Genetic algorithms for computational materials discovery accelerated by machine learning

C. Jennings, Paul; Lysgaard, Steen; Hummelshøj, Jens Strabo; Vegge, Tejs; Bligaard, Thomas

Published in:
n p j Computational Materials

Link to article, DOI:
10.1038/s41524-019-0181-4

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Materials discovery is increasingly being impelled by machine learning methods that rely on pre-existing datasets. Where datasets are lacking, unbiased data generation can be achieved with genetic algorithms. Here a machine learning model is trained on-the-fly as a computationally inexpensive energy predictor before analyzing how to augment convergence in genetic algorithm-based approaches by using the model as a surrogate. This leads to a machine learning accelerated genetic algorithm combining robust qualities of the genetic algorithm with rapid machine learning. The approach is used to search for stable, compositionally variant, geometrically similar nanoparticle alloys to illustrate its capability for accelerated materials discovery, e.g., nanoalloy catalysts. The machine learning accelerated approach, in this case, yields a 50-fold reduction in the number of required energy calculations compared to a traditional “brute force” genetic algorithm. This makes searching through the space of all homotops and compositions of a binary alloy particle in a given structure feasible, using density functional theory calculations.

**RESULTS AND DISCUSSION**

Icosahedral particles

The chemical ordering of atoms is optimized for a 147-atom Mackay icosahedral template structure. Searches elucidate the full convex hull of possible Pt_{x}Au_{147−x} for x ∈ [1, 146] compositions. The convex hull is defined as the line connecting the lowest excess energy of each stable composition. A composition can be unstable if the homotop with the lowest excess energy lies above the line. The number of homotops for each particle rises combinatorially toward the 1:1 composition. The number of possible homotops is given by Eq. 1.

\[ H_{n} = \frac{N!}{N_{1}!N_{2}!} \]

There are a total of 1.78 × 10^{64} homotops for all 146 compositions. The total number homotops for each composition

---

**ARTICLE OPEN**

Genetic algorithms for computational materials discovery accelerated by machine learning

Paul C. Jennings1,2, Steen Lysgaard3, Jens Strabo Hummelshøj4, Tejs Vegge5 and Thomas Bligaard1,2

Materials discovery is increasingly being impelled by machine learning methods that rely on pre-existing datasets. Where datasets are lacking, unbiased data generation can be achieved with genetic algorithms. Here a machine learning model is trained on-the-fly as a computationally inexpensive energy predictor before analyzing how to augment convergence in genetic algorithm-based approaches by using the model as a surrogate. This leads to a machine learning accelerated genetic algorithm combining robust qualities of the genetic algorithm with rapid machine learning. The approach is used to search for stable, compositionally variant, geometrically similar nanoparticle alloys to illustrate its capability for accelerated materials discovery, e.g., nanoalloy catalysts. The machine learning accelerated approach, in this case, yields a 50-fold reduction in the number of required energy calculations compared to a traditional “brute force” genetic algorithm. This makes searching through the space of all homotops and compositions of a binary alloy particle in a given structure feasible, using density functional theory calculations.

**RESULTS AND DISCUSSION**

Icosahedral particles

The chemical ordering of atoms is optimized for a 147-atom Mackay icosahedral template structure. Searches elucidate the full convex hull of possible Pt_{x}Au_{147−x} for x ∈ [1, 146] compositions. The convex hull is defined as the line connecting the lowest excess energy of each stable composition. A composition can be unstable if the homotop with the lowest excess energy lies above the line. The number of homotops for each particle rises combinatorially toward the 1:1 composition. The number of possible homotops is given by Eq. 1.

\[ H_{n} = \frac{N!}{N_{1}!N_{2}!} \]

There are a total of 1.78 × 10^{64} 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. A small number of PtAu compositions will preferentially distort to form rosette-icosahedral instead of the Mackay icosahedral structure. Other distortions have previously been observed for the AuPt system but not in this study. The GA locates the rosette distorted structures in a number of cases, though as structure optimization is not the focus of these benchmarks, when the distortion occurs, the calculations are prevented from entering the population preserving the template structure.

Traditional GA

We first run a traditional GA (described in detail in the Methods section “Traditional GA”) 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 candidate minimizations. 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 \(1.78 \times 10^{44}\). 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 ML selection. A GP regression model is used to predict excess energies of nanoparticles before employing electronic structure calculations. A discussion of the GP model is given in the Methods section “GP regression model.”

ML-accelerated GA (MLaGA)

Within the 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 Fig. 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 by employing different acceptance criteria. Tournament acceptance was particularly efficient at reducing the number of required energy minimizations, reducing to <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 calculation. 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 minimizations required to search the convex hull is approximately 310. When training a new model for every energy calculation, it is also possible to estimate and take advantage of the model prediction uncertainty, as discussed in the Methods section “GP regression model.” Utilizing the cumulative distribution function (Eq. 6) as a candidate’s fitness 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.

DFT verification
To ensure that advantages of the methodologies discussed above were not an artifact of utilizing the less accurate effective-medium theory (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 set-up. 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 shown in Fig. 4a. The shaded region shows the difference in stability between the distorted structures and the most stable icosahedral structures located. The complete core–shell Au$_{33}$Pt$_{33}$ 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. Figure 4b 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 composition 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 has been utilized to reduce the total number of energy minimizations required to fully search the convex hull of local minima from 16,000 to around 300.

We conclude this section with a discussion of how to include geometrical optimization in this framework. First one would add operators that move the atoms and use a fingerprint able to invariantly describe the local geometrical environment. However, if any appreciable rearrangement takes place during relaxation the fingerprint is no longer reliable, leaving the ML-predicted energies wrong. It is possible to limit the relaxation to maintain reliable fingerprints; this would require that the GA runs through more candidates as the steps taken on the PES will be smaller. Another possibility is to utilize ML schemes that can also take care of relaxation, some examples already exist for coupling these with a GA.

**METHODS**

**Computational details**

The EMT potential is used as the energy calculator for initial benchmarking. The fast inertial relaxation engine 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. GPAW is run in the linear combination of atomic orbital mode with a double zeta basis set and RPBE exchange correlation functional. Calculations are run spin-polarized with a Fermi smearing of 0.05 eV in a non-periodic 32 × 32 × 32 Å unit cell.

**Traditional GA**

The GA implemented within the Atomic Simulations Environment (ASE) software package has been utilized. The excess energy is used when determining fitness within the GA, as in Eq. 2.

$$E_{\text{excess}} = E_{AB} - \frac{E_A - n_A}{N} - \frac{E_B - n_B}{N}$$

(2)

For particles containing a total of $N$ atoms, $n_A$ and $n_B$ are the number of atoms of types A and B, respectively. $E_A$ is the total energy of the mixed particle, while $E_A$ and $E_B$ are the energies of the pure particles. To efficiently search across the full compositional convex hull, we employ a fitness function based on a niching routine. The candidates are grouped according to the composition and the fitness assigned within each composition niche is based on the excess energy (Eq. 2). The fittest
individuals for each composition, i.e., across niches, are given equal fitness. This negates the energy penalties that would otherwise prevent the algorithm from searching for minima of all possible compositions and bias the search toward a narrow composition window comprising the lowest excess energies.

When initializing the traditional GA, the population size is set to 150 candidates. This will, due to niching, keep all compositions in the population. 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, 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. The convergence criterion 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, and when there was no observed change in the population for two generations, the search is concluded.

GP regression model

The squared exponential kernel was utilized for the mapping function, as in Eq. 3.

\[
k(x, x') = \exp \left(-\frac{1}{2\sigma^2} \|x - x'\|^2\right)
\]

(3)

The kernel is applied to determine relationships between the fingerprint vectors \(\mathbf{x}\) of two candidates, where \(\|x\|\) is the Euclidean L2-norm and \(\sigma\) 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 Eq. 4.

\[
f_d = \frac{\text{#A} - \text{#A}}{N} - \frac{\text{#B} - \text{#A}}{N} - \frac{\text{#B} - \text{#B}}{N}
\]

(4)

where, e.g., \(\text{#A} - \text{#A}\) 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 species. A random swap mutation is also employed with a call probability of 0.2, where one atom type is swapped for another. The convergence criterion is assigned through a lack of progression in the population, e.g., swapping the positions of two random atoms of different elemental types. A similar \(\Delta\)-learning method, has been discussed by von Lilienfeld et al.48

Within the pool-based GA operation, it is possible to retrain the model after every evaluation. We exploit this to calculate the uncertainty within the variance distribution on the predicted mean as in Eq. 5.

\[
s\sigma(x') = \lambda + K^{-1}K^{-1}K
\]

(5)

where a new candidate has the fingerprint vector \(x'\), \(K\) 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 \(\lambda\) is the regularization hyperparameter added in order to evaluate the uncertainty of the full model.49 In order to progress the search as efficiently as possible, the cumulative distribution function \(CDF\), as in Eq. 6, is used as the fitness of a candidate.

\[
\tilde{P}(E_{x'} < E_{best}) = \frac{1}{\pi} \int_{-\infty}^{E_{x'}} e^{-((E-E_{best})^2)/2\sigma^2} \, dx
\]

(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 currently known information in the model or to explore unknown regions of the search space.50 The CDF is calculated up to the currently known fittest candidate in the composition.

DATA AVAILABILITY

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

ACKNOWLEDGEMENTS

The authors acknowledge support of the European Commission under the FP7 Fuel Cells and Hydrogen Joint Technology Initiative grant agreement FP7-2012-ITI-FCCH-325327 (SMARTCarat) and V-Sustain: THE VILLUM CENTRE FOR THE SCIENCE OF SUSTAINABLE FUELS AND CHEMICALS (NO. 9455) FROM VILLUM FONDEN. Support from the US Department of Energy Office of Basic Energy Science to the SUNCAST Center for Interface Science and Catalysis is gratefully acknowledged. The authors acknowledge the Toyota Research Institute Accelerated Materials Design and Discovery Program and the AiMade (Autonomous Materials Discovery) Program at DTU Energy.