



ELSEVIER

Contents lists available at ScienceDirect

Minerals Engineering

journal homepage: www.elsevier.com



## Towards Intelligent Mining for Backfill: A genetic programming-based method for strength forecasting of cemented paste backfill

Chongchong Qi<sup>a,\*</sup>, Xiaolin Tang<sup>b,\*</sup>, Xiangjian Dong<sup>a,\*</sup>, Qiusong Chen<sup>c</sup>, Andy Fourie<sup>a</sup>, Enyan Liu<sup>c</sup>

<sup>a</sup> School of Civil, Environmental and Mining Engineering, University of Western Australia, Perth 6009, Australia

<sup>b</sup> Business School, University of Western Australia, Perth 6009, Australia

<sup>c</sup> School of Resources and Safety Engineering, Central South University, Changsha 410083, China

### ARTICLE INFO

#### Keywords:

Cemented paste backfill  
Genetic programming  
Uniaxial compressive strength  
Forecasting  
Intelligent Mining for Backfill

### ABSTRACT

As cemented paste backfill (CPB) plays an increasingly important role in minerals engineering, forecasting its mechanical properties becomes a necessity for efficient CPB design. Machine learning (ML) techniques have previously demonstrated remarkable successes in such task by providing black-box predictions. To express the non-linear relationship in an explicit and precise way, we employed genetic programming (GP) for the uniaxial compressive strength (UCS) prediction of CPB. The influence of sampling method, training set size and maximum tree depth on the GP performance was investigated. A detailed analysis was conducted on a representative GP model and the relative variable importance was investigated using the relative variable frequency, partial dependence plots and relative importance scores. The statistical parameters show that a satisfactory performance was obtained by the GP modelling ( $R^2 > 0.80$  on the testing set). Results of this study indicate that cement-tailings ratio, solids content and curing time were the most three important variables for the UCS prediction. The predictive performance of GP modelling was comparable to well-recognised ML techniques, and the trained GP model can be generalised to entirely new tailings with satisfactory performance. This study indicates that the GP-based method is capable of providing explicit and precise forecasting of UCS, which can serve as a reliable tool for quick, inexpensive and effective assessment of UCS in the absence of adequate experimental data.

### Nomenclature

CPB	cemented paste backfill
CTR	cement-tailings ratio
DT	decision tree
ET	expression tree
GBM	gradient boosting machine
GP	genetic programming
GP_5, GP_10 and GP_15	GP models with a 5, 10 and 15 maximum tree depth
IA	Willmott's index of agreement
IRS	iterative random sampling
KS	Kennard-Stone sampling
ML	machine learning
N_mean	normalised mean value

(N)SIV	(Non-)significant input variables in Eq. (7)
PSD	particle size distribution
RF	random forest
RGP_5	representative GP_5 model
RMSE	root-mean-square error
SC	solids content
SD and N_SD	standard deviation and Normalised SD
T	curing time
UCS	uniaxial compressive strength
XRD	X-ray diffraction
$C_u$ and $C_c$	coefficient of uniformity and curvature
$C_v$	coefficient of variance
$G_s$	specific gravity
$k$	slope of regression lines
$R$	correlation coefficient
$R^2$	coefficient of determination

\* Corresponding authors.

Email addresses: 21948042@student.uwa.edu.au (C. Qi); xiaolin.tang@research.uwa.edu.au (X. Tang); xiangjian.dong@research.uwa.edu.au (X. Dong)

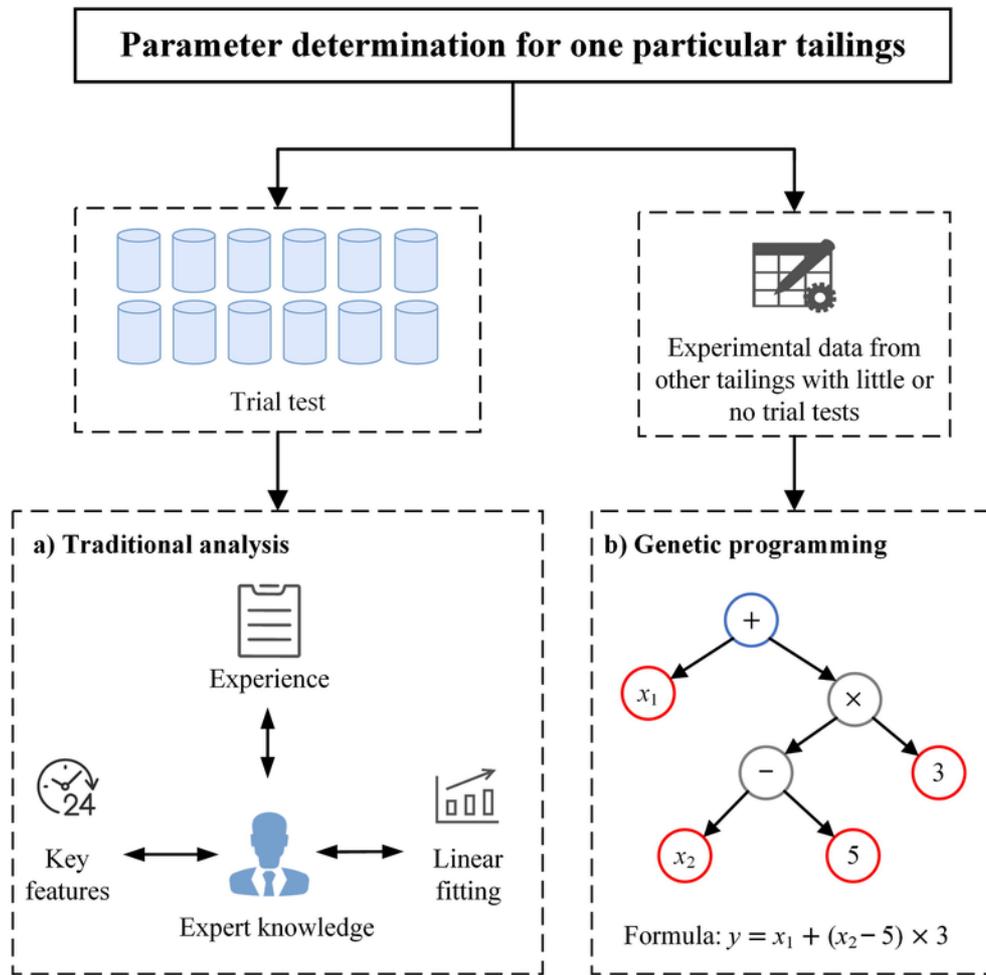


Fig. 1. Traditional and genetic programming methods in UCS analysis.

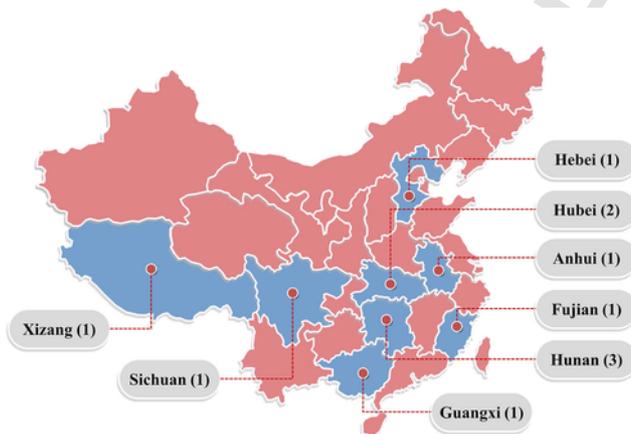


Fig. 2. Distribution of mine sites in China for tailings collection.

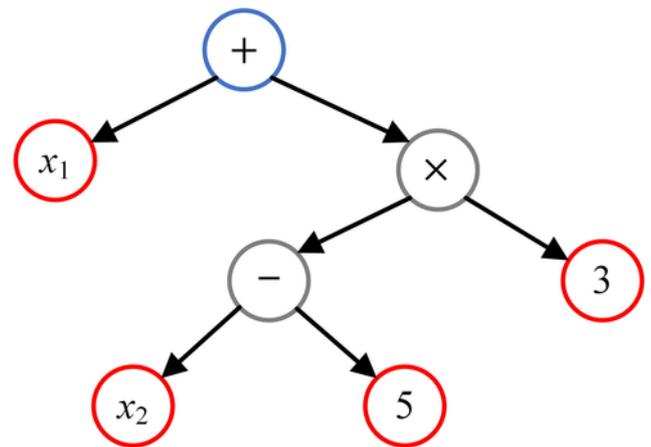


Fig. 3. A sample ET of GP.

$x'_i$  and  $y'_i$  input and output values for the  $i$ th sample  
 $x_i$  and  $y_i$  normalised  $x'_i$  and  $y'_i$   
 $x'_{min}$ ,  $x'_{max}$ ,  $y'_{min}$ , and  $y'_{max}$  minimum and maximum values

**1. Introduction**

Cemented paste backfill (CPB) is gaining attention in both academia and industry because of its potential as an environmental way of tail-

ings management (Yilmaz and Fall, 2017). In addition, it can stabilise rock mass and provide ground support for subsequent mining operations (Liu et al., 2019; Qi et al., 2018a). Another interesting feature of using CPB is to increase the ore recovery rate by reducing ore pillars. Created by mixing dewatered tailings, binders and water, CPB has been

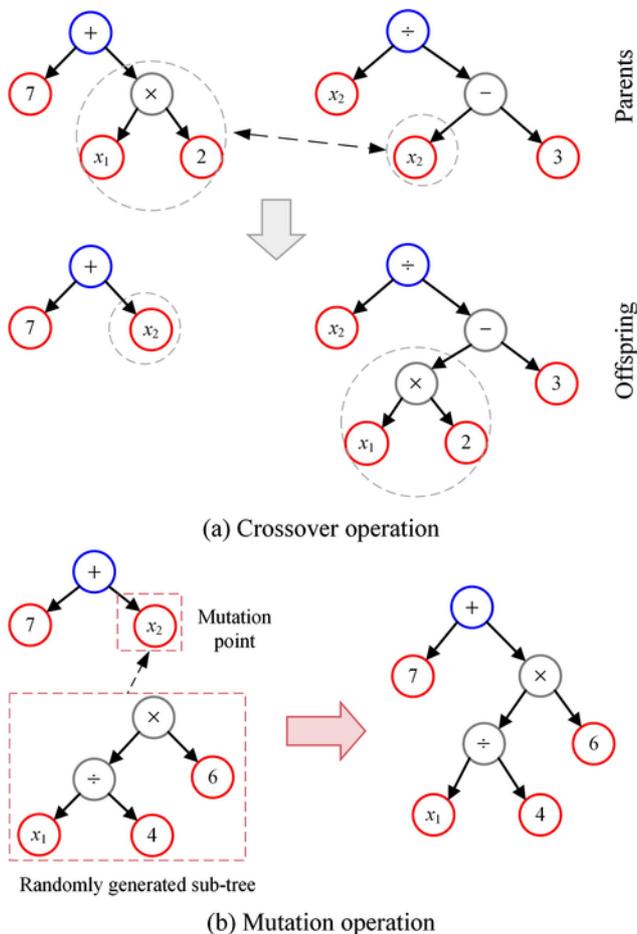


Fig. 4. Illustration of crossover and mutation procedures.

**Table 1**  
GP parameters for UCS modelling.

GP parameters		Parameter values
Initial parameters	Population size	500
	Maximum number of generations	1000
	Initialization method	Half
	Fitness function	$R^2$
	Selection method	Tournament
	Maximum tree depth	5, 10, 15
Genetic operators	Stopping criteria	Maximum generation
	Crossover probability	90%
	Mutation probability	15%

increasingly employed worldwide since its introduction (Chen et al., 2018; Cihangir et al., 2015; Kesimal et al., 2003; Liu et al., 2018; Lu et al., 2018; Mangane et al., 2018; Qi et al., 2019; Qi et al., 2018c; Yilmaz et al., 2009).

In spite of the importance of CPB and its exciting advantages, a throughout understanding of its mechanical properties is still lacking. This research gap is especially evident when all chemical and physical variables of CPB mechanical properties are considered, such as the mineralogy and type of the processing tailings. For example, the uniaxial compressive strength (UCS) is one of the most important mechanical properties of CPB, which plays a crucial role in its engineering designs and successful applications. However, cumbersome and costly experiments are required for each type of tailings and the CPB process parameters (i.e. cement-tailings ratio) are often determined by inspection (see Fig. 1a). A large number of experimental data has been generated for various tailings worldwide, which remains unexploited, at

least in most cases, due to the absence of appropriate techniques to manipulate extensive data, analyse high-dimensional inputs and possess good generalisation capability.

Recent progress in machine learning (ML) suggests that it is possible to construct an ML model that can learn the relationship between CPB mechanical properties and their influencing variables (Orejarena and Fall, 2010a; Orejarena and Fall, 2010b; Qi et al., 2018b; Qi et al., 2018d). Our previous study has shown that it is promising to predict the UCS, yield strength, Young's modulus and uniaxial tensile strength from physical and chemical properties of tailings, cement-tailings ratio, solids content, and curing time (Qi et al., 2018a). Though these ML techniques learn the relationship from a large amount of experimental data and generalise past experience to new situations with satisfactory performance, none of them put forward an explicit formulation. Consequently, these approaches are too complex to be interpreted, especially for practitioners with little or no ML knowledge. Software development can solve this problem to some extent but is still restricted by its engineering availability due to license considerations. Therefore a great need arises for developing forecasting methods with both acceptable accuracy and explicit nature to fulfil this requirement.

Genetic programming (GP) has emerged as an explicit and precise modelling technique, which is still a missing component in CPB design. It expresses the non-linear relationship as an expression tree (ET) and is regarded as the developed version of genetic algorithms (Khandelwal et al., 2017b). Utilization of GP in the field of minerals engineering has been limited to a few studies in the recent years. Ross et al. (2005) used GP to evolve mineral identification functions for hyperspectral images and found that the GP-evolved mineral classifiers are encouraging in identifying the existence or absence of particular minerals. Nazari (2012) employed GP to forecast the water absorption percentage of geopolymers created by mixing fly ash and rice husk-bark ash. Hoseinian et al. (2017) developed a GP-based method to predict the power of the semi-autogenous mill for minerals grinding.

The objective of this study is to apply GP modelling to forecast the UCS of CPB samples (Fig. 1b), which will contribute to current value in the following ways: (i) An enlarged dataset was collected using 1545 UCS tests performed on 11 types of tailings; (ii) A comprehensive sensitivity study was performed for sampling method, training set size and maximum tree depth; (iii) Results of the GP modelling were analysed in detail and compared to the results from well-recognised ML techniques; (iv) The generalisation capability of the trained GP model to entirely new tailings was investigated. The authors believe CPB property prediction is of vital significance for establishing the 'Intelligent Mining for Backfill (IMB)' system, in which artificial intelligence (AI) techniques are used in CPB design to promote its application. The developed model can serve as a reliable tool for quick, inexpensive and effective assessment of UCS in the absence of adequate experimental data, and can work as a benchmark study for future application of GP in minerals engineering.

## 2. Specification of the study

### 2.1. Dataset preparation

A dataset with accurate experimental values and wide distribution is a prerequisite for the successful application of the GP modelling. In this paper, extensive UCS experiments were performed on 11 types of tailings. The distribution of mine sites is shown in Fig. 2 and the experimental design for each type of tailings is presented in Table S1 (supplementary material). The authors note that different experimental designs were used for various tailings, which were determined by inspection and engineering requirements.

No. 325 Portland cement and tap water were used as the binder agent and the mixing water, respectively. The CPB materials, including

**Table 2**  
Descriptive statistics of inputs and output.

Variables	No.	Min	Max	Mean	SD	Skewness	C <sub>v</sub>	N <sub>mean</sub>	N <sub>SD</sub>
G <sub>s</sub>	x <sub>1</sub>	2.08	3.27	2.84	0.29	-0.66	0.10	0.64	0.24
D10	x <sub>2</sub>	0.00	0.25	0.03	0.06	2.75	1.97	0.12	0.26
D50	x <sub>3</sub>	0.01	1.44	0.18	0.35	3.11	1.90	0.12	0.25
C <sub>u</sub>	x <sub>4</sub>	4.65	30.74	16.59	10.01	0.29	0.60	0.46	0.38
C <sub>c</sub>	x <sub>5</sub>	0.87	2.04	1.08	0.27	2.04	0.25	0.18	0.23
SiO <sub>2</sub>	x <sub>6</sub>	29.11	72.50	46.07	14.32	0.67	0.31	0.39	0.33
CaO	x <sub>7</sub>	0.37	32.65	15.09	11.40	0.11	0.76	0.46	0.35
Al <sub>2</sub> O <sub>3</sub>	x <sub>8</sub>	0.37	25.04	8.32	6.50	1.18	0.78	0.32	0.26
MgO	x <sub>9</sub>	0.13	13.02	3.44	3.73	1.67	1.09	0.26	0.29
Fe <sub>2</sub> O <sub>3</sub>	x <sub>10</sub>	0.49	10.67	5.49	3.83	0.29	0.70	0.49	0.38
CTR	x <sub>11</sub>	0.05	0.33	0.15	0.07	0.90	0.49	0.36	0.26
SC	x <sub>12</sub>	59.00	81.00	72.96	5.12	-0.91	0.07	0.63	0.23
T	x <sub>13</sub>	3.00	60.00	17.86	16.14	1.22	0.90	0.26	0.28
UCS	y	0.01	10.94	1.09	1.41	2.95	1.30	0.10	0.13

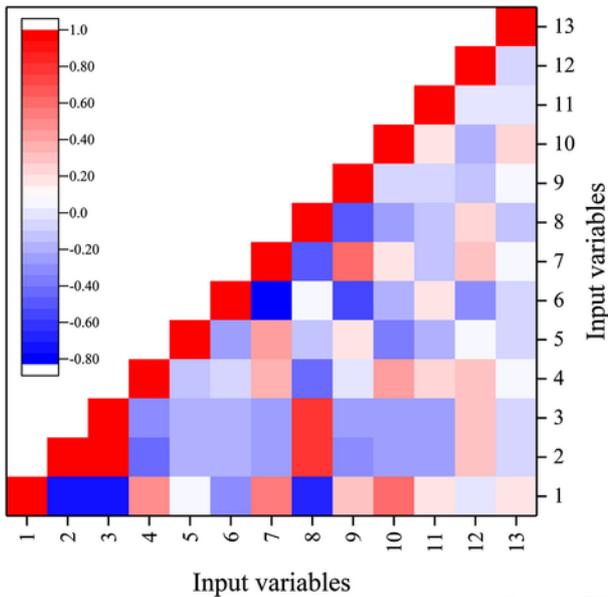


Fig. 5. Correlation coefficient plot of input variables. Serial number 1–13 corresponds to x<sub>1</sub>–x<sub>13</sub> in Table 2.

mine tailings, binder agent and mixing water, were homogeneously mixed using a concrete mixer (JJ-5, Hongda, Hebei). The mixed CPB materials were poured into bottom-drained plastic moulds (50 mm diameter and 100 mm height) and then placed in a curing box (YH-40B, Qingdao, Tianjin) at 25 °C and 90% humidity. Three replicates were conducted for each experimental scheme, leading to a total of 1545 CPB specimens. The authors note that ‘specimen’ was used to represent CPB specimens prepared for UCS tests while ‘sample’ was used to represent a pair of inputs and output in the dataset for GP modelling.

Based on the literature review, the main influencing variables were selected to be tailings type, cement-tailings ratio (CTR), solids content (SC) and curing time (T). As the tailings type is a nominal variable, namely a discrete data type with 11 categories, it was represented by its physical and chemical properties. The particle size distribution (PSD) was determined using a Malvern Mastersizer 2000 laser analyser and the X-ray diffraction (XRD) was measured using a Bruker SIMENS D500. The qualification of mineralogical compositions from XRD results was performed using the Rietveld method. The physical and chemical properties of tailings are provided in Table S2. An SHT-4206 SANS testing machine was used to perform UCS tests following the procedures in ASTM C39 (ASTM, 2001). Detailed description about the experimental procedure can be as easily found in our previous papers (Qi et al., 2018a).

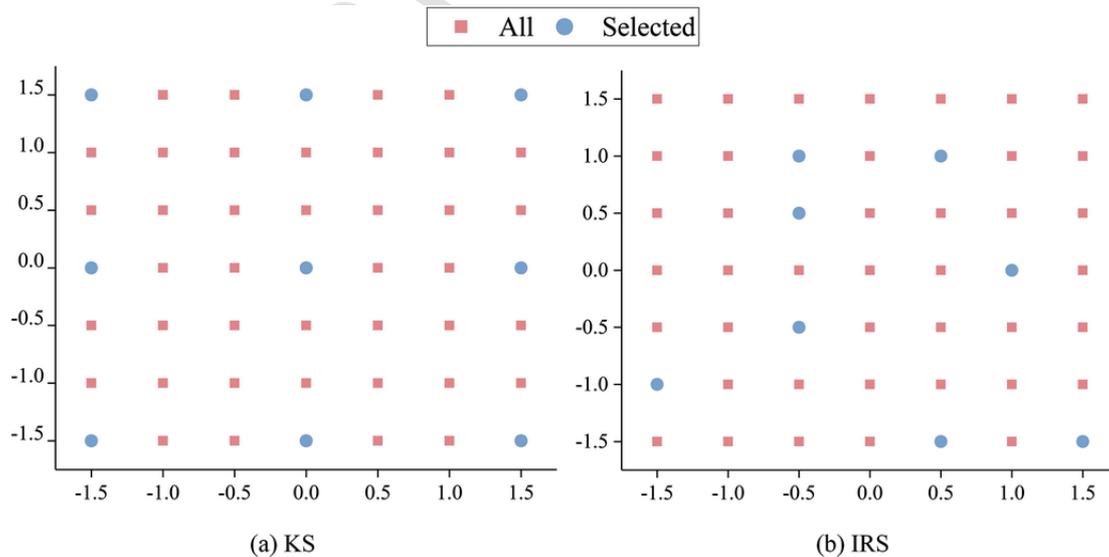


Fig. 6. Comparison of IRS and KS in a two-dimensional dataset.

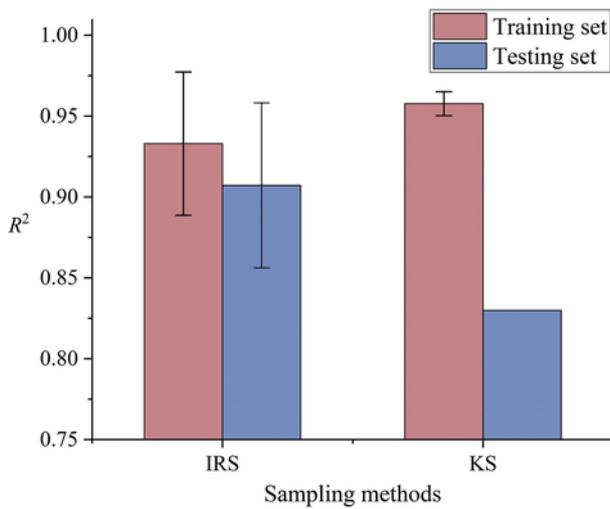


Fig. 7. The influence of IRS and KS on GP modelling performance.

2.2. Genetic programming

GP is one of the most popular data-driven algorithms that can be used to find the relationship between inputs and outputs automatically. It is proposed and modified by Cramer and Koza (Cramer, 1985; Koza, 1994) based on the inspiration of biological evolution and natural selection. In GP, each potential solution to the problem is represented by an ET consisting of branches and leaves. Fig. 3 depicts a sample ET with a simple mathematical equation:  $y = 4x + x$ , where  $y$  represents the output and  $x$  represents inputs. As shown in Fig. 3, the branches in an ET are represented by different functions, including arithmetic functions (i.e. +, -, and \*), trigonometric functions (i.e. sin and cos), exponential and logarithmic functions and special functions (i.e. integral). In contrast, terminals are used to represent leaves and the most widely used terminals are input variables and constant. Prior to the GP modelling, functions and terminals need to be pre-determined by the designer for each specific type of problem.

The GP modelling starts from an initial population generated randomly by grow, half (ramped-half-and-half) or full method (Koza, 1994). Then, the fitness of each ET is determined based on the designed fitness measure (i.e. coefficient of determination  $R^2$ ). Mimicking Darwinian evolution, the ETs with the best performances are selected

and used for the creation of the following generations (offspring). Crossover and mutation are two methods for the offspring generation using the ETs from the last generation (shown in Fig. 4). The generation-selection-generation process is iterated until the stopping criterion (such as the maximum number of generation) is reached.

The preceding introduction about the GP procedure suggests that there are several parameters for the GP modelling, including: (i) a function set and a terminal set; (ii) a fitness measure; (iii) GP parameters for the generation and evolution of ETs. In this paper, the arithmetic, trigonometric, exponential and logarithmic functions were included in the function set and the input variables and the constant were included in the terminal set. The remaining GP parameters are provided in Table 1, which were determined by trial tests and recommendations in the literature (Assimi et al., 2017; Khandelwal et al., 2017a).

2.3. Modelling procedure

As discussed in Section 2.1, 1545 UCS tests were conducted for the dataset preparation, leading to a total of 515 samples in the dataset. For each sample, the inputs included five variables for tailings physical properties (specific gravity and PSD), five variables for tailings chemical properties, CTR, SC and T while the output was the UCS value. Thus, there were 13 input variables and one output variable for each sample. The statistical description of the dataset is illustrated in Table 2. As shown, there was an evident difference in the data distribution, such as the data scope, for each variable. Towards this end, all inputs and output were normalised to (0, 1) based on their maximum and minimum values as follows:

$$x_i = \frac{x'_i - x_{min}'}{x_{max}' - x_{min}'} \tag{1}$$

$$y_i = \frac{y'_i - y_{min}'}{y_{max}' - y_{min}'} \tag{2}$$

where  $x_i$  and  $y_i$  represent normalised input and output values;  $x'_i$  and  $y'_i$  represent experimental input and output values;  $x_{min}'$ ,  $x_{max}'$ ,  $y_{min}'$ ,  $y_{max}'$  represent their corresponding minimum and maximum values.

Fig. 5 shows the correlation coefficient ( $R$ ) between input variables. As shown, most  $R$  values were less than 0.5, indicating there was a relatively weak correlation between most input variables (Koo and Li, 2016). A strong correlation was observed among  $x_1$ - $x_3$ , such as the  $R$  values between  $x_1$  and  $x_2$  (denoted as  $R_{x_1-x_2}$ ) and  $x_1$  and  $x_3$  (denoted as

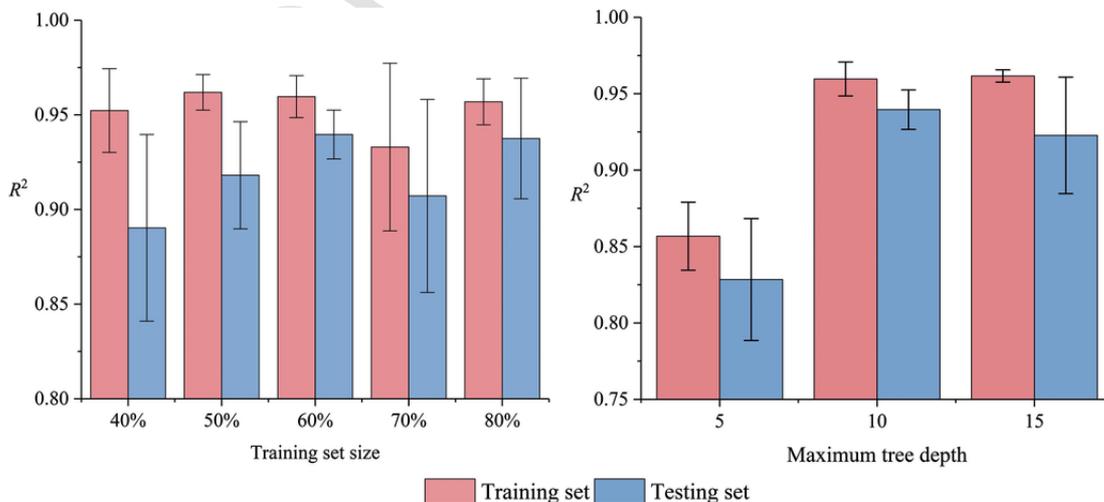


Fig. 8. The influence of training set size and maximum tree depth on GP modelling performance.

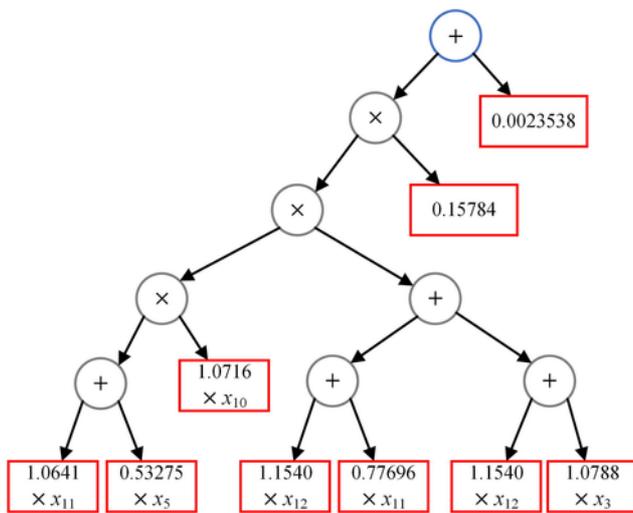


Fig. 9. Expression tree structure of the developed GP model.

$R_{x1-x3}$ ) were both  $-0.74$ . This result is straightforward as the specific gravity ( $G_s$ ) of tailings is strongly correlated to the PSD of tailings. Furthermore,  $R = -0.74$  indicates that  $G_s$  was negatively correlated with D10 and D50. The authors note that such a result could be an occasional relationship according to the utilised dataset. A better understanding of the relationship between  $G_s$  and PSD requires a comprehensive knowledge of particle interaction and movement, which will be a future topic. Strong correlations were also observed among several chemical properties, such as  $R_{x6-x7} = -0.83$  and  $R_{x7-x9} = 0.63$ . The  $x_{11}$ ,  $x_{12}$ , and  $x_{13}$  exhibited no strong correlations to any other input variables as their values were determined primarily by inspection considers limited information about tailings properties.

The entire dataset is divided into two subsets after the dataset normalisation, in which one is used for GP training (the training set) and the other is used for verification (the testing set). In this paper, two different dataset splitting methods were compared, namely the iterative random sampling (IRS) and the Kennard-Stone (KS) sampling. The IRS picks the data points randomly in the whole dataset and the dataset splitting is performed for several times to reduce the randomness. In contrast, the KS sampling is designed to select the data points evenly in the space of interest (Kennard and Stone, 1969). A comparison between IRS and KS in a two-dimensional dataset is shown in Fig. 6. In this paper, the comparison between IRS and KS was performed with a training set size of 70%.

A sensitivity study was conducted to investigate the influence of training set size (40–80% with a 10% interval) and maximum tree depth (5, 10 and 15) on the performance of GP modelling. A detailed performance evaluation was then conducted on one representative GP model, followed by a relative importance investigation of the input variables. The authors note that all dataset splitting using the IRS was performed ten times to reduce randomness. Similarly, all GP modelling using the same training and testing sets was performed three times as there is also randomness involved in GP modelling (i.e. initialisation and mutation).

#### 2.4. Statistical assessment of the results

To evaluate the reliability of the trained GP models, the following statistical descriptors were calculated between the predicted and experimental UCS values.

- Coefficient of determination ( $R^2$ ):  $R^2$  measures how well the output is predicted by the GP model using the proportion of total variation of output explained by the GP model (Draper and Smith, 2014).  $R^2$  is defined as:

$$R^2 = \left[ \frac{\sum_{i=1}^n (y_i^* - \bar{y})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (y_i^* - \bar{y}^*)^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \right]^2 \quad (3)$$

where  $N$  is the number of samples,  $y_i$  and  $y_i^*$  are the experimental and predicted UCS value of the  $i$ th sample, and  $\bar{y}$  is the average UCS value in the subset.

- Root-mean-square error (RMSE): RMSE is the standard deviation of the error between experimental and predicted UCS values. It is often used to find the unexpected large differences. RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - y_i^*)^2} \quad (4)$$

- Slope of regression lines ( $k$ ):  $k$  is calculated to be the slope of the linear regression equation fitted between experimental and predicted UCS values. A perfect prediction will produce a linear regression

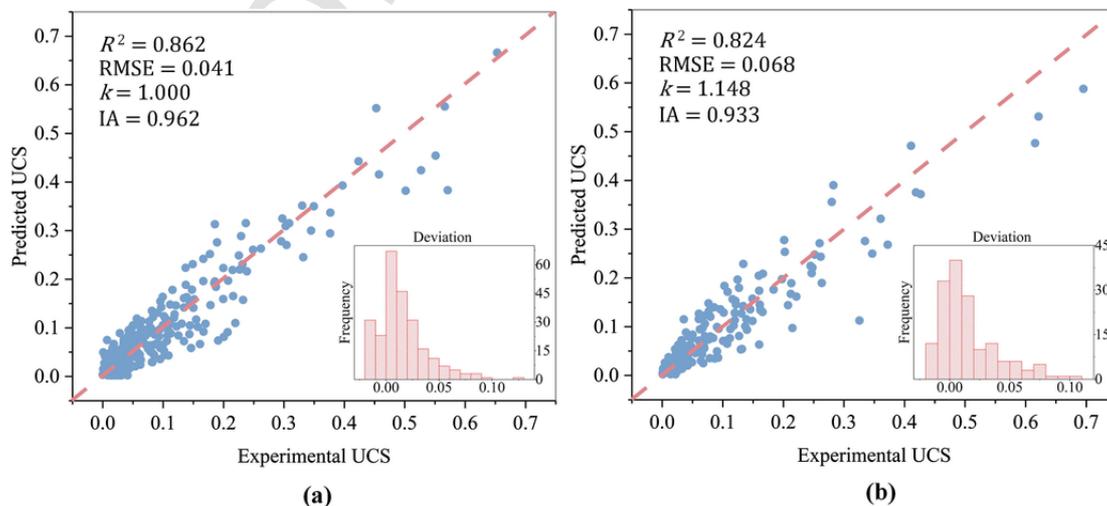


Fig. 10. Comparison of experimental and predicted UCS values: (a) Training set and (b) Testing set.

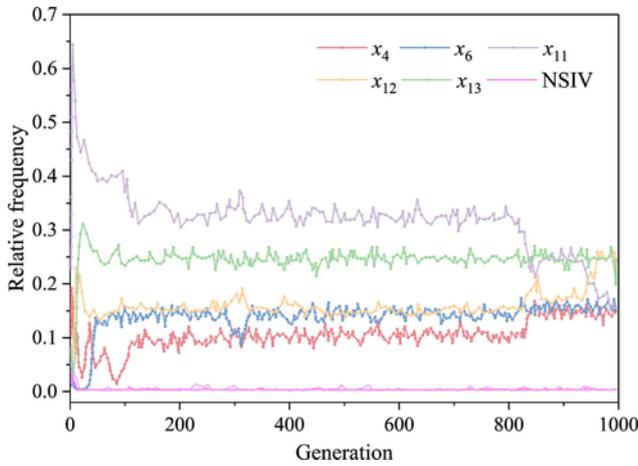


Fig. 11. Relative frequency of input variables with GP evolving generations.

with  $k = 1$  and intercept = 0. The calculation expression for  $k$  is given below:

$$k = \frac{\sum y_i y_i^*}{\sum (y_i^*)^2} \quad (5)$$

- Willmott's index of agreement (IA): IA is widely used to measure the degree of model prediction error. The IA value is ranged between [0, 1], in which 1 indicates a perfect prediction and 0 indicates an ineffective prediction (no better than randomly guessing). IA is defined as:

$$IA = 1 - \frac{\sum_{i=1}^N (y_i - y_i^*)^2}{\sum_{i=1}^N (|y_i^* - \bar{y}| + |y_i - \bar{y}|)^2} \quad (6)$$

### 3. Application and results

#### 3.1. Comparison of IRS and KS

Fig. 7 shows the influence of sampling method on the performance of GP modelling. The authors note here the comparison was conducted with a training set size of 70% as it is widely used in the literature (Qi et al., 2018e). It is found that the KS method has a higher average  $R^2$  value (0.96) on the training set compared to the IRS method (0.93). However, the average  $R^2$  value on the testing set obtained by the KS method was much lower than that obtained by the IRS method (0.83 compared to 0.91). Such a comparison indicates that the trained GP model with the KS method was prone to overfitting compared with the IRS method. This result agrees well with the findings in (Lee et al., 2018) and the reliability of the KS method might suffer from different data distribution in the training and testing sets. As indicated in (Golbraikh and Tropsha, 2002; Roy and Roy, 2008), a prediction can be considered as satisfactory if the  $R^2$  value between actual and predicted values is higher than 0.64 ( $R > 0.8$ ). Though the GP performance with the KS method was more stable compared to the IRS method (the KS method produced a smaller SD), the IRS method was employed in the following sections for its high generalisation capability.

#### 3.2. Influence of training set size and maximum tree depth

The results of the sensitivity study about the training set size and the maximum tree depth are illustrated in Fig. 8. As shown, there was no evident performance improvement on the training set with an increase in the training set size. The average  $R^2$  value on the training set was 0.95, 0.96, 0.96, 0.93 and 0.96, respectively, when the training set size was increased from 40% to 80%. However, the performance of GP modelling on the testing set was gradually improved stabilised when the training set size was increased from 40% to 60%. Above 60%, there was no clear trend along the increase of training set size. Therefore, 60% was selected to be the training set size in the following discussion.

The maximum tree depth is one of the most crucial GP parameters that control the complexity of each ET. In this paper, 5, 10 and 15

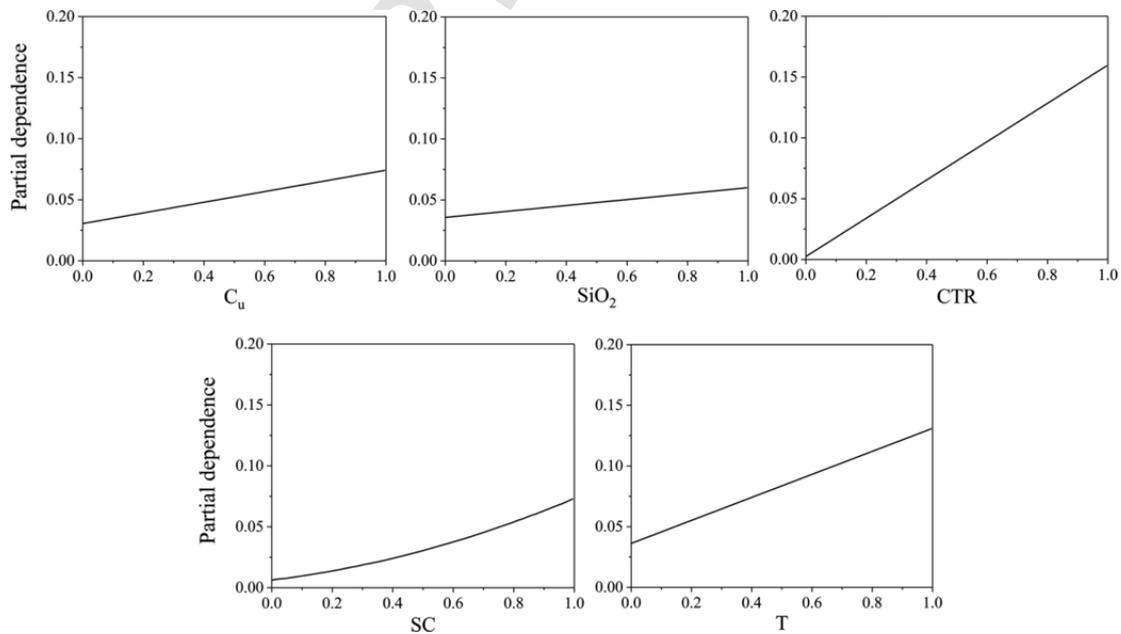


Fig. 12. Partial dependence plots of SIV from the RGP model.

**Table 3**  
Relative importance score and its SD calculated from GP and GBM modelling.

Variables	GP		GBM	
	Average	SD	Average	SD
$x_1$	0.005	0.0016	0.009	0.0021
$x_2$	0.004	0.0003	0.020	0.0078
$x_3$	0.004	0.0003	0.018	0.0051
$x_4$	0.030	0.0394	0.016	0.0046
$x_5$	0.036	0.0409	0.019	0.0038
$x_6$	0.087	0.0552	0.015	0.0021
$x_7$	0.012	0.0213	0.013	0.0013
$x_8$	0.004	0.0000	0.015	0.0023
$x_9$	0.054	0.0638	0.017	0.0016
$x_{10}$	0.005	0.0022	0.018	0.0039
$x_{11}$	0.330	0.0419	0.339	0.0147
$x_{12}$	0.231	0.0710	0.261	0.0242
$x_{13}$	0.200	0.0522	0.241	0.0215

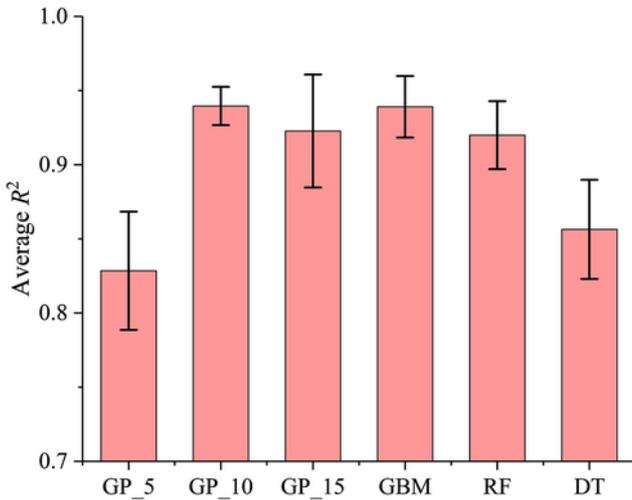


Fig. 13. Performance comparison between GP modelling and ML algorithms.

were selected to be the maximum tree depth to investigate its influence on the modelling performance. GP models with a 5, 10 and 15 maximum tree depth were referred to GP\_5, GP\_10 and GP\_15 in this paper. An evident performance improvement was observed when the maximum tree depth was increased from 5 to 10. To be more specific, the average  $R^2$  value on the training set was increased from 0.86 to 0.96 and the average  $R^2$  value on the testing set was increased from 0.83 to

0.94. Moreover, the SD values on the training and testing sets were 0.022 and 0.040, respectively, in the case of maximum tree depth = 5, which was decreased to 0.011 and 0.013 when the maximum depth was increased to 10. This result indicates the prediction became more stable when the maximum tree depth was increased to 10.

When the maximum tree depth was further increased from 10 to 15, a subtle increase (0.21%) was observed of the average  $R^2$  value on the training set. In contrast, the average  $R^2$  value on the testing set began to decline from 0.94 to 0.92. Further analysis shows that SD values decreased on the training set while increased on the testing set, implying the GP model was more stable on the training set while less stable on the testing set. Therefore, GP modelling became a little bit overfitting when the maximum tree depth exceeded 10.

In general, the results from Sections 3.1 and 3.2 suggest the GP modelling on the UCS dataset can achieve better performance with the IRS method, a training set size of 60% and a maximum tree depth of 10.

### 3.3. Performance evaluation of GP modelling

Having discussed the influence of sampling method, training set size and maximum tree depth on the performance of GP modelling, we now turn our attention to the detailed performance evaluation of the developed GP models. To make the analysis more straightforward, we used a GP\_5 model with a training set size of 60% as an example. It should be noted that 10 GP\_5 models have been constructed with the IRS method to reduce the randomness in Fig. 8. Here, we selected one representative GP\_5 (RGP\_5) model with a similar performance to the average values from 10 GP\_5 models. The RGP\_5 model had a 0.86  $R^2$  value on the training set and a 0.82  $R^2$  value on the testing set, compared with the average  $R^2$  value of 0.86 and 0.83 on the training and testing set, respectively. The analysis for the other GP\_5 models can be as easily conducted following the same procedure for the RGP\_5.

Fig. 9 shows the ET structure of the RGP\_5 model. As shown, only arithmetic functions (+, -, \* and /), input variables and constant were involved in the ET, resulting in good interpretability of the solution. The mathematical phrase of the ET structure in Fig. 9 is given as follows:

$$UCS = ((c_0 \times x_{12} + c_1 \times x_6) \times c_2 \times x_{11} \times ((c_3 \times x_{13} + c_4 \times x_{12}) + (c_5 \times x_{13} + c_6 \times x_4)) \times c_7 + c_8) \quad (7)$$

where  $c_0$ - $c_8$  were 1.0641, 0.5328, 1.0716, 1.1540, 0.7770, 1.1540, 1.0788, 0.1579 and 0.0024 respectively.

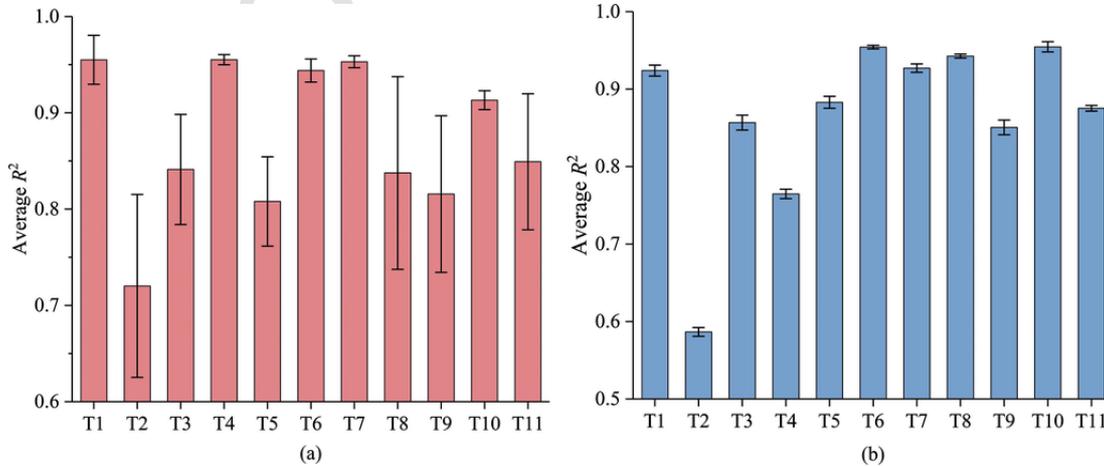


Fig. 14. Generalisation capability to a completely new tailings: (a) GP\_5 and (b) GBM.

Fig. 10 compares UCS values from experiments and the RGP\_5 model. As we can see, the RGP\_5 successfully learned the non-linear relationships between the UCS and its influencing variables. The predictive performance evaluated using  $R^2$ , RMSE,  $k$ , and IA was also shown in Fig. 10. In the case of the training set, the statistical parameters obtained by the RGP\_5 were:  $R^2 = 0.86$ ,  $RMSE = 0.04$ ,  $k = 1.00$ , and  $IA = 0.96$ . Based on the statistical recommendation, a good prediction can be evaluated with  $R^2 > 0.64$ ,  $0.85 < k < 1.15$ , or  $IA > 0.80$  (Golbraikh and Tropsha, 2002; Roy and Roy, 2008; Willmott, 1981). Therefore, the predictive performance of the RGP\_5 was quite satisfactory on the training set. Similarly, the statistical parameters on the testing set were:  $R^2 = 0.83$ ,  $RMSE = 0.07$ ,  $k = 1.15$ , and  $IA = 0.93$ , indicating a good performance on the testing set. The performance of the RGP\_5 on the training set was superior to that on the testing set, which was straightforward as the RGP\_5 was built using the training set.

The performance of the RGP\_5 model can be further demonstrated by the histogram plot of the deviation between the predicted and experimental UCS values (Fig. 10). As we can see, the peak frequency for both the training set and the testing set was around zero. This result indicates that the experimental and predicted UCS values were almost equal for most samples in the dataset. Another interesting finding is that the histogram was right-skewed for both the training and testing sets, implying the RGP\_5 model tended to predict slightly bigger UCS values, especially on several samples, than experimental values.

### 3.4. Relative importance of input variables

As indicated in (Qi et al., 2018d), exploring the relative importance of input variables could provide a better understanding of the UCS development of CPB, which might further suggest promising experimental studies. In this paper, the relative importance was investigated using the relative variable frequency and partial dependence plots from the RGP\_5 model. The relative variable frequency calculates/compares the appearance frequency associated with each variable during the GP evolution, which can provide us with the information about the relative importance of each variable (the more important a variable the more frequent it may appear in the GP evolving generations). Partial dependence plots are another way to investigate the dependence nature of the prediction on the input variables (Friedman, 2001). To obtain partial dependence plots, a number of values of the input variables are chosen in the first place. Then, the output is predicted using each of those values for all cases of other input variables. Finally, the average output is calculated, which is then plotted against its corresponding input value.

Fig. 11 illustrates the relative frequency of input variables during the evolution of the RGP\_5 model. In the following discussion, input variables not included in Eq. (7) were referred as non-significant input variables (NSIV) while input variables included in Eq. (7), namely  $x_4$ ,  $x_6$ , and  $x_{11} - x_{13}$ , were referred as significant input variables (SIV). As we can see, there was a quick adjustment for all input variables in the first 30 generations, implying the GP modelling could identify the UCS dependence on its influencing variables rapidly. For example, the relative frequency of NSIV has been decreased to zero or near zero at generation 30. Above results also demonstrate the efficiency of GP modelling in investigating the relative importance of input variables. The relative frequency was continuously adjusted, even though with small amplitudes, between generation 100 to generation 850. During this period, the relative frequency of input variables could be ranked in the following order:  $x_{11} > x_{13} > x_{12} \sim x_6 > x_4 > NSIV$ . Above generation 850, a relatively large adjustment was observed for  $x_4$ ,  $x_{11}$ , and  $x_{12}$  until generation 1000 and the frequency ranking has been adjusted to:  $x_{12} \sim x_{13} > x_4 \sim x_6 \sim x_{11} > NSIV$ .

As discussed before, partial dependence plots are calculated by measuring the output variation with the change of input values for

each variable. That means the change of NSIV values would not modify the predicted output as NSIV were not included in Eq. (7). Therefore, partial dependence was only plotted for SIV, as shown in Fig. 12. It is found that the UCS positively correlated with SIV. Furthermore, a perfect-positive linear relationship was observed between the UCS and  $x_4$ ,  $x_6$ ,  $x_{11}$ , and  $x_{13}$ . The UCS growth seems to be accelerated with the increasing of  $x_{12}$ . This phenomenon was also discovered by Qi et al. (2018d) using GBM modelling and they concluded that the growth of CPB strength was accelerated after the SC reached 70%. It should be noted here this accelerating phenomenon has not been confirmed by experiments, which would be a promising topic in the future. Based on partial dependence plots, the relative importance could be ranked as follows:  $x_{11} > x_{13} > x_{12} \sim x_6 > x_4 > NSIV$ , which agrees well with the relative frequency results.

The importance score was calculated based on the relative variable frequency from 10 GP\_5 models. As the optimum GP\_5 model was not obtained, at least in most cases, at the maximum generation, we took into account the relative frequency from all generations. Moreover, we employed GBM method using the same procedure as provided in (Qi et al., 2018d) so that the results from this study can be compared to the conclusions in the literature. In GBM modelling, the training set size was 60% selected by the IRS method and particle swarm optimisation (PSO) with an  $R^2$  fitness function was utilised for hyper-parameters tuning. Table 3 details the relative importance score from GP\_5 and GBM modelling. All relative importance scores have been scaled so that the sum of all 13 importance scores were one.

The GP\_5 and GBM both ranked  $x_{11}$ ,  $x_{12}$ , and  $x_{13}$  as the top-three influencing variables based on their importance scores ( $x_{11} = 0.33$ ,  $x_{12} = 0.23$  and  $x_{13} = 0.20$  from GP\_5 modelling while  $x_{11} = 0.34$ ,  $x_{12} = 0.26$  and  $x_{13} = 0.24$  from GBM modelling). The particular importance of  $x_{11}$ ,  $x_{12}$ , and  $x_{13}$  found in this paper agrees well with experimental conclusions in the literature (Fall et al., 2008; Yilmaz et al., 2014; Yin et al., 2012). An interesting finding is that the GP\_5 modelling ranked  $x_6$ ,  $x_9$ ,  $x_5$ , and  $x_4$  as the 4th–7th influencing variables with an importance score of 0.087, 0.054, 0.036 and 0.030, respectively, which were much higher than the other influencing variables in  $x_1 - x_{10}$ . However, the importance score of  $x_1 - x_{10}$  showed subtle difference from GBM modelling. A possible reason for this result would be the restricted complexity of GP modelling imposed by the maximum tree depth. In such case, the constructed GP\_5 model cannot take into account all influencing variables, resulting in a bias to several influencing variables during evolution. Another interesting finding is the SD value from GBM was generally smaller than the SD value from GP\_5, indicating the importance score from GBM was more stable than from GP\_5. This could also be due to the restricted complexity of GP\_5 modelling. Therefore, the investigation of relative importance score using GP with a small maximum tree depth should be conducted several times to get a representative result.

## 4. Discussion

In this section, we address two important questions that involved with GP modelling or ML algorithms in the UCS prediction. The first one is whether the performance of GP modelling is comparable to well-recognised ML techniques? The second, and probably the most important, one is that can the trained model be generalised/used to entirely new tailings? In this paper, an entirely new tailings is the one without any data samples from this tailings being used during model training.

To address the first question, we compared the predictive performance of the GP modelling with three ML techniques (DT, GBM, and RF) employed in the literature. We followed the same procedure in (Qi et al., 2018a) and used the same parameter setting for ML techniques as the GP modelling, such as the same training set size. The optimised hyper-parameters for three ML techniques are detailed in Table S3 and

Fig. 13 shows the comparison of average  $R^2$  values on the testing set between GP and ML techniques.

The average  $R^2$  values of GP\_5, GP\_10, GP\_15, GBM, RF, and DT were 0.83, 0.94, 0.92, 0.94, 0.92, and 0.86, respectively. This result indicates that the predictive performance of GP\_5 was relatively poor compared with that of GBM and RF, but was close to that of DT. In contrast, the GP\_10 and DP\_15 were as robust as the GBM and RF. In the case of the SD, it is found that the GP\_10 was the most stable model on the testing set with the smallest SD value of 0.013. Considering the average  $R^2$  values and the SD on the testing set, the investigated techniques can be ranked in the following order: GP\_10 > GBM > RF > GP\_15 > DT > GP\_5. Overall, the GP modelling has been verified to be comparable, or even better, than well-recognised ML techniques.

The second question is about the generalisation capability of trained GP or ML models to entirely new tailings. Though there has been a series of papers about the UCS prediction using ML techniques (Qi et al., 2018a; Qi et al., 2018b; Qi et al., 2018d), no attempts have been made to address this question. Certainly, this question is the key concern regarding each prediction model and its potential utility will suffer a lot if this question is not properly addressed. The authors believe this question is not only for CPB prediction but for any other predictions that involve different types of materials. For example, if a prediction model is aimed to be built for mechanical properties of concrete using different binder type, the generalisation of the trained model to the concrete with an entirely new binder needs to be clearly verified. Therefore, an essential question for such kind of prediction will be: Can the trained model be generalised to situations from where no samples have been used during the model training.

In this following, we turn our attention to the generalisation capability of the trained GP model to an entirely new tailings. To achieve this objective, we performed another 11 modelling scenarios using the GP\_5. In each scenario, one type of tailings from T1–T11 was selected and all its data samples were included in the testing set. The data samples in the training set came from all other types of tailings (except the selected one). In other words, we trained the GP model using ten types of tailings and tested its generalisation capability using the remaining one type of tailings in each scenario. The generalisation capability of GBM to an entirely new tailings was also investigated for comparison purposes.

Fig. 14 shows the average  $R^2$  values and the SD from each scenario. The authors note here T1–T11 in Fig. 14 represents which type of tailings was used in the testing set. For example, T1 means all data samples from T1 were used in the testing set and data samples from T2–T11 were used in the training set. It can be seen that the trained GP\_5 models could be well generalised to most tailings (except T2) with an average  $R^2$  value larger than 0.8. For some tailings, such as T1 and T4, the average  $R^2$  value reached up to 0.96, which were very satisfactory predictions in term of  $R^2$  values. Similar results were obtained for GBM modelling, which also shows a good generalisation capability to entirely new tailings. It needs to note that the generalisation capability of GBM modelling to T2 was far from satisfactory with an average  $R^2$  value of 0.59, implying the GP\_5 modelling might be more stable during the generalisation to entirely new tailings.

As a whole, above discussion shows that the performance of GP modelling was similar or even better compared with well-recognised ML techniques on CPB UCS prediction. The trained GP model could be generalised/used to entirely new tailings with satisfactory performance.

## 5. Limitations and outlooks

Though it has been shown that the UCS prediction using GP or ML techniques is quite promising, challenges still remain. First, the dataset

was collected from mine sites in China, resulting in a high possibility that the trained model cannot be generalised to CPB materials from other countries. This challenge is mainly due to the difference in cement classification and CPB preparation procedures. Another important challenge is about improving the accuracy and reliability of predictions, which involves delicate feature engineering, careful model selection and efficient algorithms proposing.

For the past 30 years, the UCS development of CPB has been investigated by experts from both industry and academia. The utilised methodology is still mainly experimental study and a large number of trial tests are needed for each CPB application. Instead, we anticipate that artificial intelligence techniques that can autonomously learn from the rich history of experiments will be crucial for efficient and intelligent CPB design in the near future. To realize this objective, a multinational, continuously-updated and highly-accessible database would be the 'Armstrong step'. The next concern would be how to incorporate such prediction in the whole backfill system and whether ML techniques can be applied to other aspects during CPB design. The final step would be the establishment of an 'Intelligent Mining for Backfill (IMB)' system by integrating all standalone applications together.

## 6. Conclusions

In this study, a genetic programming-based method was proposed for the UCS prediction of CPB. An enlarged dataset was collected using the results of 1545 UCS tests performed on 11 types of tailings. Two sampling methods, namely the IRS and the KS, were compared for dataset partitioning. A sensitivity study was performed to investigate the influence of training set size and maximum tree depth on the performance of GP modelling. We conducted a detailed analysis of the predictive performance of the RGP\_5 and investigated the relative variable importance using multiple methods. Moreover, we compared the predictive performance of GP modelling with well-recognised ML techniques and discussed the generalisation capability of such prediction to entirely new tailings. Based on the results of this study, the following conclusions can be drawn:

1. The IRS was found to be more suitable for UCS prediction as it produced a higher average  $R^2$  value on the testing set (0.91).
2. The best performance of the GP modelling was obtained using a training set size of 60% and a maximum tree depth of 10.
3. The statistical parameters obtained by the RGP\_5 were  $R^2 = 0.86$ , RMSE = 0.04,  $k = 1.00$ , and IA = 0.96 on the training set and  $R^2 = 0.83$ , RMSE = 0.07,  $k = 1.15$ , and IA = 0.93 on the testing set, indicating a satisfactory performance with the RGP\_5 model.
4. The GP\_5 modelling ranked CTR, SC and T as the top-three influencing variables with their corresponding importance scores of 0.33, 0.23 and 0.20 respectively.
5. The predictive performance of GP modelling was comparable to well-recognised ML techniques and the trained GP model could be generalised/used to entirely new tailings with satisfactory performance.

## Acknowledgement

This study was financially supported by the 12th Five Years Key Programs for Science and Technology Development of China (No. 2012BAC09B02) and the Fundamental Research Funds for the Central Universities of Central South University (No. 2016zzts092). The first and second authors would also thank China Scholarship Council for the financial support during Ph.D. studies (grant number: 201606420046 & 201706420058).

## Conflicts of interest

None.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.mineng.2019.01.004>.

## References

- Assimi, H., Jamali, A., Nariman-Zadeh, N., 2017. Sizing and topology optimization of truss structures using genetic programming. *Swarm Evol. Comput.* 37, 90–103.
- Astm, C., 2001. Standard test method for compressive strength of cylindrical concrete specimens. ASTM Int.
- Chen, Q., Zhang, Q., Qi, C., Fourie, A., Xiao, C., 2018. Recycling phosphogypsum and construction demolition waste for cemented paste backfill and its environmental impact. *J. Cleaner Prod.* 186, 418–429.
- Cihangir, F., Ercikdi, B., Kesimal, A., Devenci, H., Erdemir, F., 2015. Paste backfill of high-sulphide mill tailings using alkali-activated blast furnace slag: effect of activator nature, concentration and slag properties. *Miner. Eng.* 83, 117–127.
- Cramer, N.L., 1985. A representation for the adaptive generation of simple sequential programs. In: *Proceedings of the First International Conference on Genetic Algorithms*. pp. 183–187.
- Draper, N.R., Smith, H., 2014. *Applied Regression Analysis*. John Wiley & Sons.
- Fall, M., Benzaazoua, M., Saa, E.G., 2008. Mix proportioning of underground cemented tailings backfill. *Tunn. Undergr. Space Technol.* 23, 80–90.
- Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 1189–1232.
- Golbraikh, A., Tropsha, A., 2002. Beware of  $q^2$ !. *J. Mol. Graph. Model.* 20, 269–276.
- Hoseinian, F.S., Faradonbeh, R.S., Abdollahzadeh, A., Rezai, B., Soltani-Mohammadi, S., 2017. Semi-autogenous mill power model development using gene expression programming. *Powder Technol.* 308, 61–69.
- Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. *Technometrics* 11, 137–148.
- Kesimal, A., Ercikdi, B., Yilmaz, E., 2003. The effect of desliming by sedimentation on paste backfill performance. *Miner. Eng.* 16, 1009–1011.
- Khandelwal, M., Faradonbeh, R.S., Monjezi, M., Armaghani, D.J., Majid, M.Z.B.A., Yagiz, S., 2017. Function development for appraising brittleness of intact rocks using genetic programming and non-linear multiple regression models. *Eng. Comput.* 33, 13–21.
- Khandelwal, M., Shirani Faradonbeh, R., Monjezi, M., Armaghani, D.J., Majid, M.Z.B.A., Yagiz, S., 2017. Function development for appraising brittleness of intact rocks using genetic programming and non-linear multiple regression models. *Eng. Comput.* 33, 13–21.
- Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropractic Med.* 15, 155–163.
- Koza, J.R., 1994. Genetic programming as a means for programming computers by natural selection. *Statist. Comput.* 4, 87–112.
- Lee, L.C., Liong, C.-Y., Jemain, A.A., 2018. Iterative random vs. Kennard-Stone sampling for IR spectrum-based classification task using PLS2-DA. *AIP Conference Proceedings*. AIP Publishing.
- Liu, L., Fang, Z., Qi, C., Zhang, B., Guo, L., Song, K.-I., 2018. Experimental investigation on the relationship between pore characteristics and unconfined compressive strength of cemented paste backfill. *Constr. Build. Mater.* 179, 254–264.
- Liu, L., Fang, Z., Qi, C., Zhang, B., Guo, L., Song, K.I.I.L., 2019. Numerical study on the pipe flow characteristics of the cemented paste backfill slurry considering hydration effects. *Powder Technol.* 343, 454–464.
- Lu, H., Qi, C., Chen, Q., Gan, D., Xue, Z., Hu, Y., 2018. A new procedure for recycling waste tailings as cemented paste backfill to underground stopes and open pits. *J. Cleaner Prod.* 188, 601–612.
- Mangane, M.B.C., Argane, R., Trauchessec, R., Lecomte, A., Benzaazoua, M., 2018. Influence of superplasticizers on mechanical properties and workability of cemented paste backfill. *Miner. Eng.* 116, 3–14.
- Nazari, A., 2012. Experimental study and computer-aided prediction of percentage of water absorption of geopolymers produced by waste fly ash and rice husk bark ash. *Int. J. Miner. Process.* 110–111, 74–81.
- Orejarena, L., Fall, M., 2010. Artificial neural network based modeling of the coupled effect of sulphate and temperature on the strength of cemented paste backfill. *Can. J. Civ. Eng.* 38, 100–109.
- Orejarena, L., Fall, M., 2010. The use of artificial neural networks to predict the effect of sulphate attack on the strength of cemented paste backfill. *Bull. Eng. Geol. Environ.* 69, 659–670.
- Qi, C., Chen, Q., Fourie, A., Tang, X., Zhang, Q., Dong, X., Feng, Y., 2019. Constitutive modelling of cemented paste backfill: a data-mining approach. *Constr. Build. Mater.* 197, 262–270.
- Qi, C., Chen, Q., Fourie, A., Zhang, Q., 2018. An intelligent modelling framework for mechanical properties of cemented paste backfill. *Miner. Eng.* 123, 16–27.
- Qi, C., Fourie, A., Chen, Q., 2018. Neural network and particle swarm optimization for predicting the unconfined compressive strength of cemented paste backfill. *Constr. Build. Mater.* 159, 473–478.
- Qi, C., Fourie, A., Chen, Q., Tang, X., Zhang, Q., Gao, R., 2018. Data-driven modelling of the flocculation process on mineral processing tailings treatment. *J. Cleaner Prod.* 196, 505–516.
- Qi, C., Fourie, A., Chen, Q., Zhang, Q., 2018. A strength prediction model using artificial intelligence for recycling waste tailings as cemented paste backfill. *J. Cleaner Prod.* 183, 566–578.
- Qi, C., Fourie, A., Ma, G., Tang, X., 2018. A hybrid method for improved stability prediction in construction projects: a case study of stope hangingwall stability. *Appl. Soft Comput.* 71, 649–658.
- Ross, B.J., Gualtieri, A.G., Fueten, F., Budkewitsch, P., 2005. Hyperspectral image analysis using genetic programming. *Appl. Soft Comput.* 5, 147–156.
- Roy, P.P., Roy, K., 2008. On some aspects of variable selection for partial least squares regression models. *QSAR Comb. Sci.* 27, 302–313.
- Willmott, C.J., 1981. On the validation of models. *Phys. Geogr.* 2, 184–194.
- Yilmaz, E., Benzaazoua, M., Belem, T., Bussière, B., 2009. Effect of curing under pressure on compressive strength development of cemented paste backfill. *Miner. Eng.* 22, 772–785.
- Yilmaz, E., Fall, M., 2017. *Paste Tailings Management*. Springer.
- Yilmaz, T., Ercikdi, B., Karaman, K., Külekçi, G., 2014. Assessment of strength properties of cemented paste backfill by ultrasonic pulse velocity test. *Ultrasonics* 54, 1386–1394.
- Yin, S., Wu, A., Hu, K., Wang, Y., Zhang, Y., 2012. The effect of solid components on the rheological and mechanical properties of cemented paste backfill. *Miner. Eng.* 35, 61–66.