Physics-Informed Machine Learning Enabled Virtual Experimentation for 3D Thermoplastics Printing

Zhenru Chen[1], Yuchao Wu[1], Yunchao Xie[2], Kianoosh Sattari[1], and Jian Lin[1*]

[1]Department of Mechanical and Aerospace Engineering, University of Missouri, Columbia, Missouri 65211, United States

[2]Department of Mechanical and Manufacturing Engineering, Miami University, Oxford, Ohio 45056, United States

*Email: linjian@missouri.edu

Abstract

Performance of 3D printed thermoplastics largely depends on the ink formulation, which is composed of tremendous chemical space as increased number of monomers, making it very difficult to identify an optimum one with desired properties. To tackle this challenge, we demonstrate a virtual experimentation platform that is enabled by a physics-informed machine learning algorithm. As a case study, the algorithm was trained based on a multilayer perceptron (MLP) model to predict the experimental stress-strain curves of the 3D printed thermoplastics given the ink compositions made of six monomers. To solve the issue of experimental data scarcity, we first reduced the dimensions of the curves to eight principal components (PCs), which serve as the outputs of the model. In addition, we incorporated the physics-informed descriptors into the input dataset. These two strategies afford the model with prediction accuracy of R^2 of 0.97 and RMSE values of 1.01 for fracture strength, R^2 of 0.95 and RMSE of 0.40 for toughness. To perform virtual experimentation, the well-trained model was then utilized to predict 100,000 sets of the PCs from the randomly given 100,000 ink formulations. The PC sets were then converted back to the corresponding stress-strain curves. To validate the prediction results, some of the virtual experiments were performed. The results showed a good match between the predicted and experimental curves. This methodology offers a general and efficient pathway to virtual experimentation for establishing the correlation between the complex input variables and the output performance metrics of new materials.

Keywords: physics-informed, machine learning, 3D printing, thermoplastics, virtual experimentation

1. Introduction

Virtual experimentation represents a pivotal advancement in scientific research, enabling extensive pre-experimental screening that refines the scope of physical trials, thus saving resources on the most promising inquiries.[1] Such preliminary simulations are especially critical in fields where experimental setups are costly and time-intensive. A prime example is 3D printing, which offers rapid prototyping and manufacturing capabilities that have become indispensable across industries—from aerospace to healthcare—due to their ability to cost-effectively create objects with complex geometries.[2,3] Despite these advantages, the development and testing of new materials for 3D printing, especially thermoplastics, present significant challenges. The mechanical properties of thermoplastics, crucial for their functional applications, depend heavily on the precise formulation of inks. The complex interactions and subsequent polymerization of various monomers profoundly impact their mechanical properties.[4,5] The traditional experimental process, which involves the exploration of vast ink formulations to pinpoint the desired mechanical properties of 3D printed thermoplastics, requires extensive experimentation. This process becomes particularly laborious and time-consuming as the combinational chemical space dramatically increases. Virtual experimentation provides a significant advantage over traditional methods by allowing researchers to bypass the initial phases of testing, where intuition alone may not suffice to optimize experimental conditions.

Virtual experimentation has often relied on theoretical calculations or computational simulation techniques, such as Density Functional Theory (DFT) and molecular dynamics (MD). These methods have been extensively applied in fields such as materials science[6,7] and chemical engineering[8] to predict properties of materials at various scales. However, these approaches often face challenges in accurately scaling predictions to complex macroscopic phenomena. For instance, in the 3D printing processes, while molecular simulations are adept at modeling the intricate interactions between monomers,[9] they struggle to extend these predictions to the overall mechanical properties of materials. This limitation suggests that alternative approaches, such as data-driven methods that bypass detailed microstructural modeling, may be necessary. Additionally, deriving practical characteristic curves, such as stress-strain (S-S) ones, poses another significant challenge because physics-based simulation results often rely on idealized material systems under conceptualized conditions, which may not accurately reflect real-world behaviors.

In contrast, data-driven algorithms, such as machine learning (ML), have recently emerged as a complementary approach,[10] increasingly pervading the materials science in design,[11,12] property prediction,[13-15] synthesis planning,[16-18] and automated data analysis.[19-21] They are forming a new paradigm for the virtual experimentation.[1] For instance, in predicting material performance, integration of vast datasets with advanced algorithms has enabled more precise and efficient predictions than ever before. By leveraging extensive data obtained from DFT simulations, ML algorithms can now be applied to predict the performance of composites[22,23] and metamaterials[24] with unprecedented accuracy and efficiency. To mitigate the data scarcity issue, physics-informed ML (PIML) by incorporating known physical laws into the ML training have been developed.[5,25,26] This hybrid approach not only enhances prediction accuracy with limited amount of data but also extends the capability of simulations to cover unexplored material systems. For example, our group incorporated the chemical and physical properties of metal salts and organic linkers as physics-informed descriptors to unravel complex synthesis parameters for accurately predicting the crystallization propensity of metal–organic nanocapsules.[12] In our another work, we trained a scientific ML model that includes intermediate reaction variables obtained by simulations for predicting the reaction outcomes.[27] Du and coworkers utilized six mechanistic variables that represent the physics of balling defects to train a ML model for predicting defects formed during the 3D printing processes.[28] Use of PIML in virtual experimentation holds vast potential, particularly in refining the design and optimization processes in 3D printing, where understanding the detailed physical and chemical interactions crossing the multiple scales is often impractical. Despite the vast potentials and recent research progress, in most literature reports that involve ML algorithms for property prediction, typically singular numerical features (e.g. strength and fractural strain) rather than a total performance profile were reported. In our recent work, we employed a multi-objective Bayesian optimization method to identify materials for 3D printing of thermoplastics that are both strong and tough, focusing specifically on optimizing these two singular numerical values.[5] In contrast, the current study utilizes Physics-Informed Machine Learning (PIML) to conduct virtual experiments that simulate the complete mechanical performance of materials, thereby providing a more comprehensive understanding of their behaviors under various conditions.

Herein, to tackle the challenge, we demonstrate a PIML for predicting full stress-strain curves of 3D printed thermoplastics, which serves as an efficient virtual experimentation platform for screening ink formulations that lead to thermoplastics with desired mechanical properties. To realize this goal, a total of 216 S-S curves were first collected from thermoplastics that were 3D printed using six monomers. Then, dimensions of these S-S curves were reduced by principal component analysis (PCA) into eight principal components (PCs). After that, the compositions of the six monomers together with the physics-informed descriptors (including Molecular Weight, Lipophilicity, Hbond Donor/Acceptor, Rotatable Bonds, Polar Surface Area, Heavy Atoms, Complexity, Total Energy and several Solubility scores) serve as the inputs while the corresponding sets of PCs serve as the outputs to train a multiple layer perceptron (MLP) model. Given 100,000 sets of the hypothesized ink compositions, the MLP can predict the new PCs, which were then converted back to the corresponding S-S curves. Among them, the six ink formulations featuring three different types of the mechanical profiles were chosen for experimental validation. The obtained S-S curves from these experiments fell within the ranges predicted by the virtual experiments. Quantitative study shows that the model achieves prediction accuracy with satisfactory R^2 of 0.97 and root mean squared error (RMSE) of 1.01 for fracture strength, R^2 of 0.95 and RMSE of 0.40 for toughness. These results affirm the success of the virtual experimentation for large scale screening, opens a way to designing new thermoplastics with desired properties.

2. Results and Discussion

Workflow. **Figure 1** illustrates the workflow of developing a PIML based virtual experimentation platform for 3D thermoplastics printing. First, 2-Hydroxy-3-phenoxypropyl acrylate (HA), iso-octyl acrylate (IA), N-vinylpyrrolidone (NVP), acrylic acid (AA), N-(2-hydroxyethyl) acrylamide (HEAA) and isobornyl acrylate (IBOA) were selected as the six monomers.[5] This diverse selection was strategically chosen to demonstrate the robustness and adaptability of our machine learning model across a complex chemical space, showcasing the necessity and effectiveness of the proposed virtual experimentation workflow. Then, inks were prepared via mixing these six monomers in different weight ratios for printing by a liquid crystal display (LCD) printer. After that, the S-S curves of the resulting thermoplastics were collected by a tensile testing machine (Mark-10) according to American Society for Testing and Materials standards. The collected curves were preprocessed and reduced in dimensions by PCA detailed as follows. Following this, a multiple layer perceptron (MLP) model was trained by using the ink compositions together with the physics-informed descriptors as the input to predict these dimension-reduced representations. The culmination of this process employed an inverse PCA technique to reconstruct the S-S curves from the predicted PCs.

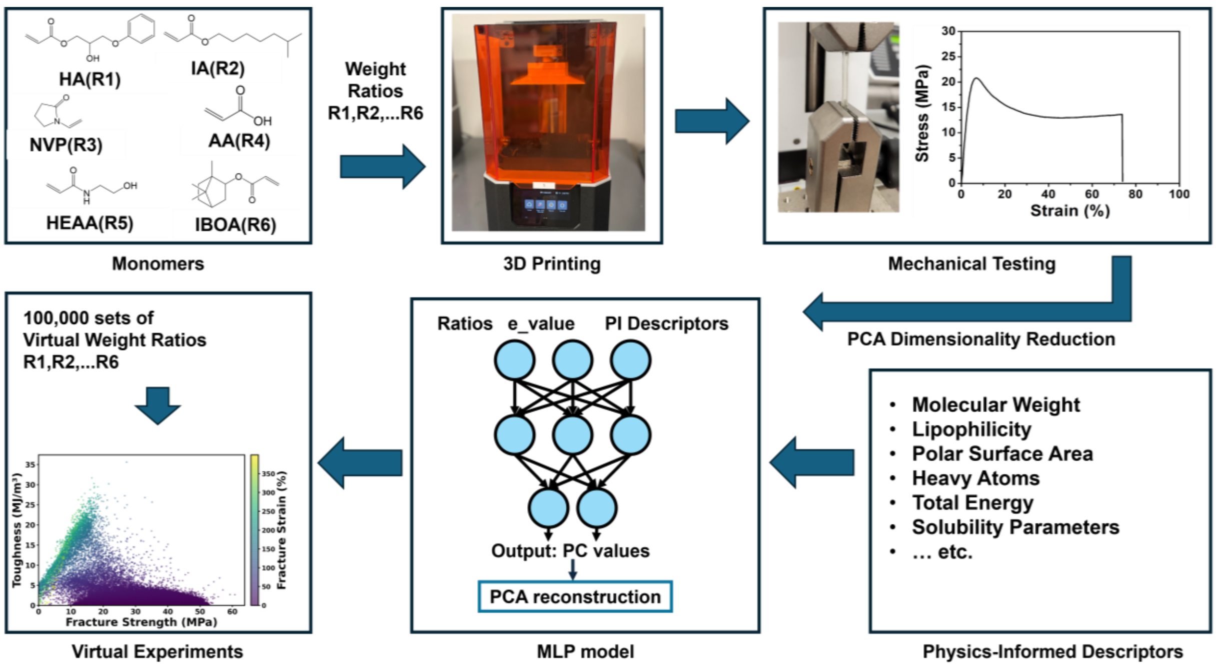

Figure 1. Workflow of developing a PIML based virtual experimentation platform.

Data collection and preprocessing. Experimental datasets were collected from 62 ink formulations, with each formulation represented by 2-4 individual S-S curves. 3D thermoplastics printed from the six monomers involve enormous chemical space. Training ML models with only the ratio of the six monomers to predict the high dimensional outputs could suffer from a serious overfitting issue. To overcome this issue, additional thirteen physics-informed descriptors were chosen as the inputs. They are the molecular weight, lipophilicity, *h*-bond donor, *n*-bond acceptor, rotatable bonds, polar surface area, heavy atoms, complexity, total energy, and solubility parameters.[29-32] After normalization, these physics-informed descriptors were multiplied by the ratios of six monomers, leading to 78 cross-features.[25] Details on these descriptors and more information about the methodology can be found in **Supplementary Note S1** and **Table S1**.

The S-S curves of the specimens with the same ink formulation underwent analysis to ensure the high quality of training data. As depicted in **Figure 2a**, the three stress-strain curves of three specimens exhibit variation even though they were printed from the same ink formulation, indicating the unavoidable experimental uncertainty. If using the ink formulation and the corresponding S-S curves as the input and output for the model training, a 'one-to-many' prediction issue may arise, where each input corresponds to multiple outputs.[33,34] It underscores the importance of using a model capable of adeptly handling such inherent data variability. To address this uncertainty, an e_value based on a normal distribution was introduced to encapsulate the inherent experimental variation. This e_value, analogous to the Z-score in a normal distribution, quantifies how many standard deviations that experimental data point deviates from the mean. Implementation of the e_value is elaborated in the Methods section. The e_value is combined with the ratios of the six monomers and the 78 cross-features to form a total of 85 features into the model.

Depending on different monomer ratios of the inks from which the samples were printed, these S-S curves represent four distinct soft/elastic, soft/tough, strong/tough, and hard/brittle samples, presenting the diversity of the training data, which imposes additional challenge for the model training (**Figure 2b**). The stress-strain curve for the soft/elastic sample shows typical elastomer behavior, with minimal stress at low strains and a constant stress level during significant elongation. The soft/tough and strong/tough samples begin with a steep initial slope, indicating stiffness, but as strain increases, the curves show continuous stress rise without peaking, reflecting substantial plastic deformation. Conversely, the hard/brittle sample's curve displays a linear increase followed by a sharp stress drop, characteristic of minimal plastic deformation before fracture. Due to significant variations in the length of data collected, preprocessing steps such as trimming, and interpolation were necessary to standardize the datasets for model training. Detailed descriptions of these preprocessing methods are provided in **Methods Section 4.4**.

Further observation shows that the numerical range of the strain axis varies considerably, even though both the strain and stress axes consisting of 50 data points each in the standardized data format. Given the limited datasets and a 100-dimension output, a concern known as the 'curse of dimensionality' arises, a phenomenon where the volume of the space increases so fast that the available data become sparse.[35] This sparsity is problematic as it can severely impact the performance of machine learning models by making it difficult to extract meaningful patterns without overfitting. Given the limited datasets and the high-dimensional output, dimension reduction becomes essential to mitigate these issues. Previous studies adopted a manual extraction strategy to identify five feature points, i.e., linear limit, maximum yielding, strain softening end, steady flow limit, and fracture points.[24,33] In our research, however, the S-S curves in our dataset are more diverse, making the manual extraction of these critical points either cumbersome or inconsistent. To address those concerns, PCA, a powerful dimension reduction technique, was employed.[36] PCA is an unsupervised method that does not require predefined criteria for extracting information. It simplifies the dataset by transforming it into a new coordinate system, where the most significant features are summarized in the principal components (PCs). This process not only makes the data more manageable for the ML model but also preserves essential information, thereby facilitating accurate predictions. Instead of directly predicting the whole S-S curves, our model predicts the PC values, which can be then converted back to the S-S curves.

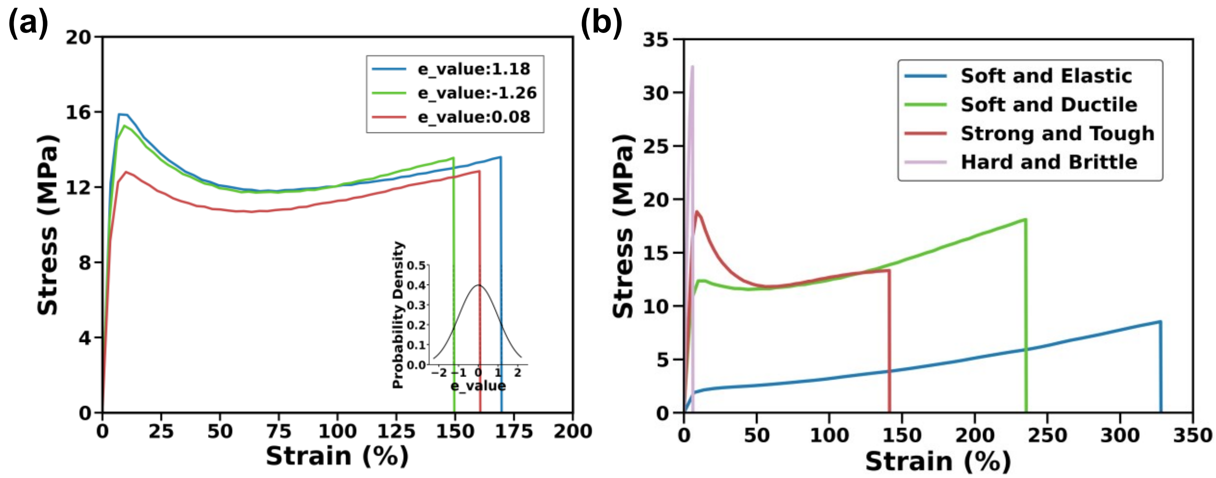

Figure 2. (a) Calculation of e_value based on normal distribution of fracture points of the S-S curves obtained from multiple samples printed with the same ink formulation. **(b)** Four typical S-S curves for the printed representative thermoplastic samples.

PCA on Stress-Strain Curves. The impact of the number of principal components (denoted as n) on the capacity of the ML model to encapsulate data variance was initially investigated, with a focus on the explained variance which refers to the cumulative proportion of the dataset variance explained with the increase of n. As shown in **Figure 3a**, the cumulative explained variance (CEV) increases sharply as n reaches 5, beyond which there is negligible change, indicating the efficacy of PCA in capturing key information from the S-S curves (see **Supplementary Note S2** for details). This trend is also evident when using the PCs to reconstruct the S-S curves (**Figure 3b-c**). The RMSE[37] was chosen to determine the difference between the reconstructed and original values of both stress and strain axes (**Supplementary Note S3**). Specifically, the strain RMSE decreases to ~0.02 % when n reaches 4, while the stress RMSE remains nearly unchanged (~0.03 MPa) at n of 7. Furthermore, the impact of n on the accuracy of the reconstructed S-S curves was also investigated visually across the collected datasets. **Figures 3d-g** show a few examples, illustrating typical representatives S-S curves as discussed in **Figure 2b**. It is found that samples show good agreement between the original and reconstructed curves when *n* reaches 6. Based on these observations, to encapsulate more subtle variations, the *n* value is set to 8 for the subsequent analysis.

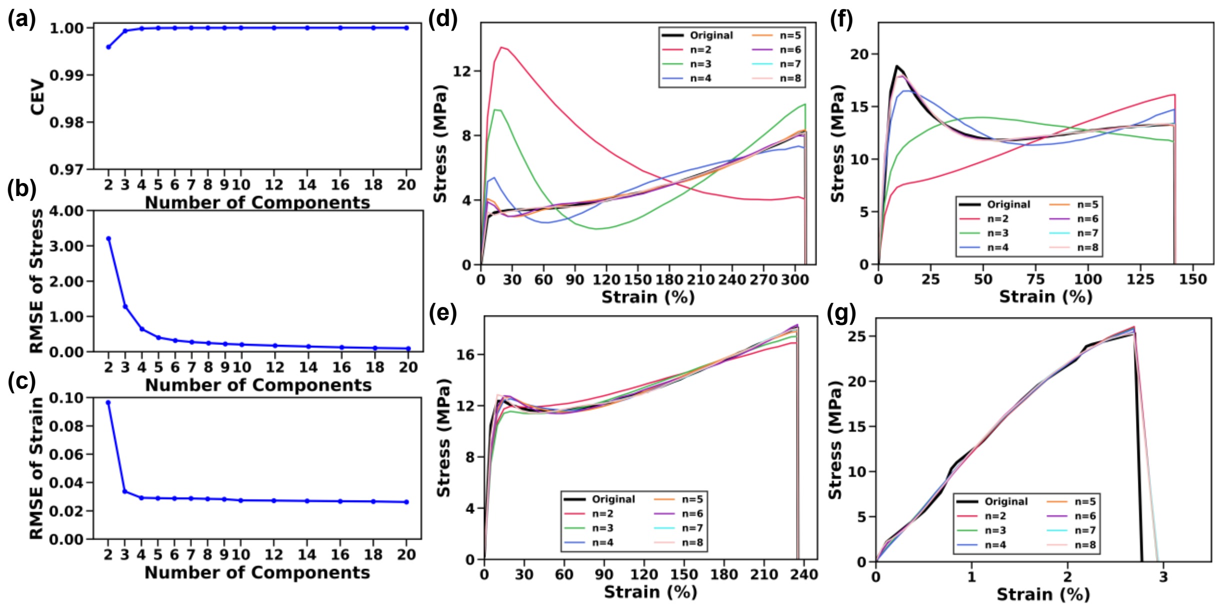

Figure 3. (a) Cumulative explained variance (CEV) with respect to different number of principal components (*n*). Change of stress RMSE **(b)** and strain RMSE **(c)** vs. *n*. The reconstructed stress-strain curves as the increase of *n* for **(d)** soft/elastic, **(e)** soft/ductile, **(f)** strong/tough, and **(g)** soft and elastic samples.

Interpretability of PCA. After exploring the influence of *n* on the reconstructed S-S curves, we thoroughly examined the interpretability of each PC during the reconstruction process. By analyzing how the PC values influence the S-S curves, we demonstrate how the PCs reflect essential features of the S-S curves. To do it, each PC is varied by ±100%, ±50%, ±20%, and ±5%, while keeping other PCs the same. As shown in **Figure 4a** and **Figure S1**, increase in PC1 prompts shift of the S-S curves towards larger strains, while increase in PC2 results in a decrease in the slope of the plastic deformation region. It is determined that PC1 has the most pronounced effect on the variations of the S-S curves. Increase in PC3 leads to a decrease in the slope of the post-yield hardening region, whereas increase in PC4 results in a decrease in the slopes of the plastic deformation region while an increase in the post-yield hardening region. Furthermore, the fractural strain remains constant regardless of the changes in PC2, PC3, and PC4. While the influence of PC5 to PC8 is not dramatically significant to be directly interpreted by material scientists analyzing the core material properties, these components still contribute to the finer details of the curves, such as minor fluctuations or inflection points in certain regions of the curves. For a brittle sample (**Figure S2**), close observation reveals that increase in PC1 leads to a shift of the curve toward smaller strain, while increase in PC2 results in the increase in the slope of the elastic deformation, fracture strength and fracture strain. There are no obvious changes in the S-S curves with the changes in PCs from PC3 to PC8. To further explore the hidden information, the relationship between PCs and mechanical properties was analyzed (**Figure S3**). Clearly, PC1 exhibits a linear relationship with the fractural strain. PC2 is proportional to toughness. PC3 is positively and negatively correlated to the fracture strength and the slope in the strain-hardening area, respectively. The observations are well aligned with the fundamental mechanical characteristics observed in the S-S curves (**Figure 4b**).

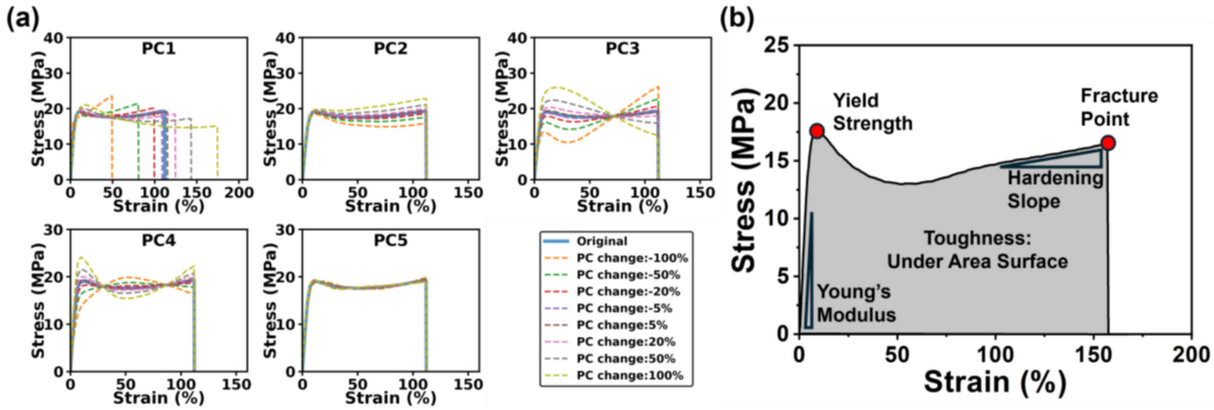

Figure 4. (a) Changes of each PC (PC1 to PC5) vs. change of the reconstructed S-S curves of a strong/tough sample. **(b)** A typical S-S curve with labeled characteristic points.

Machine Learning Model. After establishing the input and output datasets, it is about to train an MLP model. The model takes 85 distinct and cross-features as the inputs to predict outputs of the eight PCs. Given the relatively small data size, a combined approach of dropout and L1 regularization was employed to prevent overfitting. Dropout operates by randomly deactivating a subset of neurons during the training process, which is beneficial for reducing the model's dependency on specific features.[38] Meanwhile, the L1 regularization introduces a penalty to the loss function proportional to the absolute magnitude of the feature coefficients.[39] It prioritizes more influential features by pushing the coefficients of less significant ones towards zero. Both the dropout and L1 regularization work in concert to enhance the model's capacity to be generalized effectively. Furthermore, the model is designed to favor the utilization of beneficial physics informed descriptors, while reducing reliance on those with less impact. This selective approach ensures that the model not only stays accurate but also remains relevant and grounded in the practical aspects of domain science. Mean squared error between the eight predicted and true PCs is chosen as the loss function since it can effectively reflect the hierarchy of significance by preserving the original difference among the PCs.

Out of the 62 ink formulations, 50 (representing180 S-S curves) were chosen as the training datasets, while the remaining 12 (representing 36 S-S curves) were the testing datasets. Here, the test set comprises a balanced combination of materials consisting of 7 elastic ones and 5 brittle ones. Details on the model's intricacies, computation specifics, and information about the hardware and software utilized in this study are comprehensively documented in **Supplementary Note S4**.

Based on the test set, performance of the MLP model in predicting the eight PCs is presented in **Table 1**. While the specific PC values lack direct physical meanings, the R^2 values in comparison of the predicted PCs and respectively true PCs reveal the model's accuracy. The R^2 values were notably high for the first three principal components (0.97, 0.76, and 0.77 for PC1, PC2, and PC3, respectively) and gradually declined for the remaining five PCs. This trend is expected, i.e., the importance of PCs slightly decreases as the number of PCs increases. This trend also holds true for other evaluation metrics including RMSE, MAE, and MSE, indicating that the MLP model prioritizes the key PCs. RMSE exhibits an opposite trend, starting at 5.40% for PC1 and 10.94% for PC2, and then gradually increasing to 25.18 % for PC8. It also underlines the model's ability to concentrate on the most impactful PCs for balancing the accuracy by prevention of overfitting. This inherent characteristic originates from the L1 regularization and dropout to ensure a robust fit for the most significant features.

Table 1. Evaluation of the MLP model based on PCs

PC values	R^2	RMSE	MAE	MSE	Max	Min	Range	RMSE/Range
1	0.97	78.84	44.48	6215.58	1263.16	-197.45	1460.61	5.40%

2	0.76	9.79	7.58	95.82	65.84	-23.65	89.49	10.94%
3	0.77	6.47	5.32	41.86	31.38	-26.13	57.51	11.25%
4	0.58	4.63	3.82	21.43	16.21	-12.54	28.76	16.10%
5	0.29	2.17	1.78	4.69	3.4	-6.33	9.72	22.32%
6	0.21	1.21	0.97	1.47	2.84	-2.96	5.79	20.99%
7	0.41	1.25	0.9	1.55	6.96	-2.71	9.67	12.93%
8	0.19	0.69	0.53	0.48	1.13	-1.61	2.74	25.18%

Evaluating Stress-Strain Curves. The results indicate the high accuracy of the MLP model in predicting the eight PCs. We then evaluated how well the reconstructed S-S curves from these predicted PCs agree with the true ones. It is impractical to evaluate the reconstruction performance by directly calculating the difference between the reconstructed and true value at each point of the S-S curves. This is because the complexities of material behaviors and testing conditions lead to the huge variations of the S-S curves. To mitigate this issue, two critical mechanical performance matrices, i.e., fracture strength and toughness, which can be derived from the S-S curves, were deployed for evaluation. As shown in **Table 2**, the R^2 values are relatively high for fracture strength (0.97) and toughness (0.95), while RMSE and MAE of the fracture strength are 1.01 and 0.82 MPa and for toughness they are 0.40 and 0.31 MJ/m^3. After considering their ranges, RMSE of the fracture strength and toughness are relatively low, i.e., ~4% for the fracture strength and ~6% for the toughness. These results indicate the model's robust ability to account for a significant portion of the observed data variance.

Table 2. Evaluation of the ML model based on fracture strength and toughness.

Metric	R^2	RMSE	MAE	Max	Min	Range	RMSE/Range
Fracture strength	0.97	1.01	0.82	39.29	11.76	27.53	4.43%
Toughness	0.95	0.40	0.31	10.48	4.03	6.45	5.90%

To visually evaluate the model prediction performance, the true and predicted S-S curves (reconstructed from the predicted PCs by the MLP) of the four samples from test set with various fracture strength and ductility are shown in **Figure 5a**. Additionally, all 36 S-S curves from the test set are provided in **Figure S4.** The yellow lines correspond to the original S-S curves, while the blue lines represent the reconstructed S-S curves with the corresponding e_values. To effectively adapt to the variations originated from the experimental and testing conditions, the e_values varying from -2 to 2 were incorporated to reconstruct multiple S-S curves (grey lines). The grey range encompasses 95% of probability about the cases according to the Z-score definition in a normal distribution. It is found that these reconstructed S-S curves all fall within the grey areas. And their shapes and trends are matched well with the ground truth S-S curves. These results affirm the high effectiveness of the combination of the MLP model and PCA technique in predicting the S-S curves.

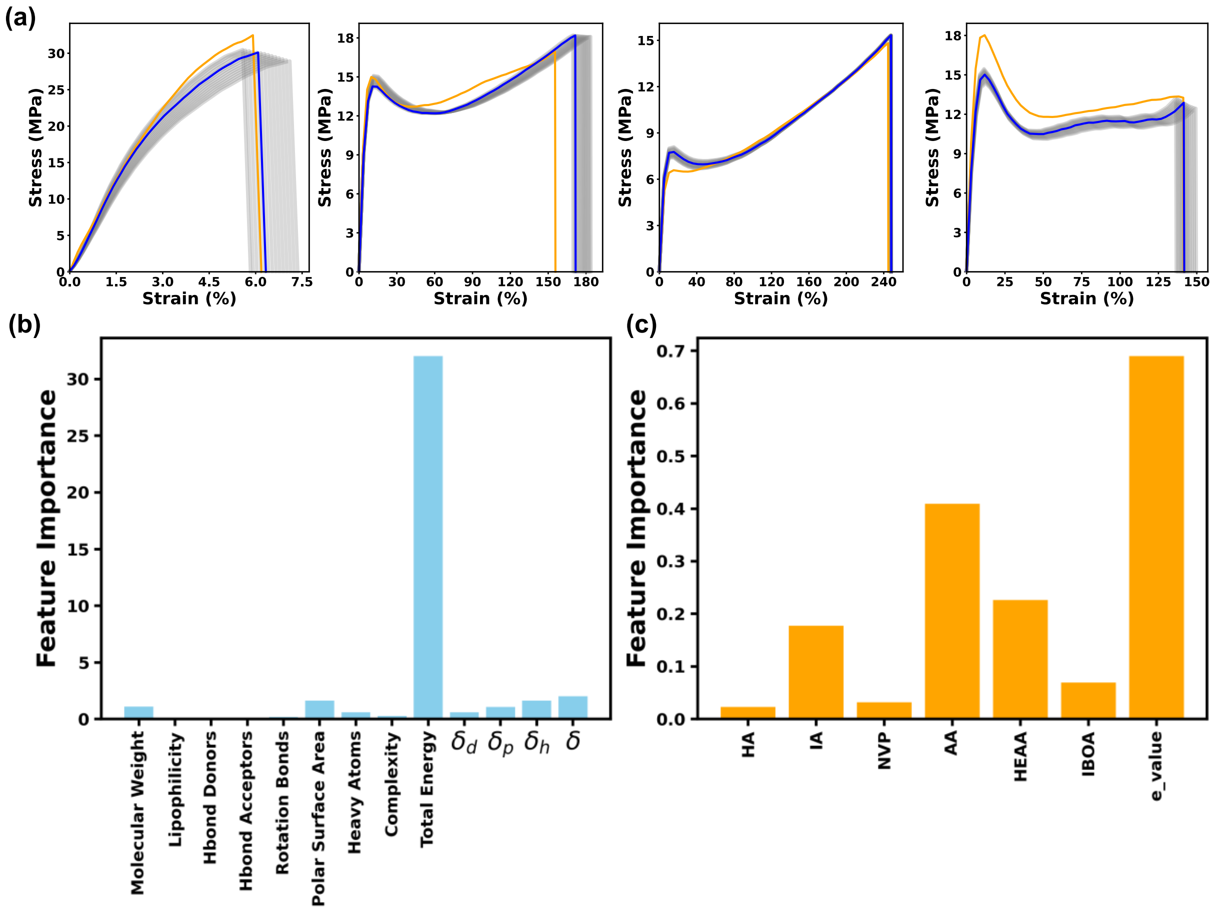

Figure 5. **(a)** The comparison between ground-truth curve (yellow) and predicted (blue) stress-strain curves of four representative samples. Considering the uncertainty, the e_values varying from -2 to 2 were used to predict the S-S curves with 95% probability (grey lines). **(b)** Importance ranking of the 13 physics-informed features. δ_d: solubility influenced by the molecule's dipole moment, δ_p: solubility parameter, δ_h: the hydrogen-bonding component of solubility, δ: solubility expressed in terms of energy density (MJ/m^3). **(c)** Importance ranking of ratios of the six monomers and e_values.

Feature importance. Importance of the physics-informed descriptors was explored via a comparative study training the MLP model using only the ratios of six monomers and the e_value without PI inputs. As shown in **Table S2**, the MLP model attained the highest R^2 value for PC1, while delivering much lower R^2 values for PC3 and PC4. This indicates that the model cannot effectively capture the underlying characteristics of the training datasets if only using PC3 and PC4. Furthermore, presence of negative R^2 values for PC2, PC5, PC6, PC7 and PC8 reveals that predictive accuracy of the MLP model is even worse than the prediction results using the average of all sampling data. This underscores a substantial limitation in the MLP model without the physics-informed descriptors. This phenomenon was also found in the predicted S-S curves (**Table S3** and **Figure S5**). The R^2 values for both true stress (0.52) and toughness (0.38) are lower than those of the MLP model trained with included physics-informed descriptors. As shown in **Figure S5**, nearly all the predicted S-S curves exhibited huge variations, revealing the poor prediction capability of the model without the physics-informed descriptors. These results indicate that the incorporation of physics-informed descriptors not only increases the predictive accuracy but also aid in accurately capturing the nuances of the S-S curves.

The significance of these physics-informed descriptors was further quantified. An integrated gradients (IG) method was applied to investigate the interpretability of the MLP model.[40] The IG method works by examining how change in the gradients of each feature influences the output. Specifically, for each PI descriptor, we calculated its interaction feature importance with each of the six monomers. To synthesize this information and provide a clearer understanding of the overall impact of each PI, we averaged the importance scores across these six monomers for every individual PI. The feature importance scores for 13 physics-informed descriptors, the ratio of six monomers, and e_values were shown in **Figure 5b** and **5c**. Detailed methodologies regarding this process are elaborated in the **Methods** section. As shown in **Figure 5b**, total energy is the primary dominant descriptor among these physics-informed descriptors, which well agrees with expertise and domain knowledge. It is reported that total energy plays a crucial role in determining the structural cohesion, arrangement, and consequent mechanical properties of polymeric materials.[41] Other physics-informed descriptors such as solubility, molecular weight, polar surface area, and the number of heavy atoms exhibit relatively lower importance. This suggests that the model effectively leverages these classical features to capture complementary information related to chain entanglement, intermolecular forces, and steric effects, which are known to influence polymer performance.[42,43] The remaining descriptors, including complexity, lipophilicity, Hbond donor, Hbond acceptor, and rotatable bonds, exhibit comparatively lower feature importance scores. These descriptors primarily pertain to molecular size, hydrophobicity, and conformational flexibility. The direct impact of these descriptors on intermolecular interactions and electronic structures, which play pivotal roles in determining the mechanical properties of polymers, may be relatively limited.

As shown in **Figure 5c**, the e_value, used to account for the experimental uncertainty, was notably discernible. This highlights the model's capability to establish a predictive range based on e_value rather than a simple one-to-one prediction. The feature importance scores for the six monomers follow the order of AA > HEAA > IA > IBOA > NVP > HA. Monomers like AA and HEAA are noteworthy for their propensity to form hydrogen bonds, significantly impacting the intermolecular interactions of the 3D printed thermoplastics.[44] Presence of IA can be attributed to its function as a softer segment than HA, contributing significantly to the flexibility and toughness of 3D printed thermoplastics, despites the potential of HA to form hydrogen bonds.[5] These feature importance scores well agree with the empirical understanding of the experiments, thus reinforcing the significance and practical applicability of these descriptors in the MLP model. This method underscores the effectiveness of combining data-driven machine learning with domain-specific expertise, paving the way to more sophisticated and accurate predictive models in materials science.

Virtual experimentation for screening new ink formulation candidates. We expect that the developed MLP can be used as a surrogate model to virtually explore the combination space to accelerate the ink formulation to make the thermoplastics that show desired S-S curves. First of all, 100,000 virtual ink formulations were randomly generated using the Dirichlet distribution method since it ensures a uniform distribution of each monomer.[45] This approach guarantees an equitable representation of all possible monomer ratios, providing a balanced and comprehensive exploration of the design space. Details on generating virtual ink formulations are provided in Method. After that, a pre-trained random forest model that we previously demonstrated was employed to predict the printability of these ink formulations.[5] Only the printable ink formulations were fed into the MLP model to predict the corresponding eight PCs. It is noteworthy that the prediction of these ink formulations took only 1 minute, highlighting the exceptional speed and efficiency of the virtual screening. Then, the S-S curves were reconstructed from the predicted PCs. Then, the fracture strength, maximum strain and toughness were extracted from these reconstructed S-S curves and plotted in **Figure 6a**. It was observed that most datapoints were clustered in the region associated with lower toughness, possibly because out of six monomers, four of them are harder monomers including NVP, HA, HEAA and IBOA. If they are dominant in the ratio combinations, they considerably favor formation of brittle thermoplastics with low toughness.

Following the virtual screening guided by MLP model, new experiments were conducted to validate the prediction results. We chose these experiments with an aim of identifying the ink formulations leading to three types of thermoplastics (strong/tough, strong/brittle, and soft/elastic). For each type, two ink formulations were randomly selected to print three specimens. **Figures 6b-g** show the profiles and trend of the predicted S-S curves by the MLP model.

The first one showing the strong/tough S-S curve has a fracture strength in the range of 15-20 MPa and a toughness in the range of 15-20 MJ/m^3. As a result, a total of 143 ink formulations were screened, from which two ink formulations with HA: IA: NVP: AA: HEAA: IBOA weight ratios 0.16: 0.39: 0.25: 0.13: 0.02: 0.05 (**Figure 6b**) and 0.34: 0.32: 0.21: 0.09: 0.02: 0.02 (**Figure 6c**) were randomly selected for experiments. As depicted in **Figures 6b-c** the resulting S-S curves from these two selections conform to the trend predicted by the MLP model, in which both cases exhibited an instance of premature fracture. Moreover, to further support our mechanical testing data and elucidate the failure mechanisms, we conducted microstructural analysis of the fracture surfaces for the sample corresponding to **Figure 6b**. For this more ductile formulation, digital microscope observations reveal plastic deformation at the fracture points (**Supplementary Figure S6a**). These microstructural observations robustly support our claims regarding the mechanical properties of the material and provide deeper insights into the fracture behavior. The second type is the strong/brittle one with a fracture strength exceeding 35 MPa and a fracture strain of 2-5%, resulting in > 10,000 ink formulations. This is because lots of formulations in the virtual experiments show hard and brittle behaviors due to dominant compositions of NVP, HA HEAA or IBOA in the formula. The experimental S-S curves of the six specimens from the selected two ink formulations with HA: IA: NVP: AA: HEAA: IBOA weight ratios 0.16: 0.18: 0.05: 0.42: 0.18: 0.01 and 0.26: 0.29: 0.05: 0.29: 0.03: 0.08 are within the predicted range (**Figures 6d-e**). For the more brittle formulation represented in **Figure 6d**, the fracture surfaces are notably smoother (**Supplementary Figure S6b**), indicating a different failure mechanism. These microstructural observations further validate our experimental results and provide deeper insights into the different fracture behaviors. The third type is the soft/elastic one. The ink formulations with a predicted fractural strain of > 250% and a fracture strength in the range of 10-15 MPa were screened, resulting in 148 formulations. The selected two ink formulations with HA: IA: NVP: AA: HEAA: IBOA weight ratios of 0.4: 0.28: 0.01: 0.0: 0.09: 0.22 and 0.35: 0.38: 0.02: 0.18: 0.07: 0.0 led to the soft/elastic thermoplastics. Their S-S curves are shown in **Figures 6f-g**. We can see that the predicted S-S profiles agree well with the experimental ones despite the little discrepancy in their fractural strains. They are out of the range of the predicted uncertainty range. These experimental validation results show that the developed MLP for virtual experiment is reliable and rapid because the prediction of 100,000 ink formulations is within one minute. This rapid and efficient virtual experimentation process can significantly facilitate the exploration of design space for identification of ink formulations that lead to materials with desired properties, thus accelerating the development of new materials.

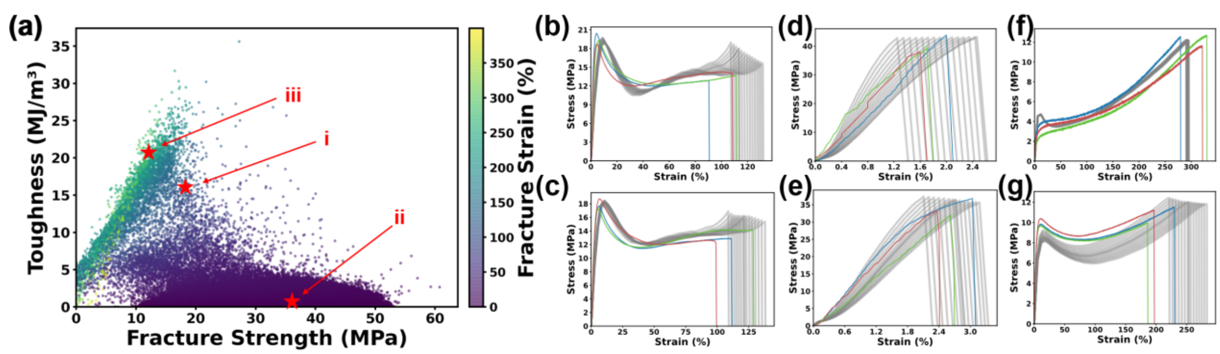

Figure 6. (a) Plot of fracture strength, fracture stain, and toughness extracted from the predicted S-S curves. Red stars i, ii, and iii indicate the chosen ink formulations shown in Panel b-c, d-e, and f-g, respectively. The S-S curves for the three samples (red, green, and blue) printed with the ink formulations that are predicted to result in the **(b-c)** strong/tough, **(d-e)** hard/brittle, and **(f-g)**

soft/elastic type of thermoplastics. The grey areas represent the uncertainty range of the predicted S-S curves.

3. Conclusions

In this study, a PIML model was developed for virtual experimentation to accelerate the discovery of 3D printed thermoplastics. The collected 216 S-S curves from 62 ink formulations were dimensionally reduced into eight PCs. Meanwhile, 13 physics-informed descriptors were included using domain knowledge to increase the robustness and generalization of the model. The developed physics informed MLP model achieved superior R^2 and RMSE values when predicting the values of the eight PCs. The reconstructed S-S curves from the predicted PCs were well matched with the true ones. Feature importance analysis confirmed the importance of physics-informed descriptors, showing that the total energy is the most important one. After mapping the mechanical properties of 100,000 ink formulations by the MLP model, six representative ink formulations that are expected to lead to three different types of thermoplastics were chosen. Validation experiments demonstrated a strong agreement between the predicted and experimental S-S curves. The methodologies and workflow can be readily extended to other materials for predicting other performance curves such as Raman, electrochemistry curves. This underscores the versatility and potential of this approach in a range of material science and chemical research scenarios, offering a robust framework for expedited and accurate material and chemical analysis.

4. Materials and Methods

4.1 Materials. 2-Hydroxy-3-phenoxypropyl acrylate (HA), isooctyl acrylate (IA, > 90%), and acrylic acid (AA, 98%) were purchased from Sigma Aldrich (St. Louis, MO, U.S.). Diphenyl(2,4,6-trimethylbenzoyl) phosphine oxide (TPO, >97%), isobornyl acrylate (IBOA, > 90%), N-vinylpyrrolidone (NVP, > 99%), and N-(2-hydroxyethyl) acrylamide (HEAA, >98%) were purchased from Fisher Scientific (Pittsburgh, PA, U.S.).

4.2 3D Printing and Mechanical Testing. In this study, the LCD 3D printing process was executed using a resin mixture comprising six monomers: HA, IA, AA, IBOA, NVP, and HEAA with carefully measured weight ratios. Each monomer's ratio in the mixture can vary continuously from 0 to 1. For the sake of experimental precision, the ratios have two decimal places. The total sum of the ratios for all monomers equals 1. To make the mixture, a photoinitiator, diphenyl(2,4,6-trimethylbenzoyl) phosphine oxide (TPO), was added at a concentration of 2 wt%. The mixture was then subjected to magnetic stirring for one minute to ensure thorough and uniform mixing. The resulting homogenized resin was used in an Anycubic Photon Mono 4K printer, operating at a 405 nm irradiation wavelength. The printing parameters included a power density of about 5 mW/cm^2, a layer thickness of 50 µm, and an exposure time of 15 seconds per layer. Following the printing process, the samples were further cured under 405-nm UV light for 60 seconds. For the mechanical assessment of the 3D-printed samples, tensile testing was carried out using a Mark-10 universal testing machine at a loading rate of 50 mm/min. To ensure a comprehensive statistical analysis, a minimum of 5 samples were printed and tested for each monomer ratio.

4.3 S-S Curve Collection. 326 S-S curves were collected from 62 distinct formulations, each of which was subjected to 5-7 independent mechanical tensile tests. To ensure the reliability and quality, the S-S curves with significant errors such as measurement inconsistencies, premature breakage, or excessive mechanical testing noise were excluded. Consequently, a refined dataset comprising 216 S-S curves was obtained, with each thermoplastic represented by 2-4 individual curves. To demonstrate the diversity and balance of the dataset, when considering a maximum strain of 10% as threshold, the data showed a distribution where approximately half of the materials displayed brittle properties (106 samples), while the other half exhibited higher ductility (80 samples).

4.4 Data Processing of S-S Curves. The preliminary cleaning of the raw data from the tensile testing machine involves trimming the initial segments of each S-S curve to eliminate any measurements taken before the machine commenced operation by standardizing the starting points to a baseline of zero stress and zero strain (0,0). Then, a critical aspect of the preprocessing involves identifying the point of failure within each sample's S-S curve. By pinpointing and marking the exact location of sample failure on each curve, the final data point is represented the moment of fracture by capturing the complete mechanical profile of each specimen. The last step in the data preprocessing routine is to apply an interpolation technique to standardize the data representation. Each S-S curve is interpolated to consist of 50 data points uniformly distributed in the x axis (strain).

4.5 Experimental Uncertainty. In this study, the e_value is calculated based on normal distribution to capture the inherent uncertainties in the S-S data at the fracture point. This refinement involves analyzing the final strain values at fracture for each dataset as illustrated in **Figure 2a**. By aggregating these values, a comprehensive picture of the strain behavior at fracture across various samples was obtained. To encapsulate the variability in the fracture strains of the materials, first their means are calculated, providing a reference for the average material behavior under stress. Then the standard deviation is computed to quantify the dispersion among these values, a crucial step in highlighting the heterogeneity in material responses. This approach normalizes each fracture strain relative to this mean, adjusting for variance. This process results in the e_values, the standard deviations indicating the deviation of each sample's fracture point from the average, Mathematically, this normalization is expressed as:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

where: μ is the mean of the fracture strain for all samples, n is the number of samples (S-S curves) and x_i is the fracture strain value for each sample.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

where: σ is the standard deviation, $(x_i - \mu)$ represents the deviation of each sample's fracture strain value.

$$e\_value = \frac{x - \mu}{\sigma}$$

4.6 Uncovering Features' Importance. Due to the inherent complexity and 'black box' characteristic of the MLP model, we utilized the Integrated Gradients (IG) method for interpretability study.[40] This approach is particularly adept to illuminate the contribution of each input feature to the model's output. It works by calculating the gradient of the model's prediction with respect to each input feature. It then integrates these gradients along a path from a baseline input (a zero vector) to the actual input. This process effectively captures the importance of each feature in the model's prediction, highlighting both linear and non-linear relationships within the model. To do that, the analysis was expanded to include the entire dataset (both training and testing datasets) to ensure a comprehensive assessment on the feature importance. The IG method, applied to each data point, calculated the significance of every feature in relation to the model's predictions, thereby providing a quantitative measure of each feature's contribution. This process involved aggregating importance scores across all samples to derive an average importance for each feature. Additionally, focused analysis was conducted on cross-features: where Physics-Informed (PI) descriptors interact with monomer ratios. For each PI descriptor, the average importance across all its interactions was calculated, allowing for an assessment of the overall influence of each PI descriptor on the model's predictions.

4.7 Virtual Experiments Ratio Generation Details. In the generation of random experiment formulations within our study, we employed the Dirichlet distribution. This distribution is commonly utilized for generating random proportions under specific constraints, like that the sum of the monomer ratios equals to 1, making it particularly suitable for simulating a diverse range of monomer mixtures.[45] Additionally, an important characteristic of the Dirichlet distribution is its uniformity and symmetry, when the parameters of the distribution, known as 'alpha', are all set equal to 1. This equal setting means that each component of the distribution has an equal chance of being sampled, leading to an evenly spread of probabilities across all ratios. For each generated combination, the first five ratios were rounded to two decimal places. The sixth ratio in each combination was then determined by subtracting the sum of these first five rounded ratios from one.

Acknowledgment

J. L. thanks the financial support from National Science Foundation (award number: 2154428), U.S. Army Corps of Engineers, ERDC (grant number: W912HZ-21-2-0050).

Author Contributions

Z. C. designed and implemented data preprocessing, model training and testing, and data analysis. Y. W. conducted all the experiments and analyzed the model interpretability aspects. K.S used DFT calculations for physics-informed descriptor data collection. Y. X. contributed to the feature-important analysis of the model. J. L. conceived the idea, managed the research progress, and provided regular guidance. Z. C., Y.W. and Y. X. drafted the first manuscript, which was thoroughly revised by J. L. All authors commented and agreed on the final version of the manuscript.

Data and Code Availability

The code supporting the findings of this study is available on GitHub at https://github.com/linresearchgroup/VirtualEXP_3Dprinting.

The repository also includes original experimental datasets used in this study.

Declaration of Interests

The authors declare no competing interests.

References

1 Gómez-Bombarelli, R. *et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature materials* **15**, 1120-1127 (2016).

2 Penumakala, P. K., Santo, J. & Thomas, A. A critical review on the fused deposition modeling of thermoplastic polymer composites. *Composites Part B: Engineering* **201**, 108336 (2020).

3 Awasthi, P. & Banerjee, S. S. Fused deposition modeling of thermoplastic elastomeric materials: Challenges and opportunities. *Additive Manufacturing* **46**, 102177 (2021).

4 Wu, Y. *et al.* A photocured Bio-based shape memory thermoplastics for reversible wet adhesion. *Chemical Engineering Journal*, 144226 (2023).

5 Sattari, K. *et al.* Physics-constrained multi-objective bayesian optimization to accelerate 3d printing of thermoplastics. *Addit. Manuf.* **86**, 104204 (2024).

6 Cox, B. & Yang, Q. In Quest of Virtual Tests for Structural Composites. *Science* **314**, 1102-1107 (2006). https://doi.org:doi:10.1126/science.1131624

7 Zhang, B., Yang, Z., Sun, X. & Tang, Z. A virtual experimental approach to estimate composite mechanical properties: Modeling with an explicit finite element method. *Comput. Mater. Sci.* **49**, 645-651 (2010).

8 Zhang, N., Lu, B., Wang, W. & Li, J. Virtual experimentation through 3D full-loop simulation of a circulating fluidized bed. *Particuology* **6**, 529-539 (2008). https://doi.org:https://doi.org/10.1016/j.partic.2008.07.013

9 Xue, Y. L., Huang, J., Lau, C. H., Cao, B. & Li, P. Tailoring the molecular structure of crosslinked polymers for pervaporation desalination. *Nature communications* **11**, 1461 (2020).

10 Xie, Y., Sattari, K., Zhang, C. & Lin, J. Toward autonomous laboratories: Convergence of artificial intelligence and experimental automation. *Prog. Mater Sci.* **132**, 101043 (2023).

11 Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73-76 (2016). https://doi.org:10.1038/nature17439

12 Xie, Y. *et al.* Machine learning assisted synthesis of metal–organic nanocapsules. *Journal of the American Chemical Society* **142**, 1475-1481 (2019).

13 Dong, Y. *et al.* Bandgap prediction by deep learning in configurationally hybridized graphene and boron nitride. *npj Computational Materials* **5**, 26 (2019). https://doi.org:10.1038/s41524-019-0165-4

14 Rao, Z. *et al.* Machine learning–enabled high-entropy alloy discovery. *Science* **378**, 78-85 (2022). https://doi.org:doi:10.1126/science.abo4940

15 Koscher, B. A. *et al.* Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back. *Science* **382**, eadi1407 (2023). https://doi.org:doi:10.1126/science.adi1407

16 Mikulak-Klucznik, B. *et al.* Computational planning of the synthesis of complex natural products. *Nature* **588**, 83-88 (2020). https://doi.org:10.1038/s41586-020-2855-y

17 Rinehart, N. I. *et al.* A machine-learning tool to predict substrate-adaptive conditions for Pd-catalyzed C–N couplings. *Science* **381**, 965-972 (2023). https://doi.org:doi:10.1126/science.adg2114

18 Coley, C. W. *et al.* A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019). https://doi.org:doi:10.1126/science.aax1566

19 Correa-Baena, J.-P. *et al.* Accelerating materials development via automation, machine learning, and high-performance computing. *Joule* **2**, 1410-1420 (2018).

20 Ren, Z. *et al.* Machine learning–aided real-time detection of keyhole pore generation in laser powder bed fusion. *Science* **379**, 89-94 (2023). https://doi.org:doi:10.1126/science.add4667

21 Chen, Z. *et al.* An interpretable and transferrable vision transformer model for rapid materials spectra classification. *Digital Discovery* **3**, 369-380 (2024). https://doi.org:10.1039/D3DD00198A

22 Yang, C., Kim, Y., Ryu, S. & Gu, G. X. Prediction of composite microstructure stress-strain curves using convolutional neural networks. *Materials & Design* **189**, 108509 (2020).

23 Tsai, M.-L., Huang, C.-W. & Chang, S.-W. Theory-inspired machine learning for stress–strain curve prediction of short fiber-reinforced composites with unseen design space. *Extreme Mechanics Letters* **65**, 102097 (2023).

24 Ha, C. S. *et al.* Rapid inverse design of metamaterials based on prescribed mechanical behavior through machine learning. *Nature Communications* **14**, 5765 (2023).

25 Karniadakis, G. E. *et al.* Physics-informed machine learning. *Nature Reviews Physics* **3**, 422-440 (2021).

26 Sattari, K. *et al.* A scientific machine learning framework to understand flash graphene synthesis. *Digital Discovery* **2**, 1209-1218 (2023).

27 Sattari, K. *et al.* De novo molecule design towards biased properties via a deep generative framework and iterative transfer learning. *Digital Discovery* **3**, 410-421 (2024).

28 Du, Y., Mukherjee, T. & DebRoy, T. Physics-informed machine learning and mechanistic modeling of additive manufacturing to reduce defects. *Applied Materials Today* **24**, 101123 (2021).

29 Chin, K. C., Cui, J., O'Dea, R. M., Epps III, T. H. & Boydston, A. J. Vat 3D printing of bioderivable photoresins–toward sustainable and robust thermoplastic parts. *ACS Sustainable Chemistry & Engineering* **11**, 1867-1874 (2023).

30 Kim, S. *et al.* PubChem 2023 update. *Nucleic acids research* **51**, D1373-D1380 (2023).

31 Bertz, S. H. The first general index of molecular complexity. *Journal of the American Chemical Society* **103**, 3599-3601 (1981).

32 O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *Journal of cheminformatics* **3**, 1-14 (2011).

33 Liu, H., Wu, F.-Y., Zhong, G.-J. & Li, Z.-M. Predicting the complex stress-strain curves of polymeric solids by classification-embedded dual neural network. *Materials & Design* **227**, 111773 (2023).

34 Kościuszko, A., Marciniak, D. & Sykutera, D. Post-processing time dependence of shrinkage and mechanical properties of injection-molded polypropylene. *Materials* **14**, 22 (2020).

35 Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B. & Liao, Q. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing* **14**, 503-519 (2017).

36 Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemometrics and intelligent laboratory systems* **2**, 37-52 (1987).

37 Chai, T. & Draxler, R. R. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development* **7**, 1247-1250 (2014).

38 Baldi, P. & Sadowski, P. J. Understanding dropout. *Advances in neural information processing systems* **26** (2013).

39 Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58**, 267-288 (1996).

40 Qi, Z., Khorram, S. & Li, F. in *CVPR Workshops* Vol. 2 1-4 (2019).

41 Wu, Y. *et al.* Photocuring three-dimensional printing of thermoplastic polymers enabled by hydrogen bonds. *ACS Appl. Mater. Interfaces* **13**, 22946-22954 (2021). https://doi.org:10.1021/acsami.1c02513

42 Pugar, J. A., Childs, C. M., Huang, C., Haider, K. W. & Washburn, N. R. Elucidating the physicochemical basis of the glass transition temperature in linear polyurethane elastomers with machine learning. *The Journal of Physical Chemistry B* **124**, 9722-9733 (2020).

43 Van Krevelen, D. W. T. N., K. in *Properties of polymers (fourth edition)* Ch. Chapter 7 - Cohesive Properties and Solubility, 189-227 (Elsevier, 2009).

44 Wu, Y. *et al.* H-bonds and metal-ligand coordination-enabled manufacture of palm oil-based thermoplastic elastomers by photocuring 3D printing. *Addit. Manuf.* **47**, 102268 (2021). https://doi.org:https://doi.org/10.1016/j.addma.2021.102268

45 Briggs, A. H., Ades, A. & Price, M. J. Probabilistic sensitivity analysis for decision trees with multiple branches: use of the Dirichlet distribution in a Bayesian framework. *Medical Decision Making* **23**, 341-350 (2003).

Supporting Information for

Physics-Informed Machine Learning Enabled Virtual Experimentation for 3D Thermoplastics Printing

Zhenru Chen[1], Yuchao Wu[1], Yunchao Xie[2], Kianoosh Sattari[1], and Jian Lin[1*]

[1]Department of Mechanical and Aerospace Engineering, University of Missouri, Columbia, Missouri 65211, United States

[2]Department of Mechanical and Manufacturing Engineering, Miami University, Oxford, Ohio 45056, United States

*Email: linjian@missouri.edu

Supplementary Notes

Supplementary Note S1: Detailed information of physics informed descriptors.

Molecular Weight (MW): This descriptor represents the molecular weight of the compounds, a fundamental property affecting various material characteristics.

Lipophilicity (xlogp3): A measure of lipophilicity, xlogp3 indicates the distribution coefficient of the compound between water and a non-aqueous phase, impacting solubility and material interactions.

Hbond_donor and Hbond_acceptor: The number of hydrogen bond donors and acceptors in a molecule, crucial for understanding molecular interactions and binding capabilities.

Rotatable Bonds (Rot_bond): This parameter denotes the flexibility of a molecule, which can influence its mechanical and physical properties.

Polar Surface Area: Relating to the molecule's ability to interact with other molecules, the polar surface area is key in determining solubility and reactivity.

Heavy Atoms (HA): The count of heavy atoms within a molecule, providing insight into the molecular size and complexity.

The above descriptors are sources from PubChem. [1]

Complexity: A descriptor of the molecule's structural complexity, which can affect its physical behavior and interactions.[2]

Total Energy (dft_sp_E_RB3LYP): Calculated using Density Functional Theory (DFT), this value represents the total energy of the molecule, indicative of its stability and reactivity.[3]

Solubility_dipole: It refers to the solubility influenced by the molecule's dipole moment, a measure of the separation of positive and negative charges. It affects the interaction of the molecule with polar solvents like water.

Solubility_p: This indicates the solubility parameter, representing the cohesive energy density of a material. Substances with similar solubility parameters are generally soluble in each other, following the 'like dissolves like' principle.

Solubility_h: This descriptor relates to the hydrogen-bonding component of solubility, reflecting the compound's capacity to form hydrogen bonds and its consequent solubility in hydrogen-bonding solvents.

Solubility_sqrt_MJperm3: This is a measure of solubility expressed in terms of energy density (MJ/m^3). The square root transformation is applied for normalization or to linearize relationships in the data. The total solubility was calculated from Eq. 1 and Eq. 2.

$$\delta^2 = \delta_d^2 + \delta_p^2 + \delta_h^2$$

(1)

$$\delta_d = \frac{\sum \boldsymbol{F}_{di}}{V} \quad ; \quad \delta_p = \frac{\sqrt{\sum \boldsymbol{F}_{pi}^2}}{V} \quad ; \quad \delta_h = \sqrt{\frac{\sum \boldsymbol{E}_{hi}}{V}}$$

(2)

F_{di}, F_{pi}, and E_{hi} for different functional groups were extracted from Table 7.10 in the book by Krevelen.[4] V is the molar volume of the monomers.

The above solubility parameters were predicted from monomers' group contributions.[4]

Supplementary Note S2: Discussion in High Explained Variance of the First Principal Component (PC1)

In the datasets, each sample is composed of two-dimensional data, consisting of X (Strain) and Y (Stress) values. During the PCA process, the data is initially reshaped into a single row before the PCA analysis is applied. Similarly, to reconstruct the S-S curves from PCs, the single row of data is reshaped back into two rows. Given the dimensionality of 100 data points, there is a significant degree of freedom involved.

During PCA, the model first identifies a "collinearity" structure in the data. In this context, "collinearity" refers to the linear dependency between variables commonly encountered in statistics and machine learning, where one variable can be well predicted by a linear combination of another. For example, the first 50 values of the 100 data points, which correspond to the strain component, are continuously increased. Thus, the fact that PC1 accounts for 99% of the variance can be intuitively understood because these 100 values adhere to a foundational structure akin to a S-S curve. This interpretation is supported by a related work, where Yang et al. conducted PCA on the stress component of the S-S curves and found that the first three PCs could explain > 85% of the variance.[5] This suggests a generalization of the data's underlying structure by PCA.

The fact that PC1 explains 99% of the variance does not imply that other PC values are unimportant. Since explained variance is a relative measure, it merely highlights that the variances of PCs following PC1 are comparatively smaller. In **Figure 4**, we conduct an interpretability analysis of the PCA results to further elucidate these observations.

Supplementary Note S3: Root Mean Squared Error

Root mean squared error (RMSE) is a standard metric used in statistical modeling to evaluate the differences between values predicted by a model and the observed values. RMSE represents the square root of the average of the squared differences between the predicted values and the actual values. This metric is particularly sensitive to large errors, as it disproportionately weighs these errors more heavily than smaller ones, making it a useful tool for highlighting significant prediction errors. Additionally, compared to Mean Squared Error (MSE), RMSE has a scale that is closer to the original data, making it easier to be interpreted in the context of the problem domain.

The mathematical formula for RMSE is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where:

n is the number of observations,

y_i represents the actual observed values,

$\hat{y}_i$ represents the predicted values.

Supplementary Note S4: Details of the MLP Model

In this study, we utilized a Multilayer Perceptron (MLP) model to process a wide array of 85 inputs, encompassing both independent and cross-features. The model is structured with four hidden layers, each featuring a descending number of neurons (200, 100, 50, 25 respectively), ultimately leading to an output of 8 principal component (PC) values. These PC values are then reconstructed from these PCs to generate the corresponding stress-strain (S-S) curves. The model's training was facilitated using the Adam optimizer, characterized by a learning rate of 0.001 and a L1 regularization factor of 0.1. This configuration ensures effective learning and regularization to achieve accurate and reliable predictions of material properties. The architecture of the model, along with its dropout rate, learning rate, and L1 regularization, was fine-tuned through a process of grid search optimization. The performance metrics presented in Table 1 and Supplementary Table S2 represent the averages obtained over 10 experimental runs.

All computational tasks in this study were performed on a desktop computer configured with an Intel Core i7-12700K processor, an NVIDIA GeForce 2080 GPU, and 64GB of RAM. The operating system used was Ubuntu 22.04.2. Programming and implementation were carried out in Python 3.7.9. For handling data processing, we employed NumPy (version 1.19.2), Scikit-learn (version 1.0.2), and Pandas (version 1.2.1). The MLP model was developed using PyTorch version 1.13.1+cu117.

Given the relatively small size of the dataset and the simplicity of the model, training could be conducted using either CPU or GPU, with each training session taking less than 15 seconds to complete. The demonstration of 100,000 virtual experiments conducted in this study was performed using CPU inference, with the entire process taking less than 1 minute to complete.

Supplementary Figures

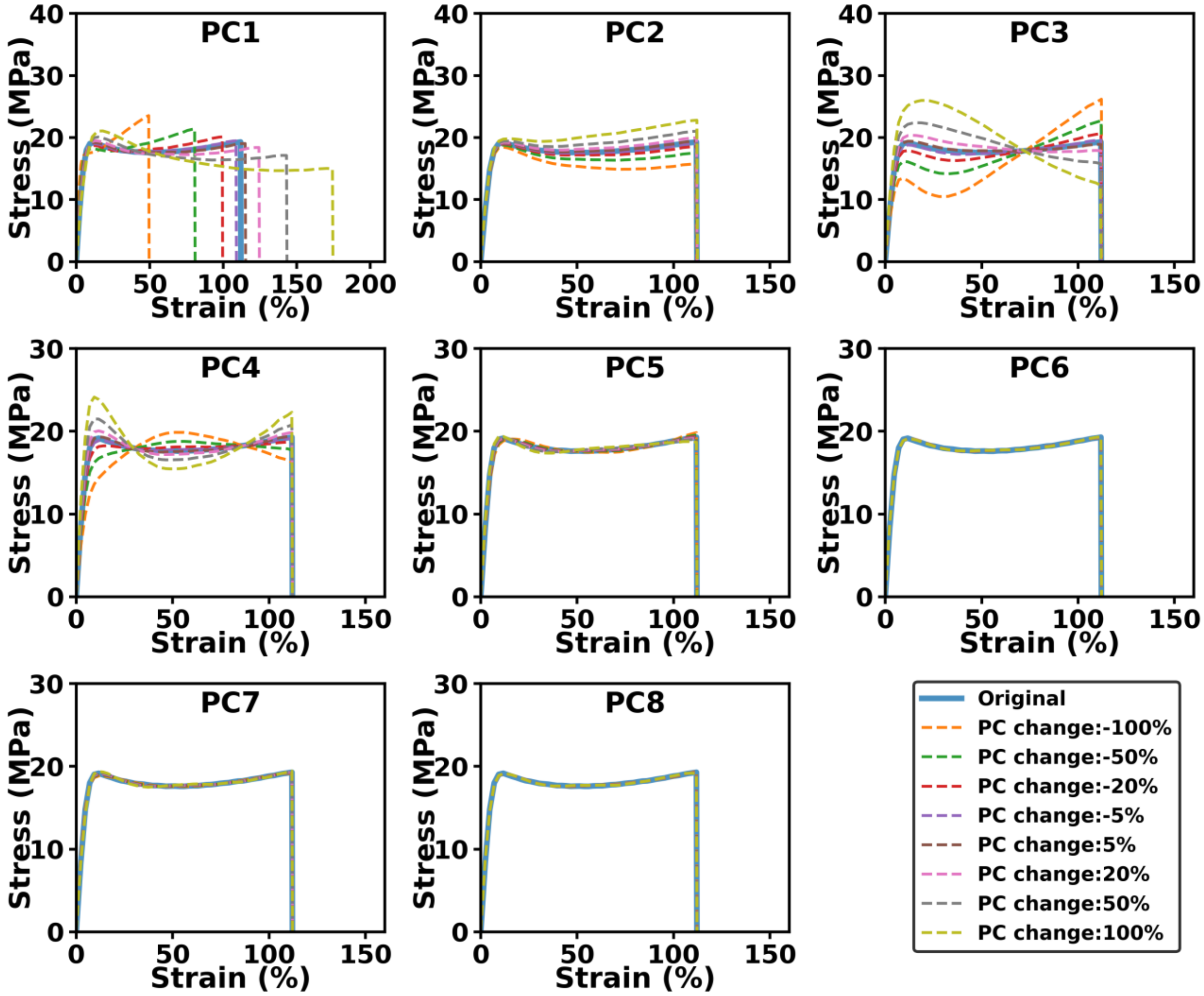

Figure S1. The impact of the changes in each principal component (PC1 to PC8) on the reconstructed stress-strain curves of a strong/tough sample.

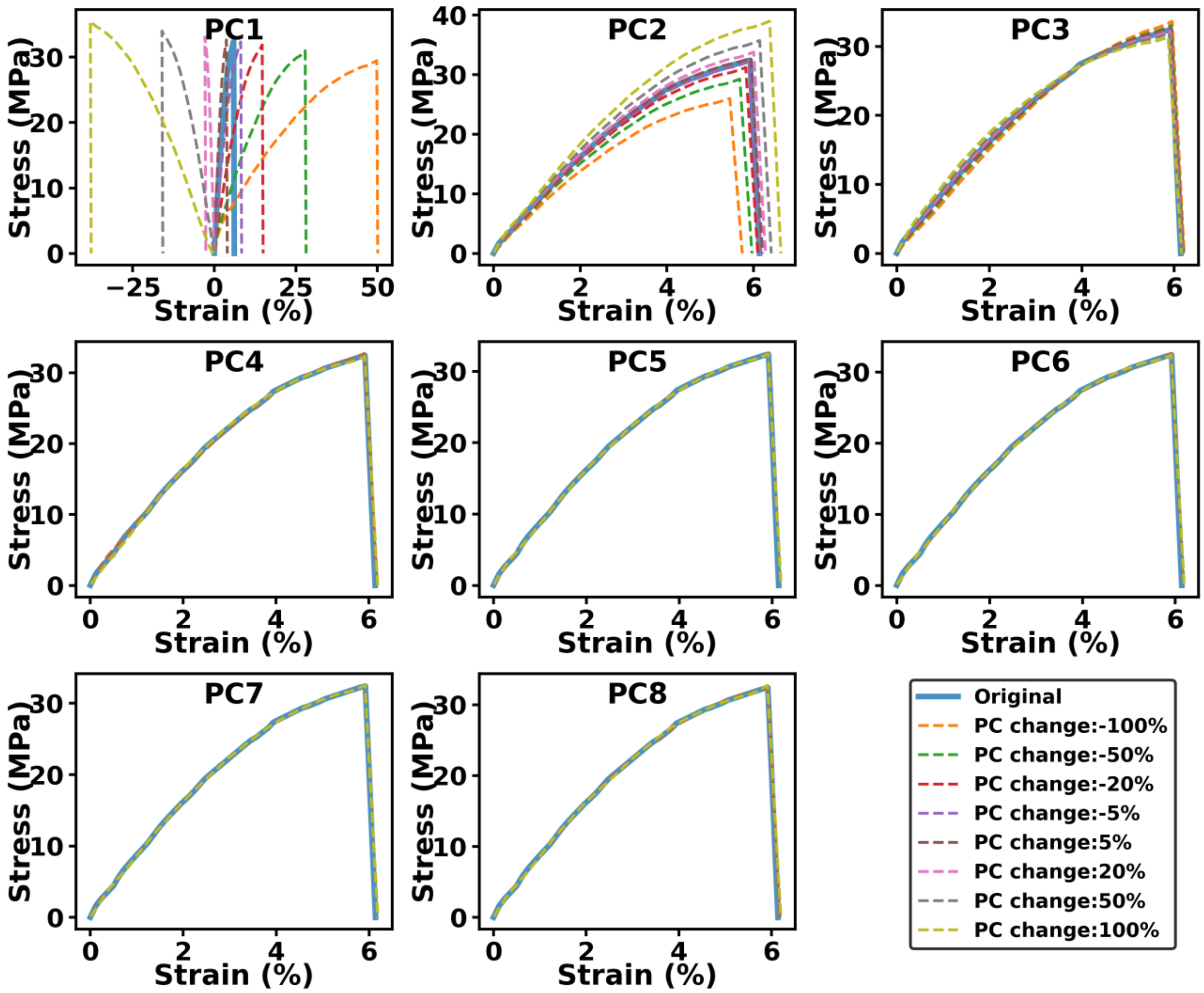

Figure S2. The impact of the changes in each principal component (PC1 to PC8) on the reconstructed stress-strain curves of a hard/brittle sample.

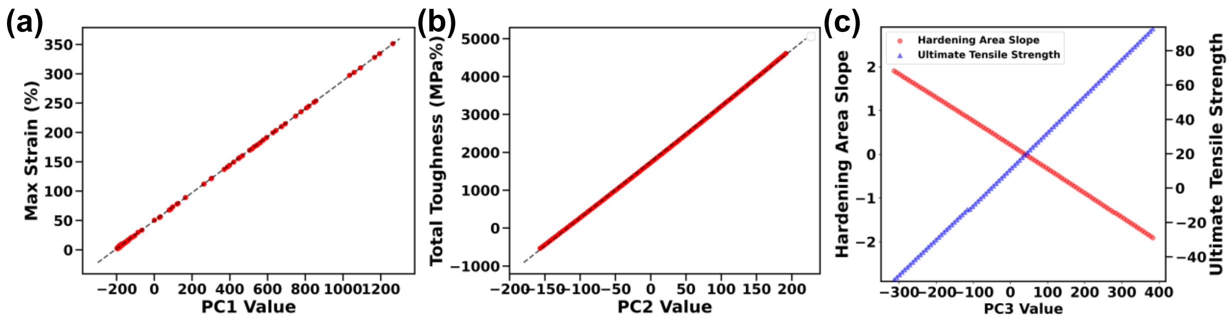

Figure S3. **(a)** Relationship between various PC1 values and the corresponding maximum strain across the entire dataset. **(b)** Relationship between different PC2 values and the corresponding toughness. Based on S-S curves in **Figure 4** and **Figure S1**, we control the other PC values and modify PC2 values to calculate the corresponding toughness. **(c)** Relationship between different PC3 values vs. the corresponding slope of the strain Hardening Area (red) and Yield Strength (blue). Based on S-S curve data in **Figure 4** and **Figure S1**, control the other PC values and modify PC3 values to calculate the corresponding features.

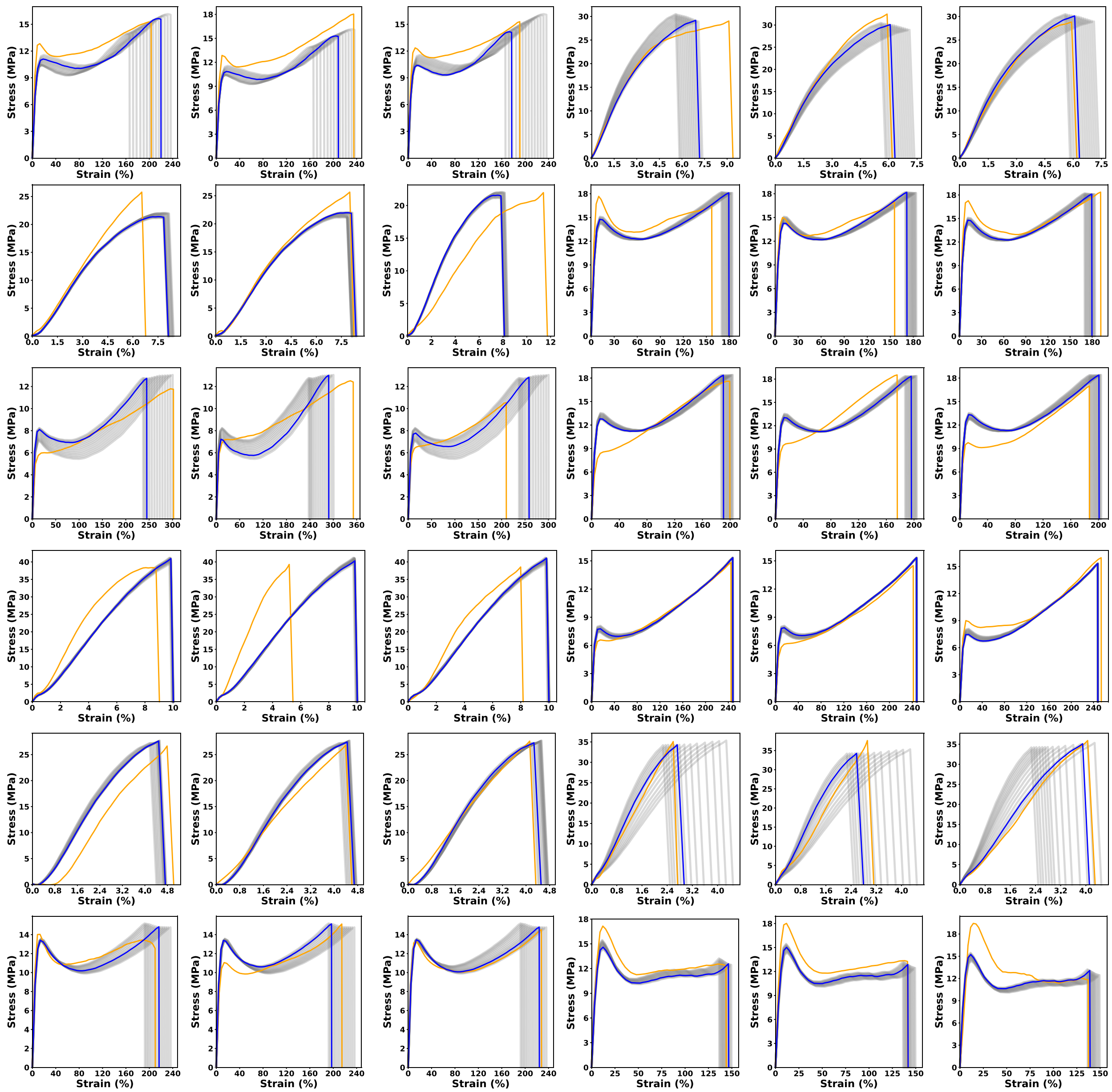

Figure S4. Comparison between 36 original (yellow) and the predicted (blue) S-S curves of the stress-strain curves (12 ink formulations). The predicted curves were reconstructed from the predicted PCs by the MLP model with the physics-informed descriptors. To effectively adapt to the variations originated from the experimental and testing conditions, the e_values were varied from -2 to 2 to reconstruct the S-S curves (grey lines).

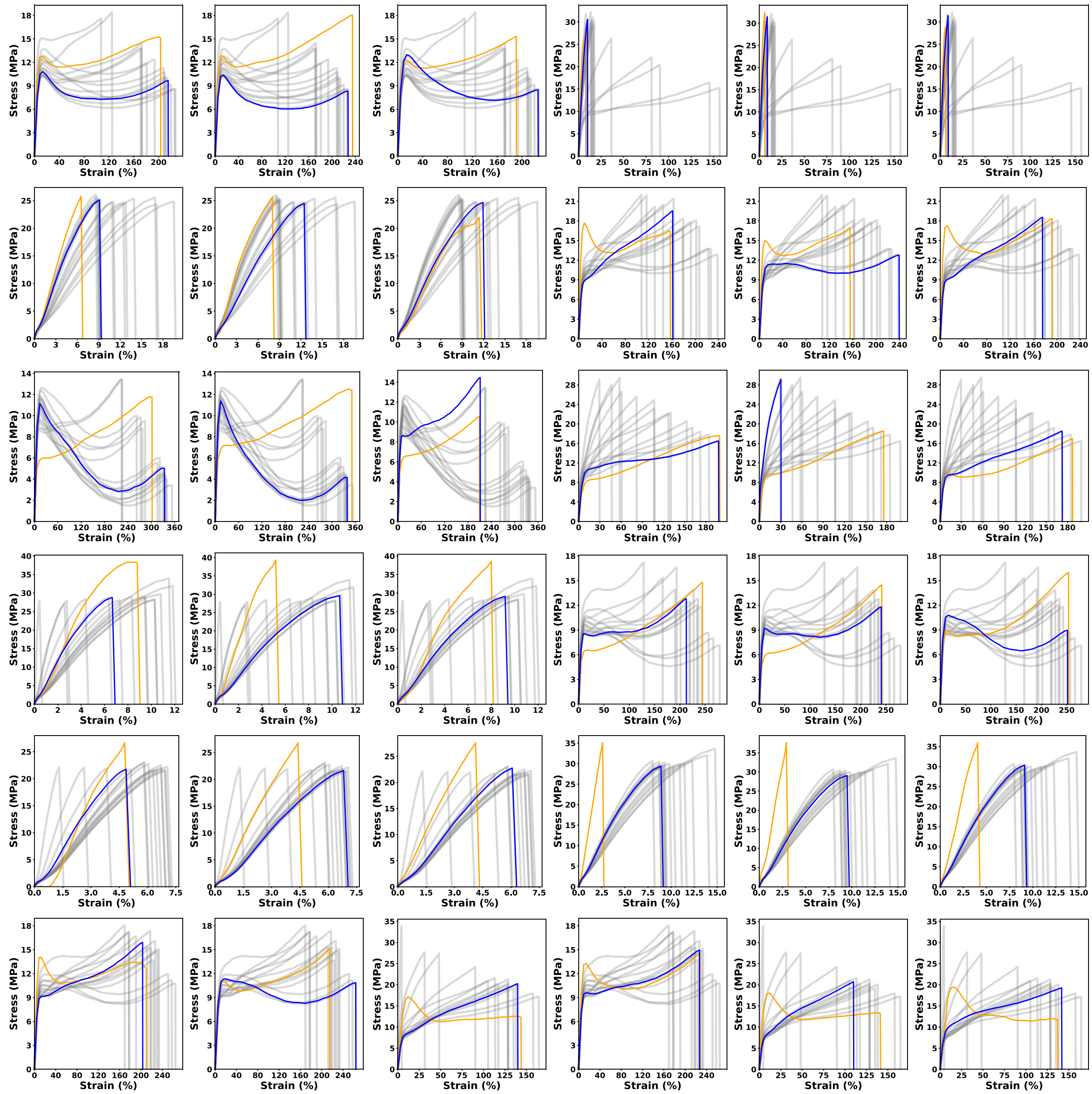

Figure S5. The comparison between original curve (yellow) and reconstructed (blue) stress-strain curves of the test set (12 ink formulation consisting of 36 stress-strain curves) using MLP model without the physics-informed descriptors. To effectively adapt to the variations originated from the experimental and testing conditions, the e_value varying from -2 to 2 were further incorporated to reconstruct the stress-strain curves (grey lines).

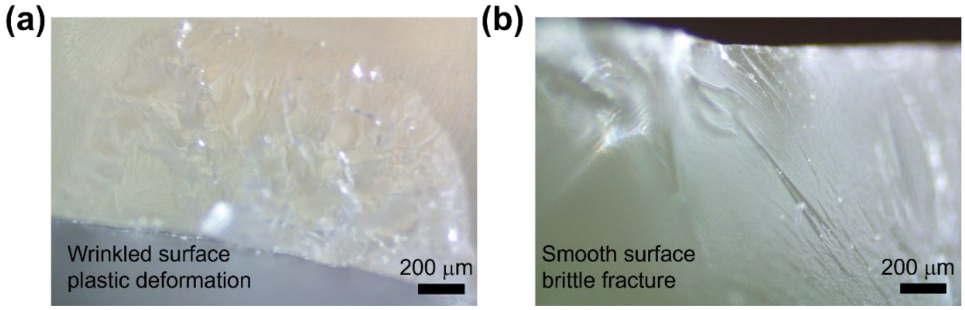

Figure S6. Digital microscope images of fracture surfaces for samples corresponding to the formulations shown in Figures 6b and 6d. (a) Fracture surface of the more ductile formulation (Figure 6b), revealing plastic deformation at the fracture points. (b) Fracture surface of the more brittle formulation (Figure 6d), exhibiting smoother surfaces indicative of a brittle failure mechanism.

Supplementary Tables

Table S1: Extracted and calculated descriptors for the six monomers.

Physics-informed descriptors	HA	IA	NVP	AA	HEAA	IBOA
Molecular Weight	222.24	184.27	111.14	72.06	115.3	208.3
Lipophilicity	1.8	4.2	0.4	0.3	-0.6	3.9
Hbond_donor	1	1	0	1	2	0
Hbond_acceptor	4	2	1	2	2	2
Rot_bond	7	1	1	1	3	3
Polar Surface Area	55.8	37.3	20.3	37.3	49.3	26.3
Heavy Atoms	16	5	8	5	8	15
Complexity	221	55.9	120	55.9	90.4	306
Total Energy	-766.6	-581.7	-364	-267.2	-401.2	-657.9
Solubility_dipole	17.49	15.57	17.87	16.04	17.34	17.99
Solubility_p	5.66	4.14	10.39	13.39	9.07	4.10
Solubility_h	12.09	4.89	8.09	17.91	15.55	4.86
Solubility_sqrt_MJperm3	22.00	16.83	22.20	27.52	25.00	19.08

Table S2: MLP model performance evaluation based on PCs without physics-informed descriptors.

PCs	R^2	RMSE	MAE	Max	Min	Range	RMSE/Range
1	0.91	132.66	63.36	1263.16	-197.45	1460.61	9.08%
2	-0.11	21.03	15.94	65.84	-23.65	89.49	23.50%
3	0.37	10.77	8.79	31.38	-26.13	57.51	18.72%
4	0.13	6.66	5.16	16.21	-12.54	28.76	23.15%
5	-0.37	3	2.11	3.4	-6.33	9.72	30.86%
6	-0.37	1.29	0.99	2.84	-2.96	5.79	22.27%
7	-0.1	1.71	1.17	6.96	-2.71	9.67	17.68%
8	-0.4	0.67	0.46	1.13	-1.61	2.74	24.45%

Table S3: MLP model performance evaluation based on the tensile strength and toughness without physics-informed descriptors.

Metric	R^2	RMSE	MAE	Max	Min	Range	RMSE/Range
Fracture strength	0.52	6.27	4.91	39.29	11.76	27.53	28.70%
Toughness	0.38	1.50	1.21	10.48	4.03	6.45	21.95%

References

(1) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2023 update. *Nucleic Acids Research* **2022**, *51* (D1), D1373-D1380. DOI: 10.1093/nar/gkac956 (acccessed 7/6/2023).

(2) Bertz, S. H. The first general index of molecular complexity. *Journal of the American Chemical Society* **1981**, *103* (12), 3599-3601. DOI: 10.1021/ja00402a071.

(3) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, *3* (1), 33. DOI: 10.1186/1758-2946-3-33.

(4) Van Krevelen, D. W.; Te Nijenhuis, K. Chapter 7 - Cohesive Properties and Solubility. In *Properties of Polymers (Fourth Edition)*, Van Krevelen, D. W., Te Nijenhuis, K. Eds.; Elsevier, 2009; pp 189-227.

(5) Yang, C.; Kim, Y.; Ryu, S.; Gu, G. X. Prediction of composite microstructure stress-strain curves using convolutional neural networks. *Materials & Design* **2020**, *189*, 108509.