

Inclusion of more physics leads to less data: Learning the interaction energy as a function of electron deformation density with limited training data

Kaycee Low,[†] Michelle L. Coote,^{*,‡} and Ekaterina I. Izgorodina^{*,†}

*Monash Computational Chemistry Group, School of Chemistry, Monash University, Clayton,
Victoria, Australia, and Research School of Chemistry, Australian National University, Canberra,
Australian Capital Territory, Australia*

E-mail: michelle.coote@anu.edu.au; katya.pas@monash.edu

Abstract

Machine learning (ML) approaches to predicting quantum mechanical (QM) properties have made great strides towards achieving the computational chemist’s holy grail of structure-based property prediction. In contrast to direct ML methods, which encode a molecule with only structural information, in this work we show that QM descriptors improve ML predictions of dimer interaction energy, both in terms of accuracy and data efficiency, by incorporating electronic information into the descriptor. We present the electron deformation density interaction energy machine learning (EDDIE-ML) model, which predicts the interaction energy as a function of Hartree-Fock electron deformation density. We compare its performance with leading direct ML schemes and modern DFT methods for the prediction of interaction energies for dimers of varying charge type, size, and intermolecular separation. Under a low-data regime, EDDIE-ML outperforms other direct ML schemes, and is the only model readily transferable to larger, more complex systems including base pair trimers and porous cages. The underlying

physical connection between the density and interaction energy enables EDDIE-ML to reach an accuracy comparable to modern DFT functionals in fewer training data points compared to other ML methods.

1 Introduction

Through all fields of chemistry, an understanding of non-covalent interactions (NCIs) is the key to controlling molecular structure and properties.^{1–3} From a theoretical perspective, accurate quantification of NCIs can be achieved through quantum mechanical (QM) calculations, via well-established computational chemistry methods such as coupled-cluster or perturbation theory.^{4,5} *Ab initio* methods (i.e., wavefunction-methods) are capable of achieving high-accuracy results, though the cost for rigorous ‘gold-standard’ benchmark methods becomes prohibitive for systems containing more than several dozen atoms.⁶ Alternatively, density functional theory (DFT) evaluates the electronic Schrödinger equation using the electron density $\rho(\mathbf{r})$ as a starting point within the Kohn-Sham formalism.⁷ In principle, $\rho(\mathbf{r})$ contains all necessary information to characterise a chemical system, but as the exact form of the functional relationship between density and energy is unknown, various approximations must be made.⁸

^{*}To whom correspondence should be addressed

[†]Monash Computational Chemistry Group, School of Chemistry, Monash University, Clayton, Victoria, Australia

[‡]Research School of Chemistry, Australian National University, Canberra, Australian Capital Territory, Australia

Nonetheless, the increases in speed and allowed system size, and relative accuracy of existing approximations within the Kohn-Sham framework have seen DFT become a modern workhorse for computing many molecular properties.⁹

Recently, applications of machine learning (ML) techniques to atomistic systems has allowed for fast and accurate predictions of molecular and material properties,^{10,11} with model accuracy limited mainly by the data that it was trained upon. With ever-increasing amounts of high-quality computational data,^{12,13} ML models have quickly reached chemical accuracy while requiring only a fraction of the time and computational resources necessary for standard QM calculations.^{14–16} ML has been applied to molecular systems which would be too large for high-level *ab initio* theory,^{17,18} and has also been used to enhance the discovery of new compounds in chemical space.^{19,20} ML-parameterized force fields are now routinely used to perform simulations of large systems with QM accuracy at a standard molecular dynamics cost.^{21–24} It is now clear that, given the right combination of regression method and structural descriptor, ML-based predictions of certain properties (including but not limited to ground- and excited-state energy,^{11,25} forces,^{26,27} and atomization energy²⁸) can readily reach QM accuracy. ML has enabled chemists to reach closer to the ultimate holy grail of rapid, reliable property prediction based on structural information alone.

Applications of ML to the prediction of dimer interaction energies have mainly focused on small and uncharged organic dimers: IPML (intermolecular potentials from machine learning),²⁹ AP-Net (atomic-pairwise neural network),³⁰ and CLIFF (component-based machine-learned intermolecular force field)³¹ are several ML models which have been applied to predicting these types of systems. All models were shown to have errors within “chemical accuracy” (*i.e.*, 1 kcal mol^{−1}), on their respective test data. A similar error was achieved using LODE (long-distance equivariant representation),³² which also works for charged dimers, periodic systems, and tensorial properties.³³ These models use structural descriptors, requiring only atomic coordinates as the input. The description of intermolecular interactions is often treated in a physics-based manner, *e.g.*, CLIFF estimates the

electrostatic interaction energy through an expansion of the atomic multipoles parametrised with machine learning. Methods using only structural information are known as ‘direct’ ML approaches, and are clearly faster for interaction energy predictions than electronic structure calculations – provided sufficient training data are available to train the model in the first place. In this work, we take a different approach and show that use of a QM descriptor allows interaction energy predictions to be made with accuracy comparable or better than direct ML models, using a fraction of the number of training points compared to other methods. That is, providing the model with a descriptor based in physics reduces the amount training data required as input to achieve prediction accuracy comparable to QM. Given a smaller training dataset size, our electron density-based descriptor is the only model that predicts accurate shaped binding curves for test dimers. To our knowledge, our electron deformation density interaction energy machine learning (EDDIE-ML) model is the first use of the electron deformation density as a quantitative machine learning descriptor to predict interaction energy. We employ the Hartree-Fock (HF) method to generate the deformation density, and propose that ML can be used as a proxy to learn the interaction energy calculated at a correlated level of theory from a descriptor based on the uncorrelated density. This approach is closely related to the multiconfigurational wave function approach first introduced in 2014, where the exchange and correlation contribution to the total electronic energy are computed from a functional of the total density ρ and an additional “on-top” density functional.^{34,35} Recently, the authors used a machine-learned functional to correct the reference energy of a multiconfigurational wave function, demonstrating that the density and on-top density hold the necessary information to correct singlet-triplet energy splittings of multiconfigurational wave functions.³⁶

Other studies have also looked to obtain descriptors based on low-cost electronic structure calculations to improve learning efficiency and transferability of ML potentials.^{15,37–40} Like EDDIE-ML, these examples can be viewed as a midway approach between direct ML using only structural information, and more expensive *ab initio* calcu-

lations including electron correlation. Evidently, any QM-based descriptor will incur a greater cost compared to structure-based descriptors. It is thus important to examine and compare the accuracy of ML predictions when these two different descriptor types are used. Furthermore, we present a comparison of the timing and accuracy of several modern DFT functionals with our ML results, showing that the efficiency and accuracy of our ML approach is most apparent for large molecular systems.

1.1 Theory and Computational Detail

1.1.1 Overview

EDDIE-ML belongs to the category of delta machine learning (Δ -ML) models,⁴¹ which predict the energy difference between a cheap (*e.g.*, Hartree-Fock (HF)) and expensive (*e.g.*, coupled cluster theory) quantum chemical method.

Examples include DeePHF⁴⁰ and MOB-ML,⁴² which both use inputs based on HF molecular orbitals to predict the coupled-cluster electronic energy, and NeuralXC,³⁹ which uses the electron density from a low-level DFT calculation to predict more accurate molecular energies. While this approach is clearly more expensive than direct-ML models,^{11,43} Δ -ML schemes have been found to learn more efficiently with smaller data sizes.³⁷ To maintain transferability associated with use of an atomic descriptor, EDDIE-ML projects the Hartree-Fock electron deformation density onto atomic positions and uses this as input to learn the total interaction energy as calculated at a highly correlated level of theory, SRS-MP2. We will show that the addition of the HF field in EDDIE-ML is vital for reasonable accuracy in predicting the interaction energy in polar and charged systems due to polarization effects between molecules, which is already properly captured by the electron deformation density calculated with HF.

1.1.2 Gaussian process regression

The ML algorithm chosen to infer intermolecular interaction energies in this work is Gaussian process regression (GPR), which has been extensively

reviewed elsewhere.^{21,44,45} When combined with an atomic descriptor, GPR assumes that the predicted property E (here, the total interaction energy) for some molecule A outside of the training data can be decomposed into an average over its atomic (local) components. Each local environment is calculated in a basis of kernel functions, which are placed on the atoms of all chemical compounds in the training data.

$$E_{pred}^A = \sum_{a \in A} E_{local}(\mathbf{x}_A) \quad (1)$$

$$= \sum_{a \in A} \sum_B \sum_{b \in B} k(\mathbf{x}_A, \mathbf{x}_B) \alpha_B \quad (2)$$

where a and b are the atoms in molecules A and B , represented with descriptors $\mathbf{x}_A$ and $\mathbf{x}_B$, respectively. α_B is the B th regression coefficient. The kernel function k measures the similarity between descriptors $\mathbf{x}_A$ and $\mathbf{x}_B$. To evaluate the kernel function between molecules A and B , the sum of each atomic environment in the molecule is taken:

$$\mathbf{K}(A, B) = \sum_{a \in A, b \in B} k(\mathbf{x}_a, \mathbf{x}_b) \quad (3)$$

As numerous studies have shown that incorporating domain knowledge into the GPR kernel function improves learning efficiency and accuracy,^{46–51} we introduce the dimer kernel, a modified kernel function that distinguishes between atoms belonging to each molecule in the dimer. The inclusion of prior domain knowledge into chemical ML descriptors (including rotational and permutational invariance, symmetry, *etc.*) has been a crucial enabler of models reaching an accuracy comparable to high-level QM calculations.^{49,52–54}

The same applies when incorporating domain knowledge into the kernel function used in GPR, which quantifies similarity between descriptors used. The appropriate quantification of similarity between different molecular structures is critical for achieving accurate predictions.^{51,55} As such, several authors have modified the kernel function,^{46–50} for example by including a symmetric covariance function which incorporates the permutation of atomic coordinates directly into the kernel.⁴⁸

Here we introduce the dimer kernel function that

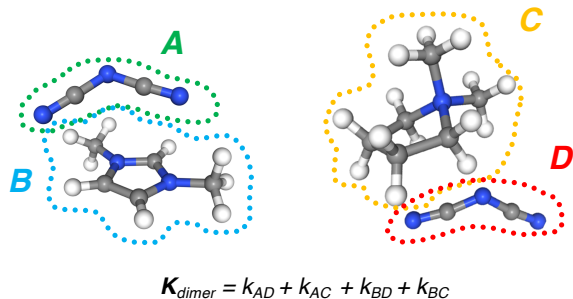


Figure 1: Graphical representation of the dimer kernel for comparison of two ionic liquids. $\mathbf{K}_{\text{dimer}}$ maximises similarity between each dimer by comparing the similarity of same-element atoms in molecule A and B with same-element atoms in molecules C and D using a pairwise kernel function k .

compares the similarity between local atomic environments of molecules within each dimer. Standard kernel functions for comparing molecular structures (*e.g.*, average and REMatch kernels⁵¹) do not distinguish between atoms belonging to different molecules as they were designed for single molecules or periodic systems. By incorporating this information into the kernel, the learning of intermolecular interaction energies should take place at a faster rate and with greater accuracy compared to the standard functions. The formulation of the dimer kernel is straightforward: for comparison of two dimers X_1 containing molecules A and B and X_2 containing molecules C and D, the similarity $\mathbf{K}(X_1, X_2)$ between the atomic environments of X_1 and X_2 is the linear combination of kernels comparing the atomic environments of A with D, A with C, B with C, and B with D. That is, $\mathbf{K}_{\text{dimer}}(X_1, X_2) = k_{AC} + k_{AD} + k_{BC} + k_{BD}$ where k is any standard kernel function which runs pairwise over the atomic environments of A, B, C, and D. The sum kernel as given in Eq. 3 is used. A graphical representation of the dimer kernel for two ionic liquid dimers is shown in Figure 1. For the pairwise similarity function k , we have used a modified polynomial kernel which includes elemental screening, ensuring that atomic environments are only compared if the atoms are of the same element:

$$k(\mathbf{x}_a, \mathbf{x}_b) = \delta_{Z_a Z_b} (\gamma \mathbf{x}_a^\top \mathbf{x}_b)^3 \quad (4)$$

where δ is the Kronecker delta function and Z_a

and Z_b refer to the nuclear charge of atoms a and b , respectively. This approach is commonly used in multicomponent systems. The elemental screening function improves learning by ensuring that only relevant pairs of atoms are compared. It also decreases the cost of evaluating the kernel since many combinations of atom pairs may be skipped.⁴⁹ This is advantageous given the extra cost associated with calculating the dimer kernel. Kernel functions were used as implemented in a modified version of `dscribe`.⁵⁶

1.1.3 Electron deformation density

The deformation density for an intermolecular dimer $\Delta\rho(\mathbf{r})$ is defined as the difference in electron density between the dimer AB and its constituent monomers, $\Delta\rho(\mathbf{r}) = \rho_{AB} - \rho_A - \rho_B$. This is an extension of the idea of density difference, which is widely used in molecular bond analysis theory,^{57,58} here applied to dimer complexes. It can be thought of as the perturbation experienced by the electron cloud of each monomer in forming an intermolecular dimer complex.

To translate a real-space electron deformation density into a form acceptable for ML, $\Delta\rho(\mathbf{r})$ was projected onto a set of atom-centred basis functions, inspired by the format of the smooth overlap of atomic positions (SOAP) descriptor created by Bartók *et al.*⁵² Angular basis functions are given by spherical harmonics $Y_l^m(\theta, \phi)$, and radial basis functions by

$$\tilde{\zeta}_n(r) = \begin{cases} \frac{1}{N} r^2 (r_0 - r)^{n+2} & \text{for } r < r_0 \\ 0 & \text{else} \end{cases} \quad (5)$$

where N is a normalization factor determined by numerical integration and r_0 is a radius cutoff around each atom. Following orthogonalization of the radial functions, the full basis is given by $\psi_{nlm}(\vec{r}) = Y_l^m(\theta, \phi) \zeta_n(r)$. For atom I of chemical element α_I at position $\mathbf{R}_I$, the descriptor coefficients c_{nlm}^I are obtained by projecting the electron deformation density $\Delta\rho(\mathbf{r})$ onto the corresponding basis functions $\psi_{nlm}^{\alpha_I}$.

$$c_{nlm}^I \equiv c_{nlm}[\Delta\rho(\mathbf{r}), \mathbf{R}_I, \alpha_I] = \int_r \Delta\rho(\mathbf{r}) \psi_{nlm}^{\alpha_I}(\mathbf{r} - \mathbf{R}_I) \quad (6)$$

The number of angular and radial basis functions

used for each chemical element was 3 and 4 respectively, with a cutoff radius of 2.5 Å.

To incorporate rotational invariance into the descriptor, the basis is aligned with a unique local coordinate system defined by the deformation density.⁵⁹ The final form of the ML input is given by these rotated descriptors,

$$c_{nlm} = \sum_{m'} (D^l(\alpha, \beta, \gamma))_{mm'}^{-1} \tilde{c}_{nlm'} \quad (7)$$

where D^l is the Wigner-D matrix for angular momentum.⁶⁰ For full details of the descriptor derivation, see Ref. 61, where $\rho(\mathbf{r})$ has been replaced with $\Delta\rho(\mathbf{r})$. All electron deformation densities were computed at the HF/cc-pVTZ level of theory in PySCF⁶² using the density fitting approximation.^{63,64} Comparison with direct ML descriptors (Coulomb matrix, SLATM) used the descriptors as available from the qml⁶⁵ package.

Data/code availability. The models, related code, and data used for training and testing are available as open-source code on GitHub: <https://github.com/lowkc/eddie-ml>.

2 Results and discussion

2.1 Neutral molecules

We initially applied EDDIE-ML to predict binding curves of small neutral dimers commonly used in interatomic ML potential studies. 60 structures in total were sourced from the S66⁶⁶ and BioFragment database¹³ (BFDf), randomly selecting 42 and 18 dimers from each database respectively. To generate binding curves, the distance between two monomers in equilibrium geometry (typically between 3 and 4 Å separation) was increased to approximately 15 Å over 14 steps, as defined by the distance between each monomer’s geometric center defined as the mean of all atomic positions. We focus on long-range interactions in this work, but note that LODE³³ was successfully trained on and used to predict interaction energies of close range separated dimers (intermolecular distance < 4.0 Å). Unrelaxed interaction energies were computed for all geometries using the SRS-MP2/cc-pVTZ level of theory to scale the

correlated energy contribution with counterpoise-corrected HF/aug-cc-pVTZ in PSI4.⁶⁷ These interaction energies were considered benchmark in this study. For neutral and ionic dimers, SRS-MP2 interaction energies were shown to have an average error of 0.3 kcal mol⁻¹ with respect to ‘gold-standard’ CCSD(T)/CBS values, with the advantage of being readily applied to larger-sized systems.^{68,69} 12 dimer binding curves (20% of the total dataset) were randomly selected to be held out from the training data and used as a test set to assess model prediction accuracy.

Mean absolute error (MAE) and root mean square error (RMSE) as defined in equations S1 and S2 are used as metrics to evaluate the performance of models. We compare our model with predictions from ML models of Grisafi *et al.* (LODE)³³ and Schriber *et al.* (CLIFF)³¹ in Table 1, with some caveats: CLIFF was trained to be a general-purpose intermolecular ML model for neutral dimers only, whereas we look at a range of charged interaction types. On the other hand, LODE is a general and flexible approach for predicting any property based on a multipolar expansion of the electrostatic energy, and predicting the dimer interaction energy is just one possible application. Here, we focus on a mixed range of intermolecular interactions which are not strongly tied to permanent electrostatics, including dimers with charged, dispersion-dominated, polar, and nonpolar interactions. Thus, the comparison of each model’s performance in this work is more for general interest to examine how each model performs and results are reflective of our selected training and test data specifically. We note that CLIFF is available as a pre-trained model, while the same data were used to train EDDIE-ML and LODE as a pre-trained model was not available for the latter.

A comparison of the timing and accuracy of EDDIE-ML predictions with several modern DFT functionals (ω B97X-D,⁷⁰ M06-2X,⁷¹ and revPBE-D3(op)^{72,73}) considered to be accurate for NCIs⁷³⁻⁷⁵ is also presented. All DFT calculations were used with a cc-pVTZ basis set⁷⁶ except for ω B97X-D, which was used with an augmented basis set aug-cc-pVTZ as this has been shown to produce superior results for non-covalent interactions.⁷⁴ Counterpoise-corrected DFT calculations were performed in PSI4⁶⁷ (revPBE) and Gaus-

sian16⁷⁷ (ω B97X-D, M06-2X).

Overall, EDDIE-ML, LODE,³³ and CLIFF³¹ all produce errors close to or below chemical accuracy (*i.e.* 1 kcal mol⁻¹) when considering interaction energy predictions for small, neutral organic dimers across the whole binding curve range. Since most of the interaction energy values at long separation distances are below 1 kcal mol⁻¹, Table 1 also shows error metrics for equilibrium distance interaction energies only. It can be seen that all the error for all three ML methods increases when only considering equilibrium distance values. EDDIE-ML is the only method which has both MAE and RMSE within chemical accuracy for data across all separation ranges, and an MAE within chemical accuracy and RMSE close to 1.0 kcal mol⁻¹ for equilibrium data points only.⁷⁸

As shown in Figure 2A, all models predict the correct asymptotic decay of interaction energy with increasing intermolecular distance, but the scale of LODE and CLIFF predictions are either over-bound (CLIFF) or both over- and under-bound (LODE). From inspection of the correlation plot (Figure 2B), EDDIE-ML predictions are more aligned with the DFT values, whereas CLIFF and LODE show significant outliers for multiple values. Considering specific examples, the SRS-MP2 equilibrium distance interaction energy for the dispersion-dominated cyclopentane...neopentane dimer is -1.6 kcal mol⁻¹, which is correctly predicted by EDDIE-ML. Both LODE and CLIFF overestimate this energy, predicting values of

Table 1: Test errors in kcal mol⁻¹ for predicted interaction energies of neutral dimers using different ML methods. N train = 48 dimers (672 points) or for EDDIE-ML and LODE, N test = 12 (168 points). ‘All’ refers to the error calculated over all points in the binding curve, ‘Equil.’ refers to error calculated for values at equilibrium distance points only.

Method	Reference	Data	MAE	RMSE
LODE	33	All	1.6	3.0
		Equil.	2.4	4.4
CLIFF	31	All	0.9	2.8
		Equil.	6.0	9.0
EDDIE-ML	This work	All	0.3	0.4
		Equil.	0.8	1.2

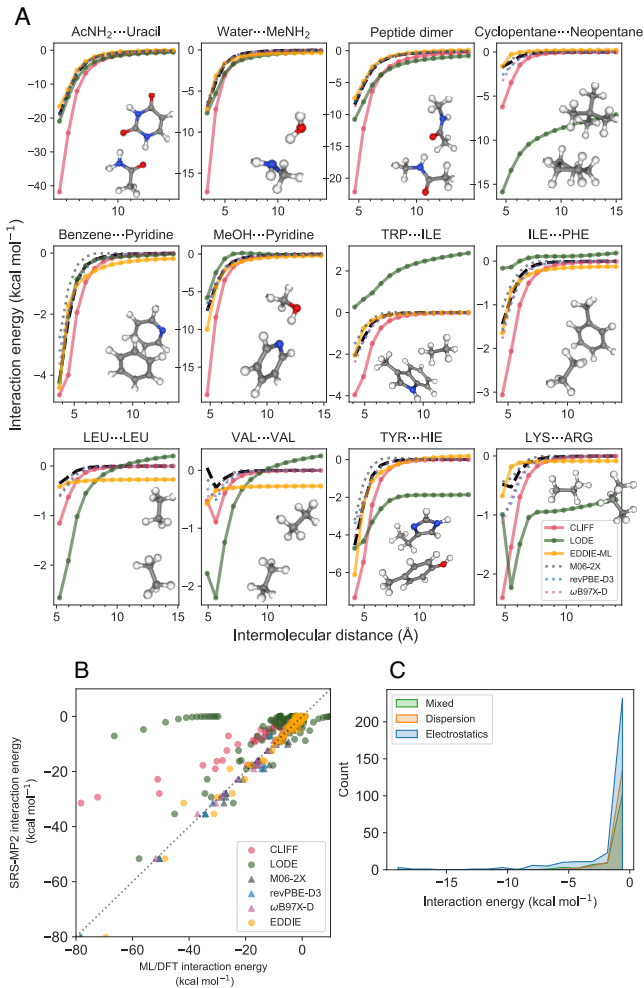


Figure 2: (A) Comparison of predicted interaction energy from CLIFF, LODE, and EDDIE-ML for neutral organic molecule dimers from the test set. Benchmark binding curves shown as dashed black lines. DFT interaction energies are shown as dotted lines. (B) Correlation plot of SRS-MP2 versus ML/DFT interaction energies. (C) Histogram of interaction energy types in S66 dimers used in training data.

-15.9 and -6.2 kcal mol⁻¹ respectively. Predictions are slightly improved for the mostly electrostatic water...MeNH₂ dimer, (reference value -7.0 kcal mol⁻¹): LODE performs (-7.7) better than CLIFF (-17.3), though both are still over-bound in comparison to EDDIE-ML (-7.2). It is possible that the lack of short-range separated dimers in the training dataset used here impacts the performance of LODE; furthermore the original paper³³ examined 2291 dimers from the BFDdb, highlighting the utility of EDDIE-ML within a low-data regime. The excellent performance of

EDDIE-ML on the test dimers demonstrates the value of incorporating QM information into a descriptor when given a small training dataset; here containing only 48 dimers or 672 points. CLIFF and LODE both achieved chemically accurate results on their respective test datasets when using training data containing approximately 100 and 1000 times more samples, respectively.^{31,32} Several authors have noted the data efficiency of Δ -ML methods,^{38,39,42} with which our results clearly agree. As shown by the learning curves in Figure S2, EDDIE-ML reaches an RMSE below chemical accuracy with just 100 training points or fewer than 10 dimer binding curves, which is an astounding result. Reducing the number of training points needed is valuable given the cost of computing reference training data points, and somewhat compensates the additional cost incurred by calculating $\Delta\rho$.

Two of the test dimers in Figure 2A that are predicted relatively poorly (though still within chemical accuracy) compared with the remaining test cases are the LEU $\cdots$ LEU and VAL $\cdots$ VAL dimers, where the EDDIE-ML predicted binding curve is over-bound at all points on the binding curve (*e.g.*, the equilibrium distance LEU $\cdots$ LEU reference intermolecular energy is -0.04 kcal mol $^{-1}$, compared with the EDDIE-ML prediction of -0.6 kcal mol $^{-1}$). This can be rationalised by examining the composition of the dataset interaction energies types, classifying each interaction as electrostatics-dominated, dispersion-dominated, or mixed-type following the division of S66 by the original authors.⁶⁶ As shown in Figure 2C, the dataset is dominated by electrostatic interactions which make up 52% of the S66 data, whereas mixed-type and dispersion-dominated interactions make up the remaining 21 and 26%, respectively. In addition, use of a non-normalized sum kernel (Equation 3) as the comparison metric across dimer environments inherently focuses the model on learning chemical environments for larger molecules, hence more accurate predictions are seen for the larger dispersion-dominated dimers such as cyclopentane $\cdots$ neopentane and ILE $\cdots$ PHE. It is likely that with the addition of more dispersion-based interactions in the training data, specially from molecules closer in size to the VAL $\cdots$ VAL dimers, the accuracy of EDDIE-

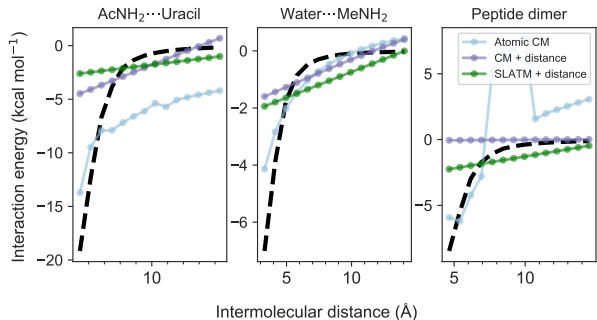


Figure 3: Comparison of predicted interaction energy on three neutral test dimers using geometric descriptors. Remaining test dimer predictions are shown in Figure S2.

ML predictions for these types of dimers would increase. This is one of the main advantages of ML models in that more (high-quality) data can be added until an arbitrary level of accuracy is achieved.

The advantage of using the EDDIE-ML descriptor encompassing density information over purely structure-based descriptors is demonstrated in Figure 3, which compares the prediction of the first three test dimers using two geometric descriptors encoding molecular structural information such as the Coulomb matrix (CM)²⁸ and SLATM.⁷⁹ The intermolecular separation between dimers is concatenated to the feature vectors of the individual monomers and used with a standard Gaussian kernel function. To encode a global representation of the dimers and evaluate the role of the dimer kernel, the atomic Coulomb matrix⁸⁰ is also tested with the same kernel as EDDIE-ML. From the predicted binding curves (remaining test dimers are shown in Figure S2), it is clear that the interaction energy cannot be captured by the simplistic representations, as most predictions show a linear decay of the interaction energy with increasing intermolecular separation. Evidently, more information than simply monomer geometries and intermolecular separation is needed to predict a reasonable value of interaction energy when using a low-data regime, as direct ML models based on geometry alone cannot reproduce the reference binding curves. Unsurprisingly, the geometric representations are outperformed by CLIFF and LODE, which both include a more sophisticated representation of the interaction energy by

decomposing the interaction into its component functional forms (CLIFF),³¹ or incorporating non-local representations into the descriptor with the same asymptotic behaviour as permanent electrostatics (LODE).^{32,33} However, only EDDIE-ML reproduces both the correct shape and magnitude of interactions for all test dimers due to its use of the electron density. These results display the utility in learning the total interaction energy as a function of a non-correlated density.

To further emphasise the notion that the non-correlated deformation density is a useful input to learn interaction energy between two molecules calculated at a correlated level of theory, we also trained EDDIE-ML models on post-HF densities using MP2, and semilocal and hybrid DFT densities. Table 2 shows the test set predictions of EDDIE-ML using HF, MP2, PBE0, and revPBE-D3 (all with cc-pVTZ basis set) densities, reported as the average over 100 random train/test splits. It can be seen that there is no significant difference between the densities used, whether they be correlated or uncorrelated. There is a slightly higher RMSE associated with use of the two PBE functionals, though the errors are still small compared to the errors using HF and MP2. Thus, the use of DFT densities, particularly with the cheap semi-local functional revPBE-D3, could be appealing for potential applications in large molecular systems or periodic systems. This will be investigated in future work. Overall, it is not surprising that there is little difference between the densities – in this work, ML is used as a tool to learn the total interaction energy as a function of the density, hence the learning process should correct for any differences in the density which may arise due to the level of theory used in its generation.

In terms of the performance compared to modern DFT functionals, Table 3 shows that EDDIE-ML has slightly higher errors than the chosen functionals. This is to be expected as GPR is an interpolative method, opposed to the exact nature of electronic structure calculations. However, the results are very similar and what is notable is the low cost of computing $\Delta\rho$: the EDDIE-ML descriptor is on average cheaper to compute than a counterpoise-corrected interaction energy calculation for all reported DFT functionals. The speed up afforded by the density-fitting approximation⁶⁴

Table 2: Average error in kcal mol⁻¹ for neutral dimers in the test set using different densities to generate $\Delta\rho$ descriptor for EDDIE-ML. Average and standard deviation taken over 100 random train/test splits.

Method	MAE	RMSE
HF	0.35 $\pm$ 0.05	0.56 $\pm$ 0.11
MP2	0.33 $\pm$ 0.06	0.55 $\pm$ 0.15
PBE0	0.34 $\pm$ 0.08	0.70 $\pm$ 0.27
revPBE-D3	0.33 $\pm$ 0.06	0.58 $\pm$ 0.18

makes generating the deformation density similar in cost to a revPBE calculation. In addition, being an ML scheme, EDDIE-ML has the advantage of being able to make predictions for multiple dimers at once.

Table 3: Average wall time (minutes) and test errors for neutral dimers in kcal/mol (with respect to SRS-MP2/cc-pVTZ energies) for EDDIE-ML and DFT functionals.

Method	Time	MAE	RMSE
EDDIE-ML	0.8 $\pm$ 0.7	0.3	0.4
revPBE-D3(op)/TZ	1.2 $\pm$ 0.8	0.1	0.2
M06-2X/TZ	2.8 $\pm$ 2.4	0.1	0.3
ω B97X-D/aTZ	4.9 $\pm$ 3.2	0.1	0.2

All calculations for neutral and charged dimer datasets were performed on 16-core Intel Xeon E5-2680v3 @ 2.50 GHz central processing units (CPUs) using 64 GB RAM.

2.1.1 Charged systems

Many ML interaction energy studies thus far (with the exception of LODE³³) have focused on predicting neutral systems at close-range separations,²⁹⁻³¹ neglecting attractive interactions between oppositely charged molecules which are ubiquitous in biological systems. The focus on equilibrium distance interactions disregards the importance of asymptotic decay effects which are essential for the treatment of large molecules. For instance, under physiological conditions, charge-attractive interactions have been shown to have a substantial contribution which can persist over more than 10 Å separation.⁸¹ Hence, we also trained a separate EDDIE-ML model on a dataset

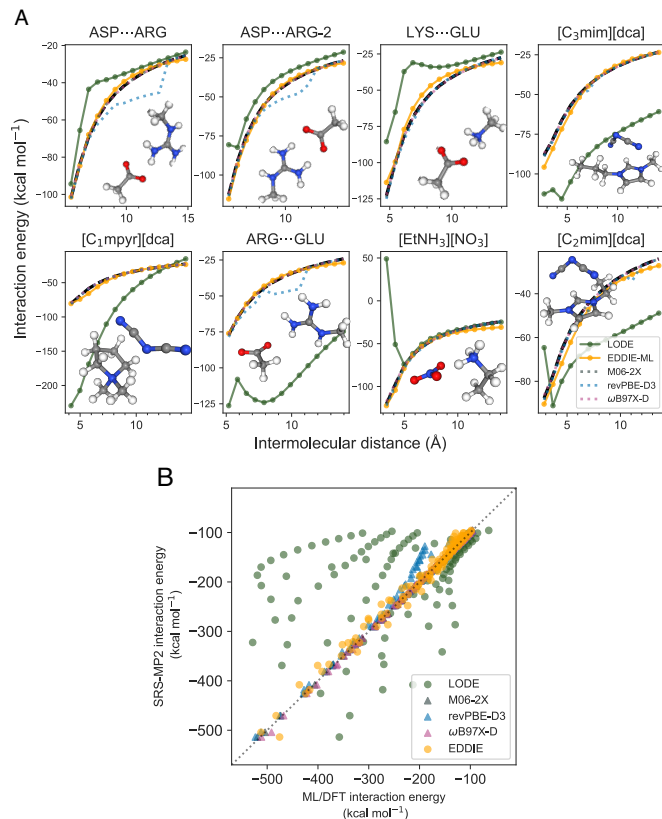


Figure 4: (A) Comparison of predicted interaction energy from LODE and EDDIE-ML for charge-charge attractive test dimers. Reference quantum mechanical binding curves are shown as dashed black lines, DFT functionals as dotted lines. (B) Correlation plot of SRS-MP2 versus ML/DFT interaction energies.

combining 22 dimer geometries from BFDdb and 24 from our ionic liquid databases^{82,83} to evaluate its ability to learn charge-attractive interactions. Binding curves and training/test data for the charged dataset were generated in the same manner as described for the neutral dataset (38 total training dimers, 8 test dimers).

Predicted binding curves from LODE and EDDIE-ML for charge-attractive test dimers are shown in Figure 4 A (CLIFF is currently applicable to neutral dimers only). Unlike for neutral systems, several LODE predictions neither have the correct shape nor magnitude, whereas EDDIE-ML predictions closely follow the reference binding curves, which can also be seen in the correlation plot (Figure 4 B). This is somewhat surprising, given that LODE was developed to predict long-distance electrostatic systems, and was

shown to reproduce B3LYP binding curves for several dimers from SSI with at least one charged monomer.³² However, in the original study the training data used to reach QM accuracy contained 600 dimer binding curves, compared with only 38 used here. In a later study, the authors also noted a degradation of the descriptor performance as the link of the target property with permanent electrostatics became less direct, which could be remedied through larger data requirements.³³ While electrostatics are the largest contributor to total interaction energy in ionic liquids, we previously noted that the dispersion component can contribute up to 20% for non-hydrogen bonded systems.⁸⁴ There seems to be some correlation between hydrogen bonded systems (such as ASP...ARG) having a lower LODE test error compared to the non-hydrogen-bonded systems (*e.g.*, [C₁mpyr][dca]) which generally have a higher error as seen in Figure 4. It is highly likely that, as discussed by the authors, LODE predictions would be improved with more training data.

The excellent predictive ability of EDDIE-ML in the low-data regime can be attributed to using ML to learn the functional relating deformation density to interaction energy. This relationship is straightforward to learn compared with the relationship between atom-density potential and interaction energy, which is the basis of LODE. Since atomic density potential can vary greatly between dimer interactions (*e.g.*, nitrogen atoms in NO₃⁻ are likely to have a different potential to nitrogen atoms in an NH₃⁺ group) larger amounts of training data are required for ML methods to learn these differences. Accounting for this difference by including the electron density accelerates this process, allowing EDDIE-ML to make accurate predictions given a small number of training examples.

Differences in electron density between dimers are pronounced in polar and charged systems, in which effects such as charge transfer and exchange may contribute significantly to the electrostatic interaction.^{1,85} These effects rapidly decay with increasing intermolecular distance. As the authors noted in their study,³² a model based on LODE features cannot describe the physics of polarisation or charge transfer as there is no explicit component to represent how the field generated by far-

away atoms depends on their neighbouring structure. As seen in Figure 4, LODE predictions of the $[\text{EtNH}_3][\text{NO}_3]$ ionic liquid are underestimated for the two closest points on the dissociation curve, but essentially identical to QM results for the remaining 12 points, where the larger intermolecular distance corresponds to a smaller contribution from exchange and charge transfer to the total interaction. The fluctuation of charge transfer and exchange effects with increasing separation cannot be encompassed in any descriptor based on atomic position or atomic density. Instead, large datasets containing thousands of reference points allow the regression model to capture differences in atomic environment and model the relationship between structure and interaction energy. As charge transfer and polarization effects occur to a lesser extent in neutral systems,¹ one can achieve reasonably accurate predictions for neutral dimers when disregarding these effects, as shown in the results for CLIFF and LODE in the previous sections. When it comes to the limited dataset sizes shown here however, only EDDIE-ML is able to predict reasonable binding curves for charged dimers due to its use of deformation density as a descriptor that encodes the underlying physics and chemistry.

When it comes to the geometric descriptors, ML predictions using SLATM and the atomic Coulomb Matrix (CM) are much more reasonable for the test dimers (see Figure 5) compared to the predictions for the neutral dimers. Models using the monomer CM perform as poorly as seen with the neutral test dimers, suggesting the the monomer geometry as represented through the CM and intermonomer distance has poor correla-

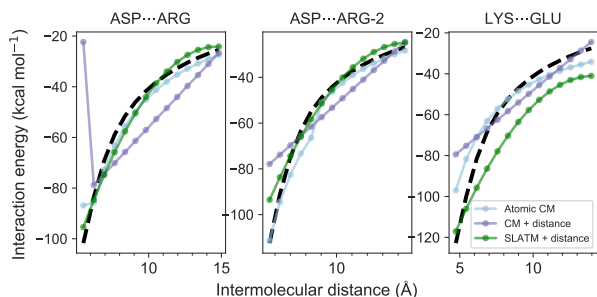


Figure 5: Comparison of predicted interaction energy on three charged test dimers using geometric descriptors. Remaining test dimer predictions are shown in Figure S2.

tion to the total interaction energy. On the other hand, the atomic CM and SLATM representations manage to mostly capture the expected asymptotic decay shape of Coulombic interactions, with some caveats. The predicted interaction energy for the SLATM model is noticeably poor (both under- and over-predicted for several test dimers) at short- and long-range separations ($\text{MAE} = 8.3 \text{ kcal mol}^{-1}$), and atomic CM predictions are non-smooth ($\text{MAE} = 14 \text{ kcal mol}^{-1}$). The improved performance of the SLATM descriptor for these systems can be attributed to its formulation which is based on the charge density distribution of the system partitioned onto atomic positions.⁸⁶ In this way, the descriptor is similar to the deformation density, except EDDIE-ML uses an SCF-converged density opposed to the coarse charge density used in SLATM. These results indicate that for Coulombic systems with a typical $1/r$ decay of interaction energy, such as electrostatics-dominated dimers, a low cost or non-converged description of the system’s electron density may (roughly) suffice as a ML input. However, the same cannot be said when the training and test dimers have more variation in the shapes of their binding curves, as shown for the neutral dimers in Figure 3, where all geometric representations fail. In non-charged systems, the interaction energy decays as roughly $1/r^6$ and varies depending on the polarisability of the molecules, meaning there is greater variation in the shape of the interaction energy decay for the neutral dimers depending to the presence of polar and apolar groups. Hence, geometric models fail to account for this diversity in binding curve shapes. Ultimately, EDDIE-ML achieves the lowest test error of all ML models for the charged test dimers ($\text{MAE} = 2.2 \text{ kcal mol}^{-1}$), clearly showing the merit to including the deformation density in improving prediction accuracy.

Comparing next EDDIE-ML with the DFT methods (Figure 4), the revPBE functional, which is of comparable cost to EDDIE-ML, shows poor performance for several of the charged systems, particularly for the $\text{ASP} \cdots \text{ARG}$ and $\text{ARG} \cdots \text{GLU}$ dimers. This is not surprising as all GGA-level functionals have well-known issues with charge localization.⁸⁷ revPBE in particular is known to have convergence issues with increasing anion-cation distance.⁸⁸ The remaining two function-

Table 4: Average wall time (minutes) and test errors for charged dimers in kcal/mol (with respect to SRS-MP2/cc-pVTZ energies) for EDDIE-ML and DFT functionals. N train = 38 dimers (532 points) for EDDIE-ML, N test = 8 (112 points). Equilibrium-only test errors are shown in brackets.

Method	Time	MAE	RMSE
EDDIE-ML	1.1 ± 0.8	2.2 (3.2)	2.9 (4.5)
revPBE-D3(op)/TZ	14.1 ± 18.1	2.0(1.1)	4.1 (1.3)
M06-2X/TZ	3.9 ± 3.2	0.2 (0.5)	0.3 (0.7)
ω B97X-D/aTZ	35.3 ± 28.2	0.3 (1.1)	0.3 (1.4)

als are within chemical accuracy of the benchmark results (Table 4), though this comes at a higher cost than EDDIE-ML. As with the neutral dimers, the performance of EDDIE-ML is thus slightly worse than DFT with similar errors to revPBE. However, unlike EDDIE-ML, several revPBE curves are non-smooth and do not show the correct asymptotic decay with increasing intermolecular distance. For the charged dimers, the difference in time required for generating $\Delta\rho(\mathbf{r})$ versus the DFT computation is significant, with the EDDIE-ML descriptor being faster to calculate than all DFT methods including revPBE due to aforementioned convergence issues. Although there is a cost associated with making GPR predictions which scales with the size of the training dataset, due to the small size of the datasets used here, the improvement in speed over DFT calculations when using EDDIE-ML is apparent. The increased cost of GPR with dataset size could be remedied by using a deep neural network as the regression method, which is the subject of future research.

2.2 Transferability to large systems

2.2.1 Dispersion-bound complexes

Moving on to larger, more practical examples, as opposed to the small molecule dimers from previous sections, we now test EDDIE-ML for complexes from the L7 dataset.^{89,90} Here we extend the application of EDDIE-ML to dimers containing 100 plus atoms, showing the transferability and size-extensivity of EDDIE-ML trained on the

entire neutral dimer dataset (60 dimers in total) and then applied to larger and more complex systems. We also present the timing of the selected DFT functionals, showing that it comes to these larger systems containing 100 plus atoms, EDDIE-ML offers a better compromise between accuracy and computational cost. For the larger dimers, calculations were performed using 128 GB RAM, except for ω B97X-D calculations, which required upwards of 600 GB RAM.

Figure 6 shows the predicted interaction energies for dimers from the L7 dataset.⁸⁹ CLIFF and LODE predictions are given in Table S4. For the L7 dimers, CLIFF predictions could only be made for C2C2PD due to convergence issues, and most LODE predictions were strongly positive (indicating a repulsive interaction), hence these results are excluded from Figure 6. SRS-MP2/cc-pVTZ values are used as the benchmark for consistency with previous sections. These values show good agreement with the established QCISD(T)/CBS benchmark⁸⁹ (see Table S4). For L7 dimers, EDDIE-ML outperforms both current ML methods, and has smaller errors than all DFT functionals except ω B97X-D. Excluding the highly extended and delocalised systems CBH and C2C2PD (which have no similar compounds in the training data due to the nature of small dimers) the average absolute error for EDDIE-ML on the remaining 5 complexes

Table 5: Average walltime in hours (and standard deviation) for DFT computed or ML calculated interaction energy for the L7 complexes. Errors (kcal/mol) are calculated with respect to SRS-MP2/cc-pVTZ values. Number of training points N is shown for ML methods.

Method	N	Time	MAE	RMSE
CLIFF ^a	7884	N/A	–	–
LODE	840	N/A	178.0	270.8
EDDIE-ML	840	1.8 ± 1.8	3.8	5.4
revPBE-D3(op)/TZ	N/A	3.5 ± 2.1	10.3	10.8
M06-2X/TZ	N/A	1.9 ± 2.1	5.9	6.5
ω B97X-D/aTZ	N/A	17.0 ± 15.3	1.8	2.4

^a CLIFF was unable to converge the calculations for 6 out of 7 complexes in the L7 dataset.

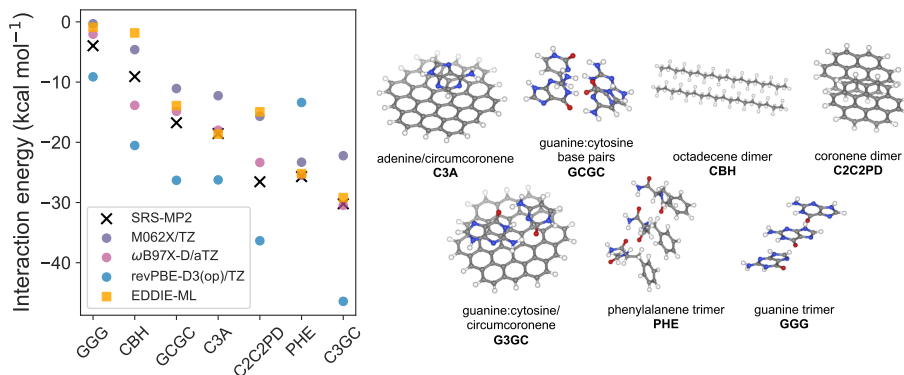


Figure 6: Comparison of EDDIE-ML predicted interaction energy with DFT-calculated interaction energies (left) for supramolecular complexes from the L7 dataset⁸⁹ (right).

is 1 kcal mol⁻¹. We expect that the larger errors for predictions of CBH and C2C2PD would be reduced by including larger dispersion-dominated dimers in the training data, as this would give the model some opportunity to learn the non-additive nature of dispersion interactions. Given the present training data, the model underpredicts the interaction energy for CBH and C2C2PD as the most similar dimers in the training set (such as cyclopentane...neopentane and VAL...VAL as shown in Figure 3) have very weak interaction energy values in the range of 1 – 3 kcal mol⁻¹. The current neutral dimer set comprises mostly electrostatic- and mixed-type dimers, and we can see that the relationship between the smaller- and larger-sized dimer interaction energies is learned successfully for these interaction types from the accuracy of predicted interaction energy for C3A, GCGC, G3GC, PHE, and GGG. The overall MAE for EDDIE-ML is 3.7 kcal mol⁻¹, which sits between the lower accuracy of M06-2X (5.9 kcal mol⁻¹) and the most accurate functional ωB97X-D (1.8 kcal mol⁻¹) (Table 5). The level of accuracy is reflected in the time taken for each DFT calculations compared to the time for the HF calculation to generate the EDDIE-ML descriptor: ωB97X-D, a functional which encapsulates both short- and long-range electron correlation effects, has an average walltime several times larger than those of all other methods, indicative of its complexity. On the other hand, the other two functionals, revPBE and M06-2X, are comparable or faster than EDDIE, though fall short in accuracy when applied to the L7 complexes.

2.2.2 Host-guest interactions

Lastly, EDDIE-ML is used to predict small molecule binding to the porous organic cage CC3, a cyclohexane functionalized imine-based cage, which has shown promise in gas separation and adsorption studies.^{91,92} Selective separation of molecules based on differential interactions within the cage has been greatly aided by computational methods which calculate relative binding energies of adsorbants.⁹³ Four small molecules (CH₄, C₂H₂, C₂H₄, CO₂) in one CC3 cage were chosen for analysis. One molecule was placed in the central cavity of CC3, giving systems with between 168 and 173 atoms. Three DFT-optimized structures were taken from the work of McKee and McKee⁹³ (Figure 7), and the structure of C₂H₄ was modified from that of C₂H₂.

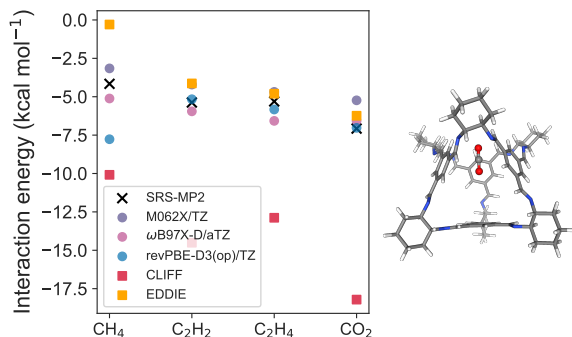


Figure 7: (left) Interaction energies for small molecules adsorbed in CC3 calculated using DFT and ML methods. Reference SRS-MP2 values are shown in black. (right) Structure of CC3 shown as licorice representation with adsorbed CO₂ molecule shown as ball and stick.

Comparing ML predicted and QM computed interaction energies for the CC3 complexes in Figure 7, it is clear that EDDIE-ML makes accurate predictions for the three larger gas molecules C_2H_2 , C_2H_4 , and CO_2 but struggles to compute the binding for the weakly interacting CH_4 . Interestingly, the three DFT methods disagree in the calculated interaction energy for CH_4 -CC3 by over 6 kcal mol^{-1} , indicating the level of difficulty in treating this weak dispersion-bound complex. DFT and EDDIE-ML predictions for the remaining three molecules are similar in magnitude; only CLIFF overbinds all four complexes though performs better than LODE, which predicted even more strongly bound values (Table S5).

The accuracy and timing of DFT and ML methods for CC3 systems is shown in Table 6. In comparison to the two direct ML schemes CLIFF and LODE, EDDIE-ML is the only scheme able to give reasonable predictions using far fewer training examples. Overall, these two large systems highlight the application of EDDIE-ML to systems which are much larger than molecules seen in the training data, where results are on par with modern DFT functionals and outperform other current ML algorithms for the prediction of interaction energy in chemical systems of any size and type.

Table 6: Average walltime in hours (and standard deviation) for DFT computed or ML calculated interaction energy for the CC3 complexes. Errors (kcal/mol) are calculated with respect to SRS-MP2/cc-pVTZ values. Number of training points N is shown for ML methods.

Method	N	Time	MAE	RMSE
CLIFF ³¹	7884	N/A	8.4	8.7
LODE ³³	840	N/A	191.8	192.1
EDDIE-ML	840	11.4 ± 1.7	1.6	2.1
revPBE-D3(op)/TZ	N/A	7.7 ± 0.5	1.1	1.8
M06-2X/TZ	N/A	8.8 ± 2.0	1.2	1.2
ω B97X-D/aTZ	N/A	65.9 ± 2.1	0.8	0.9

3 Conclusion

This work has presented EDDIE-ML, a new application of the electron deformation density as a quantitative descriptor in Δ -ML. The method predicts the interaction energy as a function of the electron density. EDDIE-ML makes accurate predictions of interaction energy for chemically diverse dimers with greater data efficiency and transferability compared to several existing direct ML models, due to the projection of deformation density onto atomic positions. The merits of including the electron density are shown by comparison with standard geometric descriptors such as the Coulomb matrix and SLATM, which cannot encapsulate the relationship between geometry and interaction energy given small training dataset sizes. The connection of the deformation density-based descriptor to the underlying physics of interaction energy means that EDDIE-ML readily predicts reasonable interaction energies for large ‘real-life’ complexes such as stacked base pairs and porous cages when trained on only small molecule dimers, whereas direct ML schemes fail in these applications. Though there is a cost associated with generating EDDIE-ML descriptors via a Hartree-Fock calculation, the speed increase over a standard DFT counterpoise-corrected interaction energy is significant when looking at small molecule dimers. Furthermore, the amount of reference data needed to train the model is reduced by over an order of magnitude compared to other ML schemes, reducing the cost associated with generating training data. The combination of DFT-level accuracy with a relatively low-cost descriptor opens future possibilities of EDDIE-ML to be trained on experimental datasets, which typically contain several hundred examples or fewer, enabling discovery of strongly interacting compounds within a given chemical domain.

Acknowledgement The authors acknowledge a generous allocation of computational resources from the Monash eResearch Centre and the National Computational Infrastructure. M. L. C. gratefully acknowledges an Australian Research Council (ARC) Laureate Fellowship (FL170100041). K.L. was supported through an Australian Government Research Training Pro-

gram scholarship.

Supporting Information Available: Supporting information available: detailed theoretical and computational methods, database generation and train-test splits, error metrics, learning curves, results using atomic Coulomb matrix descriptor and different densities, detailed results from L7 and CC3 complexes. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

Notes and References

- (1) Stone, A. *The Theory of Intermolecular Forces*; Oxford University Press, 2013.
- (2) Gromiha, M. M.; Saraboji, K.; Ahmad, S.; Ponnuswamy, M.; Suwa, M. Role of non-covalent interactions for determining the folding rate of two-state proteins. *Biophys. Chem.* **2004**, *107*, 263–272.
- (3) Woodley, S. M.; Catlow, R. Crystal structure prediction from first principles. *Nat. Mater.* **2008**, *7*, 937–946.
- (4) Bartlett, R. J.; Musiał, M. Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys.* **2007**, *79*, 291–352.
- (5) Jeziorski, B.; Moszynski, R.; Szalewicz, K. Perturbation theory approach to intermolecular potential energy surfaces of van der Waals complexes. *Chem. Rev.* **1994**, *94*, 1887–1930.
- (6) Rezac, J.; Hobza, P. Benchmark calculations of interaction energies in noncovalent complexes and their applications. *Chem. Rev.* **2016**, *116*, 5038–5071.
- (7) Kohn, W.; Sham, L. J. Self-consistent equations including exchange and correlation effects. *Physical review* **1965**, *140*, A1133.
- (8) Parr, R. G. *Horizons of quantum chemistry*; Springer, 1980; pp 5–15.
- (9) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.
- (10) von Lilienfeld, O. A.; Burke, K. Retrospective on a decade of machine learning for chemical discovery. *Nat. Commun.* **2020**, *11*, 1–4.
- (11) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192–3203.
- (12) Blum, L. C.; Raymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732.
- (13) Burns, L. A.; Faver, J. C.; Zheng, Z.; Marshall, M. S.; Smith, D. G.; Vanommeslaeghe, K.; MacKerell Jr, A. D.; Merz Jr, K. M.; Sherrill, C. D. The BioFragment Database (BFDdb): An open-data platform for computational chemistry analysis of noncovalent interactions. *J. Chem. Phys.* **2017**, *147*, 161727.
- (14) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; Von Lilienfeld, O. A. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264.
- (15) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nat. Commun.* **2020**, *11*, 1–11.
- (16) Huang, B.; von Lilienfeld, O. A. Ab Initio Machine Learning in Chemical Compound Space. *Chem. Rev.* **2021**, *121*, 10001–10036, PMID: 34387476.
- (17) McGibbon, R. T.; Pande, V. S. Learning kinetic distance metrics for Markov state models of protein conformational dynamics. *J. Chem. Theory Comput.* **2013**, *9*, 2900–2906.

- (18) Deringer, V. L.; Caro, M. A.; Csányi, G. A general-purpose machine-learning force field for bulk and nanostructured phosphorus. *Nat. Commun.* **2020**, *11*, 1–11.
- (19) Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G. Finding nature’s missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **2010**, *22*, 3762–3767.
- (20) Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **2018**, *4*, eaap7885.
- (21) Krems, R. Bayesian machine learning for quantum molecular dynamics. *Phys. Chem. Chem. Phys.* **2019**, *21*, 13392–13410.
- (22) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **2018**, *9*, 1–10.
- (23) Bartók, A. P.; Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051–1057.
- (24) Dragoni, D.; Daff, T. D.; Csányi, G.; Marzari, N. Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron. *Physical Review Materials* **2018**, *2*, 013808.
- (25) Ye, S.; Hu, W.; Li, X.; Zhang, J.; Zhong, K.; Zhang, G.; Luo, Y.; Mukamel, S.; Jiang, J. A neural network protocol for electronic excitations of N-methylacetamide. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 11612–11617.
- (26) Unke, O. T.; Muwly, M. PhysNet: a neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678–3693.
- (27) Li, Z.; Kermode, J. R.; De Vita, A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.* **2015**, *114*, 096405.
- (28) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
- (29) Bereau, T.; DiStasio Jr, R. A.; Tkatchenko, A.; Von Lilienfeld, O. A. Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *J. Chem. Phys.* **2018**, *148*, 241706.
- (30) Glick, Z. L.; Metcalf, D. P.; Koutsoukas, A.; Spronk, S. A.; Cheney, D. L.; Sherrill, C. D. AP-Net: An atomic-pairwise neural network for smooth and transferable interaction potentials. *J. Chem. Phys.* **2020**, *153*, 044112.
- (31) Schriber, J. B.; Nascimento, D. R.; Koutsoukas, A.; Spronk, S. A.; Cheney, D. L.; Sherrill, C. D. CLIFF: A component-based, machine-learned, intermolecular force field. *J. Chem. Phys.* **2021**, *154*, 184110.
- (32) Grisafi, A.; Ceriotti, M. Incorporating long-range physics in atomic-scale machine learning. *J. Chem. Phys.* **2019**, *151*, 204105.
- (33) Grisafi, A.; Nigam, J.; Ceriotti, M. Multi-scale approach for the prediction of atomic scale properties. *Chem. Sci.* **2021**, *12*, 2078–2090.
- (34) Carlson, R. K.; Truhlar, D. G.; Gagliardi, L. Multiconfiguration pair-density functional theory: a fully translated gradient approximation and its performance for transition metal dimers and the spectroscopy of Re₂Cl₈–. *Journal of chemical theory and computation* **2015**, *11*, 4077–4085.
- (35) Sharma, P.; Bao, J. J.; Truhlar, D. G.; Gagliardi, L. Multiconfiguration Pair-Density Functional Theory. *Annual Review of Physical Chemistry* **2021**, *72*, 541–564.
- (36) King, D. S.; Truhlar, D. G.; Gagliardi, L. Machine-Learned Energy Functionals for Multiconfigurational Wave Functions. *The*

- (37) Husch, T.; Sun, J.; Cheng, L.; Lee, S. J.; Miller III, T. F. Improved accuracy and transferability of molecular-orbital-based machine learning: Organics, transition-metal complexes, non-covalent interactions, and transition states. *J. Chem. Phys.* **2021**, *154*, 064108.
- (38) Margraf, J. T.; Reuter, K. Pure non-local machine-learned density functional theory for electron correlation. *Nat. Commun.* **2021**, *12*, 1–7.
- (39) Dick, S.; Fernandez-Serra, M. Machine learning accurate exchange and correlation functionals of the electronic density. *Nat. Commun.* **2020**, *11*, 1–10.
- (40) Chen, Y.; Zhang, L.; Wang, H.; E, W. Ground State Energy Functional with Hartree–Fock Efficiency and Chemical Accuracy. *J. Phys. Chem. A* **2020**, *124*, 7155–7165.
- (41) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big data meets quantum chemistry approximations: the Δ -machine learning approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.
- (42) Welborn, M.; Cheng, L.; Miller III, T. F. Transferability in machine learning for electronic structure via the molecular orbital basis. *J. Chem. Theory Comput.* **2018**, *14*, 4772–4779.
- (43) Schütt, K.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R.; Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **2019**, *10*, 1–10.
- (44) Rasmussen, C. E. *Summer school on machine learning*; 2003; pp 63–71.
- (45) Kamath, A.; Vargas-Hernández, R. A.; Krems, R. V.; Carrington Jr, T.; Manzhos, S. Neural networks vs Gaussian process regression for representing potential energy surfaces: A comparative study of fit quality and vibrational spectrum accuracy. *J. Chem. Phys.* **2018**, *148*, 241702.
- (46) Scherer, C.; Scheid, R.; Andrienko, D.; Bereau, T. Kernel-based machine learning for efficient simulations of molecular liquids. *J. Chem. Theory Comput.* **2020**, *16*, 3194–3204.
- (47) Dai, J.; Krems, R. V. Interpolation and extrapolation of global potential energy surfaces for polyatomic systems by Gaussian processes with composite kernels. *J. Chem. Theory Comput.* **2020**, *16*, 1386–1395.
- (48) Uteva, E.; Graham, R. S.; Wilkinson, R. D.; Wheatley, R. J. Interpolation of intermolecular potentials using Gaussian processes. *J. Chem. Phys.* **2017**, *147*, 161706.
- (49) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole von Lilienfeld, O. FCHL revisited: Faster and more accurate quantum machine learning. *J. Chem. Phys.* **2020**, *152*, 044107.
- (50) Ferré, G.; Haut, T.; Barros, K. Learning molecular energies using localized graph kernels. *J. Chem. Phys.* **2017**, *146*, 114107.
- (51) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769.
- (52) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.
- (53) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106.
- (54) Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Physics-inspired structural representations for molecules and materials. *Chemical Reviews* **2021**, *121*, 9759–9815.

- (55) Sadeghi, A.; Ghasemi, S. A.; Schaefer, B.; Mohr, S.; Lill, M. A.; Goedecker, S. Metrics for measuring distances in configuration spaces. *J. Chem. Phys.* **2013**, *139*, 184118.
- (56) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. Dscribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **2020**, *247*, 106949.
- (57) Bratož, S.; Daudel, R.; Roux, M.; Allavena, M. Some recent results concerning the electronic density and the force constants of small molecules. *Rev. Mod. Phys.* **1960**, *32*, 412.
- (58) Bader, R.; Henneker, W. The ionic bond. *J. Am. Chem. Soc.* **1965**, *87*, 3063–3068.
- (59) We note that use of a local coordinate system may fail for certain systems due to discontinuities in the density. As such, the option to include rotational invariance by taking the power spectrum of the descriptors (as in SOAP) is also included in our code.
- (60) Wigner, E. On unitary representations of the inhomogeneous Lorentz group. *Annals of mathematics* **1939**, 149–204.
- (61) Dick, S.; Fernandez-Serra, M. Learning from the density to correct total energy and forces in first principle simulations. *J. Chem. Phys.* **2019**, *151*, 144102.
- (62) Sun, Q.; Berkelbach, T. C.; Blunt, N. S.; Booth, G. H.; Guo, S.; Li, Z.; Liu, J.; McClain, J. D.; Sayfutyarova, E. R.; Sharma, S.; Wouters, S.; Chan, G. K.-L. PySCF: the Python-based simulations of chemistry framework. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*, e1340.
- (63) Whitten, J. L. Coulombic potential energy integrals and approximations. *J. Chem. Phys.* **1973**, *58*, 4496–4501.
- (64) Weigend, F. A fully direct RI-HF algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285–4291.
- (65) Christensen, A.; Faber, F.; Huang, B.; Bratholm, L.; Tkatchenko, A.; Muller, K.; von Lilienfeld, O. QML: A Python toolkit for quantum machine learning. *URL* <https://github.com/qmlcode/qml> **2017**,
- (66) Rezáč, J.; Riley, K. E.; Hobza, P. S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures. *J. Chem. Theory Comput.* **2011**, *7*, 2427–2438.
- (67) Turney, J. M.; Simmonett, A. C.; Parrish, R. M.; Hohenstein, E. G.; Evangelista, F. A.; Fermann, J. T.; Mintz, B. J.; Burns, L. A.; Wilke, J. J.; Abrams, M. L.; Russ, N. J.; Leininger, M. L.; Janssen, C. L.; Seidl, E. T.; Allen, W. D.; Schaefer, H. F.; King, R. A.; Valeev, E. F.; Sherrill, C. D.; Crawford, T. D. Psi4: an open-source ab initio electronic structure program. *WIREs Computational Molecular Science* **2012**, *2*, 556–565.
- (68) Tan, S. Y.; Wylie, L.; Begic, I.; Tran, D.; Izgorodina, E. I. Application of spin-ratio scaled MP2 for the prediction of intermolecular interactions in chemical systems. *Phys. Chem. Chem. Phys.* **2017**, *19*, 28936–28942.
- (69) Tan, S.; Barrera Acevedo, S.; Izgorodina, E. I. Generalized spin-ratio scaled MP2 method for accurate prediction of intermolecular interactions for neutral and ionic species. *J. Chem. Phys.* **2017**, *146*, 064108.
- (70) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.
- (71) Wang, Y.; Verma, P.; Jin, X.; Truhlar, D. G.; He, X. Revised M06 density functional for main-group and transition-metal chemistry. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 10257–10262.

- (72) Zhang, Y.; Yang, W. Comment on "Generalized gradient approximation made simple". *Phys. Rev. Lett.* **1998**, *80*, 890.
- (73) Witte, J.; Mardirossian, N.; Neaton, J. B.; Head-Gordon, M. Assessing DFT-D3 damping functions across widely used density functionals: Can we do better? *J. Chem. Theory Comput.* **2017**, *13*, 2043–2052.
- (74) Burns, L. A.; Mayagoitia, Á. V.; Sumpter, B. G.; Sherrill, C. D. Density-functional approaches to noncovalent interactions: A comparison of dispersion corrections (DFT-D), exchange-hole dipole moment (XDM) theory, and specialized functionals. *J. Chem. Phys.* **2011**, *134*, 084107.
- (75) Klimeš, J.; Michaelides, A. Perspective: Advances and challenges in treating van der Waals dispersion forces in density functional theory. *J. Chem. Phys.* **2012**, *137*, 120901.
- (76) Dunning Jr, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *The Journal of chemical physics* **1989**, *90*, 1007–1023.
- (77) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Menucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16 Revision A.03. 2016; Gaussian Inc. Wallingford CT.
- (78) Default parameters were applied for predictions using CLIFF and LODE, so it is likely that lower errors could be achieved for these models with some parameter tuning. CLIFF was originally fit to the SAPT2+(3) δ MP2/aug-cc-pVTZ interaction energies of SSI dimers, but as we found very little difference between these values and the total SRS-MP2/cc-pVTZ interaction energies for the considered dimers, we expect this to have little effect on predictions. LODE was fit to the same SRS-MP2 training data as EDDIE-ML. It is likely that LODE errors would have decreased with larger training data set size, as model accuracy was shown to increase with more benchmark data (see Refs. 32,33).
- (79) Faber, F. A.; Christensen, A. S.; Huang, B.; Von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **2018**, *148*, 241717.
- (80) Barker, J.; Bulin, J.; Hamaekers, J.; Mathias, S. *Scientific Computing and Algorithms in Industrial Simulations*; Springer, 2017; pp 25–42.
- (81) Lee, K. K.; Fitch, C. A.; García-Moreno E, B. Distance dependence and salt sensitivity of pairwise, coulombic interactions in a protein. *Protein science* **2002**, *11*, 1004–1016.
- (82) Low, K.; Kobayashi, R.; Izgorodina, E. I. The effect of descriptor choice in machine learning models for ionic liquid melting point prediction. *J. Chem. Phys.* **2020**, *153*, 104101.
- (83) Rigby, J.; Izgorodina, E. I. New SCS- and SOS-MP2 coefficients fitted to semi-

- coulombic systems. *J. Chem. Theory Comput.* **2014**, *10*, 3111–3122.
- (84) Low, K.; Tan, S.; Izgorodina, E. I. An ab initio study of the structure and energetics of hydrogen bonding in ionic liquids. *Frontiers in chemistry* **2019**, *7*, 208.
- (85) Fedorov, D. G.; Nagata, T.; Kitaura, K. Exploring chemistry with the fragment molecular orbital method. *Physical Chemistry Chemical Physics* **2012**, *14*, 7562–7577.
- (86) Huang, B.; von Lilienfeld, O. A. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nature Chemistry* **2020**, *12*, 945–951.
- (87) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Insights into current limitations of density functional theory. *Science* **2008**, *321*, 792–794.
- (88) Grimme, S.; Hujo, W.; Kirchner, B. Performance of dispersion-corrected density functional theory for the interactions in ionic liquids. *Physical Chemistry Chemical Physics* **2012**, *14*, 4875–4883.
- (89) Sedlak, R.; Janowski, T.; Pitonak, M.; Rezac, J.; Pulay, P.; Hobza, P. Accuracy of quantum chemical methods for large noncovalent complexes. *J. Chem. Theory Comput.* **2013**, *9*, 3364–3374.
- (90) Al-Hamdani, Y. S.; Tkatchenko, A. Understanding non-covalent interactions in larger molecular complexes from first principles. *J. Chem. Phys.* **2019**, *150*, 010901.
- (91) Kewley, A.; Stephenson, A.; Chen, L.; Briggs, M. E.; Hasell, T.; Cooper, A. I. Porous organic cages for gas chromatography separations. *Chem. Mater.* **2015**, *27*, 3207–3210.
- (92) Little, M. A.; Cooper, A. I. The chemistry of porous organic molecular materials. *Advanced Functional Materials* **2020**, *30*, 1909842.
- (93) McKee, N. A.; McKee, M. L. Evaluation of packing single and multiple atoms and molecules in the porous organic cage CC3-R. *Phys. Chem. Chem. Phys.* **2021**,

Graphical TOC Entry

