Article

Dislocation Substructures Evolution and an Informer Constitutive Model for a Ti-55511 Alloy in Two-Stages High-Temperature Forming with Variant Strain Rates in β Region

Shen Tan [1,2], Daoguang He [2,3,*], Yongcheng Lin [2,3,*], Bingkun Zheng [2,3] and Heyi Wu [2,3]

1   School of Automation, Central South University, Changsha 410083, China; shiina1024@csu.edu.cn
2   School of Mechanical and Electrical Engineering, Central South University, Changsha 410083, China; kakigaku1024@gmail.com (B.Z.); shiina1024@163.com (H.W.)
3   State Key Laboratory of Precision Manufacturing for Extreme Service Performance, Changsha 410083, China
*   Correspondence: daoguanghe@csu.edu.cn (D.H.); yclin@csu.edu.cn (Y.L.);
    Tel.: +86-015273177698 (D.H.); +86-013469071208 (Y.L.)

Abstract: The high-temperature compression characteristics of a Ti-55511 alloy are explored through adopting two-stage high-temperature compressed experiments with step-like strain rates. The evolving features of dislocation substructures over hot, compressed parameters are revealed by transmission electron microscopy (TEM). The experiment results suggest that the dislocations annihilation through the rearrangement/interaction of dislocations is aggravated with the increase in forming temperature. Notwithstanding, the generation/interlacing of dislocations exhibit an enhanced trend with the increase in strain in the first stage of forming, or in strain rates at first/second stages of a high-temperature compressed process. According to the testing data, an Informer deep learning model is proposed for reconstructing the stress–strain behavior of the researched Ti-55511 alloy. The input series of the established Informer deep learning model are compression parameters (compressed temperature, strain, as well as strain rate), and the output series are true stresses. The optimal input batch size and sequence length are 64 and 2, respectively. Eventually, the predicted results of the proposed Informer deep learning model are more accordant with the tested true stresses compared to those of the previously established physical mechanism model, demonstrating that the Informer deep learning model enjoys an outstanding forecasted capability for precisely reconstructing the high-temperature compressed features of the Ti-55511 alloy.

Keywords: hot deformation; informer deep learning model; microstructural evolution; titanium alloy

1. Introduction

Due to its outstanding properties consisting of mechanical properties, anticorrosive performance, and thermal treatability, near-β titanium alloy is comprehensively applied in the crucial manufacture of load-bearing aircraft components [1,2]. Usually, hot deformation is necessarily utilized to improve the microstructures and further optimize the practical performance of titanium alloys [3–5]. The coupling effects of multiple forming parameters induce intricate evolving characteristics of microstructures and high-temperature flow behavior of titanium alloys [6–10]. Hence, investigations on the microstructural evolution and accurately modeling the true stress–strain characteristics of titanium alloys are significant.

To this day, numerous investigations have been devoted to exploring the microstructural evolution mechanisms of titanium alloys [11–15]. Some reports [16,17] revealed the substructural evolving features for multiple titanium alloys in thermal forming and detected that the substructural nucleated/migration mechanisms were substantially affected by processing parameters. Meanwhile, it was found that the evolution of substructures could exert a prominent effect on the nucleated/coarsening of dynamic recrystallization

(DRX) [18–21]. Additionally, the transformation mechanisms of phases (i.e., α phase globularization [22,23], α phase conversion into β phase [24,25]) were intensively analyzed. As mentioned in previous investigations, intricate microstructural variation/interaction characteristics frequently emerge and notably affect the thermal forming features of titanium alloys.

Describing high-temperature flow characteristics of alloys is a current research subject and obtained tremendous achievements with various constitutive models [26–33]. First, multiple phenomenological models were constructed/improved for reproducing the thermal flow features of alloys [34–38]. Moreover, according to microstructural variations over processing parameters, multitudinous physical mechanism correlation models were constructed for reproducing the thermal flow characteristics of alloys [39–42]. Usually, the above two types of models can score decent prediction results, but it is challenging to formulate appropriate expressions and determine accurate material constants. Therefore, numerous machine learning models were established to simplify the conducting process and had overall superior forecasted results. For instance, the e-insensitive support vector regression (e-SVR) obtained decent results for forecasting flow characteristics [43,44]. Furthermore, complex artificial neural network (ANN) models were leveraged to predict the flow stress of titanium alloys, such as Ti600, Ti60, Ti40, and Ti-2Al-9.2Mo-2Fe β alloys [45–47]. Specifically, Ge et al. [48] leveraged the artificial neural network to propose the accurate constitutive model for the β-γ TiAl alloy. In recent years, various ANN-based deep learning models (DLMs) were developed to be applied in forecasting tasks, e.g., the recurrent neural network (RNN) [49,50] and long short-term memory (LSTM) [51–53]. However, the overfitting issue and long-term predicting performance degradation make them difficult to apply in practical usage. To tackle the problems, the Transformer-based Informer [54] deep learning model was proposed and showed an excellent capability in lithium-ion battery estimation [55]. Therefore, in this study, the two-stage high-temperature forming with variant strain rates in the β region of a Ti-55511 alloy is investigated. The Informer deep learning model was established for characterizing the microstructures and flow features of the Ti-55511 alloy.

Despite the comprehensive investigation of evolving characteristics of flow behaviors and microstructures for titanium alloys in thermal deformation at constant strain rates, systematic investigations of thermally compressed features of titanium alloys under variant strain rates remain lacking. Owing to the influences of sophisticated die structure as well as friction conditions between the die and component, the component commonly undergoes thermal forming with varying strain rates in the actual manufacturing process. Thereby, the stress–strain features for a Ti-55511 alloy in thermal compression with step-like strain rates were investigated. Furthermore, the evolving features of substructures are earnestly analyzed. Additionally, an Informer deep learning model is proposed for reconstructing the thermally compressed features of the Ti-55511 alloy.

2. Experimental Material and Procedure

The commercial near-β titanium alloy was employed in the present investigation. The chemical composition (wt.%) for the researched titanium alloy was 5.16Al-4.92Mo-4.96V-1.10Cr-0.98Fe-(bal.) Ti. Cylindrical specimens (Φ8 mm × 12 mm) for thermal compression were manufactured. The Gleeble-3500 device was employed for constructing the two-stage thermally compressed experiments. Figure 1 reveals the explicit experimental procedures. Distinctly, all forming processes contain two compressed stages (I as well as II). The compressed temperature (T) and the total strain ($\varepsilon_{\text{total}}$) were consistent in two stages. Here, three compressed temperatures (890 °C, 920 °C as well as 950 °C) and the constant value of $\varepsilon_{\text{total}}$ (1.2) were adopted. Still, discrepant strain rates were exploited in each compressed stage. The representative complete compressed experimental step is that the specimen was thermally compressed under the strain rate of the first compression stage ($\dot{\varepsilon}_{\text{I}}$) until the strain of stage I (ε_{I}) was finished, and then thermal compression was executed under the

strain rate of the second compression stage ($\dot{\varepsilon}_{II}$). Correspondingly, three values of ε_I (0.3, 0.6, as well as 0.9) were adopted.
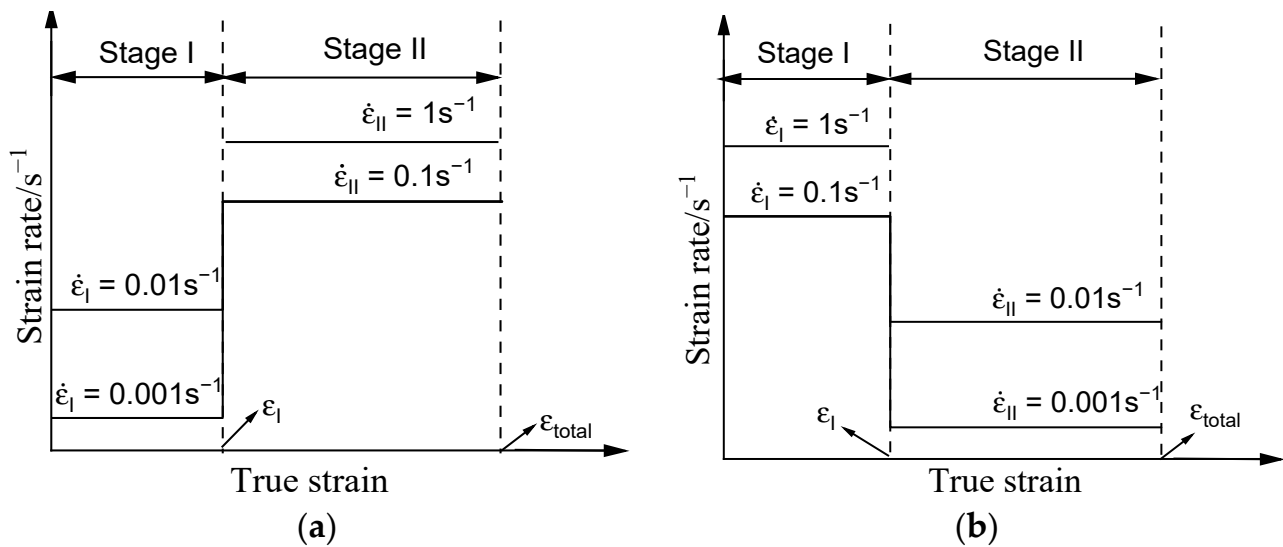

Figure 1. Tested steps of the received titanium alloy: (**a**) type A: the strain rates altered from the relative low values ($\dot{\varepsilon}_I$) to high values ($\dot{\varepsilon}_{II}$); (**b**) type B: the strain rates altered from the relative high values ($\dot{\varepsilon}_I$) to low values ($\dot{\varepsilon}_{II}$).

Before thermal compression, each sample was heated to the compressed temperatures under 10 °C/s and remained at 300 s. When the thermal compressed process was finished, the compressed blocks were directly cooled utilizing water (about 25 °C). To dissect the evolving features of substructures in thermal compression, transmission electron microscopy (TEM) was adopted. To dissect the original microstructure, electron backscatter microscopy (EBSD) was chosen. For analyzing using TEM as well as EBSD, the thermally compressed samples were axially machined for acquiring cross-sections. Afterwards, these sections were ground, polished, and etched in a solution (10 mL $HClO_4$ + 70 mL $C_4H_{10}O$ + 120 mL CH_3OH). Figure 2 displays the original grain structures, and most of the initial β grains are equiaxed grains.
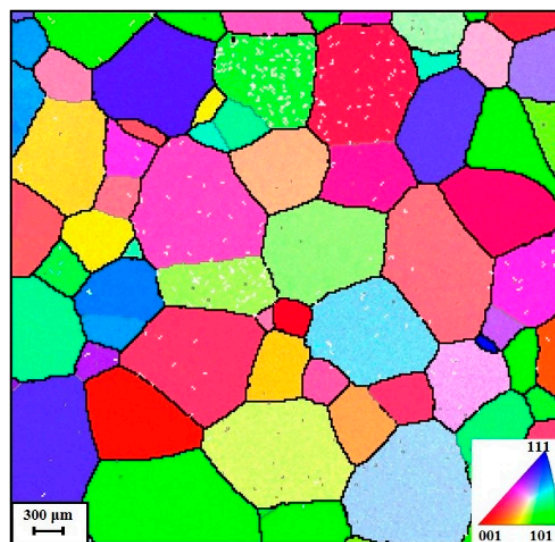

Figure 2. EBSD map of original grain structures in the received titanium alloy.

3. High-Temperature Compression Features and Substructural Evolution

The prime hot flow features of the researched titanium alloy in double-stage hot compression with stepped-strain rates are displayed in Figure 3. Clearly, the high-temperature compression behaviors are markedly affected by compression parameters. As revealed in Figure 3a, the true stresses at the first and second stages of hot compression exhibit a diminishing tendency with rising compression temperature. One principal reason for this experimental result is that the DRX behavior dramatically proceeds as the compressed temperature (T) ascends [6]. Moreover, the visible evolution of substructures occurs with the elevated compression temperature, as depicted in Figure 4a,b. For the compressed temperature of 920 °C and strain rate of 0.01 s^{-1}, the formation of high-density dislocation clusters can be detected (Figure 4a). Then, the prominent work-hardening (WH) effect is inspired owing to the acute interaction of adjacent substructures, and the rise in true stress occurs quickly [6,16]. When the compressed temperature is elevated from 920 °C to 950 °C, the intensive migration/interaction of dislocations and grain boundary occurs, and the substructures are apparently consumed (Figure 4b). Then, the reinforced dynamic softening feature emerges with a rising incompression temperature, and a decrease in true stress appears. Furthermore, the true stress at the second stage of high-temperature compression exhibits a relative increasing trend along with the rise in the strain of the first-stage compression (ε_I), as displayed in Figure 3b. This tested result is primarily ascribed to the weakened DRX development occurring at large values of ε_I, as the strain rate is transferred from a high value ($\dot{\varepsilon}_I = 0.1$ s^{-1}) to a low value ($\dot{\varepsilon}_{II} = 0.001$ s^{-1}) [6]. Meanwhile, the variations of ε_I exerting a significant influence on the substructural evolution are depicted in Figure 4c. From Figure 4a,c, it can be detected that the generation/accumulation of substructures (subgrain, dislocation network, etc.) is promoted with increasing ε_I. Owing to the formation of high-density dislocation networks, the resistance of dislocation slippage, and grain boundary motion is raised, inducing the rise in true stress at the second stage of high-temperature compression.
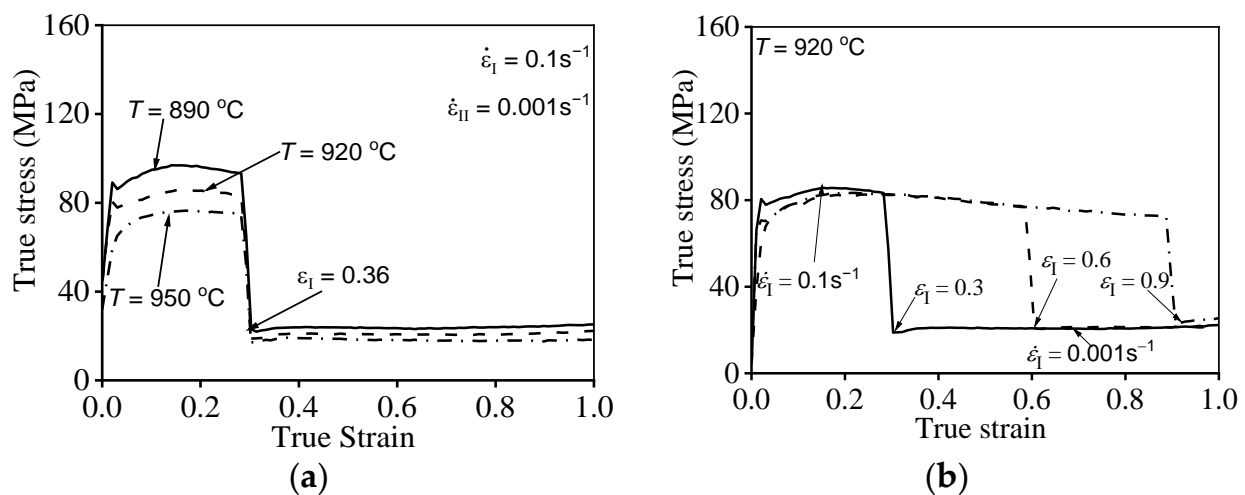

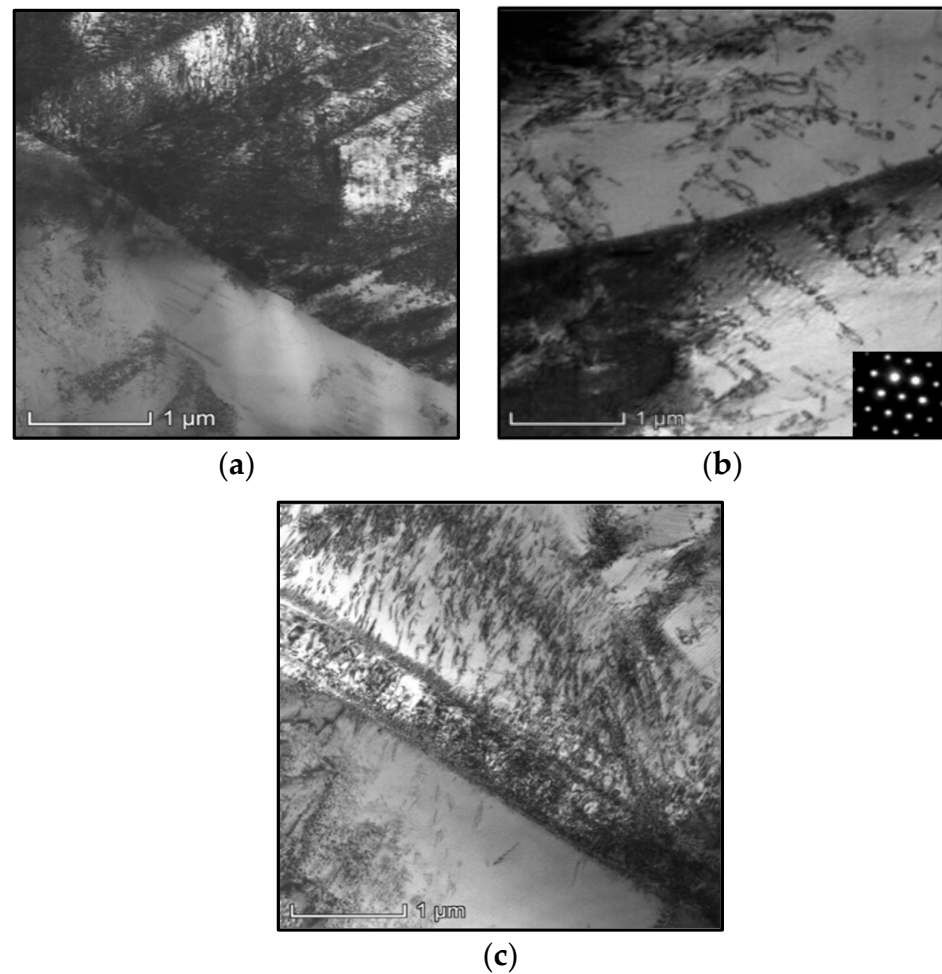**Figure 3.** Representative flow characteristic at variation of: (**a**) T, (**b**) ε_I [6].

Figure 4. TEM figures at: (**a**) T = 920 °C/$\dot{\varepsilon}_I$ = 0.1 s^{-1}/ε_I = 0.36/$\dot{\varepsilon}_{II}$ = 0.001 s^{-1}, (**b**) T = 950 °C/ $\dot{\varepsilon}_I$ = 0.1 s^{-1}/ε_I = 0.36/$\dot{\varepsilon}_{II}$ = 0.001 s^{-1}, (**c**) T = 920 °C/$\dot{\varepsilon}_I$ = 0.1 s^{-1}/ε_I = 0.6/$\dot{\varepsilon}_{II}$ = 0.001 s^{-1}.

4. The Informer Deep Learning Model for Forecasting Hot Flow Features of a Ti-55511 Alloy

In contrast to existing models with lengthy process limitations, the Transformer model demonstrates the operational potential for long sequence prediction, owing to its innovative architecture and self-attention mechanism [56]. Although the canonical self-attention mechanism is capable of processing large-scale data with impressive performance, the high computational complexity and significant memory consumption in stacking layers of the model impede its practical application. To address such a deficiency, optimized models such as the LogSparse Transformer model [57] and similar models [58] were proposed to reduce the original self-attention mechanism complexity, but their efficiency remained limited. Moreover, the Reformer model was embedded with locally sensitive hashing updated self-attention to reduce the complexity in the exceptionally long-term series for each layer [59]. In certain situations, the complexity growth rate of the Informer model was optimized to be linear, but the model could potentially experience degradation in practical long-term prediction [60]. More recently, a continuous-space attention mechanism was deployed in the Infinite Memory Transformer model to free the complexity from input length, but the prediction accuracy was decreased [61].

In summary, previous Transformer models focused on optimizing the complexity of the attention mechanism for each layer and obtained important findings. However, simultaneously cutting down the complexity and breaking the scalability bottleneck of stacking layers is rarely addressed. Therefore, the Informer deep learning model is proposed to address these limitations and accelerate its computing speed [54]. In the present research,

the Informer deep learning model is applied as a practical method for forecasting the flow characteristics of the studied titanium alloy. Specifically, the Informer deep learning model leverages the proposed ProbSparse self-attention mechanism and distilling operation to reduce the memory usage and time complexity of the dependency alignment to $\mathcal{O}(L \log L)$ and the space complexity to $\mathcal{O}((2 - \varepsilon)L \log L)$. During the inference phase, the model utilizes a generative decoder form to avoid cumulative error spreading and optimize long-series output. The Informer deep learning model architecture is shown in Figure 5.
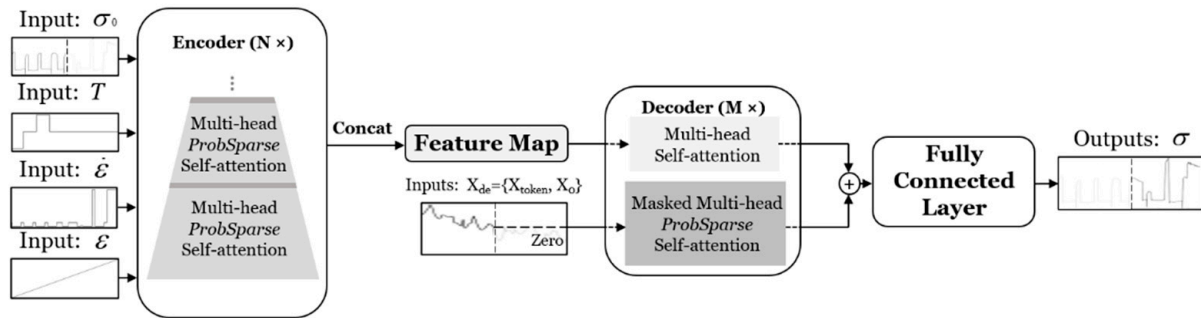

Figure 5. Architecture of the Informer deep learning model.

4.1. ProbSparse Self-Attention Mechanism

With inputs as query, key and value, the original self-attention mechanism is defined as [56],

$$\mathcal{A}(Q, K, V) = \text{Softmax}\left(QK^\top / \sqrt{d}\right)V \tag{1}$$

where $Q \in \mathbb{R}^{L_Q \times d}$, $K \in \mathbb{R}^{L_K \times d}$, $V \in \mathbb{R}^{L_V \times d}$, and d denotes the input dimension.

Derived by [62], the i-th query's attention can be defined with kernel smoothing as,

$$\mathcal{A}(q_i, K, V) = \sum_j \frac{k(q_i, k_j)}{\sum_l k(q_i, k_l)} v_j = \mathbb{E}_{p(k_j|q_i)}\left[v_j\right] \tag{2}$$

where q_i, k_i, v_i stand for the i-th row in Q, K, V, respectively, and $k(q_i, k_j) = \exp\left(q_i k_j^\top / \sqrt{d}\right)$. The part $p(k_j|q_i) = k(q_i, k_j) / \sum_l k(q_i, k_l)$ is conducted to obtain the probability, which entails a large $\mathcal{O}(L_Q L_K)$ memory usage. Therefore, the Informer deep learning model proposed the query sparsity measurement to tackle this major defect of self-attention.

The similarity between p and q can be used to distinguish the importance, which can be conducted through Kullback–Leibler divergence as,

$$KL(q\|p) = \ln \sum_{l=1}^{L_K} e^{q_i k_l^\top / \sqrt{d}} - \frac{1}{L_K} \sum_{j=1}^{L_K} q_i k_j^\top / \sqrt{d} - \ln L_K \tag{3}$$

The measurement of the i-th query is defined by dropping the constant as,

$$M(q_i, K) = \ln \sum_{j=1}^{L_K} e^{\frac{q_i k_j^\top}{\sqrt{d}}} - \frac{1}{L_K} \sum_{j=1}^{L_K} \frac{q_i k_j^\top}{\sqrt{d}} \tag{4}$$

where the formula calculates the Log-Sum-Exp (LSE) and the arithmetic mean of all keys [63]. If $M(q_i, K)$ grows larger, the probability p becomes more principal factor alterable, thus having a superior differentiating capability.

According to the above measurement, the ProbSparse self-attention mechanism can be further conducted by distributing keys to Top-u queries as

$$\mathcal{A}(Q, K, V) = \text{Softmax}\left(\frac{\overline{Q}K^\top}{\sqrt{d}}\right)V \tag{5}$$

where $\overline{Q}$ is the q-size sparse matrix. When $u = c \cdot \ln L_Q$, the layer memory usage is reduced to $\mathcal{O}(L_K \ln L_Q)$ due to the lessened calculation for each key.

Nevertheless, the query sparsity measurement needs quadratic $\mathcal{O}(L_Q L_K)$ calculation, and the LSE implement is not constantly numerically stable. Hence, an empirical approximation is conducted.

For each q_i, the discrete keys can be converted to continuous ones as vector k_j. In addition, the first term of the $M(q_i, K)$ becomes the LSE of the inner product of a fixed query q_i and all the keys, and define

$$f_i(K) = \ln \sum_{j=1}^{L_K} e^{q_i k_j^\top / \sqrt{d}} \tag{6}$$

From the Log-Sum-Exp network and relative studies [63,64], the convex function $f_i(K)$ combines linear k_j for q_i, making $M(q_i, K)$ convex. Hence, the measurement can be conducted to a derivation form with each vector k_j as follows,

$$\frac{\partial M(q_i, K)}{\partial k_j} = \frac{e^{q_i k_j^\top / \sqrt{d}}}{\sum_{j=1}^{L_K} e^{q_i k_j^\top / \sqrt{d}}} \cdot \frac{q_i}{\sqrt{d}} - \frac{1}{L_K} \cdot \frac{q_i}{\sqrt{d}} \tag{7}$$

Let $\vec{\nabla} M(q_i) = \vec{0}$ to reach the minimum value; the condition can be listed as,

$$q_i k_1^\top + \ln L_K = \cdots = q_i k_j^\top + \ln L_K = \cdots = \ln \sum_{j=1}^{L_K} e^{q_i k_j^\top} \tag{8}$$

The minimum value $\ln L_K$ can be obtained when $k_1 = k_2 = \cdots = k_{L_K}$. Therefore, the measurement can be written as

$$M(q_i, K) \geq \ln L_K \tag{9}$$

Hence, by picking the largest inner-product $\max_j \{ q_i k_j^\top / \sqrt{d} \}$, the inequation can be derived as

$$\begin{cases} \mathrm{M}(q_i, K) = \ln \sum_{j=1}^{L_K} e^{\frac{q_i k_j^\top}{\sqrt{d}}} - \frac{1}{L_K} \sum_{j=1}^{L_K} \left(\frac{q_i k_j^\top}{\sqrt{d}} \right) \\ \qquad \leq \ln \left(L_K \cdot \max_j \left\{ \frac{q_i k_j^\top}{\sqrt{d}} \right\} \right) - \frac{1}{L_K} \sum_{j=1}^{L_K} \left(\frac{q_i k_j^\top}{\sqrt{d}} \right) \\ \qquad = \ln L_K + \max_j \left\{ \frac{q_i k_j^\top}{\sqrt{d}} \right\} - \frac{1}{L_K} \sum_{j=1}^{L_K} \left(\frac{q_i k_j^\top}{\sqrt{d}} \right) \end{cases} \tag{10}$$

Eventually, by combining the above equations, the bound can be denoted as

$$\ln L_K \leq M(q_i, K) \leq \max_j \{ q_i k_j^\top / \sqrt{d} \} - \frac{1}{L_K} \sum_{j=1}^{L_K} \{ q_i k_j^\top / \sqrt{d} \} + \ln L_K \tag{11}$$

where $q_i \in \mathbb{R}^d$ and $k_j \in \mathbb{R}^d$ are in the keys set K.

From the above deductions, the max–mean measurement can be defined as

$$\overline{M}(q_i, K) = \max_j \left\{ \frac{q_i k_j^\top}{\sqrt{d}} \right\} - \frac{1}{L_K} \sum_{j=1}^{L_K} \frac{q_i k_j^\top}{\sqrt{d}} \tag{12}$$

Specifically, a long-tail distribution pattern of the self-attention mechanism was observed by performing a qualitative assessment [54]. In this case, only a few dot product pairs contribute to the major attention. Hence, $\overline{M}(q_i, K)$ only requires $U = L_K \ln L_Q$ dot product pairs of random sampling, and the remaining pairs are filled with zero values. Therefore, the operation has a weaker sensitivity and remains numerically stable. Eventually, in practical application, the relatively equivalent input length $L_Q = L_K = L$ in self-attention computation can reduce the complexity to $\mathcal{O}(L \ln L)$.

4.2. Encoder

The Informer deep learning model utilizes the encoder architecture to extract the long-term dependency of input series, where the t-th input X^t is reshaped as matrix $X_{\text{en}}^t \in \mathbb{R}^{LX \times d_{\text{model}}}$ [56]. The encoder is composed of multiple identical layers stacked on top of each other. Specifically, the architecture of a single stack in the encoder of the Informer deep learning model is given in Figure 6.
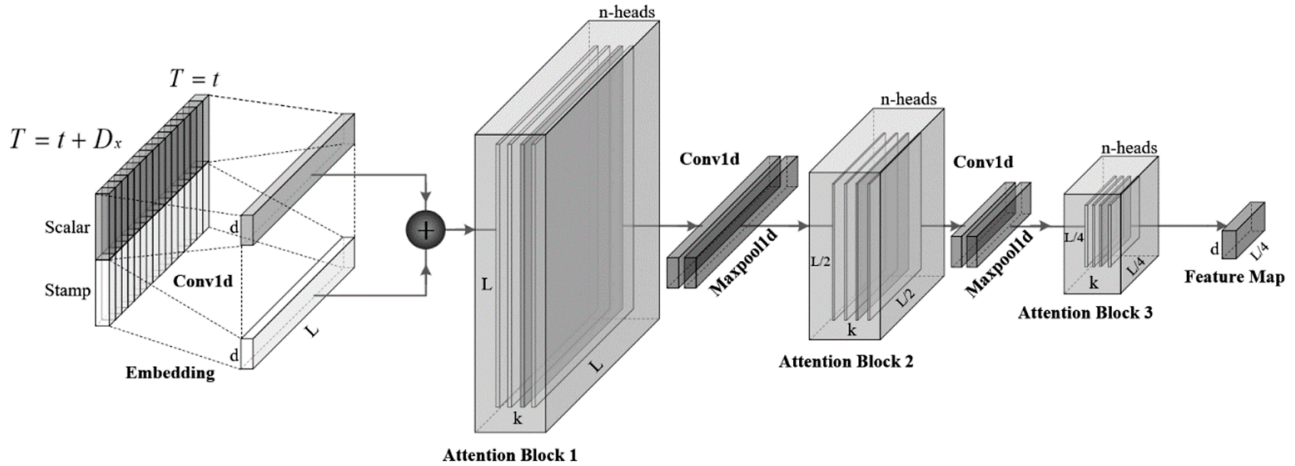

Figure 6. The architecture of a single stack in the encoder of the Informer deep learning model.

Due to the processing of the ProbSparse self-attention mechanism, the encoder is loaded with redundant value V combinations. Hence, self-attention distilling is proposed to concentrate self-attention mechanisms for the next layer.

Based on the dilated convolution [65], the distilling operation feeds forwards the $(j + 1)$-th layer as,

$$X_{j+1}^t = \text{MaxPool}\left(\text{ELU}\left(\text{Conv1d}\left(\left[X_j^t\right]_{\text{AB}}\right)\right)\right) \tag{13}$$

where $[\cdot]_{\text{AB}}$ denotes the attention block, and $\text{Conv1d}(\cdot)$ generates a 1D convolutional filter with $\text{ELU}(\cdot)$ activation function [66].

The max-pooling layer is added to reduce the total memory usage to $\mathcal{O}((2 - \varepsilon)L \log L)$. Furthermore, a pyramid-like processing structure (shown in Figure 6) is established where inputs are halved to serve as the replication of the main stack and the distilling layers drop gradually. In this case, the operation has a better robustness, and the resulting dimensions of different layers are consistent.

4.3. Decoder

The canonical decoder structure is optimized with generative inference to mitigate the long-term speed descent. The decoder mechanism is defined as

$$X_{\text{de}}^t = \text{Concat}\left(X_{\text{token}}^t, X_0^t\right) \in \mathbb{R}^{(L_{\text{token}} + L_y) \times d_{\text{model}}} \tag{14}$$

where $X_{\text{token}}^t \in \mathbb{R}^{L_{\text{token}} \times d_{\text{model}}}$ is the start token, and $X_0^t \in \mathbb{R}^{L_y \times d_{\text{model}}}$ is the placeholder for target sequences.

Extended from dynamic decoding [67], the procedure is innovated to sample a L_{token} series in the input sequence as a start token then feed it to the decoder as $X_{\text{de}} = \{X_{\text{L}}, X_0\}$. Afterwards, the decoder obtains outputs through a single forward procedure, and thus it can process with less time consumption than a normal encoder-decoder architecture.

4.4. Identification for the Parameters of the Informer Deep Learning Model

The inputs of the Informer deep learning model are temperature T = {890, 920, 950} °C, true strain ε = {0~1}, and strain rate $\dot{\varepsilon}$ = {0.001, 0.01, 0.1, 1} s^{-1}. The input sequences were preprocessed by concatenating experimental data of true stress values under different temperatures, true strains, and strain rates. The corresponding temperature, true strain, and strain rate values were also concatenated in the sequences. Then, these sequences were applied as training inputs. The experimental data are shuffled using 7/10 of the total amount for training and the rest for testing and validating the model.

As discussed above regarding the architecture of the Informer deep learning model in Sections 4.1–4.3 and the features of general deep neural networks, the Informer deep learning model should first be established by tuning hyper-parameters such as learning rate, input batch size, dropout, etc. To obtain the optimal parameters, the correction coefficient R, average absolute relative error $AARE$, mean squared error MSE, and root-mean squared error $RMSE$ assessment criteria are employed for evaluating the results.

$$R = \frac{\sum_{i=1}^{N}\left(M_i - \overset{\bullet}{M}\right)\left(P_i - \overset{\bullet}{P}\right)}{\sqrt{\sum_{i=1}^{N}\left(M_i - \overset{\bullet}{M}\right)^2 \sum_{i=1}^{N}\left(P_i - \overset{\bullet}{P}\right)^2}} \tag{15}$$

$$AARE(\%) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{M_i - P_i}{M_i}\right| \times 100\% \tag{16}$$

$$MSE = \sum_{i=1}^{N}\frac{(M_i - P_i)^2}{N} \tag{17}$$

$$RMSE = \sqrt{\sum_{i=1}^{N}\frac{(M_i - P_i)^2}{N}} \tag{18}$$

where N notes the total amount of result data, and M_i and P_i stand for the measured and predicted results when $\overset{\bullet}{M}$ and $\overset{\bullet}{P}$ are the mean values, respectively.

Generally, the accuracy and generalization ability of deep learning models are affected by various hyper-parameters. In the case of forecasting, the batch size of input sequences and the initial learning rate of the model play crucial roles. On the one hand, a larger batch size allows faster training but may result in worse model accuracy and an unstable training process [68]. On the other hand, a smaller batch size is beneficial for generalization but can lead to a longer computation time [69]. Additionally, both theoretical and empirical evidence have proven that the batch size and learning rate significantly impact the generalization ability and accuracy of the deep learning model [70–72]. To further explore the relationship between the two parameters and the results, experimental curves are displayed in Figure 7. The five curves represent the effect of the learning rate on validation loss under different batch sizes. Specifically, the learning rate is tested in a uniformly spaced range from 10^{-1} to 10^{-6} with batch sizes of 8, 16, 32, 64, and 128, respectively. The model accuracy is evaluated by the validation loss.

It is clear that the validation loss of the Informer deep learning model drops to a minimum value and then starts to fluctuate when the learning rate increases from 10^{-4} to 2×10^{-3}. As the learning rate further ascends, the fluctuation of validation loss becomes intense, and thus the optimal learning rate can be chosen as 1.2929×10^{-3}. Specifically, the curve fluctuations in Figure 7 demonstrate an appropriate balance of model accuracy and training stability under the batch size of 64. Hence, the batch size is determined as 64.
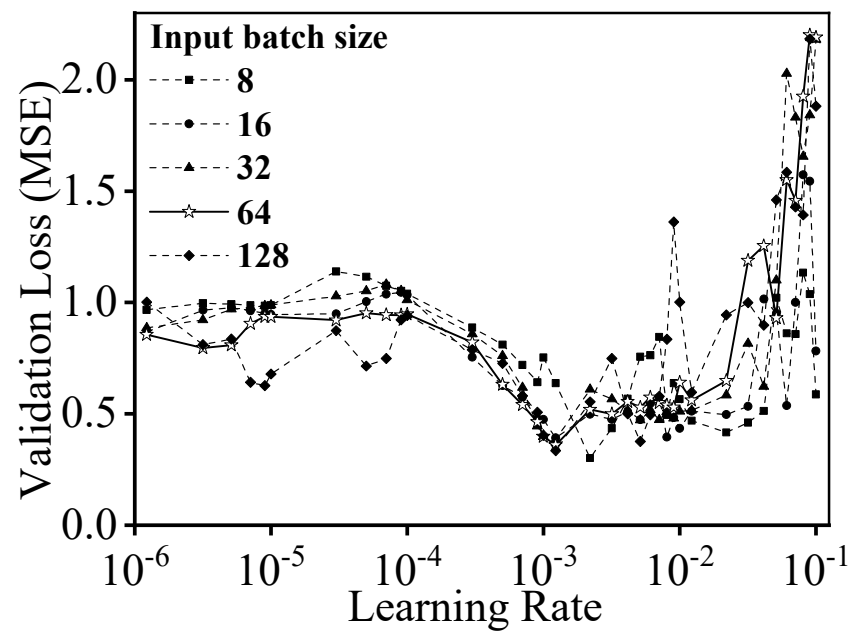
Figure 7. Variations of validation loss under different tested batch sizes and learning rates.

In addition, it is important to note that the parameters of sequence length and label length also have a significant impact on accuracy. Based on experimental results, the optimal sequence length and label length are identified as two and one, respectively.

Eventually, the values of *R*, *AARE*, and *RMRE* can be computed as 0.9986, 4.191%, and 2.2016, respectively. According to the results, the performance of the Informer deep learning model is shown in Figure 8. It illustrates good consistency between the experimental data and the modeled results, demonstrating the great capability of the Informer deep learning model to describe the high-temperature deformation features of the researched titanium alloy.
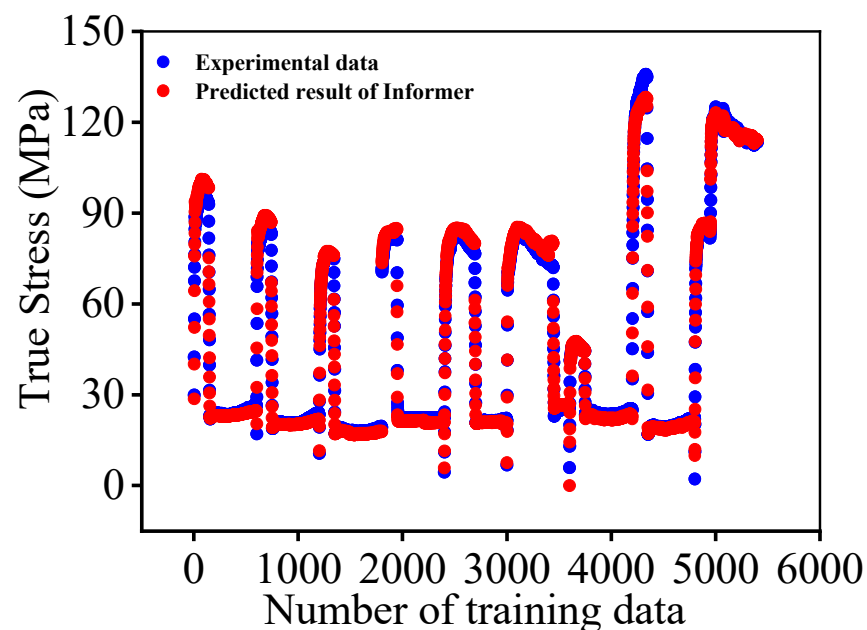

Figure 8. Performance of the Informer deep learning model.

4.5. Comparisons and Discussion

As shown in the above sections, the Informer deep learning model exhibits a strong forecasting ability for the true stress of the researched titanium alloy. According to the author's previous investigation [6], a physical mechanism (PM) model was constructed for forecasting the true stress of the researched titanium alloy, i.e.,

$$
\begin{cases}
\sigma = \sigma_{\mathrm{y}} + \sigma_{\rho} \\
\sigma_{\mathrm{y}} = 1.589\left(\dot{\varepsilon}\exp\left(\frac{205{,}800}{RT}\right)\right)^{0.2052} \\
\sigma_{\rho} = 4.38 \times 10^{-10}(21.8847 - 0.0153T)\sqrt{\rho_i} \\
\dot{\rho}_i = \dot{\rho}_i{}^{+} - \dot{\rho}_i{}^{\mathrm{DRV}} - \dot{\rho}_i{}^{\mathrm{DRX}} \\
\dot{\rho}_i{}^{+} = \frac{1}{2.86 \times 10^{-10}\Lambda}\dot{\varepsilon} \\
\frac{1}{\Lambda} = \frac{1}{s} + \frac{1}{d_i} \\
s = \frac{F_{\mathrm{s}}}{\sqrt{\rho_i}} \\
F_{\mathrm{s}} = 4.1797\ \left(\dot{\varepsilon}\exp\left(\frac{-6.4278}{RT}\right)\right)^{-0.2843} \\
\dot{d}_{\mathrm{g}} = 2.3866d^{\,0.4468} \\
\dot{d}_{\mathrm{x}} = -0.7803d^{0.0072}\dot{X}^{0.9906} \\
\dot{d}_{\mathrm{i}} = \dot{d}_{\mathrm{x}} + \dot{d}_{\mathrm{g}} \\
\dot{\rho}_i{}^{\mathrm{DRV}} = 47.0321\left(\dot{\varepsilon}\exp\left(\frac{0.04718}{RT}\right)\right)^{-0.1259}\rho \\
\dot{\rho}_i{}^{\mathrm{DRX}} = \frac{0.7313\left(\dot{\varepsilon}\exp\left(\frac{-9.1687}{RT}\right)\right)^{0.0137}\dot{X}\rho_i}{(1-X)^{2.0210}} \\
\dot{X} = \frac{6.0877M_{\mathrm{b}}P[X(1-X)]^{-1.4294}\dot{\varepsilon}^{-2.6513}}{d^{0.5733}} \\
M_{\mathrm{b}} = \frac{1.54 \times 10^{-26}}{\mathrm{k}T}\left[\dot{\varepsilon}\exp\left(\frac{-0.1418}{RT}\right)\right]^{0.0045} \\
\mathrm{P} = \frac{8.18 \times 10^{-20}\rho_i(21.8847 - 0.0153T)}{2}
\end{cases}
\tag{19}
$$

where σ is the flow stress, σ_y is the short-range component, and σ_{ρ} is the dislocation interaction stress. $\dot{\varepsilon}$ is the strain rate, R is the gas constant, T is the absolute temperature, ρ_i is the dislocation density, $\dot{\rho}_i{}^{+}$ is the dislocation density emergence rate under WH, and $\dot{\rho}_i{}^{\mathrm{DRV}}$ and $\dot{\rho}_i{}^{\mathrm{DRX}}$ are dislocation density variation rate of DRV and DRX, respectively. Λ is the mean-free path of dislocation, and d_i is the average grain size. X is the DRX fraction and the rate $\dot{X}$. M_{b} is the grain boundary movement rate. P is the driving force. D_{ob} is the factor of self-diffusion, and δ is the grain boundary thickness.

Figure 9 unveils the comparative analysis of forecasting performances between the PM and Informer deep learning model. Compared to that of the PM model, the Informer deep learning model enjoys a smaller forecasting error of true stresses, particularly for the researched titanium alloy at lower compressed temperature (890 °C) or higher $\dot{\varepsilon}_{\mathrm{I}}/\dot{\varepsilon}_{\mathrm{II}}$. To validate the forecasting capability, the correlation results of forecasted true stresses and tested ones are plotted in Figure 10. Clearly, the scatters of the PM constitutive model are more dispersed, while those of the Informer deep learning model are more centralized. Meanwhile, the values of *R*, *AARE,* and *RMSE* are determined, as noted in Table 1. Distinctly, the relative larger *R* as well as the smaller *AARE* and *RMSE* values imply that the established Informer deep learning model can accurately depict the hot compressed features of the Ti-55511 alloy.
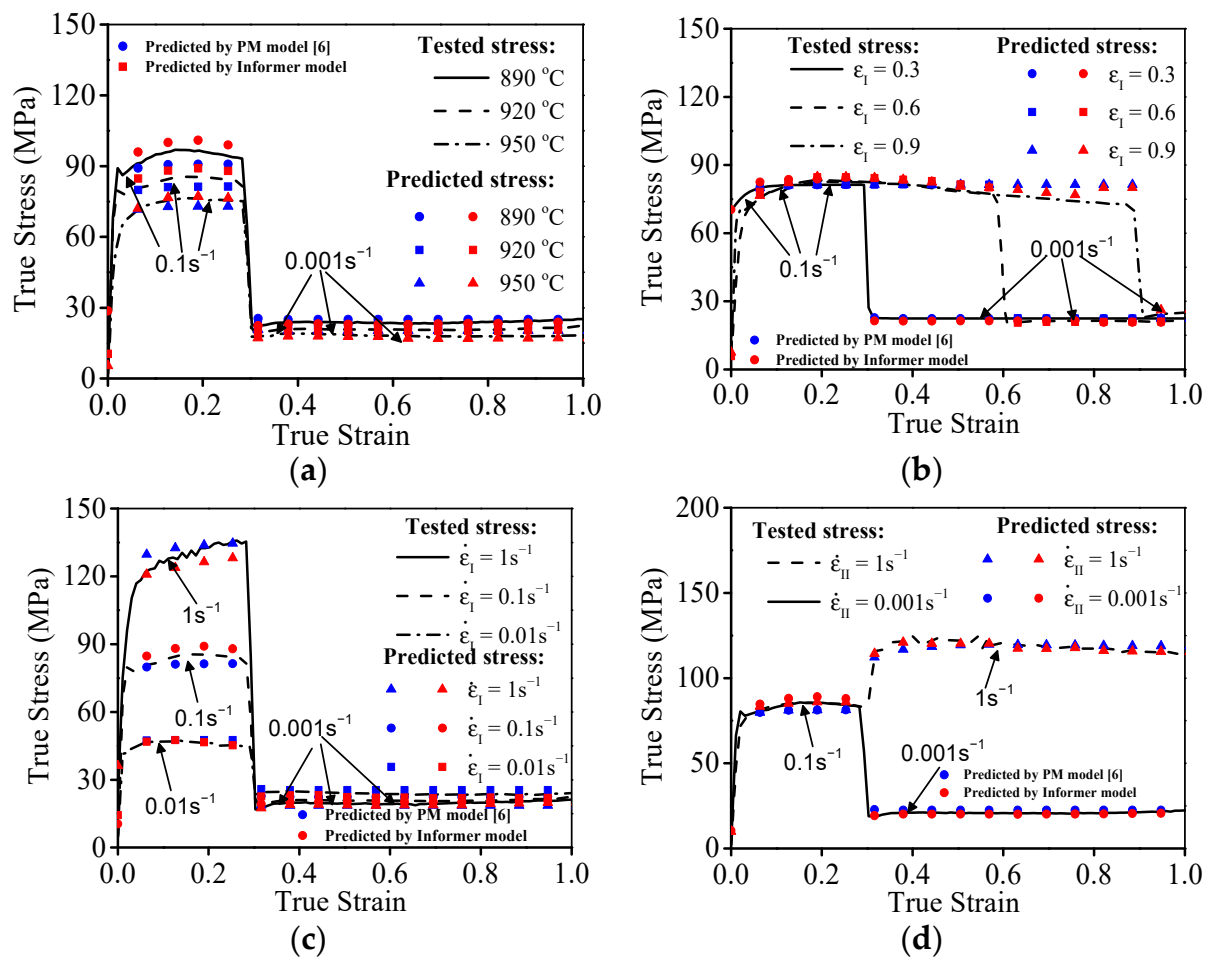
Figure 9. Comparisons of tested true stress and predicted results at: (**a**) T, (**b**) ε_I, (**c**) $\dot{\varepsilon}_I$, (**d**) $\dot{\varepsilon}_{II}$.
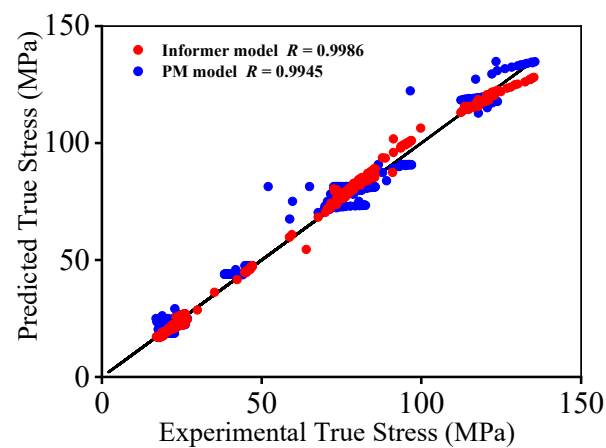

Figure 10. Correlation of tested true stresses and predicted values.

Table 1. Calculated assessment values of the Informer deep learning model and PM constitutive model.

Model	R	$AARE$(%)	$RMSE$
PM model [6]	0.9945	6.181%	3.7448
Informer deep learning model	0.9986	4.191%	2.0615

5. Conclusions

The evolving characteristics of microstructures as well as flow behavior for a Ti-55511 alloy in two-stage thermal compression experiments with step-like strain rates are researched. The decisive conclusions are drawn as:

(1) In high-temperature compression, the influences of forming parameters on the flow behaviors of the researched Ti-55511 alloy are significant. Flow stresses are reduced with the increase in compressed temperature. Notwithstanding, flow stresses at stage II of thermal compression display an increase trend with the descent of ε_{I} or increase in $\dot{\varepsilon}_{\mathrm{I}}/\dot{\varepsilon}_{\mathrm{II}}$;

(2) The formation of high-density networks/clusters through dislocation concentration/interaction is suppressed with the increase in compressed temperature. Nevertheless, the dislocation nucleation/concentration is enhanced with the increase in ε_{I};

(3) The Informer deep learning model is developed to reconstruct the thermal compressed characteristics of the researched Ti-55511 alloy. The considerable agreement between the predicted true stresses and experimental results demonstrates the high prediction accuracy of the Informer deep learning model.

References

1. Lin, Y.C.; Pang, G.-D.; Jiang, Y.-Q.; Liu, X.-G.; Zhang, X.-Y.; Chen, C.; Zhou, K.-C. Hot compressive deformation behavior and microstructure evolution of a Ti-55511 alloy with basket-weave microstructures. *Vacuum* **2019**, *169*, 108878. [CrossRef]

2. Tan, K.; Li, J.; Guan, Z.; Yang, J.; Shu, J. The identification of dynamic recrystallization and constitutive modeling during hot deformation of Ti55511 titanium alloy. *Mater. Des.* **2015**, *84*, 204–211. [CrossRef]

3. Bobbili, R.; Madhu, V. Constitutive modeling of dynamic flow behavior of Ti-5553 alloy. *J. Alloys Compd.* **2019**, *787*, 260–266. [CrossRef]

4. Li, C.W.; Xie, H.; Mao, X.N.; Zhang, P.S.; Hou, Z.M. High Temperature Deformation of TC18 Titanium Alloy. *Rare Metal. Mat. Eng.* **2017**, *46*, 326–332.

5. Zhang, J.; Wang, Y. Tension Behavior of Ti–6.6 Al–3.3 Mo–1.8 Zr–0.29 Si Alloy over a Wide Range of Strain Rates. *Mater. Lett.* **2014**, *124*, 113–116. [CrossRef]

6. He, D.-G.; Su, G.; Lin, Y.-C.; Jiang, Y.-Q.; Li, Z.; Chen, Z.-J.; Yan, X.-T.; Xia, Y.-C.; Xie, Y.-C. Microstructural Variation and a Physical Mechanism Model for a Ti-55511 Alloy during Double-Stage Hot Deformation with Stepped Strain Rates in the β Region. *Materials* **2021**, *14*, 6371. [CrossRef] [PubMed]

7. Quan, G.; Pu, S.; Wen, H.; Zou, Z.; Zhou, J. Quantitative Analysis of Dynamic Softening Behaviors Induced by Dynamic Recrystallization for Ti-10V-2Fe-2Al Alloy. *High Temp. Mater. Process.* **2015**, *34*, 549–561. [CrossRef]

8. Liang, H.; Guo, H. The integrated influence on hot deformation of dual-phase titanium alloys incorporating dynamic recrystallization evolution and α/β phase transformation. *Mater. Lett.* **2015**, *151*, 57–60. [CrossRef]

9. Li, C.; Zhang, X.-Y.; Li, Z.-Y.; Zhou, K.-C. Hot Deformation of Ti-5Al-5Mo-5V-1Cr-1Fe Near β Titanium Alloys Containing Thin and Thick Lamellar α Phase. *Mater. Sci. Eng. A* **2013**, *573*, 75–83. [CrossRef]

10. Kar, S.K.; Ghosh, A.; Fulzele, N.; Bhattacharjee, A. Quantitative microstructural characterization of a near beta Ti alloy, Ti-5553 under different processing conditions. *Mater. Charact.* **2013**, *81*, 37–48. [CrossRef]

11. Ning, Y.Q.; Xie, B.C.; Liang, H.Q.; Li, H.; Yang, X.M.; Guo, H.Z. Dynamic Softening Behavior of TC18 Titanium Alloy during Hot Deformation. *Mater. Des.* **2015**, *71*, 68–77. [CrossRef]

12. Lin, Y.C.; Zhao, C.-Y.; Chen, M.-S.; Chen, D.-D. A novel constitutive model for hot deformation behaviors of Ti–6Al–4V alloy based on probabilistic method. *Appl. Phys. A* **2016**, *122*, 716. [CrossRef]

13. Kotkunde, N.; Krishna, G.; Shenoy, S.K.; Gupta, A.K.; Singh, S.K. Experimental and theoretical investigation of forming limit diagram for Ti-6Al-4 V alloy at warm condition. *Int. J. Mater. Form.* **2017**, *10*, 255–266. [CrossRef]

14. Yang, Z.; Xu, W.; Zhang, W.; Chen, Y.; Shan, D. Effect of power spinning and heat treatment on microstructure evolution and mechanical properties of duplex low-cost titanium alloy. *J. Mater. Sci. Technol.* **2023**, *136*, 121–139. [CrossRef]

15. Lin, Y.C.; Jiang, X.-Y.; Shuai, C.-J.; Zhao, C.-Y.; He, D.-G.; Chen, M.-S.; Chen, C. Effects of initial microstructures on hot tensile deformation behaviors and fracture characteristics of Ti-6Al-4V alloy. *Mater. Sci. Eng. A* **2018**, *711*, 293–302. [CrossRef]

16. Wu, C.; Zhou, Y.J.; Liu, B. Experimental and simulated investigation of the deformation behavior and microstructural evolution of Ti6554 titanium alloy during an electropulsing-assisted microtension process. *Mater. Sci. Eng. A* **2022**, *838*, 142745. [CrossRef]

17. Li, L.; Liu, J.; Ding, N.; Li, M. Substructure evolution in two phases based constitutive model for hot deformation of TC18 in α + β phase region. *Chin. J. Aeronaut.* **2023**, *36*, 573–588. [CrossRef]

18. Li, C.; Huang, L.; Zhao, M.; Guo, S.; Su, Y.; Li, J. Characterization of hot workability of Ti-6Cr-5Mo-5V-4Al alloy based on hot processing map and microstructure evolution. *J. Alloys Compd.* **2022**, *905*, 164161. [CrossRef]

19. Lu, T.; Dan, Z.-H.; Li, K.; Yi, D.-Q.; Zhou, L.; Chang, H. Hot deformation behaviors and dynamic recrystallization mechanism of Ti-35421 alloy in β single field. *Trans. Nonferrous Met. Soc. China* **2022**, *32*, 2889–2907. [CrossRef]

20. Huang, L.; Li, C.-M.; Li, C.-L.; Hui, S.-X.; Yu, Y.; Zhao, M.-J.; Guo, S.-Q.; Li, J.-J. Research progress on microstructure evolution and hot processing maps of high strength β titanium alloys during hot deformation. *Trans. Nonferrous Met. Soc. China* **2022**, *32*, 3835–3859. [CrossRef]

21. Abbasi, S.; Momeni, A.; Lin, Y.C.; Jafarian, H. Dynamic softening mechanism in Ti-13V-11Cr-3Al beta Ti alloy during hot compressive deformation. *Mater. Sci. Eng. A* **2016**, *665*, 154–160. [CrossRef]

22. Kumar, V.A.; Murty, S.; Gupta, R.; Rao, A.G.; Prasad, M. Effect of boron on microstructure evolution and hot tensile deformation behavior of Ti-5Al-5V-5Mo-1Cr-1Fe alloy. *J. Alloys Compd.* **2020**, *831*, 154672. [CrossRef]

23. Liu, H.; Wang, Q.; Zhang, J.; Xu, K.; Xue, Y. Effect of multi-pass deformation on hot flow behavior and microstructure evolution mechanism of Ti–6Al–4V alloy fabricated by hot isostatic pressing. *J. Mater. Res. Technol.* **2022**, *17*, 2229–2248. [CrossRef]

24. Yu, Y.; Yan, H.; Chen, J.; Xia, W.; Su, B.; Ding, T.; Li, Z.; Song, M. Flow behavior and dynamic transformation of bimodal TC17 titanium alloy during high strain rate hot compression. *J. Alloys Compd.* **2022**, *912*, 165260. [CrossRef]

25. Chen, X.; Tang, B.; Wei, B.; Zhang, X.; Li, J. Investigation on recrystallization behavior of Ti-47Al-1.5Re-X (Cr, Mn, V, Nb) alloy during hot deformation. *Mater. Lett.* **2023**, *331*, 133484. [CrossRef]

26. Mirzadeh, H. Constitutive Description of 7075 Aluminum Alloy During Hot Deformation by Apparent and Physically-Based Approaches. *J. Mater. Eng. Perform.* **2015**, *24*, 1095–1099. [CrossRef]

27. He, D.-G.; Lin, Y.C.; Wang, L.-H.; Wu, Q.; Zu, Z.-H.; Cheng, H. Influences of pre-precipitated δ phase on microstructures and hot compressive deformation features of a nickel-based superalloy. *Vacuum* **2019**, *161*, 242–250. [CrossRef]

28. Khodashenas, H.; Mirzadeh, H.; Malekan, M.; Emamy, M. Constitutive Modeling of Flow Stress during Hot Deformation of Sn–Al–Zn–Cu–Mg Multi-Principal-Element Alloy. *Vacuum* **2019**, *170*, 108970. [CrossRef]

29. Xia, Q.; Yuan, S.; Xiao, G.; Long, J.; Cheng, X. Meso-modelling study of the mechanical response and texture evolution of magnesium alloy during hot compression. *Mater. Today Commun.* **2021**, *27*, 102469. [CrossRef]

30. Long, J.; Xia, Q.; Xiao, G.; Qin, Y.; Yuan, S. Flow characterization of magnesium alloy ZK61 during hot deformation with improved constitutive equations and using activation energy maps. *Int. J. Mech. Sci.* **2021**, *191*, 106069. [CrossRef]

31. Wen, D.; Gao, C.; Zheng, Z.; Wang, K.; Xiong, Y.; Wang, J.; Li, J. Hot tensile behavior of a low-alloyed ultrahigh strength steel: Fracture mechanism and physically-based constitutive model. *J. Mater. Res. Technol.* **2021**, *13*, 1684–1697. [CrossRef]

32. Tang, C.; Liu, W.; Chen, Y.; Liu, X.; Deng, Y. Hot Deformation Behavior of a Differential Pressure Casting Mg-8Gd-4Y-Nd-Zr Alloy. *J. Mater. Eng. Perform.* **2016**, *26*, 383–391. [CrossRef]

33. Tian, X.; Chen, F.; Jiang, J.; Wu, G.; Cui, Z.; Qian, D.; Han, X.; Wang, B.; Wang, H.; Wang, H.; et al. Experimental analyses and numerical modeling of the microstructure evolution of aluminum alloy using an internal state variable plasticity-based approach coupled with the effects of second phase. *Int. J. Plast.* **2022**, *158*, 103416. [CrossRef]

34. Chen, X.-M.; Lin, Y.C.; Hu, H.-W.; Luo, S.-C.; Zhou, X.-J.; Huang, Y. An Enhanced Johnson–Cook Model for Hot Compressed A356 Aluminum Alloy. *Adv. Eng. Mater.* **2021**, *23*, 2000704. [CrossRef]

35. Lin, Y.C.; Huang, J.; Li, H.-B.; Chen, D.-D. Phase transformation and constitutive models of a hot compressed TC18 titanium alloy in the α+β regime. *Vacuum* **2018**, *157*, 83–91. [CrossRef]

36. He, D.; Chen, S.-B.; Lin, Y.C.; Xie, H.; Li, C. Hot tensile behavior of a 7046-aluminum alloy: Fracture mechanisms and constitutive models. *Mater. Today Commun.* **2023**, *34*, 105209. [CrossRef]
37. Pang, G.D.; Lin, Y.C.; Qiu, Y.L.; Jiang, Y.Q.; Xiao, Y.W.; Chen, M.S. Dislocation Density–Based Model and Stacked Auto-Encoder Model for Ti-55511 Alloy with Basket-Weave Microstructures Deformed in A+ β Region. *Adv. Eng. Mater.* **2021**, *23*, 2001307. [CrossRef]
38. Wen, D.; Yue, T.; Xiong, Y.; Wang, K.; Wang, J.; Zheng, Z.; Li, J. High-temperature tensile characteristics and constitutive models of ultrahigh strength steel. *Mater. Sci. Eng. A* **2021**, *803*, 140491. [CrossRef]
39. Yu, Z.; Ma, Q.; Su, X.; Lai, X.; Tibbenham, P. Constitutive modeling for large deformation behavior of thermoplastic olefin. *Mater. Des.* **2010**, *31*, 1881–1886. [CrossRef]
40. Fan, X.; Yang, H. Internal-state-variable based self-consistent constitutive modeling for hot working of two-phase titanium alloys coupling microstructure evolution. *Int. J. Plast.* **2011**, *27*, 1833–1852. [CrossRef]
41. He, D.; Yan, X.-T.; Lin, Y.C.; Zhang, S.; Chen, Z.-J. Microstructure evolution and constitutive model for a Ni-Mo-Cr base alloy in double-stages hot compression with step-strain rates. *Mater. Charact.* **2022**, *194*, 112385. [CrossRef]
42. Chen, F.; Wang, H.; Zhu, H.; Zhu, H.; Ren, F.; Cui, Z. High-temperature deformation mechanisms and physical-based constitutive modeling of ultra-supercritical rotor steel. *J. Manuf. Process.* **2019**, *38*, 223–234. [CrossRef]
43. He, D.-G.; Lin, Y.C.; Chen, J.; Chen, D.-D.; Huang, J.; Tang, Y.; Chen, M.-S. Microstructural evolution and support vector regression model for an aged Ni-based superalloy during two-stage hot forming with stepped strain rates. *Mater. Des.* **2018**, *154*, 51–62. [CrossRef]
44. Quan, G.-Z.; Zhang, Z.-H.; Zhou, Y.; Wang, T.; Xia, Y.-F. Numerical Description of Hot Flow Behaviors at Ti-6Al-2Zr-1Mo-1V Alloy By GA-SVR and Relative Applications. *Mater. Res.* **2016**, *19*, 1253–1269. [CrossRef]
45. Zhao, J.; Ding, H.; Zhao, W.; Huang, M.; Wei, D.; Jiang, Z. Modelling of the hot deformation behaviour of a titanium alloy using constitutive equations and artificial neural network. *Comput. Mater. Sci.* **2014**, *92*, 47–56. [CrossRef]
46. Sun, Y.; Zeng, W.D.; Zhao, Y.Q.; Zhang, X.M.; Shu, Y.; Zhou, Y.G. Modeling constitutive relationship of Ti40 alloy using artificial neural network. *Mater. Des.* **2011**, *32*, 1537–1541. [CrossRef]
47. Mosleh, A.; Mikhaylovskaya, A.; Kotov, A.; Pourcelot, T.; Aksenov, S.; Kwame, J.; Portnoy, V. Modelling of the Superplastic Deformation of the Near-α Titanium Alloy (Ti-2.5 Al-1.8 Mn) Using Arrhenius-Type Constitutive Model and Artificial Neural Network. *Metals* **2017**, *7*, 568. [CrossRef]
48. Ge, G.; Wang, Z.; Zhang, L.; Lin, J. Hot deformation behavior and artificial neural network modeling of β-γ TiAl alloy containing high content of Nb. *Mater. Today Commun.* **2021**, *27*, 102405. [CrossRef]
49. Hu, C.; Martin, S.; Dingreville, R. Accelerating phase-field predictions via recurrent neural networks learning the microstructure evolution in latent space. *Comput. Methods Appl. Mech. Eng.* **2022**, *397*, 115128. [CrossRef]
50. Kautz, E.J. Predicting material microstructure evolution via data-driven machine learning. *Patterns* **2021**, *2*, 100285. [CrossRef]
51. Khandelwal, S.; Basu, S.; Patra, A. A Machine Learning-based surrogate modeling framework for predicting the history-dependent deformation of dual phase microstructures. *Mater. Today Commun.* **2021**, *29*, 102914. [CrossRef]
52. Mei, H.; Lang, L.; Yang, X.; Liu, Z.; Li, X. Study on Constitutive Relation of Nickel-Base Superalloy Inconel 718 Based on Long Short Term Memory Recurrent Neural Network. *Metals* **2020**, *10*, 1588. [CrossRef]
53. Benabou, L. Development of LSTM networks for predicting viscoplasticity with effects of deformation, strain rate and temperature history. *J. Appl. Mech.* **2021**, *88*, 071008. [CrossRef]
54. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 11106–11115. [CrossRef]
55. Zou, R.; Duan, Y.; Wang, Y.; Pang, J.; Liu, F.; Sheikh, S.R. A novel convolutional informer network for deterministic and probabilistic state-of-charge estimation of lithium-ion batteries. *J. Energy Storage* **2023**, *57*, 106298. [CrossRef]
56. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
57. Li, S.Y.; Jin, X.Y.; Xuan, Y.; Zhou, X.Y.; Chen, W.H.; Wang, Y.X.; Yan, X.F. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
58. Child, R.; Gray, S.; Radford, A.; Sutskever, I. Generating Long Sequences with Sparse Transformers. *arXiv* **2019**, arXiv:1904.10509.
59. Kitaev, N.; Kaiser, Ł.; Levskaya, A. Reformer: The Efficient Transformer. *arXiv* **2020**, arXiv:2001.04451.
60. Wang, S.N.; Li, B.Z.; Khabsa, M.; Fang, H.; Ma, H. Linformer: Self-Attention with Linear Complexity. *arXiv* **2020**, arXiv:2006.04768.
61. Martins, P.H.; Marinho, Z.; Martins, A.F.T. ∞-former: Infinite Memory Transformer-former: Infinite Memory Transformer. In Proceedings of the the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; pp. 5468–5485. [CrossRef]
62. Tsai, Y.-H.H.; Bai, S.; Yamada, M.; Morency, L.-P.; Salakhutdinov, R. Transformer Dissection: An Unified Understanding for Transformer's Attention via the Lens of Kernel. *arXiv* **2019**, arXiv:1908.11775. [CrossRef]
63. Calafiore, G.C.; Gaubert, S.; Possieri, C. Log-Sum-Exp Neural Networks and Posynomial Models for Convex and Log-Log-Convex Data. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 827–838. [CrossRef]

64. Calafiore, G.C.; Gaubert, S.; Possieri, C. A Universal Approximation Result for Difference of Log-Sum-Exp Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 5603–5612. [CrossRef] [PubMed]

65. Yu, F.; Koltun, V.; Funkhouser, T. Dilated residual networks. *arXiv* **2017**, arXiv:1705.09914.

66. Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (Elus). *arXiv* **2015**, arXiv:1511.07289.

67. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.

68. Smith, L.N. A Disciplined Approach to Neural Network Hyper-Parameters: Part 1–Learning Rate, Batch Size, Momentum, and Weight Decay. *arXiv* **2018**, arXiv:1803.09820.

69. Masters, D.; Carlo, L. Revisiting small batch training for deep neural networks. *arXiv* **2018**, arXiv:1804.07612.

70. He, F.X.; Liu, T.L.; Tao, D.C. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. *Adv. Neural. Inf. Process. Syst.* **2019**, *32*.

71. Keskar, N.S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; Tang, P.T.P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv* **2016**, arXiv:1609.04836.

72. Hoffer, E.; Hubara, I.; Soudry, D. Train longer, generalize better: Closing the generalization gap in large batch training of neural networks. *Adv. Neural. Inf. Process. Syst.* **2017**, *30*.