Computational Prediction of Muon Stopping Sites

Leandro Liborio,[1] Simone Sturniolo,[1] and Dominik Jochym[1]

Scientific Computing Department, Rutherford Appleton Laboratory, STFC

(Dated: 2 February 2018)

The stopping site of the muon in a muon-spin relaxation experiment (μ+SR) is in general unknown. There are some techniques that can be used to guess the muon stopping site, but they often rely on approximations and are not generally applicable to all cases. In this work, we propose a purely theoretical method to predict muon stopping sites in crystalline materials from first principles. The method is based on a combination of ab initio calculations, random structure searching and machine learning, and it has successfully predicted the Mu_T and Mu_{BC} stopping sites of muonium in Si, Diamond and Ge, as well as the muonium stopping site in LiF, without any recourse to experimental results. The method makes use of Soprano, a Python library developed to aid ab-initio computational crystallography, that was publicly released and contains all the software tools necessary to reproduce our analysis.

I. INTRODUCTION

In a muon-spin relaxation experiment (μ+SR) spin-polarized positive muons are implanted in a sample to probe its local static and dynamic magnetic properties. μ+SR is a sensitive probe of magnetism, but one of its limitations is not knowing the site of implantation of the muon, which limits its use for measuring magnetic moments or for comparing different magnetic structures.

There are some techniques that can determine the muon stopping sites, but these techniques are limited to certain specific cases. For instance, in some materials, the determination of the muon stopping site was possible by using accurate experimental studies of the muon frequency shift in an applied magnetic field[1]. In particular, the so-called channelling and blocking techniques have produced some experimental information on the location of the muon site in semiconductor materials[2].

Nonetheless, the number of examples where the muon site can be determined by experimental means alone is limited, and in materials such as Fe_3O_4a combination of experiments and calculations has been used to determine the muon stopping sites[3]. In these examples, theoretical calculations are a cheap way of testing potential muon stopping sites. For instance, the muon is placed in a particular site and one of the hyperfine coupling constants (HFCC) of muonium in that site is calculated and compared with the experimental results to check the validity of the muon stopping site. However, in most of these examples there are a few starting guess sites for the muon, which limits the number of test calculations. This limiting of the potential stopping sites is not possible in most systems and, therefore, it is not always possible to determine the muon stopping site using these combined techniques.

Whenever the candidate muon sites cannot be assigned by an educated guess that uses experimental data, we need to explore all the possible interstitial sites using a theoretical method. This can be a radically different process depending on whether the muon remains in its charged state (μ^+) or captures an electron forming the pseudo atom muonium (Mu). One of the currently most popular methods for predicting the stopping site is called the Unperturbed Electrostatic Potential (UEP) method. This method relies on the analysis of the electrostatic potential of the host material, which is obtained from DFT simulations. The UEP method is not fully reliable. For instance, it has predicted the stopping sites of charged muons in materials such as RFeAsO (R=La, Ce, Pr, and Sm)[4] and LaCoPO[5], but could not predict the stopping sites of neither μ^+ nor Mu in fluorides[6,7]. In the case of muonium, it has been suggested that the screening provided by the electron makes the muon less sensitive to the electrostatic potential of the host[6,8,9] . With regards to charged muons, it has been proposed that the inability to account for the muon-host interactions is what decreases the efficiency of the UEP method in predicting the muon stopping sites[6,7]. Hence, we believe that an alternative method is needed, and we propose one that combines machine learning methods with a computational technique known as Ab Initio Random Structure Searching (AIRSS)[10,11].

The AIRSS approach makes use of random structure generation to perform an unbiased search of stable crystal structures for a given stoichiometry[12]. Constraints encoding whatever knowledge we have on the system, like symmetries, can be included. Our aim is to automate as much as possible the process of extracting information about not only the absolute minimum of the system, but also the relative ones. In this work, we focus on the paramagnetic states formed by muons in semiconductors. In particular, we revisit the case of muons in pure Si, Ge, Diamond and LiF[6], and use a combination of computational methods to retrieve the known muon stopping sites for these systems. We first follow the established AIRSS methodology by generating a number of muoniated random structures and running ab-initio geometry optimization calculations on them. Then we apply our new machine learning methods to analyse the resulting structures and use them to predict the muon stopping sites.

II. METHODOLOGY

A. Structure Generation

Ab-Initio Random Structure Search, or AIRSS[13], has been demonstrated as a successful approach to many complex problems in crystalline structure detection; by comparison with most of them, the problem of finding the optimal stopping site of muonium in a crystal is much simpler, since it only involves a configuration space defined by three degrees of freedom, rather than dozens or hundreds. For this reason we decided to apply this method as a starting point for our calculations.

The AIRSS approach consists of generating random initial configurations and then using CASTEP (or another equivalent DFT code) to perform a geometry relaxation on each of them. The rationale for this approach is that studies of potential energy surfaces for crystals shows that the attraction basins of the various local minima tend to have a greater hypervolume the lower the minimum's energy is[14–16]. In other words, by sampling in a uniform random way the entire configuration space, it is statistically more likely that the great majority of starting configurations will relax to a relatively stable energy minimum. However, given the expense of performing these calculations, it is sensible to trim down some of the most unreasonable starting configurations to avoid wasting time. This was done in two steps:

- First, all configurations in which the muonium collided with an existing atom were eliminated. This is part of the standard AIRSS methodology;

- Second, all configurations were examined, and when a pair was found with starting positions closer than a given limit, one of the two was eliminated. This was done to avoid running redundant calculations on configurations that would likely slide into the same local minimum, by approximating what

Atom	r^{min}_{A-Mu} (Å)	r^{min}_{Mu-Mu} (Å)
C	0.7	0.7
Si	1.1	0.7
Ge	1.0	0.7
Li	1.8	0.8
F	0.7	0.8

TABLE I: Distances (in Angstroms) used to discard AIRSS generated starting configurations for diamond, silicon, germanium and lithium fluoride. r^{min}_{A-Mu} refers to the minimum distance between a muonium and an atom of the host crystal, while r^{min}_{Mu-Mu} refers to the minimum distance between two muonium starting positions.

in computer graphics is known as a "Poisson sphere" distribution[17]. This was done with a in house Python script.

The radii used to apply both these filtering processes can be found in table I. Particular care was taken to take into account the effects of periodicity, so that every time a distance between a pair of atoms was calculated, this was reduced to the distance between the two closest periodic copies of those atoms, avoiding artefacts due to the particular choice of unit cell representation.

B. Ab Initio Computational Details

We performed Density functional theory (DFT) calculations with the CASTEP[18] code within the generalized gradient approximation (GGA)[19] and using ultrasoft pseudopotentials[20]. These calculations were performed in Si, Ge, Diamond and LiF $2\times2\times2$ supercells based on the materials' conventional cubic unit cells and which contained one muonium each that had been placed in a random position. These supercells were big enough to prevent interactions between the muons placed in the periodic images of the supercell created by the DFT calculations. All calculations were spin polarised, with the initial spin placed on the muonium atom, and all were performed with a wavefunction cutoff of 450 eV for Si, 550 eV for Ge, 600 eV for Diamond and 700 eV for LiF. We used Monkhorst-Pack grids[21] of 2x2x2, for sampling the reciprocal space of Si, Ge and Diamond, and of 3x3x3 for sampling the reciprocal space of LiF. All the ions were allowed to relax until the total energy change and forces in all ions had fallen below convergence thresholds of 1×10^{-8} eV and 1×10^{-4} $\frac{eV}{\text{Å}}$ respectively. The relaxations were performed constraining the lattice parameters of the cells to their experimental values of Si=5.430Å, Ge=5.652Å, Diamond=3.567Å and LiF=4.02Å.

C. Cluster Analysis

A common way of analysing AIRSS results is to simply classify them based on their energy. However, in this work we pioneer a more advanced method based on machine learning techniques, by which we try to classify the output structures combining energy and geometric parameters, with the aim of extracting more information that otherwise might go unnoticed. This is motivated by our interest in all potential stopping sites, not just the lowest energy one, which requires us to be able to recognise different configurations in the higher energy range. Intuitively, we expect that if suitable parameters are chosen to describe their key properties, all final configurations that represent random fluctuations around a specific stopping site will look far more similar to each other than to those around a different site. This intuition, that would make it easy for a human eye to recognize the sites by looking at a 3D representation of the various structures, is what we try to automate in a way that might allow us to sift through dozens or hundreds of candidates far more efficiently and quickly than any human would.

The technique used here is implemented in the Python library Soprano. Soprano has been developed with funding from the CCP for NMR Crystallography, is licensed under the GNU LGPL and can be downloaded for free[22]. The main purpose of Soprano is to provide the users with a set of tools to create, manipulate and analyse large amounts of chemical structures, building upon the well known Atomic Simulation Environment (ASE)[23] library. In this specific case we make use of the 'phylogenetic' analysis tool, which characterises each structure with an array of user-defined properties and then clusters them by similarity, using the algorithms implemented in the `cluster` package of the Scipy library[24]. These algorithms constitute the simplest type of unsupervised learning for pattern classification. Their purpose is to split an ensemble of points in an N-dimensional space (here the dimensionality is controlled by the number of parameters we choose to use for classification) based on their distance; different metrics can be used sometimes, but in this work we stick to the traditional Euclidean one, $\mathbf{r}^2 = \sum_i x_i^2$. There are two principal clustering algorithms implemented in Scipy:

- the *hierarchical* clustering method forms clusters by iteratively clumping together points. It will look for the closest point-point, cluster-cluster, or cluster-point pair, join them together, then repeat the procedure until the shortest distance exceeds a user-defined 'tolerance' parameter. This tolerance is usually referred to as t and is normalised to the maximum distance in the system, so that for $t = 1$ only one cluster will exist, and for $t = 0$ all points will constitute their own clusters. For point-cluster distances, the cluster is represented by its closest point to the other point (in the case of two clusters, all possible combinations are considered and the shortest one is picked). When using this method it is possible to build a dendrogram showing how the various points join up in clusters as t increases; the lowest the value of t at which a cluster is formed, the more similar its member elements will be;

- the *k-means* clustering method instead takes as an initial parameter a guess for the expected number of clusters k. It then proceeds to create k centres for the clusters, and to attribute the points each

to a cluster based on which centre is the closest. After that, the centres are moved to minimise the overall root sum square of distances, and the assignment is repeated. The procedure continues self-consistently until the clusters' composition stops changing and the RSS is minimised.

Our method has been to use both these algorithms in sequence. First, the hierarchical method is employed, and a dendrogram is plotted to have a bird's eye view of the structure of the set. If strongly distinct clusters are present, as it should be the case if the AIRSS run did indeed produce examples of multiple stopping sites, then this should be easily visible. We can then use the dendrogram to estimate how many clusters we could expect, and input that parameter in the k-means algorithm. If the classification is meaningful, there should be reasonable agreement between the size and composition of the clusters found with either method.

Information on how the arrays identifying each system are built in Soprano is provided in Appendix A. Here we focus instead on the parameters chosen as physically significant for this specific case study. Our choice fell on two of the 'genes' implemented by default in Soprano. The first was simply the energy. The second was a collection of Steinhardt bond order parameters,[25]. These are known rotationally invariant functions which describe the local environment of a given atom with the use of spherical harmonics; they can be considered a sort of power spectrum for angular frequencies instead of spatial ones. Traditionally, these are used for probing local order in disordered systems such as liquids or glasses, and that makes them especially sensitive to the shape of the site in which a muon sits, even accounting for the small possible disorder due to the randomised starting conditions. Specifically, the parameters of interest to us are the ones defined in equation (1.3) of the cited paper, namely:

$$Q_l = \left(\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |\overline{Q}_{lm}|^2 \right)^{1/2} \tag{1}$$

for any integer angular momentum channel l. However, at a difference with the original definition, here we slightly alter the way $\overline{Q}_{lm}$ is defined in order to include a smooth cutoff that only evaluates the local environment:

$$\overline{Q}_{lm} = \langle Q_{lm}(\vec{\mathbf{r}}) \rangle = \frac{\langle S(\vec{\mathbf{r}}) Y_{lm}(\theta(\vec{\mathbf{r}}), \phi(\vec{\mathbf{r}})) \rangle}{\langle S(\vec{\mathbf{r}}) \rangle} \tag{2}$$

with a sigmoid weighing function

$$S(\vec{\mathbf{r}}) = \frac{1}{2} \left[\frac{r_0 - |\vec{\mathbf{r}}|}{\delta} \left[\left(\frac{r_0 - |\vec{\mathbf{r}}|}{\delta} \right)^2 + 1 \right]^{-1/2} + 1 \right] \tag{3}$$

where we adopted $r_0 = 2$Å and $\delta = 0.05$Å. In this case, equation 1 was computed averaging over all bonds between the muon and the rest of the atoms in the unit cell (reduced to their closest periodic copy)

for angular momentum channels ranging from $l = 1$ to $l = 6$, included. Therefore, the resulting gene had a length of six, which gave us a 7-dimensional array overall for each structure once the energy was included: $[Q_1, ..., Q_6, E]$. Higher values of l would increase sensitivity to small changes in the shape of the surrounding environment at the expense of a higher computational cost, which was deemed not necessary in this case. Therefore, this choice of genes is meant to highlight which output structures are similar to each other in both energy and local environment experienced by the muon.

III. RESULTS

A. Hierarchical Clustering

Silicon, Diamond and Germanium

Figure (1) shows the dendrogram results of hierarchical clustering for the silicon, diamond and germanium supercells. The labels on the x axis correspond to all the structures that resulted from the filtering process, which have been sorted in accordance with their relative energies. On the y axis, it can be seen that the clustering starts at very small values of t and, at $t \geq 0.1$, the systems can be clearly divided into 3 clusters for silicon and 2 clusters for diamond and germanium. To analize the clustering behaviour at values of $t \leq 0.1$, we have zoomed in the regions in the x axis highlighted in grey, red and green rectangles: the branches of the dendrogram included in these rectangles are shown as insets in Figures (1a), (1b) and (1c).

The low threshold for t for cluster definition indicates that each of the clusters is composed of structures that look similar to each other in parameter space. The fact that the optimizations have converged, from wildly different initial structures, to large sets with very small internal variability suggests that these likely correspond to muon stopping sites. We will revisit this in the results of k-means clustering.

In the insets of figure (1a) show the detailed structures of the clusters in silicon. In the grey rectangle on the left hand side we can see a very well-defined cluster formed by three structures that is defined to a level of tolerance of $t \approx 0.002$. This cluster consists of the structures with indices 0, 1 and 2, which are also the three most stable structures and thus are likely to be defining a candidate stopping site. The next set of 27 structures highlighted in the red rectangle have higher energies. However, their distribution in the parameter space is such that they are also defining a cluster to a level of tolerance of $t \approx 0.002$. Finally, the 50 structures highlighted in green define a cluster to a level of tolerance of $t \approx 0.02$, but most of the structures in this cluster are already clustered at $t \approx 0.003$. This distribution of structures indicates that some of the structures in the green cluster -the ones near the sides of the green rectangle- are slightly more 'dispersed' in the parameter space than the structures that compose the other two clusters.

Figures (1b) and (1c) show the dendrogram for the diamond and germanium supercells. It shows 71 and 83 structures that at $t \geq 0.1$, can be clearly divided into 2 clusters, whose detailed branch structure for values of $t \leq 0.1$, are also shown in the insets of these figures. The clusters highlighted in green are defined for $t \approx 0.01$ and are composed of 59 structures in diamond and 20 structures in germanium, while the clusters highlighted in red are defined for $t \approx 0.1$ and are composed of 12 structures for diamond and 63 structures for germanium. Here the threshold values of t for clustering differ in an order of magnitude, indicating a clear distinction between the structures of these clusters in the parameter space. This distinction is clearly indicated in the results of k-means clustering.

Lithium Fluoride

Figure (2) shows the dendrogram for the lithium fluoride supercell. It shows 61 structures that at $t \geq 0.2$ can be separated into 2 clusters. The insets in the figure show the detailed branch structure of these clusters. The cluster highlighted in green is defined for $t \approx 0.1$ and is composed of 19 structures, while the cluster highlighted in red is defined for $t \approx 0.2$ and is composed of 42 structures.

B. K-means Clustering

Silicon, Diamond and Germanium

Figure (3a) shows the k-means clustering in Si with a guess of n = 3. The three clusters are represented by circles whose diameter is proportional to the number of structures contained in each cluster. On the x axis we indicate the average total energy of the structures belonging to each cluster relative to the lowest energy found in the system. On the y axis we indicate the standard deviation of the energy in each of the clusters. Relatively small values for the standard deviations indicate consistent clusters, which are more likely to represent physical local energy minima.

Clusters Si_2 and Si_3 in figure (3a) contain structures where the muonium is within the tetrahedral/isotropic site (Mu_T) of the cubic cell structure type of Si, but with an important distinction. On one hand, cluster Si_2 is composed by the 27 structures showed in the red rectangle in figure (1a). In these structures, the muon is bonded to one of the four Si atoms that define the tetrahedral site and is, therefore, away from the centre of the tetrahedron. On the other hand, cluster Si_3, composed of the 50 structures highlighted in green in figure (1a), has its structures slightly displaced from the centre of the tetrahedron, but not bonded to any of the Si at the edges of the tetrahedral site. According to these results, the absolute minimum for the muon stopping site in the tetragonal site of Si is not in the tetrahedral centre, but in a

spherical shell surrounding a very low potential hill that is located at the tetrahedron centre. This hill is low enough that the quantum effects associated with the muon would likely lead to a delocalisation of the muon that would stabilise the site and avoid symmetry breaking. In Appendix B there is a more detailed discussion of these two alternative tetragonal sites for the muonium in Si. Regarding cluster Si_1 in figure (3a), it contains the 3 structures highlighted in grey in figure (1a). These structures are clustered around the axially symmetric bond-centred site (Mu_{BC}). The regions of a generic cubic cell structure corresponding to these muonium stopping sites, Mu_T and Mu_{BC}, for isotropic and axially symmetric paramagnetic muoniums in Si, Diamond and Ge, are schematically shown in Figure (4).

Figures (3b) and (3c) show the k-means clusters obtained for diamond and germanium with a guess of n = 2. Clusters Diam_1 and Ge_2 contain 12 and 63 structures clustered around the axially symmetric bond-centred site (Mu_{BC}), and clusters Diam_2 and Ge_1 contain 59 and 20 structures which have the muonium located near the centre of the tetrahedron and not bonded to any of the carbon or germanium atoms forming the tetrahedral site. For diamond there is a difference of ≈1.4 eV between the relative average energies of the two clusters and of ≈0.05 eV between their corresponding standard deviations, with cluster Diam_1 being the most disperse cluster in the parameter space. Regarding germanium, cluster Ge_2 has an average energy ≈0.41 eV larger than that of Ge_1 cluster. The standard deviation of both Ge_1 and Ge_2 is below 0.02 eV, indicating a low dispersion for both clusters in the parameter space.

Lithium Fluoride

Figure (5) show the k-means clusters obtained for lithium fluoride with a guess of n = 2. Clusters LiF_1 and LiF_2 contain 42 and 19 structures respectively. Cluster LiF_1 has an average energy of ≈ 0.38 eV and a standard deviation of ≈0.23 eV, and its structures are clustered around the octahedral site. On the other hand, cluster LiF_2 has an average energy of ≈8.0 eV and an standard deviation of ≈0.43 eV: all the structures in this cluster are much more dispersed in the space and have a much larger energy. The image shown in figure (5) indicates the structure with the minimum energy in the cluster, which has lower symmetry than the other. In general, due to the high energies involved, it seems reasonable to consider this cluster as non physical, or at least not experimentally relevant.

IV. ANALYSIS

Using only ab initio simulations and data analysis techniques we were able to identify a small number of candidate stopping sites for muonium in crystalline silicon, germanium, diamond and lithium fluoride. Each site is represented by a cluster of size ranging from 3 to 60 optimised structures, identified by en-

Cluster	Rel. Aver. Energy (eV)	Structures in Cluster	Standard Dev. (eV)
Si_1	0	3	0.0002
Si_2	0.23	27	0.0012
Si_3	0.26	50	0.0015
Diam_1	0.02	12	0.051
Diam_2	1.4	59	0.01
Ge_1	0.42	20	0.005
Ge_2	0.01	60	0.015
LiF_1	0.38	42	0.23
LiF_2	0.79	19	0.43

TABLE II: Main properties of all of the identified clusters in Si, Diamond, Ge and LiF.

ergetic and geometric similarities. The full list of their properties is listed in Table II. The low values of the standard deviation of energy among these clusters reinforce the case that they represent indeed small fluctuations around a single energy minimum, and are not just flukes. If, for example, two non-equivalent minima were grouped under the same cluster due to accidental geometric similarities, we would expect a much higher dispersion of the energy values.

The predicted sites, that can be seen in figures (3a), (3b), (3c) and (5), closely match what is known from the literature about muonium defects in diamond, silicon, germanium and lithium fluoride crystals from both experiments and theoretical calculations[2,6,26], which have identifyed a bond-centred (Mu_{BC}) and a tetragonal(Mu_T) stopping site diamond, silicon and germanium, and an octahedral site in lithium fluoride.

Regarding the sites in silicon, diamond and germanium, we know from the literature[2,27] that both the (Mu_{BC}) and (Mu_T) sites were experimentally observed. The data from TableII however suggests that the tetrahedral sites (represented by clusters Si_2 and Si_3, Diam_2, and Ge_2) all have higher formation energies than the bond centred ones. The lowest energy difference between tetrahedral and bond centred sites, ΔE, is 0.23 eV. If we assume that the system is in thermal equilibrium, this value of ΔE implies that a temperature of more than 2000 K is needed to transition from one stopping site to the other. Hence, the order of magnitude alone seems to completely refute the possibility that the tetrahedral site could be observed in thermal equilibrium at low temperatures. The prediction that the Mu_{BC} site is lower in energy than the Mu_T one is also in qualitative agreement with previous theoretical results[26]. This reinforces the commonly held view that during an experiment muons do not have the time to effectively relax to their equilibrium state and can therefore occupy metastable sites[27].

Furthermore, our calculations support the prediction of delocalisation of the muon in the Mu_T site of Si that was advanced in previous studies[8,28]. In order to clarify this point, we carried out further calculations on many configurations surrounding the site to plot the local energy landscape, reported in Appendix B. We observed for the Mu_T site in Si a flat potential with a slight maximum in the centre and a minimum approximately distributed in a radial shell surrounding it; this is the most likely reason for the observation of two clusters corresponding to that site, Si_2 and Si_3, with different positions for the muonium itself with respect to the centre of the tetrahedra. Experimentally, the delocalisation of the muon in the tetragonal site was first proposed by Holzschuh *et. al.*[29]. In this model, the muon hops between different sites in the tetragonal region, which helped to explain the anomalous temperature dependence of the isotropic component of the hyperfine coupling constant in Si.

Regarding the stopping sites in lithium fluoride, our methodology predicted the octahedral stopping site for muonium, which is the site that has been both experimentally and theoretically predicted[6,30]. The other local minimum we found has an average energy too large to be a physically meaningful muonium stopping site.

V. CONCLUSIONS

We have proposed a purely theoretical method to predict muon stopping sites in crystalline materials. The method is based on a combination of ab initio random structure searching and machine learning, and it has successfully predicted the Mu_T and Mu_{BC} stopping sites of muonium in Si, Diamond and Ge, and the octahedral stopping site in lithium fluoride, purely from first principles. The process is easily reproducible and requires little human input to analyse dozens or even hundreds of structures. Soprano, a Python library containing the tools used for this analysis as well as many others designed for different systems, has been released publicly[22] and will be fully documented in a future work.

VI. ACKNOWLEDGEMENTS

The authors would like to thank Barbara Montanari, from the Scientific Computing Department at RAL, and Francis Pratt and Stephen Cottrell, from the Muons Group at ISIS, for the useful discussions. The authors are also grateful for the computational support provided by: (a) STFC Scientific Computing Department's SCARF cluster; (b) the UK national high performance computing service, ARCHER, for which access was obtained via the UKCP consortium and funded by EPSRC grant ref EP/K013564/1; and (c) the UK Materials and Molecular Modelling Hub for computational resources, which is partially funded by EPSRC (EP/P020194/1). Funding for this work was provided by STFC-ISIS muon source and by the CCP for NMR Crystallography, funded by EPSRC grants EP/J010510/1 and EP/M022501/1.

Appendix A: The `phylogen` module in Soprano

Soprano's `phylogen` module contains classes and functions for clustering similarity analysis of large populations of structures. It owes its name to the analogy between this approach and the construction of phylogenetic maps or trees connecting living species based to the similarity of their DNA. Similarly, in this module, the user can pick a number of 'genes' by which the structures should be characterised, and these genes will be chained together into a single array that uniquely defines each structure. These arrays will then constitute the points whose similarity relationships are evaluated through clustering.

In Soprano, a gene can be any scalar or vectorial property of a structure. For example, the energy of a structure can be a gene of only one element; its lattice parameters a, b, c a gene of three; and so on. When performing a clustering operation the user can also pick a weight for each gene, thus determining the relative importance between them, and a normalising interval. By default, all genes are weighted equally, and they're normalised to $[0, 1]$ over the entire collection being analysed. So, for example, if we indicate the raw value for element j of gene i as x_i^j in an analysis using g genes each with n_i elements we'll have that the 'DNA' array for a given structure is defined as:

$$[X_1^1, X_1^2, ..., X_1^{n_1}, X_2^1, ..., X_2^{n_2}, ..., X_g^{n_g}] \tag{A1}$$

The normalised and weighted values X_i^j are defined as:

$$X_i^j = \frac{w_i}{\sqrt{n_i}} \frac{[x_i^j - \min(x_i^j)]}{[\max(x_i^j) - \min(x_i^j)]} \tag{A2}$$

where w_i is the user-defined weight, minima and maxima refer to the whole collection, and the square root term has the purpose of normalising for size, to prevent longer genes from dominating the classification.

Appendix B: Tetragonal/isotropic Mu_T site in Si

We investigated the nature of the tetrahedral site Mu_T by performing CASTEP phonon calculations at the Γ point on a Si supercell with the muon perfectly located in the tetragonal site. This muonated Si supercell has 195 vibrational modes, and we assumed that the modes ω_i associated with the muon are mutually perpendicular and are decoupled from all the other modes as a consequence of its much smaller mass. However, we found that all three the main vibrational frequencies for Mu_T are negative, which indicates that the associated mode is imaginary, i.e.: the muon in the tetragonal position is in a maximum of its corresponding Born-Oppenheimer (BO) potential. To estimate the size of this maximum, we took the calculated eigenvectors corresponding to those negative frequencies and displaced the muon along them, in positive and negative directions, and then calculated at equally spaced points the total DFT energy. Figure (6) shows the results of these calculations. The shape depends only minimally from the specific eigenvector

chosen, which suggests that the potential is close to a radially symmetric quartic with a minimum shell around $r = 1\,\text{Å}$.

This maximum for the BO potential in the Tetragonal site of muonium has been observed previously using DFT calculations[26,28], and there is experimental and theoretical evidence that the muon in the T site of Silicon is delocalised[28,29]. To the best of the author's knowledge, there is only one theoretical paper that reports a minimum of the BO potential in the tetragonal site[31], but this result could not be confirmed by other DFT calculations. Furthermore, we conducted computational tests to rule out that the maximum in the BO potential is an artificial result arising from our calculations. We performed full geometrical relaxations and phonon calculations at the Γ point, for Si with Mu in the tetragonal site using the LDA and GGA functionals, and obtained the results indicated in Table III.

Si	Exp.	DFT-LDA	DFT-GGA	DFT-GGA (a from LDA)
a (Å)	5.43	5.40	5.47	5.40
f_1 (cm^{-1})	N/A	-182.13	-410.0	-409.9
f_2 (cm^{-1})	N/A	-182.1	-409.9	-409.9
f_3 (cm^{-1})	N/A	-182.1	-409.9	-409.8

TABLE III: Lattice parameters and negative (imaginary) phonon frequencies for muonium in the tetragonal site of Si.

REFERENCES

[1]R. De Renzi, P. Guidi, P. Podini, R. Tedeschi, C. Bucci, and S. F. J. Cox. Magnetic properties of MnF_2 and CoF_2 determined by implanted positive muons. I. Localization studies. *Physical Review B*, 30(1):186, 1984.

[2]Bruce D. Patterson". Muonium states in semiconductors. *Reviews of Modern Physics*, 60(1):69, 1988.

[3]M. Bimbi, G. Allodi, R. De Renzi, C. Mazzoli, and H. Berger. Muon spin spectroscopy evidence of a charge density wave in magnetite below the verwey transition. *Phys. Rev. B*, 77:045115, Jan 2008.

[4]H. Maeter, H. Luetkens, Yu. G. Pashkevich, A. Kwadrin, R. Khasanov, A. Amato, A. A. Gusev, K. V. Lamonova, D. A. Chervinskii, R. Klingeler, C. Hess, G. Behr, B. Büchner, and H.-H. Klauss. Interplay of rare earth and iron magnetism in rFeAsO (r = La, ce, pr, and sm): Muon-spin relaxation study and symmetry analysis. *Phys. Rev. B*, 80:094524, Sep 2009.

[5]G. Prando, P. Bonfà, G. Profeta, R. Khasanov, F. Bernardini, M. Mazzani, E. M. Brüning, A. Pal, V. P. S. Awana, H.-J. Grafe, B. Büchner, R. De Renzi, P. Carretta, and S. Sanna. Common effect of chemical and external pressures on the magnetic properties of $r$copo (r = la, pr). *Phys. Rev. B*, 87:064401, Feb 2013.

[6]JS Möller, D Ceresoli, T Lancaster, N Marzari, and SJ Blundell. Quantum states of muons in fluorides. *Physical Review B*, 87(12):121108, 2013.

[7]F. Bernardini, P. Bonfà, S. Massidda, and R. De Renzi. Ab initio. *Phys. Rev. B*, 87:115148, Mar 2013.

[8]A. R. Porter, M. D. Towler, and R. J. Needs. Muonium as a hydrogen analogue in silicon and germanium: Quantum effects and hyperfine parameters. *Phys. Rev. B*, 60:13534, Nov 1999.

[9]Rolf H. Luchsinger, Yu Zhou, and Peter F. Meier. Gradient corrections in first-principles calculations of hyperfine parameters in semiconductors. *Phys. Rev. B*, 55:6927–6937, Mar 1997.

[10]Andrew J. Morris, Chris J. Pickard, and R. J. Needs. Hydrogen/silicon complexes in silicon from computational searches. *Phys. Rev. B*, 78:184102, Nov 2008.

[11]Andrew J. Morris, Chris J. Pickard, and R. J. Needs. Hydrogen/nitrogen/oxygen defect complexes in silicon from computational searches. *Phys. Rev. B*, 80:144112, Oct 2009.

[12]Chris J Pickard and RJ Needs. Ab initio random structure searching. *Journal of Physics: Condensed Matter*, 23(5):053201, 2011.

[13]Chris J Pickard and R J Needs. Ab initio random structure searching. *Journal of Physics: Condensed Matter*, 23(5):053201, 2011.

[14]Jonathan PK Doye, David J Wales, and Mark A Miller. Thermodynamics and the global optimization of lennard-jones clusters. *The Journal of Chemical Physics*, 109(19):8143–8153, 1998.

[15]Jonathan PK Doye and Claire P Massen. Characterizing the network topology of the energy landscapes of atomic clusters. *The Journal of chemical physics*, 122(8):084105, 2005.

[16]Claire P Massen and Jonathan PK Doye. Power-law distributions for the areas of the basins of attraction on a potential energy landscape. *Physical Review E*, 75(3):037101, 2007.

[17]Ares Lagae and Philip Dutré. Poisson sphere distributions. In *Vision, Modeling, and Visualization*, pages 373–379, 2006.

[18]Stewart J Clark, Matthew D Segall, Chris J Pickard, Phil J Hasnip, Matt I J Probert, Keith Refson, and Mike C Payne. First principles methods using CASTEP. *Zeitschrift für Kristallographie - Crystalline Materials*, 220(5/6), 2005.

[19]John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple . *Physical Review Letters*, 78(7):1396–1396, February 1997.

[20]J S Lin, A Qteish, M C Payne, and V Heine. Optimized and Transferable Nonlocal Separable ab initio Pseudopotentials. *Physical Review B*, 47(8):4174–4180, 1993.

[21]Hendrik J. Monkhorst and James D. Pack. Special points for brillouin-zone integrations. *Phys. Rev. B*, 13:5188–5192, 1976.

[22]Simone Sturniolo. Soprano - a library to crack crystals. `https://ccpforge.cse.rl.ac.uk/gf/project/soprano/`.

[23]Ask Hjorth Larsen, Jens Jrgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Du?ak, Jesper Friis, Michael N Groves, Bjrk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jn Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka,

Andrew Peterson, Carsten Rostgaard, Jakob Schitz, Ole Schtt, Mikkel Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W Jacobsen. The atomic simulation environment?a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017.

[24]Travis E Oliphant. Python for scientific computing. *Computing in Science & Engineering*, 9(3), 2007.

[25]Paul J. Steinhardt, David R. Nelson, and Marco Ronchetti. Bond-orientational order in liquids and glasses. *Phys. Rev. B*, 28:784–805, Jul 1983.

[26]Rolf H Luchsinger, Yu Zhou, and Peter F Meier. Gradient corrections in first-principles calculations of hyperfine parameters in semiconductors. *Physical Review B*, 55(11):6927, 1997.

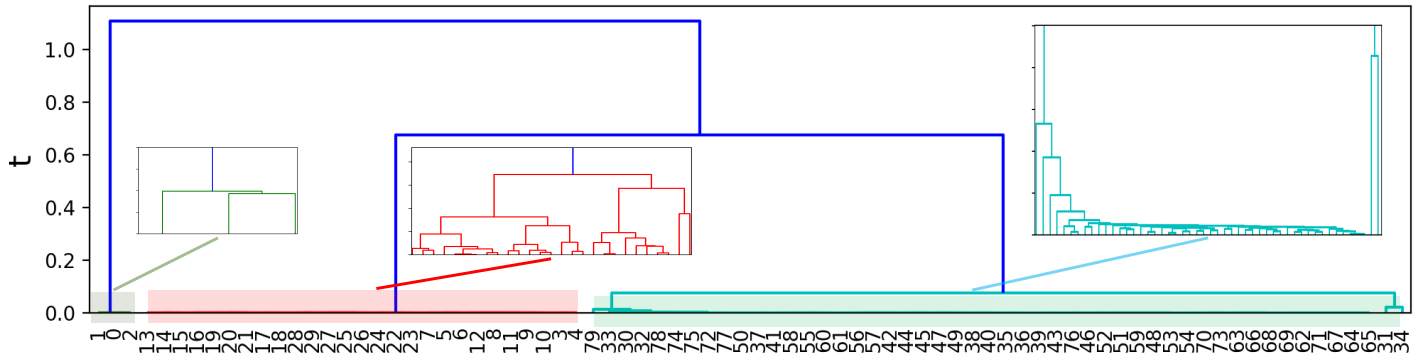
[27]J S Lord, S F J Cox, M Charlton, D P Van der Werf, R L Lichti, and A Amato. The muon spin response to intermittent hyperfine interaction: modelling the high-temperature electrical activity of hydrogen in silicon. *Journal of Physics: Condensed Matter*, 16(40):S4739, 2004.

[28]Takashi Miyake, Tadashi Ogitsu, and Shinji Tsuneyuki. Quantum distributions of muonium and hydrogen in crystalline silicon. *Physical Review Letters*, 81(9):1873–1876, 1998.

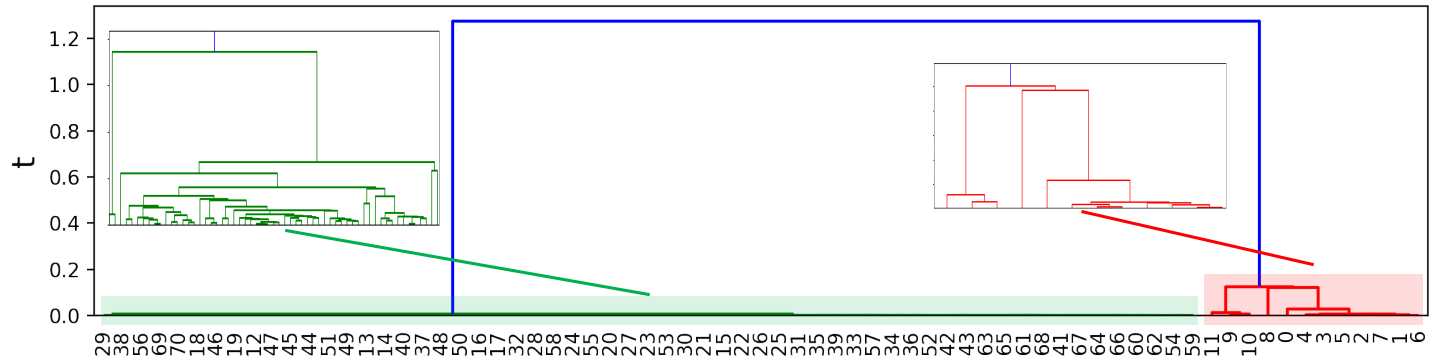
[29]Eugen Holzschuh. Direct measurement of muonium hyperfine frequencies in si and ge. *Physical Review B*, 27(1):102, 1983.

[30]H. Baumeler, R.F. Kiefl, and H. et al. Keller. Muonium centers in the alkali halides. *Hyperfine Interactions*, 32:659, 1986.

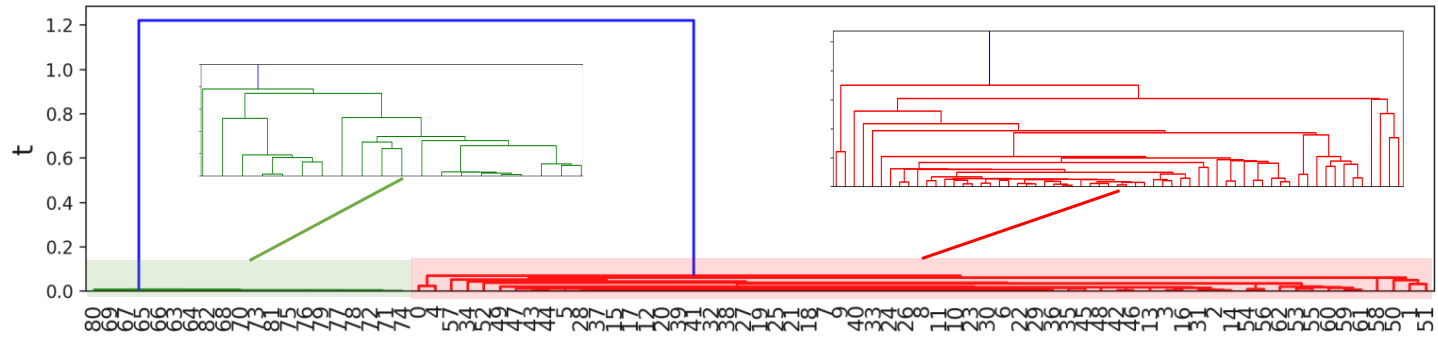
[31]AR Porter, MD Towler, and RJ Needs. Muonium as a hydrogen analogue in silicon and germanium: Quantum effects and hyperfine parameters. *Physical Review B*, 60(19):13534, 1999.



(a) Silicon



(b) Diamond



(c) Germanium

FIG. 1: (1) shows the hierarchical clustering trees for silicon, diamond and germanium, with colours applied for truncation at $t = 0.2$ in the y axis. The blue lines above $t = 0.2$ indicates the clusters into which the structures can be classified. The structures that resulted from the filtering process are labeled and placed along the x axis in accordance with their relative energies. The indexing starts with 0 for the lowest energy structure, but the values of these indexes are not correlated with their positions in the x axis: the structures with the lowest energies are closer to x=0.

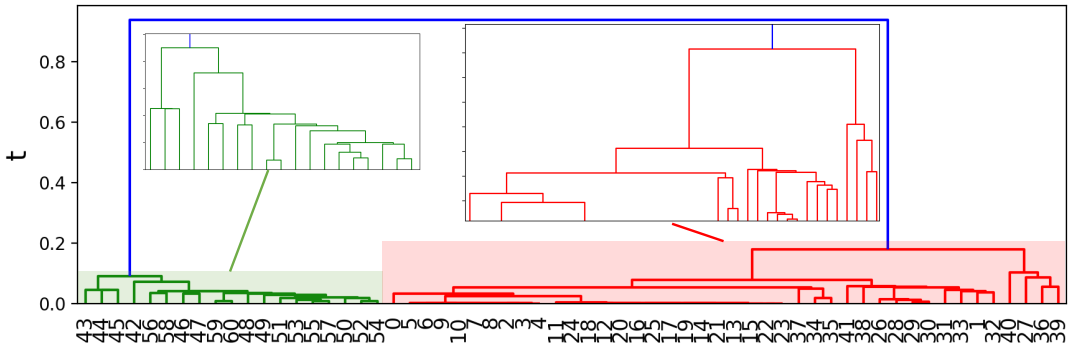
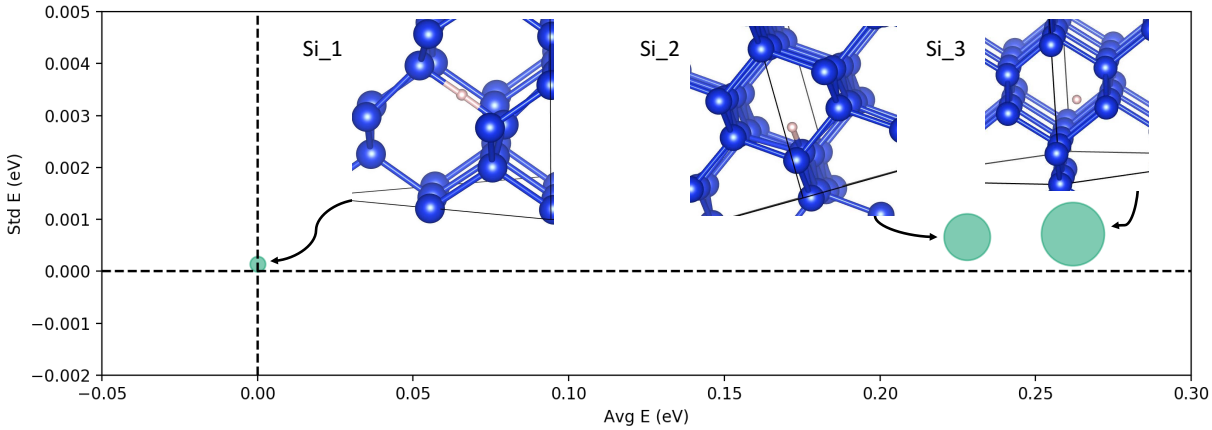
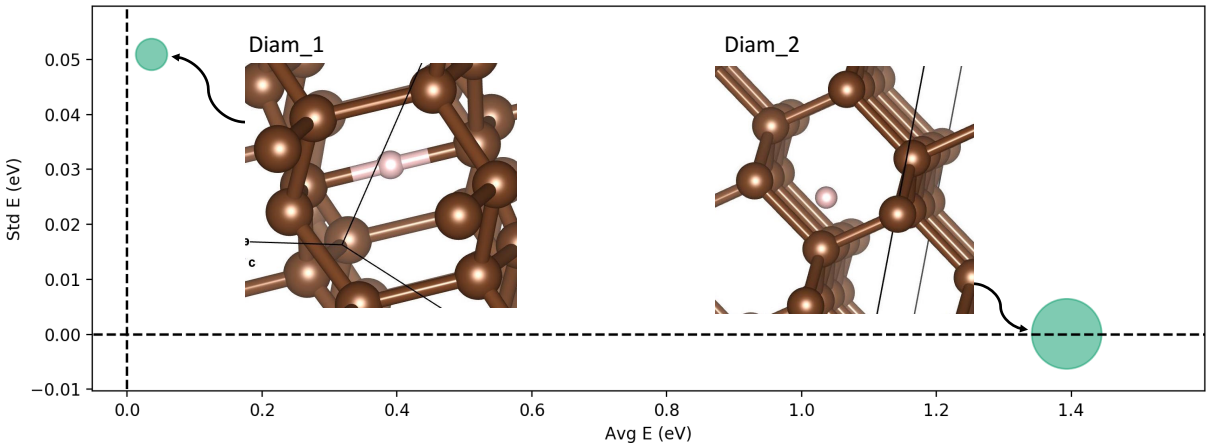


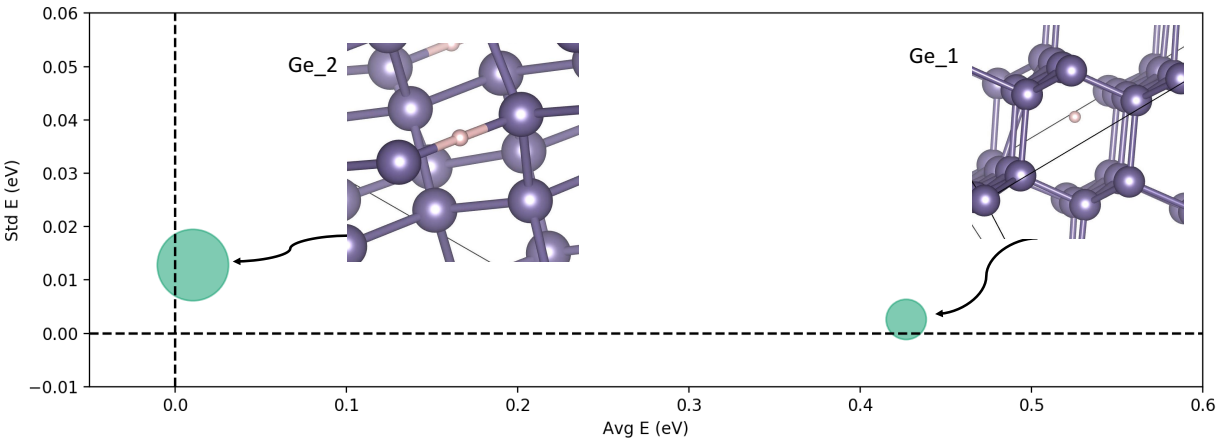
FIG. 2: (2) shows the hierarchical clustering tree for lithium fluoride, with colours applied for truncation at $t = 0.2$ in the y axis. The blue lines above $t = 0.2$ indicates the clusters into which the structures can be classified. The structures that resulted from the filtering process are labeled and placed along the x axis in accordance with their relative energies. The indexing starts with 0 for the lowest energy structure.



(a) Silicon



(b) Diamond



(c) Germanium

FIG. 3: Figure (3) shows circles representing the clusters obtained via the k-means clustering method. The crystalline structures correspond to the most stable structure in each of the clusters. The diameter of each circle represents the number of structures contained in each cluster. The x coordinate of the centre of each one of the circles indicates the average energy of the corresponding cluster -relative to the lowest energy structure in the cluster-, while the y coordinate of the centre indicates the standard deviation of the average energy of that cluster.

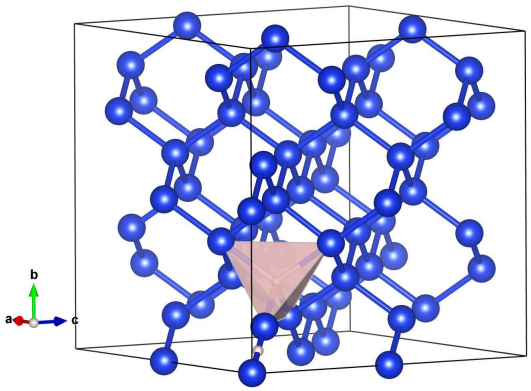


FIG. 4: Tetragonal Mu_T and bond-centred Mu_{BC} sites in a generic conventional unit cell of a material with a cubic structure like Si, Diamond or Ge. The space group is Fd3m and the only difference between these structures is that they have different lattice parameters.

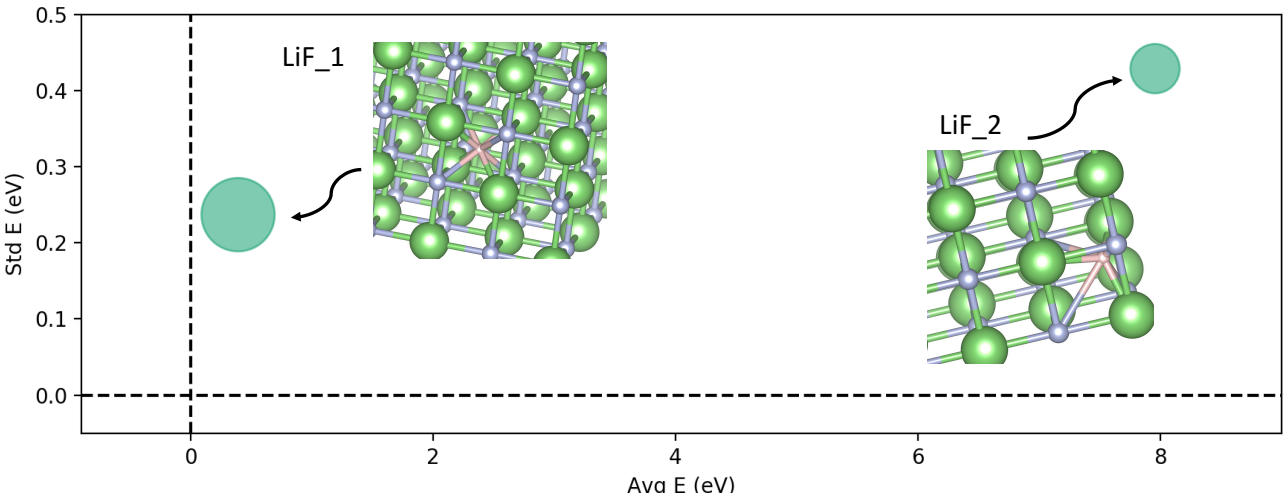


FIG. 5: Figure (5) show the two clusters obtained via the k-means clustering method in LiF. The x coordinate of the centre of each one of the circles indicates the average energy of the corresponding cluster, while the y coordinate indicates the standard deviation of the average energy of that cluster. Clearly, cluster LiF_2 is too high in energy to be representing a physically meaningful stopping site.

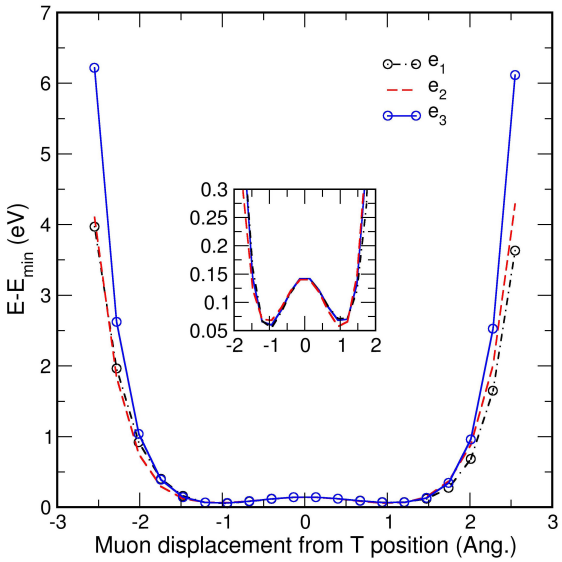


FIG. 6: DFT total energies calculated at equally spaced points along the three eigenvectors corresponding to the calculated negative frequency for Mu_T