Linear Atomic Cluster Expansion Force Fields for Organic Molecules: beyond RMSE

Dávid Péter Kovács,[*,†] Cas van der Oord,[†] Jiri Kucera,[†] Alice E. A. Allen,[‡]
Daniel J. Cole,[¶] Christoph Ortner,[§] and Gábor Csányi[†]

†*Engineering Laboratory, University of Cambridge, Cambridge, CB2 1PZ UK*
‡*Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg City, Luxembourg*
¶*School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom*
§*Department of Mathematics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2*

E-mail: dpk25@cam.ac.uk

Abstract

We demonstrate that accurate linear force fields can be built using the Atomic Cluster Expansion (ACE) framework for molecules. Our model is built from body ordered symmetric polynomials which makes it a natural extension of traditional molecular mechanics force fields, and the large number of free parameters allows sufficient flexibility that it reaches the accuracy typical of recently proposed machine learning based approaches. We test our model on the MD17 and ISO17 data sets and also on a larger, more flexible molecule, and compare to leading machine learning models as well as refitted empirical force fields. We show that the linear body ordered ACE model has excellent transferability for properties beyond raw energy and force RMSE, both for molecular dynamics at different temperatures and for configurations very far from the training set including dihedral scans and even bond breaking.

1 Introduction

The efficient simulation of the dynamics of molecules and materials based on first principles electronic structure theory is a long standing challenge in computational chemistry and materials science. There is a trade-off between the accuracy of describing the Born-Oppenheimer potential energy surface (PES)[1] and the length and time scales that are accessible in practice. A convenient way to measure this trade-off is by considering the *total number of simulated atoms*, which can be a result of either generating a few configurations consisting of many atoms, or many configurations (e.g. a long molecular dynamics trajectory) each consisting of fewer atoms. Explicit electronic structure simulations are extremely accurate and systematically improvable. They can treat on the order of a million simulated atoms in total using either cubic scaling methods and molecular dynamics, or linear scaling algorithms on larger systems. Alternatively, in order to simulate many orders of magnitude more atoms, the PES can be parametrized in terms of the nuclear coordinates only. In this way, the electrons do not have to be treated explicitly, which simplifies the simulations considerably. These methods can routinely model a trillion (10^{12}) or more simulated atoms.

When parametrizing the PES, it is natural to decompose the total energy of the system into body ordered contributions, which can then be resummed into local atomic (or *site*) energies. The site energy of atom i is written as

$$E_i = V^{(1)}_{z_i} + \frac{1}{2}\sum_j V^{(2)}_{z_i z_j}(\mathbf{r}_{ij}) + \frac{1}{3!}\sum_{j,k} V^{(3)}_{z_i z_j z_k}(\mathbf{r}_{ij}, \mathbf{r}_{ik}) + \dots \tag{1}$$

where indices j, k run over all neighbors of atom i (either unrestricted, or within a cutoff distance r_{cut}), z_i denotes the chemical element of atom i and $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$ the relative atomic positions.

The traditional approach to the parametrization of the body ordered terms for molecular systems is to use physically motivated simple functional forms with few parameters, leading to “empirical force-fields”. These models typically require a pre-determined topology, meaning that the parameters describing the interactions of a certain atom depend on its neighbors in the bonding graph that is specified before the simulation and is not allowed to change.[2–5] The potential energy is then written as a sum of body-ordered bonded and non-bonded terms, for example:

$$E = \sum_{bonds} k_r(r - r_{eq})^2 + \sum_{angles} k_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{v_n}{2}(1 + \cos(n\phi - \gamma)) + E_{non-bonded} \tag{2}$$

where r, θ and ϕ describe the intramolecular bond lengths, angles and dihedral angles in the molecule, and $E_{non-bonded}$ contains a Lennard-Jones (LJ) term accounting for van der Waals and short-range repulsive interactions and a Coulomb term to describe the long-range electrostatics. The bonded terms can be made equivalent to the body order in eq (1) by rewriting the sum over atom-tuples into sums over sites. The advantage of the simple functional form of the bonded terms is very fast evaluation and ease of fitting due to the small number of free parameters.[2,6–8] On the other hand, this simplicity limits the achievable accuracy[9] and requires significant modification to incorporate reactivity.[10] Note that while in the most widely used force fields, the non-bonded interactions are two-body, this is not the case for polarizable force fields, such as Amoeba.[11] Moreover, the direct evaluation of terms beyond 3-body contributions is computationally expensive, in general growing exponentially with the body order, which severely limits the possibility of systematically improving force fields by adding higher body order terms.

Over the past ten years a new approach has emerged, employing machine learning (ML) methods to parametrize the PES. Instead of the body order expansion, the site energy is approximated by a neural network or a Gaussian process regressor (GPR) both of which are extremely flexible functional forms, proven to be universal approximators.[24] Due to this flexibility there is no need to specify topology or atom types beyond the identity of the chemical element, and much higher model accuracy can be achieved given an appropriate (typically rather large) training set. On the other hand, this flexibility comes also at a cost: there is no guarantee that the behavior of these ML models remains chemically sensible in regions of configuration space where there is not enough training data. Spurious local minima or even wildly wrong atomization energies are par for the course.[25] The most prominent examples of ML models are Atom Centred Symmetry Function based feed forward neural networks introduced by Behler and Parinello[26] that also includes the family of ANI force fields,[17,27] the atomic neighborhood density based GPR models like Gaussian Approximation Potentials (GAP)[14,28] and FCHL,[15] the gradient domain kernel based sGDML,[16] and message passing graph neural network based Schnet,[22] Physnet[23] and DimeNet[20] and most recently the covariant or equivariant neural network based Cormorant[21] and PaiNN.[18]

There is also a third family of methods, which expands the PES as a linear combination of body-ordered symmetric polynomial basis functions. The origins of this approach can be traced back to the work of Bowman and Braams[29,30] (Permutationally Invariant Polynomials (PIPs)), which approximated the PES

Table 1: **Comparison of different force field fitting approaches**. Molecular mechanics (e.g. AMBER,[2] CHARMM[12] and OPLS[13]), machine learning: Kernels (GAP,[14] FCHL[15] and sGDML[16]), Neural Networks (ANI,[17] PaiNN,[18] GMsNN,[19] DimeNet,[20] Cormorant,[21] Schnet[22] and Physnet[23]).

	Molecular Mechanics	**Machine learning**	**Atomic Cluster Expansion**
Functional form	fixed	flexible	polynomial
Parametrization	nonlinear	nonlinear	linear
Number of parameters	100s	10^4-10^5	10^4-10^5
Body ordered?	yes	no	yes
Topology-free?	no	yes	yes
Fit to	QM and experiment	QM only	QM only

of small molecules to extremely high accuracy, albeit with exponential scaling in the number of atoms. Introducing finite distance cutoffs reduces this scaling to linear, and the resulting atomic body-ordered permutationally invariant polynomials (aPIPs) have been shown to achieve high accuracy and better extrapolation compared to the above nonlinear machine learning based approaches in both molecular and materials systems.[25,31] The main limitation of the aPIPs approach is that the evaluation time of the site energy increases quickly with body order, making it essentially impossible to go above body-order 5 (certainly when the five atoms are of the same element). More recently, the Atomic Cluster Expansion (ACE)[32,33] (and the earlier Moment Tensor Potentials[34]) are formulations of symmetric polynomial approximations that remove the steep scaling of the evaluation of the site energy with the number of neighbors independently of body order, resulting in highly efficient interatomic potentials for materials.[35]

Table 1 compares the main features of the classical force fields, machine learning based potentials and the linear Atomic Cluster Expansion force fields. In one sense, the linear ACE constitutes a middle ground between the other two: it retains the chemically natural body order, but lifts the limitations of fixed topology and inflexible functional form embodied in eq (2).

The purpose of the present paper is to demonstrate the performance of linear ACE force fields for small organic molecules. After briefly reviewing the general ACE framework and outlining the necessary choices that go into fitting our linear models, we start with the MD17[36] and ISO17[22] benchmark data sets. We are particularly interested in going *beyond the RMSE* (or MAE) of energies and forces (the typical target of the loss function in the fit), because practically useful force fields have other desirable properties too: chemically sensible extrapolation, good description of vibrational modes, and accuracy on trajectories self-generated with the force field, just to name a few. The insufficient nature of mean error metrics has been pointed out before.[37–39] In addition to the above data sets, we also demonstrate the use of ACE on a slightly larger, significantly more flexible molecule that is more representative of the needs of medicinal chemistry applications.

The programme of tests as we outlined is designed to explore the capabilities and properties of different approaches to making force fields. We emphasize here that we are not making or testing force fields that are in and of themselves generally useful to others. That is a significant undertaking and it is to be attempted once we better understand these capabilities and properties, and are able to select which approach has the best prospects. Therefore, in addition to quoting literature results for recently published ML schemes, we refit a number of them, where the necessary software is available (sGDML, ANI and GAP in particular), so that we can show their performance on our tests. We also refit a classical empirical force field (eq (2)) to exactly the same training data to more rigorously quantify the anticipated accuracy gains of the ML and ACE approaches.

2 Methods

2.1 Atomic Cluster Expansion basis functions

The atomic cluster expansion (ACE) model[32,33] keeps the body ordering of terms defined in eq (1), but reduces the evaluation cost by eliminating the explicit summation over atom-tuples. This is accomplished by projecting the atomic neighbor density onto isometry invariant basis functions. This idea, detailed below, is referred to as the "density trick", and was introduced originally to construct the power spectrum (also known as SOAP) and bispectrum descriptors[14,40] (which are in fact equivalent to the 3- and 4-body terms in ACE, respectively, so in a sense the ACE invariants can be considered a generalization of these to arbitrary body order).

We start by defining the neighborhood density of atom i as

$$\rho_i^z(\mathbf{r}) = \sum_j \delta_{zz_j}\delta(\mathbf{r}-\mathbf{r}_{ji}); \tag{3}$$

where ρ_i^z denotes the density of atoms of element z in the neighborhood of atom i. This density is projected onto a set of *1-particle basis functions*, which we choose to be a product of a radial basis and real spherical harmonics:

$$\phi_{nlm}^{z_i z_j}(\mathbf{r}) = R_{nl}^{z_i z_j}(r) Y_l^m(\hat{\mathbf{r}}). \tag{4}$$

Here the "1-particle" refers to the single sum over neighbors, with the central atom i serving as the center of the expansion. There is considerable flexibility in the choice of the radial basis; the specifics for this work are documented at the end of this subsection. We then define the *atomic base* as the projection of the neighborhood density onto the 1-particle basis functions

$$A_{z_i,znlm} = \langle \rho_i^z | \phi_{nlm}^{z_i z}\rangle = \sum_{\substack{j \\ \text{where } z_j=z}} \phi_{nlm}^{z_i z}(\mathbf{r}_{ji}) \tag{5}$$

where the index z_i refers to the chemical element of atom i. For notational convenience, we collect the rest of the 1-particle basis indices into a multi-index,

$$(znlm) \equiv v. \tag{6}$$

From the atomic base $A_{z_i v}$, we obtain permutation-invariant basis functions, which we will call the "A-basis", by forming the products,

$$A_{z_i \mathbf{v}} = \prod_{t=1}^{\nu} A_{z_i v_t}, \quad \mathbf{v} = (v_1, \ldots, v_\nu). \tag{7}$$

The product containing ν factors gives a basis function that is the sum of terms each of which depends on the coordinates of at most ν neighbors, and we refer to it either as a ν-correlation or as a $(\nu+1)$-body basis function (the extra $+1$ comes from the central atom i). A graphical illustration of this construction is shown in fig 1 for the special case where the two factors are the same. For many (different) factors, taking products of the atomic base (left side of fig 1) takes a lot less time to evaluate than the explicit sum of all possible products (right side of fig 1). This is the key step that we referred to as the density trick.

The A-basis is not rotationally invariant. We therefore construct a fully permutation and isometry-invariant overcomplete set of functions, which we call the B-basis (technically not a basis but a spanning set), by averaging the A-basis over the three dimensional rotation group,

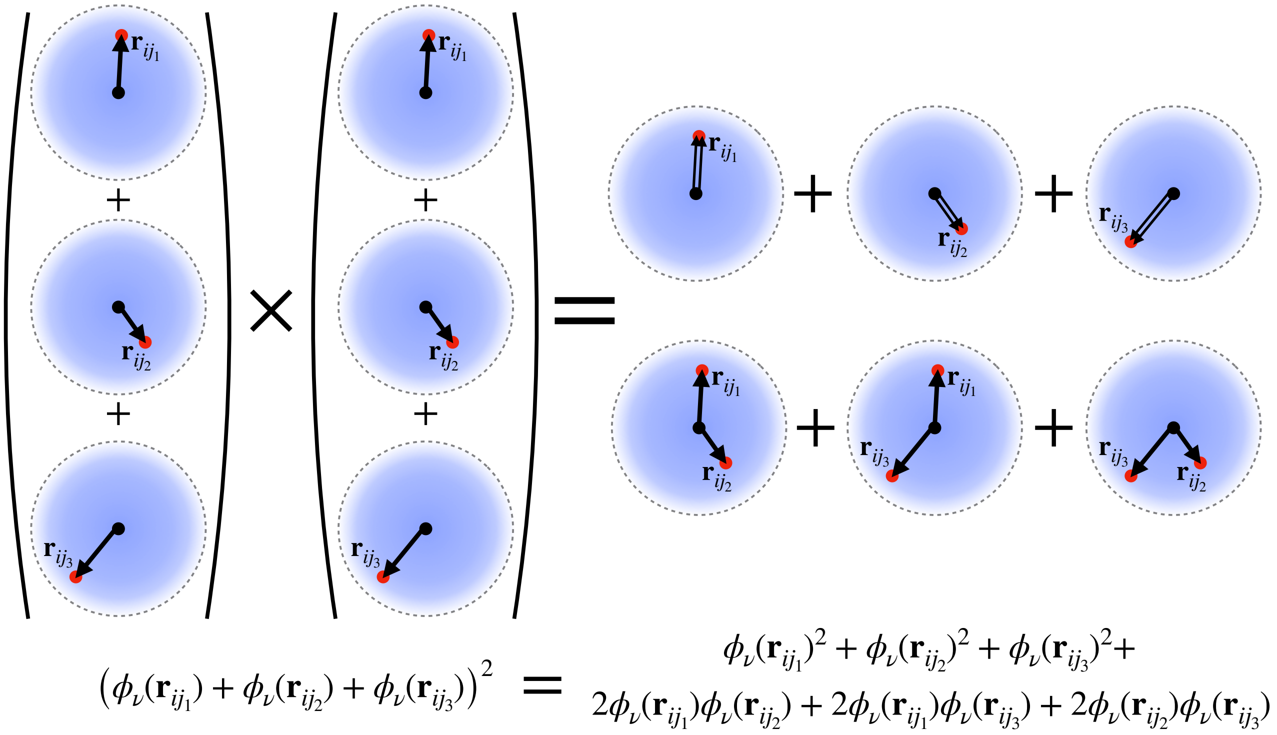
$\mathbf{r}_{ij_1}$
$\mathbf{r}_{ij_2}$
$\mathbf{r}_{ij_3}$
$\left(\phi_\nu(\mathbf{r}_{ij_1})+\phi_\nu(\mathbf{r}_{ij_2})+\phi_\nu(\mathbf{r}_{ij_3})\right)^2 = \phi_\nu(\mathbf{r}_{ij_1})^2+\phi_\nu(\mathbf{r}_{ij_2})^2+\phi_\nu(\mathbf{r}_{ij_3})^2+ 2\phi_\nu(\mathbf{r}_{ij_1})\phi_\nu(\mathbf{r}_{ij_2})+2\phi_\nu(\mathbf{r}_{ij_1})\phi_\nu(\mathbf{r}_{ij_3})+2\phi_\nu(\mathbf{r}_{ij_2})\phi_\nu(\mathbf{r}_{ij_3})$

Figure 1: **Construction of high body order invariant basis functions.** A graphical illustration showing how higher body-order basis functions can be constructed as products of the projected neighborhood density. The evaluation cost of the basis functions scales linearly with the number of neighbors rather than exponentially by doing the density projection first and than taking the products to obtain higher order basis functions. The figure (and expression) also makes explicit the occurrence of self-interaction terms in the ACE basis. They are automatically corrected through the inclusion of lower-order correlations in the basis.

O(3),

$$B_{z_i\mathbf{v}} := \int_{\hat{R}\in O(3)} \prod_{t=1}^{\nu} A_{z_i v_t}(\{\hat{R}\mathbf{r}_{ij}\})\, d\hat{R} \tag{8}$$

$$= \sum_{\mathbf{v}'} C_{\mathbf{v}\mathbf{v}'} A_{z_i\mathbf{v}'}, \tag{9}$$

where the matrix of Clebsch-Gordan coupling coefficients $C_{\mathbf{v}\mathbf{v}'}$ is extremely sparse. Many of the resulting basis functions will be linearly dependent (or even zero), but it is relatively straightforward to remove these dependencies in a pre-processing step, to arrive at an actual basis set. We refer to Dusson et al.[33] for the details of the procedure outlined up to this point.

The B-basis in eq (8) is complete in the sense that any function of the neighboring atoms that is invariant to permutations and rotations can be expanded as a linear combination of the basis functions. We therefore write the site energy of ACE as

$$E_i = \sum c_{z_i\mathbf{v}} B_{z_i\mathbf{v}} = \mathbf{c}\cdot\mathbf{B}. \tag{10}$$

The above equation makes it clear that the model is linear in its free parameters, the c coefficients. The B-basis functions are polynomials of the atomic coordinates, and in order to show that the explicit body ordering has been retained, we can switch back to using the A-basis (with the product explicitly written out),

$$E_i = \sum_{v} \tilde{c}^{(1)}_{z_i v} A_{z_i v} + \sum_{v_1 v_2}^{v_1\geq v_2} \tilde{c}^{(2)}_{z_i v_1 v_2} A_{z_i v_1} A_{z_i v_2} + \sum_{v_1 v_2 v_3}^{v_1\geq v_2\geq v_3} \tilde{c}^{(3)}_{z_i v_1 v_2 v_3} A_{z_i v_1} A_{z_i v_2} A_{z_i v_3} + \ldots \tag{11}$$

where the $\tilde{c}$ can be obtained as linear combinations of the c coefficients appearing in eq (10), using the transformation defined in eq (9).

Now the body-ordering is readily identified. Each term corresponds precisely to a sum of ν-correlations, i.e. $(\nu+1)$-body terms as in the traditional body-order expansion, eq (1). In practice, we use a recursive scheme[33] that leads to an evaluation cost that is $O(1)$ per basis function, independent of body-order. The number of basis functions does grow with body order, at a rate that has an exponent ν.

The construction outlined so far yields infinitely many polynomials $B_{z_i\mathbf{v}}$, which can be characterized by their correlation-order ν, and their (modified) polynomial degree $D = \sum_t^\nu n_t + w_Y l_t$, where n_t and l_t come from the multi-index v_t and the weight w_Y is used to trade-off the radial and angular resolution of the basis set. When it comes to defining a model in practice the expansion is truncated both in the body-order and in the maximum polynomial degree at each body-order.

2.2 Choice of radial basis

In the models in this paper we will not use much of the flexibility of the ACE framework, and simply take $R_{nl}^{z_i z_j}(r) = R_n(r)$, where

$$R_n(r) = p_n(x(r)) f_{\text{cut}}(x), \tag{12}$$

$r \mapsto x(r)$ is a one dimensional radial transformation, f_{cut} is a cutoff or envelope function and p_n are orthogonal polynomials. For the radial transform we take

$$x(r) = \frac{1}{(1 + r/r_0)^2}, \tag{13}$$

which amplifies the effect of neighbors closer to the central atom. For the cutoff function we specify both inner and outer cutoffs, $r_{\text{in}} < r_{\text{out}}$, and define

$$f_{\text{cut}}(x) = (x - x(r_{\text{in}}))^2 (x - x(r_{\text{out}}))^2, \tag{14}$$

The polynomials p_n are then defined recursively by specifying that $p_0(x) = 1, p_1(x) = x$, and the orthogonality requirement

$$\int_{x(r_{\text{in}})}^{x(r_{\text{out}})} R_n(r(x)) R_{n'}(r(x))\, x^2 dx = \delta_{nn'}, \tag{15}$$

where we have used the *inverse* of the radial transform, $x \mapsto r(x)$. Eq (15) implies that the radial basis R_n and not the polynomials p_n are orthonormal in x-coordinates.

The introduction of an inner cut-off is necessary to prevent wildly oscillating behaviour in high energy regions of configuration space where pairs of atoms are very close to one another and little or no training data is available. Alternatively, one could introduce such training data, but that would unnecessarily complicate the construction of training data sets and this inner cutoff mechanism is sufficient. To ensure short range repulsion we augment the large multi-body ACE basis by a small auxiliary basis set, consisting only of low-polynomial-degree pair interaction (two-body) functions. The construction is exactly the same as before, but we change the cut-off function to

$$f_{\text{cut}}^{\text{rep}} = (x - x(r_{\text{out}}))^2. \tag{16}$$

2.3 Basis Selection

Before we can parametrize the ACE force field we need to select a specific finite basis set chosen from the complete ACE basis constructed in the previous section. There are three approximation parameters: the cutoff radius ($r_{\text{cut}} = r_{\text{out}}$), the maximum correlation order $\nu^{\max}$, and the maximum polynomial degrees $D_\nu^{\max}$ corresponding to order ν basis functions. We have already specified the cut-off radius in the definition of the radial basis in eq (12). The basis is then chosen as (a linearly independent subset of) all possible basis functions $B_{i\mathbf{v}}$ with correlation order at most $\nu^{\max}$ and polynomial degree at most $D_\nu^{\max}$.

In all models for molecules with three or fewer distinct elements we take $\nu^{\max} = 4$, which corresponds to a general 5-body potential. In models for molecules with four or more distinct elements we reduce this to $\nu^{\max} = 3$ (4-body potential). The weight w_Y specifies the relative importance of the radial and angular basis components; here we choose $w_Y = 2$. The maximum polynomial degrees $D_\nu^{\max}$ can be adjusted to balance the size of the basis set against fit accuracy and evaluation time; the precise parameters we choose for each molecule are given in Table S1. The basis truncation we specified here is just one, rather simple, way to obtain a finite basis. There may very well be more sophisticated methods to choose an optimal subset of the complete basis.

2.4 Parametrization of the linear ACE potentials

We define the total energy of a linear ACE model with parameters $\mathbf{c}$ corresponding to a spatial configuration of atoms (denoted by X, e.g. a molecule in a particular configuration) as the sum of the site energies

$$E(\mathbf{c}; X) = \sum_{i \in X} E_i(\mathbf{c}) \tag{17}$$

where E_i is a site energy defined in eq (10). Optimal parameters are obtained by minimizing the loss function

$$L(\mathbf{c}) := \sum_X \Big(w_X^E \big| E(\mathbf{c}; X) - E_{\mathrm{QM}}(X) \big|^2 + w_X^F \big| F(\mathbf{c}; X) - F_{\mathrm{QM}}(X) \big|^2 \Big), \tag{18}$$

where the E_{QM} and F_{QM} are energies and forces, respectively, in the training data, obtained from electronic structure calculations. The sum is taken over all configurations in the training set, and w_X^E, w_X^F are weights specifying the relative importance of energies and forces. Since the model energy and force are both linear in the free parameters, the loss can be written in a linear least squares form,

$$L(\mathbf{c}) := \|\Psi \mathbf{c} - \mathbf{t}\|^2, \tag{19}$$

where the vector $\mathbf{t}$ contains the QM energy and force observations, and the design matrix Ψ contains the values and gradients of the basis evaluated at the training geometries. Ψ has a number of rows equal to the total number of observations (energies and force components) in the training set, and a number of columns equal to the total number of basis functions.

The least squares problem has to be regularized, especially when the basis contains high degree polynomials.[31] One option is to apply Tychonov regularization, where the loss function is modified as

$$\|\Psi \mathbf{c} - \mathbf{t}\|^2 + \lambda \|\Gamma \mathbf{c}\|^2. \tag{20}$$

This is widely used to regularize linear regression, often by taking Γ as just the identity matrix, or alternatively in the case of kernel ridge regression (and Gaussian process regression) as the square root of the kernel matrix.[41] In the present case, we use a diagonal Γ with entries corresponding to a rough estimate for the p-th derivative of the basis functions,

$$\|\nabla^p B_{z\mathbf{v}}\|_2 \approx \sum_{t=1}^{\mathrm{len}(\mathbf{v})} (n_t)^p + (l_t)^p, \tag{21}$$

where n_t and l_t are part of the elements of the multi-index vector $\mathbf{v}$ (cf. eq (6)). This scales down high degree basis functions, encouraging a smooth potential, which is crucial for extrapolation, and is loosely analogous to the smooth Gaussian prior of GPR. The actual solutions are then found using the standard iterative LSQR solver,[42] for the details see the SI.

In the other approach we used for solving the least squares problem the same Γ matrix is introduced, but without a Tychonov term,

$$L(\mathbf{c}) := \|\big(\Psi\Gamma^{-1}\big)\big(\Gamma\mathbf{c}\big) - \mathbf{t}\|^2, \tag{22}$$

and the solution is found using the rank revealing QR factorisation[43] (RRQR), in which we perform a QR factorization of the scaled design matrix $\Psi\Gamma^{-1}$, and truncate the small singular values below some tolerance parameter λ. For more details of the exact implementation see Refs. 25,43. We found that when the linear system is not underdetermined, RRQR gave somewhat better solutions than LSQR. All parameters of the optimization (w_X^E, w_X^F, p, λ) are given in the SI.

The last modelling choice that needs to be made is the 1-body term, that is the energies of the isolated atoms of each element in our model. One can use the energy of the isolated atoms evaluated with the reference electronic structure method, which ensures the correct behavior of the model in the dissociation limit. In other words, that the force field is modelling the *binding energy* of the atoms. An alternative approach, often used in the ML fitting of molecular energies, is to take the average energy of the training set, divided by the number of atoms in the molecule, and assign the result to each element. In this case, the fitted model has zero mean energy. This usually improves the fit accuracy slightly, by reducing the variance of the function that we need to fit, in case the data spans a narrow energy range around its average, e.g. because it came from samples of moderate temperature molecular dynamics.

A third option is to not use any reference potential energies for the fit, but only forces. Once the coefficients are determined, the potential can be shifted by a constant energy chosen to minimize the training set energy error. In the current work, we evaluated all three strategies for ACE and found that using the isolated atom energies for the 1-body term gives slightly higher RMS errors, but leads to far superior extrapolation. The other two strategies (using the average energy for the 1-body term, and fitting only to forces) result in similar somewhat lower test set errors, but inferior physical extrapolation properties.

As mentioned in the introduction, we view tests on data sets such as MD17 and ISO17 as *proxies*: the models thus created are not useful for any scientific purpose. The promise of ML force fields is greatest when the intention is to describe a very wide variety of compounds and conformations, perhaps including chemical reactions. With this in mind, the most natural choice for the 1-body term is to choose it to match the energy of the isolated atom in vacuum. This choice is independent of any particular data set, and the apparent advantages of the other choices in terms of lower errors are expected to diminish in the limit of a large and wide ranging data set.

3 Results

3.1 MD17

The original MD17 benchmark data set consists of configurations of 10 small organic molecules in vacuum sampled from density functional theory (DFT) molecular dynamics simulations at 500 K.[36] It has recently been recognized, that some of the calculations in the original data set did not properly converge, in particular, many of the forces are noisy. A subset of the full data set was recomputed with very tight SCF convergence settings and is called the rMD17 (revised MD17) data set.[44] We have used this new version of the data set and the five train-test splits as reported in Ref. 44. These revised training sets consist of 1,000 configurations to avoid the problem of correlated training and test sets: when more than 1,000 configurations are used from the full published trajectory, some of the test set configurations will necessarily fall between two neighboring training set data points that are separated by a much smaller time difference than the decorrelation time of the trajectory, resulting in an underestimation of the generalization error.[44]

Table 2 shows the Mean Absolute Error (MAE) of the different force field models trained on 1,000 configurations. The models on the left were trained (by us, except for FCHL) using the exact train-test splits of rMD17, whereas the models on the right are from the literature and were trained on the original MD17 data set using different train-test splits. The precise details of the fitting procedures and parameters can be found in the SI.

Of the descriptor based models, sGDML, FCHL and our linear ACE have the lowest MAE for some molecules. Overall, based on the per atom energy and force the ACE model achieves the lowest errors averaged across the entire data set, improving on the state of the art for several individual molecules as well. It is interesting to note, that of the neural network models, the PaiNN equivariant neural network achieves very low force errors, but its energy errors are almost three times higher compared to ACE and FCHL. In our view, the energy er-

Table 2: **Mean Absolute Error of MD17 molecules.** Energy (meV) and force (meV/Å) errors of different models trained on 1,000 samples. The models on the left were trained and tested using the same train-test splits of rMD17, whereas models on the right use MD17. The best model for each molecule (on the left and the right) are shown in bold font. The average energy MAE is calculated per atom. For reference 43 meV = 1 kcal / mol.

		ACE	**sGDML**	**FCHL**[44]	**GAP**	**ANI**	**FF**	**PaiNN**[18]	**GMsNN**[19]	**DimeNet**[20]
Aspirin	E	**6.1**	7.2	6.2	17.7	16.6	93.2	6.9	16.5	8.8
	F	17.9	31.8	20.9	44.9	40.6	260	**16.1**	29.9	21.6
Azobenzene	E	3.6	4.3	**2.8**	8.5	15.9	112	-	-	-
	F	10.9	19.2	**10.8**	24.5	35.4	246	-	-	-
Benzene	E	**0.04**	0.06	0.35	0.75	3.3	13.2	-	3.5	3.4
	F	**0.5**	0.8	2.6	6.0	10.0	105	-	9.1	8.1
Ethanol	E	1.2	2.4	**0.9**	3.5	2.5	42.1	2.7	4.3	2.8
	F	7.3	16.0	**6.2**	18.1	13.4	208	10.0	14.3	10.0
Malonaldehyde	E	1.7	3.1	**1.5**	4.8	4.6	45.9	3.9	5.2	4.5
	F	11.1	18.8	**10.3**	26.4	24.5	234	13.8	19.5	16.6
Naphthalene	E	0.9	**0.8**	1.2	3.8	11.3	65.3	5.1	7.4	5.3
	F	5.1	5.4	6.5	16.5	29.2	292	**3.6**	15.6	9.3
Paracetamol	E	4.0	5.0	**2.9**	8.5	11.5	93.9	-	-	-
	F	12.7	23.3	**12.3**	28.9	30.4	248	-	-	-
Salicylic acid	E	**1.8**	2.1	**1.8**	5.6	9.2	68.4	4.9	8.2	5.8
	F	9.3	12.8	9.5	24.7	29.7	263	**9.1**	21.2	16.2
Toluene	E	1.1	**1.0**	1.7	4.0	7.7	36.9	4.2	6.5	4.4
	F	6.5	6.3	8.8	17.8	24.3	183	**4.4**	14.7	9.4
Uracil	E	1.1	1.4	**0.6**	3.0	5.1	43.3	4.5	5.2	5.0
	F	6.6	10.4	**4.2**	17.6	21.4	233	6.1	14.3	13.1
Average MAE	E*	**0.12**	0.16	**0.12**	0.37	0.50	3.9	**0.33**	0.49	0.36
	F	**8.0**	12.8	8.6	22.5	24.1	227	**8.0**	17.3	13.0

Table 3: **Mean Absolute Error of neural network models on the original MD17.** MAE of energy (E, meV) and force (F, meV/Å) predictions of models recently published that were trained on 50,000 geometries of the original MD17 data set.

		Cormorant[21]	**Schnet**[22]	**Physnet**[23]	**GMdNN**[19]
Aspirin	E	**4.2**	5.2	5.2	5.6
	F	-	14.3	**1.7**	5.2
Benzene	E	**1.0**	3.0	3.0	3.0
	F	-	7.4	**6.1**	**6.1**
Ethanol	E	**1.3**	2.2	2.2	2.2
	F	-	2.2	**0.9**	1.7
Malonaldehyde	E	**1.8**	3.5	3.0	3.0
	F	-	3.5	**1.3**	2.2
Naphthalene	E	**1.3**	4.8	5.2	4.8
	F	-	4.8	**1.3**	3.5
Salicylic acid	E	**2.9**	4.3	4.8	4.8
	F	-	8.2	**1.3**	3.5
Toluene	E	**1.5**	3.9	4.3	3.9
	F	-	3.9	**1.3**	2.6
Uracil	E	**1.0**	4.3	4.3	4.3
	F	-	4.8	**1.3**	1.7
Average MAE	E*	**0.13**	0.29	0.29	0.28
	F	-	6.1	**1.9**	3.3

ror is important because even though a molecular dynamics trajectory is only affected directly by the forces, the stationary probability distribution that MD is used to sample is *solely a function of the energy* through the Boltzmann weight, and so errors in predicted energy translate into errors of the stationary distribution and thus of all equilibrium observables.

The ANI model in this table refers to our reparametrization of the ANI architecture with pre-training, that is the neural network weights were initialized from those of the published ANI-2x model.[17] This was crucial for achieving the errors shown. When the weights were initialized randomly, the errors are higher by factor of 2 (Table S2). The GAP model, using SOAP features to describe the atomic geometry (which are similar to ANI's features), achieves similar errors to the ANI model with pre-training. The fact that ANI is only competitive with GAP if it is pre-trained can be rationalized by the relative sample efficiency of kernel models compared to neural networks. The FCHL kernel models also use 2- and 3-body correlations as features, but they have been more carefully optimized for molecular systems and hence are able to achieve very low errors.[15]

The classical force field (FF) refers to a reparametrization of the GAFF functional form[2,45] using the ForceBalance program[6,45] and the rMD17 training set. This model gives at least an order of magnitude higher errors compared to the ML force fields. This is not a huge surprise, but is nevertheless a quantitative characterization of the limitations of the fixed functional form for a situation in which the empirical force fields are designed to do well.

For completeness, in Table 3 we show the MAEs of the neural network models reported in the literature that were trained on 50,000 structures from the original MD17 trajectories. The test set errors of these models are probably underestimating the true generalization error, because the large training set contains configurations that are correlated with the test set, as discussed above.[44] It is still interesting to note that the Cormorant equivariant neural network[21] achieves very low energy errors compared to PaiNN, even though it was trained on energy labels only, but the force errors for this model were not reported. On the other hand, the PhysNet[23] graph neural network achieves remarkably low force errors compared to the other models. But similarly to the other equivariant graph neural network models, this comes at the expense of having close to 3 times larger energy errors compared to ACE and FCHL.

3.1.1 Learning curves

The first property to consider beyond the raw energy and force errors is the learning curve, showing how a model's performance improves with additional training data. For kernel models such as FCHL and sGDML, the "kernel basis" grows precisely together with the training data, which is why these methods are universal approximators. Subject to the radial cutoff, the infinite set of Atomic Cluster Expansion basis functions forms a complete basis for invariant functions, so in principle they can also be used to approximate the potential energy surface to arbitrary accuracy.[33] In this case however, the size of the training set and the size of the basis are decoupled. One advantage is that the evaluation cost is independent of training set size, but we have to choose a finite basis set to work with by selecting a maximum body order and the truncation of the 1-particle basis. In order to motivate our choice, we show in fig 2 the force accuracy of ACE as a function of basis set size and the corresponding evaluation time, trained on 1,000 azobenzene configurations (the largest molecule in MD17).

The timings were obtained using a 2.3 GHz Intel Xeon Gold 5218 CPU. For context, we show the accuracy and evaluation time of the other ML models we trained, each called in their native environment: ACE in julia, GAP via the fortran executable, and sGDML and ANI directly from their respective Python packages. (Note that in the case of ANI some speed up could be achieved by using a GPU, though our results are in agreement with the timings reported in the original ANI paper[17]). The solid part of the ACE curve corresponds to 4-body potentials ($\nu = 3$) and we varied only the polynomial degrees, whereas for the last point (dashed), we increased the body order to 5, because the 4-body part of the curve showed saturating accuracy. Increasing the body order further is likely to bring the error down even more, however, the cost of evaluation would also grow unacceptably if all basis functions for the given body and polynomial degree are retained. In the future, effective sparsification strategies need to be developed that would allow the inclusion of some high body order basis functions without the concomitant very large increase of the overall basis set size. For the purposes of the present paper, for each molecule in MD17 we selected a basis set size such that the evaluation cost was roughly comparable with the other ML models. (Note however that in a real ML force field application, one might very well choose a much smaller basis, e.g. 10K, to take advantage of the sub-millisecond evaluation times.)

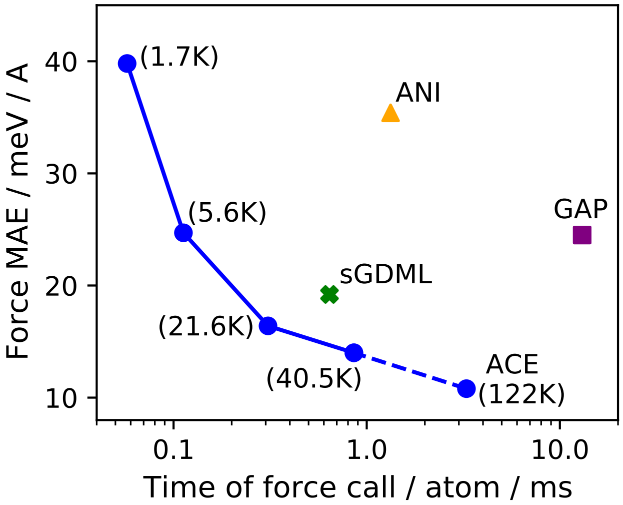

Figure 2: **Force evaluation times.** The timing of force calls per atom for the azobenzene molecule. In the case of ACE the number of basis functions is shown in parentheses.The classical force field has a timing of about 1 μs, which would not fit on this scale.

In fig 3 we show the learning curves for linear ACE and sGDML (the best models we trained from Table 2) and compare to the literature results of FCHL.[44] The low body order linear ACE is equal or better than the other many-body kernel models in the low data limit, but

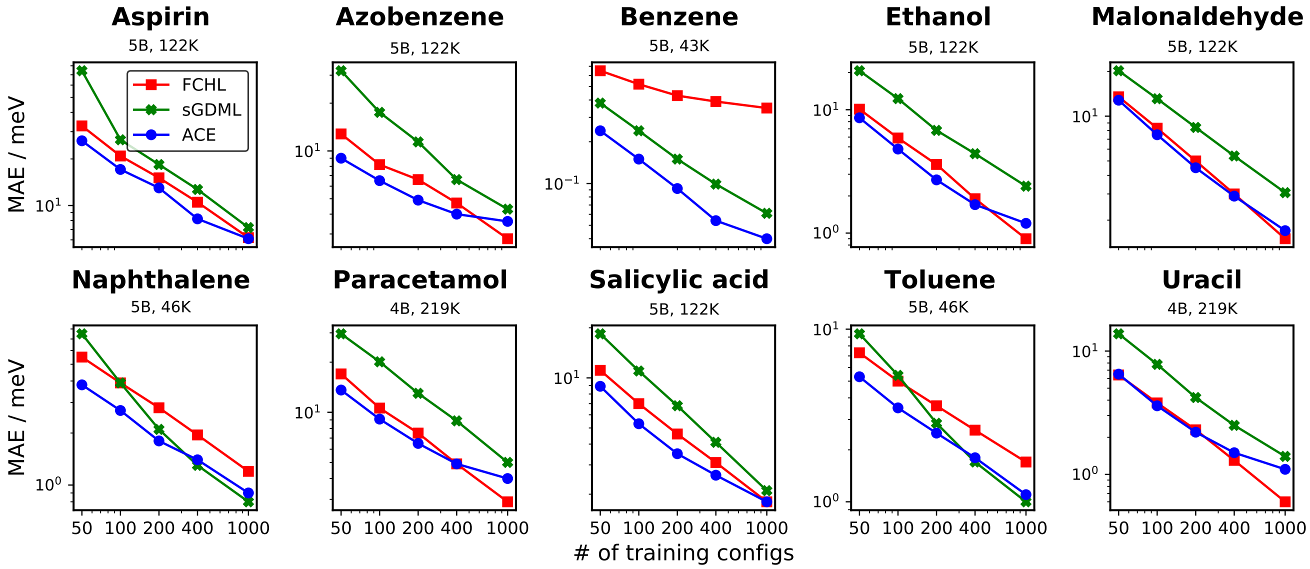

Figure 3: **Energy learning curves.** The learning curves of the best performing models on the rMD17 data set. The body order and basis set size for the ACE models are given under the title of each panel

with additional training data the kernel models overtake ACE in several cases. The latter also saturates, showing the limitations of the relatively low body order model. The learning curves for the forces are given in fig S1, and show a broadly similar trend, with less pronounced saturation for ACE.

3.1.2 Normal mode analysis

The normal modes and their corresponding vibrational frequencies characterize the potential energy surface near equilibrium. This is interesting in the context of the MD17 models because their training set contains geometries sampled at 500 K which means they are, in general, far from the equilibrium geometry. The ability of the models to describe the minima of the PES, even if it is not in the training set, is particularly important when considering larger systems with potentially many local minima, where finding all the different local minima at the target level of theory can be infeasible.

To test how well the different models infer the normal modes we took the DFT optimized geometry of each of the 10 molecules and re-relaxed them with the force field models. At the force field minima we carried out a vibrational analysis to find the normal modes and their corresponding vibrational frequencies.

Fig 4 shows the errors in the predicted normal mode vibrational frequencies for each of the 10 MD17 molecules. The ACE model achieves the lowest error for all 10 molecules, surprisingly even for those for which sGDML has lower errors based on the 500 K MD test set of Table 2. For example, for toluene sGDML has both lower energy and force errors, but at the same time the ACE model has significantly lower errors in predicting the vibrational frequencies, achieving a MAE of 1.0 cm^{-1} compared to sGDML with an error of 1.4 cm^{-1}. Observing the individual molecules in Fig 4 it is notable that the ACE model has the lowest fluctuation in the errors of the normal modes, achieving nearly uniform accuracy across the entire spectrum. The case of benzene also shows the limitations of characterizing the models by the force MAE alone. The linear ACE model has only slightly lower force MAE than sGDML (0.5 meV/Å compared to 0.8 meV/Å) but the normal mode frequency prediction is more than 3 times more accurate: 0.2 cm^{-1} compared to 0.7 cm^{-1}. The linear ACE model has very low errors for all normal modes, whereas sGDML has much higher errors for the high frequency modes.

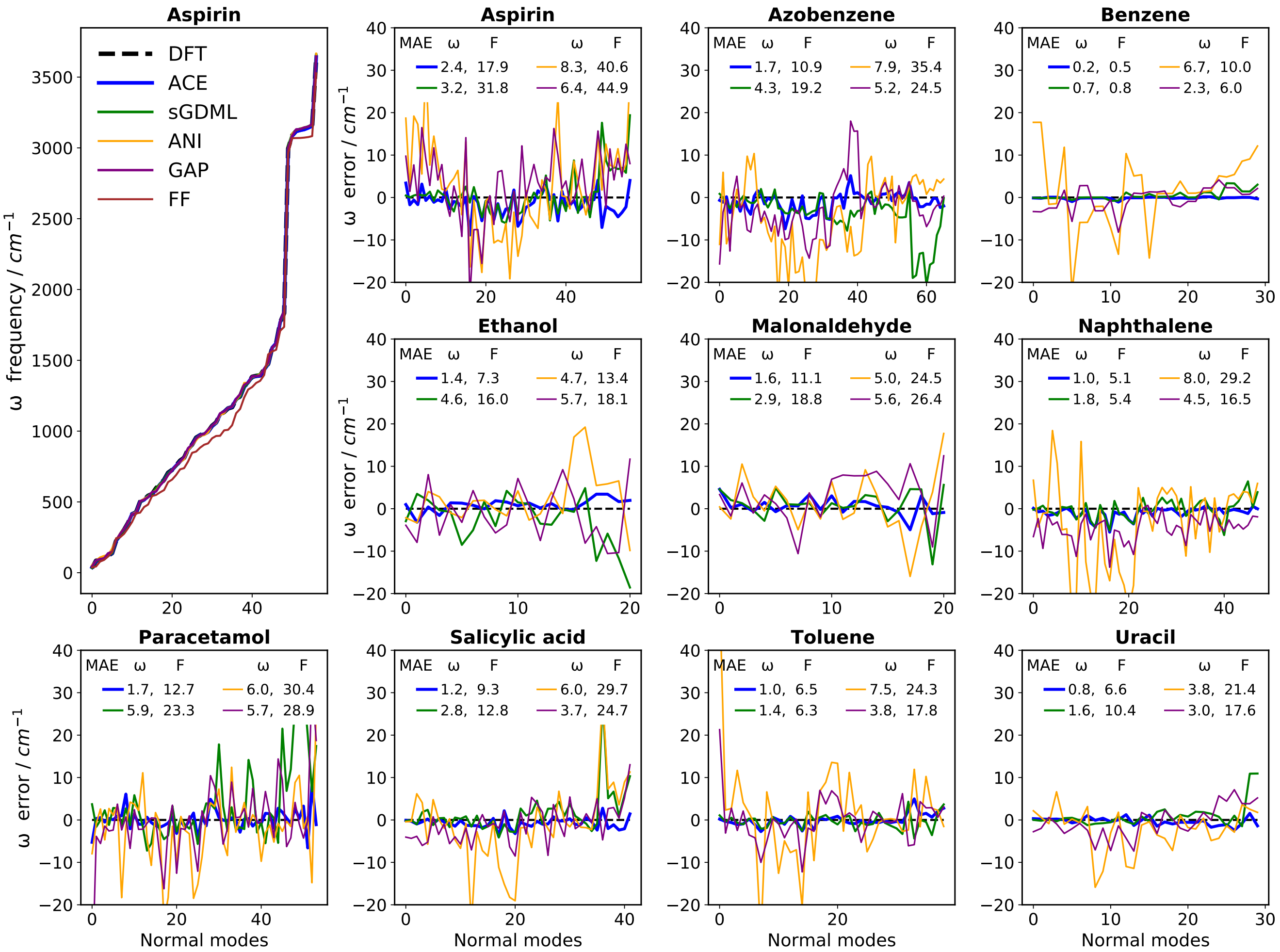

Figure 4: **Normal mode frequency test.** The frequency error of the normal modes of each of the MD17 molecules. The legend shows the frequency (ω) MAE (in cm^{-1}) and also the force (F) MAE (in meV/Å) from Table 2 for each model

Similarly, in the case of aspirin, even though the ANI model has lower MAE on the test set both for energies and forces than the GAP model, its vibrational frequency error is significantly larger than those of GAP (8.3 cm^{-1} compared to 6.4 cm^{-1}). We also compared the models to the accuracy of a classical force field. The normal mode frequency errors of the empirical FF are about 10 times higher than the errors of the ML force fields. These errors do not fit on the scale of Fig 4 but are reported in Fig S2.

3.1.3 Extrapolation in temperature

When building a new force field for a molecule, beyond high accuracy, we also need robustness, by which we mean that there should not be areas of accessible configuration space where the model predictions are unphysical or nonsensical. Sometimes called "holes" in the potential energy surface, these can be remedied by regularization[31] or by iterative fitting[37] and additional data.[46] In the context of the MD17 benchmark, with its fixed training set, we test the robustness of the models we fitted by running short molecular dynamics (MD) simulations with each model. Separate MD simulations were run at several temperatures between 350 and 950 K using a Langevin thermostat and a timestep of 0.3 fs. (Higher temperatures were not considered because most organic molecules undergo thermal decomposition at temperatures above 1000 K.) Five independent MD runs were initialized starting from different configurations. After equilibrating for 500 steps, 10 samples were taken 200 timesteps

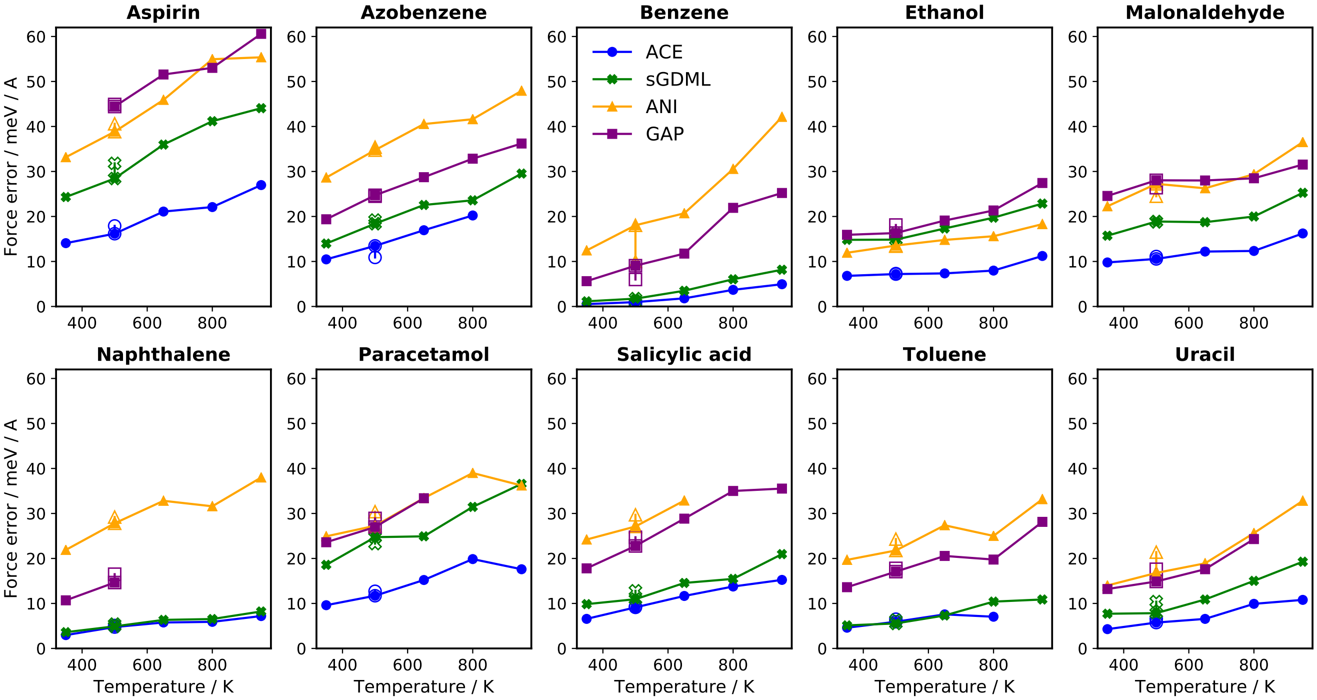

Figure 5: **Extrapolation force error.** The mean absolute force error as a function of temperature sampled from five independent MD trajectories driven by each model, at several temperatures. The empty marker corresponds to the MAE on the 500 K test set of Table 2.

apart from each of trajectories. These constitute the new test set specific for each molecule, model and temperature.

The energies and forces of the new test configurations were recomputed with DFT to estimate the accuracies of the different models at each temperature along their own MD trajectories. The force errors are shown in fig 5 whilst the energy errors showing the same trends are in fig S5. Where a point is missing, the model hit a hole in the potential and the MD run was terminated. This happened most often with the GAP model, indicating that this potential was the least regular. The linear ACE and ANI models can also be prone to hitting holes in the potential at the highest temperatures. Of all models sGDML was the most stable, it always kept the molecule intact even at 950 K for the duration of the simulations. Such extreme stability is not necessarily chemically realistic (see the next section on extrapolation to bond breaking).

Looking at the increase in errors with temperature for the different models we can see that the linear ACE often keeps the errors low with a small slope whereas the other models show a clearer increase as the temperature increases. This can be best observed for ethanol, malonaldehyde and uracil. It is notable that the model that works best at lower temperatures (in the training regime) also works best at higher temperatures confirming that the models are able to smoothly extrapolate away from the training data. Furthermore, we can see a good agreement of the test set force MAE in Table 2 with the force MAEs estimated from the models' own trajectories. This hints that the models explore similar regions of the configuration space as the original *ab initio* trajectories.

3.1.4 Extrapolation far from the training set

To test the extrapolation properties of the different models further we looked at two tests probing the torsional profile of azobenzene and O-H bond breaking in ethanol. Both of these tests probe how far away the models can smoothly extrapolate from the training data.

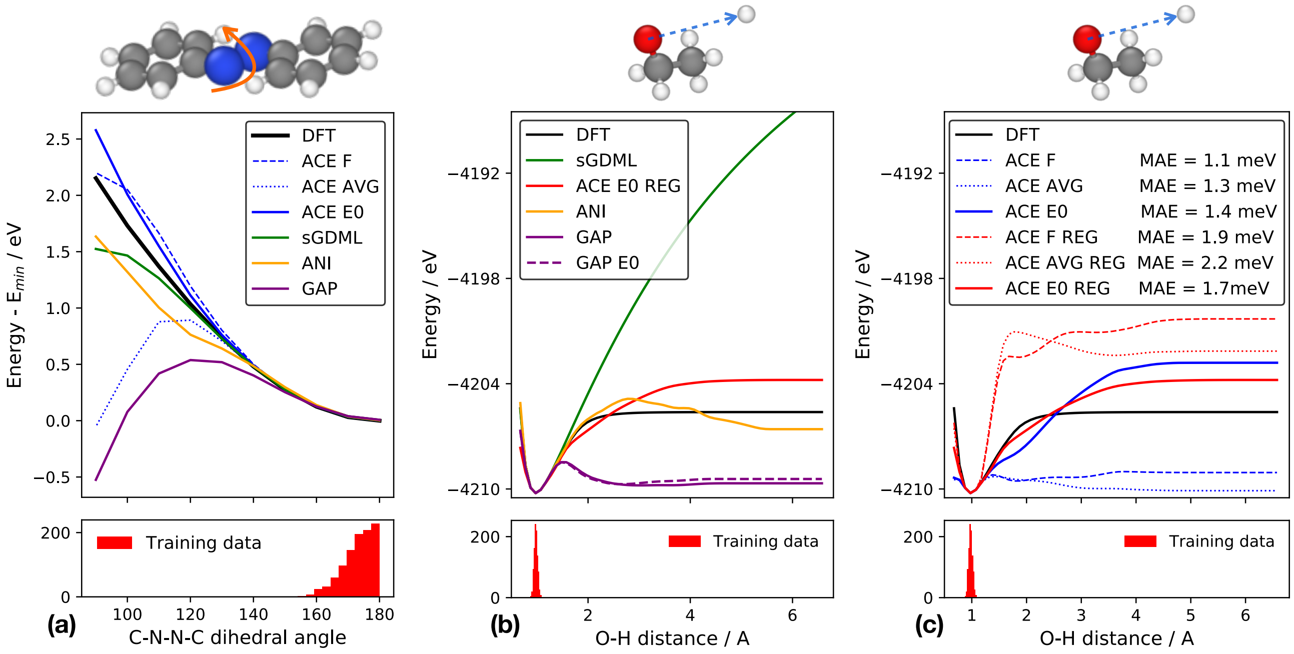

Figure 6: **Extrapolation test far away from training data.** The bottom panels show the histogram of the variables in the training set. **(a)** the energy change predicted by the different models as the C-N-N-C dihedral angle of azobenzene is decreased from the equilibrium 180 degrees. `ACE F` refers to training on forces only, `ACE AVG` refers to using average per atom energy as the 1-body term and `ACE E0` refers to using the isolated atom energy as the 1-body term. **(b)** change in energy as the O-H bond distance of ethanol is extended from equilibrium, as predicted by the different models. **(c)** comparison of ACE models with (i) lowest force MAE and (ii) with slightly stronger regularization, the latter indicated with the `REG` label.

We carried out these tests with several different versions of the linear ACE models differing in the definition of their 1-body terms, because we expect this choice to make a significant difference in how chemically reasonable the fitted models are far from the training set. We denote the ACE models fitted using force data only by `ACE F`. This has the lowest force error on the test set (comparison shown in Table S3). For the other two ACE models, energies were also included in the training. They differ in the 1-body term only, the model using average per-atom training set energy is denoted as `ACE AVG`, whereas the model using the isolated atom energies as the 1-body term is denoted `ACE E0`. The third option is the natural choice, as this ensures that if all atoms are separated from each other the predicted energy will correctly correspond to the sum of the isolated atom energies.

Fig 6(a) shows the torsional energy profile of the azobenzene molecule. The `ACE E0` model with the isolated atom 1-body term is able to extrapolate furthest, somewhat overestimating the energy, while the ANI and sGDML models also extrapolate smoothly, but slightly underestimate the energy. The linear ACE model with the average energy 1-body term and the GAP model fail to extrapolate and predict a completely nonphysical drop in energy for smaller values of the dihedral angle.

Fig 6(b) shows the energy profile as the O-H distance is varied starting from the equilibrium geometry of ethanol. The only force field that shows qualitative agreement with DFT is the `ACE E0` model. (Note that we do not expect any of the fitted models to quantitatively reproduce the DFT energy profile, even when the isolated H atom is described correctly by design, because the $C_2H_5O^{\cdot}$ radical is not.) We attribute this success to the explicit body ordered nature of the linear ACE model, including using the isolated atom as the 1-body term, and careful regularization - as was the case in a similar test for other polynomial models.[25] Fig 6(c) shows a detailed comparison of the different ACE models together with their test set MAE value. This shows that having the lowest possible test set error does not coincide with the most physically reasonable model, and using stronger regularization can lead to much smoother extrapolation. The more strongly regularized ACE models with relatively higher force error are still significantly more accurate than sGDML, ANI, GAP or the classical force field.

Interestingly, having the isolated atom as the 1-body term is not sufficient for good extrapolation. This is shown by the two different GAP models in fig 6(b), which show essentially no difference to the extrapolation, presumably due to the very poor description of the radical. GAP is not an explicitly body ordered model.

3.2 Fitting multiple molecules

Apart from sGDML, whose descriptor is tied to a given molecule with fixed topology, the models under consideration can all be fitted to multiple molecules simultaneously. Therefore, having evaluated their capacity to approximate individual potential energy surfaces one by one, it is interesting to see how the they cope with describing all of the rMD17 data set pooled together.

Table 4 shows the energy and force errors for the combined fit with linear ACE, GAP, ANI and the empirical force field. GAP and ANI errors only go up by around 30%, reflecting the fact that these are very flexible functional forms. The ANI model (which is pre-trained by starting from ANI-2x neural network weights) is now distinctly better than GAP. The empirical force field error increases by even less. In this case that is due to the use of atom-types, which help to separate the energy contribution of different functional groups. The increase in the error is largest for ACE, about a factor of two, although for most molecules it is still the combined ACE model that has the lowest error amongst these models.

In addition, we also show the performance of the original unmodified ANI-2x model (its energies and forces were tested against values recomputed with exactly the same electronic structure method and parameters that were used in its fitting[47]). Its energies and forces are better than those of the empirical force fields by factors of around 2–3 and 5, respectively. (The exception is azobenene, for which its energies are worse). The difference between ANI-2x and the re-trained ANI is about a factor of 2–4 for energies (the average over all the molecules is at the high end) and a factor of two for forces.

The other commonly used benchmark data set for machine learning based molecular force fields that contains multiple molecules is ISO17.[22] The full data set contains 5000-step *ab initio* molecular dynamics simulation trajectories of 129 molecules, all with the same chemical formula $C_7H_{10}O_2$. The standard task is to train a force field using a randomly selected 4000 configurations of 103 molecules (so about 400K configurations altogether, although these are highly correlated) and evaluate it on the remaining 1000 structures of the trajectory ("known molecules") and on the full trajectories of the "unknown molecules". We note that when all 400K training configurations are used, the conformations of "known molecules" that are usually reported as a test set are very close to the training set, at most 1 or 2 MD steps away on the trajectory from the actual training set, so the error measured on these is essentially the same as the training error.

We trained a linear ACE model on only a total of 5,000 configurations and a GAP model on only a total of 10,000 configurations sampled uniformly from the training set and evaluated them on both the known and unknown molecules. The results in Table 5 show that the linear ACE model performs significantly better than GAP, achieving errors in the same ballpark as the other methods for the unknown molecules, but using orders of magnitudes less training data. In particular, the ACE model matches the energy error of the state of the art GM-sNN[19] on the unknown molecules, demonstrating its excellent extrapolation capabilities.

Table 4: **Combined fitting of rMD17.** Mean Absolute Error of the energies (E, meV) and forces (F, meV/Å) of different models when fitting all 10 MD17 molecules together. The average energy error (E^*) is calculated on a per-atom basis.

		ACE	**GAP**	**ANI**	**FF**	**ANI-2x**
Aspirin	E	**9.9**	19.7	13.8	105.7	44.7
	F	**27.0**	46.2	35.3	287.3	76.6
Azobenzene	E	**5.7**	10.8	12.9	115.1	144.4
	F	**16.6**	30.1	30.1	243.5	115.5
Benzene	E	**0.8**	1.6	2.4	16.9	10.4
	F	**3.5**	9.3	10.6	125.6	24.4
Ethanol	E	4.6	9.2	**2.8**	42.4	22.1
	F	21.5	35.7	**15.1**	220.6	29.0
Malonaldehyde	E	5.1	9.6	**4.6**	48.8	24.2
	F	**24.1**	46.2	24.6	278.5	58.9
Naphthalene	E	**3.5**	8.3	9.5	72.8	19.5
	F	**13.3**	30.8	26.2	306.7	54.1
Paracetamol	E	**6.5**	12.5	10.7	102.9	29.7
	F	**20.8**	37.8	29.5	275.1	60.2
Salycilic acid	E	**4.6**	9.6	7.2	83.3	21.3
	F	**18.9**	35.9	26.1	310.6	68.7
Toluene	E	**3.5**	7.7	6.9	37.1	20.1
	F	**13.1**	29.7	23.3	184.2	42.5
Uracil	E	**2.8**	5.1	4.5	50.6	14.8
	F	**15.6**	26.1	21.9	265.3	48.9
Average MAE	E^*	**0.31**	0.62	0.46	4.2	2.1
	F	**17.4**	32.8	24.3	249.7	57.9

Table 5: **ISO17 test.** Mean Absolute Error in energies (E) and forces (F) of models trained on ISO17. The results on the "known molecules" are essentially training errors (see text). The bold indicates the lowest error. The linear ACE is trained on 5,000 configurations, GAP on 10,000 configurations, the neural networks on 400K (quite correlated) configurations.

		ACE	**GAP**	**Schnet**	**Physnet**	**GM-sNN**	**GM-dNN**
known molecules	E	16	54	16	4	17	7
	F	43	102	43	5	28	12
unknown molecules	E	**85**	169	104	127	**85**	118
	F	75	128	95	**60**	72	85

For all the neural network models, the error on known molecules is quite a bit lower than that for the unknown molecules, which we consider to be a sign of overfitting. For ACE and GAP, the error is still lower but by a much smaller factor, helped by the explicit regularization. Tellingly, the most similar ratio is for GM-sNN, which is a shallow neural network.

3.3 Flexible molecule test: 3BPA

Finally, noting that all the MD17 molecules are rather rigid, our last test is to assess the capabilities of the different force field models on a more challenging system that has relevance for medicinal chemistry applications. We created a new benchmark data set for the flexible drug-like molecule 3-(benzyloxy)pyridin-2-amine (3BPA).[48] Though smaller than typical drug-like molecules, with a molecular weight of 200, this molecule has three consecutive rotatable bonds, as shown in fig 7. This leads to a complex dihedral potential energy surface with many local minima, which can be challenging to approximate using classical or ML force fields.[49]

3.3.1 Preparation of the data set

To prepare a suitable training data set we started by creating a grid of the three dihedral angles (α, β and γ) removing only the configurations with atom overlap. From each of the configurations corresponding to the grid points, we started short (0.5 ps) MD simulations using the ANI-1x force field.[27] This time scale is sufficient to perturb the structures towards lower potential energies, but is not enough to significantly equilibrate them. In this way we obtained a set of 7000 configurations as shown in the left panel of Fig 7. From the distribution of dihedral angles, five different densely populated pockets were identified in the space of the three dihedral angles. One random configuration was selected from each of the 5 pockets and a long 25 ps MD simulation was performed at three different temperatures (300 K, 600 K, 1200 K) using the Langevin thermostat and 1 fs timestep. We sampled 460 configurations from each of the trajectories starting after a delay of 2 ps. In this way the final data set of 2300 configurations was obtained. The configurations were re-evaluated using ORCA[50] at the DFT level of theory using the ωB97X exchange correlation functional[51] and the 6-31G(d) basis set. (These settings are similar to that used in the creation of the ANI-1x data set[47]). From the total data set we created two training sets, one using 500 randomly selected geometries from the 300 K set, and another one, labelled "mixed-T", selecting 133 random configurations from each of the trajectories at the three temperatures. The rest of the data in each case makes up the three test sets, each corresponding to a different temperature. The right hand panels of Fig 7 show the distribution of dihedral angles in the test sets. At 300 K the separate pockets of the configuration space are sampled mostly individually, whereas at 1200 K the distribution widens significantly, and the sampling connects the pockets across multiple barriers with ease.

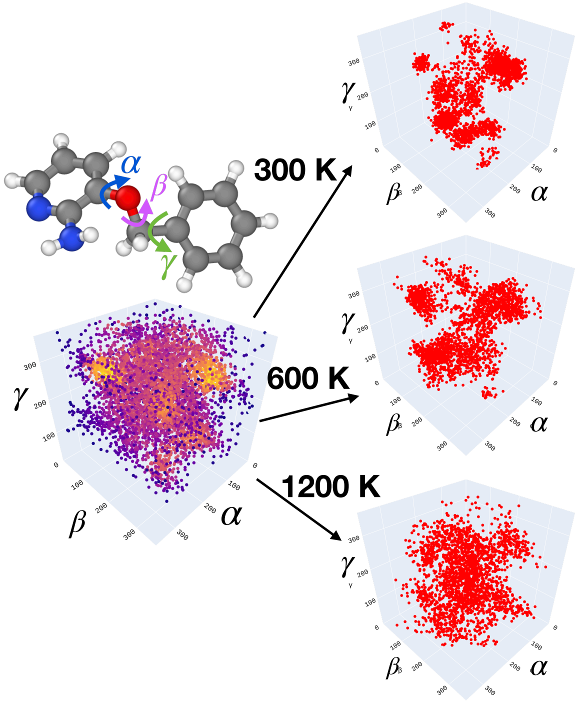

Figure 7: **3BPA data set.** The three freely rotating angles of the 3BPA molecule together with a characterization of the three different data sets sampled at different temperatures showing how the phase space sample increases significantly with temperature.

Table 6: Root mean squared error of the energy (meV) and force (meV/Å) predictions of different models of the flexible 3BPA molecule.

		ACE	sGDML	GAP	FF	ANI	ANI-2x
Fit to 300K							
300 K	E	**7.4**	8.4	20.2	59.2	22.8	38.6
	F	**25.8**	41.3	85.8	302.0	41.8	84.4
600 K	E	**27.3**	478.6	61.2	136.6	37.5	54.5
	F	**64.2**	437.6	152.0	408.3	71.7	102.8
1200 K	E	90.4	774.0	166.2	332.6	**77.5**	88.8
	F	166.6	708.8	305.0	678.1	**129.5**	139.6
Fit to mixed-T							
300 K	E	**9.5**	11.3	26.8	84.8	21.4	38.6
	F	**30.7**	53.0	85.2	306.9	56.3	84.4
600 K	E	**21.6**	24.9	48.4	115.2	39.2	54.5
	F	**53.6**	91.2	122.9	391.9	81.0	102.8
1200 K	E	**50.2**	76.3	99.2	275.0	75.2	88.8
	F	**109.4**	171.3	215.3	641.8	130.0	139.6

3.3.2 Comparison of force fields models

We trained linear ACE, sGDML, ANI and GAP force fields, and re-parametrized the bonded terms of a classical force field (FF), using the 300 K and the mixed-T training sets. Table 6 shows the energy and force RMSEs of the different models alongside the general purpose ANI-2x force field errors on the same configurations. Just as before, the weights of the re-trained ANI model were initialized form the ANI-2x weights, giving it a considerable advantage over the other models, especially because the DFT functional and basis set that we use are the same as that of the underlying DFT method of the ANI-2x model.

For the case of training on the 300 K configurations the linear ACE and sGDML models are able to achieve very low errors when tested at the same temperature, but the ACE model shows significantly better extrapolation properties to the configurations sampled at higher temperatures. The model extrapolating most accurately to 1200 K is the re-trained ANI force field, but the linear ACE is not far behind, especially considering how poor the extrapolation of the other models are. Just as for the smaller molecules, the fitted empirical force field shows much higher errors, about a factor of 2–4 for energies and a factor of 4 for forces compared with the ANI-2x force field. Only at 1200 K does ANI-2x become competitive with the ACE trained at 300 K.

Training on the mixed-T training set leads to a significant drop in the errors at the higher temperature test sets for all ML models, but not for the empirical force field. The linear ACE model achieves the lowest error in every case, showing approximately 40% decrease in the error for the high temperature test set. The other ML models improve also, by even bigger factors (because their extrapolation power was less). The gains over the general ANI-2x force field, nearly a factor of two in energies for all three test sets, show the potential scope for parametrizing such custom force fields in medicinal chemistry applications. The errors in the empirical force field are mostly unchanged, quantifying the limitations of the fixed functional form when describing the anharmonic high energy parts of the potential energy surface.

To look beyond the energy and force RMSE, we performed a constrained geometry optimization using the different force field models and DFT to map out the dihedral potential energy surface of the molecule. The complex energy landscape is visualized in Fig 8(a) at three different fixed values of β, in the α-γ plane, limiting the range to avoid overlapping atoms. Figure 8(b) shows a comparison of the ML and

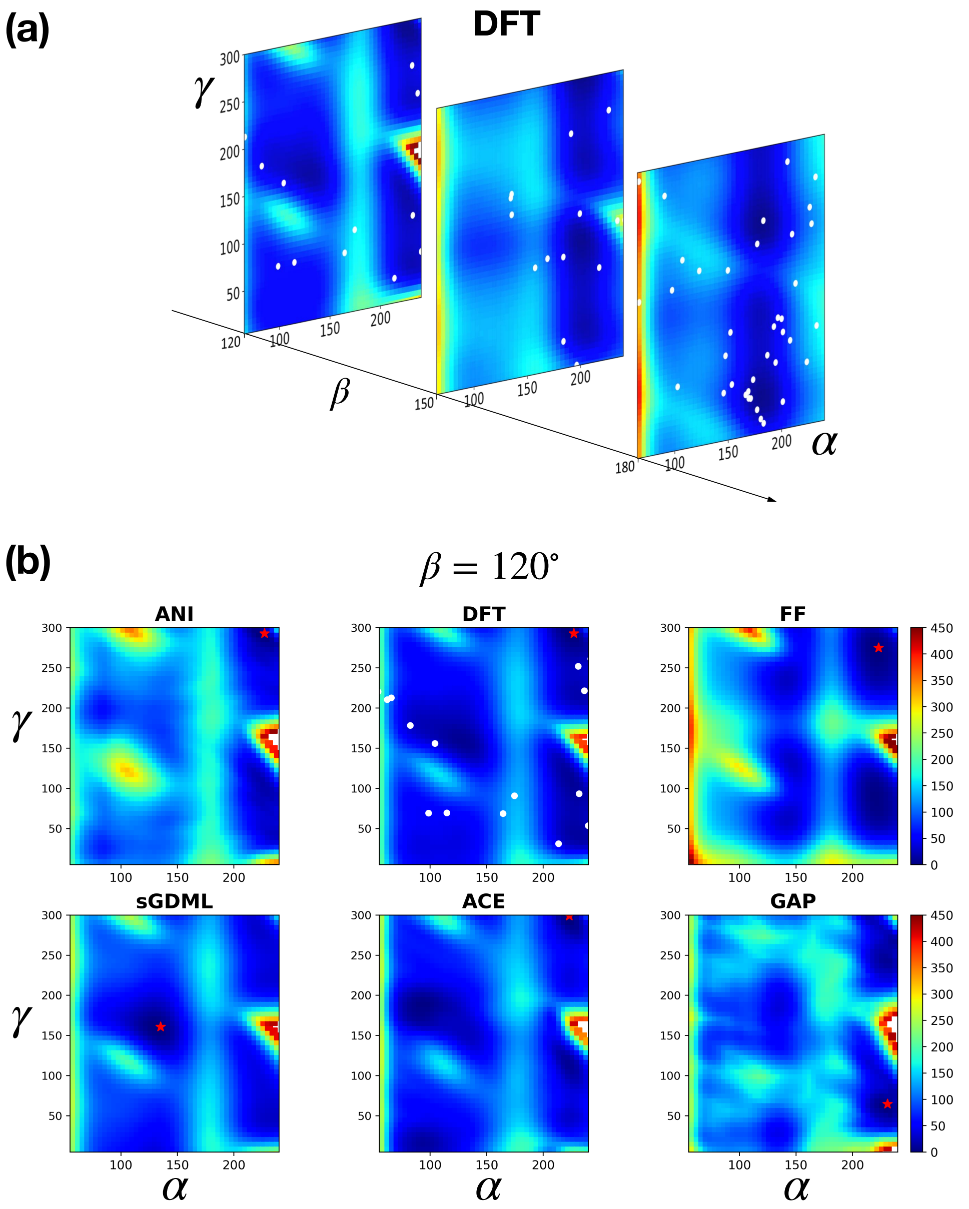

Figure 8: **Dihedral PES of 3BPA.** (a) The dihedral potential energy landscape of 3BPA for different fixed β, as predicted by DFT. (b) The $\beta = 120^o$ section of the PES for the different force field models. The white dots on the DFT PES correspond to configurations from the training set that lie within $\pm 10^o$ of the planes considered here and the red star shows the position of the energy minimum on each slice.

empirical force fields with DFT for the case of $\beta = 120°$, the plane with the fewest training data points. Analogous results for the other two values of β are reported in Figs S6 and S7. The energy landscape of the empirical force field has most of the features of the DFT landscape and is even correctly predicting the position of the lowest energy minimum in the $\beta = 120°$ plane. Some of the potential energies on this plane are clearly too high however. On the other hand the landscape of the GAP model is quite irregular, some of the most basic features are either missing or blurred together. The ANI landscape is also quite irregular, somewhat less than GAP, and some of the high energy peaks are too high and too broad. This is an example where the fixed functional form of the classical force field gives better extrapolation behavior to parts of the configuration space where there is little training data. The RMSE results clearly do not give a full, perhaps not even a very useful distinction between these models.

The ACE and sGDML models reproduce the landscape much more closely (and indeed these are the models with the lowest RMSE as well). Some differences include the sGDML getting the position of the lowest energy minimum wrong and ACE having too high a peak at $\alpha = 230°$, $\gamma = 150°$.

4 Conclusions

In this paper we have demonstrated how the Atomic Cluster Expansion framework can be used as linear models of molecular force fields. We showed that body ordered linear models built using the ACE basis are competitive with the state of the art short range ML models on a variety of standard tests. Furthermore we carried out a number of "beyond RMSE" tests to compare the ML approaches, and to study the smoothness and extrapolation properties of the fitted force fields: vibrational frequencies, force-field driven molecular dynamics and extrapolation to bond-breaking.

We also introduced a data set on a flexible drug-like molecule, with the idea that testing the performance on it is more predictive of the quality of the model for medicinal chemistry applications. The linear ACE model was significantly smoother than other transferable models and was able to extrapolate to higher potential energy regions than all other models.

We showed that the ACE framework allows us to build accurate force fields with very low evaluation cost. Together with competing approaches that are in the recent literature and in our comparison tables, the prospects are good for being able to carry out large scale biomolecular simulations with electronic structure accuracy in the near future. A number of bottlenecks remain for ACE, which include the steep increase in the number of basis functions as new chemical elements are added to the model. This can be tackled via sparsification strategies, which is the focus of our future work. Furthermore the inclusion of long range electrostatics and charge transfer are essential for the simulation of biomolecular systems and an integration of these into the ACE framework is also underway. Currently ACE is implemented in the Julia language, but can readily be called from Python via the Atomic Simulation Environment (ASE). The fitted models can also be evaluated via LAMMPS.

Acknowledgement

The authors thank Graeme Robb for valuable discussions, D.P.K. acknowledges support from AstraZeneca and the Engineering and Physical Sciences Research Council, D.J.C. acknowledges financial support from a UKRI Future Leaders Fellowship grant MR/T019654/1.

References

(1) Born;, M.; Oppenheimer, J. R. Zur Quantentheorie der Molekeln. *Annalen der Physik* **1927**, *389*, 457–484.

(2) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general Amber force field. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.

(3) Hagler, A. T. *Journal of Computer-Aided Molecular Design*; Springer International Publishing, 2019; Vol. 33; pp 205–264.

(4) Riniker, S. Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. *Journal of Chemical Information and Modeling* **2018**, *58*, 565–578.

(5) Mobley, D. L.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Slochower, D. R.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K. Escaping Atom Types in Force Fields Using Direct Chemical Perception. *Journal of*

Chemical Theory and Computation **2018**, *14*, 6076–6092.

(6) Wang, L. P.; Martinez, T. J.; Pande, V. S. Building force fields: An automatic, systematic, and reproducible approach. *Journal of Physical Chemistry Letters* **2014**, *5*, 1885–1891.

(7) Dahlgren, M. K.; Schyman, P.; Tirado-Rives, J.; Jorgensen, W. L. Characterization of biaryl torsional energetics and its treatment in OPLS all-atom force fields. *Journal of Chemical Information and Modeling* **2013**, *53*, 1191–1199.

(8) Horton, J. T.; Allen, A. E.; Dodda, L. S.; Cole, D. J. QUBEKit: Automating the Derivation of Force Field Parameters from Quantum Mechanics. *Journal of Chemical Information and Modeling* **2019**, *59*, 1366–1381.

(9) Mobley, D. L.; Lim, V. T.; Hahn, D. F.; Tresadern, G.; Bayly, C. I. Benchmark assessment of molecular geometries and energies from small molecule force fields. *F1000Research* **2020**, *9*, 1–36.

(10) Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. M.; Verstraelen, T.; Grama, A.; Van Duin, A. C. The ReaxFF reactive force-field: Development, applications and future directions. *npj Computational Materials* **2016**, *2*.

(11) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; Distasio, R. A.; Head-Gordon, M.; Clark, G. N.; Johnson, M. E.; Head-Gordon, T. Current status of the AMOEBA polarizable force field. *Journal of Physical Chemistry B* **2010**, *114*, 2549–2564.

(12) Lindahl, E.; Bjelkmar, P.; Larsson, P.; Cuendet, M. A.; Hess, B. Implementation of the charmm force field in GROMACS: Analysis of protein stability effects from correction maps, virtual interaction sites, and water models. *Journal of Chemical Theory and Computation* **2010**, *6*, 459–466.

(13) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **1996**, *118*, 11225–11236.

(14) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters* **2010**, *104*, 1–4.

(15) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole Von Lilienfeld, O. FCHL revisited: Faster and more accurate quantum machine learning. *Journal of Chemical Physics* **2020**, *152*.

(16) Chmiela, S.; Sauceda, H. E.; Poltavsky, I.; Müller, K. R.; Tkatchenko, A. sGDML: Constructing accurate and data efficient molecular force fields using machine learning. *Computer Physics Communications* **2019**, *240*, 38–45.

(17) Devereux, C.; Smith, J. S.; Davis, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *Journal of Chemical Theory and Computation* **2020**, *16*, 4192–4202.

(18) Schütt, K. T.; Unke, O. T.; Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. **2021**,

(19) Zaverkin, V.; Kästner, J. Gaussian Moments as Physically Inspired Molecular Descriptors for Accurate and Scalable Machine Learning Potentials. *Journal of Chemical Theory and Computation* **2020**, *16*, 5410–5421.

(20) Klicpera, J.; Groß, J.; Günnemann, S. Directional Message Passing for Molecular Graphs. **2019**,

(21) Anderson, B.; Hy, T. S.; Kondor, R. Cormorant: Covariant molecular neural networks. *Advances in Neural Information Processing Systems* **2019**, *32*.

(22) Schütt, K. T.; Kindermans, P. J.; Sauceda, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K. R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems* **2017**, *2017-Decem*, 992–1002.

(23) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *Journal of Chemical Theory and Computation* **2019**, *15*, 3678–3693.

(24) Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks* **1991**, *4*, 251–257.

(25) Allen, A. E. A.; Dusson, G.; Ortner, C.; Csányi, G. Atomic permutationally invariant polynomials for fitting molecular force fields. *Machine Learning: Science and Technology* **2021**, *2*, 025017.

(26) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters* **2007**, *98*, 1–4.

(27) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science* **2017**, *8*, 3192–3203.

(28) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine learning unifies the modeling of materials and molecules. *Science Advances* **2017**, *3*, 1–9.

(29) Brown, A.; McCoy, A. B.; Braams, B. J.; Jin, Z.; Bowman, J. M. Quantum and Classical Studies of Vibrational Motion of CH5 on a Global Potential Energy Surface Obtained from a Novel Ab Initio Direct Dynamics Approach. *J. Chem. Phys.* **2004**, *121*, 4105–4116.

(30) Braams, B. J.; Bowman, J. M. Permutationally Invariant Potential Energy Surfaces in High Dimensionality. *Int. Rev. Phys. Chem.* **2009**, *28*, 577–606.

(31) van der Oord, C.; Dusson, G.; Csanyi, G.; Ortner, C. Regularised Atomic Body-Ordered Permutation-Invariant Polynomials for the Construction of Interatomic Potentials. **2019**,

(32) Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B* **2019**, *99*, 1–15.

(33) Dusson, G.; Bachmayr, M.; Csanyi, G.; Drautz, R.; Etter, S.; van der Oord, C.; Ortner, C. Atomic Cluster Expansion: completeness, efficiancy and stability. **2019**, 1–41.

(34) Shapeev, A. V. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Modeling and Simulation* **2016**, *14*, 1153–1173.

(35) Lysogorskiy, Y.; van der Oord, C.; Bochkarev, A.; Menon, S.; Rinaldi, M.; Hammerschmidt, T.; Mrovec, M.; Thompson, A.; Csányi, G.; Ortner, C.; Drautz, R. Performant implementation of the atomic cluster expansion (PACE): Application to copper and silicon. **2021**, 1–16.

(36) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K. R. Machine learning of accurate energy-conserving molecular force fields. *Science Advances* **2017**, *3*.

(37) Young, T.; Johnston-Wood, T.; Deringer, V.; Duarte, F. A Transferable

Active-Learning Strategy for Reactive Molecular Force Fields. **2021**, 1–24.

(38) Fonseca, G.; Poltavsky, I.; Vassilev-Galindo, V.; Tkatchenko, A. Improving molecular force fields across configurational space by combining supervised and unsupervised machine learning. *The Journal of Chemical Physics* **2021**, *154*, 124102.

(39) Kovács, D. P.; McCorkindale, W.; Lee, A. A. Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias. *Nature Communications* **2021**, *12*, 1–9.

(40) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Physical Review B - Condensed Matter and Materials Physics* **2013**, *87*, 1–16.

(41) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; The MIT Press, 2006.

(42) Paige, C. C.; Saunders, M. A. LSQR: An Algorithm for Sparse Linear Equations and Sparse Least Squares. *ACM Transactions on Mathematical Software (TOMS)* **1982**, *8*, 43–71.

(43) Hong, Y. P.; Pan, C.-T. Rank-Revealing QR factorizations and the singular value decomposition. *Mathematics of Computation* **1992**, *58*, 213–232.

(44) Christensen, A. S.; Anatole von Lilienfeld, O. On the role of gradients for machine learning of molecular energies and forces. *Machine Learning: Science and Technology* **2020**, *1*.

(45) Wang, L. P.; Chen, J.; Van Voorhis, T. Systematic parametrization of polarizable force fields from quantum chemistry data. *Journal of Chemical Theory and Computation* **2013**, *9*, 452–460.

(46) Bernstein, N.; Csányi, G.; Deringer, V. L. De novo exploration and self-guided learning of potential-energy surfaces. *npj Computational Materials* **2019**, *5*, 1–9.

(47) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific Data* **2020**, *7*, 1–10.

(48) Cole, D. J.; Mones, L.; Csányi, G. A machine learning based intramolecular potential for a flexible organic molecule. *Faraday Discussions* **2020**, *224*, 247–264.

(49) Vassilev-Galindo, V.; Fonseca, G.; Poltavsky, I.; Tkatchenko, A. Challenges for Machine Learning Force Fields in Reproducing Potential Energy Surfaces of Flexible Molecules. **2021**, 1–25.

(50) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA quantum chemistry program package. *The Journal of chemical physics* **2020**, *152*, 224108.

(51) Chai, J. D.; Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *Journal of Chemical Physics* **2008**, *128*.