JPhys Materials

PAPER

Managing uncertainty in data-derived densities to accelerate density functional theory

Andrew T Fowler[1] , Chris J Pickard[1,2] and James A Elliott[1,3]

1   Department of Materials Science and Metallurgy, University of Cambridge, 27 Charles Babbage Road, Cambridge, CB3 0FS, United Kingdom
2   Advanced Institute for Materials Research, Tohuku University, 2-1 1 Katahira, Aoba, Sendai, 980-8577, Japan
3   Author to whom any correspondence should be addressed.

E-mail: jae1001@cam.ac.uk

Abstract

Faithful representations of atomic environments and general models for regression can be harnessed to learn electron densities that are close to the ground state. One of the applications of data-derived electron densities is orbital-free density functional theory (DFT). However, extrapolations of densities learned from a training set to dissimilar structures could result in inaccurate results, which would limit the applicability of the method. Here, we show that a non-Bayesian approach can produce estimates of uncertainty which can successfully distinguish accurate from inaccurate predictions of electron density. We apply our approach to DFT where we initialise calculations with data-derived densities only when we are confident about their quality. This results in a guaranteed acceleration to self-consistency for configurations that are similar to those seen during training and could be useful for sampling-based methods, where previous ground state densities cannot be used to initialise subsequent calculations.

1. Introduction

Density functional theory (DFT) has seen widespread adoption in many areas of research spanning the natural sciences due to its high predictive capability at modest computational cost and transferability across different systems [1]. The staggering number of applications and papers that exploit DFT are a testament to its value in materials science [2].

The foundations of DFT are the Hohenberg–Kohn theorems [3]. The first of these expresses the total energy of a many-electron system as a functional, $F[n]$, of the ground state electron density, $n(\mathbf{x})$, where $\mathbf{x}$ denotes a location in real space. The second theorem tells us the ground state density is found by minimising $F[n]$ with respect to $n(\mathbf{x})$. Although an exact form for $F[n]$ has not been established, the unknown components can be separated into a kinetic energy contribution, $T[n]$ and a term called the exchange correlation functional, $E_{xc}[n]$ [4]. The magnitude of contributions from $E_{xc}[n]$ to the total energy are known to be relatively small and so the exchange correlation term can be approximated to some extent by approaches such as the local density approximation and the generalised gradient approximation [5, 6]. The kinetic energy term cannot however be so well approximated and a universally applicable functional is still unknown [7]. This forces many applications to an alternative paradigm, Kohn–Sham (KS) DFT [5]. Here, $T[n]$ is replaced by an expectation over independent electron wave functions. In many cases, this vastly improves the accuracy of the kinetic energy contribution to $F[n]$ but it introduces a significant increase in the computational expense [8].

With the recent renewed interest in machine learning, theoretical attempts to learn $T[n]$ have been supplemented with data-driven inferences [9]. These are hampered by difficulties in approximating gradients $\partial T[n]/\partial n(\mathbf{x})$, an evaluation which is necessary in finding the ground state density [10]. Recently, an approach to circumvent this issue was proposed, stimulating a new wave of interest in data-driven orbital free (OF) DFT [11–13]. The alternative route to evaluating data-derived OF functionals on the ground state density is to

empirically infer the ground state density itself, removing the variational optimisation of $F[n]$ completely. Two possible issues with this approach ultimately stem from the availability of data. While $T[n]$ and $n(\mathbf{x})$ may be very accurate for structures similar to those seen during training, when extrapolating for unfamiliar structures, either $T[n]$ or $n(\mathbf{x})$ may give predictions that are far from the true values.

The key contribution that we make in this work is to show that predictive uncertainty can be harnessed to prevent poor extrapolations of $n(\mathbf{x})$ for structures that are dissimilar to those seen during training. We illustrate how such a measure of confidence can be applied to accelerate KS DFT by initialising calculations with a data-driven contribution only when we are confident about its quality. We note that such an application is most suited to sampling methods such as nested sampling, where subsequent structures are not guaranteed to be similar [14]. To the best of our knowledge, *ab initio* nested sampling has yet to be realised due to the prohibitive computational requirements of standard KS DFT. This work may contribute, in some part, to realising such calculations. For other applications like molecular dynamics or geometry optimisation, a temporary history of ground state densities can be applied to subsequent configurations in the calculation. This results in successive calculations being initialised fairly close to their ground state, rendering any improvements made from a data-derived density much less significant.

2. Quantifying uncertainty

Evaluating an error or measure of confidence in a data-driven prediction like $n(\mathbf{x})$ is a well studied problem [15, 16]. Applications of uncertainty quantification have recently begun appearing in materials science, with some even in DFT, such as the linear model exchange correlation functional of Aldegunde *et al* [17–21]. In this work, we show that useful applications of a predictive uncertainty in $n(\mathbf{x})$ can be realised for just one of many possible approaches. By illustrating a proof-of-concept application to accelerating KS DFT, we hope to encourage a greater awareness of the advantages of quantifying uncertainty and to stimulate interest in alternative methods and applications such as in OF DFT.

2.1. Non-Bayesian regression
In the following we adopt the notation that n and $\mathbf{x}$ refer to a known ideal model contribution to electron density and a corresponding representation for the environment of that density point, respectively. Specifically, we adopt the bispectrum representation for $\mathbf{x} = (\mathbf{x}^{\,\text{local}}, \mathbf{x}^{\,\text{global}})$, which is a concatenation of local and global contributions [22, 23]. We refer the reader to appendix A for further detail and also note that explicit dependence of n upon $\mathbf{x}$ has been dropped in this section to improve clarity. In this work, we use a non-Bayesian approach to quantify uncertainty. Although a Bayesian method to parametric regression will give a more reliable measure of uncertainty, evaluating uncertainty from the predictive distribution for nonlinear models is not a simple task and often sampling is involved which can incur significant computational overhead [15].

We propose a model in which observations of the true ground state density n are prone to random error which is distributed normally about the model predictions $\mu(\mathbf{x}, \mathbf{w})$:

$$p(n|\mathbf{x}, \mathbf{w}) = \mathcal{N}(n|\mu(\mathbf{x}, \mathbf{w}), \sigma(\mathbf{x}, \mathbf{w})^2). \tag{1}$$
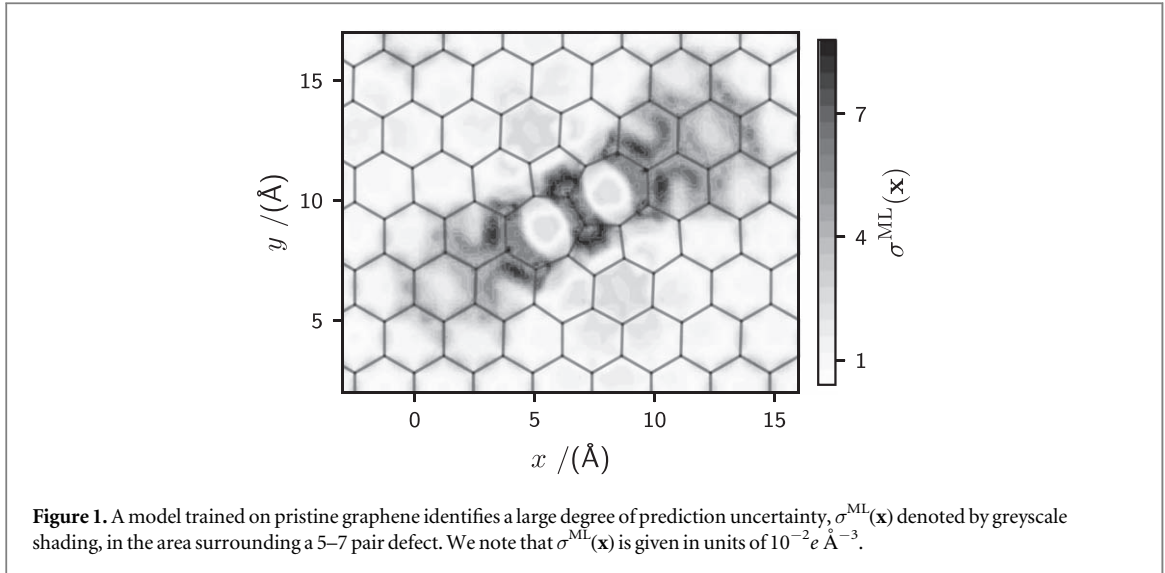
We also introduce a dependency of the variance of this random error, $\sigma(\mathbf{x}, \mathbf{w})^2$, on the environment $\mathbf{x}$, which is known as a *heteroskedastic* model for noise [24]. We use a fully connected feed-forward neural network with hidden network weights $\mathbf{w}$, to calculate $\mu(\mathbf{x}, \mathbf{w})$ and $\sigma(\mathbf{x}, \mathbf{w})^2$. Observations of n are treated as independent and identically distributed random variables and to infer $\mathbf{w}$, we calculate the maximum likelihood estimate by maximising the product $\prod_{n,\mathbf{x}} p(n|\mathbf{x}, \mathbf{w})$ over all observations in the training set.

To quantify error in the predictions of n given a new $\mathbf{x}$, we adopt an ensemble of N_{ens} neural networks, each with network weights $\mathbf{w}_i$. Adopting a uniformly weighted Gaussian mixture, the likelihood of the ensemble is then:

$$p(n|\mathbf{x}, \mathbf{W}) = \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} \mathcal{N}(n|\mu(\mathbf{x}, \mathbf{w}_i), \sigma(\mathbf{x}, \mathbf{w}_i)^2)$$
$$= \mathcal{N}(n|n^{\text{ML}}(\mathbf{x}, \mathbf{W}), \sigma^{\text{ML}}(\mathbf{x}, \mathbf{W})^2), \tag{2}$$

where $\mathbf{W} = (\mathbf{w}_1, \ldots, w_{N_{\text{ens}}})$ and $p(n|\mathbf{x}, \mathbf{W})$ is also a normal distribution [25]. Uncertainty in our prediction of n is given by the variance of $p(n|\mathbf{x}, \mathbf{W})$, $\sigma^{\text{ML}}(\mathbf{x}, \mathbf{W})^2$ and can be evaluated as:

$$\sigma^{\text{ML}}(\mathbf{x})^2 = \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} \mu(\mathbf{x}, \mathbf{w}_i)^2 - n^{\text{ML}}(\mathbf{x})^2 + \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} \sigma(\mathbf{x}, \mathbf{w}_i)^2$$

$$n^{\text{ML}}(\mathbf{x}) = \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} \mu(\mathbf{x}, \mathbf{w}_i). \tag{3}$$

Figure 1. A model trained on pristine graphene identifies a large degree of prediction uncertainty, $\sigma^{\mathrm{ML}}(\mathbf{x})$ denoted by greyscale shading, in the area surrounding a 5–7 pair defect. We note that $\sigma^{\mathrm{ML}}(\mathbf{x})$ is given in units of $10^{-2} e\,\mathring{A}^{-3}$.

2.2. Doing no harm

To apply our model for prediction uncertainty $\sigma^{\mathrm{ML}}(\mathbf{x})^2$ in (3) to accelerate KS DFT, we need to evaluate a global measure of uncertainty for an entire structure. We call this measure $H[p_{\sigma^{\mathrm{ML}}}]$, where an unknown dependency on the empirical prior distribution $p_{\sigma^{\mathrm{ML}}}$ of σ^{ML} is shown explicitly. In this work, we adopt the very simple measure that:

$$H[p_{\sigma^{\mathrm{ML}}}] = \mathbb{E}_{p_{\sigma^{\mathrm{ML}}}}[\ln(\sigma^{\mathrm{ML}})] = \frac{1}{N}\sum_{i=1}^{N}\ln(\sigma^{\mathrm{ML}}(\mathbf{x}_i)) \tag{4}$$

for N densities in a crystal. We now abbreviate $H[p_{\sigma^{\mathrm{ML}}}] = H$ and introduce a tapering function $\Gamma(H)$, which is essentially a step function with a controllable transition point and length scale. For details of the specific form of Γ used in this work, we refer the reader to appendix C. With $\Gamma(H)$, we can control the empirical contribution $n^{\mathrm{ML}}(\mathbf{x})$ to an initial density estimate:

$$n(\mathbf{x}) = n_0(\mathbf{x}) + \Gamma(H)n^{\mathrm{ML}}(\mathbf{x}). \tag{5}$$

$n_0(\mathbf{x})$ represents any standard initialisation technique for the density in DFT but typically, this is a combination of the radial components of electron density for atoms assumed to be in vacuum. The ideal model contribution n from section 2.1 is the difference of the true ground state density and the standard initial contribution, $n(\mathbf{x}) - n_0(\mathbf{x})$ from (5).
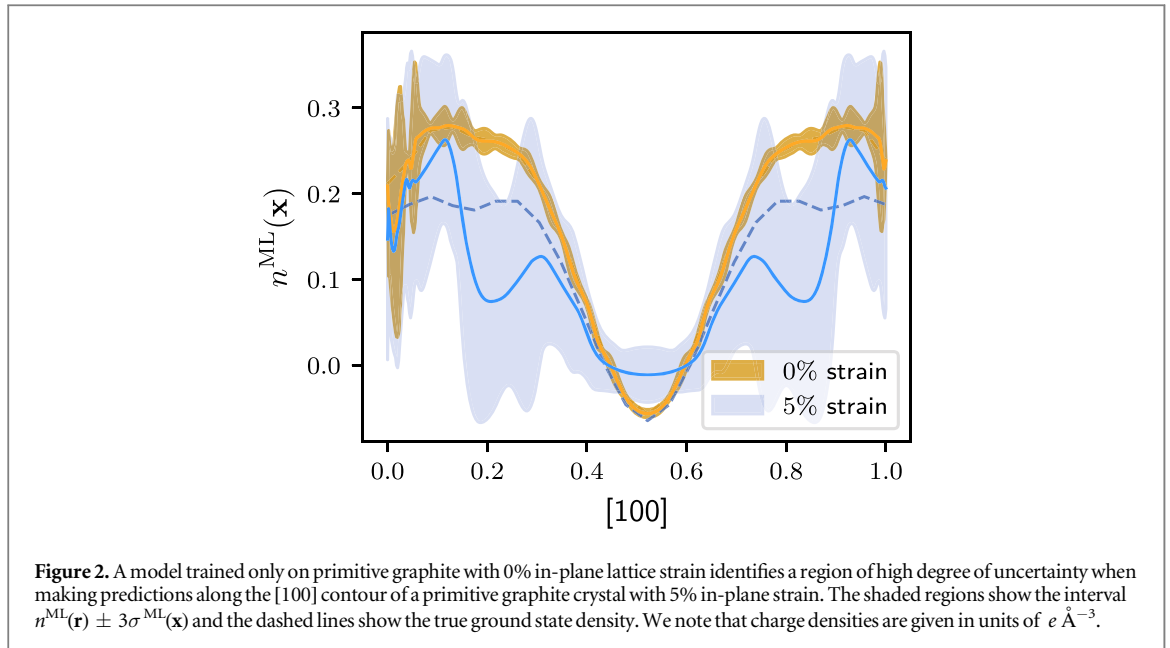
We note that an alternative strategy could be to taper empirical contributions locally at each grid point, but we choose a global approach to discourage spurious non-smoothness in $n^{\mathrm{ML}}(\mathbf{x})\Gamma(\sigma^{\mathrm{ML}}(\mathbf{x}))$ that might occur if $|\sigma^{\mathrm{ML}}(\mathbf{x} + \delta\mathbf{x}) - \sigma^{\mathrm{ML}}(\mathbf{x})| \gg 0$, for a small perturbation in environment $\delta\mathbf{x}$. We also note that the effects of any random error in $\sigma^{\mathrm{ML}}(\mathbf{x})$ are significantly reduced by considering distribution averages. While we found that the simple choice of H used in (4) worked very well at identifying uncertain predictions for the applications in this work, a more informative measure of the distribution $p_{\sigma^{\mathrm{ML}}}$ may improve this distinction further. Higher order moments of $p_{\sigma^{\mathrm{ML}}}$ such as the distribution variance for example could be utilised, in addition to knowledge about the distribution mean.

3. Results

In this section we illustrate how the non-Bayesian approach to uncertainty quantification adopted in this work can qualitatively distinguish accurate from inaccurate values of the data-derived contribution $n^{\mathrm{ML}}(\mathbf{r})$. We also show how the number of self-consistent field iterations needed to reach self-consistency in a KS DFT calculation can be reduced as the initial density tends to the exact ground state density. For environments dissimilar to those seen during training, we expect a larger predictive uncertainty.

3.1. 5–7 defect in graphene

Figure 1 shows $\sigma^{\mathrm{ML}}(\mathbf{x})$ for a single layer of graphene with a 5–7 pair (or Dienes) topological defect [26]. Only densities from a single pristine layer of graphene were used during training. In the area immediately surrounding the defect, predictive uncertainties increase (denoted by dark shading), identifying this region as an environment dissimilar to the defect-free layer.

Figure 2. A model trained only on primitive graphite with 0% in-plane lattice strain identifies a region of high degree of uncertainty when making predictions along the [100] contour of a primitive graphite crystal with 5% in-plane strain. The shaded regions show the interval $n^{ML}(\mathbf{r}) \pm 3\sigma^{ML}(\mathbf{x})$ and the dashed lines show the true ground state density. We note that charge densities are given in units of $e\ \text{Å}^{-3}$.

3.2. In-plane strain in graphite

In figure 2, we compare the prediction uncertainty of graphite with 0% and 5% in-plane strain. Specifically, we show the [100] lattice vector contour and find that predictions are significantly more certain for the 0% contour which was seen during training than the 5% contour that was not. Further details of the bispectrum and KS DFT calculations for figures 1 and 2 can be found in appendix C.

3.3. Accurate initial densities

To motivate our application of uncertainty quantification to KS DFT, we examine the convergence of single point KS DFT calculations to self-consistency as we perturb initial densities away from the exact ground state via changes to the ideal model contribution, $n(\mathbf{x}) - n_0(\mathbf{x})$ in (5).

We study a non-metallic crystal, graphite, and calculate the ground state density for several hundred primitive cell configurations sampled from a NPT molecular dynamics trajectory. The components of the discrete Fourier transform of the ideal model contribution are perturbed by additive Gaussian noise. By taking the inverse transform, we have a continuously deformed version of the true density. We measure deformation by the root-mean-squared error (RMSE) of the perturbed and true ground state density.

In figure 3, the self-consistent calculation is initialised with charge densities which are increasingly deformed versions of the ground state. We see that as the magnitude in deformation from the ground state, the RMSE, increases, so too does the number of iterations needed to reach self-consistency. The quantity displayed on the abscissa, dSCF, is the improvement, in the number of iterations, relative to a calculation with the standard initial density. As deformations increase, the improvement decreases. The hashed and shaded areas in figure 3 represent confidence intervals of 67%, showing that the relation between RMSE and convergence to self-consistency is stochastic to some degree.

To ensure that any empirical method for initialising KS DFT densities does not negatively affect convergence to self-consistency in regions where the empirical densities extrapolate poorly, uncertainty quantification is clearly needed. For further details of the DFT calculations in figure 3, see appendix C.

4. Applying global uncertainty

As we saw by the calculations in figure 3, a measure of confidence in density is necessary if we are to use empirical densities in DFT in a 'safe' manner. Not wanting to leave things worse than how we found them, we hope to ensure that every calculation initialised by a data-driven density does no worse than its ordinary counterpart.

To illustrate that a global measure of confidence in predictions, H from (4), can be applied to accelerate KS DFT, we first consider using empirical densities without using knowledge of their uncertainty. After training on five unit cell graphite configurations from a NVT molecular dynamics simulation at $T = 300$ K, we predict densities for all 300 crystals in our data set. Details of the empirical model and KS DFT calculations can be found in appendix C. Without applying any information about uncertainty, we blindly initialise Broyden density mixing (DM) DFT calculations and record the reduction in the number of iterations to self-consistency, dSCF, relative to a calculation with a standard initial density. Next, we calculate the global confidence measure
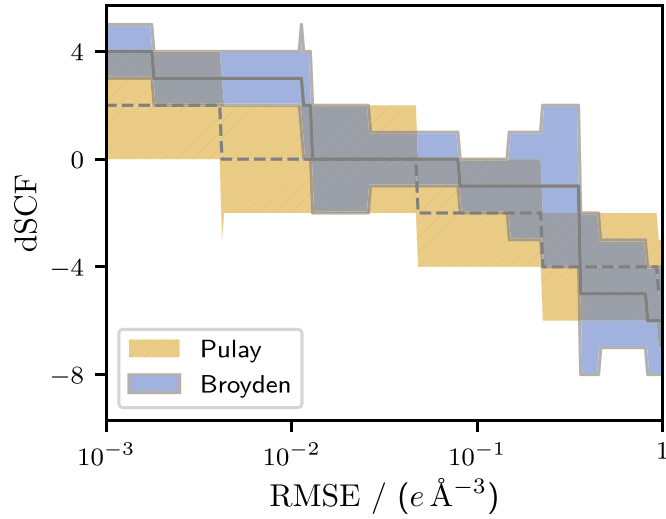
Figure 3. As the initial KS DFT densities are deformed away from the true ground state, the number of iterations necessary to reach self-consistency increases and the improvement from standard DFT initialisation, dSCF, decreases. Pulay and Broyden density mixing schemes result in a very similar convergence for all calculations.
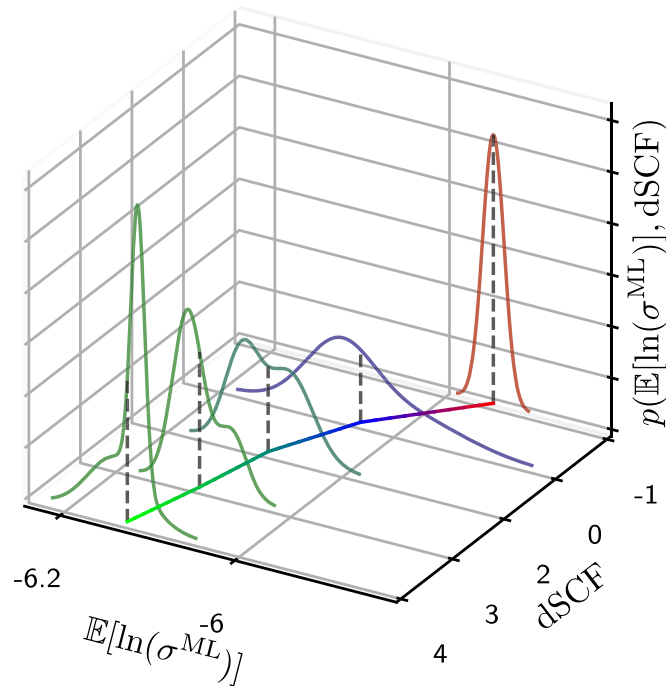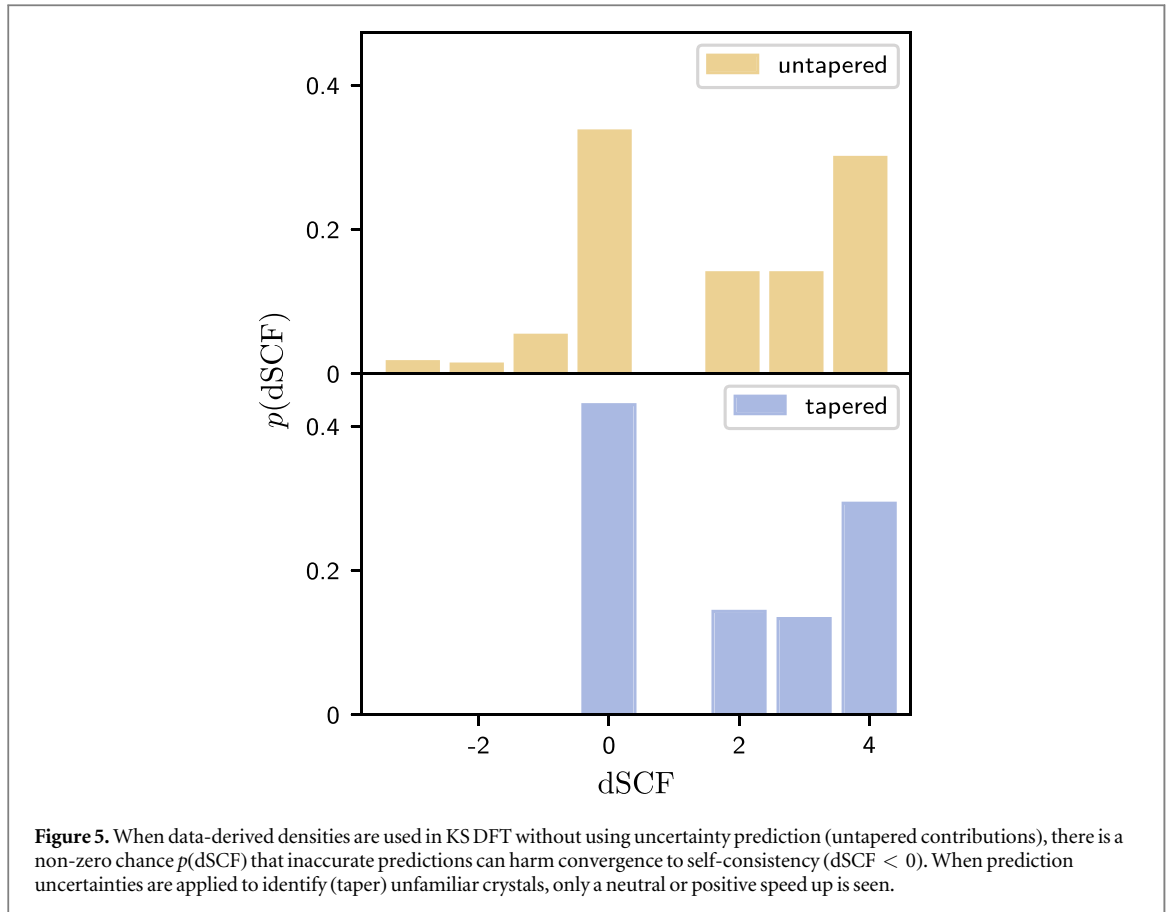

Figure 4. The empirical joint distribution $p(H, \text{dSCF})$ evaluated for a set of primitive graphite configurations shows separable categorisations of performance into good ($\text{dSCF} > 0$) and poor ($\text{dSCF} < 0$) predictions.

$H = \mathbb{E}[\ln(\sigma^{\text{ML}})]$ for each crystal and categorise crystals into discrete sets according to their dSCF score. We show the corresponding empirical joint distribution $p(\mathbb{E}[\ln(\sigma^{\text{ML}})], \text{dSCF})$ in figure 4.

We can expand the empirical joint distribution

$$p(H, \text{dSCF}) = p(\text{dSCF}|H)p(H) \tag{6}$$

in terms of the unknown conditional distribution $p(\text{dSCF}|H)$ from which we want to decide if a given prediction, H, is good enough to initialise a KS DFT calculation. Taking the prior $p(H)$ as constant, $p(H, \text{dSCF}) \propto p(\text{dSCF}|H)$ and we look for a gap in H between $p(H, \text{dSCF} \geqslant 0)$ and $p(H, \text{dSCF} < 0)$. It is here that a transition point can be set in the tapering function $\Gamma(H)$, to reduce uncertain predictions to zero. This can be visualised by comparing the peak at $\text{dSCF} = -1$ with that at $\text{dSCF} = 0$. Although there is some overlap between these two peaks, there is almost zero overlap between $\text{dSCF} = -1$ and all other peaks. This means that H can be used to identify the quality of density predictions before any DFT calculation is made. We note that the

Figure 5. When data-derived densities are used in KS DFT without using uncertainty prediction (untapered contributions), there is a non-zero chance $p(\text{dSCF})$ that inaccurate predictions can harm convergence to self-consistency (dSCF < 0). When prediction uncertainties are applied to identify (taper) unfamiliar crystals, only a neutral or positive speed up is seen.

joint distribution shown in figure 4 is a smoothed approximation of the true empirical distribution but that important properties such the width of each conditional distribution $p(\text{dSCF}|H)$, are preserved.

In fact we see that for the small study here, the expectation of H conditioned on dSCF

$$\mathbb{E}_{p(H|\text{dSCF})}[H] = \int \mathrm{d}H p(H|\text{dSCF}) H \tag{7}$$
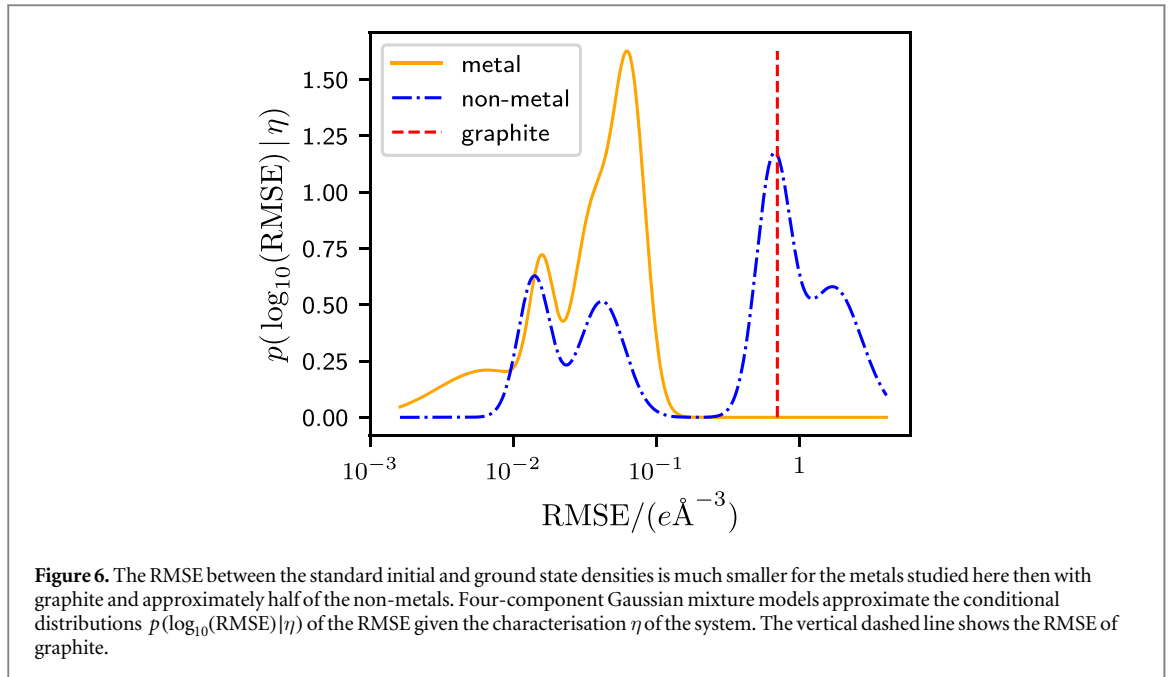
which is the dashed line in figure 4, follows a monotonic relation with dSCF. This shows that predicted uncertainty really does correspond in a monotonic way to actual error. Using the distribution of predicted uncertainties over a crystal, we can identify model predictions which are poor and will harm convergence to self-consistency. By effectively turning off poor predictions using $\Gamma(H)$, empirical corrections to the initial KS density can be applied only for crystals which are similar to those seen during training.

4.1. Accelerating self-consistency

Now that we have established a mechanism to detect global uncertainty in density, we can apply this to single point KS DFT calculations to accelerate convergence to self-consistency. In figure 5, we compare the empirical distributions $p(\text{dSCF})$ for Broyden DM DFT calculations performed using data-derived densities with and without tapering. The upper half of figure 5 shows a number of extrapolations where poor predictions of density have a negative effect upon convergence (dSCF < 0). In the lower half, uncertain predictions have been identified and reduced to zero, increasing the peak at dSCF $= 0$.

Crucially, the computation time required to evaluate our data-derived density estimate is just less than the time taken to evaluate a single SCF iteration. For further details of the calculations in figure 5 involving KS DFT parameters, see appendix C. Despite our model being trained only on 5 configurations, a large proportion of crystals exhibit a speed up in convergence to self-consistency. Such an effect could also arise by poorly choosing a test set of crystals, whereby all atom positions remain in almost identical positions. A trivial approach of applying the ground state density from a random crystal, or an average of ground state densities over all crystals, would therefore achieve similar, or better results. However, in fact this is not the case. When such simulations were run, we found that almost all (~95%) of predictions obtained dSCF $= 0$. Our test set is in fact a rather dissimilar collection of configurations, most of which involve significant shifts in registry across the basal plane, as configurations jump from one AB-stacked state to another. We attribute the ability of our model to infer useful predictions from such an incredibly small number of configurations to the fact that each crystal in the training

Figure 6. The RMSE between the standard initial and ground state densities is much smaller for the metals studied here then with graphite and approximately half of the non-metals. Four-component Gaussian mixture models approximate the conditional distributions $p(\log_{10}(\mathrm{RMSE})|\eta)$ of the RMSE given the characterisation η of the system. The vertical dashed line shows the RMSE of graphite.

set contributes $\mathcal{O}(10^4)$ grid points. Simply put, more data leads to a better inference, even when a large number of data points come from the same crystal.

4.2. Wider applicability

To this point, all calculations in this work have been made to illustrate that data-derived densities can be applied to a single system to improve the standard analytical initial densities that are used in KS DFT. We consider the wider implications of this work beyond graphite by comparing the dissimilarity of ground state and standard initial densities for a collection of 29 metals and 37 non-metals under both low and high pressure. We find that all of the metals we considered here have initial densities that are much closer to the ground state density than with graphite, while the converse is true for approximately half of the non-metals studied here. We use the RMSE of all density grid points within a crystal as a measure of dissimilarity between these two densities. To classify metals and non-metals we use the density of states at the Fermi level. We use a value of 0.2 $e(\mathrm{eV})^{-1}$ which is just above the density of states for the metalloid As to classify the two classes.

We show in figure 6 a smoothed approximation of the conditional distribution $p(\log_{10}(\mathrm{RMSE})|\eta)$ for metals or non-metals η along with a dashed vertical line showing the RMSE between the standard initial and ground state densities for graphite. The logarithm of the RMSE illustrates that the RMSE differs by almost two orders of magnitude between graphite and some of the metals. We note that the approximate representation of $p(\log_{10}(\mathrm{RMSE})|\eta)$ is a four-component Gaussian mixture model of the true data [15]. Based upon the systems studied here, we summarise that data-derived densities may in general be more suitably applied to non-metals than metals. Details of these systems as well as and the KS DFT calculations that were used to calculate the RMSE in figure 6 can be found in appendix C.

5. Discussion

We have shown that uncertainty quantification can be applied to accelerate KS DFT for sampling methods like nested sampling, attaining a maximum speed up of 57%[4] for the systems studied in this work. We view the approach taken in this work as more a proof of concept than a final solution, confident that exciting developments and insights are accessible to future work. To this end, we note that the approach taken in this work is just one of many possible methods. We use this section to discuss what we believe to be the most prominent disadvantages of this approach and outline a few ideas that could address these short comings.

While our parametric approach leads to a computation time for data-derived densities that is smaller than a single self-consistent field cycle, the time needed to train densities from a single crystal is orders of magnitude larger than a full DFT calculation. Although sampling methods require thousands of crystal configurations, the time to train or refine a data-derived density should ideally be as close to a single KS DFT calculation as possible.

---

[4] See appendix D for the definition of speed up that we adopt here.

Some heuristic techniques, such as maximising the sample variance of observations in a smaller training subset, may give some reduction in this, but a more promising avenue could be to use approximate Bayesian inference, such as deterministic variational inference [27]. A Bayesian approach, even when the posterior distribution is approximate, could also lead to more reliable uncertainty estimates. The non-Bayesian approach adopted in this work does not guarantee that 'false positives' cannot occur when determining if confidence should be placed in a data-derived density or not. In addition, Bayesian online learning could allow for an incremental approach to learning densities such that refinements are made during sampling only to crystals which are dissimilar to all of those that were previously seen during training [28].

The approach that we use in this work to make decisions about confidence in the density does not take account of the type of crystal. For applications like nested sampling, where several different phases are sampled from, it may become essential for our decision process to include knowledge about the global environment, such as from a global bispectrum representation of the crystal. An unsupervised method, such as a Gaussian mixture model, may be necessary to associate crystals with nearby clusters and to apply decisions using a predetermined set of distinct confidence thresholds for each cluster.

An aspect that we have not considered in this work is the question of which method of minimising the KS Hamiltonian, given an initial density, gives the lowest computation time. Although this is a well studied problem, perhaps new insights are possible when an estimate of confidence is available in the initial density [29].

We note that our discussion of KS DFT and the application of data-derived densities to accelerate convergence to self-consistency in this work has so far ignored spin. For many systems and processes such as radicals, transition metal complexes or homolytic bond breaking, the spatial wave functions of opposing spin states are not equal ($\psi^\alpha(\mathbf{r}) \neq \psi^\beta(\mathbf{r})$) [30–32]. Spin-unrestricted KS DFT is a generalisation of the spin-restricted form, where $\psi^\alpha(\mathbf{r}) \neq \psi^\beta(\mathbf{r})$ is possible and the variational minimisation of total energy $E[n, Q]$ is performed with respect to both the total electron density $n(\mathbf{r})$ and the spin density $Q(\mathbf{r}) = \sum_i |\psi^\alpha(\mathbf{r})|^2 - \sum_i |\psi^\beta(\mathbf{r})|^2$ [33]. Initial densities for unrestricted spin therefore require $Q(\mathbf{r})$ in addition to $n(\mathbf{r})$. A generalisation of the data-derived densities used in this work to unrestricted spin DFT could be realised by adopting $p(\mathbf{t}|\mathbf{x}, \mathbf{w}) = \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ for $\mathbf{t} = (n(\mathbf{r}), Q(\mathbf{r}))$. A parametric model would then represent $\mathbf{x} \rightarrow (\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ rather then $\mathbf{x} \rightarrow (\mu, \sigma^2)$ as in (1). The generalisation of (2) leads to $p(\mathbf{t}|\mathbf{x}, \mathbf{W}) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{\mu}^{\mathrm{ML}}, (\boldsymbol{\Lambda}^{\mathrm{ML}})^{-1})$ where $\boldsymbol{\mu}^{\mathrm{ML}} = (n^{\mathrm{ML}}, Q^{\mathrm{ML}})$ and the covariance matrix $(\boldsymbol{\Lambda}^{\mathrm{ML}})^{-1}$ represents uncertainty in the initial data-derived total (n^{ML}) and spin (Q^{ML}) densities. The simplest way to apply $\boldsymbol{\Lambda}^{\mathrm{ML}}$ to identify uncertain predictions might be to sum the diagonal components of $(\boldsymbol{\Lambda}^{\mathrm{ML}})^{-1}$ to define a scalar measure analogous to σ^{ML} in (4). We also note that $E[n, Q]$ is well known to exhibit a number of stationary points and in the absence of any knowledge about the ground state of Q, some form of approximate global optimisation must be utilised. If the data-derived densities are sufficiently accurate then global optimisation for spin-unrestricted DFT could be abandoned altogether, providing significant reductions to the computation required.

6. Conclusions

We have shown that a non-Bayesian treatment of predictive uncertainty can be applied to electron density regression to identify structures that are dissimilar to those seen during training. We have applied this approach to KS DFT where we have been able to identify and prevent unfamiliar structures from negatively effecting convergence to self-consistency. For the systems studied in this work, where confident predictions were made, we saw a maximum speed up in convergence to self-consistency of 57% (see footnote 4) and cautiously note that further improvements could be made with a more in depth study of the approach to minimise the KS Hamiltonian. Crucially, our predictions can be evaluated in less time than a single self-consistent field iteration for a primitive crystal, meaning that our application to KS DFT could be useful for methods like nested sampling.

We view this work as a proof of concept. Quantifying uncertainty in predicted densities is shown to be a fruitful endeavour and we hope our work will encourage further applications and alternative approaches, for example in OF DFT. More generally, this work motivates more sophisticated treatments of interpolation, or caching, which are currently treated deterministically to accelerate high performance plane wave DFT codes [34, 35]. We anticipate that a paradigm shift towards 'probabilistic caching', or regression, will lead to the efficient use of previously computed data.

Acknowledgments

Appendix A.

In the bispectrum approximation, elements of local and global contributions to the representation of environment, $\mathbf{x}(\mathbf{r})$, are determined by the projections c_{nlm} of local and global environment into radial (n) and spherical harmonic (lm) bases.

$$c_{nlm}^{\text{local}}(\mathbf{r}) = \sum_{i \in \Omega_{\mathbf{r}}} g_n(\mathrm{d}r_i) \, Y_{lm}(\mathrm{d}\theta_i, \, \mathrm{d}\phi_i)$$

$$c_{nlm}^{\text{global}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in \Omega_{\mathbf{r}_i}} g_n(\mathrm{d}r_{ij}) \, Y_{lm}(\mathrm{d}\theta_{ij}, \, \mathrm{d}\phi_{ij}). \tag{A.1}$$

$\Omega_{\mathbf{r}}$ is a spherical volume of finite radius surrounding a point in real space $\mathbf{r}$. Indices i and j denote pairs of the N atoms contained in the primitive cell of a crystal. $\mathrm{d}r_i = |\mathbf{r}_i - \mathbf{r}|$, $\mathrm{d}r_{ij} = |\mathbf{r}_j - \mathbf{r}_i|$ and $\mathrm{d}\theta$ and $\mathrm{d}\phi$ are the polar and azimuth projections respectively of $\mathbf{r}_i - \mathbf{r}$ and $\mathbf{r}_j - \mathbf{r}_i$.

Appendix B.

There are many equally adequate tapering functions which could be chosen. We used:

$$\Gamma(\sigma(\mathbf{r}), \, \sigma^*) = \begin{cases} \dfrac{\tilde{\sigma}^4}{1 + \tilde{\sigma}^4}, & \tilde{\sigma} > 0 \\ 0, & \tilde{\sigma} < 0 \end{cases} \qquad \tilde{\sigma} = \frac{\sigma^* - \sigma(\mathbf{r})}{\sigma_{\text{scale}}} \tag{B.1}$$

simply because it has property that every nth derivative $\partial^{(n)}\Gamma/\partial\sigma(\mathbf{r})^{(n)}$ remains continuous.

Appendix C.

For all of the KS DFT calculations in table C1, a Fermi surface smearing width of 250 K, an energy tolerance of 10^{-6} eV/atom, the PBE exchange correlation functional and Broyden DM were used, except for the calculations in figure 3 which used both Broyden and Pulay DM and the calculations in figure 6 which used a smearing width of 300 K. All graphite and graphene configurations except for the NPT calculations of figure 3 had an in-plane C–C spacing of 1.42 Å and an inter-layer spacing of 3.34 Å. In addition, a vacuum corresponding to a unit cell of 20 Å in the plane-normal axis was adopted for the graphene layer in figure 1 and in figure 5 a tapering function of the form in (B.1) was used with the threshold and scaling factor $(\sigma^*, \sigma_{\text{scale}}) = (-6.83, \, 10^{-3})$. *denotes use of the power spectrum rather than bispectrum representation for the calculations in figure 2. The notation adopted in table C1 regarding the number of nodes used in each neural network, is that $x \times y$ denotes a neural network of x node layers, each containing y nodes. Table C2 lists the database, unique identifying number and characterisation of each system used to generate the calculations in figure 6. We supply input files for all data sets within this work at https://github.com/andrew31416/densityregression/tree/master/data_sets.

Table C1. Calculation parameters.

Calculation	Bispectrum			Neural network		KS DFT	
	r_{cut} (Å)	n_{max}	l_{max}	N_{ens}	nodes	E_{cut} (eV)	k-point grid
Figure 1	4	3	3	5	2×100	400	[36 36 1]
Figure 2	6	10^*	8^*	10	2×50	400	[36 36 6]
Figure 3	—	—	—	—	—	800	[36 36 6]
Figure 4	6	6	6	10	2×200	300	[18 18 4]
Figure 5	4	4	4	5	2×150	300	[18 18 4]
Figure 6	—	—	—	—	—	800	$(0.1, 0.1, 0.1)$Å^{-1}

Table C2. The collection of metals and non-metals from figure 6 were taken from the crystallography open database (COD) [36] and the inorganic crystal structure database (ICSD) [37]. A unique identifying number is given for each crystal along with its characterisation as discussed in section 4.2.

Database	Identifier	Characterisation	Database	Identifier	Characterisation
COD	9008572	Non-metal	COD	9008531	Metal
COD	9008594	Non-metal	COD	9008468	Metal
COD	9008595	Non-metal	COD	9008544	Metal
COD	9008569	Non-metal	COD	9008482	Metal
COD	9008577	Non-metal	COD	9008478	Metal
COD	9008561	Non-metal	COD	9008485	Metal
COD	1011098	Non-metal	COD	9008463	Metal
COD	9008568	Non-metal	COD	9008458	Metal
ICSD	193853	Non-metal	COD	9008552	Metal
ICSD	26158	Non-metal	COD	9008501	Metal
ICSD	9863	Non-metal	COD	9008490	Metal
ICSD	18154	Non-metal	COD	9008522	Metal
ICSD	16516	Non-metal	COD	9008470	Metal
ICSD	15598	Non-metal	COD	9008549	Metal
ICSD	41440	Non-metal	COD	9008513	Metal
ICSD	2130	Non-metal	COD	9008536	Metal
ICSD	411857	Non-metal	COD	9008514	Metal
ICSD	27249	Non-metal	COD	9008546	Metal
ICSD	20904	Non-metal	COD	9008584	Metal
ICSD	22156	Non-metal	COD	9008570	Metal
ICSD	84461	Non-metal	COD	9008543	Metal
ICSD	15390	Non-metal	COD	9008525	Metal
ICSD	39566	Non-metal	COD	9008512	Metal
ICSD	27798	Non-metal	COD	9008557	Metal
ICSD	40914	Non-metal	COD	9008477	Metal
ICSD	16428	Non-metal	COD	9008558	Metal
ICSD	165592	Non-metal	ICSD	15535	Metal
ICSD	22157	Non-metal	ICSD	63670	Metal
ICSD	19079	Non-metal	ICSD	653014	Metal
ICSD	29128	Non-metal			
ICSD	107946	Non-metal			
ICSD	60559	Non-metal			
ICSD	16262	Non-metal			
ICSD	187642	Non-metal			
ICSD	22158	Non-metal			
ICSD	18012	Non-metal			
ICSD	77378	Non-metal			

Appendix D.

We adopt the convention that the speed up

$$\tau = \left(\frac{N_{\text{old}} - N_{\text{new}}}{N_{\text{new}}} \right) \times 100\% \tag{D.1}$$

for data-derived initial densities that require N_{new} self-consistent field iterations to reach self-consistency. N_{old} is the number of iterations required for a standard calculation that uses a non data-derived initial density. τ is defined such that a data-derived density that halves then required number of self-consistent field iterations corresponds to a 100% speed up.

ORCID iDs

Andrew T Fowler ⓘ https://orcid.org/0000-0002-7360-3078
Chris J Pickard ⓘ https://orcid.org/0000-0002-9684-5432
James A Elliott ⓘ https://orcid.org/0000-0002-4887-6250

References

[1] Burke K 2012 Perspective on density functional theory *J. Chem. Phys.* **136** 150901
[2] Jones R O 2015 Density functional theory: its origins, rise to prominence, and future *Rev. Mod. Phys.* **87** 897–923

[3] Hohenburg P and Kohn P 1973 Inhomogeneous electron gas *Phys. Rev.* B **7** 1912–9

[4] Cohen A J, Mori-Sánchez P and Yang W 2012 Challenges for density functional theory *Chem. Rev.* **112** 289–320

[5] Kohn W and Sham L J 1965 Self-consistent equations including exchange and correlation effects *Phys. Rev.* **140** A1133–8

[6] Becke A D 1986 Density functional calculations of molecular bond energies *J. Chem. Phys.* **84** 4524–9

[7] Witt W C, Del Rio B G, Dieterich J M and Carter E A 2018 Orbital-free density functional theory for materials research *J. Mater. Res.* **33** 777–95

[8] Payne M C, Teter M P, Allan D C, Arias T A and Joannopoulos J D 1992 Iterative minimization techniques for *ab initio* total-energy calculations: molecular dynamics and conjugate gradients *Rev. Mod. Phys.* **64** 1045–97

[9] Snyder J C, Rupp M, Hansen K, Müller K R and Burke K 2012 Finding density functionals with machine learning *Phys. Rev. Lett.* **108** 253002

[10] Li L, Snyder J C, Pelaschier I M, Huang J, Niranjan U N, Duncan P, Rupp M, Müller K R and Burke K 2016 Understanding machine-learned density functionals *Int. J. Quantum Chem.* **116** 819–33

[11] Brockherde F, Vogt L, Li L, Tuckerman M E, Burke K and Müller K R 2017 Bypassing the Kohn–Sham equations with machine learning *Nat. Commun.* **8** 872

[12] Sinitskiy A V and Pande V S 2018 Deep neural network computes electron densities and energies of a large set of organic molecules faster than density functional theory (DFT) arXiv:1809.02723

[13] Grisafi A, Wilkins D M, Fabrizio A, Meyer B, Corminboeuf C and Ceriotti M 2018 Transferable machine-learning model of the electron density *ACS Cent. Sci.* **5** 57–64

[14] Brewer B J, Pártay L B and Csányi G 2011 Diffusive nested sampling *Stat. Comput.* **21** 649–56

[15] Bishop C M 2006 *Pattern Recognition and Machine Learning* (Berlin: Springer)

[16] Ghahramani Z 2015 Probabilistic machine learning and artificial intelligence *Nature* **521** 452–9

[17] Aldegunde M, Kermode J R and Zabaras N 2016 Development of an exchange-correlation functional with uncertainty quantification capabilities for density functional theory *J. Comput. Phys.* **311** 173–95

[18] Aldegunde M, Zabaras N and Kristensen J 2016 Quantifying uncertainties in first-principles alloy thermodynamics using cluster expansions *J. Comput. Phys.* **323** 17–44

[19] Mo S, Zhu Y, Zabaras N, Shi X and Wu J 2018 Deep convolutional encoder-decoder networks for uncertainty quantification of dynamic multiphase flow in heterogeneous media *Water Resour. Res.* **55** 703–28

[20] Longbottom S and Brommer P 2019 Uncertainty quantification for classical effective potentials: an extension to potfit *Modelling Simul. Mater. Sci. Eng.* **27** 044001

[21] Parks H L, McGaughey A J H and Viswanathan V 2019 Uncertainty quantification in first-principles predictions of harmonic vibrational frequencies of molecules and molecular complexes *J. Phys Chem.*C **123** 4072–84

[22] Bartók A P 2010 Gaussian approximation potential: an interatomic potential derived from first principles quantum mechanics *PhD Thesis* University of Cambridge

[23] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments *Phys. Rev.* B **87** 184115

[24] Pearson K 1896 Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia *Phil. Trans. R. Soc.* A **187** 253–318

[25] Lakshminarayanan B, Pritzel A and Blundell C 2017 Simple and scalable predictive uncertainty estimation using deep ensembles *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (Red Hook, NY: Curran Associates, Inc.)

[26] Monthioux M and Charlier J C 2014 Giving credit where credit is due: the stone-(thrower)-wales designation revisited *Carbon* **75** 1–4

[27] Wu A, Nowozin S, Meeds E, Turner R E, Hernández-Lobato J M and Gaunt A L 2018 Fixing variational Bayes: deterministic variational inference for Bayesian neural networks arXiv:1810.03958

[28] Opper M 1999 A Bayesian approach to on-line learning *On-Line Learning in Neural Networks* ed D Saad (Cambridge: Cambridge University Press) pp 363–78

[29] Woods N 2018 On the nature of self-consistency in density functional theory *Masters Thesis* University of Cambridge

[30] Gräfenstein J, Kraka E, Filatov M, Cremer D, Gräfenstein J, Kraka E, Filatov M and Cremer D 2002 Can unrestricted density-functional theory describe open shell singlet biradicals? *Int. J. Mol. Sci.* **3** 360–94

[31] Power P P 2012 Stable two-coordinate, open-shell (d1–d9) transition metal complexes *Chem. Rev.* **112** 3482–507

[32] Yamanaka S, Kawakami T, Nagao H and Yamaguchi K 1994 Effective exchange integrals for open-shell species by density functional methods *Chem. Phys. Lett.* **231** 25–33

[33] Jacob C R and Reiher M 2012 Spin in density-functional theory *Int. J. Quantum Chem.* **112** 3661–84

[34] Clark S J, Segall M D, Pickard C J, Hasnip P J, Probert M I, Refson K and Payne M C 2005 First principles methods using CASTEP *Z. Kristallogr.* **220** 567–70

[35] Hafner J 2008 *Ab-initio* simulations of materials using VASP: density-functional theory and beyond *J. Comput. Chem.* **29** 2044–78

[36] Gražulis S, Daškevič A, Merkys A, Chateigner D, Lutterotti L, Quirós M, Serebryanaya N R, Moeck P, Downs R T and Le Bail A 2012 Crystallography open database (COD): an open-access collection of crystal structures and platform for world-wide collaboration *Nucleic Acids Res.* **40** D420–7

[37] Hellenbrandt M 2004 The inorganic crystal structure database (ICSD)–present and future *Crystallogr. Rev.* **10** 17–22