Machine Learning of Dynamic Electron Correlation Energies from Topological Atoms

James L. McDonagh,† Arnaldo F. Silva,† Mark A. Vincent,‡ and Paul L. A. Popelier*†‡

†Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester M1 7DN, Great Britain
‡School of Chemistry, The University of Manchester, Oxford Road, Manchester M13 9PL, Great Britain

ABSTRACT: We present an innovative method for predicting the dynamic electron correlation energy of an atom or a bond in a molecule utilizing topological atoms. Our approach uses the machine learning method Kriging (Gaussian Process Regression with a non-zero mean function) to predict these dynamic electron correlation energy contributions. The true energy values are calculated by partitioning the MP2 two-particle density-matrix via the Interacting Quantum Atoms (IQA) procedure. To our knowledge, this is the first time such energies have been predicted by a machine learning technique. We present here three important proof-of-concept cases: the water monomer, the water dimer, and the van der Waals complex H₂···He. These cases represent the final step toward the design of a full IQA potential for molecular simulation. This final piece will enable us to consider situations in which dispersion is the dominant intermolecular interaction. The results from these examples suggest a new method by which dispersion potentials for molecular simulation can be generated.

1. INTRODUCTION

Dynamic electron correlation is a quantum mechanical phenomenon with profound ramifications for intermolecular interactions.1-3 Typical quantum chemical methods commonly employed, such as DFT, generally do not completely include many of these correlation effects, and in some cases they are not included at all. To include them completely requires expensive quantum mechanical methods. However, a plethora of methods now exists to obtain Electron Correlation Energies (ECE). These methods range from empirical potentials in simulation methods4-12 to a posteriori corrections in DFT or ab initio wave function methods.13-18 In addition, more recently, many-body correction terms have been employed to supplement DFT16,19.

These methods have been applied to a wide variety of important chemical systems including the solid state,20-22 biological materials,23 and more recently nanostructures.24-26 Recent work has seen much improvement in methods16,27 to account for ECE. Although electron correlation is proportionally a small contribution energetically, it is ubiquitous and has been shown to be responsible for macroscopic phenomena, such as an explanation of the gecko’s ability to adhere to a variety of surfaces.28,29

Many of these proposed corrections rely upon the use of so-called damping functions to attenuate the interactions between two atoms, which would become unrealistically large without such a mathematical device. The use of these damping functions has been questioned (Section 2.11 in ref 29). Ideally, such purely mathematical constructs are not required, and indeed, they are not if one works with topological atoms,30,31 as we do here. These are essentially regions that naturally appear in a system’s electron density without using any parameters. They are space-filling, which means that there are no gaps between them, and they do not overlap. The latter property is important because topological atoms do not penetrate each other (see ref 33 and Figure S1 in the Supporting Information (SI)), and hence, no associated damping function is required. The space-filling nature of the topological atoms also avoids the well-known polarization catastrophe (occurring with overlapping charge clouds) and, hence, avoids the corresponding damping functions. Consequently, when using topological atoms, there is no need to work out how such damping functions (for either penetration or polarization catastrophe) change with atom type.

Machine learning (ML) is a popular tool in many areas of science and has found a variety of applications in chemistry.34-40 Here, we present a novel application in which models capable of accurate predictions of dynamic electron correlation energy are employed. This means that, in exchange for a relatively small number of initial expensive correlation calculations, we can make accurate predictions of unknown correlation energies for related geometries in a fraction of the...
time. This method holds the promise of a new way to derive van der Waals energies for use in simulations.

In order to generate such machine learning models, the chemical system must be appropriately represented. There are many approaches to representing a chemical system for input to machine learning methods. Here, we use our own method, which employs the Atomic Local Frame (ALF) approach. This technique defines a local axis system, in spherical coordinates, centered on an atom of interest. This leads to a relative portrayal of the system’s geometry as viewed from the atom of interest. An ALF is generated for each atom of the system, and the respective atomic ECE is then linked with it as a target “true” value, for that given geometry. The ML models task is to find a relationship between the ALF and the ECE value for each geometry.

Our ultimate aim is to apply this method to the novel force field FFLUX, which we are currently developing. FFLUX is a novel next-generation force field based upon a ML method called Kriging. This approach has previously showcased the ability of such Kriging models to predict, with promising accuracy, the Coulombic and exchange energies for chemical systems in a variety of molecular arrangements. Kriging models have been applied successfully on a variety of systems, including ethanol, tetrahydrofuran, cyclohexane, and histidine, the four aromatic (peptide-capped) amino acids, all naturally occurring amino acids, helical decana-amides, water clusters, cholesterol, and carbohydrates. Although the roots of FFLUX lie in electrostatics, particularly multipolar electrostatics, non-electrostatic energy contributions were also successfully trained for using Kriging: for example, the atomic kinetic energy, exchange energies, which were also successfully Kriged for the B3LYP functional. In the current article we present the approach has previously showcased the ability of such Kriging method as implemented in the in-house code MORPHY, a local derivative of MORPHY. A quantum mechanically rigorous prescription to partition the electron density, and hence the energy, of a molecule (or complex) into topological atoms, is calculated via a six-dimensional quadrature, implemented as a succession of two three-dimensional integrals carried out over the respective volumes Ω of atoms A and B:

\[
\begin{align*}
I^{\text{AB}}_{\text{ee, corr}} &= \sum_{j=1}^{N_A} \sum_{k=1}^{N_B} \sum_{l=1}^{N_j} \sum_{m=1}^{N_k} k_{ijlm}^{\text{corr, prim}} \int d \Omega_1 \int d \Omega_2 \frac{1}{r_{ij}} G_{ij}(\mathbf{r}_i - \mathbf{R}_j) \times \\
& \times \int d \Omega_3 d \Omega_4 \frac{1}{r_{jk}} G_{jk}(\mathbf{r}_j - \mathbf{R}_k) \\
I^{\text{AB}}_{\text{ee, corr}} &= \sum_{j=1}^{N_A} \sum_{k=1}^{N_B} \sum_{l=1}^{N_j} \sum_{m=1}^{N_k} k_{ijlm}^{\text{corr, prim}} \int d \Omega_1 d \Omega_2 d \Omega_3 d \Omega_4 \end{align*}
\]

The fact that the product of two Gaussian functions in eq 1 is a third Gaussian function has been employed to reduce the number of Gaussians to two from the four primitive ones, corresponding to j, k, l, and m indices. Note that A and B can refer to the same atom, in which case eq 1 returns the ECE within a single atom (i.e., the intra-atomic dynamic electron correlation). In eq 1, k_{ijlm} are prefactors from the producing of the primitive Gaussian basis functions. The upper bounds in the summations over indices j and k, or over l and m, mean that only the lower triangle and the diagonal terms of k_{ijlm} need to be considered. Full details of this method are available elsewhere.

2.2. Correlation Energy. For each of TYCHE’s generated geometries the ab initio program GAUSSIAN09 provides the correlated part of the two-particle reduced density-matrix (2PDM), arising from an MP2/uncon-6-31+G(d,p) calculation, where uncon emphasizes that the basis set is uncontracted. We stress that the 6-31+G(d,p) basis set is uncontracted and thus provides a better description of the systems considered here compared to the normal contracted form.

The 2PDM, which is a four-dimensional matrix, is then normalized to the primitive atomic orbital basis, and its elements are formally denoted d_{\text{corr, prim}}. A detailed derivation of this matrix is provided in Section 2 of the SI. The four-dimensional matrix d_{\text{corr, prim}} is provided as input for the IQA method as implemented in the in-house code MORPHY. A quantum mechanically rigorous prescription to partition the electron density, and hence the energy, of a molecule (or complex) into topological atoms, is calculated via a six-dimensional quadrature, implemented as a succession of two three-dimensional integrals carried out over the respective volumes Ω of atoms A and B:

\[
\begin{align*}
I^{\text{AB}}_{\text{ee, corr}} &= \sum_{j=1}^{N_A} \sum_{k=1}^{N_B} \sum_{l=1}^{N_j} \sum_{m=1}^{N_k} k_{ijlm}^{\text{corr, prim}} \int d \Omega_1 \int d \Omega_2 \frac{1}{r_{ij}} G_{ij}(\mathbf{r}_i - \mathbf{R}_j) \times \\
& \times \int d \Omega_3 d \Omega_4 \frac{1}{r_{jk}} G_{jk}(\mathbf{r}_j - \mathbf{R}_k) \\
I^{\text{AB}}_{\text{ee, corr}} &= \sum_{j=1}^{N_A} \sum_{k=1}^{N_B} \sum_{l=1}^{N_j} \sum_{m=1}^{N_k} k_{ijlm}^{\text{corr, prim}} \int d \Omega_1 d \Omega_2 d \Omega_3 d \Omega_4 \end{align*}
\]

The fact that the product of two Gaussian functions in eq 1 is a third Gaussian function has been employed to reduce the number of Gaussians to two from the four primitive ones, corresponding to j, k, l, and m indices. Note that A and B can refer to the same atom, in which case eq 1 returns the ECE within a single atom (i.e., the intra-atomic dynamic electron correlation). In eq 1, k_{ijlm} are prefactors from the producing of the primitive Gaussian basis functions. The upper bounds in the summations over indices j and k, or over l and m, mean that only the lower triangle and the diagonal terms of k_{ijlm} need to be considered. Full details of this method are available elsewhere.

2.3. Machine Learning (ML). After having partitioned the MP2 energy, which arises from each of the TYCHE generated geometries, a training set file is produced. This file contains geometries as input and energies as output. In particular, this file defines the molecule in terms of ML input variables, which are called features, and are in this case the ALF coordinates. Because we are only interested in training for internal geometry changes, there are 3N−6 features for nonlinear systems, where N is the number of atoms. The training set files are automatically generated using a Python script (see Section 3 of the SI) and in-house programs.

The three ML methods Kriging, Random Forest (RF), and Support Vector Regression (SVR) are employed in the present study and are discussed in detail in Section 4 of the SI. Briefly, Kriging, which is the machine learning (ML) method of
choice as an interpolator, maps a set of input data to an output response value, via a Gaussian process conditioned by a kernel function. This method is implemented in our program FEREBUS. While Kriging has several variants, we have applied only one form and that is simple Kriging. In simple Kriging the output value \( y \) depends on a set of \( n \) input points \( x \) and is modeled by the mean of the output \( \mu \) plus an “error” or deviation \( \epsilon \) term, which relates to, and is explicitly dependent on, the geometry-specific effects (marked by \( x \)) on the ECE (or \( y \)). Second, RF is an ensemble learning method based upon the concept of decision trees. A forest of decision trees is generated, in which each tree is “grown” using a subset of the training data and features. The nodes, branching these trees, are decided on based upon minimizing an error metric, which is the Root Mean Square Error (RMSE) in this case. Finally, SVR searches for an optimal regression within a margin of acceptable error \( \epsilon \) in a higher order feature space via the “kernel trick”. The RF and SVR codes are used as implemented in the R package, Caret. Hyperparameter information for all models is provided in Section 4 of the SI, and final Kriging hyperparameters for each model are reported in the SI in Section 5. Note that due to the different training methods and hyperparameter optimization techniques the Kriging results are not directly comparable with the RF and SVR results. We have chosen the RF and SVR algorithms, with a 10-fold cross-validation, as these provide examples of widely accessible methods. Previously, such methods have been successfully applied to property prediction.

3. RESULTS

3.1. Water Monomer. 3.1.1. Introduction. The first proof-of-concept example we present is that of a single water molecule. This case study tests the ability of our ML methods to predict the atomic correlation energy contributions of a molecular system. We refer to these contributions as intra-molecular correlation energies, which are rarely studied yet are important to fully comprehend a chemical system’s electronic structure. We note that dispersion energies are traditionally considered inter-molecular in nature and determined by Symmetry Adapted Perturbation Theory, but in our approach they can be intra-molecular as well. However, the current method is not limited to through-space interactions only and can even calculate correlation within an atom bound inside a molecule. Thus, the intramolecular correlation energy can be divided up into intra-atomic correlation energy (i.e., the correlation energy within a single atom in the molecule of interest) and interatomic correlation energy (i.e., the correlation energy between two atoms in the molecule of interest). We note that these different atoms do not have to be directly bonded to each other. This full partitioning enables us to study the effects of electron correlation on both chemical bonds and atoms individually. For clarity, we note that the term interatomic may refer to intermolecular or intramolecular interactions.

3.1.2. Energy Partitioning Results. The program TYCHE generated 100 distorted water geometries around one energy minimum. HERMES determined the O–H bond had a distortion range of 0.2 Å (0.86–1.06 Å) with an average bond length of 0.95 Å, while the H–O–H angle had an average of 105.5° and a range of 80.7°. A visualization of the distortion is shown in Section 5 of the SI. A locally modified version of GAUSSIAN09 generated the \( d_{\text{ intra-atomic total}} \) matrix for each of these distorted geometries. Subsequently, MORFI partitioned the molecular energy into inter- and intra-atomic contributions. To provide a sense of scale, Table 1 gives the partitioned and total ECE of an arbitrarily distorted water molecule. The SI gives an overlay of all the distorted water geometries (Figure S4).

### Table 1. Partitioned and Total ECE Calculated by MORFI for a Distorted Water Molecule with an MP2/uncon-6-31+G(d,p) Wave Function

| intra-atomic | interatomic | total    |
|--------------|-------------|----------|
| O            | H1          | H2       | O–H1   | O–H2   | H1–H2   | H2O     |
| -521.0       | -22.5       | -25.6    | 15.9   | 18.4   | -5.1    | -539.8  |

*All values are quoted in kJ/mol. This table provides the quantitative details of the energy partitioning of a single geometry of the water monomer. These values are provided to give an indication of the magnitude of the differences in ECE between individual components of the total ECE.*

We can see clearly from Table 1 that the oxygen’s intra-atomic correlation energy is the dominant term. In addition, it is clearly noticeable from these energies that this is a distorted geometry with asymmetric hydrogen intra-atomic energies and interaction energies with oxygen. However, the energy-minimized structure, used as a seed for the distorted structures, was symmetric. Note that the total molecular dynamic ECE of \(-539.8 \text{ kJ/mol}\) is still relatively very small (0.3%) compared to the total energy for water, which is of the order of \(-200,000 \text{ kJ/mol}\). Considering all of the geometries, the average absolute error over the data set was 1.6 kJ/mol with a maximum absolute error of 1.7 kJ/mol (Figure S8).

3.1.3. Kriging Results. A training set of 40 geometries was created for the construction of Kriging models, which were tested on 60 geometries. Here we discuss the ML of seven energy terms: the O, H1, and H2 intra-atomic energies, the O–H1, O–H2, and H1–H2 interatomic energies, and finally the total ECE.

Table 2 summarizes the statistics of the predictions of the various ECE contributions with respect to the MORFI reconstruction energies.

Table 2 shows that the models are of excellent quality with all \( q_\text{fi} \) values of better than 0.99 and all RMSE values less than 0.5 kJ/mol. We note that the largest maximum absolute error originates from the oxygen intra-atomic energy and the same error is found in the total ECE, but even this error represents only 1.5% of the range of the predictions. From this point onward we focus our attention on the total ECE predictions. Below we display (Figure 2A) a so-called S-curve of the total ECE, which is a plot showing the percentage (y-axis) of total ECE predictions below a certain absolute error (x-axis). Table 2 and Figure 2A display excellent predictive accuracy, with all total ECE predictions below an absolute error of 0.5 kJ/mol and 90% of these predictions below an absolute error of 0.01 kJ/mol, over an energy range of 33.9 kJ/mol. This level of accuracy is sufficient for use in a molecular force field. This is particularly promising given the training set required consisted of only 40 geometries and was able to predict 60 varying geometries. Previous results for the IQA multipolar electrostatics and exchange have required much larger training sets (at least 5–10 times) to achieve similar levels of accuracy.

As a final test, it is useful to compare three MP2 correlation energies: (i) the original one from GAUSSIAN09 (MP2/uncon-6-31+G(d,p)), (ii) the “reconstructed” one, obtained from MORFI, which introduces a small numerical error via the...
grid used in its energy partitioning, and (iii) the Kriged one, obtained from FEREBUS. For convenience, we take the correlation of the first geometry in the test set as a reference. Figure S9 then shows relative energy from each calculation technique. We can see clearly that the relative energies are consistent between all three methods of calculation. In terms of chemistry, we are generally interested in the relative energy difference between configurations. Hence, it is vital for any calculation method to produce this accurately. Here we clearly demonstrate the utility of the Kriging models to meet this criterion.

3.1.4. RF and SVR Results. The same input features that were employed in the Kriging models were provided to the RF and SVR ML methods. Both RF and SVR were applied to predict the total ECE only, with the results summarized in Table 3.

Table 3. Total ECE Predictions by RF and SVR ML Methods for the Water Monomer

| measure          | O intra | H1 intra | H2 intra | O−H1 inter | O−H2 inter | H1−H2 inter | total ECE |
|------------------|---------|----------|----------|------------|------------|-------------|----------|
| max absolute error | 0.52    | 0.07     | 0.12     | 0.23       | 0.09       | 0.03        | 0.49     |
| 99% absolute error | 0.47    | 0.04     | 0.09     | 0.20       | 0.07       | 0.02        | 0.48     |
| 90% absolute error | 0.10    | 0.01     | 0.02     | 0.03       | 0.04       | 0.00        | 0.09     |
| $q^2$            | 0.9998  | 0.9999   | 0.9999   | 0.9997     | 0.9999     | 0.9999      | 0.9998   |
| $r^2$            | 0.9999  | 0.9999   | 0.9999   | 0.9998     | 0.9999     | 0.9999      | 0.9999   |
| mean absolute error (MAE) | 0.05     | 0.00     | 0.01     | 0.02       | 0.01       | 0.00        | 0.05     |
| root mean squared error (RMSE) | 0.12     | 0.01     | 0.03     | 0.05       | 0.02       | 0.00        | 0.12     |
| range            | 33.90   | 9.16     | 9.03     | 11.42      | 11.65      | 5.93        | 33.67    |

$^a$All energies are quoted in kJ/mol. These statistics display the quality and predictive power of the Kriging models produced for each individual ECE term and the total ECE.

The results in Table 3 clearly show that the total ECE is a predictable quantity in this case, for both RF and SVR approaches. In both these cases reasonable $r^2$ coefficients are matched by fairly low RMSE values for the predictions. Additionally the standard deviation of the predictions over the 10-fold cross-validations is low, showing that the results are of a reasonable accuracy independent of the training set selected.

3.2. H$_2$−He. The second example is that of the van der Waals complex H$_2$−He, which includes intermolecular as well as intramolecular predictions.

3.2.1. Energy Partitioning Results. The details of the distorted geometries for H$_2$−He are given in Section 5 of the SI. Table 4 gives the breakdown of the ECE for the first distorted geometry as an example. These values are provided to give an indication of the magnitude of the differences in ECE between individual components of the total ECE.

Table 4 shows that the helium intra-atomic energy is the largest single contribution, but it is not as singularly dominant as the oxygen intra-atomic correlation energy in the water examples. The distortions produce geometries for this complex ranging from nearly linear to asymmetric T-shape (see Section S.2 of the SI). We also note that the interaction energies between H and He show different signs. This suggests that one of the interactions is stabilizing, while the other is destabilizing in this geometric arrangement. However, taken together, the overall intermolecular interaction is neutral in this case although generally stabilizing. Similar observations have been made previously.$^{30,83}$

For all the distorted geometries of the H$_2$−He complexes, the maximum error in MORFI’s reconstruction of the correlation energy, for a given geometry, was 1.33 kJ/mol with an average error of 1.29 kJ/mol over a range of approximately 4 kJ/mol. It is shown in Figure S13 that, unlike for the water example, the correlation energy was more consistent over the range of geometries, with over half of these geometries having an ECE in the range −136 to −137 kJ/mol. Additionally, we note that the reconstruction is similar in accuracy to that of the H$_2$O monomer with a reasonably constant offset of approximately 1 kJ/mol. We have additionally tested the applicability of this method to the atomization of the H$_2$−He complex. These results are presented in Section 7.2 of the SI and show comparable accuracy levels to the current example but over a larger energy range of several hundreds of kJ/mol.

3.2.2. Kriging Results. For the H$_2$−He van der Waals complex a total of 15 geometries were used as training data, and 35 geometries were employed as prediction points. There are once again seven IQA ECE terms in this system. Table 5 summarizes the statistics arising from the Kriging models predictions of the test set with errors measured against the MORFI energy reconstructions.

The statistics in Table 5 show the same trends as in Table 2. The quality of the models is excellent as displayed by all $q^2$ values being over 0.98 and MAE and RMSE errors being less than 0.1 kJ/mol. The S-curve in Figure 2B displays the predictive quality of this model. The maximum absolute prediction errors are largest for the total ECE energy, with the single largest max absolute error for a single component being for H1−H2 interatomic ECE. Numerically, these energy values are all notably smaller than that of the water molecule. As with the water monomer example we will now focus on the total ECE. It is clear from Table 5, Figure 2B, and Figure S14, that
the accuracy of the prediction is very good for this intermolecular case. All predictions of the total ECE have an absolute error below 0.5 kJ/mol, and 90% of predictions have an absolute error below 0.12 kJ/mol. Once again this example required very few training points, 15 in total, and yet has a wide domain of applicability, predicting 35 test set points from the resultant models.

In terms of relative energy, Figure S14 shows the relative energy differences from the three calculation methods explained above. Despite the increased MORFI reconstruction error the relative errors remain very consistent across the three methods although there are small deviations. The fact that the Kriging models are very close to the energy landscape computed from MORFI demonstrates the ability and utility of these models as a basis for the dispersion energy component of a molecular force field.

3.2.3. RF and SVR Results. As with the water monomer example, we provided the same features to the RF and SVR ML models, for predictions of the total ECE. The results are summarized in Table 6.

Table 6. Total ECE Predictions by the RF and SVR ML Methods for the H$_2$···He van der Waals Complex$^{a}$

| measure       | H$_1$ intra | H$_2$ intra | He intra | H$_1$–H$_2$ inter | He–H$_1$ inter | He–H$_2$ inter | total ECE |
|---------------|-------------|-------------|----------|-------------------|----------------|----------------|----------|
| max absolute error | 0.09        | 0.08        | 0.00     | 0.10              | 0.00           | 0.00           | 0.43     |
| 99% absolute error | 0.06        | 0.06        | 0.00     | 0.09              | 0.00           | 0.00           | 0.20     |
| 90% absolute error | 0.01        | 0.01        | 0.00     | 0.06              | 0.00           | 0.00           | 0.12     |
| $q_r^2$         | 0.9999      | 0.9999      | 0.9963   | 0.9998            | 0.9905         | 0.9962         | 0.9853   |
| $r^2$           | 0.9999      | 0.9999      | 0.9963   | 0.9999            | 0.9925         | 0.9973         | 0.9867   |
| mean absolute error (MAE) | 0.01        | 0.01        | 0.00     | 0.02              | 0.00           | 0.00           | 0.04     |
| mean root mean squared error (RMSE) | 0.00        | 0.00        | 0.02     | 11.59             | 0.02           | 0.04           | 3.60     |

$^a$These statistics display the quality and predictive power of the Kriging models produced for each individual ECE term and the total ECE.

Figure 1. Water dimer atom labeling.

Again the total ECE is a predictable quantity by the RF and SVR methods. Both achieve reasonable $r^2$ values and low RMSEs. We note the standard deviations over the 10-fold cross-validation are once again low, showing a consistency over different choices of training and test set.

3.3. H$_2$O--H$_2$O. 3.3.1. Introduction. The third case study is the water dimer. We employed TYCHE to generate 100 distorted water dimer geometries from a reference optimized geometry. To provide a sense of the magnitude of the displacements generated by TYCHE, the following statistics were determined by HERMES with reference to the atom labeling in Figure 1.

The O–H bonds had an average bond length of 0.97 Å with distortion ranges for the bonds H$_2$–O$_1$, H$_3$–O$_1$, H$_5$–O$_4$, and H$_6$–O$_4$ being 0.21, 0.27, 0.24, and 0.18 Å, respectively. The two water molecules (i.e., H$_2$–O$_1$–H$_3$ and H$_5$–O$_4$–H$_6$), respectively, had average bond angles of 105.48°, over a range of 32.72°, and 104.61° over a range of 37.14°.

3.3.2. Energy Partitioning Results. Table S3 gives the magnitude of the ECE for an arbitrary example of a distorted water dimer described at the MP2/uncon-6-31++G(d,p) level. As we observed for the water monomer, the oxygen’s intra-atomic correlation energy is the largest contribution to the total ECE. The second largest contributions come from hydrogen’s intra-atomic ECE and the O–H interatomic terms. This is followed by the intramolecular H···H interactions and the intermolecular O··O interaction. Notice that there is no symmetry in either the inter- or intra-atomic energies due to the distorted nature of the geometry generated by TYCHE, although the minimum of the dimer has $C_1$ symmetry. Considering all of the geometries, the maximum absolute error in MORFI’s reconstruction of the total ECE is 4.0 kJ/mol with an average absolute error of 3.6 kJ/mol. Figure S23 presents the MORFI reconstruction of the total MP2 energy, along with the MP2/uncon-6-31++G(d,p) energy.

3.3.3. Kriging Results. A training set of 70 geometries was used for the construction of the Kriging models, which were tested for the prediction of 30 geometries. The number of features used for each model was $3N–6 = 3 \times 6 – 6 = 12$, which is notably larger than for the water monomer and H$_2$···He complex (3 features each). We present the results for all 21 IQA ECE components in Table S4. Note that the O$_1$···O$_4$ (O···O intermolecular) and the O$_1$···H$_6$ (O···H intermolecular) terms are directly involved in the hydrogen bond.$^{a}$

Table S4 shows that the quality of the models is similar to those of the water monomer and the H$_2$···He complex with the $q_r^2$ values still being very high as most values are larger than 0.98, with the only exception being H$_6$···O$_4$ interaction. For this interaction in particular the maximum absolute error is 0.5 kJ/mol. Note that all RMSE values are below 0.2 kJ/mol. The models for the oxygen atoms and the total ECE are still very strong with $q_r^2$ values being larger than 0.99. We now focus the discussion on the total ECE, as for the water monomer and H$_2$···He before. Figure 2C confirms the accuracy seen in Table S4 for the prediction of the total ECE. All predictions have a maximum absolute error less than 1.0 kJ/mol, which is below
4.0 kJ/mol (commonly taken as chemical accuracy), and 90% of the predictions have an absolute error below 0.25 kJ/mol. This level of accuracy is appropriate for application to molecular force field development. A training set of 70 geometries predicted 30 geometries with an excellent level of accuracy. Even though the results for the water monomer were slightly better, the number of features necessary to describe the water dimer is four times larger (12 versus 3), demonstrating the increased conformational freedom of the water dimer complex. It is therefore unsurprising that the accuracy is slightly reduced.

In terms of the relative energies of the predictions, Figure S24 shows the relative energy for the test set geometries calculated in the three possible ways (i.e., original MP2 energy, MORFI-reconstructed energy and FEREBUS predicted energy). It is clear that the MORFI and MP2/uncon-6-31++G(d,p) methods track each other very closely in terms of relative energies. It is also clear that the Kriging predictions track closely to the other methods’ correlation energies and, hence, reproduce the relative energies with excellent accuracy. This test clearly shows the use of such models in a molecular force field that then enables an accurate representation of the relative energetic differences between conformations, as demonstrated for all three systems. We have additionally calculated the dissociation energy of the water dimer at the same level of theory that our Kriging model was developed at, which results in a dissociation energy of 26.7 kJ/mol. This dissociation energy was obtained as the difference between the water dimer and two isolated water molecules, at the MP2/uncon-6-31++G(d,p) level. This energy partitioning methodology has recently been applied to larger water clusters.84 It may be possible, through refinements to the numerical integration grid, to further improve the results by reducing the MORFI reconstruction error. This will require further investigations.

3.3.4. RF and SVR Results. The same input features, as for the Kriging ML, were provided for the RF and SVR ML methods, but only the total ECE was predicted. The results are summarized in Table 7.

Table 7. Prediction of the Total ECE by RF and SVR for the Water Dimer85

| measure | RF | SVR |
|---------|----|-----|
| $r^2 \pm \sigma$ | $0.87 \pm 0.01$ | $0.91 \pm 0.01$ |
| RMSE $\pm \sigma$ | $4.44 \pm 0.1$ | $3.89 \pm 0.26$ |

85The RMSE is expressed in kJ/mol.

The results in Table 7 show that the total ECEs predictability by RF and SVR has slightly diminished compared to the water monomer. For both RF and SVR the $r^2$ coefficient and RMSE are good, with RF RMSE slightly over the gold standard 4 kJ/mol, which is considered chemical accuracy. Both have seen a small reduction in predictive accuracy compared to the water monomer. We note that the standard deviations of the predictions have remained fair, suggesting the methods are consistent in their predictive accuracy with respect to the training and test sets selected.

As we noted previously for the Kriging models, one would expect a reduction in the accuracy on going from the water...
monomer to the water dimer, as the system size, and thus chemical complexity, increases. This is also seen for the RF and SVR methods. The raw results, in terms of RMSE and $r^2$, show that Kriging has maintained a higher level of predictive power. We note that previous work has enabled the parameters of the Kriging method to be highly optimized to the current task and thus a direct comparison of Kriging, RF, and SVR should not be made based on these data.

**DISCUSSION**

These results provide a promising start to the exploration of Kriging predicted ECE. Previous work has shown that the task of accurately predicting electrostatic and exchange energies requires larger training data sets. The correlation energy partitioning is significantly more costly in computational time than the Coulombic and exchange energy analogues, but a much-reduced training set size can compensate for this additional cost. In other words, a smaller number of geometries are needed in order to generate a widely applicable model. Furthermore, the Kriging models are of excellent quality ($q^2 \sim 0.99$), and as the system size increases, it appears to retain this behavior for the total ECEs investigated here. However, further work is required to test that this continues to be the case for larger systems and to improve the reconstruction errors from MORFI. The fact that fewer training points are required suggests that ECE has less of a geometric dependence when contrasted with the other energy components. Hence, there is an advantage in knowing the partitioned ECE separately as the alternative approach training for a total atomic energy (i.e., including ECE, Coulomb, exchange and intra-atomic energy) would waste computational resources, as the most expensive energy contribution (ECE) requires the smallest training set. A second advantage of separate knowledge of ECE terms is access to the chemical energetics (i.e., dispersion) that QCT and hence FF.LUX offers.

The systems presented here, although small in size, represent important case studies. Thus, our study of the water monomer and dimer is the first step toward modeling the ECE effects for condensed phases. In addition, our consideration of the van der Waals complex $\text{H}_2\text{He}$ is important, as it enables us to focus on predicting van der Waals dispersion energy. The dispersion energy is a key component of the ECE with important physical consequences, such as inert gas condensation, hydrocarbon liquid, and solid structure as well as some aspects of protein conformation.

The predictive quality of the Kriging models is clearly shown by their S-curves (Figure 2). These plots display the ability of all of the Kriging models presented in this work to predict the total ECE correctly to within a few kJ/mol of the exact value over a range of test sets. All of the predictions are well within 4 kJ/mol (commonly asserted as chemical accuracy) of the MORFI reconstruction energies.

For the RF and SVR models, the predictions made are of a good quality for all systems. The majority of the models achieves a prediction with an RMSE of approximately 4 kJ/mol or less, which is regularly cited as chemical accuracy. This is another demonstration of the opportunity in using ML methods from standard packages to learn quantum chemical energies and predict them. The Kriging models have maintained an exceptionally high accuracy level, which enables us to suggest a new path toward correlation energy corrections for simulation methods. We also note that other recent work has seen IQA partitioning applied to CCSD energies, providing an opportunity to consider ML of correlation energies from other ab initio methods.

**CONCLUSION**

The total MP2/uncon-6-31++G(d,p) dynamic electron correlation energy is a readily predictable quantity for small molecular systems by the Kriging machine learning method, which provides excellent results. Kriging can predict, for all the test geometries, their total ECE within 0.5 kJ/mol of the actual MORFI reconstruction energy. Additionally, the training set sizes required to achieve this level of accuracy are considerably smaller (order of tens), compared to those required for the Coulomb and exchange energy counterparts (hundreds or even thousands). These results also represent the last proof-of-concept step for obtaining the energy components of the FFLUX force field. In its current state, FFLUX already contains well-defined Coulombic and exchange energy terms. It is also of note that the correlation energy is calculated from the same wave function as the other Coulombic and exchange energy terms. From this viewpoint, the method presents a highly coherent way to incorporate ECE, requiring no additional functions or corrections, which also means that we do not regard any interactions to be “special cases”. In a wider context, these results suggest a novel method to include ECE into molecular simulations. In the future, this work may also be applicable to generate energy corrections for other methods, such as DFT.

**ASSOCIATED CONTENT**

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.7b01157.

Figures S1–S49 and Tables S1–S20 including two examples of the topological atom, explanation of the correlated part of the two particle density matrix, Python scripts used to set up training set files for ML, data on the distortion of the systems considered, ML results for RF and SVR approaches, and data on all three systems studied with different basis sets and geometries (PDF)

**AUTHOR INFORMATION**

**Corresponding Author**

*Phone: +44 161 3064511. E-mail: pla@manchester.ac.uk.

**ORCID**

James L. McDonagh: 0000-0002-3232-6898

Paul L. A. Popelier: 0000-0001-9053-1363

**Funding**

We acknowledge the EPSRC for funding through the award of an Established Career Fellowship to P.L.A.P. (grant EP/K005472). A.F.S thanks the Brazilian government’s FAPESP for the award of postdoctoral position, numbers 2014/21241-9 and 2015/22247-3.

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

We thank Dr. S. J. Davie and Dr. N. di Pasquale for their support in generating training data files for the FEREBUS Kriging engine, Dr. S. Cardamone for assistance with TYCHE, and Dr. P. I. Maxwell for assistance with the workflow process for potential generation.
(69) Popelier, P. L. A. Comput. Phys. Commun. 1996, 93, 212.
(70) Popelier, P. L. A. In The Chemical Bond - 100 years old and getting stronger; Mingos, M., Ed.; Springer: Switzerland, 2016; p 71.
(71) Popelier, P. L. A. In Challenges and Advances in Computational Chemistry and Physics dedicated to “Applications of Topological Methods in Molecular Chemistry”; Chauvin, R., Lepetit, C., Alikhani, E., Silvi, B., Eds.; Springer: Switzerland, 2016; p 23.
(72) Vincent, M. A.; Silva, A. F.; McDonagh, J. L.; Popelier, P. L. A. Int. J. Quantum Chem. 2018, e25519.
(73) Breiman, L. Mach. Learning 2001, 45, S.
(74) Cortes, C.; Vapnik, V. Machine Learning 1995, 20, 273.
(75) Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A. J.; Vapnik, V. N. Advances in Neural Information Processing Systems 1997, 9, 155.
(76) Di Pasquale, N.; Davie, S. J.; Popelier, P. L. A. J. Chem. Theory Comput. 2016, 12, 1499.
(77) Kuhn, M. J.Stat.Software 2008, 28, I.
(78) McDonagh, J. L.; van Mourik, T.; Mitchell, J. B. O. Mol. Inf. 2015, 34, 715.
(79) McDonagh, J. L.; van Mourik, T.; Mitchell, J. B. O. Mol. Inf. 2016, 35, 538.
(80) McDonagh, J.; Palmer, D. S.; van Mourik, T.; Mitchell, J. B. O. J. Chem. Inf. Model. 2016, 56, 2162.
(81) Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. J. Chem. Inf. Model. 2008, 48, 220.
(82) Nath, N.; de Ferrari, L.; McDonagh, J. L. Github, 2016. https://github.com/Jammyzx1/R-ML-RF-SVM-PLS (accessed 9/11/16).
(83) McDonagh, J. L.; Silva, A. F.; Vincent, M. A.; Popelier, P. L. A. J. Phys. Chem. Lett. 2017, 8, 1937.
(84) Silva, A. F.; Vincent, M. A.; McDonagh, J. L.; Popelier, P. L. A. ChemPhysChem 2017, DOI: 10.1002/cphc.201700890.
(85) Chávez-Calvillo, R.; García-Revilla, M.; Francisco, E.; Martín Pendás, A.; Rocha-Rinza, T. Comput. Theor. Chem. 2015, 1053, 90.
(86) Holguin-Gallego, F. J.; Chávez-Calvillo, R.; García-Revilla, M.; Francisco, E.; Martín Pendás, A.; Rocha-Rinza, T. J. Comput. Chem. 2016, 37, 1753.