



# Deep learning-based grain-size decomposition model: A feasible solution for dealing with methodological uncertainty

YUMING LIU<sup>\*†</sup> , TING WANG<sup>\*‡</sup>, TAO WEN<sup>§</sup> , JIANGUANG ZHANG<sup>\*†</sup>, BO LIU<sup>¶</sup>, YUE LI<sup>\*</sup>, HANG ZHANG<sup>\*†</sup>, XIAOQING RONG<sup>\*†</sup>, LONG MA<sup>\*\*\*</sup>, FEI GUO<sup>\*††</sup>, XINGXING LIU<sup>\*</sup> and YOUBIN SUN<sup>\*</sup>

<sup>\*</sup>State Key Laboratory of Loess and Quaternary Geology, Institute of Earth Environment, Chinese Academy of Sciences, Xi'an 710061, China (E-mail: [liuyuming2024@gmail.com](mailto:liuyuming2024@gmail.com)) (E-mail: [sunyb@ieecas.cn](mailto:sunyb@ieecas.cn))

<sup>†</sup>University of Chinese Academy of Sciences, Beijing 100049, China

<sup>‡</sup>School of Civil Engineering, Hubei Engineering University, Xiaogan 432000, China

<sup>§</sup>Department of Earth and Environmental Sciences, Syracuse University, Syracuse, NY 13244, USA

<sup>¶</sup>Xi'an Institute for Innovative Earth Environment Research, Xi'an 710061, China

<sup>\*\*</sup>Department of Geology, State Key Laboratory of Continental Dynamics, Shaanxi Key Laboratory of Early Life and Environment, Northwest University, Xi'an 710069, China

<sup>††</sup>Institute of Marine Science and Technology, Shandong University Qingdao, Qingdao 266237, China

Associate Editor – Subhasish Dey

## ABSTRACT

Terrigenous clastic sediments cover a large area of the Earth's surface and provide valuable insights into the Earth's evolution and environmental change. Sediment grain-size decomposition has been widely used as an effective approach to inferring changes in sediment sources, transport processes and depositional environments. Several algorithms, such as single sample unmixing, end-member modelling analysis and the universal decomposition model, have been developed for grain-size decomposition. The performance of these algorithms is highly dependent on parameter selections, introducing subjective uncertainty. This uncertainty could undermine the reliability of decomposition results, limit the application of grain-size decomposition techniques and reduce comparability across different studies. To mitigate the methodological uncertainty, a novel deep learning-based framework for grain-size decomposition of terrigenous clastic sediments is proposed. First, an improved universal decomposition model is used to analyse the collected grain-size data, in order to provide training sets for the end-to-end decomposers. To meet the data size requirements of supervised learning, generative adversarial networks are also trained for data augmentation. The performance of the new framework is then evaluated using a small-scale dataset (73 393 samples from 18 sites) of three sedimentary types (loess, fluvial and lake delta deposits). The decomposed grain-size results demonstrate high feasibility and great potential of the framework in constructing a robust grain-size decomposition model. Finally, it is proposed that future grain-size research should aim to establish guidelines for grain-size data sharing and produce a big grain-size database for deep learning.

**Keywords** Convolutional neural network, data augmentation, deep learning, generative adversarial network, grain-size decomposition.

## INTRODUCTION

Terrigenous clastic sediments cover a large area of the Earth's surface and are essential for understanding Earth's evolution and exploiting petroleum and gas resources (Pettijohn, 1975; Folk, 1980; Ruddiman, 2001; Galloway & Hobday, 2012; Reineck & Singh, 2012). Grain size is a basic physical property of sediments that changes during weathering, erosion, transport, deposition and post-deposition processes (Pettijohn, 1975). Hence, grain size is widely used to analyse these processes and related control factors (Folk, 1966; Visher, 1969; McLaren & Bowles, 1985; Vandenberghe, 2013; Újvári *et al.*, 2016; D'Arcy *et al.*, 2017; Vandenberghe *et al.*, 2018). Due to the fact that grain size can easily be altered by hydrogeological processes, grain-size distribution (GSD) mainly reflects the depositional environment of the most recent sedimentation process (Folk & Ward, 1957; Visher, 1969; McLaren & Bowles, 1985). However, the relationship between grain size and sedimentary environment is complex due to the inherent complexity of terrigenous clastic sedimentary systems (Folk & Ward, 1957; Visher, 1969; Reineck & Singh, 2012; Liang & Yang, 2023). Interpreting the sedimentological significance of GSD is still challenging (Liu *et al.*, 2023). Decomposing GSDs has become a popular method for extracting vital environmental and climatic information from grain-size data (Weltje, 1997; Sun *et al.*, 2002; Dietze *et al.*, 2012; Paterson & Heslop, 2015; Liu *et al.*, 2023). The fundamental assumption of grain-size decomposition is that measured multimodal GSDs result from the superposition of multiple unimodal elementary distributions (Folk & Ward, 1957; Visher, 1969; Sun *et al.*, 2002; Weltje & Prins, 2003). That is to say that the mixing of sediments from multiple sources or transported by various dynamic processes which ultimately deposit simultaneously at a location constitutes a forward process. Conversely, grain-size decomposition employs mathematical and statistical methods to reconstruct pre-mixed distributions and the mixing proportions representing a reverse process. As expected, the initial grain-size information is partially lost following mixing, rendering grain-size decomposition a highly challenging task in its nature. However, compared with traditional grain-size parameters (for example, mean, sorting coefficient, skewness and kurtosis), grain-size decomposition provides a more concise perspective for understanding the variations of GSDs (Weltje & Prins, 2007; Liu *et al.*, 2023). Moreover, it can

separate the effects of different sedimentary processes on GSDs and provide new grain-size indicators (Liu *et al.*, 2023). These advantages make grain-size decomposition a worthwhile problem and a long-standing topic for grain-size research.

Several algorithms for grain-size decomposition have been developed, including single sample unmixing (SSU) (Sun *et al.*, 2002; Qin *et al.*, 2005; Wu *et al.*, 2020; Liu *et al.*, 2021; Peng *et al.*, 2022), end-member modelling analysis (EMMA) (Weltje, 1997; Heslop *et al.*, 2007; Dietze *et al.*, 2012; Paterson & Heslop, 2015; Seidel & Hlawitschka, 2015; Yu *et al.*, 2016; Zhang *et al.*, 2018, 2020; Dietze & Dietze, 2019; Liu *et al.*, 2021) and the universal decomposition model (UDM) (Liu *et al.*, 2023). SSU focuses on individual samples and uses a curve-fitting approach to determine the mixing proportions and parameters of elementary distributions (Sun *et al.*, 2002; Qin *et al.*, 2005). However, Weltje & Prins (2003, 2007) noted that grain-size decomposition is a poorly constrained problem that leads to severe instability and ambiguity in SSU. Therefore, those authors proposed EMMA (end-member modelling analysis), which processes all samples from a specific region simultaneously and leverages geological settings to produce more stable and reasonable decomposition results (Weltje, 1997; Weltje & Prins, 2003, 2007). Subsequently, EMMA has gained significant popularity and has given rise to many variant algorithms (Liu *et al.*, 2023). However, in recent years there has been an increase in discussions regarding the constraints and limitations of the EMMA algorithm (van Hateren *et al.*, 2018; Varga *et al.*, 2019; Liu *et al.*, 2021, 2023; Dietze *et al.*, 2022). For example, van Hateren *et al.* (2018) noted substantial performance differences among various EMMA algorithms. Liu *et al.* (2021) and Dietze *et al.* (2022) reassessed EMMA's applicability and found that it performs poorly when dealing with end members with varying mode sizes. Moreover, Liu *et al.* (2023) identified inherent flaws in EMMA, arguing that end members are designed to ignore potential component changes in reality, resulting in a suboptimal design. Those authors also concluded that addressing this poorly constrained problem requires using as much *a priori* knowledge as possible from sedimentological understanding. Based on this idea, Liu *et al.* (2023) proposed UDM, which employs a more flexible mathematical model like SSU and uses as much *a priori* knowledge as possible to make the decomposition results more reasonable

with respect to sedimentology (Liu *et al.*, 2023). UDM theoretically found the inherent identity of SSU and EMMA methods, and thereby unified two schools of grain-size decomposition, largely reducing the systematic uncertainty caused by their principles.

However, this decomposition method developed along the lines of traditional methods still has some subjective uncertainties due to individual user proficiency. Traditional methods require users to judge their decomposition results and select appropriate algorithm parameters. Previous studies have also tried to use some statistical methods and automated programs to select parameters, but most of these cannot fully ensure their effectiveness (Weltje & Prins, 2007; Dietze *et al.*, 2012; Paterson & Heslop, 2015; van Hateren *et al.*, 2018; Zhang *et al.*, 2020; Liu *et al.*, 2023). Some studies even used seriously flawed algorithm settings because they did not recognize the unreliability of these parameter selection methods. For example, the method currently widely used to determine the optimal number of components (end members) was found to be unreliable (van Hateren *et al.*, 2018; Liu *et al.*, 2023), but most of the current studies still use it to select the number of components. The inappropriate number of components will not only affect the accuracy of the decomposition results, but also lead to serious bias when interpreting the geological significance of the components and related depositional processes. These subjective uncertainties may greatly affect the credibility of grain-size decomposition results and have already caused a considerable negative impact on the promotion of grain-size decomposition methods. They also seriously damage the comparability between different studies. Therefore, solving this problem is very urgent.

Deep learning has seen tremendous progress in the past decade and has produced notable results in various disciplines (e.g. Silver *et al.*, 2017; Jumper *et al.*, 2021). Deep learning techniques are well-suited for automatically learning underlying patterns in training data and provide a black-box model that eliminates the need for users to understand technical details (LeCun *et al.*, 2015). This approach allows us to construct a unified model based on deep learning technology, thereby completely removing the step of subjective algorithm parameter selection, and thus greatly reducing uncertainty caused by subjective selection differences. A novel deep learning-based framework for grain-size decomposition is proposed that utilizes various deep learning

technologies such as multilayer perceptron (MLP), convolutional neural network (CNN) and generative adversarial network (GAN). The framework integrates traditional methods (for example, UDM) to ensure the interpretability and reliability of decomposition results. It uses deep learning techniques to transfer inherent knowledge from traditional methods, training data and expert decision-making processes into a comprehensive model. This eliminates the need for users to understand the technical details of the model and makes the model highly user-friendly. The proposed framework provides a feasible solution for dealing with methodological uncertainty in grain-size decomposition. To verify its feasibility, 73 393 grain-size samples were collected from 18 sites across three types of sediment and this decomposition framework was applied.

## SAMPLING SITES AND GRAIN-SIZE MEASUREMENT

Grain-size samples of three sediment types were gathered from 18 sites, primarily located in China and Central Asia (Table 1; Fig. 1). The sampling sites included nine profiles and one borehole from Chinese loess, four profiles from arid Central Asian (ACA) loess, two boreholes from fluvial deposits in Eastern Henan Basin, China, and two boreholes from lake delta deposits in Weihe Basin, China. The majority of these sites are situated within the East Asian monsoon region, predominantly influenced by the East Asian winter and summer monsoons (An *et al.*, 2014). The climate at the ACA loess sites is primarily governed by the westerlies (Li *et al.*, 2020). Previous studies have reported on most of these sites (Ma *et al.*, 2017; Guo *et al.*, 2021; Li *et al.*, 2021, 2022, 2023; Liu *et al.*, 2021, 2023; Zhang *et al.*, 2024), with the YB19 profile and WB1 borehole being introduced for the first time in this study.

The YB19 profile (35.65°N, 103.15°E, 2145 metres above sea level – m a.s.l.), situated 3 km east of the LX borehole (Guo *et al.*, 2021), was excavated to a depth of 44 m for grain-size sample collection. A total of 879 powder samples were collected at intervals of 5 cm. The WB1 borehole (34.65°N, 109.58°E, 343 m a.s.l.) is part of the Weihe Basin Drilling Project (An *et al.*, 2020) and is located in the same basin as the HX borehole (Liu *et al.*, 2023). A 1386.5 m long core was drilled, and 27 392 grain-size samples were obtained. Positioned at the Gushi

**Table 1.** The profiles and boreholes used in this study.

Sediment type	Profile/ Borehole	Longitude	Latitude	Sampling depth (m)	Number of samples	Stratigraphic age	References
Loess	GJP	107.39	37.11	44.6	1366	<i>ca</i> 150 ka	Ma <i>et al.</i> (2017)
	BGY	107.29	36.66	27.3	2230	<i>ca</i> 150 ka	Ma <i>et al.</i> (2017)
	YC	109.94	36.62	20.0	399	<i>ca</i> 150 ka	Ma <i>et al.</i> (2017)
	LC	109.43	35.73	10.6	106	<i>ca</i> 150 ka	Ma <i>et al.</i> (2017)
	TC	108.96	34.97	12.9	258	<i>ca</i> 150 ka	Ma <i>et al.</i> (2017)
	BL	109.19	34.20	9.5	471	<i>ca</i> 150 ka	Ma <i>et al.</i> (2017)
	WN19	109.58	34.43	16.0	321	<i>ca</i> 150 ka	Liu <i>et al.</i> (2021)
	YB19	103.15	35.65	44.0	879	<i>ca</i> 150 ka	This study
	LX	103.12	35.63	203.8	4947	<i>ca</i> 650 ka	Guo <i>et al.</i> (2021)
	Osh	73.01	40.61	7.0	350	<i>ca</i> 30 ka	Li <i>et al.</i> (2021)
	FS18	115.40	40.20	15.0	750	<i>ca</i> 150 ka	Li <i>et al.</i> (2022)
	BSK	74.78	42.70	18.0	900	<i>ca</i> 150 ka	Li <i>et al.</i> (2022)
	CMG	69.83	38.39	25.0	1250	<i>ca</i> 150 ka	Li <i>et al.</i> (2022)
NLK	83.25	43.76	20.5	1026	<i>ca</i> 70 ka	Li <i>et al.</i> (2023)	
Fluvial deposits	HKG	113.89	34.64	303.5	2878	<i>ca</i> 3.2 Ma	Zhang <i>et al.</i> (2024)
	BS	113.89	34.72	320.2	3354	<i>ca</i> 3.5 Ma	Zhang <i>et al.</i> (2024)
Lake delta sediments	HX	108.71	34.13	512.2	24 516	\	Liu <i>et al.</i> (2023)
	WB1	109.58	34.65	1386.5	27 392	\	This study

depression, WB1 was characterized by finer grain size and deeper water level. Its sedimentary environments were dominated by shallow lakes and delta fronts with developed subaqueous distributary channels.

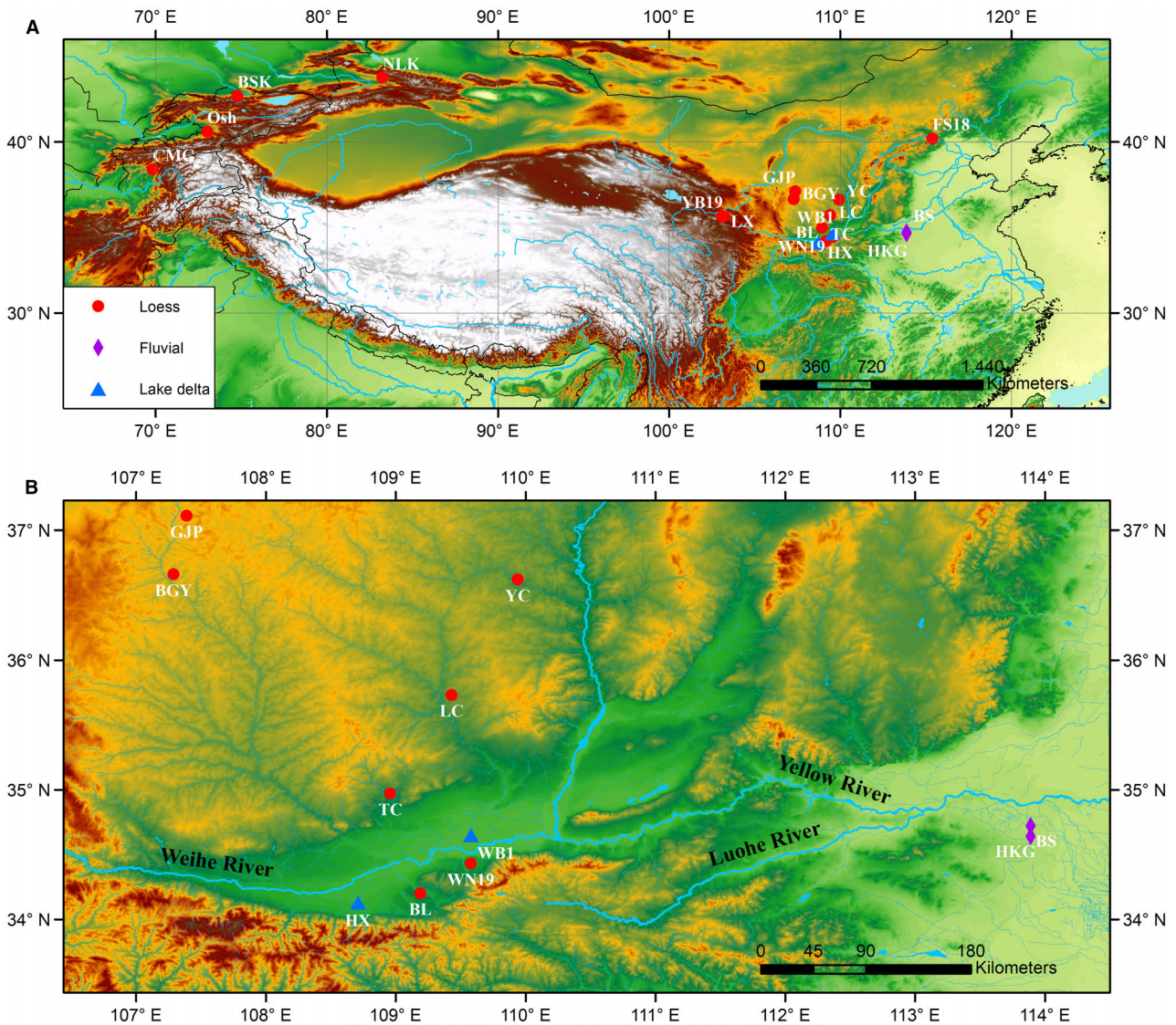
The loess sites are the most numerous and widely distributed. Although the numbers of samples of these loess sites are relatively small, loess has a simple depositional environment and grain-size characteristics. Therefore, it is easy to use loess data for training and verify the model feasibility at a basic level. The other two types of sediment have fewer sites and more complex depositional environments. Especially, the sediments in the Weihe Basin are affected by wind, river and lake actions (Liu *et al.*, 2023). This increases the training difficulty and challenges the model training with a higher difficulty level.

A total of 73 393 sediment powder samples of three sediment types from 18 locations were collected for the grain-size measurements. After the removal of organic matter and carbonate (Lu & An, 1997), the GSDs were measured using a Malvern Mastersizer 2000 laser diffraction instrument (Malvern Panalytical, Malvern, UK) at the Institute of Earth Environment, Chinese Academy of Sciences. The instrument delivered 100 classes within a size range of 0.02 to 2000  $\mu\text{m}$ .

## DECOMPOSITION FRAMEWORK

### Overview

The breakthrough of AlexNet (Krizhevsky *et al.*, 2012) in computer vision significantly accelerated the development of deep learning. As a result, supervised learning has matured and found extensive applications in various fields (LeCun *et al.*, 2015; Silver *et al.*, 2017; Alom *et al.*, 2018; Jumper *et al.*, 2021; Epuna *et al.*, 2022; He *et al.*, 2022; Wen *et al.*, 2022; Esmaeiloghli *et al.*, 2023; Wang *et al.*, 2023). Therefore, it is logical to explore whether grain-size decomposition can be reformulated as a supervised learning problem. One challenge to formulate grain-size decomposition as a supervised learning problem is the lack of training samples, i.e. datasets with known decomposed components and proportions. In contrast to image classification problems, manually completing data labelling for grain-size decomposition, such as directly providing decomposed components and proportions by observing the GSDs, is not feasible. In other words, this is the motivation of previous scholars to develop traditional decomposition algorithms. The UDM algorithm, which integrates SSU and EMMA, and exhibits reliable performance (Liu *et al.*, 2023), can address this issue by providing



**Fig. 1.** (A) The locations of the profiles and boreholes used in this study; (B) focuses on the central region where the sites are densely distributed. For more information on the sampling sites, please refer to Table 1 and [Sampling sites and grain-size measurement](#) section.

decomposed data for further training. Moreover, the UDM algorithm has been further improved by incorporating the spatial distances of samples in this study (please refer to the [Supplementary Information](#)). By assigning a uniform set of initial parameters to each dataset, the UDM algorithm can yield reliable decomposition results with the requirement of *a priori* knowledge of the dataset and site.

Supervised learning requires a large volume of training samples to ensure adequate model generalization performance. However, in many research domains, data acquisition is costly and

challenging. Therefore, data augmentation is indispensable. A prevalent data augmentation technique involves using GANs (Goodfellow *et al.*, 2014; Bergen *et al.*, 2019; Al-Najjar & Pradhan, 2021, Fig. 2). GAN, an unsupervised learning approach, relies solely on sample authenticity as labels. In principle, GANs can be directly trained using measured GSDs. However, due to potential multiple solutions in grain-size decomposition, samples generated by GANs trained in this manner may not align with sedimentological knowledge. Hence, the decomposition results of UDM are used as a reference to

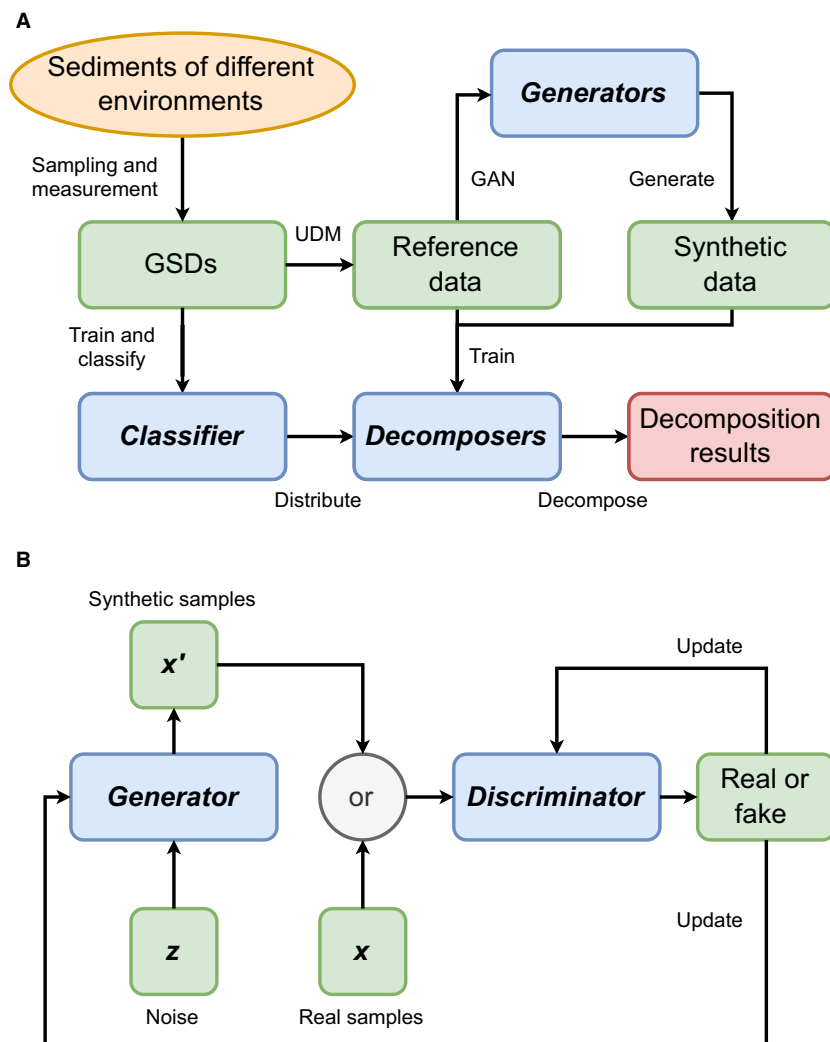


Fig. 2. (A) Deep learning-based decomposition framework proposed in this study and (B) the basic architecture of generative adversarial networks (GANs) (Goodfellow *et al.*, 2014) used in this study. For more details, please refer to the main text and the [Supplementary Information](#).

impose additional constraints on GAN training, and to enable GANs to directly learn to generate decomposed data.

Generative adversarial networks can be used to generate supplementary data, enabling the training of reliable deep learning-based decomposers which accept measured GSDs and provide decomposed components and proportions. To accommodate the variations in the number and specific conditions of components across different sediment types, data from each sediment type is processed independently. A consistent set of UDM algorithm settings is used for the grain-size data of each sediment type, and a corresponding set of GANs is trained using the UDM results. Subsequently, the decomposers are trained for each sediment type using data derived from both UDM and GANs. A classifier, essentially a sedimentary facies discriminator

based on grain-size data, is also trained to automatically assign measured GSDs to the appropriate decomposers. Finally, a detailed evaluation of the trained generators, decomposers and classifier was conducted to confirm the feasibility of this decomposition framework.

### Preparation of datasets

The measured GSDs were classified into three categories based on sediment types, and the UDM algorithm was applied with varying parameters for initial decomposition. The primary differentiation in these parameters lies in the number and initial parameters of the components. For loess, fluvial and lake delta sediments, this study employed 3, 4 and 5 components respectively, following the Weibull distribution for decomposition. The initial

parameters for each component were ascertained by applying the SSU algorithm to decompose selected representative samples. Please refer to the [Supplementary Information](#) for more details on the algorithm settings and decomposition results.

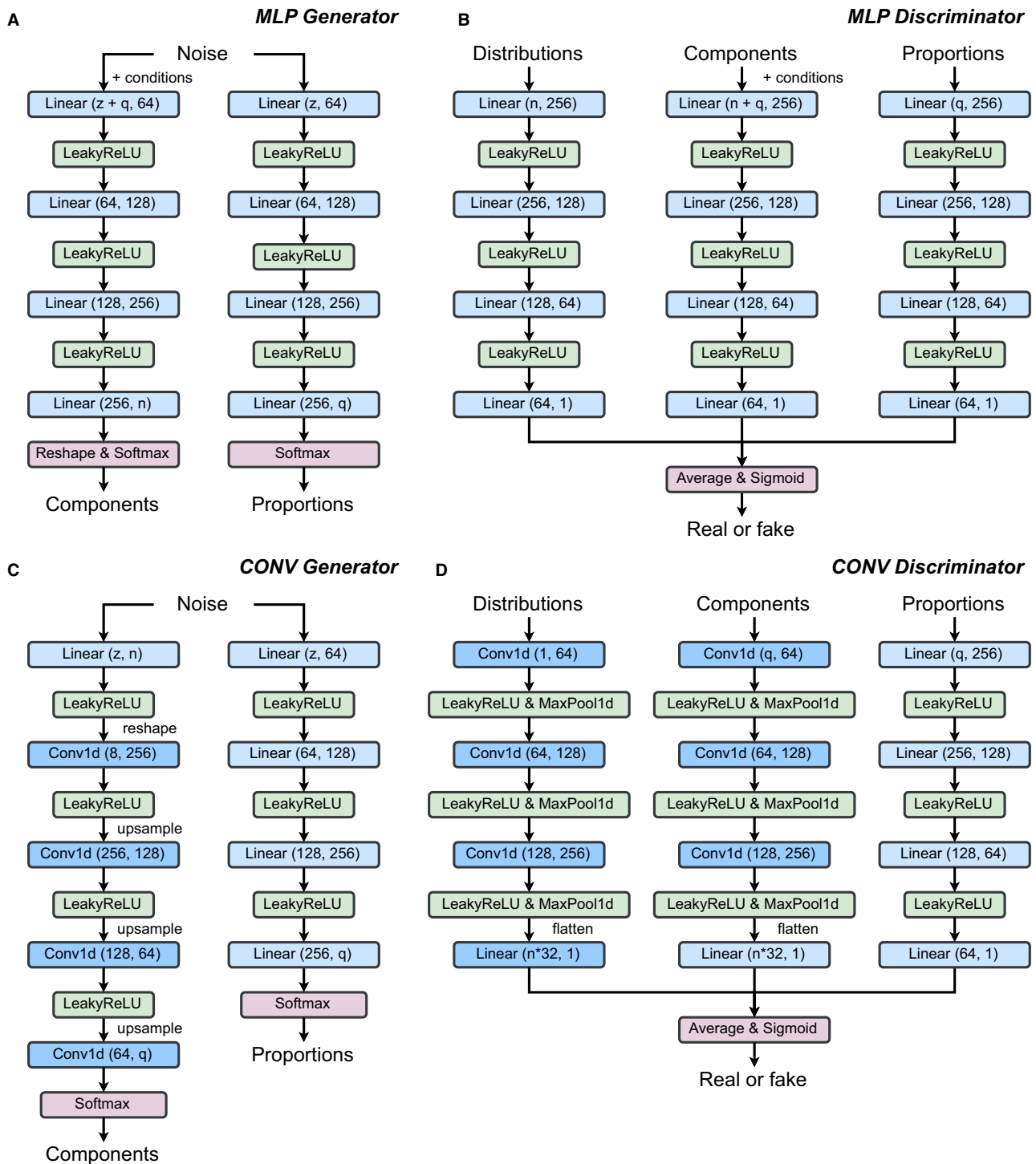
Given the abundance of available loess profiles, only eight loess datasets from the Chinese Loess Plateau (CLP) were used as the training sets for both generators and decomposers, while datasets from other loess sites were used as test sets to evaluate the generalization performances of decomposers. In the test sets, the LX borehole has a total length of approximately 203.8 m and an age of about 650 ka. It can be used to test whether the model trained using loess from the last interglacial can be used to decompose older loess samples. FS18 and four loess samples from ACA, whose grain-size characteristics differ from those of CLP's loess (Li *et al.*, 2021, 2022, 2023), can be used to test the generalization ability further. For other sediment types, all samples were included in the training of generators and decomposers. Considering that classifier training is relatively straightforward, only measured GSDs were used for training, without using data generated by GANs.

### Generative adversarial networks for data augmentation

Generative adversarial networks (GAN), inspired by game theory, simultaneously train a generator and a discriminator (Goodfellow *et al.*, 2014). The generator's role is to create realistic samples, while the discriminator evaluates whether these samples are genuine or fabricated. The training process involves a competition between the generator and the discriminator until the generator can produce data so realistic that the discriminator cannot differentiate it from real samples. Since their inception, GANs have garnered significant attention, propelling advancements in generative model research and applications (Goodfellow *et al.*, 2020). One crucial application of GANs is data augmentation, where the generator learns the distribution of real samples and generates new ones via random sampling (Al-Najjar & Pradhan, 2021). Training original GANs presents challenges, leading to the proposal of several improved models, such as the Wasserstein GAN (WGAN) (Arjovsky *et al.*, 2017). WGAN amends the loss function to tackle instability issues during training, but its

gradient clipping strategy is rudimentary (Arjovsky *et al.*, 2017; Gulrajani *et al.*, 2017; Miyato *et al.*, 2018). Subsequent research has proposed enhancements like gradient penalty (Gulrajani *et al.*, 2017) and spectral normalization (SNGAN) (Miyato *et al.*, 2018) to mitigate this limitation. Furthermore, GANs provide flexibility in designing network structures for both the generator and discriminator, facilitating the integration of various network modules from other studies, such as convolutional layers (Lecun *et al.*, 1998; Dhillon & Verma, 2020).

For grain-size decomposition, the generator accepts multidimensional noise following a Gaussian distribution as input, and outputs components and proportions. The generated mixed GSDs can be obtained through further calculation. The generated components, proportions and GSDs are then input into the discriminator for judgement. At the same time, an equal number of UDM decomposed components, proportions and measured GSDs are also input into the discriminator for judgement. This study explores the applicability of convolutional layers for data augmentation in GSDs by proposing two network architectures for the generator and discriminator. One architecture is based on the MLP (Fig. 3A and B), while the other employs the CNN (Fig. 3C and D; Lecun *et al.*, 1998). Different branch structures are designed for different types of data to perform generation and discrimination. Regardless of the branch type, the overall structure is similar, composed of four or five layers of linear or convolutional layers. Because there are no residual connections, blindly increasing the number of network layers does not necessarily improve the performance of the model (He *et al.*, 2016). The performance of the network structure using these numbers of layers is relatively good according to actual testing results. To address training instability in GANs, it is explored whether WGAN and SNGAN can enhance performance; thus, each network type is trained using GAN, WGAN and SNGAN methods separately. Six sets of generators and discriminators are trained for each sediment type using identical hyperparameters such as the learning rate across all training sessions in this study. A comparative analysis is conducted to ascertain the optimal network structure and training method. Please refer to the [Supplementary Information](#) for more details on the model design and validation.



**Fig. 3.** The network structures of the generators and discriminators based on multilayer perceptron (MLP) and convolutional neural network (CNN). Please refer to the [Supplementary Information](#) and the published codes for more details.



## Decomposers and classifier

In the presented framework, a decomposer refers to the model trained to decompose the measured GSDs and generate multiple elementary distributions (i.e. components) along with their respective proportions. By observing its input and output, it is very similar to the generator in GANs. Therefore, a basic decomposer can be built by making minor modifications to the input layer of the generator to match the dimensions of GSDs. Considering that convolutional layers significantly improve the performance of GAN generators, only the structures adopting convolutional layers are used as a basis for constructing the decomposers here. This method is consistent with the GAN design where each sediment type requires a unique decomposer. Since it is necessary to train a specific model for each sediment type, distributing data among these models is necessary. Therefore, a suitable classifier is needed for this task. The classifier's input consists of the GSDs to be decomposed, and its output provides the label corresponding to the sediment type. Similarly, this classifier is akin to the sub-module used in GANs for evaluating GSDs. Therefore, with minor modifications to its output layer, a basic classifier can be developed.

## Model validation

Assessing the quality of the generators in GANs is challenging, especially for the grain-size decomposition problem. Existing evaluation metrics designed for image generation tasks are not directly applicable. Thus, it is crucial to develop quantitative metrics for evaluating the generator's performance. This study introduces two novel metrics to assess the quality of GAN generators for the generative task of GSDs. The first metric, termed 'component precision', is designed to ascertain how effectively the generator has learned to generate a grain-size component that conforms to a specific distribution (for example, the Weibull distribution in this study). This involves understanding of the hidden relationships between the frequencies of different grain-size classes, i.e. the knowledge about the structure of grain-size components. The knowledge is not hard-coded into the generator's model, it is straightforward to verify whether the model has learned this knowledge by fitting the generator's output to a specific distribution (for example, the Weibull distribution in this study). The fitting residuals can then

quantitatively assess whether the generator has learned this knowledge. In this study, the negative logarithmic mean square error (LMSE) is used as a measure of component precision. The second metric evaluates whether the generator has effectively learned the data distributions of grain-size components from the UDM results of real samples. For this purpose, this metric is based on the Wasserstein distance, which is widely used to measure the difference between two distributions (Vaserstein, 1969; Arjovsky *et al.*, 2017; Gulrajani *et al.*, 2017). Since Wasserstein distance cannot be directly applied to vector-type data, a slight modification is needed. This study separately computes Wasserstein distances for both the mean grain size and the sorting coefficient of bulk samples and grain-size components. Then, the average of these different Wasserstein distances is used to represent the overall Wasserstein distance. In addition, several visualization charts shown in the [Supplementary Information](#) can further examine the details of the generated data.

To evaluate the performance of trained decomposers, this study, following the convention in grain-size decomposition research, uses the residual of the decomposition results as a fundamental metric (Weltje, 1997; Sun *et al.*, 2002; Qin *et al.*, 2005; Dietze *et al.*, 2012; Paterson & Heslop, 2015). While lower residuals may not necessarily indicate superior performance for an individual sample, they can reflect the overall accuracy of the model's decomposition results for a dataset as a whole to some extent. To examine how dataset size influences decomposer performance and validate the necessity of data augmentation, this study designed a set of control experiments. These experiments consist of 16 sub-experiments with identical model structures and training hyperparameters for the decomposers but varying sample quantities in their training sets. The first four sub-experiments use training data derived from UDM's decomposition results with increasing sample quantities: 512, 1024, 2048 and 4096 respectively. From Experiment 5 to 10, generated samples from one of the GAN generators (CONV SNGAN) are introduced into these experiments with quantities being 12 288, 24 576, 49 152, 98 304, 196 608 and, finally, 393 216, respectively. From Experiments 11 to 16, compared with Experiments 5 to 10, the generated samples are produced by all six different GAN generators, while the number of samples remains consistent. This design is intended to

determine whether the incorporation effect of different generators is beneficial to the training of the decomposer by enhancing the diversity of generated samples. In addition, a detailed comparison has been taken on the decomposition results of UDM and the decomposers for all sites.

Furthermore, it is essential to evaluate the performance of the classifier using established evaluation methods. Despite the limited importance of the classifier in this study, approximately 20% of samples are allocated as the test set and basic evaluation techniques that involve computing precision, recall, F1 score for various sediment types and overall accuracy are employed.

## RESULTS

### Decomposition results of the universal decomposition model

Figure 4 presents the decomposition results of UDM for several selected sites (please refer to Figs S1 to S3 for additional sites). The decomposed components of the loess datasets generally exhibit similar shapes, with the BSK profile being an exception (Fig. S2M). The decomposition results for the BSK profile reveal a distinctive feature: a high kurtosis and a notably low proportion for its component  $C_3$  (Fig. S2N and O). Upon conducting an exhaustive comparison on the data distribution of loess sites in the CLP and ACA regions, a notable difference is observed (Fig. 8). This disparity suggests that the formation mechanisms or control factors across these two regions may be different.

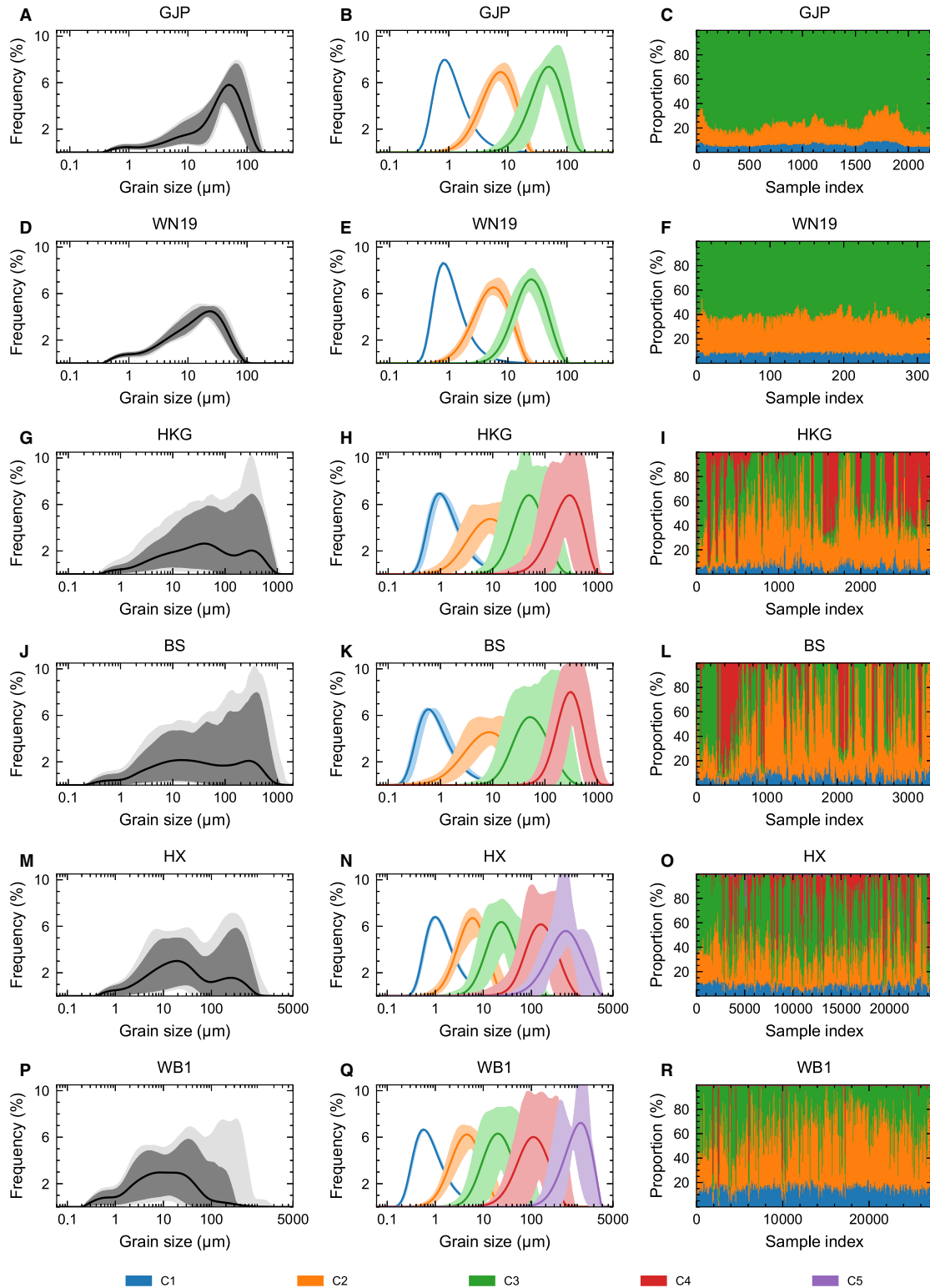
Moreover, there is spatial diversity that aligns with the aeolian deposition regime of the loess on the CLP (Liu, 1985; An *et al.*, 2014). Specifically, for sites proximal to the dust origins (for example, GJP), the component proportions display pronounced glacial–interglacial fluctuations; however, for distal sites (for example, WN19), these fluctuations are less evident. This observation suggests that the loess formation model for south-eastern loess sites deviates slightly from the classic model (Liu, 1985; Pye, 1995; An *et al.*, 2014). While the classic loess formation model is dominated by the East Asian winter monsoon, sites closer to the south-eastern CLP appear to be significantly influenced by the East Asian summer monsoon (An *et al.*, 2014; Hou *et al.*, 2021).

In the comparative analysis of the decomposition results from two fluvial boreholes, a significant consistency is observed in the shapes of their components and the patterns of component proportion fluctuations (Fig. 4). A key observation is the frequent alternation in the proportions of coarse components  $C_3$  (ca 50  $\mu\text{m}$ ) and  $C_4$  (ca 250  $\mu\text{m}$ ) (Fig. 4I and L). The sedimentological implications of component  $C_4$  are relatively straightforward. It can be readily identified in measured GSDs (Fig. 4G and J) and its grain-size attributes align with numerous channel deposits (Cheetham *et al.*, 2008; Hajek *et al.*, 2010; Zhang *et al.*, 2015). Elevated proportions of  $C_4$  in the stratum suggest channel developments. The close association between  $C_3$  and  $C_4$  implies that  $C_3$  represents the comparatively finer component deposited when floodplain water dynamics weaken following the transportation of coarse particles across the channel by runoff. Its grain-size characteristics align with floodplain deposits reported in other studies (Fan *et al.*, 2006; Pan *et al.*, 2015; Vandenberghe *et al.*, 2018). Regarding the remaining silt-sized components  $C_1$  (ca 1.4  $\mu\text{m}$ ) and  $C_2$  (ca 8.8  $\mu\text{m}$ ), they are transported by suspension via runoff, and their proportions are relatively higher in depositional environments with less riverine influence (for example, oxbow lakes and river marshes).

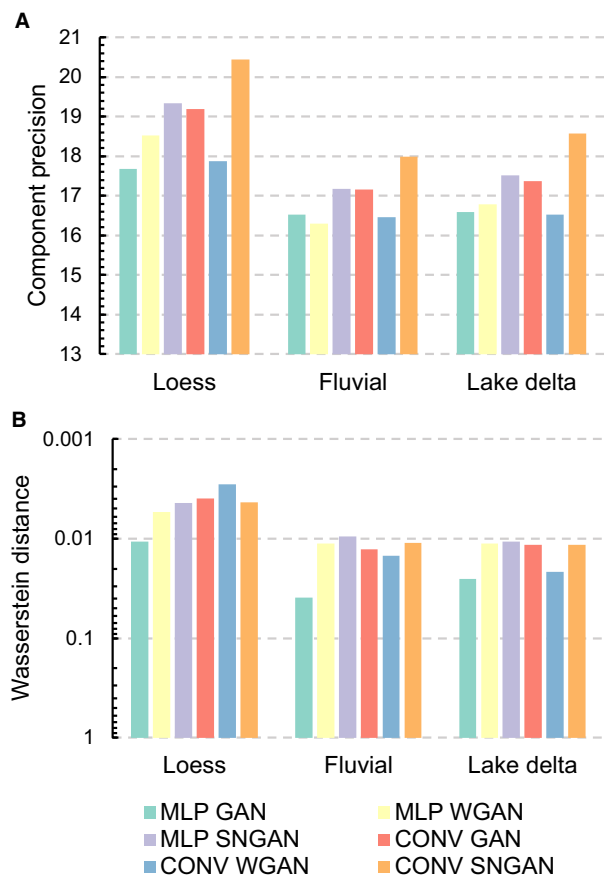
Liu *et al.* (2023) have already provided a detailed description and discussion of the decomposition results of the HX borehole. Compared to the HX borehole, the most noticeable difference in the decomposition results of the WB1 borehole is its significantly lower proportions of coarse components ( $C_4$  and  $C_5$ ) (Fig. 4O and R). This evidence suggests that the WB1 borehole had a higher lake water level than the HX borehole, with its sedimentary facies predominantly characterized by shallow lake and delta front sub-facies. In general, the decomposition results of UDM indicate that it can effectively handle grain-size data from various sediment types, whether it comes from wind-blown loess deposits with relatively simple grain-size characteristics or from complex lake delta deposits resulting from the combined depositional dynamics of wind, rivers and lakes.

### Comparison of different generators

The two proposed metrics, component precision and Wasserstein distance, demonstrated excellent convergence in actual results (Fig. S11), indicating their effectiveness in evaluating



**Fig. 4.** The measured grain-size distributions (GSDs) (left), components (middle) and proportions (right) decomposed by universal decomposition model (UDM) on several selected sites. The components from fine to coarse are C1 to C5. From top to bottom, the first and second rows show the decomposition results of two typical sites from the proximal and distal regions of the Chinese Loess Plateau (CLP), respectively. The third and fourth rows show the decomposition results of two fluvial boreholes, and the fifth and sixth rows show the decomposition results of two lake delta boreholes. For more results from other sites, please refer to the [Supplementary Information](#).



**Fig. 5.** The performance comparison of six generators trained with data from different sediment types. For the definitions of two evaluation metrics, component precision and Wasserstein distance, please refer to the main text.

generator performance. This study compared the performance of six types of GAN generators on different sediment type training datasets based on these two metrics (Fig. 5, refer to the [Supplementary Information](#) for more details). The results showed that these generators perform optimally when processing loess data due to its simpler characteristics and fewer components. Moreover, the performance of the original GAN is generally inferior, and model collapse can easily occur during its training process (Fig. S11), leading to a higher demand for reasonable hyperparameter settings. While WGAN can theoretically enhance the stability, its aggressive gradient clipping strategy makes its performance improvement unstable and occasionally leads to a sudden decrease in performance (Fig. S11). For SNGAN, it provides a very stable improvement in model performance and is not sensitive to hyperparameter settings.

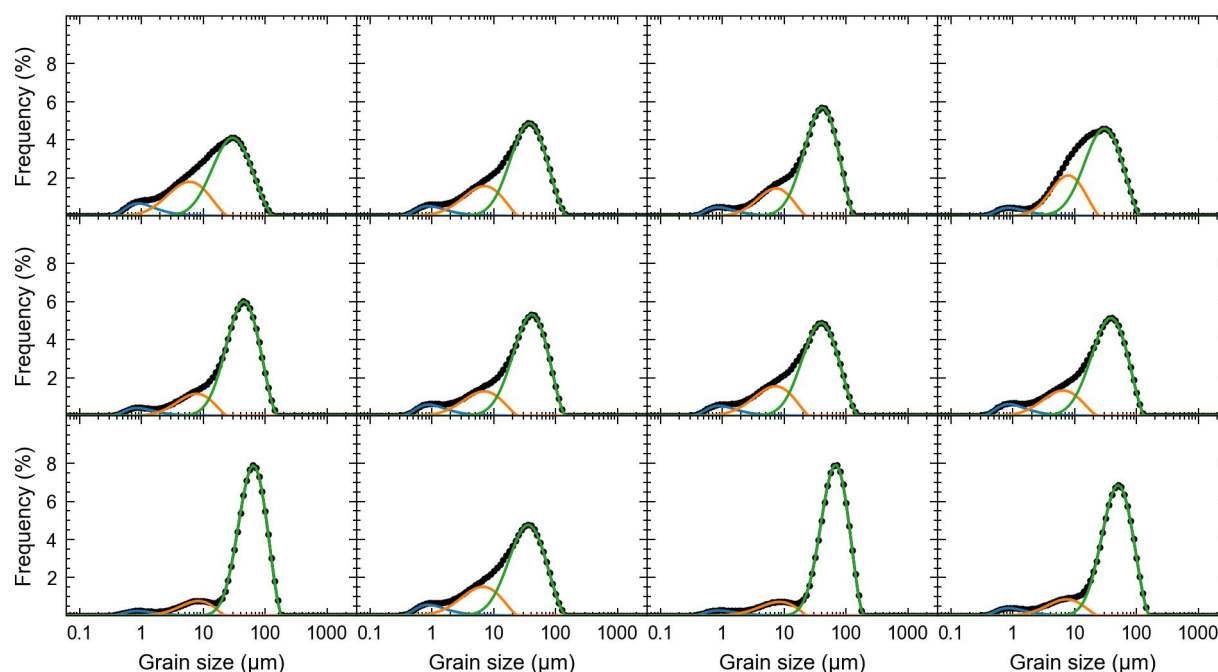
Importantly, replacing linear layers with convolutional layers significantly improves generator performance (Fig. 5). As a result, the CONV SNGAN demonstrates superior and more consistent performance (Video 1).

In terms of the diversity of data generated by the generator, except for the overall poor performance of MLP GAN, the overall level of other GANs is similar (Fig. 5B), but there are differences in the specific details of data distributions (Figs S12 to S15). These differences are relatively random among different generators and do not have an obvious pattern. This situation may be due to randomness caused by the training method of deep learning models (for example, model initialization and stochastic gradient descent). On the other hand, it may also be due to small model parameter scale or suboptimal hyperparameter settings, resulting in insufficient ability of the model to learn and memorize data distribution, and can only focus on generating a portion of the data. Considering these factors, all generators were used to generate data in the last sub-experiment when training decomposers, to ensure that the data diversity is optimal (Fig. 8).

Furthermore, the current datasets pertaining to fluvial and lake delta are limited, resulting in their data distribution appearing as isolated clusters, indicating a lack of diversity (Figs S14 and S15). Interestingly, it is observed that the extent of data distribution for the generated samples tends to be broader (Fig. S14). This suggests that certain inaccuracies in the generator's learning of real data distribution could potentially enhance the diversity of the generated dataset.

### Influence of data augmentation

A collection of decomposers was trained utilizing loess data, following the experimental design outlined in the preceding section. Subsequent performance evaluation revealed that the LMSE of the decomposer, with training sample sizes spanning from 1024 to 102 400, exhibits a gradual decrease as the quantity of training samples increases. This trend is more evident in the test sets, suggesting that an increase in the number of training samples can indeed enhance the performance of the decomposer, particularly its generalization performance. In scenarios characterized by limited training data, despite the decomposer demonstrating satisfactory performance on the training set, its LMSE on the test sets remained significantly high, indicating a deficiency in generalization performance. Moreover, when the quantity of training



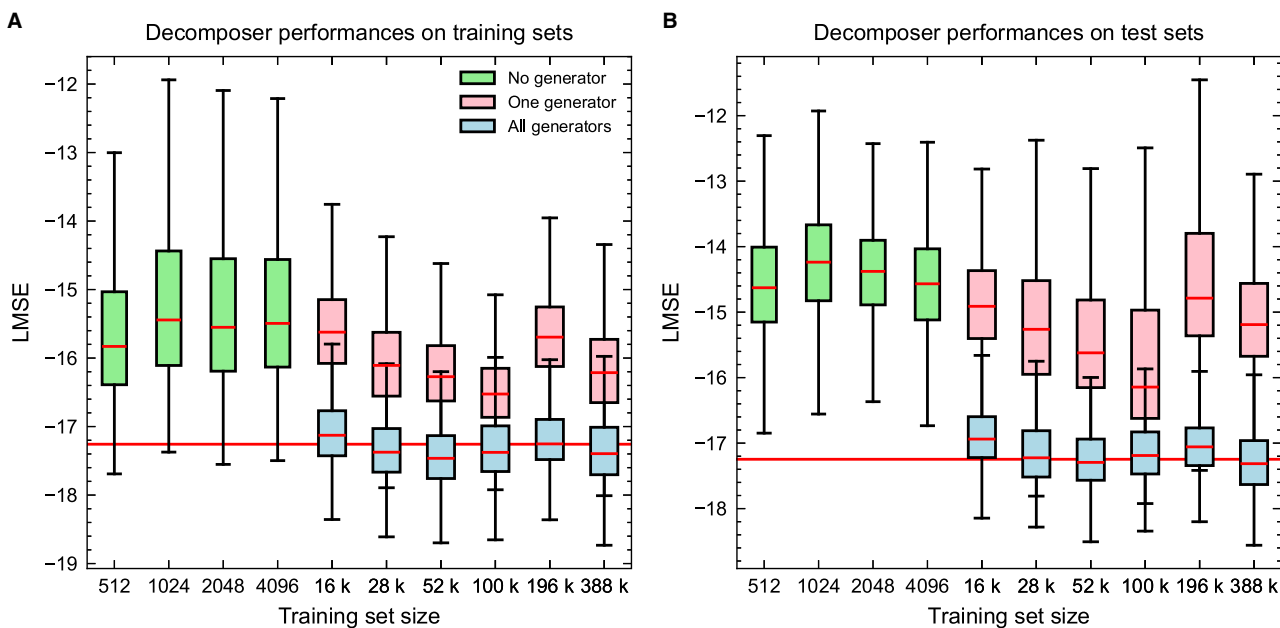
**Video 1:** The artificial loess samples generated by CONV SNGAN. This video lasts for 60 s, contains 300 frames, each frame displays the specific situation of 12 randomly generated samples, showing a total of 3600 generated samples.

samples escalated from 102 400 to 200 704, the LMSE of the decomposer unexpectedly increased, suggesting that the application of GANs for data augmentation has certain limitations. Notably, in comparison to employing a single generator, the LMSE of the decomposer trained using multiple generators is considerably lower. This could imply that the diversity of training data is of greater importance than the quantity. Furthermore, in comparison to the traditional grain-size decomposition algorithm (UDM, represented by red horizontal lines in Fig. 6), achieving a performance level comparable to it is challenging when using a single generator. However, when multiple generators are employed, only approximately 50 000 training samples are required to achieve a performance slightly superior to it, and these generators were trained with approximately 6000 actual samples. In summary, although the use of GANs for data augmentation can indeed mitigate the challenges associated with model training with a low sample volume, it imposes higher demands on the training of GANs.

### Performance of the decomposers

A detailed comparison of the decomposition results from UDM and the decomposer at each site

was conducted (Fig. 7 and Figs S20 to S22). The results indicate that the decomposer's performance on different sediment types in the training sets is very consistent with UDM, suggesting that the decomposer can accurately learn the latent knowledge embedded in UDM's decomposition results (Video 2). Furthermore, most decomposition results are reliable when examining each sample's specific decomposition situation (Figs S23 to S25). However, for some test sets – especially those whose sample data distribution significantly differs from that of training samples (Fig. 8) – the decomposition results are suboptimal. For instance, the decomposition results on the three datasets from Arid Central Asia (ACA) – CMG, Osh and BSK – are not consistent with that of UDM (Fig. 7 and Fig. S22). Although both belong to aeolian deposition, due to certain differences in formation mechanisms and grain-size characteristics between loess in the CLP and ACA, it is challenging to directly apply a decomposer trained on CLP loess to decompose grain-size data from ACA loess. Additionally, the decomposer's performance when decomposing grain-size data from FS18 is not ideal, mainly reflected in the inaccuracy of C1 (Fig. 7). This may be because C1 in the FS18 profile is much finer compared to C1 at other sites (Fig. S1K). However,



**Fig. 6.** Influence of training set size on decomposer performance. LMSE refers to the overall logarithmic mean squared error of the decomposition results. Red horizontal lines indicate the median LMSE of the universal decomposition model (UDM) algorithm on the training/test sets.

the decomposer's performance is not poor on all test sets; it performs very well on LX and NLK (Fig. 7). This indicates that the decomposer retains a certain degree of generalization capability; at a minimum, a model trained on loess grain-size data since the last interglacial can be utilized to decompose loess grain-size data from older strata.

### Evaluation of the classifier

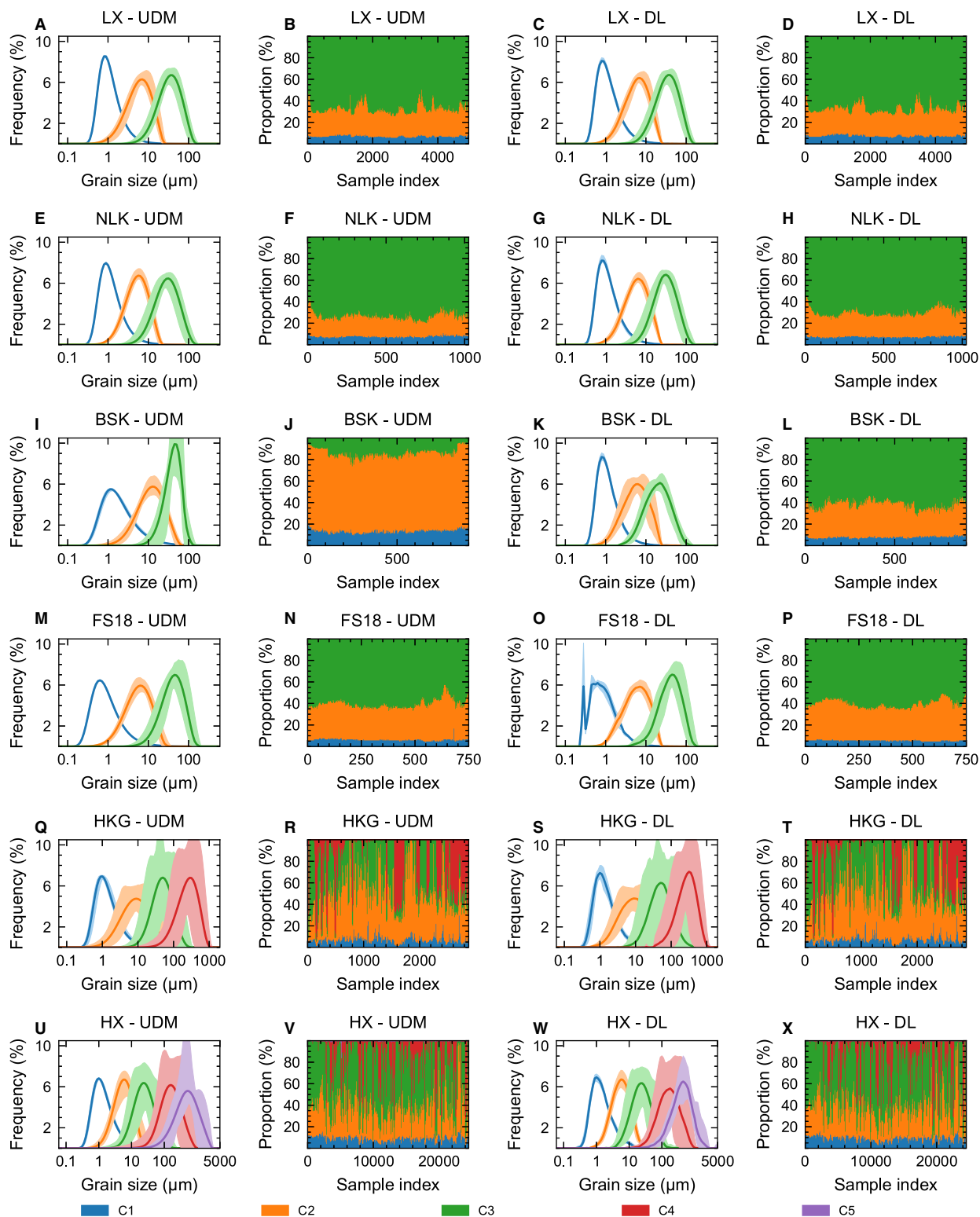
The classifier's test set, encompassing 14 656 randomly chosen samples, which constitute 20% of the total dataset size, was employed to evaluate the classifier's performance and to ascertain potential overfitting. The results obtained (Table 2) indicated an overall accuracy of 97% on the test set. Nonetheless, a marginally reduced precision was noted, specifically for samples derived from fluvial sediments. This inconsistency could be attributed to the oversimplification involved in classifying these three types of sediments, given that lake deltas also comprise sediments bearing close resemblance to fluvial sediments, thereby introducing challenges in attaining higher levels of accuracy. In summary, the trained classifier demonstrated commendable performance on the test set, as evidenced by multiple indicators, with no apparent signs of overfitting.

## DISCUSSION

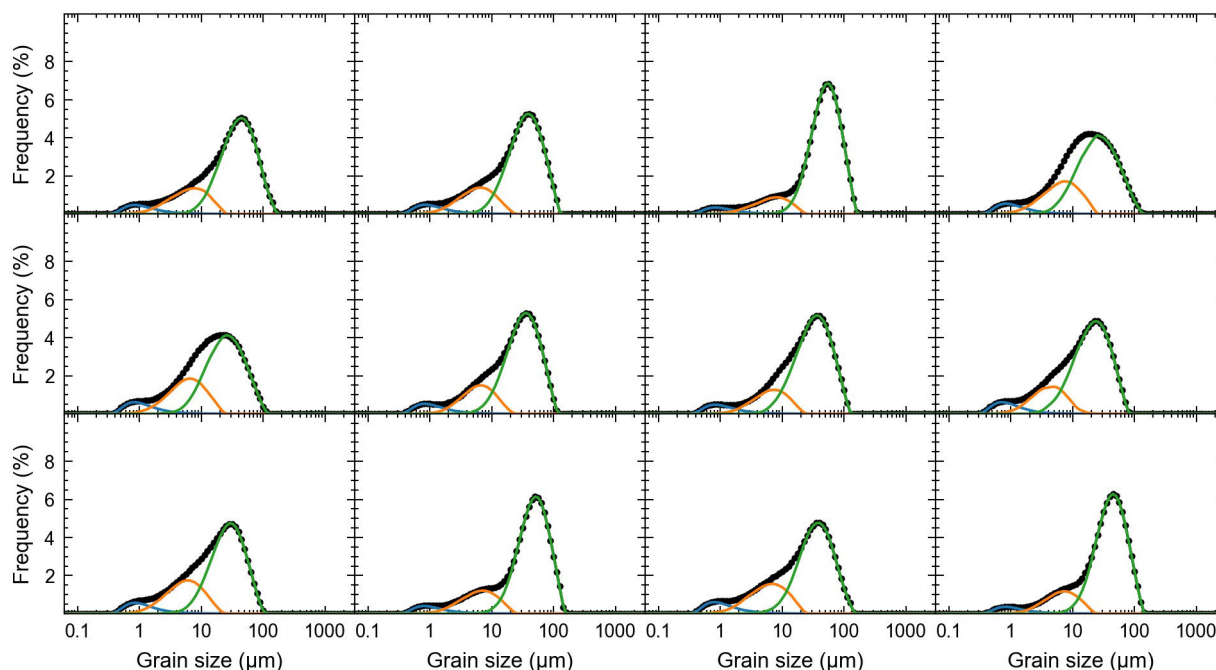
### Advantages and limitations of the proposed framework

The primary benefit of the deep learning-based decomposition framework lies in its ability to mitigate uncertainties arising from inappropriate algorithm configurations in traditional methods due to user inexperience. While traditional methods' decomposition results serve as training data within this framework, these results can be uniformly supplied by several experts with extensive professional expertise. This approach circumvents artificial uncertainties stemming from individual variations in professional knowledge backgrounds. Furthermore, decomposition models trained using deep learning techniques offer significant advantages in inferential performance, enabling rapid processing of data awaiting decomposition.

Moreover, deep learning-based models are not mere replacements for traditional methods. Particularly during initial stages when training datasets are limited and lack diversity, models may exhibit inadequate generalization capabilities. While these models cannot entirely supplant traditional methods in delivering highly accurate results, they can provide reasonable initial parameters to aid traditional methods in



**Fig. 7.** The comparison of the decomposition results on several representative sites yielded by the universal decomposition model (UDM) and the trained decomposers (denoted as DL). The first and second columns show the components and proportions obtained by UDM, respectively. The third and fourth columns show the components and proportions obtained by the trained decomposers, respectively. The first to fourth rows show the decomposition results of four representative loess test sets. The fifth and sixth rows show the decomposition results of fluvial and lake delta boreholes, respectively. For more results from other sites, please refer to the [Supplementary Information](#).



**Video 2:** The decomposition results of the loess decomposer on the training sets. This video lasts for 60 s, contains 300 frames, each frame displays the specific decomposition situation of 12 randomly selected samples, showing a total of 3600 decomposition results.

achieving superior outcomes and reducing their heavy reliance on parameter settings. Although this study employs the traditional method, UDM, for providing training data, it does not imply that the proposed deep learning framework is entirely tied to it. Future research could easily transition or incorporate multiple traditional methods should superior alternatives emerge. Importantly, comparing to the pure-deep learning approach, integrating with traditional methods can ensure the interpretability of the deep learning model.

### Possible approaches to improving the capacity of decomposers

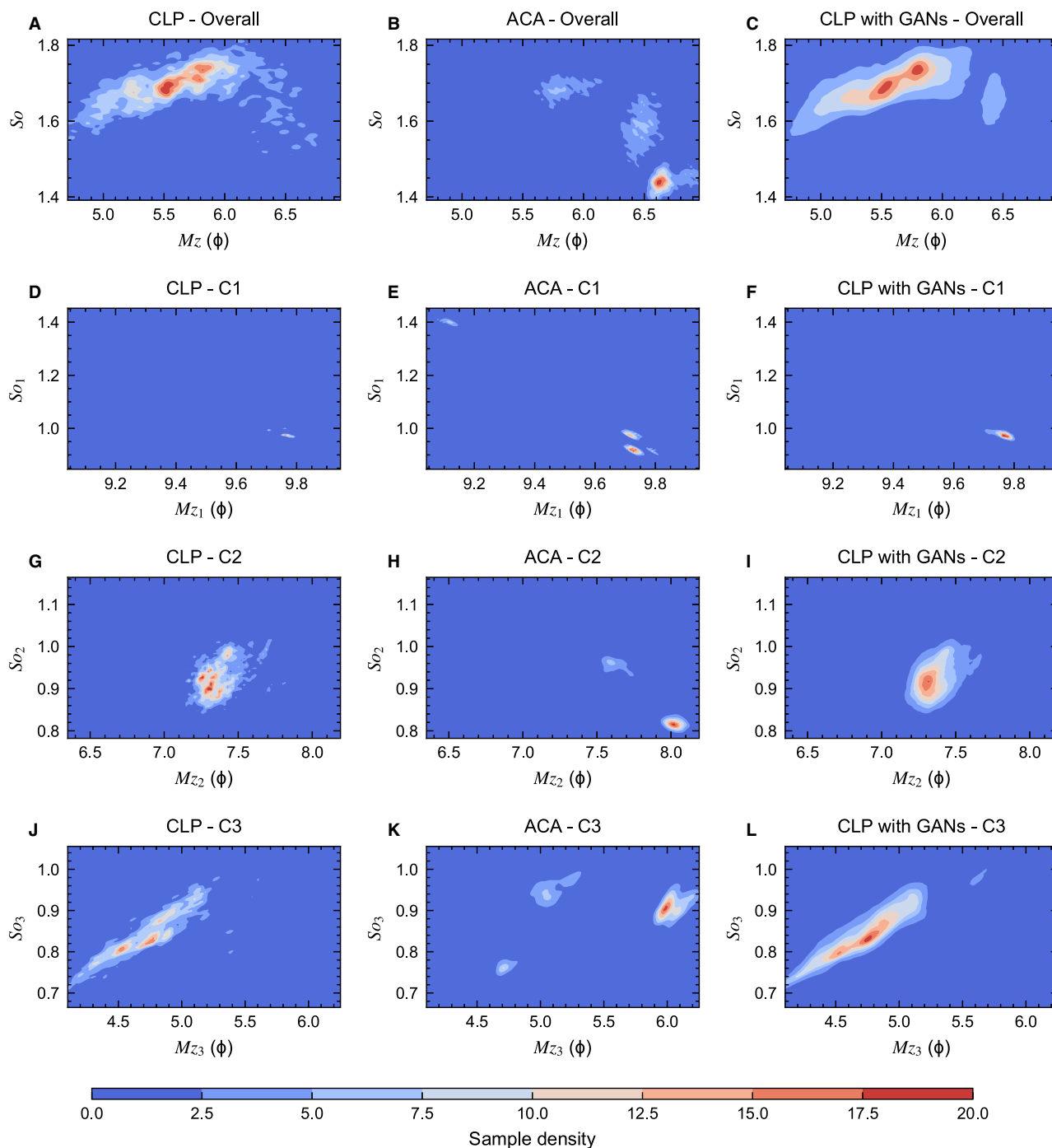
In deep learning models, our aspiration for them is to exhibit robust generalization capabilities, enabling them to handle diverse datasets. Given that deep learning currently operates on a data-driven paradigm, its generalization capabilities hinge on the diversity of data encountered during training. This has been demonstrated by the performances of our trained decomposers on the test sets. Consequently, enhancing a model's capability lies in supplying it with diverse training data. One prominent approach involves extensive data integration, which will be a focal point in future research. Alternatively, existing data can be

augmented. A key aspect of this study involved using GANs for data augmentation. The results indicate that this approach yielded positive outcomes and confirmed its viability. Besides this, conventional methods for data augmentation exist, such as preserving component shapes and making minor adjustments to their mode size. Future research could explore incorporating more conventional methods of data augmentation to further diversify training data.

### Prospects of grain-size research

Sediment grain-size analysis has been a prominent topic within sedimentology for nearly a century. The introduction of advanced tools like laser particle size analysers has substantially improved the generation of accurate and reliable grain-size data for fine particles, thereby leading to an array of new scientific challenges. A primary issue involves interpreting the geological significance tied to grain-size components and their fluctuations. This subject has been thoroughly investigated in numerous studies employing decomposition techniques, yet some interpretations remain disputed. For instance, interpretations regarding fine-grained components (*ca* 10  $\mu\text{m}$ ) in loess continue to be





**Fig. 8.** The comparison of the data distributions of loess samples from the Chinese Loess Plateau (CLP; training sets), arid Central Asian (ACA; test sets) and CLP combined with generative adversarial network (GAN)-generated data. Each subfigure uses mean grain size ( $Mz$ ) and sorting coefficient ( $So$ ) to show the sample distribution, and uses colour to indicate the sample density, with warmer colours indicating higher density. From top to bottom, the first row shows the distribution of the bulk samples, and the second to fourth rows show the distribution of components C1 to C3, respectively. By comparing the sample distributions of different datasets, the differences in their data diversity can be easily identified.

**Table 2.** The evaluation results of the classifier.

	Precision	Recall	F1-score	Support
Loess	0.98	0.98	0.98	3007
Fluvial	0.85	0.84	0.85	1184
Lake delta	0.98	0.98	0.98	10 465
Accuracy	–	–	0.97	14 656
Macro average	0.94	0.93	0.93	14 656
Weighted average	0.97	0.97	0.97	14 656

contentious (Qiang *et al.*, 2010; Vandenberghe *et al.*, 2018; Li *et al.*, 2022), despite attempts by some studies to utilize them in unveiling variability in westerlies (Sun, 2004; Prins *et al.*, 2007; Yang *et al.*, 2023). The findings of this study indicate pronounced differences in grain-size characteristics between CLP and ACA loess, with noticeable spatial heterogeneity within CLP loess profiles. Presumably, these disparities could be linked to loess formation mechanisms and associated environmental and climatic factors such as altitude, latitude and precipitation. Future research could leverage grain-size decomposition for an in-depth comparison of spatial heterogeneity among different components and their relationships with various factors, potentially leading to a more comprehensive understanding of these components' specific geological significance.

The conflicts on the interpretation have hampered the widespread use of grain-size decomposition and the potential of various grain-size indicators in sedimentology and palaeoclimatology. In general, these conflicts perhaps arise from incorrect applications of analysis methods or a lack of comprehensive understanding of the relationships between grain size and sedimentary environments. The first step towards resolving these conflicts involves addressing uncertainties associated with decomposition methods. It is essential to thoroughly compare these methods, ascertain their applicability and establish usage guidelines. Concurrently, it is crucial to draw insights from other disciplines and strive to develop innovative methods. A more precise understanding of the applicability of existing methods or re-analysis of previous data using new methods could allow for a re-evaluation of past conclusions, leading to more reliable insights. The deep learning-based decomposition framework proposed in this study aims to address uncertainties associated with current decomposition methods and establish a

reliable and user-friendly decomposition model. Although the framework has only been tested on a small-scale dataset comprising three types of sediments, it has demonstrated the feasibility of this approach. Once enough grain-size data from various sediment types are collected, an ultimate decomposition model can be trained following this framework. On another note, most current studies using grain-size decomposition are typically based on sediment samples from specific regions and lack large-scale integration and comparison. Without a global perspective, the conclusions drawn are often biased, leading to conflicting results across different studies. Hence, there is a pressing need for data integration and comparison, followed by comprehensive analysis using unified methods. This approach will facilitate a reliable understanding of fundamental issues that can guide future research.

Generally, the aforementioned solutions depend heavily on extensive grain-size data. Despite the academic community amassing a substantial amount of such data over years of research, the absence of specific guidelines for its release and sharing has hindered its full utilization. Consequently, it is urged that scholars and research communities promptly establish guidelines for grain-size data sharing, a step integral to sediment grain-size studies and their subsequent applications.

## CONCLUSIONS

In this paper, a novel deep learning-based framework for grain-size decomposition of terrigenous clastic sediments is introduced to reconcile a long-standing challenge in application of grain-size data for palaeoenvironmental reconstruction. By integrating traditional methods and deep learning techniques, a novel framework is developed that can deal with methodological uncertainty and subjectivity, providing a user-friendly, unified model for grain-size decomposition. Testing on a small-scale dataset of three sediment types shows promising results for the framework. The proposed framework can achieve comparable decomposition performance on all training sets and partial test sets. The poor performance on some test sets is due to their large differences from the training data, which implies the importance of data diversity. It is suggested that future research should focus on data collection and integration, introducing advanced deep learning techniques, and conducting global comparisons.

## ACKNOWLEDGEMENTS

We express our gratitude to the editor and anonymous reviewers for their constructive and enhancing feedback, which significantly improved the quality of our manuscript. We are also thankful to Yuxiang Shi and Jiawang Ge for their insightful comments on a preliminary version of this manuscript. Our appreciation extends to Hua Wang, Min Zhao, Yili Long and Huimin Fan for their assistance in the field and laboratory. The financial support for this study was provided by the Chinese Academy of Sciences (No. XDB40000000), the National Natural Science Foundation of China (No. 42177429), the Youth Innovation Promotion Association of the Chinese Academy of Sciences (No. 2023428), the open fund of State Key Laboratory of Loess and Quaternary Geology, IEECAS (No. SKLLQG2337) and the Xiaogan Natural Science Program (No. XGKJ2023010065).

## DATA AVAILABILITY STATEMENT

The improved UDM's algorithm implementation, in conjunction with the QGrain software, is accessible on GitHub at the following link: <https://github.com/yuriok/QGrain>. The scripts utilized for constructing and training the deep learning models, overseeing the training procedures, and assessing the training outcomes can be found in this GitHub repository: <https://github.com/yuriok/gSdecomposer>. Additional technical details are available in the [Supplementary Information](#) accompanying this paper. Additional data related to this paper may be requested from the authors.

## REFERENCES

- Al-Najjar, H.A.H. and Pradhan, B. (2021) Spatial landslide susceptibility assessment using machine learning techniques assisted by additional data created with generative adversarial networks. *Geosci. Front.*, **12**, 625–637.
- Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Van Esesn, B.C., Awwal, A.A.S. and Asari, V.K. (2018) The history began from AlexNet: a comprehensive survey on deep learning approaches. *arXiv*, <https://doi.org/10.48550/arXiv.1803.01164>.
- An, Z., Sun, Y., Zhou, W., Liu, W., Qiang, X., Wang, X., Xian, F., Cheng, P. and Burr, G.S. (2014) Chinese loess and the East Asian monsoon. In: *Late Cenozoic Climate Change in Asia* (Ed. An, Z.), Springer Netherlands, Dordrecht, **16**, 23–143.
- An, Z., Zhang, P., Vogel, H., Song, Y., Dodson, J., Wiersberg, T., Feng, X., Lu, H., Ai, L. and Sun, Y. (2020) Scientific drilling workshop on the Weihe Basin Drilling Project (WBDP): cenozoic tectonic–monsoon interactions. *Sci. Drill.*, **28**, 63–73.
- Arjovsky, M., Chintala, S. and Bottou, L. (2017) Wasserstein generative adversarial networks. In: *Proceedings of the 34th International Conference on Machine Learning*, PMLR, 214–223.
- Bergen, K.J., Johnson, P.A., de Hoop, M.V. and Beroza, G.C. (2019) Machine learning for data-driven discovery in solid Earth geoscience. *Science*, **363**, eaau0323.
- Cheetham, M.D., Keene, A.F., Bush, R.T., Sullivan, L.A. and Erskine, W.D. (2008) A comparison of grain-size analysis methods for sand-dominated fluvial sediments. *Sedimentology*, **55**, 1905–1913.
- D'Arcy, M., Roda-Boluda, D.C. and Whittaker, A.C. (2017) Glacial-interglacial climate changes recorded by debris flow fan deposits, Owens Valley, California. *Quatern. Sci. Rev.*, **169**, 288–311.
- Dhillon, A. and Verma, G.K. (2020) Convolutional neural network: a review of models, methodologies and applications to object detection. *Prog. Artif. Intell.*, **9**, 85–112.
- Dietze, E. and Dietze, M. (2019) Grain-size distribution unmixing using the R package EMMAgeo. *E&G Quat. Sci. J.*, **68**, 29–46.
- Dietze, E., Hartmann, K., Diekmann, B., Ijmker, J., Lehmkuhl, F., Opitz, S., Stauch, G., Wünnemann, B. and Borchers, A. (2012) An end-member algorithm for deciphering modern detrital processes from lake sediments of Lake Donggi Cona, NE Tibetan Plateau, China. *Sed. Geol.*, **243–244**, 169–180.
- Dietze, M., Schulte, P. and Dietze, E. (2022) Application of end-member modelling to grain-size data: constraints and limitations. *Sedimentology*, **69**, 845–863.
- Epuna, F., Shaheen, S.W. and Wen, T. (2022) Road salting and natural brine migration revealed as major sources of groundwater contamination across regions of northern Appalachia with and without unconventional oil and gas development. *Water Res.*, **225**, 119128.
- Esmaeiloghli, S., Tabatabaei, S.H. and Carranza, E.J.M. (2023) Infomax-based deep autoencoder network for recognition of multi-element geochemical anomalies linked to mineralization. *Comput. Geosci.*, **175**, 105341.
- Fan, M., Song, C., Dettman, D.L., Fang, X. and Xu, X. (2006) Intensification of the Asian winter monsoon after 7.4 Ma: grain-size evidence from the Linxia Basin, northeastern Tibetan Plateau, 13.1 Ma to 4.3 Ma. *Earth Planet. Sci. Lett.*, **248**, 186–197.
- Folk, R.L. (1966) A review of grain-size parameters. *Sedimentology*, **6**, 73–93.
- Folk, R.L. (1980) *Petrology of Sedimentary Rocks*. Hemphill Publishing Company, Austin, TX.
- Folk, R.L. and Ward, W.C. (1957) Brazos River bar [Texas]; a study in the significance of grain size parameters. *J. Sed. Res.*, **27**, 3–26.
- Galloway, W.E. and Hobday, D.K. (2012) *Terrigenous Clastic Depositional Systems: Applications to Petroleum, Coal, and Uranium Exploration*. Springer Science & Business Media, Berlin/Heidelberg, Germany.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) Generative adversarial nets. In: *Advances in Neural Information Processing Systems* (Eds Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D. and Weinberger, K.Q.), Curran Associates, Inc., **27**, 2672–2680.

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2020) Generative adversarial networks. *Commun. ACM*, **63**, 139–144.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A.C. (2017) Improved training of Wasserstein GANs. In: *Advances in Neural Information Processing Systems* (Eds Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D. and Weinberger, K.Q.), Curran Associates, Inc., **30**, 5767–5777.
- Guo, F., Clemens, S.C., Wang, T., Wang, Y., Liu, Y., Wu, F., Liu, X., Jin, Z. and Sun, Y. (2021) Monsoon variations inferred from high-resolution geochemical records of the Linxia loess/paleosol sequence, western Chinese Loess Plateau. *Catena*, **198**, 105019.
- Hajek, E.A., Huzurbazar, S.V., Mohrig, D., Lynds, R.M. and Heller, P.L. (2010) Statistical characterization of grain-size distributions in sandy fluvial systems. *J. Sediment. Res.*, **80**, 184–192.
- van Hateren, J.A., Prins, M.A. and van Balen, R.T. (2018) On the genetically meaningful decomposition of grain-size distributions: a comparison of different end-member modelling algorithms. *Sed. Geol.*, **375**, 49–71.
- He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.
- He, Y., Zhou, Y., Wen, T., Zhang, S., Huang, F., Zou, X., Ma, X. and Zhu, Y. (2022) A review of machine learning in geochemistry and cosmochemistry: method improvements and applications. *Appl. Geochem.*, **140**, 105273.
- Heslop, D., von Dobeneck, T. and Höcker, M. (2007) Using non-negative matrix factorization in the “unmixing” of diffuse reflectance spectra. *Mar. Geol.*, **241**, 63–78.
- Hou, K., Qian, H., Zhang, Y., Zhang, Q. and Qu, W. (2021) New insights into loess formation on the southern margin of the Chinese Loess Plateau. *Catena*, **204**, 105444.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P. and Hassabis, D. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
- Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.*, **25**, 1097–1105.
- Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proc. IEEE*, **86**, 2278–2324.
- LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning. *Nature*, **521**, 436–444.
- Li, Y., Song, Y., Fitzsimmons, K.E., Chen, X., Prud'homme, C. and Zong, X. (2020) Origin of loess deposits in the North Tian Shan piedmont, Central Asia. *Palaeogeogr. Palaeoclimatol. Palaeoecol.*, **559**, 109972.
- Li, Y., Song, Y., Kaskaoutis, D.G., Zan, J., Orozbaev, R., Tan, L. and Chen, X. (2021) Aeolian dust dynamics in the Fergana Valley, Central Asia, since ~30 ka inferred from loess deposits. *Geosci. Front.*, **12**, 101180.
- Li, Y., Song, Y., Fitzsimmons, K.E., Dave, A.K., Liu, Y., Zong, X., Sun, H., Liu, H. and Orozbaev, R. (2022) Investigating potential links between fine-grained components in loess and westerly airflow: evidence from East and Central Asia. *Front. Earth Sci.*, **10**. <https://doi.org/10.3389/feart.2022.901629>.
- Li, Y., Song, Y., Chen, X., Shi, Z., Kaskaoutis, D.G., Gholami, H. and Li, Y. (2023) Late Pleistocene dynamics of dust emissions related to westerlies revealed by quantifying loess provenance changes in North Tian Shan, Central Asia. *Catena*, **227**, 107101.
- Liang, P. and Yang, X. (2023) Grain shape evolution of sand-sized sediments during transport from mountains to dune fields. *J. Geophys. Res. Earth*, **128**, e2022JF006930.
- Liu, T. (1985) *Loess and the Environment*. China Ocean Press, Beijing, China.
- Liu, Y., Liu, X. and Sun, Y. (2021) QGrain: an open-source and easy-to-use software for the comprehensive analysis of grain size distributions. *Sed. Geol.*, **423**, 105980.
- Liu, Y., Wang, T., Liu, B., Long, Y., Liu, X. and Sun, Y. (2023) Universal decomposition model: an efficient technique for palaeoenvironmental reconstruction from grain-size distribution. *Sedimentology*, **70**, 2127–2149.
- Lu, H. and An, Z. (1997) The influence of pre-treatment to grain size analysis results of loess. *Chin. Sci. Bull.*, **42**, 2535–2538.
- Ma, L., Li, Y., Liu, X. and Sun, Y. (2017) Registration of precession signal in the last interglacial paleosol (S<sub>1</sub>) on the Chinese Loess Plateau. *Geochem. Geophys. Geosyst.*, **18**, 3964–3975.
- McLaren, P. and Bowles, D. (1985) The effects of sediment transport on grain-size distributions. *J. Sed. Res.*, **55**, 457–470.
- Miyato, T., Kataoka, T., Koyama, M. and Yoshida, Y. (2018) Spectral normalization for generative adversarial networks. In: *International Conference on Learning Representations*.
- Pan, B., Pang, H., Zhang, D., Guan, Q., Wang, L., Li, F., Guan, W., Cai, A. and Sun, X. (2015) Sediment grain-size characteristics and its source implication in the Ningxia–Inner Mongolia sections on the upper reaches of the Yellow River. *Geomorphology*, **246**, 255–262.
- Paterson, G.A. and Heslop, D. (2015) New methods for unmixing sediment grain size data. *Geochem. Geophys. Geosyst.*, **16**, 4494–4506.
- Peng, J., Zhao, H., Dong, Z., Zhang, Z., Yang, H. and Wang, X. (2022) Numerical methodologies and tools for efficient and flexible unmixing of single-sample grain-size distributions: application to late quaternary aeolian sediments from the desert-loess transition zone of the Tengger Desert. *Sed. Geol.*, **438**, 106211.
- Pettijohn, F.J. (1975) *Sedimentary Rocks*, 3rd edn. Harper and Row Publishers, New York.
- Prins, M.A., Vriend, M., Nugteren, G., Vandenberghe, J., Lu, H., Zheng, H. and Jan Weltje, G. (2007) Late quaternary aeolian dust input variability on the Chinese Loess Plateau: inferences from unmixing of loess grain-size records. *Quatern. Sci. Rev.*, **26**, 230–242.
- Pye, K. (1995) The nature, origin and accumulation of loess. *Quatern. Sci. Rev.*, **14**, 653–667.
- Qiang, M., Lang, L. and Wang, Z. (2010) Do fine-grained components of loess indicate westerlies: insights from observations of dust storm deposits at Lenghu (Qaidam Basin, China). *J. Arid Environ.*, **74**, 1232–1239.
- Qin, X., Cai, B. and Liu, T. (2005) Loess record of the aerodynamic environment in the east Asia monsoon area since 60,000 years before present. *J. Geophys. Res.*, **110**, B01204.
- Reineck, H.-E. and Singh, I.B. (2012) *Depositional Sedimentary Environments: With Reference to Terrigenous Clastics*. Springer Science & Business Media, Berlin.
- Ruddiman, W.F. (2001) *Earth's Climate: Past and Future*. MacMillan, New York, NY.

- Seidel, M. and Hlawitschka, M. (2015) An R-based function for modeling of end member compositions. *Math. Geosci.*, **47**, 995–1007.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T. and Hassabis, D. (2017) Mastering the game of go without human knowledge. *Nature*, **550**, 354–359.
- Sun, D. (2004) Monsoon and westerly circulation changes recorded in the late Cenozoic aeolian sequences of Northern China. *Glob. Planet. Chang.*, **41**, 63–80.
- Sun, D., Bloemendal, J., Rea, D.K., Vandenberghe, J., Jiang, F., An, Z. and Su, R. (2002) Grain-size distribution function of polymodal sediments in hydraulic and aeolian environments, and numerical partitioning of the sedimentary components. *Sed. Geol.*, **152**, 263–277.
- Újvári, G., Kok, J.F., Varga, G. and Kovács, J. (2016) The physics of wind-blown loess: implications for grain size proxy interpretations in quaternary paleoclimate studies. *Earth-Sci. Rev.*, **154**, 247–278.
- Vandenberghe, J. (2013) Grain size of fine-grained windblown sediment: a powerful proxy for process identification. *Earth-Sci. Rev.*, **121**, 18–30.
- Vandenberghe, J., Sun, Y., Wang, X., Abels, H.A. and Liu, X. (2018) Grain-size characterization of reworked fine-grained aeolian deposits. *Earth-Sci. Rev.*, **177**, 43–52.
- Varga, G., Újvári, G. and Kovács, J. (2019) Interpretation of sedimentary (sub)populations extracted from grain size distributions of Central European loess-paleosol series. *Quat. Int.*, **502**, 60–70.
- Vaserstein, L.N. (1969) Markov processes over denumerable products of spaces, describing large systems of automata. *Probl. Peredachi Inf.*, **5**, 64–72.
- Visher, G.S. (1969) Grain size distributions and depositional processes. *J. Sed. Res.*, **39**, 1074–1106.
- Wang, F., Wang, F., Zhang, W., Xu, S. and Lai, Z. (2023) A novel machine learning fingerprinting method using sparse representation for provenance detection in delta sediments. *Catena*, **227**, 107095.
- Weltje, G.J. (1997) End-member modeling of compositional data: numerical-statistical algorithms for solving the explicit mixing problem. *Math. Geol.*, **29**, 503–549.
- Weltje, G.J. and Prins, M.A. (2003) Muddled or mixed? Inferring palaeoclimate from size distributions of deep-sea clastics. *Sed. Geol.*, **162**, 39–62.
- Weltje, G.J. and Prins, M.A. (2007) Genetically meaningful decomposition of grain-size distributions. *Sed. Geol.*, **202**, 409–424.
- Wen, T., Chen, C., Zheng, G., Bandstra, J. and Brantley, S.L. (2022) Using a neural network – physics-based hybrid model to predict soil reaction fronts. *Comput. Geosci.*, **167**, 105200.
- Wu, L., Krijgsman, W., Liu, J., Li, C., Wang, R. and Xiao, W. (2020) CFLab: a MATLAB GUI program for decomposing sediment grain size distribution using Weibull functions. *Sed. Geol.*, **398**, 105590.
- Yang, S., Zhou, J., Chen, Z., Li, P., Wen, C., Xu, X. and Li, Q. (2023) Westerly Variations in the Eastern Tibetan Plateau since the last interglacial revealed by the grain-size records of the Ganzi loess. *Atmos.*, **14**, 238.
- Yu, S., Colman, S.M. and Li, L. (2016) BEMMA: a hierarchical Bayesian end-member modeling analysis of sediment grain-size distributions. *Math. Geosci.*, **48**, 723–741.
- Zhang, X., Li, Z., Li, P., Cheng, S., Zhang, Y., Tang, S. and Wang, T. (2015) A model to study the grain size components of the sediment deposited in aeolian–fluvial interplay erosion watershed. *Sed. Geol.*, **330**, 132–140.
- Zhang, X., Zhou, A., Wang, X., Song, M., Zhao, Y., Xie, H., Russell, J.M. and Chen, F. (2018) Unmixing grain-size distributions in lake sediments: a new method of endmember modeling using hierarchical clustering. *Quat. Res.*, **89**, 365–373.
- Zhang, X., Wang, H., Xu, S. and Yang, Z. (2020) A basic end-member model algorithm for grain-size data of marine sediments. *Estuar. Coast. Shelf Sci.*, **236**, 106656.
- Zhang, J., Lv, D., Chen, H., Yu, C., Zhao, K., Liu, X., Liu, Y., Zhang, H., Liu, B., Qiang, X., Kang, S. and Sun, Y. (2024) Evolution of sedimentary environment in the Eastern Henan Basin since the Late Pliocene. *Palaeogeogr. Palaeoclimatol. Palaeoecol.*, **633**, 111896.

Manuscript received 13 November 2023; revision accepted 27 March 2024

## Supporting Information

Additional information may be found in the online version of this article:

**Figure S1.** The measured grain-size distributions (GSDs) (left), components (middle) and proportions (right) decomposed by UDM.

**Figure S2.** The measured GSDs (left), components (middle) and proportions (right) decomposed by UDM.

**Figure S3.** The measured GSDs (left), components (middle) and proportions (right) decomposed by UDM.

**Figure S4.** The loss variations of the generator and discriminator in MLP (multilayer perceptron) GAN for loess.

**Figure S5.** The comparison of the GSDs, scaled components, and original components between the real loess samples (left) (decomposed by UDM) and generated loess samples by MLP GAN (right).

**Figure S6.** Eighteen grain-size samples of loess randomly generated by MLP GAN.

**Figure S7.** The probability density of each statistical parameter (mean, standard deviation, skewness, kurtosis) for generated loess samples by MLP GAN.

**Figure S8.** The comparison of the data distribution between training loess samples (left) and generated loess samples (right) by MLP GAN.

**Figure S9.** The variation in the component precision of each component as the number of training batches increases during the training processes of MLP GAN for loess.

**Figure S10.** The variation in the Wasserstein distance of each statistical parameter of each component as the

number of training batches increases during the training processes of MLP GAN for loess.

**Figure S11.** The variations in the evaluation metrics of the generators as the number of training batches increases.

**Figure S12.** Comparison of the distributions learned by the generators with the true distributions of the measured samples.

**Figure S13.** Comparison of the sample distributions learned by different generators with the observed distributions of the measured samples for the loess datasets.

**Figure S14.** Comparison of the sample distributions learned by different generators with the observed distributions of the measured samples for the fluvial datasets.

**Figure S15.** Comparison of the sample distributions learned by different generators with the observed distributions of the measured samples for the lake delta datasets.

**Figure S16.** Eighteen grain-size samples of loess randomly generated by CONV (convolutional) SNGAN.

**Figure S17.** Eighteen grain-size samples of fluvial deposits randomly generated by CONV SNGAN.

**Figure S18.** Eighteen grain-size samples of lake delta sediments randomly generated by CONV SNGAN.

**Figure S19.** The specific results of the control experiments in each profile and borehole.

**Figure S20.** The comparison of the decomposition results yielded by UDM and the trained decomposer (denoted as DL).

**Figure S21.** The comparison of the decomposition results yielded by UDM and the trained decomposer (denoted as DL).

**Figure S22.** The comparison of the decomposition results yielded by UDM and the trained decomposer (denoted as DL).

**Figure S23.** The specific decomposition results of several loess samples yielded by the trained decomposer.

**Figure S24.** The specific decomposition results of several fluvial samples yielded by the trained decomposer.

**Figure S25.** The specific decomposition results of several lake delta samples yielded by the trained decomposer.

**Table S1.** The initial parameters of each loess site for the UDM algorithm.

**Table S2.** The numbers of trainable parameters for different generators and discriminators.

**Appendix S1.** Technical details and comprehensive results of the study.