Machine Learning: Science and Technology

PAPER • OPEN ACCESS

Optimal data generation for machine learned interatomic potentials

To cite this article: Connor Allen and Albert P Bartók 2022 *Mach. Learn.: Sci. Technol.* **3** 045031

MACHINE
LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
25 July 2022

REVISED
5 September 2022

ACCEPTED FOR PUBLICATION
17 October 2022

PUBLISHED
28 December 2022

PAPER

Optimal data generation for machine learned interatomic potentials

Connor Allen[1] and Albert P Bartók[1,2,*]

[1] Department of Physics, University of Warwick, Coventry CV4 7AL, United Kingdom
[2] Warwick Centre for Predictive Modelling, School of Engineering, University of Warwick, Coventry CV4 7AL, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: apbartok@gmail.com

Keywords: database generation, interatomic potential fitting, phonon dispersion

Abstract

Machine learning interatomic potentials (MLIPs) are routinely used atomic simulations, but generating databases of atomic configurations used in fitting these models is a laborious process, requiring significant computational and human effort. A computationally efficient method is presented to generate databases of atomic configurations that contain optimal information on the small-displacement regime of the potential energy surface of bulk crystalline matter. Utilising non-diagonal supercell (Lloyd-Williams and Monserrat 2015 *Phys. Rev.* B **92** 184301), an automatic process is suggested for *ab initio* data generation. MLIPs were fitted for Al, W, Mg and Si, which very closely reproduce the *ab initio* phonon and elastic properties. The protocol can be easily adapted to other materials and can be inserted in the workflow of any flavour of MLIP generation.

1. Introduction

Modern approaches to material discovery and characterisation include the use of *ab initio* modelling. While well-established methods, such as density functional theory (DFT), reliably predict the electronic, mechanic and thermodynamic properties of materials [1, 2], most of these techniques are limited by the fact that computational effort scales as $\mathcal{O}(N^3)$ or worse with the number of atoms (N). Although linear scaling implementations of DFT exist [3, 4], large prefactors prevent efficient sampling of atomic configurations, which are required, for example, to compute thermodynamic averages. In the past decade, data driven approaches emerged as possible solutions to realise *ab initio* accuracy at an affordable computational cost, even at large length and time scales [5, 6]. Surrogate models of the Born–Oppenheimer potential energy surface (PES) can be generated in the form of (MLIPs) [7–10]. These are based on non-linear, non-parametric regression of the PES, fitted using databases of atomic configuration and their associated *ab initio* total energies and derivatives. Exploiting locality, or the nearsightedness of quantum mechanics [11], fitting can be performed on configurations containing relatively few atoms, therefore keeping the computational cost of generating the database affordable, while the resulting MLIP may be used in extended systems. Machine learning techniques in atomic modelling have evolved into a mature field, with a broad range of methods present, such as Schnet [12], MTP [13], ACE [14], NN [15], PhysNet [16] and Gaussian approximation potential (GAP)[17], among others. While the underlying principles of MLIPs can vary significantly, they all rely on carefully built databases that contain atomic configurations representative of a wide range of atomic environment that are relevant to the intended purpose of the model.

Creating such databases of atomic configurations are time consuming, both in terms of human and computational effort. Even though automated approaches, such as active learning [18, 19] can eliminate human intervention to a large extent, 'hand-crafting' parts of the database is often necessary to include specific configurations, such as various known crystalline polymorphs, defects or surfaces. Accurate modelling of the elastic and vibrational properties of bulk crystals is crucial in numerous applications, such as the finite temperature stability of different phases or defect formation energies. To provide targeted fitting data for the phonon spectrum, samples from molecular dynamics calculations [20] or specifically perturbed configurations [21] are employed routinely. In this work, we suggest a highly efficient approach based on non-diagonal supercells (NDSCs) introduced by Lloyd-Williams and Monserrat [22], which can be used to automatically generate small atomic configurations that contain optimal information to fit the PES in the small-displacement regime. As *ab initio* calculations only need to be performed on configurations containing only a handful of atoms each, data generation is efficient. We used the GAP framework [23] to fit MLIPs of bulk crystals of metallic and semiconducting elements representing different crystal structures. In our benchmarks, we obtained highly accurate phonon dispersions and elastic properties when comparing to the underlying DFT model.

2. Background

2.1. Gaussian approximation potential

The machine learned potential framework we use is GAP [8], although we emphasise that the database generation workflow is easily transferable to other approaches. GAP can be formulated as a kernel based method that predicts the total energy of a given configuration $\boldsymbol{X} = \{\mathcal{R}_1, \mathcal{R}_2, \ldots \mathcal{R}_I\}$ as:

$$E(\boldsymbol{X}) = \sum_i^I \sum_s^M \alpha_s K(\mathcal{R}_i, \mathcal{R}_s) \tag{1}$$

where $\mathcal{R}$ represents an atomic environment, s is a summation over a set of M representative environments, each associated with a weight α_s. The kernel function, $K(\mathcal{R}, \mathcal{R}')$, may be regarded as a similarity measure between two atomic environments $\mathcal{R}$ and $\mathcal{R}'$. In this work, we describe atomic environments using the Smooth Overlap of Atomic Positions (SOAP)[24, 25] descriptor, where a given atomic neighbourhood environment is initially characterised as a density:

$$\rho_i(\mathbf{r}) = \sum_{i'} f_{\mathrm{cut}}(r_{ii'}) e^{-|\mathbf{r}-\mathbf{r}_{ii'}|^2/2\sigma^2_{\mathrm{atom}}} \tag{2}$$

where a Gaussian with a width of σ_{atom} is centred on each atom up to a specified cutoff radius, whereby beyond this cutoff, f_{cut} smoothly goes to zero. This density is then expressed in terms of radial and spherical harmonics basis functions

$$\rho_i(\mathbf{r}) = \sum_{nlm} c^i_{nlm} Y_{lm}(\hat{\mathbf{r}}) g_n(\mathbf{r})$$

which are defined up to a specified complexity controlled by n_{max} and l_{max} and $m = -l, -l+1, \ldots l$. Rotationally invariant features are constructed from the power spectrum elements as

$$\tilde{\mathbf{p}}_i \equiv \sum_{m=-l}^{l} c^{i*}_{nlm} c^i_{n'lm}$$

which are normalised

$$\mathbf{p}_i = \tilde{\mathbf{p}}_i / |\tilde{\mathbf{p}}_i|.$$

Finally, we obtain an expression for our covariance evaluation between atomic neighbourhoods as

$$K(\mathcal{R}, \mathcal{R}') = \delta^2 (\mathbf{p} \cdot \mathbf{p}')^{\zeta} \tag{3}$$

where δ and ζ are hyperparameters that control the energy scaling of the descriptor and smoothness of the kernel, respectively.

To obtain the weights α_s, we minimise the loss function

$$\mathcal{L} = \sum_{n=1}^{N} \frac{[y_n - \tilde{y}_n]^2}{\sigma_n^2} + \sum_{s,s'}^{M} \alpha_s K(\mathcal{R}_s, \mathcal{R}_{s'}) \alpha_{s'} \tag{4}$$

where the quantity y can be the one of total energy, force or stress value of an atomic configuration, and σ_n is a hyperparameter, related to the weight or importance of each data point. y_n represents the reference *ab initio* values, whereas $\tilde{y}_n$ is the GAP prediction of the total energy using equation (1) or the appropriate derivatives, with respect to atomic coordinates or lattice deformations. This definition allows us to fit using the total energy observations, as well as the forces on atoms and the virial stress for each configuration. The second term in the loss function acts as a regulariser.

In algebraic form, the minimisation of this loss function with respect to α yields

$$\boldsymbol{\alpha} = (\mathbf{K}_{MM} + \mathbf{K}_{MN}\boldsymbol{\Sigma}^{-1}\mathbf{K}_{NM})^{-1}\mathbf{K}_{MN}\boldsymbol{\Sigma}^{-1}\mathbf{y}$$

where $\boldsymbol{\Sigma}$ contains a diagonal matrix containing the values of σ_n and $\mathbf{y} = [y_1, \ldots, y_N]$. The kernel matrices $\mathbf{K}$ contain all pairwise evaluations of kernel functions between atomic environments, where M denotes the representative set, and N refers to the reference database of atomic configurations.

2.2. Non-diagonal supercells

For an interatomic potential model to represent the PES near a stationary point, which in our case is the perfect bulk crystal, it needs to reproduce the force constant matrix (FCM) of an extended system, formulated as the Hessian of the Born–Oppenheimer total energy E with respect to Cartesian atomic coordinates

$$\Phi_{i\alpha j\beta} = \frac{\partial^2 E}{\partial r_{i\alpha}\partial r_{j\beta}}$$

where i, j denote atomic indices, and α, β represent Cartesian directions. Under the harmonic approximation, the total energy is expressed as a Taylor expansion with terms higher than second order truncated. Most MLIP approaches rely on the assumption of locality of the atomic interactions, i.e. for any small number $\varepsilon > 0$ there exist an r_cut such that all $\Phi_{i\alpha j\beta} < \varepsilon$ for $|\mathbf{r}_i - \mathbf{r}_j| > r_\mathrm{cut}$, corresponding to a truncated FCM. It should be noted that MLIP frameworks can be extended to represent long-range, such as Coulombic, interactions, but our current discussion is limited to the short-range term representing local, i.e. covalent or metallic, bonding.

It is customary to express the elements of the FCM such that they are indexed by labels of the basis atoms i and j within their primitive unit cells, and the displacement vector $\mathbf{R}_p$ that translates the two primitive unit cells into each other:

$$\Phi_{i\alpha j\beta}(\mathbf{R}_p) \equiv \Phi_{i\alpha j'\beta}$$

such that $\mathbf{r}_{j'} = \mathbf{R}_p + \mathbf{r}_j$. Fitting MLIPs is ultimately data driven, therefore atomic configurations should ideally contain information on as many elements of the truncated FCM as possible. As fitting data is most commonly provided as atomic configurations with the corresponding *ab initio* total energies, forces, and stresses, supercells capable of accommodating perturbations of distant atom pairs are highly desirable. A common approach is to use supercells generated such that their shape is as closely cubic as possible, as an attempt to include atom pairs isotropically. The lattice vectors $\mathbf{a}_s$, $\mathbf{b}_s$, and $\mathbf{c}_s$ of a supercell s are related to the unit cell lattice vectors $\mathbf{a}_u$, $\mathbf{b}_u$, and $\mathbf{c}_u$ as

$$\begin{pmatrix} \boldsymbol{a}_s \\ \boldsymbol{b}_s \\ \boldsymbol{c}_s \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{pmatrix} \begin{pmatrix} \boldsymbol{a}_u \\ \boldsymbol{b}_u \\ \boldsymbol{c}_u \end{pmatrix}$$

where elements of the supercell matrix $S_{ij} \in \mathbb{Z}$, and for a diagonal supercell (DSC) $S_{ij} = S_i\delta_{ij}$ with $S_i \in \mathbb{Z}^+$. For example, generating DSC of the cubic unit cell of fcc or bcc crystals, or in case of hexagonal crystals, using the orthorhombic unit cell is a convenient choice, as the unit cell lattice vectors are orthogonal.

Atoms need to be displaced in the supercell before computing the *ab initio* total energy and its derivatives. George *et al* [21] suggested using randomly perturbed atomic coordinates as well as the displacement of a single atom in the supercell. Randomising displacements with a certain amplitude is expected to result in force observations that are dominated by terms containing the largest elements of the FCM, as the α component of force of atom i may be approximated as

$$f_{i\alpha} \approx -\sum_{j\beta} \Phi_{i\alpha j\beta}(r_{j\beta} - r_{j\beta,0}) \qquad (5)$$

where $\mathbf{r}_{j,0}$ denotes the equilibrium position of atom j in the supercell. Since fitting of MLIPs assumes some degree of uncertainty on each observation as described in section 2.1, such dominance may have detrimental effect on the quality of the fit as small contributions will be indistinguishable from noise. More terms in equation (5), corresponding to larger supercells, is expected to aggravate the situation, leading to poor fit of small elements of the FCM. Alternatively, the displacement of a single atom along a Cartesian direction results in the resolution of each individual element of the FCM, but such configurations contain highly correlated atomic environments and cannot be regarded as realistic examples of configurations sampled from finite temperature simulations. Samples from finite temperature simulations, such as molecular dynamics, are an optimal solution, but only if the sampling uses a sufficiently similar PES to that of the *ab initio* model, otherwise the configurations will be practically equivalent to those generated by randomisation. The computational cost of the *ab initio* reference calculations when using DSC scales as $(S_1\ S_2\ S_3)^3$ if using plane-wave DFT, therefore the size of the supercell, and the representable elements of the FCM is severely limited.

Lloyd-Williams and Monserrat [22] demonstrated that perturbations that require a DSC constituting $S_1 \times S_2 \times S_3$ primitive cells, may be represented by a NDSC with no more than the least common multiple of S_1, S_2, and S_3 number of primitive cells. Lloyd-Williams and Monserrat suggested this method to sample the vibrational modes in the Brillouin zone (BZ) of a crystal uniformly on an $N \times N \times N$ grid. When computing the FCM using finite differences, DSC of the size $N \times N \times N$ are needed, whereas if using NDSCs, only supercells of size up to N are required. Even though more NDSCs have to be typically considered, each individual calculation incurs significantly less computational cost, while the process can benefit from trivial parallelisation. Overall, significant reductions in the computational cost associated with *ab initio* phonon dispersion calculations can be realised, and also allows one to consider more dense sampling of the BZ.

2.3. Phonon dispersion

With the force constants determined under the harmonic approximation, one method for finding the frequency of the allowed vibrational modes $\mathbf{q} \in$ BZ is done via finding the eigenvalues of the dynamical matrix $\mathbf{D}(\mathbf{q})$ whose elements are obtained via Fourier-transforming the mass-weighted FCM as

$$D_{i\alpha j\beta}(\mathbf{q}) = \frac{1}{\sqrt{m_i m_j}} \sum_{\mathbf{R}_p} \Phi_{i\alpha j\beta}(\mathbf{R}_p) e^{-i\mathbf{q}\cdot\mathbf{R}_p} \tag{6}$$

where m_i and m_j are the masses of atoms i and j. The square root of the eigenvalues at each $\mathbf{q}$ vector are the phonon frequencies. Negative eigenvalues result in imaginary frequencies, corresponding to dynamically unstable modes, along which displacements result in lowering the energy. As customary, we represent such imaginary frequencies as negative numbers on our phonon dispersion plots.

3. Methodology

3.1. Density functional theory calculations

The underlying *ab initio* calculations that were used to train the interatomic potentials as well as benchmark them was preformed using the plane-wave DFT code, `CASTEP`[26]. On-the-fly ultrasoft pseudopotentials [27] were generated for Mg, Al, Si, and W with the respective valence electronic structure: 2 $s^2 2p^6 3$ s^2, 3 $s^2 3p^1$, 3 $s^2 3p^2$, and 5 $s^2 5p^6 4f^{14} 6$ $s^2 5d^4$. In all instances a generalized gradient approximation [28] exchange-correlation functional was used. The plane-wave energy cutoff (E_{cut}), density of the electronic BZ sampling of a Monkhorst-Pack grid [29] (k-spacing) and the self-consistent field energy tolerance for convergence was set for each system to find a converged result on the total energy and derivative quantities. Geometry optimisations were then performed for all systems to find the relaxed lattice parameters for a given fixed crystal symmetry. The specific DFT parameter set and primitive cell information found from the geometry optimisations are presented in table 1.

The elements of the FCM for the phonon dispersion calculations at the *ab initio* level were determined using the finite difference method [30] as implemented in `CASTEP`, corresponding to a 4 × 4 × 4 grid in the BZ. A displacement of 0.05 Å from the ideal lattice site was used, and phonon dispersion curves were computed along high symmetry lines using Fourier interpolation.

3.2. Database generation

Our aim is to investigate a protocol that produces database configurations targeted to fit vibrational properties of crystalline materials in a computationally optimal way. We suggest basing the workflow on NDSCs, which can represent long-range perturbation of crystalline order using the configurations that contain the fewest possible atoms.

To construct a database that can explores displacements around the pristine crystal geometry corresponding to an $N \times N \times N$ supercell of the primitive unit cell, we generated NDSCs using the `FORTRAN` 90 program by Lloyd-Williams and Monserrat [22] which contain supercells formed of up to N primitive unit cells, where those cells related by symmetry are already eliminated. The NDSC configurations therefore contain information about the vibrational modes corresponding to a $N \times N \times N$ phonon $\mathbf{q}$-vector grid. In addition, we introduced deformation of the cells by homogeneous scaling of the cell vectors to capture isotropic compression and expansion. To capture the response of atoms displaced from ideal lattice sites within the different NDSC configurations, copies were made where atoms were randomly displaced via a

Table 1. DFT parameter set (planewave cutoff energy, spacing of the k-point sampling of the BZ and tolerance of the self-consistent iterations) used to perform *ab initio* data generation and benchmark comparison and the geometry optimised lattice information for each system.

	Mg	Al	Si	W
DFT:				
E_{cut} (eV)	520	800	400	600
k-spacing (Å^{-1})	0.012	0.010	0.030	0.015
SCF tol. (eV)	10^{-11}	10^{-11}	10^{-11}	10^{-10}
Lattice:				
Structure	hcp	fcc	dia.	bcc
a (Å)	3.198	2.856	3.867	2.756
c (Å)	5.179	—	—	—

Table 2. GAP hyperparameter set for each system, and associated data used for training. Virial stresses ($N_{\text{virial}} = 6N_{\text{energy}}$) and atomic forces ($N_{\text{force}} = 3N_{\text{atoms}}$) was also included on all configurations. Primitive cell vectors from a geometry optimisation using each potential are also presented.

	Mg	Al	Si	W
GAP:				
r_{cut} (Å)	8.0	10.0	6.0	6.0
n_{max}	8	10	8	8
l_{max}	6	8	6	6
σ_{atom} (Å)	0.5	0.5	0.5	0.5
Regularisation:				
Δ_{F}	10^{-2}	10^{-2}	10^{-2}	10^{-2}
$\sigma_{\text{F}}^{\text{min}}$ (eV Å^{-1})	10^{-3}	10^{-3}	10^{-3}	10^{-3}
Δ_{V}	10^{-2}	5×10^{-3}	10^{-2}	10^{-2}
$\sigma_{\text{V}}^{\text{min}}$ (eV)	10^{-3}	5×10^{-4}	10^{-3}	10^{-3}
Amount of Data:				
N_{atoms}	4276	1683	3300	1550
N_{energy}	1004	850	936	836
Lattice				
Structure	hcp	fcc	dia.	bcc
a (Å)	3.194	2.856	3.867	2.757
c (Å)	5.181	—	—	—

normal distribution with standard deviation of 0.10 Å. Finally, to inform the fitting procedure on how the PES responds to anisotropic cell deformations, random shearing was applied on the NDSC configurations. The lattice vectors, contained in **L** were transformed by a symmetrical strain matrix, ε, as:

$$\mathbf{L}_{\text{rand.}} = (\mathbf{I} + \boldsymbol{\epsilon})\mathbf{L}$$

where **I** is the identity matrix and each entry of the strain matrix is sample from a uniform distribution, $\epsilon_{ij} \sim \mathcal{U}(-0.01, 0.01)$, such that ε is symmetric.

To investigate the transferability of the proposed workflow for database generation using NDSC, four different crystal structures were considered: hexagonal close-packed (hcp) Mg, diamond (dia) Si, body-centred cubic (bcc) W and face-centred cubic (fcc) Al. The NDSCs were generated were commensurate with a 4 × 4 × 4 grid sampling of the vibrational BZ of the relaxed primitive cell of each system. All configuration manipulations were done through the Atomic Simulation Environment [31].

3.3. Fitting MLIPs

We used the GAP framework to generate MLIPs, but we stress that any other similar fitting approaches would benefit equally. For all models presented here we select 1400 sparse points through a CUR decomposition [32] and set $\delta = 2$ eV and $\zeta = 4$. Further GAP hyperparameters and details on the training data set are specified in table 2. For the Al Bain path model developed, additional data was included to capture the bcc phase and the half-way point on the Bain path as described by equation (7). This GAP was trained on 3751 atomic environments, for a total of 1455 target energies, using 1400 sparse points selected via CUR decomposition. For the minimal data case on fcc Al, 74 atomic environments (24 target energies) constituted the training set for the NDSC model, whereas 65 atomic environments (2 target energies) were considered for the DSC model. In the minimal data GAP for fcc Al, both the NDSC and DSC contained the geometry optimised primitive cell.

Based on the work of George *et al* [21], we employed adaptive regularisation, via adjusting the hyperparameter σ as described by the loss function in equation (4). In addition to scaling σ corresponding to force components of the data, we implemented a similar adjustment algorithm for virial stress components. Throughout this work we use a constant regularisation on the total energy predictions while the element wise viral regularisation and component wise force regularisation are implemented as

	σ_n	
energy	0.001 eV	
force	$\Delta_F\lvert\mathbf{F}_i\rvert$	if $\Delta_F\lvert\mathbf{F}_i\rvert > \sigma_F^{\mathrm{min}}$
	σ_F^{min}	else
virial	$\Delta_V\lvert V_{\alpha\beta}\rvert$	if $\Delta_V\lvert V_{\alpha\beta}\rvert > \sigma_V^{\mathrm{min}}$
	σ_V^{min}	else

introducing σ^{min} to define a minimum value for the regularisation and Δ to scale the value for each component of the corresponding quantity. The choice of regularisation parameters are summarised in table 2.

The FCM elements of the developed MLIPs were calculated using the finite difference method [30] using the `phonopy` package [33]. We calculated phonon frequencies along the high symmetry lines suggested by Setyavan and Curtarolo [34], and determined the phonon density of states based on a $40 \times 40 \times 40$ **q**-vector grid.

4. Results

Having fitted a series of GAP models for W, Al, Si and Mg using databases consisting of NDSC configurations, we evaluated the accuracy of each model by comparing its vibrational and elastic properties to DFT values. Our reference DFT calculations show good agreement with the literature [21, 33, 39–42]. Overall, we find that all fitted models show excellent performance in our benchmarks. The summary of geometric and elastic parameters predicted by the GAP models, and comparisons to DFT results is presented in table 3. Excellent agreement with DFT may be observed across all our test systems, with the root mean squared error (RMSE) on phonon modes below 0.5 THz.

We fitted a reduced model for Si that only contained the primitive unit cell configurations, in order to study the role of different elements of the database. Tabulated results in table 3 show excellent agreement of elastic constants for both models. As the elastic moduli are related to the slope of the acoustic phonon modes near the Γ-point [43], portions of the dispersion of phonon modes are also in good agreement for the minimal model, as shown in figure 1. However, at phonon modes corresponding to intermediate wavelengths the agreement for the minimal model is poor, confirming that deformed unit cells provide information to the GAP fitting about the elastic behaviour of a given material, but larger supercells are required to inform the fitting procedure on the full FCM. Indeed, adding NDSC configurations to the database, we recover the phonon dispersion across the BZ accurately.

The GAP model reproduces the elastic and vibrational properties of bcc W and fcc Al to a great accuracy, as shown in figure 2 and table 3. We note that for W the largest error compared to DFT in the phonon frequencies is at the H point, together with a discrepancy in the curvature along the $P-H$ direction. Since the phonon mode at H corresponds to displacing oppositely the two atoms located on neighbouring corners of the cubic cell, the corresponding elements of the FCM can be regarded as well represented in our database.

To illustrate the efficiency gains realised when using NDSC configurations in the training set, we compare the phonon dispersion curves corresponding to two GAP models, constructed to emphasise the advantage of using configurations consisting of fewer atoms. The two models were based on two separate databases, both of which required the same amount of computational time to calculate the *ab initio* reference energies, forces and virial stresses. To generate the first database, we used DSC of size commensurate with the desired q-point sampling, where only a single atom is displaced from its ideal lattice site, as suggested by George *et al* [21]. The other database contained NDSC configurations generated using the workflow described in section 3.2, such that the same amount of computational effort was needed to compute the *ab initio* data as for the DSC. The comparison of the phonon dispersion of the two models is presented in figure 3.

The model fitted on NDSC configurations performs noticeably better, with phonon modes in close agreement with the *ab initio* reference data. On the other hand, while the model fitted with DSCs captures some of the phonon branches, it predicts unphysical dynamical instabilities in fcc Al. While models fitted using databases based on DSC configurations are expected to achieve the accuracy of those based on NDSC

Table 3. Comparison of elastic constant predictions for NDSC GAP developed for each system, the underlying DFT calculations and experiment. GAP* refers to the reduced model in Si that does not contain NDSC configurations.

E.C	B	C_{11}	C_{12}	C_{13}	C_{33}	C_{44}	C_{66}
Mg							
GAP	36.2	65.6	20.7	22.4	63.6	17.4	22.1
DFT	36.5	64.2	20.2	23.7	64.7	17.1	20.9
Expr [35].	36.9	63.5	25.9	21.7	66.5	18.4	18.8
Al							
GAP	77.5	109.2	61.7	—	—	31.9	—
DFT	77.6	106.9	62.9	—	—	33.3	—
Expr. [36]	82.0	116.3	64.8	—	—	30.9	—
Si							
GAP	88.6	152.4	56.7	—	—	72.4	—
GAP*	89.7	155.8	56.7	—	—	72.8	—
DFT	88.6	152.7	56.6	—	—	73.3	—
Expr. [37]	99.1	167.5	64.9	—	—	80.2	—
W							
GAP	306.4	510.5	204.3	—	—	136.9	—
DFT	306.3	512.8	203.0	—	—	135.9	—
Expr. [38]	314.7	533.9	205.1	—	—	163.3	—

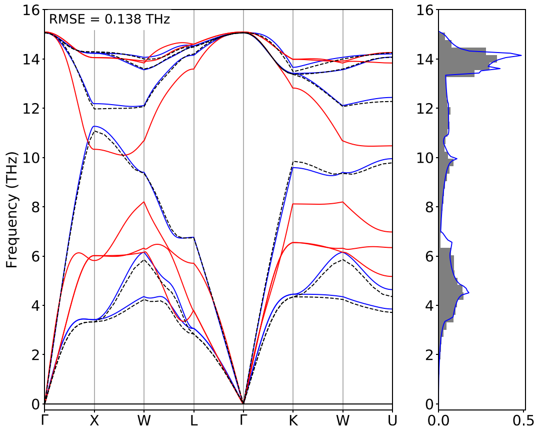

Figure 1. Phonon dispersion and density of states calculations by GAPs trained using a minimal database only containing deformed primitive unit cells (red) and the full database containing NDSC configurations (blue). Reference DFT values are shown as black dashed lines.

configurations [20, 21, 40], the computational cost of plane-wave DFT calculations scales unfavourable for the former strategy.

To establish how the performance and accuracy of our MLIP models benefit from increasing the amount of training data, we fitted a series of GAP models for Al, using different size random subsets of the NDSC data. We present our learning curves as a series of phonon dispersion diagrams in figure 4, showing two approaches: (a) keeping the set of representative atomic environments (or sparse points, set M in equation (1)) constant across the models, using the representative set selected from atomic environments in the largest data set; (b) selecting representative atomic environments from each of the fitting subset. As expected, increasing the data size leads to significant but diminishing improvements the accuracy of the model, measured as RMSE of the predicted phonon dispersion against the *ab initio* benchmark. However, it is interesting to observe that when the sparse points, which represent basis functions in the GAP framework, are selected from atomic environments not necessarily present in the training configurations, the accuracy is markedly improved even when using the same fitting targets. Therefore we suggest that GAP models may be improved by adding atomic configurations that do not need *ab initio* data associated with them, in order to increase the set of sparse points. The advantage of this approach is that significantly less computational effort is needed to generate the expensive *ab initio* data and it is possible to make improvements without the need to calculate additional target quantities at the DFT level.

We were also interested in studying how NDSC configurations may assist fitting the PES at stationary points other than minima. Aluminium at ambient conditions is dynamically unstable in the body-centred

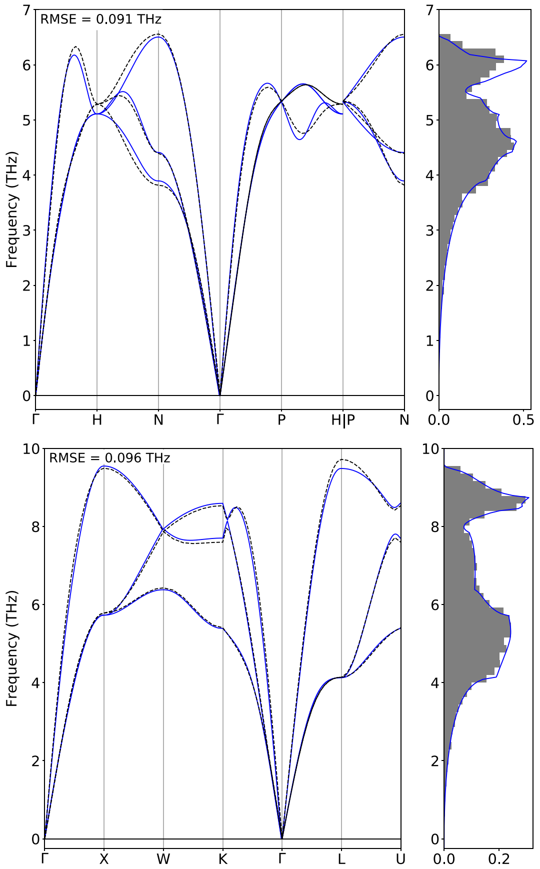

Figure 2. GAP models trained using NDSC structures (blue) versus commensurate DFT (black dashed) phonon dispersion calculations for bcc W (top panel) and fcc Al (bottom panel) with corresponding density of states.

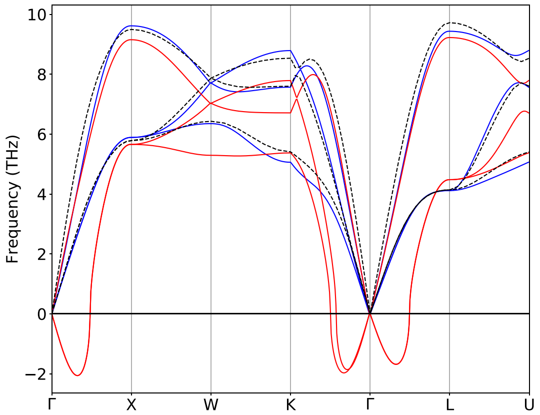

Figure 3. GAPs for Al based on NDSC (blue line, 66.1 CPU hours) and diagonal supercell (red line, 60.1 CPU hours) data as compared against underlying DFT (black dashed line) for a minimal amount of data.

cubic form, although at extreme pressures the bcc phase becomes energetically favourable [44]. We have collected training configurations along the Bain path that connects the bcc and fcc phases of Al. The lattice vectors of primitive unit cells were described as body-centred tetragonal,

$$\mathbf{L} = \begin{bmatrix} -{}^{a}\!/_{2} & {}^{a}\!/_{2} & {}^{c}\!/_{2} \\ {}^{a}\!/_{2} & -{}^{a}\!/_{2} & {}^{c}\!/_{2} \\ {}^{a}\!/_{2} & {}^{a}\!/_{2} & -{}^{c}\!/_{2} \end{bmatrix} \tag{7}$$

where the rows of $\mathbf{L}$ represent the cell vectors, and a and c change as

$$a = a_0(1 - \eta + 2^{{}^{1}\!/_{3}}\eta)$$

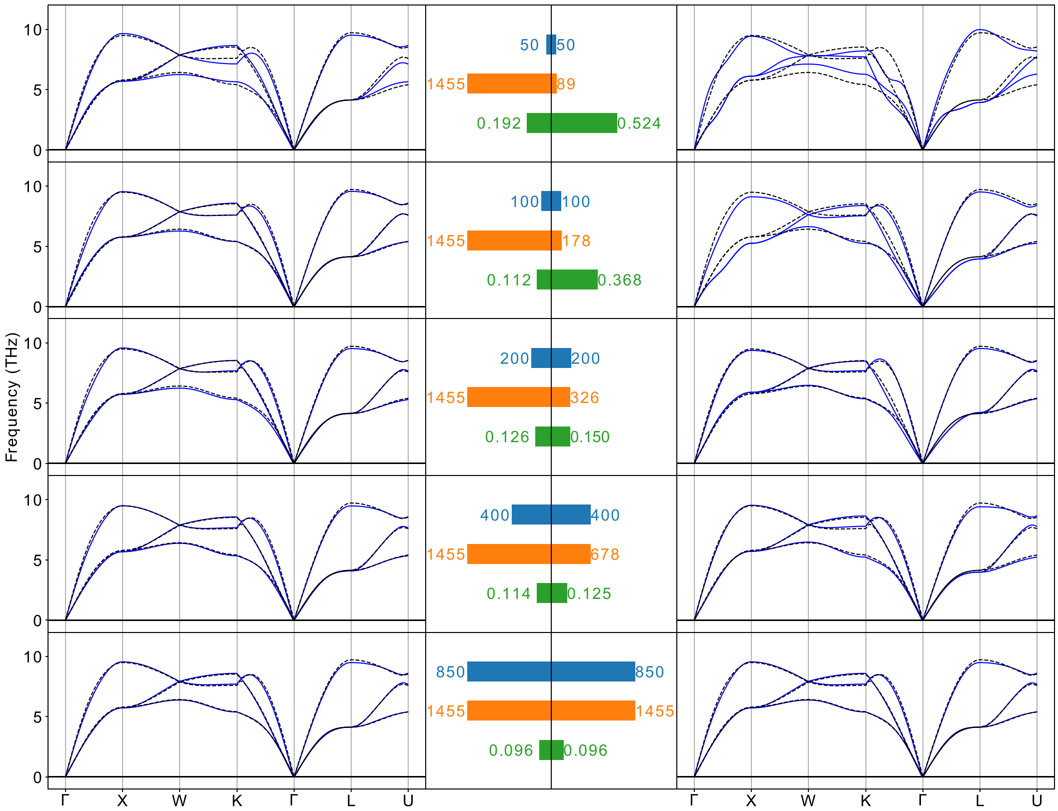

Figure 4. Left and right: Phonon dispersion curves of fcc Al GAP models (blue) compared to the DFT reference (black dashed). The set of training configurations increases from the top to the bottom, with the number of training configurations (i.e. total energy targets) displayed as blue bars (with numbers) on the centre panel. The orange bars represent the number of representative (sparse) points, as selected from a CUR decomposition, in each of the GAP models, and the green bars show the RMSE of the phonon frequencies in THz units.

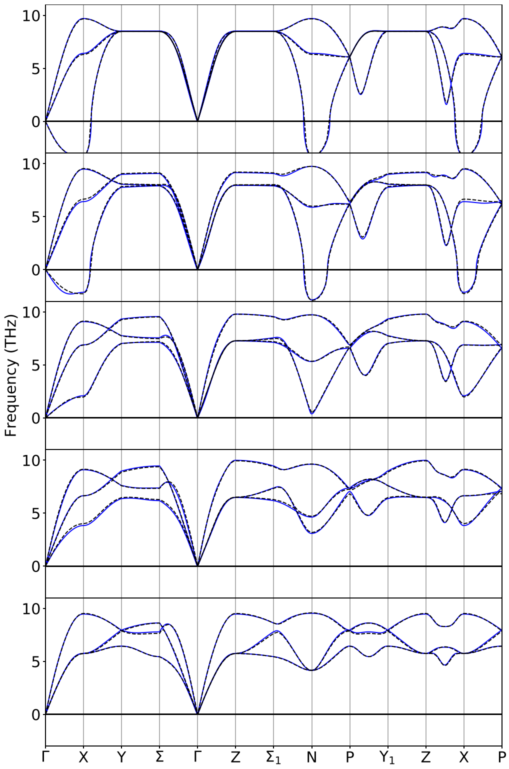

Figure 5. GAP (blue) trained using NDSC structures compared to commensurate DFT (black dashed) phonon dispersion calculations for in Al, where the lattice was transformed from the stable fcc structure (bottom) to the dynamically unstable bcc structure (top) along the Bain path.

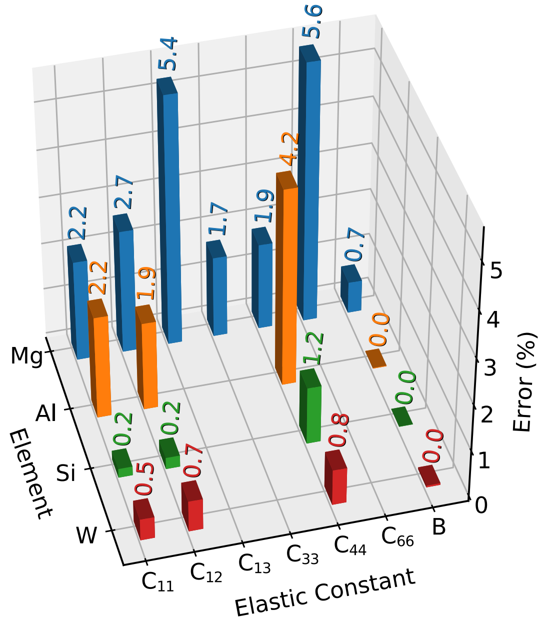

Figure 6. Relative error of elastic constants of GAP models of Mg (blue), Al (orange), Si (green) and W (red) compared to DFT results.

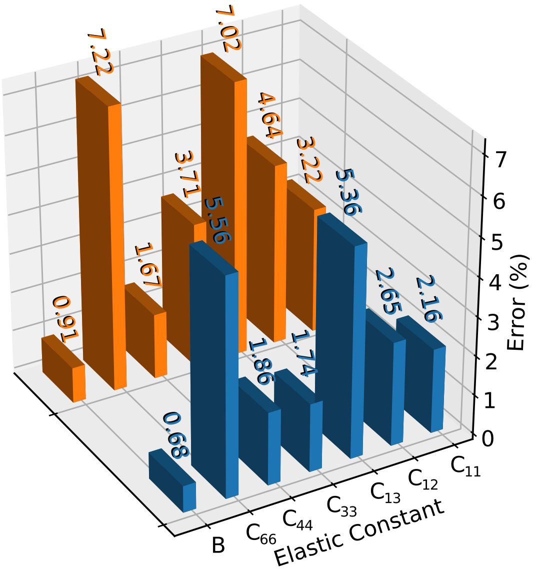

Figure 7. Errors of elastic constants of GAP models of Mg relative to DFT values in the hcp crystal structure. GAP models were fitted using static force and virial regularisation (orange bars and figures) and adaptive force and virial regularisation (blue bars and figures), as described in section 3.3.

$$c = \frac{a_0}{\sqrt{1-\eta+2^{1/3}\eta}}$$

with $0 \leqslant \eta \leqslant 1$, such that the volume of the cell $|\det \mathbf{L}|$ is conserved. With this definition, the cases $\eta = 0$ and $\eta = 1$ correspond to the bcc and fcc lattices, respectively. We fitted a GAP model based on perturbed NDSC configurations that were generated using primitive unit cells at $\eta = 0, {}^1/_2$ and 1. We benchmarked the phonon dispersion curves obtained with our model considering intermediate η values, as shown in figure 5. We found excellent agreement with DFT overall, with instabilities at the N-point reproduced highly accurately.

We have also included a non-cubic crystalline system, magnesium, in our benchmarks, whose ground state structure at ambient condition is hexagonal close packed. Using NDSC configurations to train a GAP model, we find excellent agreement between the GAP and the *ab initio* phonon dispersion curves.

While accurate phonon spectra on acoustic modes near the Γ-point indicate that the elastic properties of the crystal are well represented [43], it is insightful to directly examine the numerical values of the elastic

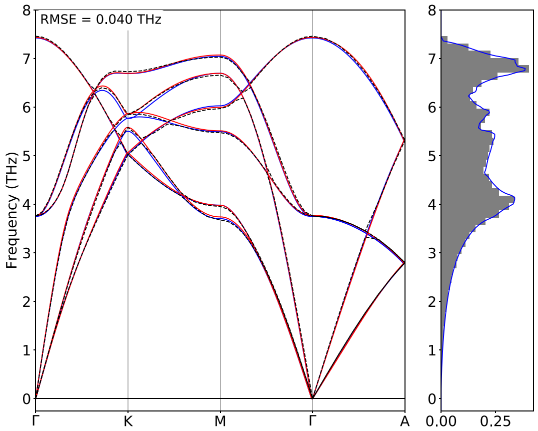

Figure 8. GAPs trained using NDSC (blue) with component-wise regularisation as informed by the force on each atom and virial stress on each configuration compared to a model where the regularisation is set to a constant value (red line) as well as DFT (black dashed) phonon dispersion curves.

constants of the MLIP models for additional benchmarking purposes. Table 3 and figure 6 summarise the elastic moduli computed at the relaxed geometries both using DFT and the GAP models, showing excellent agreement with up to 6% error.

We illustrate the effect of adaptive regularisation of individual virial stress components, introduced in section 3.2, by comparing two Mg GAP models, one of which uses a static regularisation of 0.01 eV for each virial component, while the other employing the adaptive scheme. As shown in figure 7, some of the elastic constants are only reproduced to an error of up to 7%, while introducing the adaptive virial regularisation, accuracy is significantly improved across all elastic constants without any deterioration of the quality of the phonon dispersion curves depicted in figure 8.

5. Conclusions

In conclusion, we explored a computationally efficient approach using the NDSC method introduced by Lloyd-Williams and Monserrat to generate database configurations for fitting MLIP models based on *ab initio* data. We found that NDSC configurations provide sufficient data to fit MLIPs reproducing the FCM near stationary points of bulk crystalline materials, while costing significantly less computational effort than DSCs. We have also suggested an adaptive scheme to regularise virial stress components of *ab initio* databases and demonstrated improvements of the elastic behaviour of MLIPs. The procedure described in this work can be fully automated and integrated into existing database generating workflows, allowing to save computational cost or include a greater variety of representative structures, realising savings on cost and carbon emissions associated with high-performance computations, or improved quality MLIPs. While we only used random perturbations of the atomic and lattice coordinates to sample out of equilibrium configurations, the computational advantage of using NDSCs in the *ab initio* calculations still holds if other sampling strategies, such as molecular dynamics, are used. We also envisage further use of NDSC configurations in databases used to inform models for alloy materials, as an addition or alternative to semi quasi-random structures, where the substitution of elemental species may be regarded as alchemical perturbations.

6. Software and data availability

All *ab initio* training data and the scripts used to generate the configurations are made available in a dedicated repository [45]. We used the QUIP software package with the GAP plugin [46], available under the General Public License and the Academic Source License, respectively. The Atomic Simulation Environment [47] was used to manipulate atomic configurations and we employed the phonopy package [48] to calculate the phonon dispersion curves of the GAP models. Our workflow greatly benefited from using GNU `parallel`[49].

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/ConnorSA/ndsc_tut.

Acknowledgment

We thank Noam Bernstein and Keith Refson for fruitful discussions. We acknowledge support from the NOMAD Centre of Excellence, funded by the European Commission under Grant Agreement 951786. C A is supported by a studentship jointly by the UK Engineering and Physical Sciences Research Council–supported Centre for Doctoral Training in Modelling of Heterogeneous Systems, Grant No. EP/S022848/1 and the Atomic Weapons Establishment. A B P acknowledges funding from CASTEP-USER funded by UK Research and Innovation under the Grant Agreement EP/W030438/1. Calculations were performed using the Sulis Tier 2 HPC platform hosted by the Scientific Computing Research Technology Platform at the University of Warwick. Sulis is funded by EPSRC Grant EP/T022108/1 and the HPC Midlands+ consortium. We acknowledge the University of Warwick Scientific Computing Research Technology Platform for assisting the research described within this study.

ORCID iDs

Connor Allen https://orcid.org/0000-0002-5625-7425
Albert P Bartók https://orcid.org/0000-0002-4347-8819

References

[1] Pickard C J and Needs R J 2006 *Phys. Rev. Lett.* **97** 45504
[2] Pickard C J and Needs R J 2011 *J. Phys.: Condens. Matter* **23** 053201
[3] Ordejón P, Drabold D A, Martin R M and Grumbach M P 1995 *Phys. Rev.* B **51** 1456
[4] Prentice J C A *etal* 2020 *J. Chem. Phys.* **152** 174111
[5] Deringer V L, Bernstein N, Csányi G, Mahmoud C, Ceriotti M, Wilson M, Drabold D A and Elliott S R 2020 *Nature* **589** 59
[6] Cheng B, Engel E A, Behler J, Dellago C and Ceriotti M 2019 *Proc. Natl Acad. Sci.* **116** 1110
[7] Blank T B, Brown S D, Calhoun A W and Doren D J 1995 *J. Chem. Phys.* **103** 4129
[8] Bartók A P *et al* 2010 *Phys. Rev. Lett.* **104** 136403
[9] Seko A, Takahashi A and Tanaka I 2015 *Phys. Rev.* B **92** 054113
[10] Deringer V L, Bartók A P, Bernstein N, Wilkins D M, Ceriotti M and Csányi G 2021 *Chem. Rev.* **121** 10073
[11] Prodan E and Kohn W 2005 *Proc. Natl Acad. Sci.* **102** 11635
[12] Schütt K T, Kindermans P-J, Sauceda H E, Chmiela S, Tkatchenko A and Müller K-R 2017 *Proc. 31st Int. Conf. on Neural Information Processing Systems, Series and Number NIPS'17* (Red Hook, NY: Curran Associates Inc.) p 992
[13] Shapeev A V 2016 *Multiscale Model. Simul.* **14** 1153
[14] Drautz R 2019 *Phys. Rev.* B **99** 014104
[15] Behler J 2017 *Angew. Chem.* **56** 12828
[16] Unke O T and Meuwly M 2019 *J. Chem. Theory Comput.* **15** 3678
[17] Bartók A P and Csányi G 2015 *Int. J. Quantum Chem.* **115** 1051
[18] Deringer V L and Csányi G 2017 *Phys. Rev.* B **95** 094203
[19] Deringer V L, Caro M A and Csányi G 2020 *Nat. Commun.* **11** 5461
[20] Bartók A P, Kermode J, Bernstein N and Csányi G 2018 *Phys. Rev.* X **8** 041048
[21] George J, Hautier G, Bartók A P, Csányi G and Deringer V L 2020 *J. Chem. Phys.* **153** 044104
[22] Lloyd-Williams J H and Monserrat B 2015 *Phys. Rev.* B **92** 184301
[23] Bartók A P, Payne M C, Kondor R and Csányi G 2010 *Phys. Rev. Lett.* **104** 136403
[24] Bartók A P *et al* 2013 *Phys. Rev.* B **87** 184115
[25] Musil F, Grisafi A, Bartók A P, Ortner C, Csányi G and Ceriotti M 2021 *Chem. Rev.* **121** 9759
[26] Clark S J, Segall M D, Pickard C J, Hasnip P J, Probert M I J, Refson K and Payne M C 2005 *Z. Kristallogr. Cryst. Mater.* **220** 567
[27] Vanderbilt D 1990 *Phys. Rev.* B **41** 7892
[28] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865
[29] Monkhorst H J and Pack J D 1976 *Phys. Rev.* B **13** 5188
[30] Kunc K and Martin R M 1982 *Phys. Rev. Lett.* **48** 406
[31] Hjorth Larsen A *et al* 2017 *J. Phys.: Condens. Matter* **29** 273002
[32] Mahoney M W and Drineas P 2009 *Proc. Natl Acad. Sci.* **106** 697
[33] Togo A and Tanaka I 2015 *Scr. Mater.* **108** 1
[34] Setyawan W and Curtarolo S 2010 *Comput. Mater. Sci.* **49** 299
[35] Slutsky L J and Garland C W 1957 *Phys. Rev.* **107** 972
[36] Vallin J, Mongy M, Salama K and Beckman O 1964 *J. Appl. Phys.* **35** 1825
[37] Hall J J 1967 *Phys. Rev.* **161** 756
[38] Stathis J H and Bolef D I 1980 *J. Appl. Phys.* **51** 4770
[39] Debernardi A, Alouani M and Dreyssé H 2001 *Phys. Rev.* B **63** 064305
[40] Szlachta W J, Bartók A P and Csányi G 2014 *Phys. Rev.* B **90** 104108
[41] Jiang D, Zhong S, Xiao W, Liu D, Wu M and Liu S 2020 *Int. J. Quantum Chem.* **120** e26231
[42] Zhuang H, Chen M and Carter E A 2016 *Phys. Rev. Appl.* **5** 064021

[43] Böer K W and Pohl U W 2018 *Semiconductor Physics* (Cham: Springer)

[44] Sin'ko G V and Smirnov N A 2002 *J. Phys.: Condens. Matter* **14** 6989

[45] (Available at: https://github.com/ConnorSA/ndsc_tut)

[46] (Available at: https://github.com/libAtoms/QUIP)

[47] Larsen A H *et al* 2017 *J. Phys.: Condens. Matter* **29** 273002

[48] Togo A and Tanaka I 2015 *Scr. Mater.* **108** 1

[49] Tange O 2022 *Gnu Parallel 20220322 ('МариўПолЬ')* (https://doi.org/10.5281/zenodo.6377950)