J. Mech. Phys. Solids 194 (2025) 105916

Contents lists available at ScienceDirect

Journal of the Mechanics and Physics of Solids

journal homepage: www.elsevier.com/locate/jmps

Mechanics-informed, model-free symbolic regression framework for solving fracture problems

Ruibang Yi [a], Dimitrios Georgiou [a], Xing Liu [b,c], Christos E. Athanasiou [a,*]

[a] Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA
[b] George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA
[c] Department of Mechanical and Industrial Engineering, New Jersey Institute of Technology, Newark, NJ, 07102, USA

ARTICLE INFO

Keywords:
Fracture mechanics
Machine learning
Symbolic regression
Fracture toughness

ABSTRACT

Data-driven methods have recently been introduced to address complex mechanics problems. While model-based, data-driven approaches are predominantly used, they often fall short of providing generalizable solutions due to their inherent reliance on pre-selected models. Model-free approaches, such as symbolic regression, hold promise for overcoming this limitation by extracting solutions directly from datasets. However, these approaches remain unexplored when dealing with high-dimensional fracture mechanics problems and require significant customization to be effective. In this work, we propose a new symbolic regression framework that integrates mechanics knowledge to enhance the ability to generalize solutions. This framework also includes a model-free variable separation scheme to decouple high-dimensional problems into simpler sub-problems with manageable complexity while preserving data fidelity. We demonstrate the advantages of this framework through two fracture mechanics problems, showing that it can potentially provide generalizable, analytical solutions to novel, easy-to-use fracture testing configurations.

1. Introduction

The path of scientific breakthroughs is often shaped by combining data with physical intuition to discover mathematical expressions that accurately describe datasets - a form of symbolic regression. An early example of this can be traced back to Ptolemy in 100 AD, who sought to match his astronomical observations with mathematical models to predict planetary motion. Centuries later, in the field of fracture mechanics, the development of Paris' law, the most established model for predicting the growth rate of fatigue cracks, showcases a similar approach where empirical data guided the formulation of a fundamental mechanics law (Paris and Erdogan, 1963; Paris et al., 1961).

As the field of mechanics is becoming increasingly interdisciplinary, *e.g.,* coupled with chemistry (Athanasiou et al., 2024), optics (McMillen et al., 2016), and acoustics (Mozaffari et al., 2023), the need to 'learn' physics or extract solutions from complex datasets becomes more apparent. To address this need, model-based efforts have been recently introduced, including machine learning models, such as neural networks (NNs), and Gaussian process regression (Lu et al., 2020; Briñez-de León et al., 2022; Goswami et al., 2022; Chen and Gu, 2023; Dekhovich et al., 2024; Gu et al., 2018; Bessa et al., 2017; Vlassis and Sun, 2021; Buehler, 2024; Fuhg et al., 2024; Niu and Srivastava, 2022; Agyei et al., 2023; Pantidis and Mobasher, 2023; Liu et al., 2023; Karapiperis and Kochmann, 2023; Liu

* Corresponding author.
E-mail address: athanasiou@gatech.edu (C.E. Athanasiou).

https://doi.org/10.1016/j.jmps.2024.105916
Received 27 June 2024; Received in revised form 17 September 2024; Accepted 18 October 2024
Available online 28 October 2024

et al., 2020; Liu et al., 2021; Athanasiou et al., 2023). For example, Lu et al. used NNs for extracting mechanical properties from instrumented indentation experiments (Lu et al., 2020), and Briñez-de León et al. (2022) developed a deep convolutional NN (PhotoelastNet) for photoelasticity imaging-based stress evaluation. While model-based data-driven approaches are widely used, they often fail to offer generalizable solutions due to their inherent dependence on pre-selected models. Model-free approaches hold promise for overcoming this limitation by extracting solutions directly from datasets (Karapiperis et al., 2021; Prume et al., 2023; Bahmani and Sun, 2024; Bomarito et al., 2021; Flaschel et al., 2022).

Symbolic regression is a data-driven, model-free method that can uncover intricate mathematical relationships. Using symbolic regression, Bahmani and Sun (2024) discovered new constitutive models for polyconvex incompressible hyperelastic materials, Bomarito et al. (2021) developed a new plasticity model for porous materials, and Flaschel et al. extracted plasticity models directly from full-field displacement and global force data (Flaschel et al., 2022). Nevertheless, the capability of existing symbolic regression methods to handle complex high-dimensional datasets and identify generalizable analytical solutions has not been fully explored. To address this gap, we propose a mechanics-informed symbolic regression framework based on model-free variable separation to derive generalizable analytical expressions for complex fracture problems. By applying mechanics knowledge to preprocess the training dataset and using the model-free variable separation, our framework can deconstruct complex high-dimensional problems into simpler, more manageable ones with fewer variables. Additionally, employing mechanics criteria to select the appropriate expression

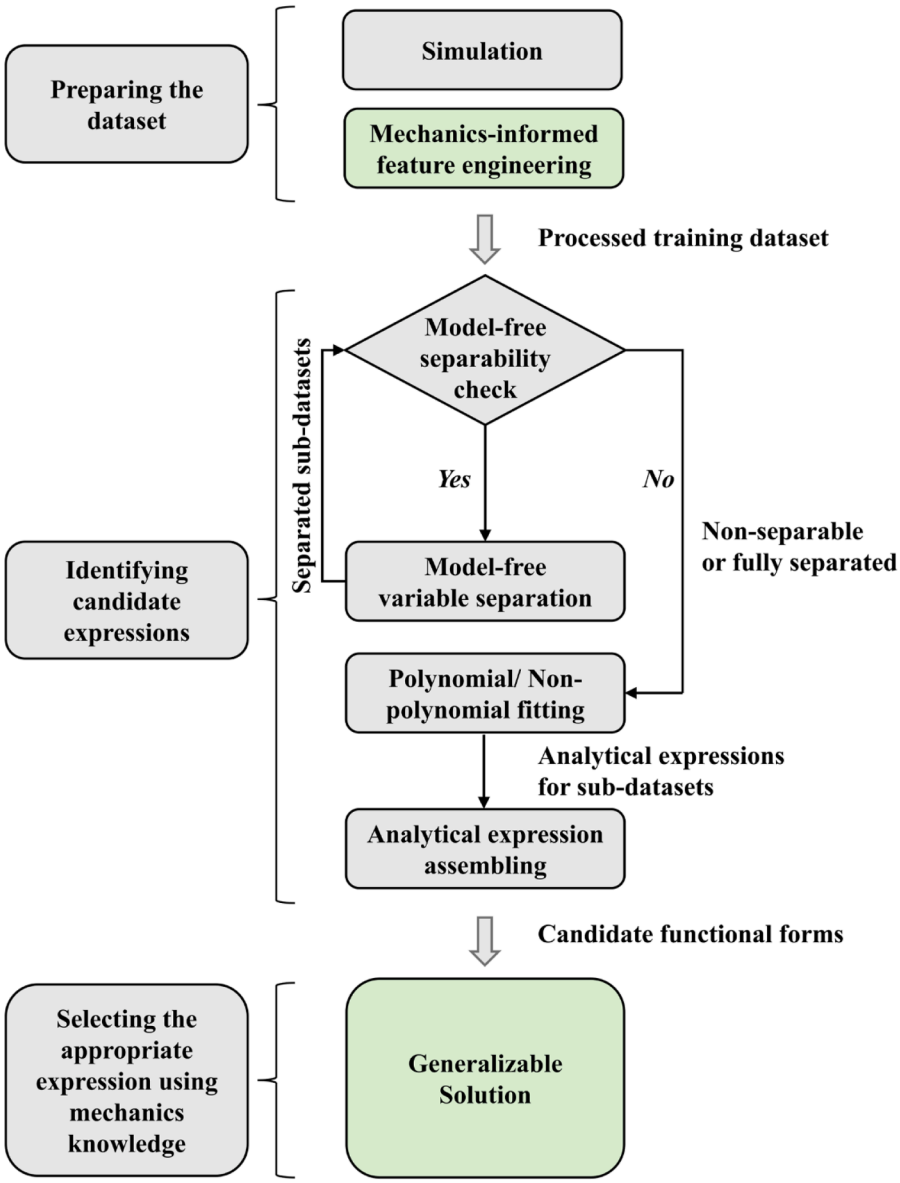


Fig. 1. Flowchart of the mechanics-informed symbolic regression algorithm. Mechanics-informed parts are green colored.

from an extensive candidate list can enhance the accuracy and generalizability of the resulting solution. We demonstrate this framework using two benchmark fracture mechanics problems: *i*) fracture of a Single Edge Notched Beam (SENB) specimen under three-point bending, and *ii*) fracture of a Crackline-Loaded Edge-Crack (CLEC) specimen. Our symbolic regression framework provides a methodological pathway for the extraction of fracture mechanics solutions and can potentially be extended to obtain analytical solutions for non-standardized fracture testing configurations.

2. Incorporating mechanics knowledge in symbolic regression

Symbolic regression is a widely used method in multiple scientific fields (Zhang et al., 2021; Davidson et al., 2003; LeCun et al., 1998; Koza, 1992; Udrescu and Tegmark, 2020); it typically consists of three key steps: dataset preparation, identification of candidate expression, and selection of the appropriate expression. It requires the use of a *dataset* of sufficient size, which can be obtained from various sources, with the most common being Finite Element Analysis (FEA) simulations or experimental data. Feature engineering is a common challenge in dataset preparation. To address this, we employ mechanics insights to reconstruct the raw data, thereby improving the performance of symbolic regression. There are two main categories of algorithms for *identifying candidate expressions*: those for polynomial cases (*e.g.*, gradient-based algorithms) (Davidson et al., 2003; LeCun et al., 1998) and those for non-polynomial cases (*e.g.*, brute-force genetic programming algorithms) (Koza, 1992). However, existing methods show poor performance when dealing with high dimensional correlations, a challenge we address by splitting the dataset into sub-datasets of reduced dimensionality. We obtain symbolic expressions for the reduced-dimensionality datasets and use them to recover the symbolic expression for the original high-dimensional problems. When *selecting the appropriate expressions*, existing methods rely purely on mathematical criteria, which can lead to overfitting. Our approach overcomes this challenge by embedding mechanics knowledge in the expression selection process.

2.1. Dataset preparation

Symbolic regression is the machine learning task aimed at discovering equations from collections of measured data. Symbolic regression methods take a dataset S consisting of multiple measurements of a set of variables $S = \{X,\ Y\}$, where X is a set of variables $\{x_1, x_2, \ldots\}$ and Y is a target value. The output of symbolic regression is an equation of the form $Y = f(X)$, where the right-hand side of the equation is a closed-form mathematical expression. The equation should provide an optimal fit for the measurements in S, *i.e.*, minimize the discrepancy between the observed values of the target variable Y and the values calculated using the equation (Udrescu and Tegmark, 2020). For example, symbolic regression can be performed on fatigue data including crack growth rate $\left(\frac{da}{dN}\right)$ and the range of stress intensity factor, ΔK, to extract the Paris law (Paris and Erdogan, 1963; Paris et al., 1961). Dataset preparation requires careful consideration of two key elements: *i*) the dataset size, and *ii*) the parameter space. Larger parameter spaces and datasets are beneficial to the performance of the symbolic regression algorithm; however, they are subject to limitations related to computational or experimental resources. Here, we start from a predefined parameter space, Λ, and construct a small training dataset. We gradually increase the size of the training dataset until it is sufficient for obtaining a convergent expression. Additionally, converting the raw data into characteristic quantities, (*e.g.*, in case of Paris law using the range of the stress intensity factor ΔK instead of applied tensile far-field loads (Paris and Erdogan, 1963; Paris et al., 1961)), can significantly enhance the performance of the algorithm. After selecting the appropriate expression using established mechanics criteria (Fig. 1), we construct an extended dataset $\{X^g, Y^g\}$ within a larger parameter space Λ^g to test the generalization error of the selected expression, *i.e.*, the maximum relative error E^g.

2.2. Identification of candidate expressions

To break down the challenge of identifying the high-dimensional correlations, we split the dataset. The algorithm examines the separability of variables and splits the original dataset into sub-datasets of lower dimensions accordingly. Then, we use polynomial and non-polynomial fitting to identify expressions for each of the fully separated datasets. Assembling the expressions obtained for each of the sub-datasets provides a set of candidate formulas able to fit the original dataset.

2.2.1. Model-free quantification of separability & creation of sub-datasets

We focus on additive and multiplicative separability of variables, as they are the most common mathematical operators. For any dataset $X = \{x_1, \ldots, x_N\}$, comprising N number of variables and n number of datapoints (*i.e.* $dim(X) = n \times N$) we define $X_i = \left\{x_{i_1}, \cdots, x_{i_k}, \cdots, x_{i_{N_i}}\right\}$ as a subset of X, and $X_j = \left\{x_{j_1}, \cdots, x_{j_k}, \cdots, x_{j_{N_j}}\right\}$ as the remaining subset of X such that $X = X_i + X_j$. Here, k indexes the variables in each subset, and ranges from 1 to the total number of variables in the respective datasets, N_i and N_j, *i.e.* $k = 1, \ldots, N_i$ $(resp.\ N_j)$. The total number of variables in X is denoted by N such that $N = N_i + N_j$. Furthermore, we denote the p-th datapoint of the k-th variable (*e.g.* x_{i_k}), or of a dataset (*e.g.* X), by $\left[x_{i_k}\right]_p$ and $[X]_p$ respectively, where p indexes the datapoints, and ranges from 1 to the total number of datapoints n, *i.e.* $p = 1, 2, \ldots, n$.

Instead of using the NN-based approach (Udrescu and Tegmark, 2020), we apply a model-free approach to quantify the additive and multiplicative separability indices of X_i ($E^{add}(X_i)$, $E^{mult}(X_i)$) by introducing error functions that calculate the error induced by separating X_i directly from the original dataset X. Such a model-free approach is expected to eliminate the uncertainty in quantifying

the separability. A lower separability index suggests better separability. The additive separability index E^{add} is given by Eq. (1)

$$E^{add}(X_i) = \frac{1}{n}\sum_{p=1}^{n}\left|\frac{\left[Y([X]_p) - Y([X_i]_p, C_j) - Y\left(C_i, [X_j]_p\right) + Y(C_i, C_j)\right]}{Y([X]_p)}\right| \tag{1}$$

Similarly, the multiplicative separability index E^{mult} is given by Eq. (2):

$$E^{mult}(X_i) = \frac{1}{n}\sum_{p=1}^{n}\left|\frac{\left[Y([X]_p) - \frac{Y([X_i]_p, C_j) \times Y\left(C_i, [X_j]_p\right)}{Y(C_i, C_j)}\right]}{Y([X]_p)}\right| \tag{2}$$

Where C_i (resp. C_j) is a set containing N_i (resp. N_j) constants $\left\{c_{i_1}, \cdots, c_{i_k}, \cdots, c_{i_{N_i}}\right\}$ (resp. $\left\{c_{j_1}, \ldots, c_{j_k}, \ldots, c_{j_{N_j}}\right\}$) where each constant corresponds to the mean value of the respective variable in X_i (resp. X_j). Specifically, the k-th constant in C_i (resp. C_j), denoted by c_{i_k} (resp. c_{j_k}), is defined as the mean value of the variable x_{i_k} (resp. x_{j_k}) across all data points as shown in Eq. (3)

$$c_{i_k} = \frac{1}{n}\sum_{p=1}^{n}\left[x_{i_k}\right]_p \tag{3}$$

We test all the possible subsets, X_i, of X, and select the one with the lowest separability index to split the dataset X. We define parent datasets $\{\widetilde{X}, \widetilde{Y}\}$ as datasets containing more than one variable, and child datasets $\{\widetilde{X}_i, \widetilde{Y}_i\}$ and $\{\widetilde{X}_j, \widetilde{Y}_j\}$ as the subset of the corresponding parent dataset with the lowest separability index and the remaining part, respectively. Moreover, we define E^t as the threshold such

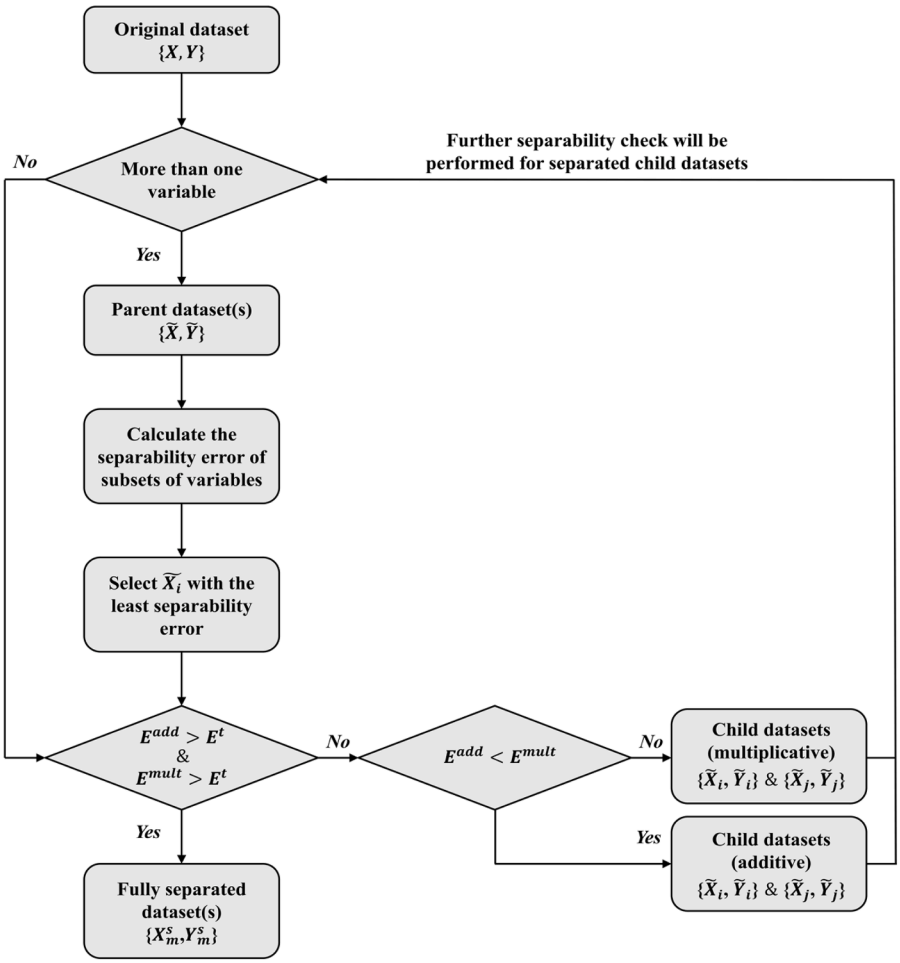


Fig. 2. Flowchart of quantifying separability and creating sub-datasets, where X are variables and Y is the target value. The use of the tilde indicates that the computation takes place within the iterative process.

that variables with separability index lower than this threshold ($E^{add} < E^t$ or $E^{mult} < E^t$) are considered separable. Separable variables should have a separability index of 0. However, separable variables may lead to a non-zero separability index due to noise in the experimental data. Additionally, variables that are separable within the domain of interest but non-separable across the entire definition space can also produce a non-zero separability index. When both separability indices are lower than the threshold ($E^{add} < E^t$ and $E^{mult} < E^t$), the separation scheme with the lower separability index is used, and the parent dataset is split to child datasets. The criterion of selecting the separability threshold, E^t, is illustrated in Appendix A. Once E^t is determined, the order of sub-datasets of the original dataset will be fixed. After we separate the dataset, the expressions for each sub-dataset will be obtained independently, and the order of sub-datasets will not influence the expression candidates. This separation process is iterative: divided datasets with more than one variable (*i.e.*, parent datasets) are evaluated in terms of their separability, and the separation process continues until fully separated sub-datasets $\{X_m^s,\ Y_m^s\}$, are constructed, where $m = 1,\ 2,\ \ldots,\ M$ with M being the total number of fully separated sub-datasets and the superscript s denoting fully separated sub-datasets. Fully separated datasets are defined as those with separability index greater than the threshold values ($E^{add} > E^t$ and $E^{mult} > E^t$), or those containing only one variable. The procedure is illustrated in Figs. 2 and 3.

The child datasets are constructed as follows: the parent dataset $\{\widetilde{X}, \widetilde{Y}\}$ is split into two additive or multiplicative child datasets $\{\widetilde{X}_i, \widetilde{Y}_i\}$ and $\{\widetilde{X}_j, \widetilde{Y}_j\}$, based on the scheme dictated by the lowest separability index, provided it falls below the threshold, as follows:

$$\widetilde{Y}_i(\widetilde{X}_i) = \widetilde{Y}(\widetilde{X}_i, \widetilde{C}_j) \ \textit{and}\ \widetilde{Y}_j(\widetilde{X}_j) = \widetilde{Y}(\widetilde{C}_i, \widetilde{X}_j) \tag{4}$$

where $\widetilde{C}_i$ and $\widetilde{C}_j$ are sets of constants such that each constant corresponds to the mean value of the respective variable in $\widetilde{X}_i$ and $\widetilde{X}_j$ respectively (Eq. (3)), and $\widetilde{Y}_i$ and $\widetilde{Y}_j$ are the target values of $\widetilde{X}_i$ and $\widetilde{X}_j$, respectively. Fitting the child datasets $\{\widetilde{X}_i, \widetilde{Y}_i\}$ and $\{\widetilde{X}_j, \widetilde{Y}_j\}$ yields a

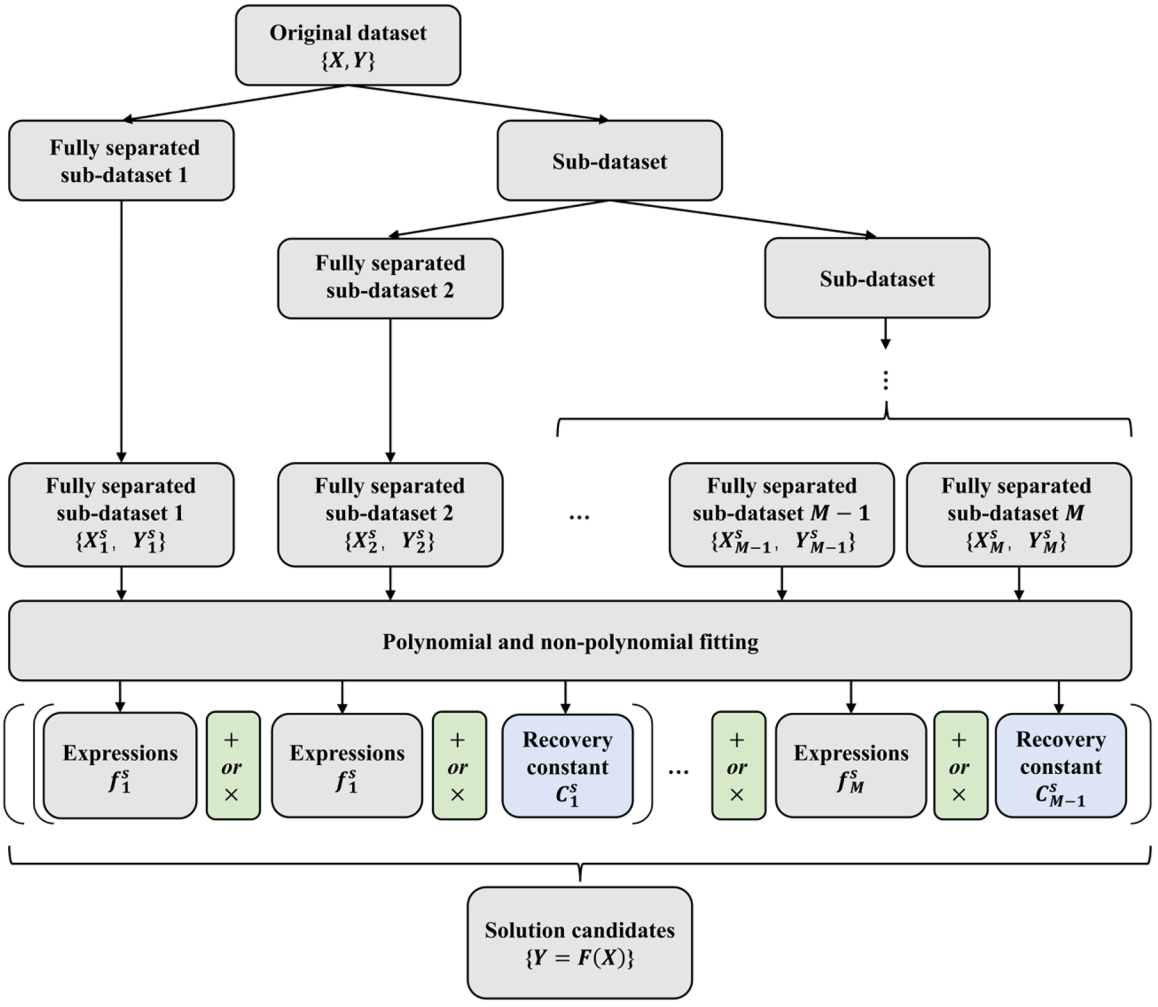


Fig. 3. Flowchart showcasing the assembly of candidate expressions. Each fully separated sub-dataset will provide a list of expressions. Recovery constants are used to ensure that when the sub-expressions are combined, the solution accurately reflects the original dataset.

list of symbolic expressions $\widetilde{Y}_i = \widetilde{f}_i(\widetilde{X}_i)$ and $\widetilde{Y}_j = \widetilde{f}_j(\widetilde{X}_j)$ respectively. For each separation a recovery constant, C_m^s, is extracted from the parent dataset as follows:

$$C_m^s = \widetilde{f}_i(\widetilde{C}_i) \ + \ or \ \times \ \widetilde{f}_j(\widetilde{C}_j) = \ \widetilde{Y}(\widetilde{C}_i, \widetilde{C}_j) \tag{5}$$

These recovery constants are required to recover the expressions of the parent dataset $\widetilde{Y} = \widetilde{f}(\widetilde{X})$. While creating the sub-datasets, we directly utilize the data points from the parent datasets, which differs from the NN-based approach in (Udrescu and Tegmark, 2020). The NN-based approach fits a NN to the parent dataset and then uses the NN to generate the sub-datasets, resulting in loss of fidelity. In contrast, our proposed approach is model-free, thereby preserving fidelity.

2.2.2. *Brute-force & polynomial fitting*

Polynomial and non-polynomial expressions, f_m^s, for each fully separated dataset $\{X_m^s,\ Y_m^s = f_m^s(X_m^s)\}$ are calculated using a gradient-based algorithm and a brute-force genetic programming algorithm, respectively. Gradient-based algorithms, including gradient descent optimization algorithms and their numerous variants, are typically used to solve unconstrained optimization problems (LeCun et al., 1998). The brute-force genetic programming algorithm is a commonly employed method for screening all possible combinations of mathematical operators within a predefined functional space of bounded complexity and obtaining the expressions that best fit the parent dataset (Koza, 1992; Udrescu and Tegmark, 2020). Both polynomial and brute-force methods are used to identify an expression, which is discussed in Appendix B.

2.2.3. *Assembling candidate expressions for the original dataset*

To recover the solution for the original dataset ($Y = f(X)$), we need to assemble the symbolic expressions found for the fully separated datasets $f_m^s(X_m^s)$, the recovery constants C_m^s, and the corresponding mathematical operators ($+$ or $\times$) in a sequential manner. As illustrated in Fig. 3, the expressions of each fully separated sub-dataset should be assembled according to the tree structure. The entire list of symbolic expressions $\widetilde{Y}_i = \widetilde{f}_i(\widetilde{X}_i)$ and $\widetilde{Y}_j = \widetilde{f}_j(\widetilde{X}_j)$ will be considered, resulting in a larger potential expression space compared to selecting the optimal single formula or constructing a Pareto front (*i.e.*, a collection of candidate expressions with increasing complexity and accuracy). A broader pool of candidate expressions contains more information about each variable and is more likely to preserve the formula that adheres to the established mechanics criteria in the subsequent selection process.

2.3. *Selecting the appropriate expression*

We aim to identify a formula with strong generalization capability (*i.e.,* a small E^g) rather than simply achieving the best fit to the training dataset (*i.e.,* minimize E^{train}). E^g and E^{train} are defined by Eq. (6a) and Eq. (6b), where $\{X^g, Y^g\}$ represents an extended dataset within a larger parameter space Λ^g and $\{X^{train},\ Y^{train}\}$ is the training dataset.

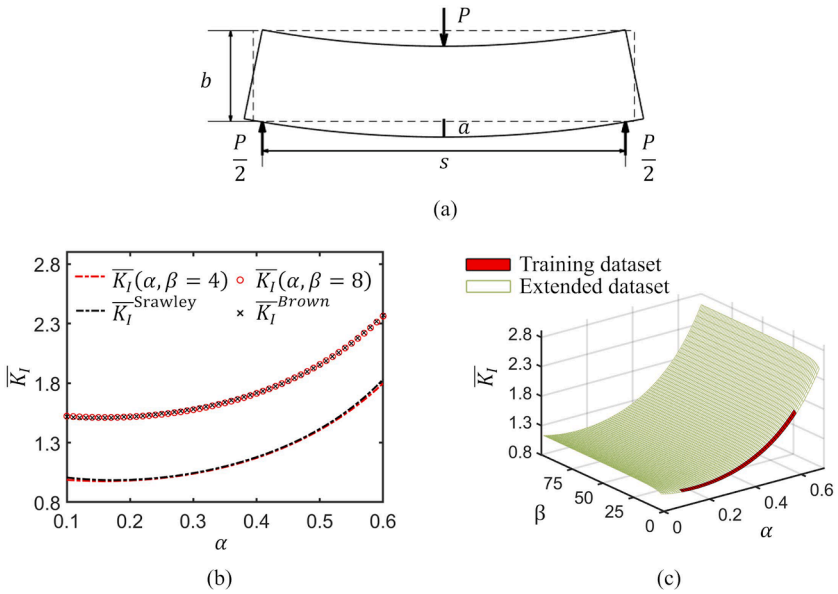


Fig. 4. (a) A SENB specimen under the three-point bend load. (b) The mechanics-informed symbolic regression result is consistent with Srawley's solution and Brown's solution with maximum relative error being $< 2\%$ for $\overline{K_I}(\alpha, \beta = 4)$ and $< 1\%$ for $\overline{K_I}(\alpha, \beta = 8)$, respectively. (c) The result obtained from the training dataset (red) can be generalized in the extended dataset (green), where the maximum relative error, E^g, of the obtained expression is lower than 10%.

$$E^g = \left(\left| \frac{F(X^g) - Y^g}{Y^g} \right| \right) \tag{6a}$$

$$E^{train} = \left(\left| \frac{F\left(X^{train}\right) - Y^{train}}{Y^{train}} \right| \right) \tag{6b}$$

Mechanics knowledge can play a crucial role in selecting an appropriate expression from the list of candidate expressions. Instead of relying solely on the performance of the training dataset, which is commonly adopted in symbolic regression (Davidson et al., 2003; LeCun et al., 1998; Koza, 1992; Udrescu and Tegmark, 2020), our new symbolic regression framework identifies the most generalizable solutions by examining their consistency with established mechanics criteria. This approach provides a more generalizable formula which does not necessarily yield the minimum E^{train}. The mechanics-informed formula selection effectively avoids overfitting issues and enhances the applicability of our solution across a broader range of parameters.

3. Case 1: Fracture of a Single Edge Notched Beam (SENB) in a three-point bending configuration

The SENB is a standard configuration for quantitative evaluation of mode-I fracture toughness (K_{Ic}) (Fig. 4). The geometry of the beam is described as follows: the distance between two support points is s, the depth of the beam is b, and the notch depth is a. The cross-section of the beam is rectangular, allowing the problem to be reduced to a two-dimensional (2D) problem. The three-dimensional (3D) beam is placed on two support mounts, with an external force per unit thickness, P, applied to its middle point on the top face. The stress intensity factor at the crack tip is K_I. The mechanics-informed symbolic regression procedure is utilized to derive a general form:

$$K_I = K_I(a, b, s, P) \tag{7}$$

We build a 2D finite element method (FEM) model and employ the J-integral (Rice, 1968) to calculate K_I for a given set of $\{a,b,s,P\}$. Using the Buckingham π theorem (Evans, 1972), we obtain three dimensionless variables $\{\alpha = \frac{a}{b}, \beta = \frac{s}{b}, \widetilde{K}_I = \frac{K_I\sqrt{b}}{P}\}$. The first two define the geometry (crack-to-depth ratio, span-to-depth ratio, respectively) while the third is the normalized stress intensity factor. We introduce $\sigma_{avg} = \frac{3Ps}{2b^2}$, which is the stress of an unnotched beam under bending, to approximate the stress at the crack tip of a notched beam. By introducing this intermediate variable, we can rewrite the normalized stress intensity factor as $\widetilde{K}_I = \frac{K_I b^2}{\sigma_{avg} s a^{\frac{3}{2}}}$. After the mechanics-informed feature engineering, the desired general form of $\widetilde{K}_I$ can be denoted as:

$$\widetilde{K}_I = \widetilde{K}_I(\alpha, \beta) \tag{8}$$

In practice, the range of α and β is $0 < \alpha \leq 0.7$ and $\beta \geq 4$. We utilize the predefined space of raw input parameters $\{\alpha, \beta\}$ which is denoted by $\Lambda = [0.1, 0.6] \times [4, 6]$ containing $11 \times 9 = 99$ uniformly spaced data points. The original dataset is separated according to the separability indices defined in Eqs. (1) and (2). The values of additive and multiplicative separability indices are 1.71% and 0.06%, respectively. Consequently, the original dataset is separated into two fully separated datasets $\left\{X^s_{m=1}, Y^s_{m=1}\right\}$ and $\left\{X^s_{m=2}, Y^s_{m=2}\right\}$ following the multiplicative rule, where $X^s_{m=1} = \{\alpha\}$, $X^s_{m=2} = \{\beta\}$, and $Y^s_{m=1}$ and $Y^s_{m=2}$ are defined by Eq. (4). The corresponding recovery constant is $C^s_{m=1} = 1.13$ (Eq. (5)). We obtain the 16 solution candidates by assembling $f^s_{m=1}$ and $f^s_{m=2}$, listed in Table 1, according to the tree diagram (Fig. 3). To select the appropriate expression, the solution of a pure bending problem is introduced based on the following considerations: the three-point-bend beam will be reduced to a pure bending problem when the span-to-depth ratio, β, is large. Theoretically, the generalizable solution should be able to recover the solution to a pure bending problem. Meanwhile, Benthem and Koiter proposed the stress intensity factor for a short crack to be $\overline{K_I} = \frac{K_I}{\sigma\sqrt{\pi a}} = 1.12$ (Benthem and Koiter, 1973). We calculate $\overline{K_I}$ for each expression we obtained at $\alpha = 0.01$ and $\beta = 100$ (approximating a short crack in a slender beam) and select the optimal one by comparing $\overline{K_I}$ with the prefactor 1.12. An extended parameter space $\Lambda^g = [0.01, 0.7] \times [4, 100]$ is used to compare the generalization performance of the symbolic regression results. For brevity, a portion of the candidate list with E^{train} lower than 3% is presented in Table 1:

Table 1
Performance of candidate solutions for the SENB configuration.

Candidate Number	$f^s_1(\alpha)$	$f^s_2(\beta)$	E^{train}	E^g	$\overline{K_I}(0.01, 100)$
1	$0.38\alpha^{-1}\ e^{\alpha(3.14\alpha\ -\ 1)}$	$e^{e^{\cos\left(\sin\left(\frac{\beta}{\beta+1}\right)\right)}} - 6.97$	0.89%	53.80%	1.69
2	$0.58\alpha^{-1}\sqrt{e^{\cos\ (2.70e^{\alpha})}}$	$-11.07\sin\left(e^{\frac{\cos(\pi+1)}{\beta}} - 1\right)$	2.07%	14.71%	1.09
3	$\frac{0.38e^{\alpha(\pi\alpha-1)}}{\alpha}$	$-11.07\sin\left(e^{\frac{\cos(\pi+1)}{\beta}} - 1\right)$	1.50%	7.05%	1.12
4	$0.38\alpha^{-1}\ e^{\alpha(3.14\alpha\ -\ 1)}$	$e^{\frac{1}{\sqrt{\beta-\frac{1}{\pi}}}+1} - 3.18$	0.94%	403.00%	− 3.32

The final solution is the third formula whose $\overline{K_I}(\ 0.01, 100\ ) = 1.12$. The expression is given by:

$$K_I(a,b,s,P) = \frac{3Ps^2a^{\frac{3}{2}}}{2b^4}\widetilde{K}_I\left(\frac{a}{b},\frac{s}{b}\right) \tag{9}$$

where

$$\widetilde{K}_I(\alpha,\beta) = -3.7194\left(\frac{e^{\alpha(\pi\alpha-1)}}{\alpha}\right)\sin\left(e^{\frac{\cos(\pi+1)}{\beta}} - 1\right) \tag{10}$$

We compare the symbolic regression results with existing closed-form expressions for K_I derived for this specific problem by the mechanics community in Table 2.

Gross and Srawley (1965) developed a solution for a specific three-point bend specimen with $\beta = 4$. Thus, we compare the symbolic regression result $\overline{K_I}(\alpha, \beta = 4)$ with Srawley's formula $\overline{K_I}^{Srawley}$ in the range $\alpha \in [0.1, 0.6]$, finding the maximum difference to be less than 2% (Fig. 4a). Brown and Srawley provided a solution for three-point bend beam specimen with $\beta = 8$ for any $\alpha \leq 0.6$ (Brown and Srawley, 1966). Similarly, we compared the outcome of symbolic regression $\overline{K_I}(\alpha, \beta = 8)$ and Brown's solution $\overline{K_I}^{Brown}$ within the range of $\alpha \in [0.1, 0.6]$, showing significant agreement with a maximum difference of 0.71% (Fig. 4b). Guinea *et al.* (Guinea et al., 1998), derived a general solution by interpolating existing solutions for $\beta = 4$ and $\beta = \infty$. This provided formula claims to fit a large parameter space, where $\alpha \in (0, 0.7]$ and $\beta \in [4, \infty)$. Considering the use of $\alpha = 0.01$ to denote a short crack and $\beta = 100$ to approximate a long beam, we utilize Λ^g to test the generalization performance of the mechanics-informed symbolic regression solution. This generalized parameter space is much larger than the predefined training parameter space, which is illustrated in Fig. 4c. The mechanics-informed symbolic regression solution shows significant consistency with Guinea's solution even in the extended dataset, with a maximum difference less than 10%. In traditional symbolic regression procedure, the best-fit symbolic regression expression would be selected, like the formula shown in Table 1 with $E^{train} = 0.89\%$. However, this overfitting result exhibits a dramatical deviation when applied to the larger parameter space, resulting in $E^g = 53.80\%$. In general, the mechanics-informed symbolic regression can provide accurate prediction of K_I for a wide range of α and β, while Srawley's and Brown's approaches are constraint to a fixed β. In contrast to Guinea's solution, symbolic regression can extract the solution directly from the raw data without relying on pre-existing analytical solutions. These outcomes demonstrate strong consistency between the mechanics-informed symbolic regression findings and established solutions, and the mechanics-informed steps significantly improve the generalization performance of symbolic regression.

4. Case 2: Fracture of Crackline-Loaded Edge-Crack (CLEC) specimen

We investigate the fracture of a single edge notched cantilever beam with a non-uniform depth (Fig. 5a). The beam is loaded at the edges of crack surfaces with a pair of line forces P. The notch size is a, and the slope of the top and bottom surfaces is m. The minimum depth is d, which is located at the left end of the cantilever. This is a 2D plane stress model of a thin plate CLEC specimen (Gross and Srawley, 1967). The mode-I stress intensity factor at the crack tip is K_I. We use mechanics-informed symbolic regression to explore the correlation between K_I and $\{d, a, m, P\}$:

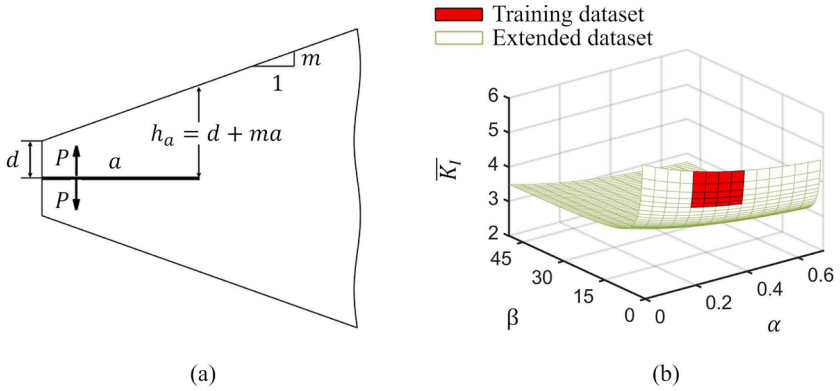
$$K_I = K_I(m, a, d, P) \tag{11}$$

Similarly to the first case (Section 3), we build a 2D FEM model and employ the J-integral to calculate K_I for a given set of $\{d,a,m,P\}$. Based on our mechanics insights, we replace d with the beam depth at the crack tip, $h_a = d + ma$, since the crack is more sensitive to h_a than d, especially for a large a. We apply the Buckingham π theorem to obtain three dimensionless variables: the slope $\alpha = m$, the crack length to crack tip depth ratio $\beta = \frac{a}{h_a}$, and the normalized stress intensity factor $\widetilde{K_I} = \frac{K_I h_a^{\frac{1}{2}}}{P}$. After the mechanics-informed feature engineering, the desired general form of $\widetilde{K_I}$ can be written as:

$$\widetilde{K_I} = \widetilde{K_I}(\alpha, \beta) \tag{12}$$

Table 2
Closed-form expressions for K_I obtained through different approaches.

Approach	$K_I(a,b,s,P)$
Srawley's (Gross and Srawley, 1965) ($s = 4b$)	$\frac{6Pa^{\frac{1}{2}}}{b}\left(\frac{1.99 - \frac{a}{b}\left(1-\frac{a}{b}\right)\left(2.15 - \frac{3.93a}{b} + 2.7\left(\frac{a}{b}\right)^2\right)}{\left(1+\frac{2a}{b}\right)\left(1-\frac{a}{b}\right)^{\frac{3}{2}}}\right)$
Brown's (Brown and Srawley, 1966) ($s = 8b$)	$\frac{12P}{b}(\pi a)^{\frac{1}{2}}\left(1.106 - 1.552\left(\frac{a}{b}\right) + 7.71\left(\frac{a}{b}\right)^2 - 13.53\left(\frac{a}{b}\right)^3 + 14.23\left(\frac{a}{b}\right)^4\right)$
Model-free	$\frac{-5.58Ps^2a^{\frac{1}{2}}}{b^3}\left(e^{\frac{a}{b}\left(\frac{\pi a}{b}-1\right)}\right)\sin\left(e^{\frac{b\cos(\pi+1)}{s}} - 1\right)$



(a) (b)

Fig. 5. (a) A thin plate CLEC specimen under splitting load P. (b) The result obtained from the training dataset (red) can be generalized in the extended dataset (green), where the maximum relative error of the obtained expression, E^g, is lower than 1%.

In practice, the range of $[0, 0.6]$ for α and values of $\beta \geq 1$ are of engineering interest (Gross and Srawley, 1967). We utilize the predefined space of raw input parameters $\{\alpha, \beta\}$, which is denoted by $\Lambda = [0.2, 0.4] \times [1, 2]$ containing $5 \times 11 = 55$ uniformly spaced data points. The multiplicative separability index is $E^{mult} = 0.06\%$ and is lower than the additive separability index (E^{add}). Thus, the original dataset is separated into two fully separated datasets $\{X_1^s, Y_1^s\}$ and $\{X_2^s, Y_2^s\}$ following the multiplicative rule. The recovery constant is $C_1^s = 6.55$. The candidate expressions are structured according to the tree diagram depicted in Fig. 3. There are 14 formulas in total and all of them fit well in the training dataset with E^{train} lower than 3%. To select a generalizable solution, we consider the special case of a notched double cantilever beam (DCB) with uniform depth ($\alpha = 0$). A classical solution to the stress intensity factor of a uniform double cantilever beam with a large β, is $\overline{K_I}^{DCB} = \frac{\widetilde{K_I}^{DCB}}{\beta} = 2\sqrt{3}$. Then we calculate $\overline{K_I}(\alpha = 0, \beta = \infty) = \frac{\widetilde{K_I}(\alpha=0,\beta=\infty)}{\beta}$ for each expression we obtained (this indicates a DCB with an infinitely long crack) and select the best fitting one by comparing it with $\overline{K_I}^{DCB}$. We utilize the extended parameter space $\Lambda^g = [0, 0.7] \times [1, 50]$, containing 1485 uniformly spaced data points to test the generalization performance of the mechanics-informed symbolic regression solution. The comparison between the training dataset and the extended dataset is illustrated in Fig. 5b. A portion of the candidate list is presented in Table 3:

The first formula gives $\overline{K_I}(0, \infty) = 3.4509$ which is close to $\overline{K_I}^{DCB}$. Thus, the final solution is given by:

$$K_I(m, a, d, P) = \frac{P}{\sqrt{d + ma}} \widetilde{K_I}\left(m, \frac{a}{d + ma}\right) \tag{13}$$

where

$$\widetilde{K_I}(\alpha, \beta) = 1.5982 \log\left(\frac{\pi + 1}{\cos((\sin(\alpha) - 1))} + 1\right)\left(\beta + \sin\left(\cos\left(\frac{1}{\sqrt{\pi - 1}}\right)\right)\right) \tag{14}$$

Although candidate solutions for the first and third formulas in Table 3 fit the training dataset well, with E^{train} less than 0.01%, the mechanics-informed symbolic regression result fits the extended dataset significantly better, with $E^g = 0.38\%$. In contrast, the third formula, which does not meet the mechanics criterion in selection, shows a significant increase in $E^g = 9.60\%$ in the extended dataset. By applying the mechanics criterion, we can avoid such plausible but less generalizable solutions.

Table 3
The performance of candidate solutions for the CLEC configuration.

Candidate Number	$f_1^s(\alpha)$	$f_1^s(\beta)$	E^{train}	E^g	$\overline{K_I}(0, \infty)$
1	$3.52 \log\left(\frac{\pi + 1}{\cos((\sin(\alpha) - 1))} + 1\right)$	$2.98\left(\beta + sin\left(cos\left(\frac{1}{\sqrt{\pi - 1}}\right)\right)\right)$	0.0002%	0.38%	3.4509
2	$10.94 log\left(log\left(\alpha + e^{\frac{1}{\alpha+1}+1}\right)\right)$	$2.08 + \beta + \frac{\beta}{e^{\pi^{-1}-1}}$	0.0033%	0.40%	3.4490
3	$\frac{16.68}{\sqrt{\alpha} + 2}$	$2.98\left(\beta + sin\left(cos\left(\frac{1}{\sqrt{\pi - 1}}\right)\right)\right)$	0.01%	9.60%	3.7919
4	$5.02 + \alpha - log(\alpha)$	$2.08 + \beta + \frac{\beta}{e^{\pi^{-1}-1}}$	0.44%	9.48%	NaN

5. Discussion and conclusions

5.1. Convergence of mechanics-informed symbolic regression with respect to dataset size

Sufficient data to obtain a convergent expression is key for using the proposed framework. We start from a small dataset and gradually increase its size until the expressions converge. For both examples, we investigated the influence of size of training datasets on convergence, by utilizing refined training datasets (See Appendix C). In the SENB case, we obtained the same solution (Eq. (10)) using a dataset consisting of 99 data points and a refined dataset comprising 1071 data points. In the CLEC case, we initially used a dataset of 55 data points and confirmed the convergence of the expression using a refined dataset of 189 data points. Starting with a smaller dataset and progressively increasing its size until the result converges is an effective method to determine the minimum number of data points needed for symbolic regression.

5.2. High-fidelity variable separation using a model-free approach

The model-free separation algorithm constructs sub-datasets directly from the original data points without invoking any specific machine learning model. Such an approach preserves the fidelity of the FEM dataset and eliminates the uncertainty in the variable separation introduced by model-based methods. For example, in NN-based approaches, the inherent training error, the variation of NN structures, and the choice of different hyperparameters can result in loss of fidelity and in uncertainty during variable separation. Both are detrimental to the subsequent symbolic regression steps and impede the identification of the generalizable solution (more information in Appendix D). To separate the training dataset, we compared our model-free with a NN-based approach (Udrescu and Tegmark, 2020). For the SENB specimen case, both approaches identified the same multiplicative separability, yet the NN-based approach introduced an additional $\sim 5\%$ error to the separated datasets (in contrast to the zero error in the model-free approach). The generalizability of the resulting expression in the case of the NN-based approach is poor, exhibiting $\sim 16\%$ error in the extended dataset, which is over twice the error of the expression obtained through the model-free approach. In the CLEC case, the separability identified through the NN-based approach is highly sensitive to the selected hyperparameters of the NNs, resulting in large uncertainty in the variable separation. The NN-based approach introduced at least $\sim 3\%$ error in separated sub-datasets, leading to an expression with inferior generalizability ($\sim 8\%$), over 20 times the error obtained through the model-free approach. The model-free approach demonstrates superior performance over model-based methods. The advantage of the model-free approach for variable separation is particularly evident when the raw data is of high quality (*e.g.*, noise-free FEM data).

5.3. Decoupling complex problems by embedding mechanics knowledge

Variable separation is an important step in using symbolic regression to solve fracture mechanics problems. When variables are separable, we can directly identify the candidate expressions using the raw data obtained from FEA simulations. However, fracture problems typically contain coupled variables. To decouple a fracture mechanics problem, we employ mechanics-informed feature engineering to identify intermediate variables that are separable. These intermediate variables can help simplify the original complex symbolic regression task into several sub-tasks that contain less variables. For the SENB case, we examined the separability of the original dataset, $\{\frac{a}{b},\ \frac{s}{b},\ \frac{K_I\sqrt{b}}{P}\}$, and found that the variables are highly coupled, as both the multiplicative and additive separability indices exceed 30%. Therefore, we constructed a new dataset, $\{\alpha = \frac{a}{b},\ \beta = \frac{s}{b},\ \widetilde{K}_I = \frac{K_I b^2}{\sigma_{avg} s a^{\frac{3}{2}}}\}$, by introducing the intermediate variable σ_{avg}, which is the approximate average stress at the crack tip. The multiplicative separability index of the newly constructed dataset is less than 1%, allowing us to fully decouple the symbolic regression task. Introducing intermediate variables can decouple complex mechanics problems which require domain knowledge. For problems relative to stress intensity factors and energy release rates, the average or maximum stress near the crack tip can be selected as intermediate variables. For a complex geometry, we can identify the potential intermediate variables by examining the solution to the geometry's simplified counterpart. For example, a uniform DCB configuration is a special case of CLEC configuration for $m = 0$ (Fig. 5). The energy release rate of a uniform DCB under splitting force is linearly proportional to $\left(\frac{a}{h}\right)^2$, where h is the depth at the crack tip. Therefore, we choose $\beta = \frac{a}{h_a}$, where $h_a = d + ma$, as the intermediate variable for the CLEC problem, which is found to be separable. For problems involving complex constitutive laws, we can identify intermediate variables from the solutions to simpler models, such as linear or nonlinear elastic models, similar to the approach of Hutchinson, Rice, and Rosengren (HRR) to derive the crack-tip singularities in an elastoplastic material (Rice and Rosengren, 1968; Hutchinson, 1968).

5.4. Selecting a generalizable expression by applying mechanics knowledge

Conventional symbolic regression algorithms cannot guarantee the generalizability of their results. We emphasize the importance of generalizability, as a generalizable solution can offer insights into the underlying mechanics and principles of the problem under investigation. In our framework, we enhance the generalizability of results by using a mechanics-informed selection step (Fig. 1). We apply mechanics knowledge to develop a selection criterion that allows for identifying the most generalizable expression from a large pool of candidates. In this work, we examine the candidate solutions by comparing them with established solutions to simplified

problems. For example, in the SENB case, the symbolic regression task is to identify the stress intensity factor solution for a finite beam with a finite crack. The simplified version of this problem consists of a slender beam with a short crack, whose parameter space is not covered by the training dataset of the symbolic regression task. The solution to the original problem needs to be consistent with the solution to the simplified problem. We employ consistency as the criterion to evaluate the generalizability of the candidate solutions. In the CLEC case, the symbolic regression task is to identify the stress intensity factor solution for a non-uniform CLEC beam under splitting forces. We consider a reduced problem of a uniform DCB with a long crack whose solution is well established. Similarly, the generalizable solution to the CLEC problem should be consistent with the solution to the reduced problem. We employ this criterion to select the appropriate expression (Eqs. (10) and (14)). The mechanics criterion based on established solutions effectively enhances the generalizability of the selected expression.

5.5. Potential of the framework in fracture mechanics and beyond

The proposed framework holds potential for advancing fracture mechanics by enabling efficient evaluation of fracture- and fatigue-related properties from unconventional test specimen configurations (Athanasiou et al., 2023; Athanasiou and Bellouard, 2015; Athanasiou et al., 2017; Nazir et al., 2022). Such evaluations are currently unattainable due to the complexity of extracting properties from experimental datasets and the heavy reliance on complicated FEM solutions. By providing easy-to-use analytical formulas, this framework could bypass the need for FEM simulations, and takes a crucial step toward democratizing fracture testing, especially with novel test specimen configurations often used in small-scale mechanical testing (Athanasiou et al., 2024). Additionally, its ability to generate generalizable solutions from limited datasets can lead to substantial resource savings, particularly in experimental testing at challenging scales or under demanding environmental conditions.

Author Statement

We the undersigned declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We understand that the Corresponding Author is the sole contact for the Editorial process. He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs.

CRediT authorship contribution statement

Ruibang Yi: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Dimitrios Georgiou:** Writing – review & editing, Formal analysis. **Xing Liu:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Formal analysis, Conceptualization. **Christos E. Athanasiou:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

R.Y., D.G., and C.E.A. acknowledge support from the National Science Foundation (NSF) CAREER Award [CMMI-2338508]. The simulations were conducted using the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA.

Appendix A. The criterion for selecting the separability threshold E^t

The separability threshold, E^t, is a parameter that determines the maximum separability index at which a variable or a set of variables should be considered separable. A larger E^t means we allow the algorithm to separate variables with a higher separability index during model-free variable separation (Fig. 1). In the model-free separability check, both additive and multiplicative separability indices will be calculated (Eqs. (1) and (2)) and sorted in ascending order $E^s = \{E_1^s, E_2^s, \dots\}$. We can start by choosing $E^t = 0$, and check if the E^{train} (Eq. (6b)) of the expressions obtained using this E^t is lower than a pre-defined acceptable error (3%). When the expression obtained by the selected E^t cannot satisfy the pre-defined acceptable error, we need to increase E^t to the next element in E^s, *i.e.* $E^t = E_1^s$, and repeat the process until the pre-defined acceptable error is satisfied. In the SENB case (Section 3), the separability indices are $E^s = \{0.06\%, 1.71\%\}$. We start from $E^t = 0$, which means only variables with $E^{add} = 0$ or $E^{mult} = 0$ will be separated. The minimum

separability index is $E_1^s = E^{mult} = 0.06\% > E^t = 0$ therefore no separability was identified in the input dataset. Consequently, the polynomial/ non-polynomial fitting are performed on the original input dataset $\{\alpha, \beta, \widetilde{K}_I\}$. We did not obtain an expression with E^{train} lower than the predetermined value of 3%, while the expression with the minimum E^{train} (5.67%) is:

$$\widetilde{K}_I(\alpha, \beta) = \frac{5.57}{\sin(\pi\alpha)(\alpha + \beta)} \tag{15}$$

Thus, we increase E^t to 0.06% (*i.e.* E_1^s), to further separate the original dataset. Some expressions with $E^{train} < 3\%$ are listed in Table 1.

Appendix B. Polynomial and non-polynomial fitting

To obtain symbolic expressions describing the given dataset, we use both a gradient descent algorithm (polynomial fitting) and a brute-force algorithm (non-polynomial fitting). The gradient descent algorithm is used to address the limitations of the brute-force algorithm, as brute-force methods can be computationally expensive when discovering high-order polynomials. In general, these fitting algorithms contain two parts: defining the search space of candidate expressions and finding the optimal candidate expressions.

The search space should include the potential solution to the original problem while being as narrow as possible to enhance search efficiency. For the polynomial fitting, we construct a search space containing polynomials up to degree 3*N*, where *N* is the total number of variables. For the non-polynomial fitting, we construct a search space including expressions consisting of compositions of 14 commonly used pre-defined functions and operators (Udrescu and Tegmark, 2020). This search space performs well when tasked to discover physically meaningful symbolic expressions (Udrescu and Tegmark, 2020; Keren et al., 2023). For different problems, the best choice of the search space can be different. For example, Kaheman *et al.* (Kaheman et al., 2020) constructed a search space using a library of candidate functions derived from measured trajectory data in nonlinear dynamics and Zhang *et al.* (Zhang et al., 2023) constructed a search space using fixed equation structures to more effectively discover parametric equations.

To find the optimal expressions in a pre-defined search space, in the polynomial fitting, we applied a gradient-descent algorithm with a learning rate of 0.001 over 10,000 iterations to obtain the polynomial, of degree under 3*N*, with the least fitting error. Regarding the non-polynomial fitting, we enumerate possible operators and functions according to the complexity order defined in (Udrescu and Tegmark, 2020), until the preset time runs out.

Appendix C. Symbolic regression results using different training dataset sizes

In the SENB case, we constructed the initial training dataset within the predefined parameter space $\Lambda = [0.1, 0.6] \times [4, 6]$, consisting of 99 data points uniformly distributed across 11×9 grid points. Subsequently, a refined dataset, also within the same space Λ, was developed with a finer mesh of 51×21 points, yielding a total of 1071 uniformly distributed data points. We found the same multiplicative separability for both the initial and refined datasets, and they were separated accordingly. We obtained a large number of candidate expressions through polynomial/non-polynomial fitting on the separated datasets. We identified the most generalizable solution using the same mechanics criterion (Section 3). The selected solution appeared to be the same as the one obtained from the initial dataset (Eq. (10)).

In the CLEC case, we began with the initial training dataset defined on the parameter space $= [0.2, 0.4] \times [1, 2]$, comprising $5 \times 11 = 55$ uniformly distributed data points. The refined dataset was collected from a denser 9×21 grid, resulting in 189 data points. We identified the same multiplicative separability in both the initial dataset and the refined one and separated them accordingly (Eqs. (4) and (5)). Through the polynomial/non-polynomial fitting, we obtained more candidate expressions from the refined training dataset. The most generalizable expression, selected according to the mechanics criterion established in Section 4, was identical to that obtained from the initial training dataset (Eq. (14)).

Appendix D. Symbolic regression performance using model-free and NN-based separation approaches.

In the SENB case, we employed the model-free and a NN-based approach to examine the separability of the same training dataset consisting of 99 data points. Both identified the multiplicative separability. The model-free approach was found to preserve the fidelity of dataset. However, in the case of the NN-based approach, we observed a 4.72% error when reconstructing the original dataset from the separated dataset, indicating a significant loss of fidelity. Subsequently, we applied the polynomial/non-polynomial fitting algorithm (Section 2.2.3) to the datasets separated using the NN-based approach, assembled the expressions for separated datasets, and selected the most generalizable expression according to the mechanics criterion established in Section 3. This expression obtained through the NN-based approach exhibits a significantly larger E^{train} error of 8.03% in the training dataset compared to the model-free approach. Additionally, the generalizability of the NN-based expression is inferior to that of the model-free expression. The former showed an E^g error of 16.55%, while the latter only 7.05% (Table B1).

Table B1
Solving the SENB problem using the model-free and the NN-based separation approaches.

Variable separation approach	Model-free	NN-Based
Error in the separated datasets	0%	4.72%
Selected expressions	*Eq.* (10)	$\widetilde{K}_I(\alpha,\beta) = \left(-0.57 + \frac{e^{(\alpha(\alpha+1))-1}}{\alpha}\right)\left(\frac{5.99}{\beta+1}\right)$
E^{train}	1.50%	8.03%
E^{g}	7.05%	16.55%

In the CLEC case, we employed the model-free and NN-based approaches to examine the separability of the same training dataset consisting of 55 data points. The model-free approach identified the multiplicative separability while the NN-based approach gave different separability depending on the selected hyperparameters of the NNs. The NN-based approach introduced 2.59% error to the separated sub-datasets while the mode-free approach preserved the data fidelity. Subsequently, we applied the polynomial/non-polynomial fitting (Section 2.2.3) to the sub-datasets separated through the NN-based approach and the mechanics criterion established in Section 4 to select the most generalizable solution from assembled candidate solutions. The most generalizable expression obtained through the NN-based approach fits well the training dataset with $E^{train} = 2.59\%$, while the model-free solution (Eq. (14)) has an error lower than 1%. We introduced the extended dataset within $\Lambda^g = [0, 0.7] \times [1, 50]$ to identify the generalizability of both solutions. Due to the error introduced in separated sub-datasets, the NN-based solution shows poor generalizability compared with the model-free solution, whose E^g values are 7.62% and 0.38% respectively (Table B2).

Table B2
Solving the CLEC problem using the model-free and the NN-based separation approaches.

Variable separation approach	Model-free	NN-Based
Error in the separated datasets	0%	2.59%
Selected expressions	*Eq.* (14)	$\widetilde{K}_I(\alpha,\beta) = 11.5\cos(\sin(\alpha+1))\left(0.36 + \beta e^{\sin(pi+1)}\right)$
E^{train}	0.0002%	2.59%
E^{g}	0.38%	7.62%

Data availability

Data will be made available on request.

References

Agyei, R., Hanhan, I., Sangid, M.D., 2023. A data-driven microstructural rationale for micro-void nucleation in discontinuous fiber composites. J. Thermoplast. Compos. Mater. 36 (4), 1694–1716. https://doi.org/10.1177/08927057211068734.

Athanasiou, C.E., Bellouard, Y., 2015. A monolithic micro-tensile tester for investigating silicon dioxide polymorph micromechanics, fabricated and operated using a femtosecond laser. Micromachines 6 (9), 1365–1386. https://doi.org/10.3390/mi6091365.

Athanasiou, C.E., Fincher, C.D., Gilgenbach, C., Gao, H., Carter, W.C., Chiang, Y.M., Sheldon, B.W., 2024. Operando measurements of dendrite-induced stresses in ceramic electrolytes using photoelasticity. Matter 7 (1), 95–106. https://doi.org/10.1016/j.matt.2023.10.014.

Athanasiou, C.E., Hongler, M.O., Bellouard, Y., 2017. Unraveling brittle-fracture statistics from intermittent patterns formed during femtosecond laser exposure. Phys. Rev. Appl. 8 (5), 054013. https://doi.org/10.1103/PhysRevApplied.8.054013.

Athanasiou, C.E., Liu, X., Gao, H., 2024. A Perspective on Democratizing Mechanical Testing: harnessing Artificial Intelligence to Advance Sustainable Material Adoption and Decentralized Manufacturing. J. Appl. Mech. 91 (11), 11080. https://doi.org/10.1115/1.4066085.

Athanasiou, C.E., Liu, X., Zhang, B., Cai, T., Ramirez, C., Padture, N.P., Lou, J., Sheldon, B.W., Gao, H., 2023. Integrated simulation, machine learning, and experimental approach to characterizing fracture instability in indentation pillar-splitting of materials. J. Mech. Phys. Solids 170, 105092. https://doi.org/10.1016/j.jmps.2022.105092.

Bahmani, B., Sun, W.C., 2024. Physics-constrained symbolic model discovery for polyconvex incompressible hyperelastic materials. Int. J. Numer. Methods Eng. e7473. https://doi.org/10.1002/nme.7473.

Benthem, J.P., Koiter, W.T., 1973. Asymptotic approximations to crack problems. In: Mechanics of fracture: methods of analysis and solutions of crack problems, 1. Springer, Dordrecht, pp. 131–178. https://doi.org/10.1007/978-94-017-2260-5_3.

Bessa, M.A., Bostanabad, R., Liu, Z., Hu, A., Apley, D.W., Brinson, C., Chen, W., Liu, W.K., 2017. A framework for data-driven analysis of materials under uncertainty: countering the curse of dimensionality. Comput. Methods Appl. Mech. Eng 320, 633–667. https://doi.org/10.1016/j.cma.2017.03.037.

Bomarito, G.F., Townsend, T.S., Stewart, K.M., Esham, K.V., Emery, J.M., Hochhalter, J.D., 2021. Development of interpretable, data-driven plasticity models with symbolic regression. Comput. Struct. 252, 106557. https://doi.org/10.1016/j.compstruc.2021.106557.

Briñez-de León, J.C., Rico-García, M., Restrepo-Martínez, A., 2022. PhotoelastNet: a deep convolutional neural network for evaluating the stress field by using a single color photoelasticity image. Appl. Opt. 61 (7), D50–D62. https://doi.org/10.1364/AO.444563.

Brown, W., Srawley, J., 1966. Plane strain crack toughness testing of high strength metallic materials. In Plane strain crack toughness testing of high strength metallic materials. ASTM International 129. https://doi.org/10.1520/STP44663S.

Buehler, M.J., 2024. MechGPT, a language-based strategy for mechanics and materials modeling that connects knowledge across scales, disciplines, and modalities. Appl. Mech. Rev. 76 (2), 021001. https://doi.org/10.1115/1.4063843.

Chen, C.-T., Gu, G.X., 2023. Physics-informed deep-learning for elasticity: forward, inverse, and mixed problems. Adv. Sci. 10, 2300439. https://doi.org/10.1002/advs.202300439.

Davidson, J.W., Savic, D.A., Walters, G.A., 2003. Symbolic and numerical regression: experiments and applications. Inf. Sci. 150 (1–2), 95–117. https://doi.org/10.1016/S0020-0255(02)00371-7.
Dekhovich, A., Tax, D.M.J., Sluiter, M.H.F., Bessa, M.A., 2024. Neural network relief: a pruning algorithm based on neural activity. Machine Learning 113, 2597–2618. https://doi.org/10.1007/s10994-024-06516-z.
Evans, J.H., 1972. Dimensional analysis and the Buckingham Pi theorem. Am. J. Phys. 40 (12), 1815–1822. https://doi.org/10.1119/1.1987069.
Flaschel, M., Kumar, S., De Lorenzis, L., 2022. Discovering plasticity models without stress data. npj Comput. Mater. 8, 91. https://doi.org/10.1038/s41524-022-00752-4.
Fuhg, J.N., Jones, R.E., Bouklas, N., 2024. Extreme sparsification of physics-augmented neural networks for interpretable model discovery in mechanics. Comput. Methods Appl. Mech. Eng. 426, 116973. https://doi.org/10.1016/j.cma.2024.116973.
Goswami, S., Yin, M., Yu, Y., Karniadakis, G.E., 2022. A physics-informed variational DeepONet for predicting crack path in quasi-brittle materials. Comput. Methods Appl. Mech. Eng. 391, 114587. https://doi.org/10.1016/j.cma.2022.114587.
Gross, B., Srawley, J.E., 1965. Stress-intensity factors for single-edge-notch specimens in bending or combined bending and tension by boundary collocation of a stress function. NASA Technical Note, TN. D-3092.
Gross, B., Srawley, J.E., 1967. Stress intensity factors for crackline-loaded edge-crack specimens. NASA Technical Note, TN. D-3820.
Gu, G.X., Chen, C.-T., Buehler, M.J., 2018. De novo composite design based on machine learning algorithm. Extreme Mech. Lett. 18, 19–28. https://doi.org/10.1016/j.eml.2017.10.001.
Guinea, G.V., Pastor, J.Y., Planas, J., Elices, M., 1998. Stress intensity factor, compliance and CMOD for a general three-point-bend beam. Int. J. Fract. 89, 103–116. https://doi.org/10.1023/A:1007498132504.
Hutchinson, J.W., 1968. Singular behaviour at the end of a tensile crack in a hardening material. J. Mech. Phys. Solids 16 (1), 13–31. https://doi.org/10.1016/0022-5096(68)90014-8.
Kaheman, K., Kutz, J.N., Brunton, S.L., 2020. SINDy-PI: a robust algorithm for parallel implicit sparse identification of nonlinear dynamics. Proc Math Phys Eng Sci 476 (2242), 20200279. https://doi.org/10.1098/rspa.2020.0279.
Karapiperis, K., Kochmann, D.M., 2023. Prediction and control of fracture paths in disordered architected materials using graph neural networks. Commun. Eng. 2, 32. https://doi.org/10.1038/s44172-023-00085-0.
Karapiperis, K., Stainier, L., Ortiz, M., Andrade, J.E., 2021. Data-Driven multiscale modeling in mechanics. J. Mech. Phys. Solids 147, 104239. https://doi.org/10.1016/j.jmps.2020.104239.
Keren, L.S., Liberzon, A., Lazebnik, T., 2023. A computational framework for physics-informed symbolic regression with straightforward integration of domain knowledge. Scientific Reports 13 (1), 1249. https://doi.org/10.1038/s41598-023-28328-2.
Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA.
LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. In: Proc. IEEE, 86, pp. 2278–2324. https://doi.org/10.1109/5.726791.
Liu, B., Ocegueda, E., Trautner, M., Stuart, A.M., Bhattacharya, K., 2023. Learning macroscopic internal variables and history dependence from microscopic models. J. Mech. Phys. Solids 178, 105329. https://doi.org/10.1016/j.jmps.2023.105329.
Liu, X., Athanasiou, C.E., Padture, N.P., Sheldon, B.W., Gao, H., 2020. A machine learning approach to fracture mechanics problems. Acta Mater 190, 105–112. https://doi.org/10.1016/j.actamat.2020.03.016.
Liu, X., Athanasiou, C.E., Padture, N.P., Sheldon, B.W., Gao, H., 2021. Knowledge extraction and transfer in data-driven fracture mechanics. Proc. Natl. Acad. Sci. U. S. A 118 (23), e2104765118. https://doi.org/10.1073/pnas.2104765118.
Lu, L., Dao, M., Kumar, P., Ramamurty, U., Karniadakis, G.E., Suresh, S., 2020. Extraction of mechanical properties of materials through deep learning from instrumented indentation. Proc. Natl. Acad. Sci. U. S. A. 117 (13), 7052–7062. https://doi.org/10.1073/pnas.1922210117.
McMillen, B., Athanasiou, C., Bellouard, Y., 2016. Femtosecond laser direct-write waveplates based on stress-induced birefringence. Opt. Express 24 (24), 27239–27252. https://doi.org/10.1364/oe.24.027239.
Mozaffari, K., Ahmadpoor, F., Deng, Q., Sharma, P., 2023. A minimal physics-based model for musical perception. Proc. Natl. Acad. Sci. U. S. A. 120 (5), e2216146120. https://doi.org/10.1073/pnas.2216146120.
Nazir, S.I., Athanasiou, C.E., Bellouard, Y., 2022. On the behavior of uniaxial static stress loaded micro-scale fused silica beams at room temperature. J. Non-Cryst Solids: X 14, 100083. https://doi.org/10.1016/j.nocx.2022.100083.
Niu, S., Srivastava, V., 2022. Ultrasound classification of interacting flaws using finite element simulations and convolutional neural network. Engineering with Computers 38, 4653–4662. https://doi.org/10.1007/s00366-022-01681-y.
Pantidis, P., Mobasher, M.E., 2023. Integrated Finite Element Neural Network (I-FENN) for non-local continuum damage mechanics. Comput. methods Appl. Mech. Eng 404, 115766. https://doi.org/10.1016/j.cma.2022.115766.
Paris, P., Erdogan, F., 1963. A critical analysis of crack propagation laws. J. Basic Eng. Trans. 85 (4), 528–534. https://doi.org/10.1115/1.3656900.
Paris, P., Gomez, M., Anderson, W., 1961. A rational analytic theory of fatigue. Trends Eng 13, 9–14.
Prume, E., Reese, S., Ortiz, M., 2023. Model-free data-driven inference in computational mechanics. Comput. methods Appl. Mech. Eng 403 (Part A), 115704. https://doi.org/10.1016/j.cma.2022.115704.
Rice, J.R., 1968. A path independent integral and the approximate analysis of strain concentration by notches and cracks. J. Appl. Mech. 35 (2), 379–386. https://doi.org/10.1115/1.3601206.
Rice, J.R., Rosengren, G.F., 1968. Plane strain deformation near a crack tip in a power-law hardening material. J. Mech. Phys. Solids 16 (1), 1–12. https://doi.org/10.1016/0022-5096(68)90013-6.
Udrescu, S.-M., Tegmark, M., 2020. AI Feynman: a physics-inspired method for symbolic regression. Sci. Adv. 6, eaay2631. https://doi.org/10.1126/sciadv.aay2631.
Vlassis, N.N., Sun, W.C., 2021. Component-based machine learning paradigm for discovering rate-dependent and pressure-sensitive level-set plasticity models. J. Appl. Mech. 89 (2), 021003. https://doi.org/10.1115/1.4052684.
Zhang, M., Kim, S., Lu, P.Y., Soljačić, M., 2023. Deep learning and symbolic regression for discovering parametric equations. IEEE Trans. Neural Netw. Learn. Syst. 1–13. https://doi.org/10.1109/TNNLS.2023.3297978.
Zhang, Q., Barri, K., Jiao, P., Salehi, H., Alavi, A.H., 2021. Genetic programming in civil engineering: advent, applications and future trends. Artif. Intell. Rev. 54, 1863–1885. https://doi.org/10.1007/s10462-020-09894-7.