further reduction in the number of calculations required, this may end up being time consuming. This is because, performing the electronic structure calculations cannot be parallelized, as would be possible with the generational population. When this methodology is utilized the number of energy calculations required to search the convex hull is approximately 310.

When training a new model for every energy calculation, it is also possible to exploit uncertainty within the variance distribution on the predicted mean, as in Equation 5.[32]

$$\sigma^2\left(\mathbf{x}^*\right) = \lambda + k\left(\mathbf{x}^*, \mathbf{x}^*\right) - \mathbf{k}^T K^{-1} \mathbf{k} \tag{5}$$

Where a new candidate has the fingerprint vector $\mathbf{x}^*$, $\mathbf{k}$ is the covariance vector between a new data point and the training data set, K is the covariance, or Gram, matrix for the training data and λ is the regularization hyperparameter. In order to progress the search as efficiently as possible, the cumulative distribution function (cdf), as in Equation 6, is used as the fitness of a candidate.

$$\tilde{P}(E_{[x]} < E_{[best]}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\left(E_{[x]} - E_{[best]}\right)^2 / \sigma_x^2} dx \tag{6}$$

When the fitness function also accounts for the variance, it is possible to utilize the inherent uncertainty within a prediction to either exploit the current known information in the model, or to explore unknown regions of the search space.[33] The cdf is calculated up to the current known fittest candidate in the composition. The pool based MLaGA is able to locate the convex hull of stable minima in approximately 280 energy calculations. A comparison of the different methods is in Fig. 3. There are clear advantages to performing the search with the augmented ML method.

To ensure that advantages of the methodologies discussed above were not an artifact of utilizing the less accurate EMT calculator, the MLaGA was tested searching directly on the DFT PES. As a significant reduction in the number of energy calculations is likely to be achieved and parallelization of calculations is favorable, the search is performed with the generational population setup. Utilizing the MLaGA methodology, while allowing the nested search to run for a greater number of generations, it is possible to locate the convex hull of minima with approximately 700 DFT calculations. When optimizing geometries with the DFT calculator, there was a 0 eV barrier to structural rearrangement for a small range of the Au deficient compositions.

The convex hull located for the DFT search is in Fig. 4. The shaded region shows the difference in stability between the distorted structures and the most stable icosahedral structures located. The complete core-shell $Au_{92}Pt_{55}$ structure is located as the most stable for both the EMT and DFT searches. There is good general agreement between the structures obtained elsewhere on the hull, aside from the region of distortion. Further there is broad agreement in the efficiency of the search routines based on the benchmarking and actual searches. Fig 4 also shows the convergence profile as a function of each subsequent DFT calculation. The abrupt bend after around 150 calculations corresponds to a particularly favorable chemical ordering that is distributed to all compositions in the following calculations. This is of course an effect of similar chemical ordering across the whole Au-Pt composition range.

Coupling ML with the GA provides significant advantages in accelerating searches. Performing a search on the surrogate model provides a cheap energy descriptor without requiring expensive electronic structure calculations to assess stability of these nanoparticles. The exact method should be optimized based on the advantages of parallelizing the execution of energy calculations and reducing the total CPU hour cost of the search. A hierarchy of methods have been utilized to reduce the total number of energy calculations required to fully search the convex hull of local minima from 16,000 to around 300.

III. METHODS

A. Computational details

The effective-medium theory (EMT) potential[12] is used as the energy calculator for initial benchmarking. The fast inertial relaxation engine[34] optimization routine is utilized to relax the structures, with forces on all individual atoms minimized to at least 0.1 eV $Å^{-1}$. DFT calculations are performed using GPAW with a real space implementation of the projector-augmented wave method.[35] GPAW is run in the linear combination of atomic orbitals mode[36] with a double zeta basis set and RPBE exchange correlation functional.[37] Calculations are run spin-polarised with a Fermi smearing of 0.05 eV in a non-periodic $32 \times 32 \times 32$ Å unit cell.

B. Traditional Genetic Algorithm

The GA implemented within the Atomic Simulations Environment (ASE) software package[38] has been utilized. A niching fitness function is employed to efficiently search across the full compositional convex hull.[39] When initializing the traditional GA, the population size is set to 150 candidates. The method for selecting parents is handled by roulette wheel selection. Selection probabilities are directly related to the ascribed fitness, which accounts for the stabilities of the nanoparticle. Offspring are created by either mating two parents, or by mutating a single candidate. The mating and mutation routines are mutually exclusive and thus are not allowed to stack i.e. performing crossover and mutation before evaluation. Cut and splice crossover functions, described by Deaven and Ho,[5]

are used to generate new candidates with a call probability of 0.6. Random permutation mutations are utilized with a call probability of 0.2, e.g. swapping the positions of two random atoms of different elemental species. A random swap mutation is also employed with a call probability of 0.2, where one atom type is swapped for another. Convergence criteria is assigned through a lack of progression in the population e.g. the fitness of the population does not change for a number of generations.[6] The GA is run with relatively loose convergence criteria, when there has no observed change in the population for two generations, the search is concluded.

IV. DATA AVAILABILITY

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

V. ACKNOWLEDGEMENTS

The authors acknowledge support of the European Commission under the FP7 Fuel Cells and Hydrogen Joint Technology Initiative grant agreement FP7-2012-JTI-FCH-325327 (SMARTCat) and V-Sustain: The VILLUM Centre for the Science of Sustainable Fuels and Chemicals (no. 9455) from VILLUM FONDEN.

Competing interests: The authors declare no competing interests.

VI. REFERENCES

[1] J. H. Holland, *Adaptation in Natural and Artificial Systems* (The University of Michigan Press, 1975) p. 211.
[2] D. E. Goldberg, *Genetic algorithms in search, optimization, and machine learning* (Addison-Wesley, 1989) p. 412.
[3] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* (Cambridge University Press, 2000) p. 189.
[4] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning* (MIT Press, 2006) p. 248.
[5] D. Deaven and K. Ho, "Molecular geometry optimization with a genetic algorithm." Phys. Rev. Lett. **75**, 288–291 (1995).
[6] R. L. Johnston, "Evolving better nanoparticles: Genetic algorithms for optimising cluster geometries," Dalton Trans. , 4193–4207 (2003).
[7] R. Ferrando, J. Jellinek, and R. L. Johnston, "Nanoalloys: from theory to applications of alloy clusters and nanoparticles." Chem. Rev. **108**, 845–910 (2008).
[8] L. O. Paz-Borbón, R. L. Johnston, G. Barcaro, and A. Fortunelli, "Structural motifs, mixing, and segregation effects in 38-atom binary clusters." J. Chem. Phys. **128**, 134517 (2008).
[9] A. Logsdail, L. O. Paz-Borbón, and R. L. Johnston, "Structures and Stabilities of Platinum-Gold Nanoclusters," J. Comput. Theor. Nanosci. **6**, 857–866 (2009).
[10] S. Lysgaard, D. D. Landis, T. Bligaard, and T. Vegge, "Genetic Algorithm Procreation Operators for Alloy Nanoparticle Catalysts," Top. Catal. **57**, 33–39 (2013).
[11] S. Lysgaard, J. S. G. Mýrdal, H. A. Hansen, and T. Vegge, "A DFT-based genetic algorithm search for AuCu nanoalloy electrocatalysts for CO_2 reduction." Phys. Chem. Chem. Phys. **17**, 28270–28276 (2015).
[12] K. W. Jacobsen, J. K. Norskov, and M. J. Puska, "Interatomic interactions in the effective-medium theory," Phys. Rev. B **35**, 7423–7442 (1987).
[13] R. Gupta, "Lattice relaxation at a metal surface," Phys. Rev. B **23**, 6265–6270 (1981).
[14] A. P. Sutton and J. Chen, "Long-range Finnis-Sinclair potentials," Philos. Mag. Lett. **61**, 139–146 (1990).
[15] J. N. Murrell and R. E. Mottram, "Potential energy functions for atomic solids," Mol. Phys. **69**, 571–585 (1990).
[16] S. Heiles and R. L. Johnston, "Global optimization of clusters using electronic structure methods," Int. J. Quantum Chem. **113**, 2091–2109 (2013).
[17] G. H. Jóhannesson, T. Bligaard, A. V. Ruban, H. L. Skriver, K. W. Jacobsen, and J. K. Nørskov, "Combined electronic structure and evolutionary search approach to materials design," Phys. Rev. Lett. **88**, 255506 (2002).
[18] N. S. Froemming and G. Henkelman, "Optimizing core-shell nanoparticle catalysts with a genetic algorithm," The Journal of Chemical Physics **131**, 234103 (2009).
[19] S. Heiles, A. J. Logsdail, R. Schäfer, and R. L. Johnston, "Dopant-induced 2D-3D transition in small Au-containing clusters: DFT-global optimisation of 8-atom Au-Ag nanoalloys." Nanoscale **4**, 1109–1115 (2012).
[20] J. B. A. Davis, A. Shayeghi, S. L. Horswell, and R. L. Johnston, "The Birmingham parallel genetic algorithm and its application to the direct DFT global optimisation of Ir_N (N = 10-20) clusters." Nanoscale **7**, 14032–8 (2015).
[21] L. B. Vilhelmsen and B. Hammer, "Systematic Study of Au_6 to Au_{12} Gold Clusters on MgO(100) F Centers Using Density-Functional Theory," Phys. Rev. Lett. **108**, 126101 (2012).
[22] U. Martinez, L. B. Vilhelmsen, H. H. Kristoffersen, J. Stausholm-Møller, and B. Hammer, "Steps on rutile TiO 2 (110): Active sites for water and methanol dissociation," Physical Review B **84**, 205434 (2011).
[23] P. C. Jennings and R. L. Johnston, "Structures of Small Ti- and V-Doped Pt Clusters: A GA-DFT Study," Comput. Theor. Chem. **1021**, 91–100 (2013).
[24] C. J. Heard and R. L. Johnston, "A density functional global optimisation study of neutral 8-atom Cu-Ag and Cu-Au clusters," Eur. Phys. J. D **67**, 34 (2013).
[25] A. Shayeghi, D. A. Götz, R. L. Johnston, and R. Schäfer, "Optical absorption spectra and structures of Ag_6^+ and Ag_8^+," Eur. Phys. J. D **69**, 152 (2015).
[26] X. Li, J. Liu, W. He, Q. Huang, and H. Yang, "Influence of the composition of core-shell au-pt nanoparticle electrocatalysts for the oxygen reduction reaction," Journal of Colloid and Interface Science **344**, 132 – 136 (2010).
[27] C. Cui, L. Gan, H.-H. Li, S.-H. Yu, M. Heggen, and P. Strasser, "Octahedral ptni nanoparticle catalysts: Exceptional oxygen reduction activity by tuning the alloy particle surface composition," Nano Letters **12**, 5885–5889 (2012).
[28] R. Ferrando, "Symmetry breaking and morphological instabilities in core-shell metallic nanoparticles." J. Phys. Condens. Matter **27**, 013003 (2015).
[29] O. Echt, K. Sattler, and E. Recknagel, "Magic numbers for sphere packings: Experimental verification in free xenon clusters," Phys. Rev. Lett. **47**, 1121–1124 (1981).
[30] A. L. Gould, K. Rossi, C. R. A. Catlow, F. Baletto, and A. J. Logsdail, "Controlling structural transitions in auag nanoparticles through precise compositional design," The Journal of Physical Chemistry Letters **7**, 4414–4419 (2016).
[31] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Big data meets quantum chemistry approximations: The Δ-machine learning approach," Journal of Chemical Theory and Computation **11**, 2087–2096 (2015).
[32] A. Girard, C. Rasmussen, J. Quiñonero-Candela, and

R. Murray-Smith, "Gaussian process priors with uncertain inputs - application to multiple-step ahead time series forecasting," in *Neural Information Processing Systems* (2003) pp. 529–536.

[33] M. S. Jørgensen, U. F. Larsen, K. W. Jacobsen, and B. Hammer, "Exploration Versus Exploitation in Global Atomistic Structure Optimization," The Journal of Physical Chemistry A **122**, 1504–1509 (2018).

[34] E. Bitzek, P. Koskinen, F. Gähler, M. Moseler, and P. Gumbsch, "Structural Relaxation Made Simple," Phys. Rev. Lett. **97**, 170201 (2006).

[35] J. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Dulak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. A. Hansen, H. H. Kristoffersen, M. Kuisma, A. H. Larsen, L. Lehtovaara, M. Ljungberg, O. Lopez-Acevedo, P. G. Moses, J. Ojanen, T. Olsen, V. Petzold, N. A. Romero, J. Stausholm-Moller, M. Strange, G. A. Tritsaris, M. Vanin, M. Walter, B. Hammer, H. Hakkinen, G. K. H. Madsen, R. M. Nieminen, J. K. Nørskov, M. Puska, T. T. Rantala, J. Schiøtz, K. S. Thygesen, and K. W. Jacobsen, "Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method," Journal Of Physics-Condensed Matter **22**, 253202 (2010).

[36] A. H. Larsen, M. Vanin, J. J. Mortensen, K. S. Thygesen, and K. W. Jacobsen, "Localized atomic basis set in the projector augmented wave method," Physical Review B **80**, 195112 (2009).

[37] B. Hammer, L. Hansen, and J. Nørskov, "Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals," Physical Review B **59**, 7413–7421 (1999).

[38] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, "The atomic simulation environment—a python library for working with atoms," Journal of Physics: Condensed Matter **29**, 273002 (2017).

[39] B. Sareni and L. Krahenbuhl, "Fitness sharing and niching methods revisited," IEEE Trans. Evol. Comput. **2**, 97–106 (1998).

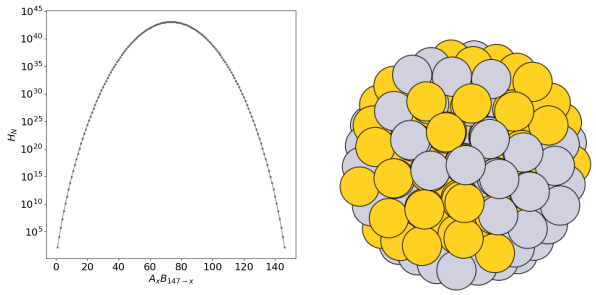

FIG. 1: The homotop optimization problem for the 147 Mackay icosahedral nanoparticle.

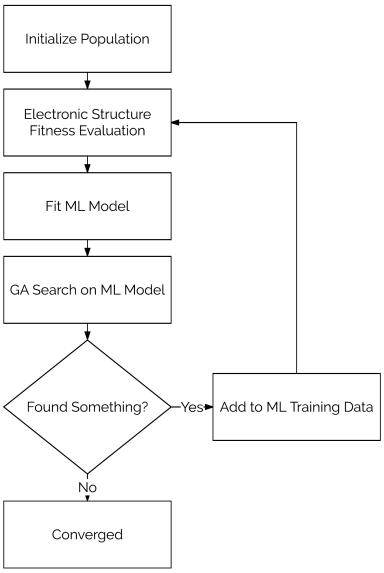

FIG. 2: Flowchart for the MLaGA method. As specified in the text the method only terminates when ML assisted GA fail to produce candidates that are predicted to improve the population.

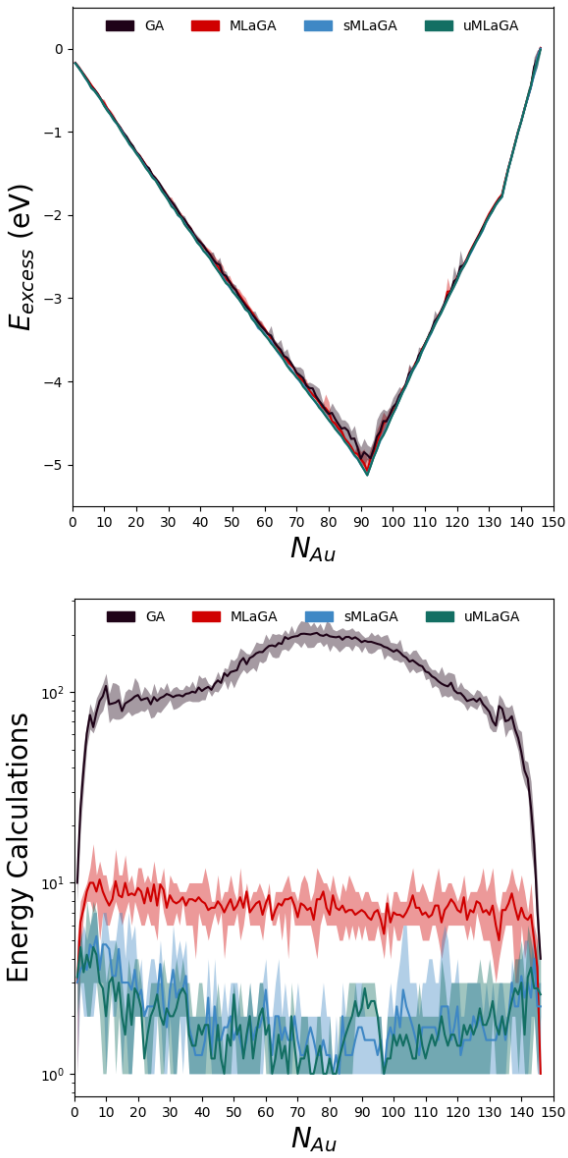

FIG. 3: (Top) The convex hull located with the MLaGA employing the EMT calculator. (Bottom) The number of energy calculations as a function of composition of the particle. Data is shown for the traditional GA (GA), the ML accelerated GA (MLaGA), the serialized MLaGA (sMLaGA) and the MLaGA utilizing uncertainty (uMLaGA). The dark lines and the shaded areas show the average and variation of five repeated searches respectively.

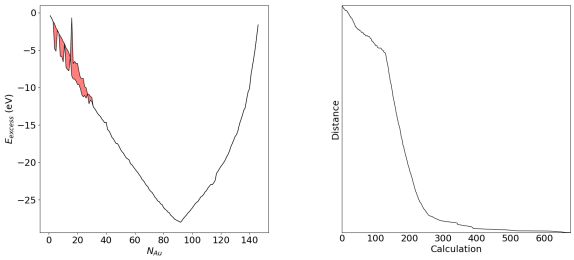

FIG. 4: (Left) The convex hull located with the MLaGA employing the DFT calculator. (Right) The convergence profile for the GA search. The distance is the cumulative energy deviation from the correct convex hull, it is plotted against each energy calculation i.e. it gives an indication of the energy gain of each calculation.