A Multi-Objective Evolutionary Framework for Formulation of Nonlinear Structural Systems

Amir H. Gandomi, *Senior Member*, *IEEE*, David A. Roke

***Abstract*— In this study, an evolutionary framework is proposed for seismic response formulation of self-centering concentrically braced frame (SC-CBF) systems. A total of 75 different SC-CBF systems were designed, and their responses were recorded under 170 earthquake records. To select the most important earthquake intensity measures, an evolutionary feature selection strategy is introduced, which tries to find the highest correlation. For the formulation of the SC-CBF response, a hybrid multi-objective genetic programming and regression analysis is implemented, considering both model accuracy and model complexity as objectives. In the hybrid approach, regression tries to connect multiple genes. Non-dominated models are presented, and the best model is selected based on the practical approach proposed here. The best model is compared with four other genetic programming models. The results show that the evolutionary procedure is highly effective for designing the SC-CBF system using a simple and accurate model for such a complex system.**

***Index Terms*— Evolutionary Computation; Genetic Programming; Feature Selection; Formulation; Self-centering concentrically braced frame; Multi-objective**

I. Introduction

EMPIRICAL data mining and formulation by artificial intelligence (AI) methods remain a highly-researched topic, particularly for modeling practical and engineering problems. Models derived by AI differ from conventional models, which are based on engineering principles like elasticity and plasticity theories. Instead of theoretical derivations, AI-based models are mainly derived from available experimental data. Although several AI modelling tools have been proposed in the literature, most of them, such as artificial neural networks and fuzzy logic, are considered gray/black box models because of their complexity [1] and lack of transparency. Therefore, these techniques are usually implemented as a part of a program, and the explicit model is not presented. In this case, not only does the applicability of the model decrease, but accessibility becomes an issue unless the program itself is made available. Due to such complexity, gray/black-box models also have more overfitting potential in comparison with simple mathematical models. In practice, a simple and white-box model is most desirable. For problems involving inference, the main goal is to understand the relationship between the predictors and response variable, and therefore, interpretability is more important than accuracy.

In practice, a simple white-box model is most desirable. Genetic programming (GP) is one of the robust AI techniques for developing data-driven models, which can develop simple explicit models [2]. GP is inspired by the principle of Darwinian natural selection as proposed by Koza [3]. GP creates a machine code that can be directly translated into a mathematical formula, making it suitable for building explicit models for nonlinear system modeling. In other words, GP creates programs to build data-driven models that can translate to prediction equations. Additionally, GP has th

e flexibility to create models without any prior structures or assumptions, which is its main advantage over other data mining tools.

Classical GP has a tree structure and, thus, is sometimes called tree-based GP. Several alternative GP-based approaches have been suggested, such as linear GP [4] and multi-stage GP [5]. Any GP can be hybridized with other data-driven approaches. For instance, Searson *et al.* [6] and Arnaldo *et al.* [7] hybridized GP with regression analysis (RA) to boost the performance of GP. GP and its variants have been successfully applied to several challenging problems [8] and have demonstrated to be a reliable tool for complex engineering modelling [9].

The complexity of GP models must be controlled, otherwise GP tends to build overly complex models during generation, which is known as bloat [10]. When the number of inputs is low and the problem landscape is not sophisticated, the complexity of GP models can be controlled. However, model complexity can become an issue for a high-dimensional problem or a problem with a fluctuated landscape. Therefore, the best models should be selected by considering both model accuracy and complexity. In other words, multi-objective optimization is necessary to find the best models in terms of complexity (in general terms) and performance [11].

Steel concentrically-braced frame (CBF) systems have been widely used as seismic lateral force-resisting systems in the United States and around the world. A recently-developed earthquake-resistant structural system, called the self-centering CBF (SC-CBF) system, was proposed by Roke *et al.* [12] to

Manuscript received February 24, 2021; revised October 10, 2021; accepted October 28, 2021. This work was supported in part by the Australian Government through the Australian Research Council (project number DE210101808). Paper no. TII-21-0835 (*Corresponding author: Amir H Gandomi*).

A.H. Gandomi is with the Faculty of Engineering & IT at University of Technology Sydney, Australia (e-mail: Gandomi@uts.edu.au).

D.A. Roke is with the Department of Civil Engineering at the University of Akron, OH, USA. (e-mail: Roke@uakron.edu)

Digital Object Identifier 10.1109/TII.2021.3126702

eliminate the residual drift and related damage in the CBF system. This system has some special elements that permit rocking behavior under lateral loads, such as those generated by an earthquake. Since earthquakes have particularly stochastic behavior, researchers have proposed many intensity measures (IMs) for characterizing them. Therefore, modelling the SC-CBF system under earthquake excitation is a very complex and nonlinear problem. Finding the best IMs to be considered in the prediction model is another challenging step of this problem, as various IMs have been used for similar purposes in the literature.

The objectives of this study were to formulate and predict the maximum response of SC-CBF systems of a specific earthquake record. For this challenging engineering problem, the goal is to develop an explicit and interpretable model that is applicable in practice and for designing such systems. The highly nonlinear nature of the structural response, combined with the stochastic nature of earthquake motion, requires a huge amount of data related to the system's behaviour to establish a substantially accurate model. In order to create a large database, 75 SC-CBF systems were designed and subjected to 170 earthquakes, for which several different IMs were evaluated for each simulation. An evolutionary approach was proposed to first select the best IM. This approach utilizes modified correlation coefficients in order to capture the nonlinear correlation between two pairs of data (e.g., an IM and the peak structural response),which is used to rank IMs as variables. Second, a multi-objective GP (MOGP) method was used to formulate this problem based on the mechanical and geometric parameters of the structures, as well as the selected IMs. The MOGP strategy consists of multiple genes that are combined using a linear RA.

The rest of the paper is organized as follows. In the next section, the problem is defined along with a prototype structure, and the process of seismic analysis of SC-CBF systems is expressed. The evolutionary approaches used for feature selection and modelling are explained in section III. Section IV discusses the results of feature selection using the proposed approach, modelling using the MOGP, the non-dominated models (Pareto set), and selection of the best model. Comparison and validation of the results are also provided in the latter section. In section V, a summary and conclusions are presented.

The objectives of this study are listed below:

- Develop a comprehensive analysis of a new structural system in the construction industry, and create a valuable dataset by applying 170 real earthquake records to 75 different designed SC-CBF systems for explicit formulation of the responses.
- Propose an evolutionary correlation coefficient to select the best features to account for Pearson's correlation coefficient, which only measures linear correlation.
- Introduce MOGP in order to find a simple yet accurate model for this complex problem – the explicit model provided can be used in the SC-CBF design procedure.
- Propose a new approach for genetic programming model selection to achieve both accuracy and complexity.

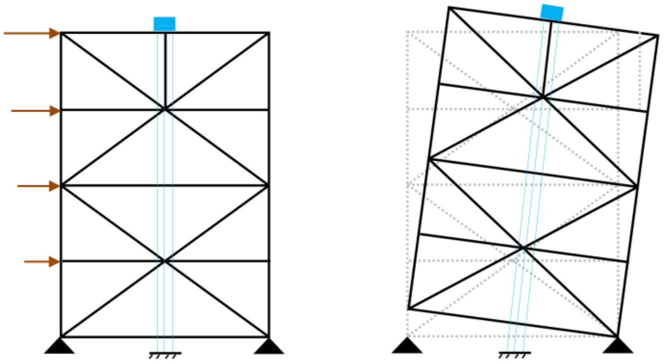


Fig. 2. Schematic of an SC-CBF system behavior under high level of lateral load.

II. Seismic Analysis of SC-CBF Systems

A. SC-CBF Systems

The structure of study is a self-centering concentrically braced (SC-CBF) system, which was originally proposed by Roke *et al.* [12]. Figure 1 shows a schematic representation of an SC-CBF system and its special elements, such as post-tensioning (PT) bars and lateral-load bearings. As shown in this figure, the SC-CBF column is separate from the main gravity columns of the structure, and the two types of columns interact through friction-based lateral-load bearing elements that act as energy dissipation elements.

Figure 2 illustrates this structure under lateral loads and its potential rocking behavior under a high level of lateral loads. One distinctive element in this system is the post-tensioning bars that connect the roof level to the ground. The PT bars are meant to limit the response in the first mode of vibration by acting as a fuse. More detail regarding this system can be found in other publications (e.g., [13]).

B. Prototype Structure

The prototype structure of this study is an office building located in Los Angeles, CA that was designed for stiff soil, as studied by Gandomi [14]. The typical floor plan of the prototype building is shown in Figure 3, where the studied SC-CBF system and its related tributary area are highlighted grey.

Two mechanical properties are considered as variables: the yield stress of structural members (F_y) and the coefficient of

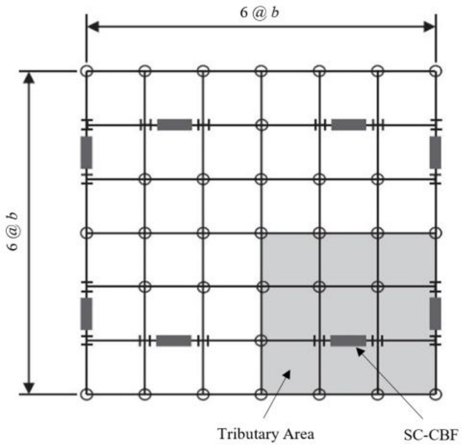


Fig. 3. Typical floor plan of the prototype building.

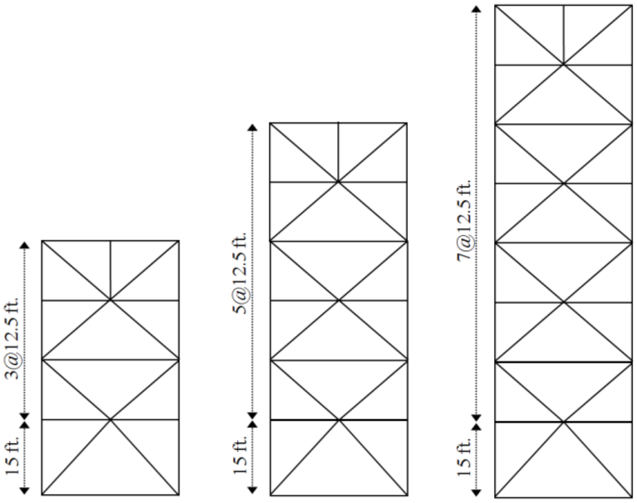


Fig. 4. Three different elevations of SC-CBF systems.

friction at the lateral-load bearings (μ). The height of the structure (h) is one of the main variables in this problem. Therefore, three different building heights (corresponding to 4-, 6-, and 8-story structures) have been considered (as shown in Figure 4). The SC-CBF's aspect ratio (calculated as h/b) is used as the geometric predictor in this study. Three different values have been considered for each mechanical and geometric variable, as presented in Table I.

In addition to the structural design parameters, earthquake intensities are important to predict structural responses to an individual earthquake. Therefore, several normalized intensity measures (IMs) are considered as candidates for the final model. Thirteen established IMs were considered, as listed in [15].

TABLE I
SUMMARY OF DESIGN VALUES

Geometrical		Mechanical	
***b*, ft (m)**	***h*, ft (m)**	***F_y*, ksi (MPa)**	***μ***
22.5 (6.9)	52.5 (16)	36 (248)	0.30
30 (9.1)	77.5 (23.6)	50 (345)	0.45
40 (12.2)	102.5 (31.2)	60 (414)	0.60

This study intended to use structural design parameters (mechanical and geometric properties of the structures) and earthquake intensity measures (IMs) to predict the peak roof drift of the system, as formulated below:

$$\theta = f(h/b,\ h,\ F_y,,\ \mu,\ IMs) \qquad (1)$$

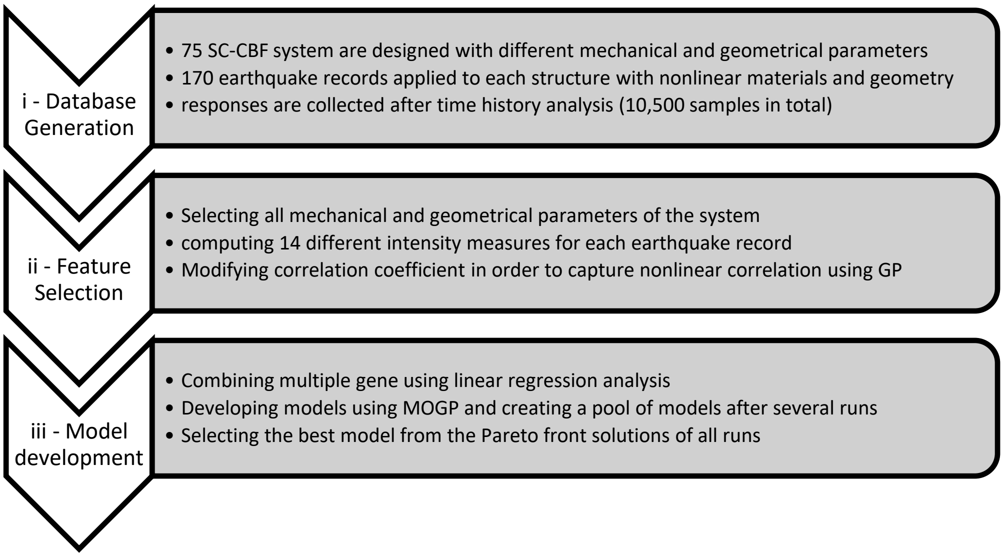
The complexity of the system's response, due to the nonlinear behavior and rocking of the SC-CBF, makes predicting the peak response difficult.

For modeling this system, GP was used to formulate the maximum response prediction as a function of structural parameters and earthquake intensity measures. Nonlinear materials and linear geometry were employed for the finite element model of the structures, and the design and analysis processes were automated [14] using MATLAB®. For the nonlinear dynamic finite element analysis, MATLAB® calls the OpenSEES software package [16].

III. Methodology

Evolutionary computation (EC) techniques are a subset of AI that slightly differ from the classical AI methods since their intelligence is mimicked from biological systems or nature in general. Bio-inspired/evolutionary computation techniques are those in which "the computational algorithms model natural phenomena" [17]. The efficiency of EC methods is owed to their exceptional ability to imitate the best features of nature, which have evolved by natural selection over millions of years. In recent years, EC methods have been widely used and have remained a highly-researched topic, particularly for complex engineering and real-world problems.

Evolutionary algorithms, like genetic algorithm, evolutionary strategy, particle swarm optimization, and differential evolution, are mostly known for their parameter optimization ability when each solution is vector of numbers.



i - Database Generation

- 75 SC-CBF system are designed with different mechanical and geometrical parameters
- 170 earthquake records applied to each structure with nonlinear materials and geometry
- responses are collected after time history analysis (10,500 samples in total)

ii - Feature Selection

- Selecting all mechanical and geometrical parameters of the system
- computing 14 different intensity measures for each earthquake record
- Modifying correlation coefficient in order to capture nonlinear correlation using GP

iii - Model development

- Combining multiple gene using linear regression analysis
- Developing models using MOGP and creating a pool of models after several runs
- Selecting the best model from the Pareto front solutions of all runs

Fig. 5. Three different elevations of SC-CBF systems.

Unlike traditional evolutionary algorithms, each genetic programming (GP) solution is a computer program that can be a combination of numbers, variables, and mathematical operators. Each program can be directly translated into a mathematical formula, which makes it suitable for building explicit models for nonlinear system modeling.

In this section, the GP-based approach proposed in this study is explained in detail. As outlined in Figure 5, the general process of the approach consists of three phases. In phase i, a database is generated, as described in section II. Phase ii includes selection of structure and earthquakes features, which were discussed in section II. In order to select the best earthquake IMs to model the SC-CBF response from the IMs established in the literature, the correlation coefficient is modified using GP, generating an evolutionary correlation coefficient. GP is explained in section III.A, and the proposed evolutionary correlation coefficient is formulated in section III.B. In phase iii, the MOGP is applied multiple times, and the Pareto front sets are extracted from a pool of all models. The MOGP is explained in sections III.C. The final model is selected based on the simple multi-objective strategy proposed in section III.D.

A. Genetic programming

GP is an evolutionary-based approach that was first proposed by Koza (1992). GP is a more recent specialization of genetic algorithms (GAs), which follow Darwinian principles to find solutions/models. GP is applied to develop computer programs as models, instead of binary strings from GAs. However, most evolutionary mechanisms that are commonly implemented in GAs can also be used in GP. GP creates programs as data-driven models, which can be translated into prediction equations. The main advantage of the GP approach over other data modelling tools is its ability to create models without any prior assumption. GP solutions can be represented as tree structures and declared in a functional programming language, whereas GA results are binary strings. Therefore, the GP solution (program) is a variable-length parse tree rather than a fixed-length binary string. The traditional GP solution, also known as tree-based GP, has a hierarchically structured tree, consisting of root, terminal, and function nodes [14].

The tree structure starts with a root point, extends to branches, and ends in one or more terminal nodes. In GP, the functional nodes are selected from a predefined set of functions. For instance, the functions can contain Boolean logic (e.g., NOT, AND, OR), basic arithmetic (+, −, ×, ÷), or other mathematical functions, such as trigonometric, logarithmic, and/or exponential functions. The terminal nodes can contain numerical constants, variables, or logical constants. The terminals and functions are randomly chosen and combined to form a tree-like structure [14].

Once an initial population is created, the GP algorithm evaluates the individuals' fitness. Then, like any other EC technique, GP employs evolutionary operators (e.g., crossover, mutation, and selection) to evolve trees/programs. For crossover, a point on a branch of each parent tree is randomly selected, then the set of functions and terminals from each parent is swapped to create two children (two new programs). Occasionally, mutation occurs based on a defined probability and randomly replaces a function or terminal from a tree (program/solution). The evolutionary process of GP is an iterative process that evolves the programs during generations (iterations). Therefore, GP continues by evaluating the new population and applying another iteration of evolutionary mechanisms.

B. Evolutionary Coefficient

When there are too many independent variables, a variable/feature selection method is necessary to select the most effective variables and reduce the dimensionality of data. Several methods, such as sequential feature selection, have been published in the literature for this purpose. The easiest and most popular feature selection method is to select the variables based on correlation to the output. In this study, a new method is proposed to evaluate the correlations between variables and output.

The well-known Pearson correlation coefficient (*R*) only shows linear relationships between parameters and has the following formula:

$$R=\frac{\sum_{i=1}^{n}\left(y_i-\overline{y_i}\right)\left(t_i-\overline{t_i}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i-\overline{y_i}\right)^2\sum_{i=1}^{n}\left(t_i-\overline{t_i}\right)^2}} \tag{2}$$

where t_i and y_i are the i^{th} predicted and actual outputs, respectively; variables with bars above them are the average values of those quantities; and *n* is the number of records.

However, two parameters can have nonlinear relationships that cannot be captured by *R* or R^2. Evolutionary computations can be used to find nonlinear relationships between two variables; therefore, an evolutionary algorithm is proposed herein to find the highest possible correlation between two parameters. A new coefficient, called the evolutionary coefficient (R_e^2), is proposed using GP. For the j^{th} variable (x_j), R_e^2 is equal to the highest R^2 value found after developing a GP model using R^2 (or $|R|$) as the fitness function. $\mathrm{R_e}$ can be mathematically expressed as:

$$R_e=\frac{\sum_{i=1}^{n}\left(y_i-\overline{y_i}\right)\left(f\left(x_j\right)_i-\overline{f\left(x_j\right)_i}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i-\overline{y_i}\right)^2\sum_{i=1}^{n}\left(f\left(x_j\right)_i-\overline{f\left(x_j\right)_i}\right)^2}} \tag{3}$$

where $f\left(x_j\right)_i$ is the i^{th} correlated output of the function for x_j; and $\overline{f\left(x_j\right)_i}$ is its average value.

For parameter selection, Ratner [18] suggested using GP functions as variables to determine the most effective parameters. Ratner also tried to predict the response using a combination of the original variables and the best GP models as the new variables. In a similar study, Ratner also proposed a new model called GenIQ as an alternative to the statistical ordinary least squares and logistic regression models [19]. Therefore, the EC proposed here could be considered as a special GenIQ model, where the correlation coefficient is the main objective.

C. Multi-Gene Genetic Programming

Multi-gene genetic programming (MGGP) is a robust variant of GP proposed by Searson *et al.* [6]. This algorithm has been successfully applied to several real-world problems. Initially, Gandomi and Alavi [20] introduced this algorithm to solve complex civil and mechanical engineering problems. Garg *et al.* [21] used this approach for material modeling, achieving competitive results with well-known algorithms. Muduli and Das [22] applied MGGP for seismic soil liquefaction potential evaluation.

Symbolic regression is typically carried out through traditional GP for the evolution of a population of programs (trees). The evolved programs predict an N × 1 vector of output (y) by using the $N \times M$ matrix of inputs (X), where N is the number of samples and M is the number of variables [6].

Each MGGP model is a weighted linear combination of sub-tress. Each tree includes multiple sub-trees, and each sub-tree can be considered a “gene.” Figure 6 displays a typical model of multi-gene expression, which predicts an output variable using three input variables (A, B, and C). This model is linear with respect to the coefficients d_0, d_1, and d_2, although the structure may contain nonlinear terms (e.g., the cosine function). In practice, the maximum number of genes (G_{max}) and sub-tree depth (D_{max}) should be user-specified, allowing a high degree of control over the complexity of the tree-based models. It should be noted that enforcing rigid sub-tree depth restrictions will result in the evolution of relatively compact models. The evolved models are linear combinations of low-order nonlinear transformations of the predictor variables [6].

In the MGGP method, the linear coefficients are derived for each model using the ordinary least squares method from the training data. MGGP has shown to have higher accuracy and computational efficiency for symbolic regression than standard GP [6, 11].

By creating individuals in MGGP with different quantities of genes (1 to G_{max}), the initial population is constructed. In addition to the traditional GP recombination operators, genes are acquired and deleted during the MGGP run using a tree crossover operator called the two-point high-level crossover, which allows the exchange of genes between individuals. For example, if the first parent individual consists of three genes (G_1, G_2, and G_3) and the second one is composed of four genes (G_4, G_5, G_6, and G_7), two randomly created crossover points are selected for each individual. Square brackets denote the genes within the crossover points in Eq. 4:

$$(G_1\ [\boldsymbol{G_2}]\ G_3)\ (G_4\ G_5\ [\boldsymbol{G_6\ G_7}]) \tag{4}$$

Then, the genes within the crossover points are exchanged, resulting in the creation of new individuals, as follows:

$$(G_1\ [\boldsymbol{G_6\ G_7}]\ G_3)\ (G_4\ G_5\ [\boldsymbol{G_2}]) \tag{5}$$

The two-point high-level crossover operation allows the acquisition of new genes for both individuals. Gene removal is also permitted, and genes are randomly deleted if an exchange of genes results in an individual containing more genes than G_{max}.

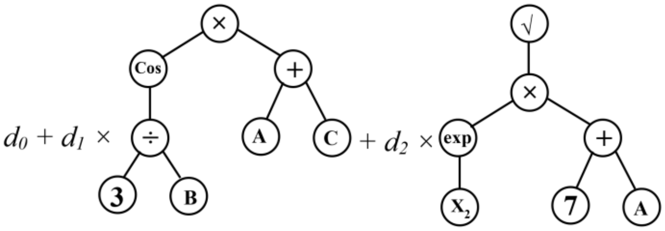
In this algorithm, a standard GP sub-tree crossover is referred to as a low-level crossover, wherein a sub-tree is chosen at random from each parent individual. After that, the standard sub-tree crossover is applied, in which the created sub-trees replace the parent sub-trees in the next generation of the otherwise unaltered individual.

The other main evolutionary mechanism is mutation, which is a random process. There are six different mutation schemes in MGGP, including sub-tree mutation, mutation of constants, substitution of an input node with another input node, setting a constant to zero, substitution of a constant with another constant, and setting a constant to one [6].

The user can set the relative likelihood of the mutation, crossover, and reproduction processes (where the probabilities of these processes must add up to 1.0). The user may also define the probabilities of event subtypes, like specifying the likelihood of a two-point high-level crossover [6].

D. Multi-Objective Strategy

In practice, the simplest possible model is the most desirable, as increasing the complexity of a model increases its overfitting potential. Note that “complexity” refers to the expressional complexity of the GP model, not the computational complexity of the process. The computational complexity of GP is not within the scope of this study, as the data for such a complex case are limited. Therefore, the simplicity and accuracy of the



$$Y = d_0 + d_1(A + C)\left(\frac{3}{B}\right) + d_2\sqrt{(7 + A)e^B}$$

Fig. 6. Representation of an expression tree with two sub-trees as a typical multi-gene GP model.

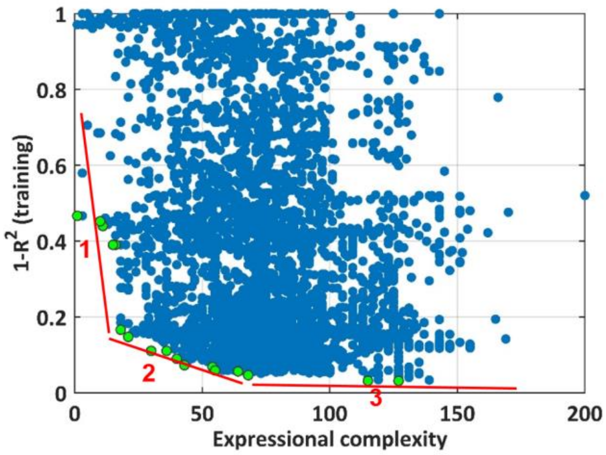


Fig. 7. Example of GP models and the Pareto set.

final model are the main concerns.

When the number of inputs for model development is low, model complexity can be easily controlled by the user. However, with an increasing number of input parameters, model complexity can become an issue. Therefore, the best models should be selected considering both model correlation and model complexity. In other words, multi-objective optimization is necessary to find the best models in terms of both complexity (in general terms) and performance [11].

The expressional complexity of the model is measured using the expression structure, which is more easily quantified than a response surface, especially when a tree-based structure is used. Smits and Kotanchek [24] proposed a strategy to measure the complexity of a GP model that depends on tree depth, tree nodes, component function nonlinearity, number of variables, and combinations of these parameters. This multi-objective genetic programming algorithm was implemented using a MATLAB®-based software platform called GPTIPS 2 [11]. More information about this platform can be found in Gandomi and Atefi [25].

These best models are determined using a trade-off surface called the Pareto front. Figure 7 shows an example of the Pareto set during optimization, where the green data points represent the Pareto set referring to the models with the best combinations of high correlation and low complexity.

The final model should be selected from the Pareto set, which includes the models that cannot dominate one another in both complexity and accuracy. In general, the Pareto set could be divided into three different regions as shown in Figure 7. In region 1, the accuracy of the models is significantly improved with a slight increase in complexity. For models in region 2, accuracy and complexity change together. In region 3, model complexity increases notably with only a small improvement in the accuracy, whereby the overfitting potential increases with increasing complexity. For maximum accuracy and limited complexity, the best model in the Pareto front set should be selected from solutions very close to the boundary of regions 2 and 3.

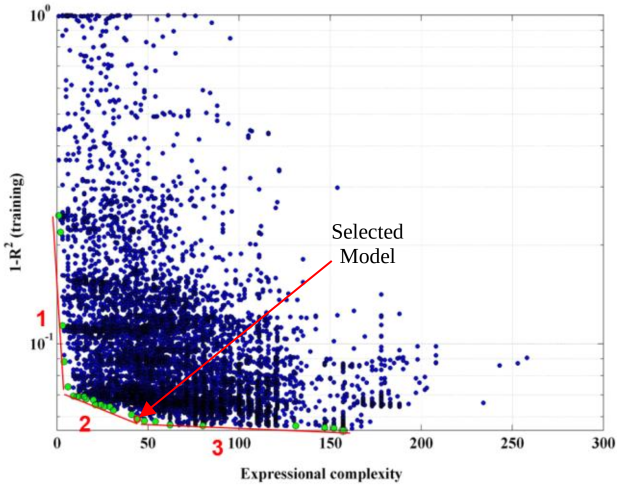


Fig. 8. MOGP Models with Highlighted Pareto front set and the selected Model.

TABLE II
EVOLUTIONARY CORRELATION AND CORRELATION RANK OF IMS

IM	R^2	Proposed			*F*-test	
		R_e^2	↑(%)	Rank	Rank	Score
$S_a(T)$	0.5589	0.7975	42.7	3	3	304.3
$S_a(2T)$	0.6709	0.8680	29.4	2	2	386.7
S_v	0.5560	0.7938	42.8	4	4	296.4
S_d	0.5147	0.7761	50.8	5	5	282.8
PGA	0.4190	0.5359	27.9	10	11	151.6
PGV	0.7765	0.9022	16.2	1	1	413.2
PGD	0.5181	0.7222	39.4	6	6	242.9
CAV	0.1890	0.5694	201.3	11	10	158.5
CAD	0.4036	0.6729	66.7	7	8	212.2
I_A	0.1461	0.6612	352.6	8	7	212.5
I_v	0.4233	0.6454	52.5	9	9	196.1
A_{rms}	0.2858	0.3235	13.2	13	13	72.9
I_c	0.2053	0.3305	61.0	12	12	74.9
T_D	0.0216	0.0881	307.9	14	14	32.6

IV. RESULTS & DISCUSSIONS

In this section, the results of phases i-iii are presented, and the selected model is validated using external validation metrics.

A. Phase i: Database Generation

Based on the design matrix in Table I, a database was developed using seismic analysis of the 75 designed SC-CBFs. Each of these structures was subjected to two ground motion suites. The first suite contains 30 recorded ground motions, scaled to the design-basis earthquake (DBE) level, which were previously used with SC-CBF systems by Roke *et al.* [13]. The second suite contains 140 ground motions from the SAC ground motion suite [26] at multiple hazard levels. All of the ground motions were applied to each of the designed SC-CBF systems to provide a database of earthquake responses in different hazard levels, so that SC-CBF responses to individual earthquakes could be predicted.

In addition to the structural design parameters, earthquake intensities are essential to predict peak response to an individual earthquake. Therefore, several normalized intensity measures (IMs) were considered as candidates for the final model (listed in Table S.1). The statistics of the 14 IMs computed for the 170 earthquake records are tabulated in Table S.2, and the summary of the earthquake characteristics are presented in Tables S.3-S.6. The distributions of these IMs are also visualized in Figures S.1(a)-S.1(n) of the Supplementary Materials.

B. Phase ii: Feature Selection

The variables considered to predict the peak roof drift for each individual earthquake include the mechanical and structural properties of the SC-CBF system and normalized IMs of the ground motion record. Before formulation, the three most effective IMs were selected based on their evolutionary coefficients (R_e^2) (as described in Section II.B), which were determined with respect to peak roof drift response. As higher R_e^2 values indicate better correlation of an IM to the peak roof drift, this reduces the number of IMs considered in the model, accelerating the process of finding a simple prediction equation. As these simulations are only for finding the most effective IMs, data with constant b, F_y, and μ values were used, arbitrarily

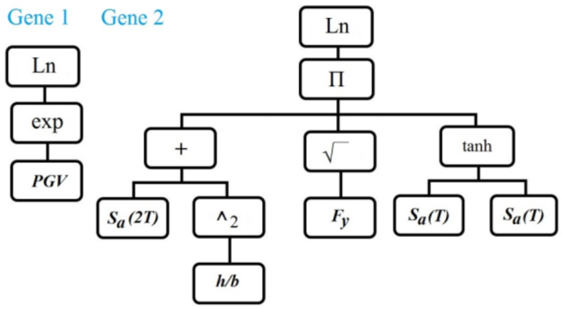


Fig. 9. Sub-trees from the selected model from the Pareto front.

set to the middle values presented in Table I. The associated R_e^2 values and ranking of the 13 normalized IMs and T_D are presented in Table II, and the correlation results are visualized in Figures S.2(a)-S.2(n) of the Supplementary Materials.

As shown in Table II, the normalized peak ground velocity, (*PGV*), the elastic spectral acceleration for double the first mode period ($S_a(2T)$), and the elastic spectral acceleration at the fundamental period of a structure ($S_a(T)$) exhibited the highest R_e^2 values (ranked from 1 to 3, respectively). Therefore, these IMs were selected for the model development in addition to the structural parameters used in the next section. The variable rankings were compared with *F*-test results, which score (–log(*p*-value)) the importance of each input variable individually for regression. The ranking is very similar to what is provided by R_e^2, validating our proposed approach. Note that the modification of R^2 (R_e^2) can measure nonlinear correlations used for feature selection, can reveal the hidden nonlinearity (presented in percentages in Table II) missed by R^2, and can provide an explicit equation to represent this nonlinearity.

It should be noted that the natural logarithm of the roof drift response ($ln(\theta)$) was used here, as the drift approximately follows a log-normal distribution, and improved the modeling results. With the selected IMs, the mathematical problem statement can be expressed as follows:

$$ln(\theta) = f(h/b, h, F_y,, \mu, PGV, S_a(T), S_a(2T)) \qquad (6)$$

Out of 10,500 datasets in each database, 8,400 datasets (80% of the data) were used for training, and the remaining 2,100 datasets (20%) were equally divided for validation and testing purposes. The next subsection discusses the use of a multi-objective strategy to find the best model for this complex problem, optimizing both accuracy and complexity at the same time.

C. Phase iii: Model Development

Multi-objective GP (MOGP) was subsequently employed in this phase to develop models with the basic mechanical and geometric variables of SC-CBF systems and three selected IMs (*PGV*, $S_a(T)$ and $S_a(2T)$). MOGP was utilized to model this highly nonlinear problem and to find a straightforward model. The best MOGP run was selected from 50 different runs with different random initial populations, which included 25,000 models in total. Out of the 25,000 possible models, 26 models were on the Pareto front (non-dominated solutions). All the solutions are presented in Figure 8, where the Pareto sets are highlighted in green. The final model should be selected from the Pareto set, which includes the models that cannot dominate each other in both complexity and accuracy. As previously discussed, the Pareto set could be divided into three different subsets and the best model in the Pareto front set should be located somewhere between regions 2 and 3. This strategy was used to select the final model. All Pareto front models are presented in Table S.7, in which the selected model is highlighted in blue. As shown in Figure 9, the selected model has two genes (sub-trees), whereby μ does not contribute to the final model. It should be noted that some models contained μ, but they did not dominate the final model in terms of accuracy and simplicity.

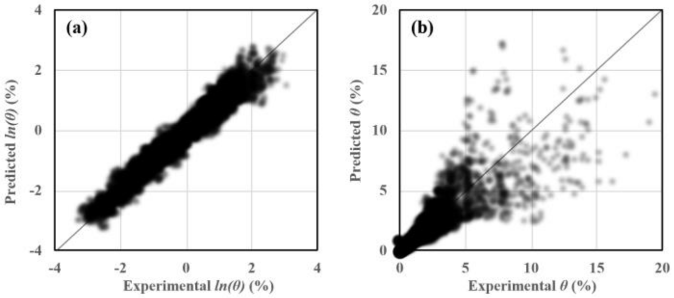


Fig. 10. Comparison of MOGP predictions and numerical analysis (experimental) results for peak roof drift prediction: (a) ln(θ); (b) θ.

A tree model can be represented by a simple formula. All mathematical models in the Pareto set are presented in Table S.7, which provides the complexity and performance characteristics of the models on the Pareto front and is sorted based on model accuracy. The mathematical formulation of the selected model from the Pareto set is as follows:

$$Ln(\theta) = 25.9PGV + 0.615\ln\left|\tanh(2S_a(T))\left(S_a(2T) + \left(\frac{h}{b}\right)^2\right)\sqrt{F_y}\right| - 1.08 \qquad (7)$$

The prediction of Eq. 7 is very accurate and straightforward, with R = 0.9700 and a complexity of 44. The prediction of θ is not as accurate as the prediction of $ln(\theta)$, whereby the correlation decreased to R = 0.8684. This is because the dispersion of error increased as θ increased. It should be noted that such large roof drifts rarely happen in practice and are difficult to accurately predict regardless of method. The prediction results of the best MOGP model for both $ln(\theta)$ and θ are shown in Figure 10, which confirms that MOGP can provide an accurate estimation of the peak structural response.

D. Comparison

The MOGP-based model was compared with three well-known approaches: traditional single-tree based GP, single objective Multi-Gene Genetic Programming (MGGP) without

TABLE III
COMPARISON WITH OTHER GPS

Metric	GP	GEP	MGGP	MOGP-X	MOGP
R^2	0.9133	0.9111	0.9426	0.9452	0.9409
RMSE[a]	0.3524	0.3577	0.2848	0.2812	0.2911
XC[b]	51	78	96	157	44

[a]*RMSE* is Root Mean Square Error; [b]*XC* is Expressional Complexity

considering complexity as an objective, and Gene Expression Programming (GEP) [27]. GEP was selected for comparison since it is another GP variant with a multi-gene structure that provides a good benchmark. Additionally, the most extreme model in terms of accuracy among Pareto-front solutions from the proposed MOGP technique (called MOGP-X) was selected for comparison. For MOGP-X, complexity was the objective during the training (evolution) but not during model selection. The comparison of these models is presented in Table III in terms of accuracy and complexity metrics. The results show that MGGP and MOGP outperformed the well-known GP and GEP models in terms of accuracy. As expected from the multi-objective strategy, the selected MOGP model exhibited the lowest complexity among the models. MGGP slightly outperformed MOGP in terms of error; comparatively, MGGP aims to select the most accurate model, while MOGP attempts to find the best trade-off between accuracy and complexity. It was found that MOGP-X is the best model in terms of accuracy, not complexity, which may be due to the fact that it takes advantage of a multi-objective strategy for diverse search but only focuses on accuracy during model selection. This could be an effective strategy when accuracy is the final and only objective. More detailed results of MOGP-X and MGGP, which outperformed MOGP (though at a significant cost of model complexity), can be found in the Supplementary Materials.

V. Summary & Conclusions

In this work, 75 SC-CBF systems with different mechanical and geometric properties were designed to develop a model for predicting peak roof drift response of self-centering concentrically braced frame (SC-CBF) systems. The peak response prediction for an SC-CBF subjected to an individual earthquake record was formulated using multi-objective genetic programming (MOGP) with multiple genes. The multiple genes in the MOGP models were connected using regression analysis.

Structural design variables and normalized earthquake intensity measures were used as model parameters. A new evolutionary method was proposed in this study to find the intensity measures most correlated to the peak roof drift response. The multi-objective strategy was applied to find the best model in terms of both accuracy and simplicity. The results of this study are presented in three phases as follows:

i. Database Generation
ii. Feature Selection
iii. Model Development

The final model in this strategy was selected from the nondominated solutions. Not only is the selected model relatively simple, but it is also accurate and can successfully predict the response of each specific earthquake.

The findings of the current research are described below:

- Extensive time-history analyses on the 75 designed SC-CBF systems under 170 earthquakes resulted in a comprehensive dataset for accurate analysis and modeling of the peak dynamic responses.
- GP models are based on experimental data instead of on assumptions regarding system behavior. GP has not been used for formulating the design process of any structural system in the literature; therefore, GP can be considered a new tool in this field.
- The effects of ground motion intensity measures (IMs) on the SC-CBF system response were ranked using a newly proposed feature selection strategy called the evolutionary correlation coefficient approach.
- The highly nonlinear SC-CBF system was investigated using a multi-objective GP approach to formulate this complex system with the aim to find a simple but accurate model.
- The Pareto front models of MOGP were presented for the SC-CBF system, and the best model was selected from the Pareto front set based on a simple decision-making procedure.
- The final model was further verified using several external validation criteria. From the results, it can be concluded that the selected model is both simple, very accurate, and can successfully predict the target in this complex problem.

Future work in this field should investigate 3D models, hybrid structural systems and high rise buildings. Also, additional structural parameters can be incorporated into the GP models such as initial stress in the PT bars.

References

[1] Suykens, J.A. and Vandewalle, J.P. eds., 2012. Nonlinear Modeling: advanced black-box techniques. Springer Science & Business Media.

[2] Gandomi, A. H., & Roke, D. A. (2015). Assessment of artificial neural network and genetic programming as predictive tools. Advances in Engineering Software, 88, 63-72.

[3] Koza, J.R. (1992). Genetic programming: on the programming of computers by means of natural selection. MIT Press, Cambridge, MA.

[4] Brameier, M. F., & Banzhaf, W. (2007). Linear genetic programming. Springer Science & Business Media.

[5] Gandomi, A. H., & Alavi, A. H. (2011). Multi-stage genetic programming: a new strategy to nonlinear system modeling. Information Sciences, 181(23), 5227-5239.

[6] Searson, D.P., Leahy, D.E. and Willis, M.J., (2010). GPTIPS: an open source genetic programming toolbox for multigene symbolic regression. Proceedings of the International Multi Conference of Engineers and Computer Scientists (IMECS 2010), Hong Kong, 17-19 March.

[7] Arnaldo, I., Krawiec, K. and O'Reilly, U.M., 2014, July. Multiple regression genetic programming. In Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation (pp. 879-886). ACM.

[8] Gandomi, Amir H., Amir H. Alavi, and Conor Ryan, eds. Handbook of genetic programming applications. Switzerland: Springer, 2015.

[9] Lensen, A., Xue, B., & Zhang, M. (2020). Genetic Programming for Evolving a Front of Interpretable Models for Data Visualization. IEEE Transactions on Cybernetics.

[10] Vanneschi, L., Castelli, M. and Silva, S., 2010, July. Measuring bloat, overfitting and functional complexity in genetic programming. In Proceedings of the 12th annual conference on Genetic and evolutionary computation (pp. 877-884). ACM.

[11] Searson D.P., GPTIPS 2: an open-source software platform for symbolic data mining. Chapter 22 in Handbook of Genetic Programming Applications, A.H. Gandomi *et al.*, (Eds.), Springer, New York, NY, 2015.

[12] Roke, D., Sause, R., Ricles, J.M., Seo, C.-Y., and Lee, K.-S. (2006). Self-Centering Seismic-Resistant Steel Concentrically-Braced Frames. Proceedings of the 8th U.S. National Conference on Earthquake Engineering, EERI, San Francisco, April 18-22.

[13] Roke, D., Sause, R., Ricles, J.M., and Chancellor, N.B. (2010). Damage-Free Seismic-Resistant Self-Centering Concentrically-

Braced Frames. ATLSS Report 10-09, Lehigh University, Bethlehem, PA, USA.

[14] Gandomi, A. (2015). Seismic Response Prediction of Self-Centering Concentrically Braced Frames Using Genetic Programming (Doctoral dissertation, The University of Akron).

[15] Huang, Q., Gardoni, P., and Hurlebaus, S. (2010). Probabilistic Seismic Demand Models and Fragility Estimates for Reinforced Concrete Highway Bridges with One Single-Column Bent. Journal of Engineering Mechanics, 136(11), 1340–1353.

[16] McKenna, F., Fenves, G. L., & Scott, M. H. (2000). Open system for earthquake engineering simulation. University of California, Berkeley, CA.

[17] Gandomi, A.H. (2017) Interview by Marjan Eggermont. NASA-VINE Biomimicry Summit for Aerospace. Zygote Quarterly, ZQ 18, pages 76-83.

[18] Ratner (2012) Finding the best variables for marketing models. Chapter 30 in Statistical and Machine-Learning Data Mining. 2nd ed., CRC Press.

[19] GenIQ http://www.geniqmodel.com/What-Is-The-GenIQ-Model.html, last access April 2, 2020.

[20] Gandomi, A. H., & Alavi, A. H. (2012). A new multi-gene genetic programming approach to nonlinear system modeling. Part I: materials and structural engineering problems. Neural Computing and Applications, 21(1), 171-187.

[21] Garg, A., Garg, A., & Lam, J. S. L. (2015). Evolving functional expression of permeability of fly ash by a new evolutionary approach. Transport in Porous Media, 107(2), 555-571.

[22] Muduli, P. K., & Das, S. K. (2014). CPT-based seismic liquefaction potential evaluation using multi-gene genetic programming approach. Indian Geotechnical Journal, 44(1), 86-93.

[23] Searson D.P., Willis, M.J., and Montague, G.A. (2007). Co-evolution of nonlinear PLS model components, Journal of Chemometr 2:592-603.

[24] Smits, G.F., and Kotanchek, M. (2005). Pareto-Front Exploitation in Symbolic Regression. Chapter 17 in Genetic Programming Theory and Practice II. U.-M., O'Reilly *et al.* (Eds.), Boston, MA: Springer. Vol. 8, pp 283-299.

[25] Gandomi, A. H., & Atefi, E. (2019). Software review: the GPTIPS platform. Genetic Programming and Evolvable Machines, 1-8.

[26] Somerville, P., Smith, N., Punyamurthula, S., and Sun, J. (1997). Development of ground motion time histories for phase 2 of the FEMA/SAC steel project, SAC/BD-97/04. Sacramento, CA: SAC joint venture

[27] Ferreira, C. (2006). Gene expression programming: mathematical modeling by an artificial intelligence (Vol. 21). Springer.

Amir H. Gandomi (Senior Member) is a Professor of Data Science and an ARC DECRA Fellow at the Faculty of Engineering & Information Technology, University of Technology Sydney. Prior to joining UTS, Prof. Gandomi was an Assistant Professor at Stevens Institute of Technology, USA and a distinguished research fellow at BEACON center, Michigan State University, USA. Prof. Gandomi has published over three hundred journal papers and 12 books which collectively have been cited 29,000+ times (H-index = 77). He has been named as one of the most influential scientific minds and Highly Cited Researcher (top 1% publications and 0.1% researchers) for five consecutive years, 2017 to 2021. He also ranked 17th in GP bibliography among more than 12,000 researchers. He has received multiple prestigious awards for his research excellence and impact, such as the 2022 Walter L. Huber Prize which is known as the highest level mid-career research award in all areas of civil engineering. He has served as associate editor, editor, and guest editor in several prestigious journals such as AE of IEEE TBD and IEEE IoTJ. Prof Gandomi is active in delivering keynotes and invited talks. His research interests are global optimisation and (big) data analytics using machine learning and evolutionary computations in particular.

David A. Roke is an Associate Professor of Civil Engineering in the College of Engineering and Polymer Science at The University of Akron. Dr. Roke has produced seminal research relating to the development of self-centering braced frame systems and continues their study with his graduate students. His primary research interests are damage-resistant structural systems for earthquake engineering, the use of recycled materials in concrete mixes, and engineering education.