

# DPK: Deep Neural Network Approximation of the First Piola-Kirchhoff Stress

Tianyi Hu<sup>1,2</sup>, Jerry Zhijian Yang<sup>2,3,1</sup> and Cheng Yuan<sup>2,3,\*</sup>

<sup>1</sup> *Institute of Artificial Intelligence, School of Computer Science, Wuhan University, Wuhan, Hubei 430072, China*

<sup>2</sup> *School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China*

<sup>3</sup> *Hubei Key Laboratory of Computational Science, Wuhan University, Wuhan, Hubei 430072, China*

Received 14 June 2022; Accepted (in revised version) 9 November 2022

---

**Abstract.** This paper presents a specific network architecture for approximation of the first Piola-Kirchhoff stress. The neural network enables us to construct the constitutive relation based on both macroscopic observations and atomistic simulation data. In contrast to traditional deep learning models, this architecture is intrinsic symmetric, guarantees the frame-indifference and material-symmetry of stress. Specifically, we build the approximation network inspired by the Cauchy-Born rule and virial stress formula. Several numerical results and theory analyses are presented to illustrate the learnability and effectiveness of our network.

**AMS subject classifications:** 68T07, 65Z05, 41A29

**Key words:** Piola-Kirchhoff stress, deep neural networks, Cauchy-Born rule.

---

## 1 Introduction

Atomistic-based constitutive relation has played a critical role in multiscale modeling. With a proper description of macroscopic stress in the form of atomistic information, we can couple models based on continuum mechanics with models at microscopic scale [1], such as molecular dynamics (MD) and molecular statics (MS). In the last few decades, many research on the formulation [2–6], application [7–10] and simplification [11, 12] of various kinds of atomistic stress has been done. Among these, a well known and widely used atomistic-based stress for equilibrium and homogeneous system is the virial stress, which is firstly studied by Clausius in 1870 [13]. Formally, the virial stress is a sum of

---

\*Corresponding author.

Emails: hutianyi@whu.edu.cn (T. Hu), zjyang.math@whu.edu.cn (J. Yang), yuancheng@whu.edu.cn (C. Yuan)

the multiplications of atoms positions and force, which can be easily computed in the molecular simulations. In practical multiscale modeling, however, we need to repeat this interatomic calculation in each coarse-grained patch [9, 10]. As a consequence, an easier-to-compute deformation-stress relation with atomic-level accuracy would make the multiscale modeling even more efficient.

On the other hand, with recent development of deep learning (DL) in traditional artificial intelligence (AI) domains, such as face recognition in computational visual (CV) and text classification in natural language processing (NLP), a lot of successes has also been achieved in the application of DL to scientific computation [14], e.g., using deep learning in speeding up the MD simulations [15, 16], solving differential equations [17–21], sampling equilibrium states in statistic mechanics [22], predicting multiphase flow [23], and identifying the constitutive law [24]. A significant difference in the utilization of DL in scientific computation from traditional AI problems is that, various constraints informed by different physical laws or boundary conditions must be considered. As a result, a physics-informed network is often wanted for the satisfaction of physical constraint. Furthermore, with a neural network enhanced by some physical knowledge, we can use less data in the training procedure.

In this article, we would build a special physics-informed network architecture for representation of the Piola-Kirchhoff stress (PK stress) in Lagrangian coordinates. Different from several existing works on the learning of constitutive relations with direct or indirect observed data by continuum model [25–28], we learn the PK stress directly from atomistic data driven by virial formulation, which supply the learning stress with an atomic-level accuracy. Furthermore, by constructing the network with symmetric operations in the crystallographic point symmetry group, this stress network is intrinsic symmetric, in the sense that the frame-indifference and material-symmetry of stress is exactly satisfied. In summary, our main contributions are as follows:

- We propose a novel deep neural network (DPK), for the representation of PK stress.
- The DPK stress is intrinsic symmetric and can be trained by virial stress.
- Several numerical results and theory analysis are given to illustrate the learnability, transferability and effectiveness of DPK.

The rest of this paper is organized as follows. In Section 2 we first review the PK stress and virial formulation, then propose the DPK stress and its simplification version. Several theorems on the symmetry, approximation and learnability will also be presented. After that, some numerical results will be shown in Section 3 to demonstrate the effectiveness and transferability of the simplified DPK stress. Finally in Section 4 we will give our main conclusions.

## 2 Symmetric neural networks for PK stress

In continuum thermodynamics, for simple elastic material the first PK stress is defined as the derivative of the Helmholtz free energy density  $\Psi(\mathbf{F}, T)$  with respect to the deformation tensor  $\mathbf{F}$ :

$$\mathbf{P}(\mathbf{F}, T) = \frac{\partial \Psi(\mathbf{F}, T)}{\partial \mathbf{F}}. \quad (2.1)$$

At zero temperature, the free energy density in above definition should be replaced as the strain energy density [29]:

$$\mathbf{P}(\mathbf{F}) = \frac{\partial W(\mathbf{F})}{\partial \mathbf{F}}. \quad (2.2)$$

Two important constraints should be satisfied by these relations [30]:

$$\mathbf{P}(\mathbf{QF}) = \mathbf{QP}(\mathbf{F}), \quad \forall \mathbf{Q} \in SO(3), \quad (2.3a)$$

$$\mathbf{P}(\mathbf{F}) = \mathbf{P}(\mathbf{FH})\mathbf{H}^T, \quad \forall \mathbf{H} \in G, \quad (2.3b)$$

where  $SO(3)$  stands for the proper orthogonal group and  $G$  refers to the material symmetry group. The first constraint (2.3a) is due to the principle of material frame-indifference, meaning that the constitutive relations should be invariant when the frame is changed. The second constraint (2.3b) results from the material symmetry, which requires that the response of the material should be unaltered if symmetric transformation was applied to reference configuration before the deformation. The Group consists of such symmetric transformations is defined as the material symmetry group  $G$ .

In the following sections, we will focus on the construction of deep symmetric network  $\mathbf{P}_\theta(\mathbf{F})$  to represent the first PK stress (DPK). By symmetric we mean the network can exactly satisfy the two constraints (2.3a) and (2.3b).

### 2.1 Cauchy-Born rule and virial stress

In order to design an intrinsic symmetric deep network to learn the stress, we first review the definition of PK stress in atomic level and verify the constraints above. Consider a material consists of  $N$  atoms, of which we denote the  $i$ -th atom's reference position, deformed position and force as  $\mathbf{X}_i$ ,  $\mathbf{x}_i$  and  $\mathbf{f}_i$  respectively. For both pair-wise interatomic potential and multi-body potential system (e.g., embedded atom method, EAM), the force  $\mathbf{f}_i$  can be decomposed as the sum of pair force form:

$$\mathbf{f}_i = \sum_{j \neq i} \mathbf{f}_{ij}, \quad (2.4)$$

where the pair force  $\mathbf{f}_{ij}$  is defined as the minus derivative of total potential  $V$  with respect to  $\mathbf{r}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ :

$$\mathbf{f}_{ij} = -\frac{\partial V(r_{12}, \dots, r_{ij}, \dots)}{\partial \mathbf{r}_{ij}} = -\frac{\partial V(r_{12}, \dots, r_{ij}, \dots)}{\partial r_{ij}} \frac{\mathbf{r}_{ij}}{r_{ij}}. \quad (2.5)$$

Here we write the norm of  $\mathbf{r}_{ij}$  as  $r_{ij}$ . According to the Cauchy-Born rule, when a small deformation  $\mathbf{F}$  is applied to this system at zero temperature, the atoms will move uniformly to the deformed states  $\mathbf{x} = \mathbf{F}\mathbf{X}$ . Denote  $\mathbf{X}_{ij} = \mathbf{X}_i - \mathbf{X}_j$  and let  $\Omega$  be the system volume in reference configuration, the atomistic PK stress of current system can be calculated from definition (2.2):

$$\mathbf{P}(\mathbf{F}) = \frac{1}{\Omega} \frac{\partial V(|\mathbf{F}\mathbf{X}_{12}|, \dots, |\mathbf{F}\mathbf{X}_{ij}|, \dots)}{\partial \mathbf{F}} = -\frac{1}{2\Omega} \sum_i \sum_{j \neq i} \mathbf{f}_{ij} \otimes \mathbf{X}_{ij}. \quad (2.6)$$

It is worth mentioning that in (2.5), the potential  $V$  depends only on  $C_n^2 = \frac{N(N-1)}{2}$  pair distance  $r_{ij, i < j}$  with  $r_{ji} = r_{ij}$ . An alternative expression of the potential is:

$$\tilde{V}(r_{12}, r_{21}, \dots, r_{ij}, r_{ji}, \dots) = V\left(\frac{(r_{12} + r_{21})}{2}, \dots, \frac{(r_{ij} + r_{ji})}{2}, \dots\right). \quad (2.7)$$

In this way the stress (2.6) can be formulated as

$$\mathbf{P}(\mathbf{F}) = \frac{1}{\Omega} \sum_{j \neq i} \left[ \frac{\partial \tilde{V}(|\mathbf{F}\mathbf{X}_{12}|, \dots, |\mathbf{F}\mathbf{X}_{ij}|, \dots)}{\partial r_{ij}} \frac{1}{r_{ij}} \mathbf{F}\mathbf{X}_{ij} \otimes \mathbf{X}_{ij} \right], \quad (2.8)$$

where  $r_{ij} = |\mathbf{F}\mathbf{X}_{ij}|$ . An advantage of (2.7) is that we can explicitly express the symmetry of the potential function: for any permutation  $\tau$  of  $I = [1, 2, \dots, N]$ , we can define  $\psi(i, j) = (\tau(i), \tau(j))$  as an one-to-one map of  $I \times I$ . With this notation, the symmetry of the potential can be expressed as

$$\tilde{V}(\dots, r_{\psi(ij)}, \dots) = \tilde{V}(\dots, r_{ij}, \dots).$$

We would use  $V$  (rather than  $\tilde{V}$ ) to stands for the alternative expression in (2.7) hereinafter unless otherwise stated.

The formula (2.6) is also known as the virial stress, which is commonly used in the modeling of homogeneous system in equilibrium. With this expression we can model the macro constitutive relations by using atomistic information. In addition to the ability of computing stress with atomic-level accuracy, formula (2.6) (or equivalently formula (2.8)) is exactly in conformity with the two constraints (2.3a) and (2.3b). More explicitly, we have the following result:

**Theorem 2.1** (Symmetry of virial stress). *For a given crystal system with periodic boundary condition, suppose the potential  $V$  is a symmetric function, let  $\mathbf{P}(\mathbf{F})$  be the stress defined by (2.8), then the constraint (2.3a) holds for any  $\mathbf{Q}$  in  $SO(3)$ , and (2.3b) holds for any  $\mathbf{H}$  in the crystallographic symmetry group (point symmetry group  $G_p$ , [30]).*

*Proof.* According to formula (2.8), for any  $\mathbf{Q}$  in  $SO(3)$ ,

$$\begin{aligned}\mathbf{P}(\mathbf{QF}) &= \frac{1}{\Omega} \sum_{j \neq i} \left[ \frac{\partial V(|\mathbf{QF}\mathbf{X}_{12}|, \dots, |\mathbf{QF}\mathbf{X}_{ij}|, \dots)}{\partial r_{ij}} \frac{1}{r_{ij}} \mathbf{QF}\mathbf{X}_{ij} \otimes \mathbf{X}_{ij} \right] \\ &= \mathbf{Q} \frac{1}{\Omega} \sum_{j \neq i} \left[ \frac{\partial V(|\mathbf{F}\mathbf{X}_{12}|, \dots, |\mathbf{F}\mathbf{X}_{ij}|, \dots)}{\partial r_{ij}} \frac{1}{r_{ij}} \mathbf{F}\mathbf{X}_{ij} \otimes \mathbf{X}_{ij} \right] \\ &= \mathbf{Q}\mathbf{P}(\mathbf{F}).\end{aligned}\quad (2.9)$$

Which validate the constraint (2.3a).

Now consider any transformation  $\mathbf{H}$  in the point symmetry group  $G_p$ . By definition the transformed configuration is invariant with the original one, namely we have a one-to-one map  $\phi_H$  such that  $\mathbf{H}\mathbf{X}_i = \mathbf{X}_{\phi_H(i)}$  holds for any  $i \leq N$ , indicating that  $\phi_H$  is a permutation of the index set  $I = [1, 2, \dots, N]$ . Furthermore we can define the permutation map of  $I \times I$  by  $\psi_H((i, j)) = (\phi_H(i), \phi_H(j))$ , which satisfies that

$$\mathbf{H}\mathbf{X}_{ij} = \mathbf{X}_{\psi_H(ij)}. \quad (2.10)$$

Applying this to right hand side of (2.3b) together with (2.8), we can obtain that

$$\begin{aligned}\mathbf{P}(\mathbf{FH})\mathbf{H}^T &= \frac{1}{\Omega} \sum_{j \neq i} \left[ \frac{\partial V(\dots, |\mathbf{FH}\mathbf{X}_{ij}|, \dots)}{\partial r_{ij}} \frac{1}{|\mathbf{FH}\mathbf{X}_{ij}|} \mathbf{FH}\mathbf{X}_{ij} \otimes \mathbf{X}_{ij} \mathbf{H}^T \right] \\ &= \frac{1}{\Omega} \sum_{j \neq i} \left[ \frac{\partial V(\dots, |\mathbf{F}\mathbf{X}_{\psi_H(ij)}|, \dots)}{\partial r_{ij}} \frac{1}{|\mathbf{F}\mathbf{X}_{\psi_H(ij)}|} \mathbf{F}\mathbf{X}_{\psi_H(ij)} \otimes \mathbf{X}_{\psi_H(ij)} \right] \\ &= \frac{1}{\Omega} \sum_{j \neq i} \left[ \frac{\partial V(\dots, |\mathbf{F}\mathbf{X}_{ij}|, \dots)}{\partial r_{\psi_H(ij)}} \frac{1}{r_{\psi_H(ij)}} \mathbf{F}\mathbf{X}_{\psi_H(ij)} \otimes \mathbf{X}_{\psi_H(ij)} \right] \\ &= \mathbf{P}(\mathbf{F}).\end{aligned}\quad (2.11)$$

The last two equalities are due to the facts that  $V$  is symmetric and  $\psi_H$  is a permutation map of  $I \times I$  respectively. Thus we have proved the second constraint (2.3b) holds for (2.6).  $\square$

Notice that different from the material symmetry group  $G$  mentioned in (2.3b), we restrict the symmetry group in Theorem 2.1 to the point symmetry group  $G_p$ , according to the hypothesis on crystalline solids given by BD Coleman [30]. This restriction is also known as the Neumann's Principle, which states that the symmetry of physical property for crystal must include the point group of the crystal [31].

In the proof of Theorem 2.1, we can see that the symmetry of virial stress results from the intrinsic symmetry of potential function and crystal structure, which leads to the following heuristic network architecture.

## 2.2 DPK: deep neural networks for PK stress

Inspired by the virial stress, we design the deep symmetry network as follows. For a given crystal material, we denote the crystallographic point symmetry group as  $G_p$ . For any finite 3-dimensional trainable weight vectors  $\mathbf{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_N\}$ , the symmetrized weights  $\tilde{\mathbf{W}}$  is defined as the image set of  $G_p$  on  $\mathbf{W}$ :

$$\tilde{\mathbf{W}} = \{\mathbf{H}_i \mathbf{w}_j | \mathbf{H}_i \in G_p, \mathbf{w}_j \in \mathbf{W}\}. \quad (2.12)$$

By construction  $\tilde{\mathbf{W}} = \{\tilde{\mathbf{w}}_1, \dots, \tilde{\mathbf{w}}_M\}$  is an invariant subset of  $G_p$ : for any  $\mathbf{H} \in G_p$ ,  $\mathbf{H}\tilde{\mathbf{W}} = \tilde{\mathbf{W}}$ . Namely  $H$  defines a permutation  $\phi_H$  of  $[1, 2, \dots, M]$  by

$$\tilde{\mathbf{w}}_{\phi_H(i)} = \mathbf{H}\tilde{\mathbf{w}}_i. \quad (2.13)$$

With this we can define the deep symmetry network as follows:

**Definition 2.1 (DPK).** Let  $\mathbf{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_N\}$  be  $N$  trainable 3-dimensional weight vectors and  $\tilde{\mathbf{W}}$  the symmetrized weights defined above, denote  $g(\mathbf{x}; \theta) = h_n \circ a_{n-1} \circ \dots \circ a_1 \circ h_1(\mathbf{x})$  as the classical dense neural network (DNN) with linear functions  $h_i(\mathbf{x}; \theta_i)$  and activation function  $a_i(\mathbf{x})$ , the deep symmetric network for PK stress is defined as

$$\mathbf{P}_{\theta, \mathbf{W}}(\mathbf{F}) = \mathbf{F} \sum_{i=1}^M \frac{\partial f}{\partial x_i} (|\mathbf{F}\tilde{\mathbf{w}}_1|, \dots, |\mathbf{F}\tilde{\mathbf{w}}_M|; \theta) \frac{1}{|\mathbf{F}\tilde{\mathbf{w}}_i|} \tilde{\mathbf{w}}_i \otimes \tilde{\mathbf{w}}_i, \quad (2.14)$$

where

$$f(\mathbf{x}, \theta) = \frac{1}{M!} \sum_{\tau \in S_M} g(\tau(\mathbf{x}); \theta)$$

with  $S_M$  stands for all the permutations of  $[1, 2, \dots, M]$ .

The architecture of this network is illustrated in Fig. 1. At first glance, formula (2.14) is quite similar with the virial stress (2.8) in form. Indeed, we can view the symmetrized trainable weights  $\tilde{\mathbf{W}}$  as configuration coordinates of some virtual atoms, and  $\frac{\partial f}{\partial x_i}$  as virtual force between such atoms. Nevertheless, with the trainable parameters  $\mathbf{W}$  and trainable network  $g$ , there are at least three advantages to learning with such architecture:

- The DPK satisfies the symmetry constraints (2.3a) and (2.3b) exactly.
- The virial stress can be well approximated by DPK stress. Meanwhile, we can learn the DPK stress by using data from both macroscopic observation and atomistic modeling.
- By training DPK with enough data, we may expect to get a virtual system  $\tilde{\mathbf{W}}$  which is much smaller than the atoms system in virial stress, resulting in a faster computation of stress compared with (2.6).

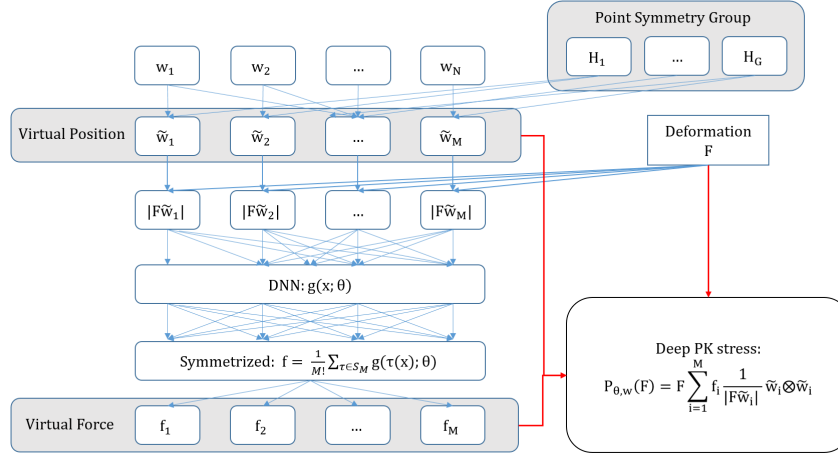


Figure 1: The network architecture of DPK.

The first two statements is proved by following theorems, while the third statement will be illustrated by the subsequent numerical results.

**Theorem 2.2** (Symmetry of DPK). *For a given crystal system with crystallographic symmetry group  $G_p$ , let  $\mathbf{P}_{\theta, \mathbf{W}}(\mathbf{F})$  be the DPK stress defined by (2.14), then the constraint (2.3a) holds for any  $\mathbf{Q}$  in  $SO(3)$ , and (2.3b) holds for any  $\mathbf{H}$  in  $G_p$ .*

*Proof.* The proof is a reprise of the argument in Theorem 2.1. Actually, with the fact that  $\frac{1}{M!} \sum_{\tau \in S_M} g(\tau(\mathbf{x}); \theta)$  is a symmetric function of  $\mathbf{x}$  and formula (2.13), we have

$$\begin{aligned}
 \mathbf{P}_{\theta, \mathbf{W}}(\mathbf{QF}) &= \mathbf{QF} \sum_{i=1}^M \frac{\partial f}{\partial x_i} (|\mathbf{QF}\tilde{\mathbf{w}}_1|, \dots, |\mathbf{QF}\tilde{\mathbf{w}}_M|; \theta) \frac{1}{|\mathbf{QF}\tilde{\mathbf{w}}_i|} \tilde{\mathbf{w}}_i \otimes \tilde{\mathbf{w}}_i \\
 &= \mathbf{QF} \sum_{i=1}^M \frac{\partial f}{\partial x_i} (|\mathbf{F}\tilde{\mathbf{w}}_1|, \dots, |\mathbf{F}\tilde{\mathbf{w}}_M|; \theta) \frac{1}{|\mathbf{F}\tilde{\mathbf{w}}_i|} \tilde{\mathbf{w}}_i \otimes \tilde{\mathbf{w}}_i \\
 &= \mathbf{Q}\mathbf{P}_{\theta, \mathbf{W}}(\mathbf{F})
 \end{aligned} \tag{2.15}$$

and

$$\begin{aligned}
 \mathbf{P}_{\theta, \mathbf{W}}(\mathbf{FH})\mathbf{H}^T &= \mathbf{F} \sum_{i=1}^M \frac{\partial f}{\partial x_i} (|\mathbf{FH}\tilde{\mathbf{w}}_1|, \dots, |\mathbf{FH}\tilde{\mathbf{w}}_M|; \theta) \frac{\mathbf{H}\tilde{\mathbf{w}}_i \otimes \tilde{\mathbf{w}}_i \mathbf{H}^T}{|\mathbf{FH}\tilde{\mathbf{w}}_i|} \\
 &= \mathbf{F} \sum_{i=1}^M \frac{\partial f}{\partial x_i} (|\mathbf{F}\tilde{\mathbf{w}}_{\phi_{\mathbf{H}}(1)}|, \dots, |\mathbf{F}\tilde{\mathbf{w}}_{\phi_{\mathbf{H}}(M)}|; \theta) \frac{\tilde{\mathbf{w}}_{\phi_{\mathbf{H}}(i)} \otimes \tilde{\mathbf{w}}_{\phi_{\mathbf{H}}(i)}}{|\mathbf{F}\tilde{\mathbf{w}}_{\phi_{\mathbf{H}}(i)}|} \\
 &= \mathbf{F} \sum_{i=1}^M \frac{\partial f}{\partial x_{\phi_{\mathbf{H}}(i)}} (|\mathbf{F}\tilde{\mathbf{w}}_1|, \dots, |\mathbf{F}\tilde{\mathbf{w}}_M|; \theta) \frac{\tilde{\mathbf{w}}_{\phi_{\mathbf{H}}(i)} \otimes \tilde{\mathbf{w}}_{\phi_{\mathbf{H}}(i)}}{|\mathbf{F}\tilde{\mathbf{w}}_{\phi_{\mathbf{H}}(i)}|} \\
 &= \mathbf{P}_{\theta, \mathbf{W}}(\mathbf{F}).
 \end{aligned} \tag{2.16}$$

We complete the proof.  $\square$

In addition of symmetry, the DPK also has a good representation ability of the virial stress under certain conditions. To show this, we first introduce the Sobolev space and some existing result on the approximation of DNN, then a theorem on the DPK approximation will be presented.

**Definition 2.2** (Sobolev space). *Let  $n \in \mathbb{N}$  and  $1 \leq p \leq \infty$ , the Sobolev space  $W^{n,p}$  on the  $M$  dimensional region  $\Omega$  is defined as*

$$W^{n,p}(\Omega) = \{f \in L^p(\Omega) \mid \|D^\alpha f\|_{L^p(\Omega)} < \infty, \forall \alpha \in \mathbb{N}^M \text{ with } |\alpha| \leq n\} \quad (2.17)$$

with the norm being defined as

$$\|f\|_{W^{n,p}} = \left( \sum_{0 \leq |\alpha| \leq n} \|D^\alpha f\|_{L^p}^p \right)^{1/p}, \quad 1 \leq p < \infty, \quad (2.18a)$$

$$\|f\|_{W^{n,\infty}} = \max_{0 \leq |\alpha| \leq n} \|D^\alpha f\|_{L^\infty}. \quad (2.18b)$$

**Lemma 2.1** ([32, Corollary 4.2]). *Let  $V \in W^{n,p}((0,1)^M)$  with  $n > 1$ ,  $n \in \mathbb{N}$  and  $1 \leq p \leq \infty$ . If  $\|V\|_{W^{n,p}((0,1)^M)} \leq B$  for some  $B$ , then for any  $0 < \epsilon < \frac{1}{2}$  and  $0 \leq s \leq 1$  there exists a dense neural network  $g(x)$  with ReLU as the activation function and constant  $c(M, n, p, B, s)$  independent of  $\epsilon$  such that*

$$\|V - g\|_{W^{s,p}((0,1)^M)} \leq \epsilon, \quad (2.19)$$

while the number of neurons  $K$  is bounded as

$$K \leq c(M, n, p, B, s) \epsilon^{-M/(n-s)} \log_2^2(\epsilon^{-n/(n-s)}). \quad (2.20)$$

**Theorem 2.3** (Approximation of DPK). *For a finite crystal system consists of  $N$  atoms with symmetric potential  $V$  satisfying  $\|\mathbf{X}_{ij}\|_2 < 1$  for any  $j \neq i$ , if there exists some  $n > 1$ ,  $B < \infty$  such that*

$$\|V\|_{W^{n,\infty}((0,1)^{N(N-1)})} < B,$$

then for any  $\epsilon \in (0, 1/2)$  there exists a DPK function  $\mathbf{P}_0$  such that

$$\|\mathbf{P}_{\text{virial}}(\mathbf{F}) - \mathbf{P}_0(\mathbf{F})\|_F \leq \epsilon, \quad (2.21)$$

if  $\|\mathbf{F}\mathbf{X}_{ij}\|_2 < 1$ . Here  $\|\cdot\|_F$  represents the Frobenius norm and  $P_{\text{virial}}(\mathbf{F})$  is the virial stress defined by (2.8).

*Proof.* To begin with, we denote  $M = N(N-1)$  and  $\tilde{\mathbf{X}}_{(i-1)*N+j} = \mathbf{X}_{ij}$  for  $j \neq i$ . Thus

$$\{\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_M\} = \{\mathbf{X}_{12}, \dots, \mathbf{X}_{N-1,N}\}.$$



By (2.8), the virial stress can be written as

$$\mathbf{P}_{\text{virial}}(\mathbf{F}) = \frac{1}{\Omega} \mathbf{F} \sum_{i=1}^M \frac{\partial V}{\partial x_i} (|\mathbf{F}\tilde{\mathbf{X}}_1|, \dots, |\mathbf{F}\tilde{\mathbf{X}}_M|) \frac{1}{|\mathbf{F}\tilde{\mathbf{X}}_i|} \tilde{\mathbf{X}}_i \otimes \tilde{\mathbf{X}}_i. \quad (2.22)$$

According to Lemma 2.1, given the potential function  $V \in W^{n,\infty}((0,1)^M)$  with norm bounded by  $B$ , for any  $\epsilon \in (0, 1/2)$ , there exists a ReLU activated DNN  $g_0(\mathbf{x}; \theta)$  such that

$$\|V - g_0\|_{W^{1,\infty}} \leq \epsilon \Omega / M. \quad (2.23)$$

Consequently, for the symmetrized function

$$f_0(\mathbf{x}, \theta) = \frac{1}{M!} \sum_{\tau \in S_M} g_0(\tau(\mathbf{x}); \theta),$$

we have

$$\begin{aligned} \|V - f_0\|_{W^{1,\infty}} &= \|V - \frac{1}{M!} \sum_{\tau \in S_M} g_0(\tau(\mathbf{x}); \theta)\|_{W^{1,\infty}} \\ &= \left\| \frac{1}{M!} \sum_{\tau \in S_M} (V(\tau(\mathbf{x}); \theta) - g_0(\tau(\mathbf{x}); \theta)) \right\|_{W^{1,\infty}} \\ &\leq \frac{1}{M!} \sum_{\tau \in S_M} \|V(\tau(\mathbf{x}); \theta) - g_0(\tau(\mathbf{x}); \theta)\|_{W^{1,\infty}} \\ &\leq \epsilon \Omega / M. \end{aligned} \quad (2.24)$$

Consider the special DPK function with  $\mathbf{w}_i = \tilde{\mathbf{X}}_i$  and  $g = \frac{1}{\Omega} g_0$ :

$$\mathbf{P}_0(\mathbf{F}) = \frac{1}{\Omega} \mathbf{F} \sum_{i=1}^M \frac{\partial f_0}{\partial x_i} (|\mathbf{F}\tilde{\mathbf{X}}_1|, \dots, |\mathbf{F}\tilde{\mathbf{X}}_M|; \theta) \frac{1}{|\mathbf{F}\tilde{\mathbf{X}}_i|} \tilde{\mathbf{X}}_i \otimes \tilde{\mathbf{X}}_i, \quad (2.25)$$

compared with the virial stress we can obtain

$$\begin{aligned} \|\mathbf{P}_{\text{virial}}(\mathbf{F}) - \mathbf{P}_0(\mathbf{F})\|_F &= \left\| \frac{1}{\Omega} \mathbf{F} \sum_{i=1}^M \frac{\partial(V - f_0)}{\partial x_i} (|\mathbf{F}\tilde{\mathbf{X}}_1|, \dots, |\mathbf{F}\tilde{\mathbf{X}}_M|; \theta) \frac{1}{|\mathbf{F}\tilde{\mathbf{X}}_i|} \tilde{\mathbf{X}}_i \otimes \tilde{\mathbf{X}}_i \right\|_F \\ &\leq \sum_{i=1}^M \frac{1}{\Omega} \left\| \frac{\mathbf{F}\tilde{\mathbf{X}}_i}{|\mathbf{F}\tilde{\mathbf{X}}_i|} \otimes \tilde{\mathbf{X}}_i \right\|_F \cdot \left| \frac{\partial(V - f_0)}{\partial x_i} \right| \\ &\leq \sum_i \frac{\|\tilde{\mathbf{X}}_i\|_2}{\Omega} \cdot \left| \frac{\partial(V - f_0)}{\partial x_i} \right| \end{aligned} \quad (2.26)$$

$$\leq \frac{M}{\Omega} \cdot \|V - f_0\|_{W^{1,\infty}} \quad (2.27)$$

$$\leq \epsilon, \quad (2.28)$$

where we have used (2.24), condition  $\|\mathbf{X}_{ij}\|_2 < 1$  and the fact  $\|\alpha \otimes \beta\|_F = \|\alpha\|_2 \|\beta\|_2$  in (2.28), (2.27) and (2.26) correspondingly.  $\square$

It should be pointed out that Theorem 2.3 illustrates that the approximation error of DPK can be arbitrarily small by construction. In practice, however, we will hope to get a network with smaller size (in the meaning of  $\mathbf{W}$ ) and simpler architecture  $f$  than (2.25), which can be calculated more efficiently.

### 2.3 A simplified network inspired by harmonic approximation

The DPK architecture introduced above supplies us with a strong approximation tool which can exactly satisfy the symmetry conditions. On the other hand, an overcomplicated architecture may lead to over-fitting and be hard to converge. More specifically, the sum term

$$f(\mathbf{x}, \theta) = \frac{1}{M!} \sum_{\tau \in \mathcal{S}_M} g(\tau(\mathbf{x}); \theta)$$

with a large  $M$  introduce an intolerable computation complexity (for every  $\mathbf{x}$ , we need  $M!$  forward propagations of  $g$  to get the value of  $f(\mathbf{x})$ ), which requires a large number of calculations in both training and inference procedure. In this section we will introduce a simplified architecture while keeping its symmetry and approximation capability.

In solid state physics, the quasi-harmonic approximation is a classical method to simplify the calculation of free energy at finite temperature. By using Taylor expansion of the potential function  $V$  around some uniform deformed state, the free energy can be written as a sum over vibrational modes. At zero temperature, we can similarly take the Taylor expansion of  $V(\mathbf{x})$  around the perfect state  $\mathbf{X} = [\mathbf{X}_1; \mathbf{X}_2; \cdots; \mathbf{X}_N]$ :

$$V(\mathbf{F} \circ \mathbf{X}) = V(\mathbf{X}) + \frac{1}{2} (\mathbf{F} \circ \mathbf{X} - \mathbf{X})^T \mathbf{D}(\mathbf{X}) (\mathbf{F} \circ \mathbf{X} - \mathbf{X}), \quad (2.29)$$

where  $\mathbf{F} \circ \mathbf{X} = [\mathbf{F}\mathbf{X}_1; \mathbf{F}\mathbf{X}_2; \cdots; \mathbf{F}\mathbf{X}_N]$  and  $\mathbf{D}(\mathbf{X})$  is the force constant matrix. The first order term vanishes since  $\mathbf{X}$  is a stable state. Let  $\mathbf{u} = \mathbf{F} \circ \mathbf{X} - \mathbf{X}$  and  $\xi$  be a point in the first Brillouin zone  $\mathcal{B}$ , by defining the Fourier transforms as

$$\hat{\mathbf{u}}(\xi) = \frac{1}{\sqrt{N}} \sum_{m=1}^N \mathbf{u}_m e^{-i\xi \cdot \mathbf{x}_m}, \quad \hat{\mathbf{D}}(\xi) = \frac{1}{\sqrt{N}} \sum_{m=1}^N \mathbf{D}_{m,1} e^{-i\xi \cdot \mathbf{x}_i}, \quad (2.30)$$

the quadratic term in (2.29) can be written as [12]

$$\frac{1}{2} \mathbf{u}^T \mathbf{D} \mathbf{u} = \frac{1}{2} \sum_{\xi \in \mathcal{B}} \hat{\mathbf{u}}(\xi)^T \hat{\mathbf{D}}(\xi) \hat{\mathbf{u}}(\xi). \quad (2.31)$$

Thus we can approximate the potential  $V(\mathbf{F} \circ \mathbf{X})$  as

$$V(\mathbf{F} \circ \mathbf{X}) = V(\mathbf{X}) + \frac{1}{2} \sum_{\xi \in \mathcal{B}} \hat{\mathbf{u}}(\xi)^T \hat{\mathbf{D}}(\xi) \hat{\mathbf{u}}(\xi), \quad (2.32)$$

namely the potential can be decomposed as the sum over finite different virtual atoms with displacement  $\hat{\mathbf{u}}(\xi)$  and single atom potential  $\frac{1}{2} \hat{\mathbf{u}}(\xi)^T \hat{\mathbf{D}}(\xi) \hat{\mathbf{u}}(\xi)$ .

Inspired by this decomposition, we could assume the virtual potential function  $f$  in (2.14) as

$$f(x_1, x_2, \dots, x_M; \theta) = \sum_{i=1}^M g(x_i; \theta), \quad (2.33)$$

where  $g(x) = h_n \circ a_{n-1} \circ \dots \circ a_1 \circ h_1(x)$  is the classical DNN function. Applying (2.33) to (2.14), we can obtain the first simplification of DPK (denoted as  $DPK_0$ ) as:

$$\mathbf{P}_{\theta, \mathbf{W}}^{DPK_0}(\mathbf{F}) = \mathbf{F} \sum_{i=1}^M \frac{dg}{dx}(|\mathbf{F}\tilde{\mathbf{w}}_i|; \theta) \frac{1}{|\mathbf{F}\tilde{\mathbf{w}}_i|} \tilde{\mathbf{w}}_i \otimes \tilde{\mathbf{w}}_i. \quad (2.34)$$

In practice, however, the loss of this form is hard to be convergent since  $|\mathbf{F}\tilde{\mathbf{w}}_i|$  is in the denominator, a small value of  $\tilde{\mathbf{w}}_i$  may lead to the inaccuracy and inefficiency in both training and predicting. Furthermore, due to the fact that derivative of ReLU always equals to zeros when  $x < 0$ , a well-known gradient vanishing phenomenon may occur during the computation of  $\frac{dg}{dx}$ , which also makes the training hard to converge (as we will see in the following numerical results). Consequently, we can embed this denominator term and the derivative operation into the DNN function  $\tilde{g}$  by representing

$$\frac{1}{x} g'(x; \theta) = \tilde{g}(x; \tilde{\theta}),$$

where  $\tilde{g}(x; \tilde{\theta})$  is a new DNN function. In this case the new simplified DPK (denoted as  $DPK_2$ ) can be written as:

**Definition 2.3** (simplified DPK). *Let  $\mathbf{W}, \tilde{\mathbf{W}}$  be defined as in Definition 2.1 and  $\tilde{g}$  be a DNN function, the simplified DPK is defined as*

$$\mathbf{P}_{\theta, \mathbf{W}}^{DPK_2}(\mathbf{F}) = \mathbf{F} \sum_{i=1}^M \tilde{g}(|\mathbf{F}\tilde{\mathbf{w}}_i|; \tilde{\theta}) \tilde{\mathbf{w}}_i \otimes \tilde{\mathbf{w}}_i. \quad (2.35)$$

Notice that this formula is indeed much simpler than (2.14), in the sense that we use the sum of an identical univariate function  $\tilde{g}(x; \tilde{\theta})$  to represent the potential. The interatomic interactions are embedded in the virtual atoms  $\tilde{\mathbf{w}}_i$ . In the meanwhile, since  $f(x)$  in (2.33) is symmetric, the simplified DPK still meets the symmetry constraints (2.3a) (2.3b).

**Theorem 2.4** (Symmetry of simplified DPK). *For a given crystal system with crystallographic symmetry group  $G_p$ , let  $\mathbf{P}_{\theta, \mathbf{W}}(\mathbf{F})$  be the simplified DPK stress defined by (2.35), then the constraint (2.3a) holds for any  $\mathbf{Q}$  in  $SO(3)$ , and (2.3b) holds for any  $\mathbf{H}$  in  $G_p$ .*

*Proof.* We have

$$\begin{aligned}
 \mathbf{P}_{\theta, \mathbf{W}}(\mathbf{QF}) &= \mathbf{QF} \sum_{i=1}^M \tilde{g}(|\mathbf{QF}\tilde{\mathbf{w}}_i|; \theta) \tilde{\mathbf{w}}_i \otimes \tilde{\mathbf{w}}_i \\
 &= \mathbf{QF} \sum_{i=1}^M \tilde{g}(|\mathbf{F}\tilde{\mathbf{w}}_i|; \theta) \tilde{\mathbf{w}}_i \otimes \tilde{\mathbf{w}}_i \\
 &= \mathbf{QP}_{\theta, \mathbf{W}}(\mathbf{F})
 \end{aligned} \tag{2.36}$$

and

$$\begin{aligned}
 \mathbf{P}_{\theta, \mathbf{W}}(\mathbf{FH})\mathbf{H}^T &= \mathbf{F} \sum_{i=1}^M \tilde{g}(|\mathbf{FH}\tilde{\mathbf{w}}_i|; \theta) \mathbf{H}\tilde{\mathbf{w}}_i \otimes \tilde{\mathbf{w}}_i \mathbf{H}^T \\
 &= \mathbf{F} \sum_{i=1}^M \tilde{g}(|\mathbf{F}\tilde{\mathbf{w}}_{\phi_{\mathbf{H}}(i)}|; \theta) \tilde{\mathbf{w}}_{\phi_{\mathbf{H}}(i)} \otimes \tilde{\mathbf{w}}_{\phi_{\mathbf{H}}(i)} \\
 &= \mathbf{P}_{\theta, \mathbf{W}}(\mathbf{F}).
 \end{aligned} \tag{2.37}$$

This completes the proof.  $\square$

The approximation ability and effectiveness of (2.35) will be illustrated numerically in the following section. Additionally, we can prove that the estimation error (generalization error) of (2.35) can be arbitrarily small with sufficient large number of training samples, i.e., the simplified DPK function space defines a learnable hypothesis class.

**Theorem 2.5** (Learnability of simplified DPK). *Let  $\mathcal{A} = \{\mathbf{P}_{\theta, \mathbf{W}}\}$  be the class of DPK functions defined by (2.35) satisfying that  $\tilde{g}(x; \tilde{\theta})$  has  $l \geq 2$  layers,  $N_w$  weights (bounded by  $V$ ), Lipschitz continuous activation function with Lipschitz coefficient  $L > 1/V$  and each unit maps into  $[-b, b]$  for some  $b > 0$ . Given the training samples  $S_m = \{(\mathbf{F}^1, \mathbf{P}^1), \dots, (\mathbf{F}^m, \mathbf{P}^m)\}$  according to independent identical distribution  $D$  on  $[-1, 1]^{3 \times 3} \times [-1, 1]^{3 \times 3}$  and loss function*

$$l(f(\mathbf{X}), \mathbf{Y}) = \frac{1}{9} \sum_{i,j} l_0(f_{ij}(\mathbf{X}), \mathbf{Y}_{ij}) \quad \text{for } f \in \mathcal{A}, \tag{2.38}$$

where  $l_0(x, y) = |x - y|$ , then for any approximate sample error minimization (SME, [33]) algorithm  $\mathbb{M}$  and  $\epsilon \leq 2\max\{b, V, \sqrt{3V}\}$ ,  $\delta > 0$ , there exists some  $m_0(\epsilon, \delta, \mathcal{A}) < \infty$  such that  $\forall m > m_0$ , with probability at least  $1 - \delta$ ,  $\mathbb{M}$  returns a function  $P^* \in \mathcal{A}$  satisfying

$$E_{(x,y) \sim D}[l(P^*(x), y)] < \inf_{f \in \mathcal{A}} E_{(x,y) \sim D}[l(f(x), y)] + \epsilon. \tag{2.39}$$

In practice, the conditions of  $\tilde{g}$  are indeed satisfied since we will use the ReLU as activation function and bound the network weights by using an  $l_1$  regularization. The proof of above theorem utilizes some classical tools in learning theory, such as fat-shattering dimension. In the appendix we will first list some existing results and then give our proof of Theorem 2.5.

### 3 Numerical results

In this section several numerical results will be presented to verify the effectiveness of our simplified DPK model (2.35). We chose two different crystalline systems, the  $\alpha$ -Fe and NiAl alloy to illustrate the learning ability of DPK.

#### 3.1 Learning the virial stress of $\alpha$ -Fe

As our first example, we use a  $15 \times 15 \times 15 \times 2$  BCC crystal system with periodic boundary condition to calculate the PK stress. The training and testing data is generated by virial formula (2.6). For the molecular simulation, the interatomic interaction is chosen as the analytic EAM potential proposed by [34], which is  $C^3$  smooth. The input deformation and output stress are set as

$$\mathbf{F} = \begin{bmatrix} 1+a_{11} & a_{12} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{P} = \mathbf{P}_{\text{virial}}(\mathbf{F}), \quad (3.1)$$

where  $a_{11}$  and  $a_{12}$  are random variables sampled from a uniform distribution over  $[-0.1, 0.1]$ . In practice, we samples 20000 data by (3.1) and split it into training and testing data equally. The target stress field is plotted in Fig. 2.

For the network architecture, we choose  $N$  (size of  $\mathbf{W}$ ) as 50, and the size of point symmetry group for BCC crystal is 48, which include the symmetry operations generated by rotation, inversion and mirror transformation [30]. Thus the total number of virtual atoms is  $50 \times 48 = 2400$ . The DNN network  $\tilde{g}(x, \tilde{\theta})$  consists of 8 hidden layers with width being 50, 80, 60, 60, 50, 40, 30, 20 and  $ReLU$  as the active function. As to the optimization procedure, we use the Adam algorithm in batch training with batch size being 40, and select the mean absolute error (MAE) together with an  $l_1$  regularization of network weights as our optimization objective. In order to better illustrate the advantage of simplified DPK (2.35) (denoted by  $DPK_2$ ) over (2.34) (denoted by  $DPK_0$ ), we also plot the training result of  $DPK_0$  and the intermediate form  $DPK_1$  which is defined as

$$\mathbf{P}_{\theta, \mathbf{W}}^{DPK_1}(\mathbf{F}) = \mathbf{F} \sum_{i=1}^M g(|\mathbf{F}\tilde{\mathbf{w}}_i|; \theta) \frac{1}{|\mathbf{F}\tilde{\mathbf{w}}_i|} \tilde{\mathbf{w}}_i \otimes \tilde{\mathbf{w}}_i. \quad (3.2)$$

According to the left part of Fig. 3, after 600 epochs the training loss of  $DPK_2$  converges almost to  $10^{-4}$  eV/ $\text{\AA}^3$ , and the testing error also descends to  $10^{-3.75}$  eV/ $\text{\AA}^3$  before overfitting occurs. In the meanwhile, we can see from Fig. 4 that the  $DPK_1$  converges much slower than  $DPK_2$ , since it takes more time to train the parameter  $\tilde{\mathbf{W}}$  at the early stage when the norm of  $\tilde{\mathbf{W}}$  is rather small, which makes the computation of loss and training inaccurate. More seriously, the loss of  $DPK_0$  never converged due to the phenomenon of gradient vanishing, which further implies the necessity of the simplification in (2.35). Fig. 5 shows the learning result of stress field  $\mathbf{P}_{11}(\mathbf{F}_{11}, \mathbf{F}_{12})$  and  $\mathbf{P}_{12}(\mathbf{F}_{11}, \mathbf{F}_{12})$  at

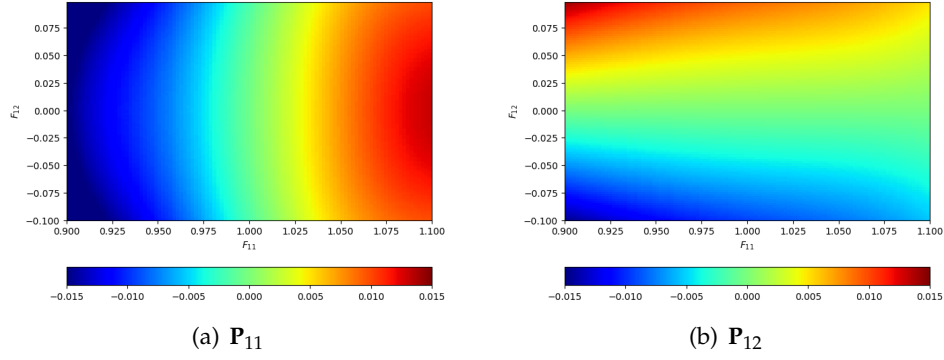


Figure 2: PK stress for  $\alpha$ -Fe by using virial stress. The unit of the pressure is  $\text{eV}/\text{\AA}^3$  ( $=160.212\text{GPa}$ ).

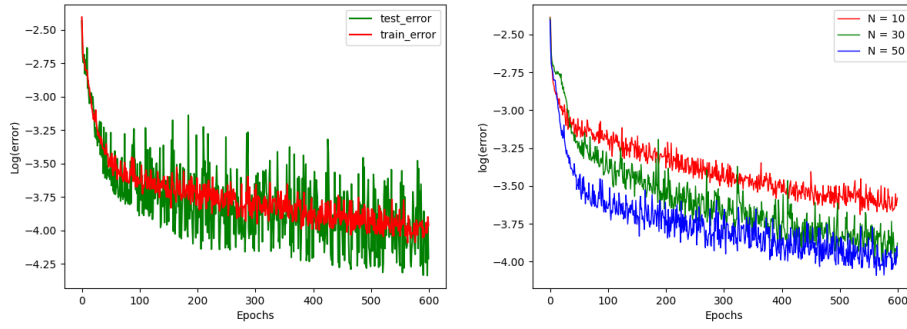


Figure 3: Training and testing error of  $DPK_2$ . Left:  $\log(\text{error})$  for both training and testing error at every epoch. Right:  $\log(\text{error})$  at each training epoch for different size of  $W$ .

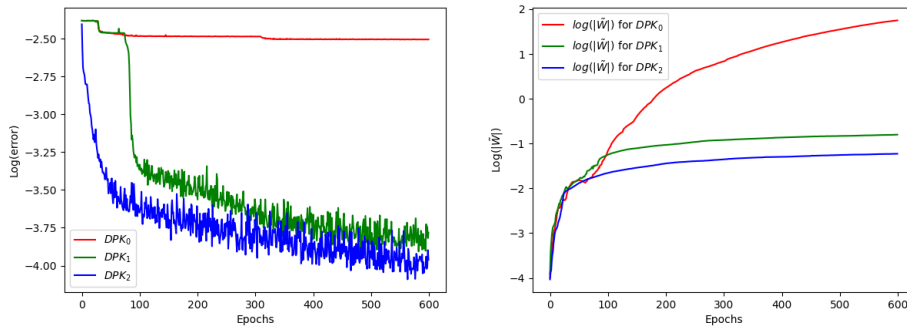


Figure 4: Comparison between different version of simplified DPK. Left:  $\log(\text{error})$  for training procedure at every epoch. Right:  $\log(\text{average norm of } \tilde{w}_i)$  at each training epoch.

epoch = 600, from which we can find the fitting error of  $DPK_2$  is rather small (the  $L_1$  norm of the error function is around  $10^{-4} \text{ eV}/\text{\AA}^3$ ) and the error of  $DPK_1$  is slightly larger than

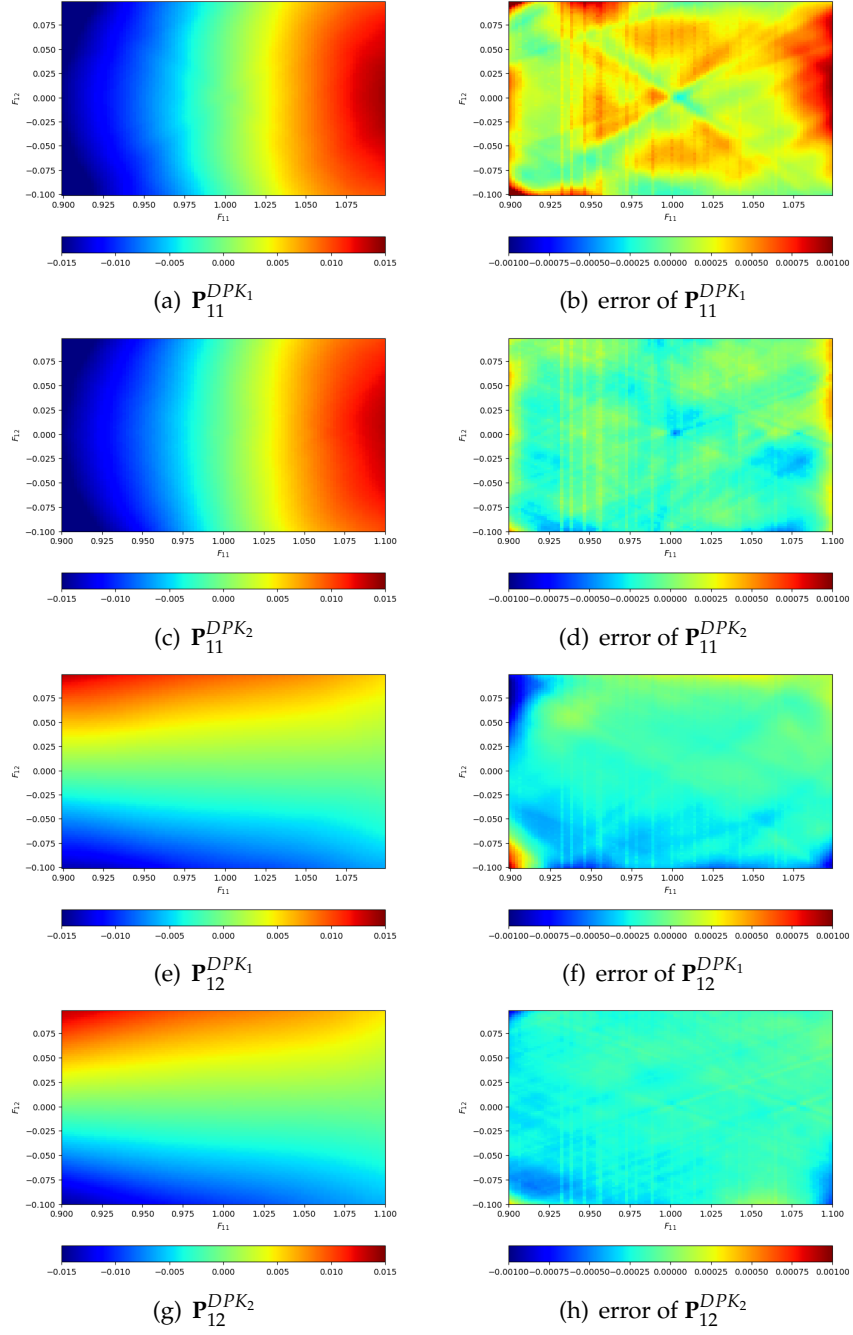


Figure 5: The numerical result of PK stress for  $\alpha$ -Fe by using different version of simplified DPK. The unit of the pressure is  $\text{eV}/\text{\AA}^3$  ( $=160.212\text{GPa}$ ).

Table 1: The MAE at epoch = 600 and time cost of simplified DPK with different  $N$ .

$N$	$M$	MAE (eV/Å <sup>3</sup> )	(batch) Time cost of simplified DPK (s) <sup>1</sup>	Time cost of simplified DPK (s) <sup>2</sup>	Time cost of virial stress (s) <sup>2</sup>
10	480	$3.04 \times 10^{-4}$	$7.9877 \times 10^{-2}$	$3.6315 \times 10^{-2}$	$6.6123 \times 10^{-1}$
30	1440	$1.61 \times 10^{-4}$	$1.4544 \times 10^{-1}$	$3.6538 \times 10^{-2}$	
50	2400	$1.29 \times 10^{-4}$	$5.8815 \times 10^{-1}$	$3.9808 \times 10^{-2}$	

<sup>1</sup> A **batch** inference (batch size = 1000) on Tesla P40 with 24GB memory.

<sup>2</sup> A single calculation on Intel i7-6600U with 16GB memory.

$DPK_2$ , since the loss of  $DPK_2$  converges faster and better than  $DPK_1$  as we have seen in Fig. 4.

As mentioned before, an advantage of the DPK method is that, we can use a smaller virtual system (2400 virtual atoms, depicted by the right part of Fig. 6) to compute the stress by forward network propagation, instead of the virial stress of a crystal system with 6750 atoms (left part of Fig. 6), resulting in a faster calculation. Indeed, as illustrated by Table 1, for a single calculation the average time cost of DPK is 16 times faster than virial stress. Furthermore, with the power of parallel computation of GPU, the time cost of DPK for a batch inference is even smaller than the average cost of a single calculation of virial stress. On the other hand, as shown in the right part of Fig. 3, with a larger size of  $\mathbf{W}$  we can achieve a better approximation of virial stress after same training time. In practice, however, a tradeoff between approximation and time cost should always be considered in the choice of  $N$ .

### 3.2 Transfer learning to NiAl alloys

In this part, we will test the ability of transfer learning [35] of simplified DPK by approximating the virial stress for NiAl, which is a B2 type alloy. All the numerical setup are kept as previous, expect the  $F_{11}$  and  $F_{12}$  of input deformation  $\mathbf{F}$  are sampled uniformly from



Figure 6: Configuration of atoms in stress calculation. Left: planform of perfect BCC crystal. Right: planform of virtual atoms.



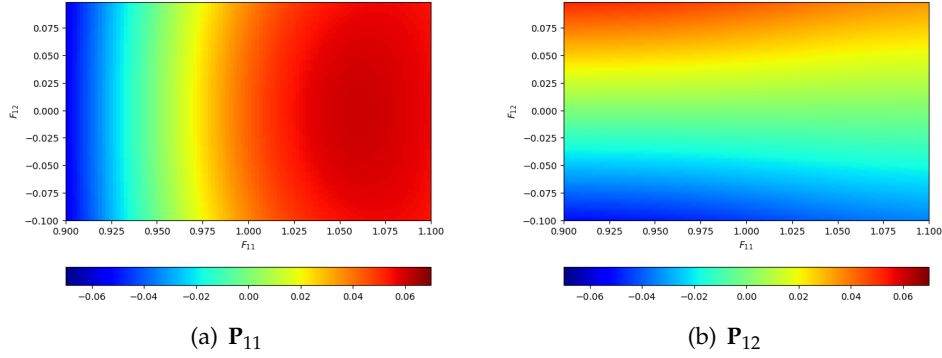


Figure 7: PK stress for NiAl alloy system by using virial stress. The unit of the pressure is  $\text{eV}/\text{\AA}^3$  ( $=160.212\text{GPa}$ ).

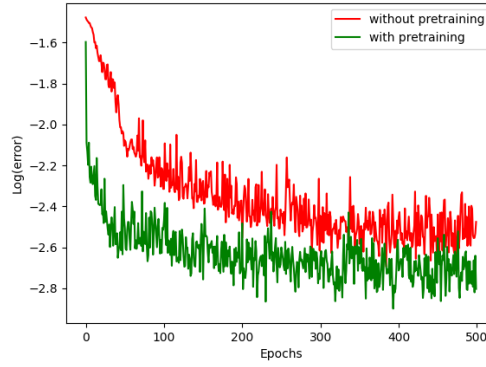


Figure 8: Training error in the learning of virial stress for NiAl alloy system.

$[0.98, 1.06]$  and  $[-0.04, 0.04]$  respectively. We adopt the EAM potential studied in [36, 37] in the molecular simulation. The stress field of training data is shown by Fig. 7.

In order to do the transfer learning as in natural language processing [38], we take the fitting of stress for  $\alpha$ -Fe as a pre-training task, while the learning of NiAl stress is treated as a downstream fine-tuning task. More precisely, we first conduct the experiment in previous section, and select  $\mathbf{W}$  and the parameters of the first 3 layers in the DNN function as initial parameters for the new learning task of NiAl. The motivation of this setting is that due to the similarity in crystal structure (The  $\alpha$ -Fe can also be regarded as a complex B2 lattice with same basis atoms), we can expect that parameters in DPK trained for  $\alpha$ -Fe provides a better initial guess than random initialization.

As shown in Fig. 8, with pre-training the loss converges much faster, compared with the learning process with random initialization. In fact, we can observe that the training loss with pre-training decreases to  $10^{-2.6} \text{ eV}/\text{\AA}^3$  after only 100 epochs, while this can not be achieved until about 400 epochs if we initialize the network randomly. The learning

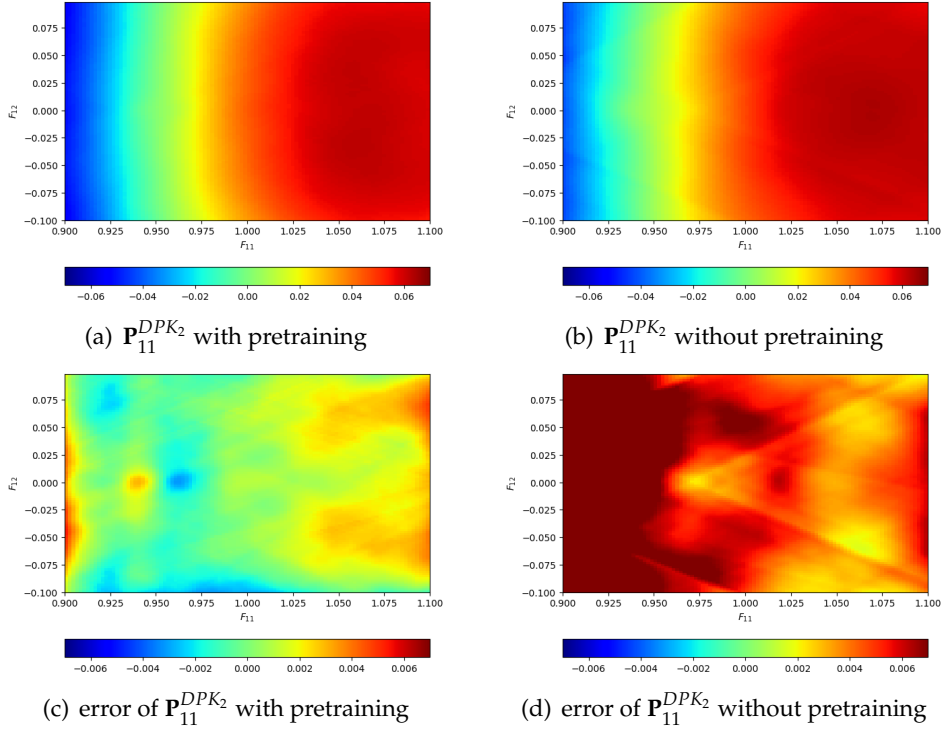


Figure 9: The numerical result of  $\mathbf{P}_{11}$  for NiAl alloy system by using simplified DPK. The unit of the pressure is  $\text{eV}/\text{\AA}^3$  ( $=160.212\text{GPa}$ ).

stress field of  $P_{11}$  and  $P_{12}$  after 500 epochs are plotted in Fig. 9 and Fig. 10 correspondingly. As presents, the fitting error with a pre-training is observably smaller in both cases.

## 4 Conclusions

We have proposed a novel neural network architecture (DPK) to learn the first Piola-Kirchhoff stress. By introducing symmetric trainable weights, our network is intrinsic consistent with the physical constraints required by the frame-indifference and material-symmetry. A further simplification of DPK is also presented to avoid an over complicated computation in both training and inference. On the other hand, with some theory analysis and numerical examples, we have shown that this simplified DPK provides a learnable, transferable and efficient approximation tool for the representation of deformation-stress relation with an atomic-level accuracy. Although the present work focuses on the learning of PK stress for elastic simple materials, this structure can also be used to represent the conservative elastic part of the stress tensor for non-hyperelastic material. Besides, the methodology in the constructing of DPK (Generating the symmetric structure in the network based on the formulation of atomistic stress) can also be applied to more

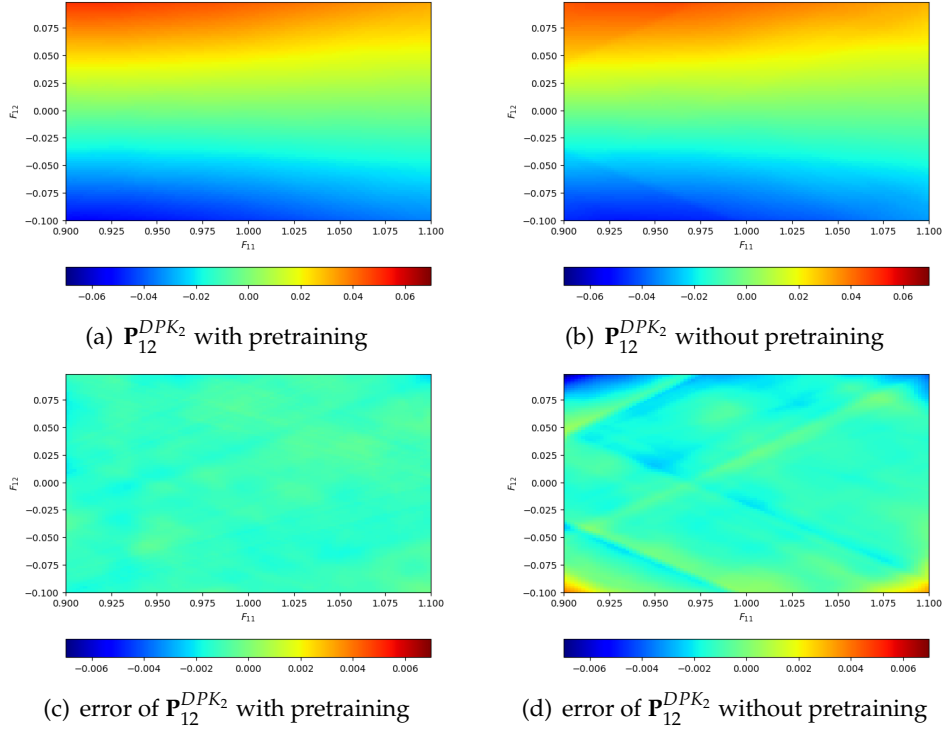


Figure 10: The numerical result of  $P_{12}$  for NiAl alloy system by using simplified DPK. The unit of the pressure is  $\text{eV}/\text{\AA}^3$  ( $=160.212\text{GPa}$ ).

general case, once the atomistic stress expression is available.

## Appendix

The proof of Theorem 2.5 consists of two parts: firstly, we review the concept of fat-shattering dimension and prove that simplified DPK owns finite fat-shattering dimension. Secondly, with a classical result claiming that function class with finite fat-shattering dimension is learnable, we can eventually prove the estimation error of our simplified DPK can be sufficient small, given a large size of learning samples.

**Definition A.1** (Fat-shattering dimension [33]). Let  $F$  be a function set mapping from  $X$  to  $\mathbb{R}$ , for a given sample set  $S = \{s_1, \dots, s_m\} \subset X$  and positive real number  $\epsilon$ , if there exists real numbers  $r_1, \dots, r_m$  such that  $\forall b \in \{0, 1\}^m$ ,  $\exists f_b \in F$  satisfying

$$f_b(s_i) \geq r_i + \epsilon \quad \text{if } b_i = 1, \quad f_b(s_i) \leq r_i - \epsilon \quad \text{if } b_i = 0, \quad i = 1, \dots, m, \quad (\text{A.1})$$

we say that  $S$  is  $\epsilon$ -shattered by  $F$ . Furthermore, we define the fat-dimension  $\text{fat}_F(\epsilon)$  as the maximum cardinality of sampling sets  $S$  shattered by  $F$ .

**Lemma A.1** ([33, Theorem 14.9]). *For each dense neural network has  $l \geq 2$  layers and  $N_w$  weights of which the  $\|\cdot\|_1$  norm are bounded by some  $V > 0$ , suppose the activation function is Lipschitz continuous with the Lipschitz coefficient  $L > 1/V$ , and there exists some  $b$  such that each unit maps into  $[-b, b]$ , then for any  $\epsilon \leq 2b$ ,*

$$fat_F(\epsilon) \leq 16N_w \left( l \cdot \ln(LV) + 2\ln(32N_w) + \ln\left(\frac{b}{\epsilon(LV-1)}\right) \right). \quad (\text{A.2})$$

**Lemma A.2** ([39, [Theorem 3]). *Let  $\epsilon > 0$ ,  $k \geq 2$ ,  $u : [-1, 1]^k \rightarrow [-1, 1]$  be uniformly continuous and  $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_k$  be function classe including  $f : \Omega \rightarrow [-1, 1]$ , then there exist  $0 < \alpha(u, k, \epsilon), \beta(u, k, \epsilon) < \infty$  such that*

$$fat_{u(\mathcal{A}_1, \dots, \mathcal{A}_k)}(\epsilon) \leq \alpha \sum_{i=1}^k fat_{\mathcal{A}_i}(\beta), \quad (\text{A.3})$$

where

$$u(\mathcal{A}_1, \dots, \mathcal{A}_k) \triangleq \{u(f_1(x), \dots, f_k(x)) | f_i \in \mathcal{A}_i, i = 1, \dots, k\}.$$

With Lemmas A.1 and A.2, we can prove that each component function  $\mathbf{P}_{\theta, \mathbf{w}}^{ij}(\mathbf{F})$  has a finite fat-shattering dimension:

**Theorem A.1.** *Let*

$$P_{ij} = \left\{ \frac{1}{M} \mathbf{e}_i^T \mathbf{P}_{\theta, \mathbf{w}}(\mathbf{F}) \mathbf{e}_j : [-1, 1]^{3 \times 3} \rightarrow \mathbb{R} \right\}$$

*be the function space generated by the  $(i, j)$  component function of DPK defined by (2.35), if  $\tilde{g}$  has  $l \geq 2$  layers and  $N_w$  weights of which the  $\|\cdot\|_1$  norm are bounded by some  $V < 1$ ,  $\|\tilde{\mathbf{w}}_i\|_1 < V$  for  $i = 1, \dots, M$ , suppose the activation function is Lipschitz continuous with the Lipschitz coefficient  $L > 1/V$ , and there exists some  $b < 1$  such that each unit of  $\tilde{g}$  maps into  $[-b, b]$ , then for any  $\epsilon \leq 2\max\{b, V, \sqrt{3V}\}$ ,  $\exists c(\epsilon, M, L, V, b) < \infty$  such that*

$$fat_{P_{ij}}(\epsilon) \leq c(\epsilon, M, L, V, b). \quad (\text{A.4})$$

*Proof.* For each  $i, j$ ,

$$P_{ij} = \frac{1}{M} \sum_{k=1}^M \tilde{g}(|\mathbf{F}\tilde{\mathbf{w}}_k|) \cdot (\mathbf{e}_i^T \mathbf{F}\tilde{\mathbf{w}}_k \otimes \tilde{\mathbf{w}}_k \mathbf{e}_j).$$

Under above conditions, we have  $P_{ij} \in [-1, 1]$ . By defining

$$h_1(\mathbf{F}) = \mathbf{F}\tilde{\mathbf{w}}_k, \quad a_1(x) = x^2, \quad h_2(\mathbf{x}) = \sum_{i=1}^3 x_i, \quad a_2(x) = \sqrt{x},$$

we may write  $\tilde{g}(|\mathbf{F}\tilde{\mathbf{w}}_k|)$  as

$$g_1(\mathbf{F}) = \tilde{g} \circ a_2 \circ h_2 \circ a_1 \circ h_1(\mathbf{F}), \quad (\text{A.5})$$

which can be regarded as an element in the new network spcae with two more layers. Notice that  $\|\tilde{\mathbf{w}}_k\|_1 < V$  indicate that the  $\|\cdot\|_1$  norm of additional layers are bounded by  $V_1 = \max\{V, 3\}$ ,

$$\begin{aligned} \|a_1 \circ h_1(\mathbf{F})\|_\infty &\leq \|\mathbf{F}\|_\infty \|\tilde{\mathbf{w}}_k\|_1^2 < V^2 < V, \\ |a_2 \circ h_2 \circ a_1 \circ h_1(\mathbf{F})| &= \|\mathbf{F}\tilde{\mathbf{w}}_k\|_2 \leq \sqrt{3\|a_1 \circ h_1(\mathbf{F})\|_\infty} < \sqrt{3V}, \end{aligned}$$

shows the range of additional units are bounded by  $b_1 = \max\{b, V, \sqrt{3V}\}$ . Thus  $g_1$  meets the condition in Lemma A.1, and we have for any  $\epsilon \leq 2b_1$ ,

$$fat_{g_1}(\epsilon) \leq 16N_w(g_1)(l(g_1)\ln(LV_1) + 2\ln\left(32(N_w(g_1)) + \ln\left(\frac{b_1}{\epsilon(LV_1-1)}\right)\right)), \quad (\text{A.6})$$

where  $N_w(g_1)$  and  $l(g_1)$  stands for the number of weights and number of layers for  $g_1$ . Similarly, let  $h_3(\mathbf{x}) = x_i \tilde{\mathbf{w}}_k^j$ ,  $a_3(x) = x$ , we can represent  $\mathbf{e}_i^T \mathbf{F} \tilde{\mathbf{w}}_k \otimes \tilde{\mathbf{w}}_k \mathbf{e}_j$  as

$$g_2(\mathbf{F}) = a_3 \circ h_3 \circ a_3 \circ h_1(\mathbf{F}),$$

which also satisfies the condition in Lemma A.1 while the computation unit and  $l_1$  norm of weights are bound by  $b_2 = V$ ,  $V_2 = V$ , respectively. Thus for any  $\epsilon \leq 2b_2$ ,

$$fat_{g_2}(\epsilon) \leq 16N_w(g_2)\left(l(g_2)\ln(LV_2) + 2\ln\left(32(N_w(g_2)) + \ln\left(\frac{b_2}{\epsilon(LV_2-1)}\right)\right)\right). \quad (\text{A.7})$$

Now considering the uniformly continuous function  $u(x, y) = xy$  with

$$\mathcal{A}_1 = \{g_1 | g_1(\mathbf{F}) = \tilde{g} \circ a_2 \circ h_2 \circ a_1 \circ h_1(\mathbf{F})\} \quad \text{and} \quad \mathcal{A}_2 = \{g_2 | g_2(\mathbf{F}) = a_3 \circ h_3 \circ a_3 \circ h_1(\mathbf{F})\},$$

according to Lemma A.2, for any  $\epsilon \leq 2b_2 < 2b_1$ , there exit  $0 < \alpha_1(u, 2, \epsilon), \beta_1(u, 2, \epsilon) < \infty$  such that

$$fat_{u(\mathcal{A}_1, \mathcal{A}_2)}(\epsilon) \leq \alpha_1 \sum_{i=1}^2 fat_{\mathcal{A}_i}(\beta_1) \leq \alpha_1 \sum_{i=1}^2 16N_w(g_i) \left( \ln \frac{32^2 b_i (LV_i)^{l(g_i)} N_w(g_i)^2}{\beta_1 (LV_i - 1)} \right). \quad (\text{A.8})$$

Further more, notice that for different  $k = 1, \dots, M$ ,  $\tilde{g}(|\mathbf{F}\tilde{\mathbf{w}}_k|)$  and  $(\mathbf{e}_i^T \mathbf{F} \tilde{\mathbf{w}}_k \otimes \tilde{\mathbf{w}}_k \mathbf{e}_j)$  share the same network architecture, by using Lemma A.2 again with

$$\bar{u}(x_1, \dots, x_M) = \frac{1}{M} \sum_{i=1}^M x_i,$$

we can obtain that there exists  $0 < \alpha_2(\bar{u}, M, \epsilon), \beta_2(\bar{u}, M, \epsilon) < \infty$  such that

$$\begin{aligned}
 fat_{P_{ij}}(\epsilon) &\leq \alpha_2 \sum_{k=1}^M fat_{u(\mathcal{A}_1, \mathcal{A}_2)}(\beta_2(\bar{u}, M, \epsilon)) \\
 &\leq \alpha_2 \sum_{k=1}^M \alpha_1 \sum_{i=1}^2 fat_{\mathcal{A}_i}(\beta_1(u, 2, \beta_2(\bar{u}, M, \epsilon))) \\
 &\leq M \alpha_1 \alpha_2 \sum_{i=1}^2 fat_{\mathcal{A}_i}(\beta_1(u, 2, \beta_2(\bar{u}, M, \epsilon))) \\
 &\leq 16M \alpha_1 \alpha_2 \sum_{i=1}^2 N_w(g_i) \left( \ln \frac{32^2 b_i (LV_i)^{l(g_i)} N_w(g_i)^2}{\beta_1(LV_i - 1)} \right). \tag{A.9}
 \end{aligned}$$

We complete the proof.  $\square$

**Lemma A.3** (see the proof of Theorem 19.1 in [33]). *Let  $\mathcal{A}$  be a class of functions mapping from  $X$  to  $[0, 1]$  with finite fat-shattering dimension. Given the training samples  $S_m = \{(x_1, y_1), \dots, (x_m, y_m)\}$  according to distribution  $D$  and loss function  $l(f(x), y)$  for  $f \in \mathcal{A}$ , for any  $\epsilon, \delta > 0$ , there exists*

$$m_0(\epsilon, \delta) = \frac{256}{\epsilon^2} \left( 18 fat_{\mathcal{A}} \left( \frac{\epsilon}{256} \right) \ln^2 \left( \frac{128}{\epsilon} \right) + \ln \left( \frac{16}{\delta} \right) \right), \tag{A.10}$$

such that  $\forall m > m_0$ :

$$P \left( \left\{ S_m \mid \exists f \in \mathcal{A}, \left| E_{(x,y) \sim D} [l(f(x), y)] - \frac{1}{m} \sum_{i=1}^m l(f(x_i), y_i) \right| \geq \frac{\epsilon}{2} \right\} \right) \leq \frac{\delta}{2}. \tag{A.11}$$

**Corollary A.1.** *Let  $\mathcal{A}$  be a class of functions mapping from  $X$  to  $[-1, 1]$  with finite fat-shattering dimension. Given the training samples  $S_m = \{(x_1, y_1), \dots, (x_m, y_m)\}$  according to distribution  $D$  and loss function  $l(f(x), y) = |f(x) - y|$ , let  $u(x) = \frac{x+1}{2}$ , then for any  $\epsilon, \delta > 0$  there exists*

$$m_0(\epsilon, \delta) = \frac{256}{\epsilon^2} \left( 18 fat_{u(\mathcal{A})} \left( \frac{\epsilon}{256} \right) \ln^2 \left( \frac{128}{\epsilon} \right) + \ln \left( \frac{16}{\delta} \right) \right), \tag{A.12}$$

such that  $\forall m > m_0$ :

$$P \left( \left\{ S_m \mid \exists f \in \mathcal{A}, \left| E_{(x,y) \sim D} [l(f(x), y)] - \frac{1}{m} \sum_{i=1}^m l(f(x_i), y_i) \right| \geq \frac{\epsilon}{2} \right\} \right) \leq \frac{\delta}{2}. \tag{A.13}$$

*Proof.* The proof is a combination of Lemmas A.2 and A.3. Let  $\mathcal{B} = \left\{ \frac{f+1}{2} \mid f \in \mathcal{A} \right\}$ , according Lemma A.2,  $\mathcal{B}$  forms class of functions mapping to  $[0, 1]$  with finite fat-shattering dimension

$$fat_{\mathcal{B}}(\epsilon) = fat_{u(\mathcal{A})}(\epsilon) < \infty.$$

The rest follows from Lemma A.3 if we consider the new regression problems with function class  $\mathcal{B}$ , samples

$$\tilde{S}_m = \left\{ \left( x_1, \frac{y_1+1}{2} \right), \dots, \left( x_m, \frac{y_m+1}{2} \right) \right\}$$

and loss function  $\tilde{l}(f(x), y) = 2|f(x) - y|$ . This completes the proof.  $\square$

Now we can give the proof of the learnability of simplified DPK as follows:

*Proof of Theorem 2.5.* For fixed  $1 \leq i, j \leq 3$ , considering the scaled real-valued function space  $\mathcal{A}_{ij} = \left\{ \frac{1}{M} \mathbf{P}_{ij} \mid \mathbf{P} \in \mathcal{A} \right\}$  ( $\mathcal{A}$  is defined in Theorem 2.5) which maps from  $[-1, 1]^{3 \times 3}$  to  $[-1, 1]$ . According to Theorem A.1, for any  $\epsilon \leq 2 \max\{b, V, \sqrt{3V}\}$ ,  $\text{fat}_{\mathcal{A}_{ij}}(\epsilon)$  is finite, then by Corollary (A.1),  $\forall \epsilon \leq 2 \max\{b, V, \sqrt{3V}\}, \delta > 0, \exists m_0$  defined by (A.12) such that  $\forall m > m_0$ ,

$$P\left(\left\{S_m : \exists f \in \mathcal{A}_{ij}, \left| E_{(\mathbf{X}, \mathbf{Y}) \sim D} \left[ l_0\left(f(\mathbf{X}), \frac{\mathbf{Y}_{ij}}{M}\right) \right] - \frac{1}{m} \sum_{k=1}^m l_0\left(f(\mathbf{X}^k), \frac{\mathbf{Y}_{ij}^k}{M}\right) \right| \geq \frac{\epsilon}{6M} \right\}\right) \leq \frac{\delta}{18}. \quad (\text{A.14})$$

As a consequence,

$$\begin{aligned} & P\left(\left\{S_m : \exists f \in \mathcal{A}, \left| E[l(f(\mathbf{X}), \mathbf{Y})] - \frac{1}{m} \sum_{k=1}^m l(f(\mathbf{X}^k), \mathbf{Y}^k) \right| \geq \frac{\epsilon}{6} \right\}\right) \\ & \leq P\left(\left\{S_m : \exists f \in \mathcal{A}, \frac{1}{9} \sum_{i,j} \left| E[l_0(f_{ij}(\mathbf{X}), \mathbf{Y}_{ij})] - \frac{1}{m} \sum_{k=1}^m l_0(f_{ij}(\mathbf{X}^k), \mathbf{Y}_{ij}^k) \right| \geq \frac{\epsilon}{6} \right\}\right) \\ & \leq P\left(\bigcup_{i,j} \left\{S_m : \exists f \in \mathcal{A}, \left| E[l_0(f_{ij}(\mathbf{X}), \mathbf{Y}_{ij})] - \frac{1}{m} \sum_{k=1}^m l_0(f_{ij}(\mathbf{X}^k), \mathbf{Y}_{ij}^k) \right| \geq \frac{\epsilon}{6} \right\}\right) \\ & \leq \sum_{i,j} P\left(\left\{S_m : \exists f \in \mathcal{A}_{ij}, \left| E\left[l_0\left(f(\mathbf{X}), \frac{\mathbf{Y}_{ij}}{M}\right)\right] - \frac{1}{m} \sum_{k=1}^m l_0\left(f(\mathbf{X}^k), \frac{\mathbf{Y}_{ij}^k}{M}\right) \right| \geq \frac{\epsilon}{6M} \right\}\right) \\ & \leq \frac{9\delta}{18} \leq \frac{\delta}{2}. \end{aligned} \quad (\text{A.15})$$

On the other hand, for the given  $\epsilon$ , by definition the approximate-SME algorithm  $M$  returns  $P^*$  such that

$$\frac{1}{m} \sum_{k=1}^m l(P^*(\mathbf{X}^k), \mathbf{Y}^k) < \inf_{f \in \mathcal{A}} \frac{1}{m} \sum_{i=1}^m l(f(\mathbf{X}^k), \mathbf{Y}^k) + \frac{\epsilon}{3}. \quad (\text{A.16})$$

Meanwhile, use the definition of infimum, we can find  $\bar{P} \in \mathcal{A}$  satisfying

$$E[l(\bar{P}(\mathbf{X}), \mathbf{Y})] \leq \inf_{f \in \mathcal{A}} E[l(f(\mathbf{X}), \mathbf{Y})] + \frac{\epsilon}{3}. \quad (\text{A.17})$$

Combining the result of (A.15)-(A.17), we can deduce that with probability at least  $1 - \delta$ ,

$$\begin{aligned}
 E[l(P^*(\mathbf{X}), \mathbf{Y})] &\leq \frac{1}{m} \sum_{i=1}^m l(P^*(\mathbf{X}^k), \mathbf{Y}^k) + \frac{\epsilon}{6} && \text{(by (A.15))} \\
 &\leq \inf_{f \in \mathcal{A}} \frac{1}{m} \sum_{i=1}^m l(f(\mathbf{X}^k), \mathbf{Y}^k) + \frac{\epsilon}{2} && \text{(by (A.16))} \\
 &\leq \frac{1}{m} \sum_{i=1}^m l(\bar{P}(\mathbf{X}^k), \mathbf{Y}^k) + \frac{\epsilon}{2} && \text{(definition of infimum)} \\
 &\leq E[l(\bar{P}(\mathbf{X}), \mathbf{Y})] + \frac{2\epsilon}{3} && \text{(by (A.15))} \\
 &\leq \inf_{f \in \mathcal{A}} E[l(f(\mathbf{X}), \mathbf{Y})] + \epsilon. && \text{(by (A.17))}
 \end{aligned}$$

Thus we have proved that for  $m > m_0$ , the conclusion in Theorem 2.5 holds.  $\square$

## Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2020YFA0714200), by the National Nature Science Foundation of China (Nos. 12125103 and 12071362), by the Natural Science Foundation of Hubei Province (Nos. 2021AAA010 and 2019CFA007), and by the Fundamental Research Funds for the Central Universities. The numerical calculations have been done at the Super Computing Center of Wuhan University.

## References

- [1] NIKHIL CHANDRA ADMAL AND ELLAD B. TADMOR, *A unified interpretation of stress in molecular systems*, Journal of Elasticity, 100(1) (2010), pp. 63–143.
- [2] D. H. TSAI, *The virial theorem and stress calculation in molecular dynamics*, The Journal of Chemical Physics, 70(3) (1979), pp. 1375–1382.
- [3] JONATHAN A. ZIMMERMAN, EDMUND B. WEBBIII, J. J. HOYT, REESE E. JONES, P. A. KLEIN, AND DOUGLAS J. BAMMANN, *Calculation of stress in atomistic simulation*, Model. Simul. Mater. Sci. Eng., 12(4) (2004), p. S319.
- [4] J. H. IRVING AND JOHN G. KIRKWOOD, *The statistical mechanical theory of transport processes. iv. the equations of hydrodynamics*, The Journal of Chemical Physics, 18(6) (1950), pp. 817–829.
- [5] ROBERT J. HARDY, *Formulas for determining local properties in molecular-dynamics simulations: Shock waves*, The Journal of Chemical Physics, 76(1) (1982), pp. 622–628.
- [6] JERRY ZHIJIAN YANG, XIAOJIE WU, AND XIANTAO LI, *A generalized Irving–Kirkwood formula for the calculation of stress in molecular dynamics models*, The Journal of Chemical Physics, 137(13) (2012), p. 134104.
- [7] TENGYUAN HAO AND ZUBAER M. HOSSAIN, *Atomistic mechanisms of crack nucleation and propagation in amorphous silica*, Phys. Rev. B, 100(1) (2019), p. 014204.



- [8] A. I. BITSANIS, J. J. MAGDA, M. TIRRELL, AND H. T. DAVIS, *Molecular dynamics of flow in micropores*, The Journal of Chemical Physics, 87(3) (1987), pp. 1733–1750.
- [9] Xiantao Li, Jerry Z. Yang, and Weinan E, *A multiscale coupling method for the modeling of dynamics of solids with application to brittle cracks*, J. Comput. Phys., 229(10) (2010), pp. 3970–3987.
- [10] Ellad B. Tadmor, Michael Ortiz, and Rob Phillips, *Quasicontinuum analysis of defects in solids*, Philosophical Magazine A, 73(6) (1996), pp. 1529–1563.
- [11] Jerry Z. Yang, Chao Mao, Xiantao Li, and Chun Liu, *On the Cauchy–Born approximation at finite temperature*, Comput. Mater. Sci., 99 (2015), pp. 21–28.
- [12] Shuyang Dai, Fengru Wang, Jerry Zhijian Yang, and Cheng Yuan, *On the Cauchy–Born approximation at finite temperature for alloys*, Discrete and Continuous Dynamical Systems-B, (2021).
- [13] Rudolf Clausius, *Xvi. on a mechanical theorem applicable to heat*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 40(265) (1870), pp. 122–127.
- [14] Weinan E, *Machine learning and computational mathematics*, Commun. Comput. Phys., 28(5) (2020), pp. 1639–1670.
- [15] Weile Jia, Han Wang, Mohan Chen, Denghui Lu, Lin Lin, Roberto Car, Weinan E, and Linfeng Zhang, *Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning*, in SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, pages 1–14, IEEE, 2020.
- [16] Han Wang, Linfeng Zhang, Jiequn Han, and Weinan E, *Deepmd-kit: A deep learning package for many-body potential energy representation and molecular dynamics*, Comput. Phys. Commun., 228 (2018), pp. 178–184.
- [17] Maziar Raissi, Paris Perdikaris, and George E. Karniadakis, *Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving non-linear partial differential equations*, J. Comput. phys., 378 (2019), pp. 686–707.
- [18] Jiequn Han and Arnulf Jentzen, et al., *Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations*, Commun. Math. Stat., 5(4) (2017), pp. 349–380.
- [19] Jiequn Han, Arnulf Jentzen, and Weinan E, *Solving high-dimensional partial differential equations using deep learning*, Proceedings of the National Academy of Sciences, 115(34) (2018), pp. 8505–8510.
- [20] Hongliang Liu, Jingwen Song, Huini Liu, Jie Xu, and Lijuan Li, *Legendre neural network for solving linear variable coefficients delay differential-algebraic equations with weak discontinuities*, Adv. Appl. Math. Mech., 13(1) (2021), pp. 101–118.
- [21] Fengru Wang, Jerry Zhijian Yang, and Cheng Yuan, *Practical absorbing boundary conditions for wave propagation on arbitrary domain*, Adv. Appl. Math. Mech., 12(6) (2020), pp. 1384–1415.
- [22] Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu, *Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning*, Science, 365(6457) (2019), eaaw1147.
- [23] Wei Feng and Haibo Huang, *Fast prediction of immiscible two-phase displacements in heterogeneous porous media with convolutional neural network*, Adv. Appl. Math. Mech., 13(1) (2021), pp. 140–162.
- [24] T. Lin, Z. Wang, R. X. Lu, W. Wang, and Y. Sui, *Characterising mechanical properties of flowing microcapsules using a deep convolutional neural network*, Adv. Appl. Math. Mech., 14(1) (2022), pp. 79–100.

- [25] KAILAI XU, DANIEL Z. HUANG, AND ERIC DARVE, *Learning constitutive relations using symmetric positive definite neural networks*, J. Comput. Phys., 428 (2021), p. 110072.
- [26] REESE E. JONES, JEREMY A. TEMPLETON, CLAY M. SANDERS, AND JAKOB T. OSTIEN, *Machine learning models of plastic flow based on representation theory*, arXiv preprint arXiv:1809.00267, (2018).
- [27] DANIEL Z. HUANG, KAILAI XU, CHARBEL FARHAT, AND ERIC DARVE, *Learning constitutive relations from indirect observations using deep neural networks*, J. Comput. Phys., 416 (2020), p. 109491.
- [28] XIN LIU, SU TIAN, FEI TAO, AND WENBIN YU, *A review of artificial neural networks in the constitutive modeling of composite materials*, Compos. Part B Eng., 224 (2021), p. 109152.
- [29] ELLAD B. TADMOR AND RONALD E. MILLER, *Modeling Materials: Continuum, Atomistic and Multiscale Techniques*, Cambridge University Press, 2011.
- [30] BERNARD D. COLEMAN AND WALTER NOLL, *Material symmetry and thermodynamic inequalities in finite elastic deformations*, Arch. Ration. Mech. Anal., 15(2) (1964), pp. 87–111.
- [31] R. E. NEWNHAM, *Properties of Materials: Anisotropy, Symmetry, Structure*, 2009.
- [32] INGO GÜHRING, GITTA KUTYNIOK, AND PHILIPP PETERSEN, *Error bounds for approximations with deep relu neural networks in  $w$   $s$ ,  $p$  norms*, Anal. Appl., 18(05) (2020), pp. 803–859.
- [33] MARTIN ANTHONY, PETER L. BARTLETT, AND PETER L. BARTLETT, ET AL., *Neural Network Learning: Theoretical Foundations*, volume 9, Cambridge University Press Cambridge, 1999.
- [34] H. CHAMATI, N. I. PAPANICOLAOU, Y. MISHIN, AND D. A. PAPACONSTANTOPOULOS, *Embedded-atom potential for fe and its application to self-diffusion on Fe(100)*, Surface Science, 600(9) (2006), pp. 1793–1803.
- [35] FUZHEN ZHUANG, ZHIYUAN QI, KEYU DUAN, DONGBO XI, YONGCHUN ZHU, HENGSHU ZHU, HUI XIONG, AND QING HE, *A comprehensive survey on transfer learning*, Proceedings of the IEEE, 109(1) (2021), pp. 43–76.
- [36] MURRAY S. DAW AND MICHAEL I. BASKES, *Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals*, Phys. Rev. B, 29(12) (1984), p. 6443.
- [37] S. M. FOILES, M. I. BASKES, AND MURRAY S. DAW, *Embedded-atom-method functions for the fcc metals Cu, Ag, Au, Ni, Pd, Pt, and their alloys*, Physical review B, 33(12) (1986), p. 7983.
- [38] JACOB DEVLIN, MING-WEI CHANG, KENTON LEE, AND KRISTINA TOUTANOVA, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805, (2018).
- [39] OHAD ASOR, HUBERT HAORYANG DUAN, AND ARYEH KONTOROVICH, *On the additive properties of the fat-shattering dimension*, IEEE Transactions on Neural Networks and Learning Systems, 25(12) (2014), pp. 2309–2312.