First-principles database for fitting a machine-learning silicon inter-atomic force-field

K. Zongo[1], L.K Béland[2], C. Ouellet-Plamondon[1]

[1]*Ecole de Technologie Supérieure, Montréal, Quebec, Canada*
[2]*Department of Mechanical and Materials Engineering, Queen's University, Kingston, Ontario, Canada*

To cite this article:
Zongo, K., Béland, L.K. and Ouellet-Plamondon, C. First-principles database for fitting a machine-learning silicon interatomic force field. MRS Advances 7, 39–47 (2022).(https://doi.org/10.1557/s43580-022-00228-z)

Abstract

Data-driven machine learning has emerged to address the limitations of traditional methods when modelling interatomic interactions in materials, such as electronic density functional theory (DFT) and semi-empirical potentials. These machine learning frameworks involve mathematical models coupled to quantum mechanical data. In the present article, we focus on the moment tensor potential (MTP) machine learning framework. More specifically, we provide an account of the development of a preliminary MTP for silicon, including details pertaining to the construction of a DFT database.

1 Introduction

Material modeling has become a powerful tool for gaining insight into materials properties and to circumvent obstacles met by experiments. It can also help to guide experimentalists and accelerate materials discovery by providing computationally predicted compounds of potential interest [1, 2, 3]. Models span different time and length scales, allowing one to choose a scale based on desired properties or available computational means. Current materials modeling methods include quantum mechanical approaches such as density functional theory (DFT), semi-empirical methods and machine learning. The latter is a data driven approach introduced to address the limitations of the first two methods such as spatio-temporal and transferability limitations respectively [4, 5]. Machine learning approaches allow for flexible and adaptive force fields for material research and simulation [6]. Indeed, their computational cost are typically several orders of magnitude lower than that of DFT [6]. They provide interpolative predictions of properties of a new atomic configuration using reference data (atomic configuration, properties) usually carefully prepared with DFT [6, 7]. The approach maintains near-chemical accuracy and is more versatile than semi-empirical potentials and faster than both DFT and even than some semi-empirical potentials [7, 8]. The machine learning method is made of three part, namely its underlying mathematical model, its database and its implementation [9]. This work is devoted to the second part, the database. Thus, in the present paper we describe the quantum mechanical database preparation in the context of clay minerals modeling using machine learning. Our long-term goal is to develop a machine learning force field that will enable a better description of interatomic forces within clay systems in the context of radioactive waste sealing and environmental applications. Our strategy is to build the database in a step-wise manner starting by one of the main constituent elements of clay systems. Namely, we start at a simple level with a single chemical element (silicon) and we will keep adding other chemical elements.

In this article, we will describe how our silicon atomic configurations database was constructed. The database is assessed by carrying out a preliminary implementation of a machine learning interatomic potential for silicon based on the moment tensor potential model [10].

2 Methods

2 .1. Database generation

Databases are one of the main components that enable machine learning potential to address the challenges faced by DFT and semi-empirical potentials. Recall that the ultimate goal of machine learning interatomic potential (ML-IAP) is to accurately handle previously unknown scenarios and structures [9]. In fact different atomic environments may occur during manufacturing and practical applications of materials [9]. Thus, truly predictive machine learning potential not only depend on the ML model but also need to be designed based on a highly diversified database either from experiments or *ab initio* calculations [11, 12]. Although we are limited to a certain number of *ab initio* calculations that can be conducted, our database is designed by sampling most of the possible configurations and phases in which the material may chemically exist. First, we include all the possible phases including solid and liquid of the material of interest. Second for solid phase we take into account all the reported crystallographic and amorphous phases. Up to 13 different crystal structures of silicon with qualitatively different bonding have been reported in the literature[13]. Thus, we build most of the configurations from experimental parameters such as lattice constant, bond length, bond angle and atomic coordinates found in the literature [14, 15, 16, 17, 18, 19, 20]. However all these experimental parameters were not available for some polymorphs including face centered cubic (FCC), body centered cubic silicon(BCC) and ST12 as well as non crystalline structures such as grain boundary. In these particular situations we use lattice constant and atomic coordinates from other DFT calculations [21, 22, 23, 24] as starting point and then compute new properties with our own optimal DFT parameters such as kinetic cut-off energy and k points. Regardless of theses parameters the re-computation is necessary as most of the configurations from others DFT calculations in the literature except the lattice constant and atomic coordinates usually lack useful properties such as energies, forces and stress which are the main ingredient of our machine learning model. We begin by computing the ground state for each crystallographic phase and then apply strain in six different modes following the Voigt notation as detailed in [25]. In other words, we apply shear, tensile and compression from the ground state configuration with strain value ranging from -10% to 10% at 2% intervals. This strain increment will ensure to get different configurations with respect to atomic positions, bonding and possibly coordination numbers. In the other words, we ensure that atomic coordinates and bonding information are highly sensitive to the applied strain and will be significantly different between two consecutive deformations.

In addition, the ground state configuration of each solid phase is replicated sufficiently and oriented before we introduce defects including point, line, planar and 3-dimensional defects. As interactions such as defect wave-function overlap, magnetic interactions, and strain field may affect the defect calculation in the supercell method using periodic boundary conditions [26], we employ replicated supercell of the diamond structure, 2x2x2 for vacancy, 3x3x3 for divacancy and interstitial and 4x4x4 for vacancy clusters as detailed in table 1 below. Configurations containing line defects such as dislocation were generated using three different replicated supercell 4x3x1, 5x4x1 and 6x5x1 in which quadrupole was introduced in order to cancel out long-range elastic strain field [27]. Both edge and screw dislocations as well as related planar defects such as generalized stacking faults(GFS) were considered in our database. The GSF is associated with dislocation properties that govern plastic deformation and fracture of crystalline solids. Here, the slab method [28] was used in two directions $[\bar{1}\,10]$ and $[11\bar{2}\,]$ representing respectively full dislocation and partial dislocation directions in the (111) plane which is the natural cleavage plane of silicon. As stacking faults have very short-range interactions (one or two atomic layer distance) [29, 30, 31], we considered an orthorhombic supercell of the diamond structure with lengths $a\frac{\sqrt{2}}{2}$, $3a\sqrt{3}$ and $a\frac{\sqrt{6}}{2}$ (a is the lattice parameter) corresponding to $\frac{1}{2}[\bar{1}\,10]$ x $3[111]$ x $\frac{1}{2}[11\bar{2}\,]$ respectively. In total the supercell consist of 36 atoms and containing nine bilayers in the [111] direction. We cut the crystal following the well known two distinct ways namely the shuffle and glide cut [32] as illustrated in figure 5 in the supplemental information. The GFS were introduced within a bilayer and between bilayers for glide and shuffle cut respectively. A vacuum of 20 Å is created on each side of the (111) faces to avoid spurious interaction between the slab and its periodic images in the calculation. During the calculation, atoms were only allowed to relaxed perpendicularly to the plane of the cut. Another important planar defect that we considered in our database is the external interface defects (surfaces). Surfaces are important in a wide range of technological field including interfaces, catalysis, semiconductor fabrication and many others [33, 34]. They are also inherent in the study and the understanding of physico-chemical processes at heterophase interfaces such as gas-liquid [35, 36], solid-liquid [37, 38, 39] and solid-gas interfaces [40]. Thus, we include (100), (110) and (111) of diamond surfaces along with (1x1), (2x1) and (2x2) reconstructions of (100), (110) and (111). Here again, we used the slab model with 8 atomic layers for (100) and (110) and 12 atomic layers for (111) in both relaxation and reconstruction. A vacuum of 20 Å is then added in the direction normal to the considered surface.

Liquid and amorphous configurations are also generated using replicated unit cells. We run an NVT *ab initio* molecular dynamics simulation (AIMD) at 300 K and 1x, 1.5x, and 2x of the melting point with a time step of 1 fs and equilibration of 20000 time steps. Because of compute resource limitations, AIMD amorphization via the melting-quenching method was not possible. Instead, we generate amorphous configurations in two steps. First, we use the semi-classical Stillinger-Weber [41] and Tersoff [42] potentials to melt and quench the silicon crystal. Second, the resulting structures are refined via DFT geometry optimization thereby improving local geometry distortions and bringing each configuration to the nearest energy minimum. Note that only atomic coordinates are optimized so that the density is not altered. In addition we add some amorphous structures from [43, 44]. These structures only have atomic coordinates and cell parameters so we recalculate associated energies, forces and stress with our own kinetic cut-off energy and k points. We also add disorders structures to the database through random displacements of the optimized bulk unit cell of each crystallographic phase. We randomly displace atoms of up to 0.6 Å in certain cases in two different modes. In the first mode, atoms are randomly displaced in each of the x, y, and z Cartesian directions while their bulk positions are maintained in the other two directions. The second mode correspond to an isotropic random displacement in which all the atoms are displaced with same magnitude in three Cartesian directions.The configurations including different phases of silicon are detailed in the table 1 in below.

Table 1: Content of the DFT database used for fitting the MTP. The first column indicates which structure types were included in the database. The second column indicates the dimension of the simulation box, in terms of unit cell length–note that a mix of primitive and conventional unit were used. The third column depicts the number of atoms in the unit cell. The fourth column gives the total number of configurations for each structure type that were included in the database. Note that 25 % of the structures were reserved as a validation set–they were not included in the training set.

Content	Replication	Atom/cell	Total
Diamond structure			
Bulk	1x1x1	2	1
Bulk deformations (tensile, compression, shear)	1x1x1	2	177
Vacancies	2x2x2 & 3x3x3	63, 213, 214, 215	58
Divacancies	3x3x3	214	44
Vacancy cluster (V_3, V_4, V_5, V_{17})	4x4x4	507, 508, 509, 495	20
Interstitials $I_1 - I_4$	3x3x3	217, 218, 219, 220	81
Diamond surfaces (100), (110), (111)	$\frac{\sqrt{2}}{2}$x$\frac{\sqrt{2}}{2}$x4, $\frac{\sqrt{2}}{2}$x1x4$\sqrt{2}$, $\frac{\sqrt{2}}{2}$x$\frac{\sqrt{2}}{2}$x4$\sqrt{3}$	8, 12, 16	41
Surface reconstruction (100), (110), (111)	$\sqrt{2}$x$\frac{\sqrt{2}}{2}$x4, $\frac{\sqrt{2}}{2}$x1x4$\sqrt{2}$, $\frac{\sqrt{2}}{2}$x1x4$\sqrt{3}$	16, 24	9
Stacking fault	$\frac{\sqrt{2}}{2}$x3$\sqrt{3}$x$\frac{\sqrt{6}}{2}$	36	168
Dislocation(screw)	5$\frac{\sqrt{6}}{2}$x4$\sqrt{3}$x$\frac{\sqrt{2}}{2}$	240	6
Grain boundaries(tilt & twist)	$\sqrt{5}$x3$\sqrt{5}$x1, $\sqrt{2}$x4$\sqrt{2}$x$\frac{\sqrt{2}}{2}$, 2x4x2	44, 86, 119, 130, 160	8
MD (300 K)	2x2x2	64	25
Random displacement	2x2x2, 3x3x3, 4x4x4	32,64, 192, 216, 512	19
Disordered phases			
Amorphous	2x2x2, 3x3x3, 4x4x4, 5x5x5	64, 216,512, 1000	42
MD-liquids (1770 K, 2530 K, 3370 K)	2x2x2	64	48
β-S_n			
Bulk	1x1x1	2	1
Bulk deformations	1x1x1	2	105
Random displacement	2x2x2	54	2
Simple hexagonal			
Bulk	1x1x1	1	1
Bulk deformations	1x1x1	1	105
Random displacement	2x2x2	64	2
Hexagonal closed packed			
Bulk	1x1x1	2	1
Bulk deformations	1x1x1	2	105
Random displacement	2x2x2	54	2
Hexagonal diamond			
Bulk	1x1x1	4	1
Bulk deformations	1x1x1	4	105
Random displacement	2x2x2	108	2
FCC			
Bulk	1x1x1	4	1
Bulk deformations	1x1x1	1	9
Random displacement	2x2x2	108	2
BCC			
Bulk	1x1x1	2	1
Bulk deformations	1x1x1	1	9
Random displacement	2x2x2	54	2

Content	Replication	Atom/cell	Total
BC8			
Bulk	1x1x1	8	1
Bulk deformations	1x1x1	8	66
Random displacement	2x2x2	64	2
ST12			
Bulk	1x1x1	12	1
Bulk deformations	1x1x1	12	66
Random displacement	2x2x2	96	2
Clathrate Structure Si_{24}			
Bulk	1x1x1	24	1
Bulk deformations	1x1x1	24	56
Random displacement	2x2x2	192	2
Clathrate Structure Si_{46}			
Bulk	1x1x1	46	1
Bulk deformations	1x1x1	46	56
Random displacement	2x2x2	46	2
Clathrate Structure Si_{136}			
Bulk	1x1x1	136	1
Bulk deformations	1x1x1	136	56
Random displacement	2x2x2	136	2
Planar and puckered silicene			
Bulk	1x1x1	2	2
Bulk deformations	1x1x1	2	15
Grand total			1427

2.2. Computational details

We perform all the calculations using the Quantum Espresso package [45]. The interactions between valence electrons and ionic cores are described with the projector augmented wave pseudopotential (PAW) [46] while the exchange-correlation interaction energy is calculated with the generalized gradient approximation (GGA) of Perdew, Burke, and Ernzerhof(PBE) [47]. The energy criterion for convergence of self-consistent field (SCF) iterations was set to 10^{-10} eV. Convergence criteria for geometry optimization including atomic relaxation were set to 10^{-6} eV and 10^{-5} eV/Å for energy and forces, respectively. Before calculation on each crystallographic phase, we performed a convergence test for both kinetic cut-off energy and k-point. Each of these parameters is varied until the total energy become stable and the change in total energy become less than 10^{-4} eV. Thus, a value of 884 eV was set as the kinetic cut-off energy for the expansion of the wave function into plane waves. We employ Monkhorst-Pack grid [48] for Brillouin zone sampling and various k-point are used for the database generation including 8x8x8 for the ordinary phase.

In this work, we describe the potential energy surface (PES) by means of the moment tensor potential (MTP) [10] . Similarly to others ML-IAP, this interatomic interaction model assume that the PES is partitioned into a sum of individual atomic energies that only depend on their local environment delimited by some cut-off radius R_c . Thus, given an atomic configuration in which long-range interaction can be neglected, the total energy is expressed as a sum of the contributions of each of the n atoms:

$$E_{Total} = \sum_{i=1}^{n} E_i = \sum_{i=1}^{n} V_{local}(u_i) \tag{1}$$

in which u_i is a collection of relative distance r_{ij} of atoms j to atom i within the sphere or circle of radius R_c. The potential energy V_{local} of an atom i is then expanded as a linear combination of basis function B_β.

$$V_{local} = \sum_{\beta} c_\beta B_\beta \tag{2}$$

The basis functions B_β depend on the atomic environment of the atom i and are determined from the scalar tensorial contraction of the local atomic environment descriptors $M_{\mu,\nu}$ designed to account for all the physical symmetries.

$$M_{\mu,\nu}(r_{ij},\tau_i,\tau_j) = \sum_{j} f_\mu(|r_{ij}|,\tau_i,\tau_j)\, \underbrace{r_{ij} \otimes \cdots\cdots\cdots \otimes r_{ij}}_{\nu \text{ times}} \tag{3}$$

where τ_i and τ_j depict the type of atoms i and j respectively. $M_{\mu,\nu}$ is a tensor of rank ν enclosing the radial distribution f_μ and the angular information $r_{ij} \otimes \cdots\cdots\cdots \otimes r_{ij}$ of the neighborhood of the central atom i. The radial part of the descriptor is further expanded in the radial basis functions $Q^{(\alpha)}$ which are expressed through the Chebyshev polynomials in the current implementation of MTP [49] .

$$f_{\mu},(|r_{ij}|,\tau_i,\tau_j) = \sum_{\alpha} c^{(\alpha)}_{\mu,\tau_i,\tau_j} Q^{(\alpha)}(|r_{ij}|) \tag{4}$$

$c^{(\alpha)}_{\mu,\tau_i,\tau_j}$ along with c_β are the MTP parameters that are computed during the fitting. The model has shown good performance and computational efficiency in various problems involving different materials [4, 50, 51, 52, 53, 54].

2.3. Reliability of the database

In the context of interatomic potential, machine learning consist of several key steps including the generation of reference data, training of the machine learning model and the using of the implemented potential to run molecular dynamic simulation for the targeted property. At each step, a numerical noise is expected. Thus, one need to handle carefully the data involve in each process mainly at the calculation step of the reference data as numerical noise are intrinsic to atomistic calculations [55]. Therefore a large number of basic properties such as lattice constant, bulk modulus, elastic constants, bond length and bond angle are computed and compared with experimental or other DFT calculation taking into account that experimental properties are usually measured at ambient conditions and others DFT calculations may use different parameters and functionals. Point defects formation energies are also compared with the literature. When these properties deviate more than 3% from experimental or theoretical results, we first recheck the structure and if necessary we change our calculations parameters such as kinetic energy cut-off and k-points. As we deal with more than 1000 configurations for each material, we do not record every deviation but table 2 below provide an example on how we do a quick check of the reliability of our data.

Table 2: Comparison between our DFT computed lattice parameters and experimental values
(first row is a and the second row is c)

	DFT	Experimental	Deviation (%)
Diamond structure	5.469	5.435 [56]	0.6256
β-S_n	4.809 2.655	4.69 2.579 [57]	2.53 2.94

2.3. Training details

The training of MTP using n atomic configurations consist of finding the parameters $\{\Theta\}$ of the model by solving the minimization problem as given below.

$$\sum_{i=1}^{n}\Big[w_e(E^{mtp}(x^{(i)},\Theta)-E^{qm}(x^{(i)}))^2+ \\ w_f\sum_{j=1}^{N_a(x^{(i)})}\left|F_j^{mtp}(x^{(i)},\Theta)-F_j^{qm}(x^{(i)})\right|^2 \\ +w_\sigma\left|\sigma^{mtp}(x^{(i)},\Theta)-\sigma^{qm}(x^{(i)})\right|^2\Big]\rightarrow min \quad (1)$$

where w_e, w_f and w_σ are non-negative weight indicating the importance of each property (energy, force, or stress) in the optimization problem; $N_a(x^i)$ is the number of atoms in the configuration $x^{(i)}$. $\{E^{qm}(x^{(i)}), F^{qm}(x^{(i)}), \sigma^{qm}(x^{(i)})\}$ and $\{E^{mtp}(x^{(i)}), F^{mtp}(x^{(i)}), \sigma^{mtp}(x^{(i)})\}$ denote the properties calculated with DFT and MTP respectively. Here we choose the silicon as a prototype material for our first implementation of moment tensor potential model. A total number of 1427 configurations has been used with a split of 68:32 i.e 68% for training and 32% for validation. We use the default setting of the MTP code such as scaling and weighting. Thus a weight of 1, 0.01 and 0.001 were used for energy, forces and stress respectively. We use a MTP functional form of level 26 with four radial basis functions including a total number of 889 free parameters. The cut off and minimum distance were set to R_{cut} = 5.0 Å and R_{min} = 1.5 Å respectively. We set the number of iterations to 2000. The average fitting errors are detailed in table 2 below.

As can be seen from table 3 above, except the energy for which the training and validation errors are closed to each other (1.4 meV atom^{-1} versus 1.9 meV atom^{-1}) , there are significant difference between those errors for forces and stress that can be a sign of underfitting or overfitting. Possible reasons for this include the large number of configurations used in the training and validation and the lack of the variation of MTP level. For instance, we used MTP level 26 in our first test which is not necessary an optimal level . Thus, optimal level including cut off distance as well as radial basis size were supposed to be found by varying simultaneously the MTP level like 18, 20, 22, 24, and the associated parameters (R_{cut} and basis size). Another important point is that MTP model parameters are randomly initialized at the beginning of the training so the training need to be repeated several times in order to quantify the uncertainty of the predictions and to circumvent overfitting issues.

Table 3: Average training and validation errors

	Energy error meV atom^{-1}	Force error meV Å^{-1} (%)	Stress error GPa (%)
Training	1.4	10 (1.82)	0.076 (0.9)
Validation	1.9	22 (0.15)	0.095 (0.17)

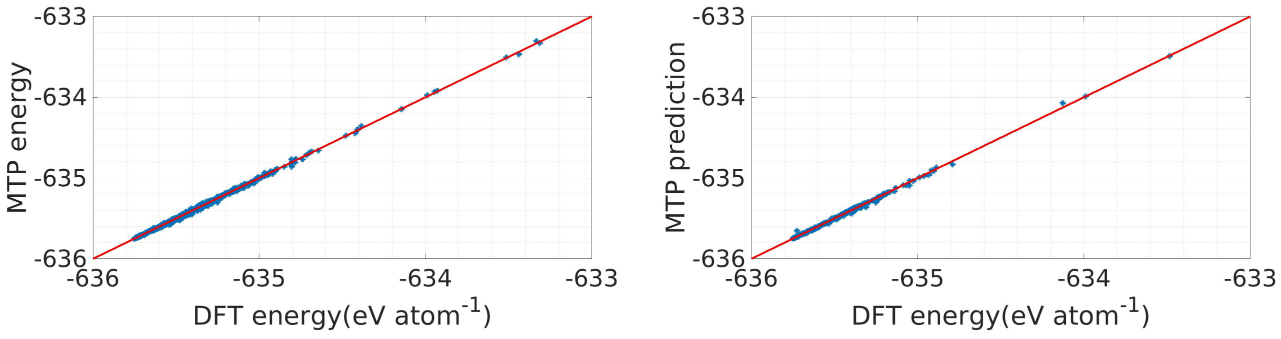
3 Results and discussion

In this section, we present the first set of results obtained by testing the MTP model on silicon. Fig. 1 (a), (b), Fig. 2 (a), (b) and Fig. 3 (a), (b) illustrates the comparison of the DFT-calculated energies, atomic forces ,and stress and those calculated by the moment tensor potential. These plots of energies, force components and stress components were obtained using training set and validation set. As can be seen, all of them show good correlation between DFT and MTP models. The excellent agreement between the MTP models and DFT from the validation set indicates that MTP can make accurate predictions within configuration space that was not explicitly employed in the training sets. We investigate generalized stacking fault (GSF) and points defects diffusion in silicon crystal both relevant to the mechanical properties and structural evolution of solids. As mentioned earlier we considered shuffle and glide cut of diamond crystal. The shuffle cut result was already reported in conference article of the Canadian nuclear society (CNS) [58]. So here, we only report the result of the glide cut of diamond lattice. As shown in figure 4 and figure 5, first our MTP stacking fault energies are in good agreement with our DFT energies. The first peak in figure 5 indicate the unstable stacking fault energy (γ_{USFE}) in the (111) planes along the [112] direction. The DFT calculated value of γ_{USFE} is 1.564 j/m^2 which is comparable to that of figure 22 in [59]. Second, MTP results compare better to DFT than semi-empirical Stillinger-Weber potential [41] and Tersoff potential [42].

The second properties we investigate was the point defect diffusion such as vacancy-, divacancy- and interstitial-type defects. As such calculations are time consuming and very expensive using regular method like nudged elastic band(neb) [60] coupled with density functional theory (DFT), we first manually moved the targeted atom along the considered path and compute the corresponding energies with DFT. Since this is not necessary a minimum energy path, the resulting energy barriers are higher, however the energy profile compare well with those found in the literature [61, 62, 63] as can be seen in the supporting information. Here the MTP reproduce with accuracy DFT energy profiles. Second, we used our implemented silicon MTP potential in LAMMPS to calculated energy barrier associated with the aforementioned point defect diffusion. The computed formation energy E_v, migration energy E_m and activation energy barrier E_a are reported in table 4 below. They compare reasonably to the DFT energies barrier that were obtained using different calculation parameters such as kinetic energy cut off, k points as well as exchange correlation functional and pseudo-potentials.

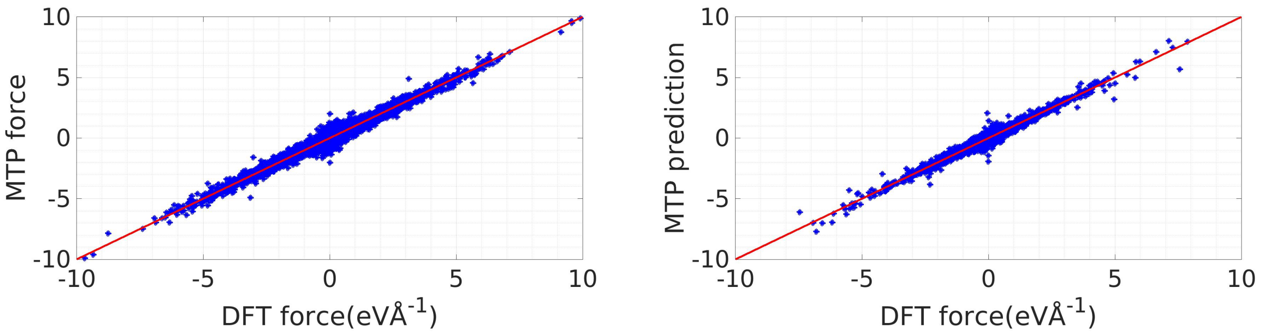
Table 4: Point defects diffusion energies barriers (eV)

	DFT	MTP
Vacancy		
E_v	3.25	3.33
E_m	0.21	0.17
E_a	3.46	3.50
Divacancy(movement)		
E_v	5.45	4.98
E_m	1.36	1.23
E_a	-	6.21
Divacancy(dissociation)		
E_v	5.45	4.98
E_m	2.00	2.07
E_a	-	7.05



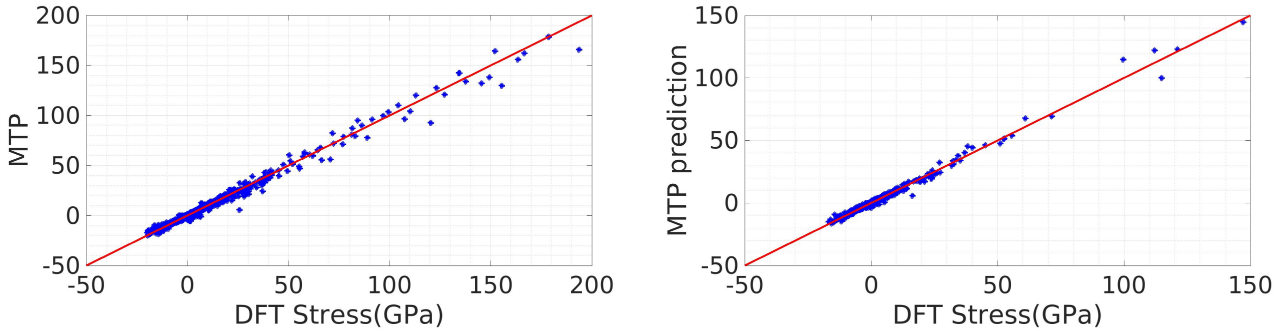
(a) (b)

Figure 1: DFT-computed and MTP-predicted total energies (a) training set- 1062 configurations (b) validation set - 365 configurations



(a) (b)

Figure 2: DFT-computed and MTP-predicted force components (a) training set - 1062 configurations(b) validation set - 365 configurations



(a) (b)

Figure 3: DFT-computed and MTP-predicted stress components (a) training set - 1062 configurations(b) validation set - 365 configurations

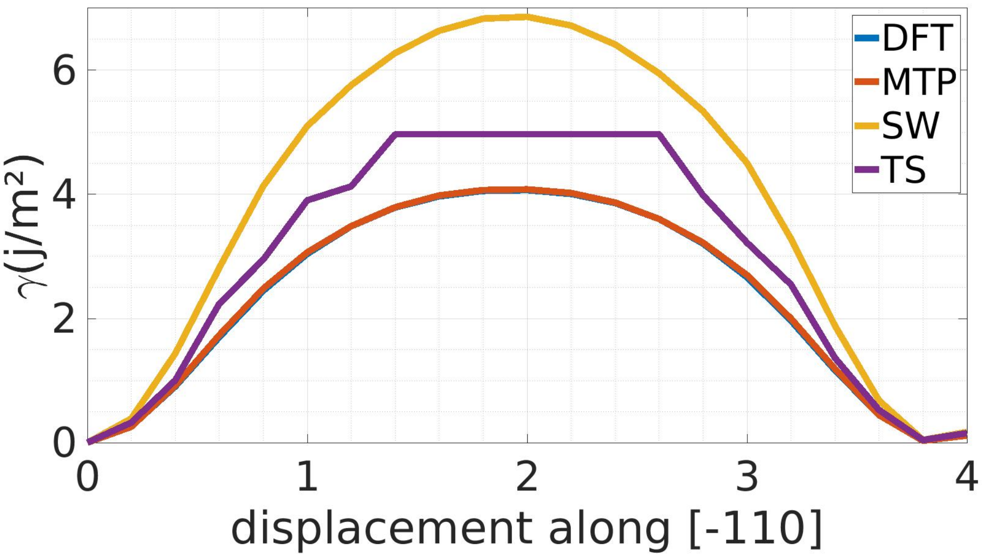


Figure 4: Generalized stacking fault curve

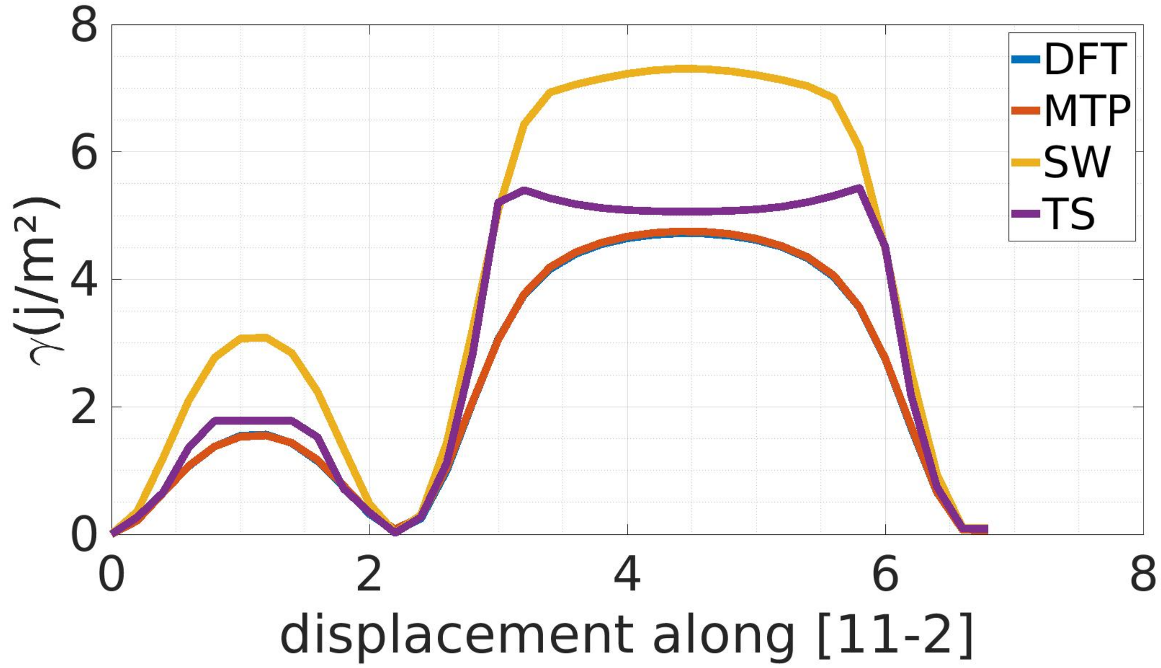


Figure 5: Generalized stacking fault curve

Conclusion and future work

The preliminary results described herein are very encouraging. Our implementation of a MTP-based model of silicon is a good demonstration of the versatility of MTP models. Benchmarks against 476 validation DFT calculations indicate that the MTP potential can correctly describe energies, forces and stresses in a wide variety of scenarios where semi-empirical potentials fail. Future work will include the extension of our database to silica and oxygen as well as the revision of the MTP training methods. We will also investigate point defects diffusion in silicon crystal, the amorphization of silica and the simulation of gas-liquid interface of oxygen using a unified fully implemented MTP potential of the silicon, oxygen and silica.

Acknowledgement

We thank Dr. Yaoting Zhang for insightful discussions. The authors wish to thank Compute Canada for generous allocation of computer resources, the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Nuclear Waste Management Organization (NWMO) for financial support. There is no conflict of interest. Data will be made available upon request.

References

[1] Geoffroy Hautier, Chris Fischer, Virginie Ehrlacher, Anubhav Jain, and Gerbrand Ceder. Data mined ionic substitutions for the discovery of new compounds. *Inorganic chemistry*, 50(2):656–663, 2011.

[2] Yue Liu, Tianlu Zhao, Wangwei Ju, and Siqi Shi. Materials discovery and design using machine learning. *Journal of Materiomics*, 3(3):159–177, 2017.

[3] Evgeny V Podryabinkin, Evgeny V Tikhonov, Alexander V Shapeev, and Artem R Oganov. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. *Physical Review B*, 99(6):064114, 2019.

[4] YX Zuo, C Chen, XG Li, Z Deng, YM Chen, J Behler, G Csanyi, AV Shapeev, AP Thompson, and MA Wood. Performance and cost assessment of machine learning interatomic potentials. *Journal of Physical Chemistry A*, 124(4):731–745, 2020.

[5] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.

[6] Tim Mueller, Aaron Gilad Kusne, and Rampi Ramprasad. Machine learning in materials science: Recent progress and emerging applications. *Reviews in Computational Chemistry*, 29:186–273, 2016.

[7] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[8] Venkatesh Botu, Rohit Batra, James Chapman, and Rampi Ramprasad. Machine learning force fields: construction, validation, and outlook. *The Journal of Physical Chemistry C*, 121(1):511–522, 2017.

[9] Volker L Deringer, Miguel A Caro, and Gábor Csányi. Machine learning interatomic potentials as emerging tools for materials science. *Advanced Materials*, 31(46):1902765, 2019.

[10] Alexander V Shapeev. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Modeling & Simulation*, 14(3):1153–1173, 2016.

[11] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, 3(1):1–13, 2017.

[12] Konstantin Gubaev, Evgeny V Podryabinkin, and Alexander V Shapeev. Machine learning of molecular properties: Locality and active learning. *The Journal of chemical physics*, 148(24):241727, 2018.

[13] S Zhao, EN Hahn, B Kad, Bruce A Remington, CE Wehrenberg, Eduardo Marcial Bringa, and Marc A Meyers. Amorphization and nanocrystallization of silicon under shock compression. *Acta Materialia*, 103:519–533, 2016.

[14] Kristin Persson. Materials data on sio2 (sg:15) by materials project, 11 2014. An optional note.

[15] Michael J Mehl, David Hicks, Cormac Toher, Ohad Levy, Robert M Hanson, Gus Hart, and Stefano Curtarolo. The aflow library of crystallographic prototypes: part 1. *Computational Materials Science*, 136:S1–S828, 2017.

[16] David Hicks, Michael J Mehl, Eric Gossett, Cormac Toher, Ohad Levy, Robert M Hanson, Gus Hart, and Stefano Curtarolo. The aflow library of crystallographic prototypes: part 2. *Computational Materials Science*, 161:S1–S1011, 2019.

[17] Oleg B Gadzhiev, Stanislav K Ignatov, Mikhail Yu Kulikov, Alexander M Feigin, Alexey G Razuvaev, Peter G Sennikov, and Otto Schrems. Structure, energy, and vibrational frequencies of oxygen allotropes o n (n 6) in the covalently bound and van der waals forms: Ab initio study at the ccsd (t) level. *Journal of chemical theory and computation*, 9(1):247–262, 2013.

[18] Robert T Downs and Michelle Hall-Wallace. The american mineralogist crystal structure database. *American Mineralogist*, 88(1):247–250, 2003.

[19] Hui Zheng, Xiang-Guo Li, Richard Tran, Chi Chen, Matthew Horton, Donald Winston, Kristin Aslaug Persson, and Shyue Ping Ong. Grain boundary properties of elemental metals. *Acta Materialia*, 186:40–49, 2020.

[20] Richard Tran, Zihan Xu, Balachandran Radhakrishnan, Donald Winston, Wenhao Sun, Kristin A Persson, and Shyue Ping Ong. Surface energies of elemental crystals. *Scientific data*, 3(1):1–13, 2016.

[21] Byeong-Joo Lee. A modified embedded atom method interatomic potential for silicon. *Calphad*, 31(1):95–104, 2007.

[22] Cong Li, Cuiping Wang, Jiajia Han, Lihui Yan, Bin Deng, and Xingjun Liu. A comprehensive study of the high-pressure–temperature phase diagram of silicon. *Journal of materials science*, 53(10):7475–7485, 2018.

[23] J Crain, SJ Clark, GJ Ackland, MC Payne, V Milman, PD Hatton, and BJ Reid. Theoretical study of high-density phases of covalent semiconductors. i. ab initio treatment. *Physical Review B*, 49(8):5329, 1994.

[24] Hui Zheng, Xiang-Guo Li, Richard Tran, Chi Chen, Matthew Horton, Donald Winston, Kristin Aslaug Persson, and Shyue Ping Ong. Grain boundary properties of elemental metals. *Acta Materialia*, 186:40–49, 2020.

[25] Maarten De Jong, Wei Chen, Thomas Angsten, Anubhav Jain, Randy Notestine, Anthony Gamst, Marcel Sluiter, Chaitanya Krishna Ande, Sybrand Van Der Zwaag, and Jose J Plata. Charting the complete elastic properties of inorganic crystalline compounds. *Scientific data*, 2(1):1–13, 2015.

[26] Christoph Freysoldt, Blazej Grabowski, Tilmann Hickel, Jörg Neugebauer, Georg Kresse, Anderson Janotti, and Chris G Van de Walle. First-principles calculations for point defects in solids. *Reviews of modern physics*, 86(1):253, 2014.

[27] Emmanuel Clouet. Ab initio models of dislocations. *Handbook of Materials Modeling: Methods: Theory and Modeling*, pages 1503–1524, 2020.

[28] Anuj Goyal, Yangzhong Li, Aleksandr Chernatynskiy, Jay S Jayashankar, Michael C Kautzky, Susan B Sinnott, and Simon R Phillpot. The influence of alloying on the stacking fault energy of gold from density functional theory calculations. *Computational Materials Science*, 188:110236, 2021.

[29] PJH Denteneer and W Van Haeringen. Stacking-fault energies in semiconductors from first-principles calculations. *Journal of Physics C: Solid State Physics*, 20(32):L883, 1987.

[30] Fu-Yang Tian, Nan-Xian Chen, Lorand Delczeg, and Levente Vitos. Interlayer potentials for fcc (1 1 1) planes of pd–ag random alloys. *Computational materials science*, 63:20–27, 2012.

[31] Wei Li, Song Lu, Qing-Miao Hu, Se Kyun Kwon, Börje Johansson, and Levente Vitos. Generalized stacking fault energies of alloys. *Journal of Physics: Condensed Matter*, 26(26):265005, 2014.

[32] Yu-Min Juan and Efthimios Kaxiras. Generalized stacking fault energy surfaces and dislocation properties of silicon: a first-principles theoretical study. *Philosophical Magazine A*, 74(6):1367–1384, 1996.

[33] Wenhao Sun and Gerbrand Ceder. Efficient creation and convergence of surface slabs. *Surface Science*, 617:53–59, 2013.

[34] David Sholl and Janice A Steckel. *Density functional theory: a practical introduction*. John Wiley & Sons, 2011.

[35] Aziz Ghoufi, Patrice Malfreyt, and Dominic J Tildesley. Computer modelling of the surface tension of the gas–liquid and liquid–liquid interface. *Chemical Society Reviews*, 45(5):1387–1409, 2016.

[36] Jean-Claude Neyt, Aurelie Wender, Veronique Lachet, and Patrice Malfreyt. Prediction of the temperature dependence of the surface tension of so2, n2, o2, and ar by monte carlo molecular simulations. *The Journal of Physical Chemistry B*, 115(30):9421–9430, 2011.

[37] Pascale Geysermans, D Gorse, and V Pontikis. Molecular dynamics study of the solid–liquid interface. *The Journal of Chemical Physics*, 113(15):6382–6389, 2000.

[38] Roland Šolc, Martin H Gerzabek, Hans Lischka, and Daniel Tunega. Wettability of kaolinite (001) surfaces—molecular dynamic study. *Geoderma*, 169:47–54, 2011.

[39] Gianfranco Ulian, Daniele Moro, and Giovanni Valdrè. Dft simulation of the water molecule interaction with the (00l) surface of montmorillonite. *Minerals*, 11(5):501, 2021.

[40] Zhi Liang, William Evans, and Pawel Keblinski. Equilibrium and nonequilibrium molecular dynamics simulations of thermal conductance at solid-gas interfaces. *Physical Review E*, 87(2):022119, 2013.

[41] Frank H Stillinger and Thomas A Weber. Computer simulation of local order in condensed phases of silicon. *Physical review B*, 31(8):5262, 1985.

[42] Jerry Tersoff. New empirical approach for the structure and energy of covalent systems. *Physical review B*, 37(12):6991, 1988.

[43] Gerard T Barkema and Normand Mousseau. High-quality continuous random networks. *Physical Review B*, 62(8):4985, 2000.

[44] Volker L Deringer, Noam Bernstein, Albert P Bartók, Matthew J Cliffe, Rachel N Kerber, Lauren E Marbella, Clare P Grey, Stephen R Elliott, and Gábor Csányi. Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics. *The journal of physical chemistry letters*, 9(11):2879–2885, 2018.

[45] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, and Ismaila Dabo. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *Journal of physics: Condensed matter*, 21(39):395502, 2009.

[46] Peter E Blöchl. Projector augmented-wave method. *Physical review B*, 50(24):17953, 1994.

[47] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.

[48] Hendrik J Monkhorst and James D Pack. Special points for brillouin-zone integrations. *Physical review B*, 13(12):5188, 1976.

[49] Ivan S Novikov, Konstantin Gubaev, Evgeny Podryabinkin, and Alexander V Shapeev. The mlip package: Moment tensor potentials with mpi and active learning. *Machine Learning: Science and Technology*, 2020.

[50] Bohayra Mortazavi, Evgeny V Podryabinkin, Stephan Roche, Timon Rabczuk, Xiaoying Zhuang, and Alexander V Shapeev. Machine-learning interatomic potentials enable first-principles multiscale modeling of lattice thermal conductivity in graphene/borophene heterostructures. *Materials Horizons*, 7(9):2359–2367, 2020.

[51] Andre Lomaka and Toomas Tamm. Linearization of moment tensor potentials for multicomponent systems with a preliminary assessment for short-range interaction energy in water dimer and trimer. *The Journal of chemical physics*, 152(16):164115, 2020.

[52] Ivan S Novikov, Yury V Suleimanov, and Alexander V Shapeev. Automated calculation of thermal rate coefficients using ring polymer molecular dynamics and machine-learning interatomic potentials with active learning. *Physical Chemistry Chemical Physics*, 20(46):29503–29512, 2018.

[53] Konstantin Gubaev, Evgeny V Podryabinkin, Gus LW Hart, and Alexander V Shapeev. Accelerating high-throughput searches for new alloys with active learning of interatomic potentials. *Computational Materials Science*, 156:148–156, 2019.

[54] II Novoselov, AV Yanilkin, AV Shapeev, and EV Podryabinkin. Moment tensor potentials as a promising tool to study diffusion processes. *Computational Materials Science*, 164:46–56, 2019.

[55] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

[56] John Francis Cannon. Behavior of the elements at high pressures. *Journal of Physical and Chemical Reference Data*, 3(3):781–824, 1974.

[57] Jing Zhu Hu, Larry D Merkle, Carmen S Menoni, and Ian L Spain. Crystal data for high-pressure phases of silicon. *Physical Review B*, 34(7):4679, 1986.

[58] K. Zongo, L.K Béland, and C. Ouellet-Plamondon. Improving atom-scale models of clay minerals using machine learning. *Canadian Nuclear Society*, 2021.

[59] Albert P Bartók, James Kermode, Noam Bernstein, and Gábor Csányi. Machine learning a general-purpose interatomic potential for silicon. *Physical Review X*, 8(4):041048, 2018.

[60] Graeme Henkelman, Blas P Uberuaga, and Hannes Jónsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of chemical physics*, 113(22):9901–9904, 2000.

[61] Gyeong S Hwang and William A Goddard III. Diffusion and dissociation of neutral divacancies in crystalline silicon. *Physical Review B*, 65(23):233205, 2002.

[62] Yaojun A Du, Stephen A Barr, Kaden RA Hazzard, Thomas J Lenosky, Richard G Hennig, and John W Wilkins. Fast diffusion mechanism of silicon tri-interstitial defects. *Physical Review B*, 72(24):241306, 2005.

[63] Fedwa El-Mellouhi, Normand Mousseau, and Pablo Ordejón. Sampling the diffusion paths of a neutral vacancy in silicon with quantum mechanical calculations. *Physical Review B*, 70(20):205202, 2004.