Large language models, physics-based modeling, experimental measurements: the trinity of data-scarce learning of polymer properties

Ning Liu[a], Siavash Jafarzadeh[b], Brian Y. Lattimer[c], Shuna Ni[d], Jim Lua[a], Yue Yu[b,*]

[a]*Global Engineering and Materials, Inc., Princeton, NJ 08540, USA*
[b]*Department of Mathematics, Lehigh University, Bethlehem, PA 18015, USA*
[c]*Department of Mechanical Engineering, Virginia Tech, Blacksburg, VA 24060, USA*
[d]*Department of Fire Protection Engineering, University of Maryland, College Park, MD 20742, USA*

Abstract

Large language models (LLMs) bear promise as a fast and accurate material modeling paradigm for evaluation, analysis, and design. Their vast number of trainable parameters necessitates a wealth of data to achieve accuracy and mitigate overfitting. However, experimental measurements are often limited and costly to obtain in sufficient quantities for finetuning. To this end, we present a physics-based training pipeline that tackles the pathology of data scarcity. The core enabler is a physics-based modeling framework that generates a multitude of synthetic data to align the LLM to a physically consistent initial state before finetuning. Our framework features a two-phase training strategy: (1) utilizing the large-in-amount while less accurate synthetic data for supervised pretraining, and (2) finetuning the phase-1 model with limited experimental data. We empirically demonstrate that supervised pretraining is vital to obtaining accurate finetuned LLMs, via the lens of learning polymer flammability metrics where cone calorimeter data is sparse.

1. Introduction

Hundreds of millions of tons of polymer materials have been crafted for use in a colossal and ever-expanding array of applications with unceasing novel material demands, spanning automotive and ship components [1], consumer packaging [2], fabrics [3], solar cells [4], etc. The quest for qualified polymer materials for specific applications hinges on the precise prediction of their properties, driving the need for swift and accurate assessment methods. In this pursuit, machine learning (ML) based models [5–8] have emerged as a promising platform that empowers more rapid and precise polymer property prediction compared to state-of-the-art analytical and numerical methods (e.g., density functional theory [9, 10]). In particular, a variety of supervised learning methods such as graph-based neural models [11–13] have excelled in predicting polymer property tasks, largely attributed to their awareness and explicit encoding of the underlying chemical topology. Nevertheless, the vast chemical space, together with the generally restricted availability of polymer property labels, has made supervised learning cumbersome. This, in turn, has fostered the necessity for generalizable molecular representation learning and further sparked the exploitation in the

---

*Corresponding author
    Email address: `yuy214@lehigh.edu` (Yue Yu)

use of attention-based large language models (LLMs) [14–16] for polymer property prediction, where one can simply convert the intricate chemical structure to a compact string representation termed Simplified Molecular-Input Line Entry System (SMILES) [17–21]. SMILES is essentially equivalent to a flattening of the chemical graph via a depth-first pre-order traversal of the spanning tree structure, and it thereby encodes the molecular graph in an implicit manner. Applying LLMs to polymer property prediction thus amounts to a chemical language learning task where the LLM endeavors to grasp the semantic and syntactic rules of the molecular structure leveraging various unsupervised learning schemes such as masked language modeling. The pretrained LLMs, akin to chemical linguists, can then be applied for downstream tasks via finetuning using labeled data [22–26]. Pioneering work [14–16] has demonstrated the exceptional power of LLMs in predicting polymer properties and the correct geometrical interpretation of the spatial connectivity across substructures in a molecule.

On the other hand, although LLMs are powerful in aiding material assessment and novel design, the current application of LLMs resorts to directly finetuning a previously pretrained LLM with (likely limited) labeled data for downstream applications [15, 16]. This task poses grand challenge, as the vast number of trainable parameters in LLMs (in the order of tens of millions [27]) necessitates a substantial amount of experimental data to finetune the LLM in order to reach a desired level of accuracy. However, this is generally infeasible in many material design and assessment tasks, where conducting a large set of experiments is either prohibitively expensive or simply unfeasible. In this context, it calls for a highly effective training method to finetune LLMs with only a handful of experimental data.

A natural pathway to overcome this dilemma is to generate a large amount of synthetic data to pretrain the LLM, either in a supervised or unsupervised manner. Prior work [15] has explored decomposing existing polymers into chemical fragments [28], followed by randomly assembling these fragments to yield hypothetical polymers. While this method shows promise in pretraining the underlying LLM in an unsupervised manner, it is hard to judge whether the generated hypothetical polymers are physically meaningful, nor can these plausible polymers be correlated with any fundamental properties instrumental in constructing a physics-based simulation model or finetuning the final LLM decoder. In contrast, we take a different approach and investigate to incorporate physical knowledge into the synthetic polymer generation process. Specifically, we design a hypothetical polymer generation method grounded on physics-based group contribution (GC) [29], based on which a variety of fundamental material properties such as the heat of combustion and heat capacity can be calculated. These computed fundamental properties can further be used to either finetune a decoder for direct downstream applications or serve as inputs to a numerical model of the physical asset to generate a large dataset of quantities of interest. An overview of the proposed trinity framework is illustrated in Figure 1.

Besides the superiority in generating physically meaningful polymers, we show for the first time that, by simultaneously pretraining the encoder and decoder of the pretrained LLMs with physics-based synthetic data, the resulting LLMs inherit a better prior from GC, thereby achieving high predictability even in

2

Current practice
Directly finetune a pretrained LLM with limited experimental data, which can easily result in the model being trapped at local minima.

Finetuned LLM
Final finetuned LLM for downstream applications.

Phase-2 finetuning
Finetune the phase-1 LLM with limited experimental data to mitigate the effect of numerical errors and simplified assumptions inherited from the simulation model and maximize prediction accuracy.

Pretrained LLM
Any chemical LLM pretrained on a large corpus of unlabeled SMILES representations of polymer chemical structures.

Group contribution
Use group contribution to generate a multitude of physically meaningful hypothetical polymers with the correlation of a variety of fundamental thermophysical properties.

Physics-based modeling
Build a physics-based model of the physical process leveraging on the fundamental properties from group contribution as input and produce synthetic data.

Supervised pretraining
Apply supervised pretraining to the LLM in phase 1 using synthetic data to arrive at a physically consistent initial state.
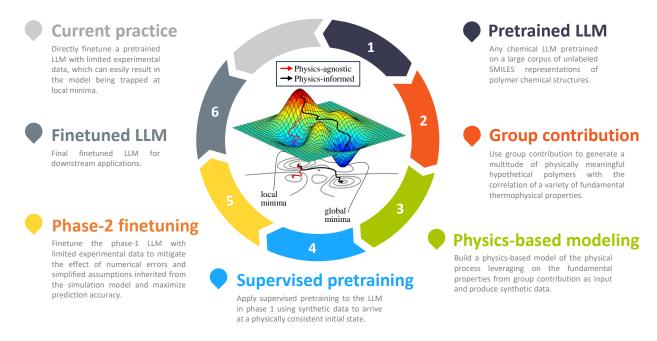
Figure 1: An overview of the proposed trinity framework for data-scarce learning of LLMs.

cases where experimental data is extremely limited. We then put forward a simple yet effective two-phase prediction-correction training strategy for finetuning pretrained LLMs towards polymer property prediction. In phase 1, we implicitly impose the physical knowledge from GC by training the encoder and decoder of pretrained LLMs using (possibly low-fidelity) physics-based synthetic data. As a result, the encoder and decoder parameters arrive at a physically consistent initial state. We then start from this initial state in phase 2 and further finetune the LLM with limited but most accurate experimental data. Analogous to pretraining the encoder of LLMs where one aims to train the encoder to learn the syntactical and grammatical rules of the underlying language [30], the rationale behind the first phase is to pretrain the encoder and decoder to understand the physical rules governing the underlying chemical formation of polymers, through a data-driven process fused with physics-based synthetic data. To distinguish our phase-1 pretraining from the unsupervised pretraining of LLM encoders, we term it *supervised pretraining*. In phase 2, the limited experimental data is leveraged to mitigate the effect of numerical errors and simplified assumptions intrinsically inherited from the simulation model and maximize prediction accuracy.

Our main contributions are:

- We propose a pathway that allows the generation of a large dataset of physically meaningful hypothetical polymers for LLM pretraining, anchored on the foundation of a physics-based GC with which a variety of fundamental properties can be correlated with the generated polymers.

- We put forth a physics-based modeling framework, leveraging on the fundamental properties obtained from GC as model input and efficiently producing a large amount of synthetic data.

- We design a two-phase prediction-correction training strategy for finetuning LLMs in scientific prob-

lems in the absence of sufficient ground truth data, where the first phase applies supervised pretraining to the encoder and decoder using the generated (possibly less accurate) synthetic dataset, and the second phase employs the limited but most accurate experimental measurements for final finetuning/correction.

- Interlocking the three pillars of LLMs, physics-based modeling and experimental measurements, we demonstrate the proposed trinity framework via the lens of learning polymer flammability metrics [31–34], where cone calorimeter data is limited to dozens, and show that, by supervisedly pretraining the LLM with synthetic data, the model's prediction accuracy can be effectively improved by more than 50%.

To the authors' best knowledge, the present work represents the first attempt that explores the generation of physically meaningful synthetic polymers for supervised pretraining of LLMs. Together with the two-phase training strategy, we present not only an effective approach that makes the best use of limited labeled data, but also, more importantly, a comprehensive LLM learning framework applicable to a broad range of scientific problems.

2. Results

Physically meaningful hypothetical polymer generation

We generate a set of structurally valid and physically meaningful hypothetical polymers based on GC [29, 35, 36] (cf. the "Methods" section). The method provides a list of groups, their contributions to a variety of physical properties, and a relationship that estimates the compound's properties from the contributing groups. To estimate a property for a compound, one typically identifies the constituent sub-molecule groups from the list and uses their contributions and the relationship to obtain an estimation. In this work, we exploit a method to generate labeled hypothetical polymers. The idea is to use various admissible combinations between groups to construct new compounds and use (4.1) to label them with associated pyrolysis kinetic parameters. An admissible combination refers to a selection of groups that can form a structurally valid polymer compound (i.e., without open valences except at polymerization points). In the current study, we sample groups with two open valences and obtain unique combinations for up to three groups, which results in a total of 3237 hypothetical polymers. Figure 2 demonstrates the formation of physically admissible polymers from constituent groups and the process of obtaining estimations for the associated pyrolysis properties.

Physics-based modeling and synthetic data generation

To generate adequate synthetic data for supervised pretraining of LLMs, a numerically accurate and computationally efficient model of the physical process needs to be constructed. In the instantiation to assessing the fire performance of polymers, we adopt the state-of-the-art fire simulation toolkit, Fire Dynamics Simulator (FDS) [37], to emulate the cone calorimeter test and predict polymer flammability metrics such as the time to ignition t_{ig} and the peak heat release rate per unit area pHRR. A brief introduction of the
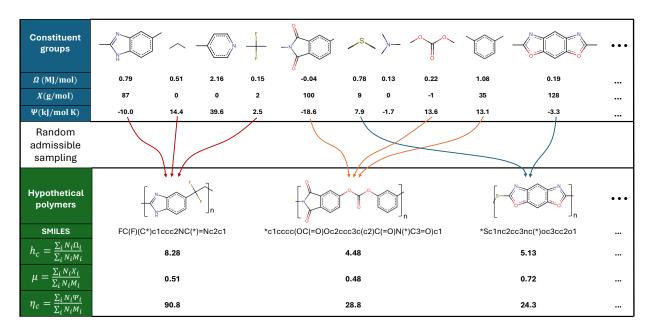
Figure 2: Using group contribution to generate structurally admissible hypothetical polymers with physically meaningful material properties.

cone calorimeter test, FDS and our modeling strategy is provided in the "Methods" section. In this context, our goal is to first build a high-fidelity FDS model that correctly simulates the thermal decomposition process of polymers and the accompanied pyrolysis gas generation that further fuels the fire. With the correct physics being captured, we can then build a reduced-order model (ROM) for rapid data generation via either simplifying the high-fidelity model or ML.

A demonstration of the workflow for physically meaningful synthetic data generation can be found in Figure 3. To simulate the cone calorimeter process in FDS, several material properties need to be defined, including thermophysical properties like density ρ, thermal conductivity κ and specific heat c_p, and kinetic pyrolysis parameters like heat release capacity η_c, heat of combustion h_c, char fraction μ, the temperature range of pyrolysis dT, and the temperature at peak mass loss rate T_p. Through GC for physically meaningful hypothetical polymer generation, the kinetic pyrolysis parameters of η_c, h_c and μ are implicitly correlated. By further assuming single reaction, the temperature range of pyrolysis dT is dependent on η_c and h_c, however without an explicit algebraic connection [29]. Additionally, previous work [29] has shown that the temperature at peak mass loss rate T_p is a function of η_c, h_c, and dT. We therefore employ two random forest regressors (RFRs) to learn these relationships, one from (η_c, h_c) to dT and the other from (η_c, h_c, dT) to T_p. The two trained RFRs reach a R^2 accuracy of 0.99/0.92 and 0.99/0.88 for training/testing, respectively. Note that GC is a low-fidelity method and can only provide estimates. To improve the quality of the synthetic labeled data, we filter out nonphysical instances. In particular, we only keep the compounds with $\mu < 1$, $h_c < 50$, $\eta_c > 0$, and $1 < dT < 300$ and discard otherwise. After filtering, we come in possession of 3237 physically meaningful hypothetical polymers labeled with their kinetic pyrolysis parameters.
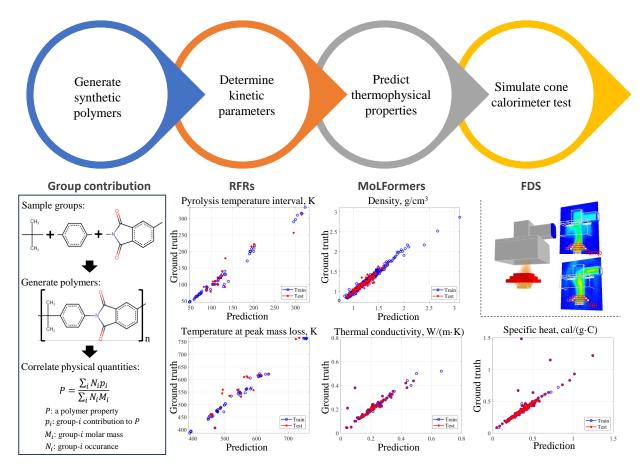
Figure 3: Demonstration of the synthetic data generation workflow.

On the other hand, since the thermophysical properties are relatively more available, we resort to the online polymer database, PolyInfo [38], to build three training datasets for density ρ, thermal conductivity κ and specific heat c_p, covering various polymer types including homopolymers and copolymers. A summary of the employed datasets is listed in Table 1. With these three datasets, we take the SMILES representations of polymers as input and train three large chemical language models to predict ρ, κ and c_p, respectively. Although the proposed framework is model-agnostic, in this work we employ MoLFormer [14, 22], an open-source pretrained molecular transformer, for demonstration. We report the test errors in terms of the relative mean squared error (MSE), and the best MoLFormers obtain 3.27%, 16.35%, and 5.82% for ρ, κ and c_p, respectively. Note that only the density dataset is randomly split into 932/100 for training and testing, respectively. Due to the relatively small datasets for κ and c_p and the good coverage of polymer types, we use all the data for training and only use the homopolymer data for testing, so as to get an accurate model for interpolation without considerably compromising generalizability. Leveraging GC, two RFRs and three MoLFormers, all the needed FDS input parameters can be computed.

With the above capability to define the input parameters for FDS, we subsequently build a high-fidelity FDS model to simulate cone calorimeter tests at a constant heat flux exposure of 50 kW/m^2, taking into

account the full solid and gas phases of the process. This model, albeit validated against available polymer cone calorimeter data and showed high accuracy [39], is computationally slow in that it takes 1-4 hours to run one simulation. We then simplify the model by only considering the solid phase of the decomposition process with an additional simulation that includes a step increase in the heat flux of 25 kW/m^2 after ignition to account for the heat flux from the flames back down onto the polymer sample. This ROM produces comparable results to the high-fidelity model and can be completed in approximately a few seconds, thus enabling rapid generation of a large synthetic dataset for supervised pretraining.
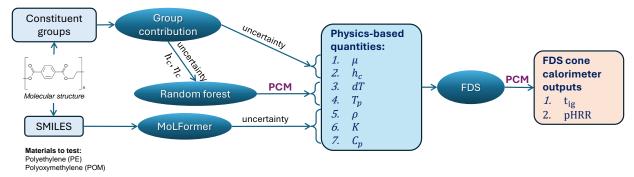
Table 1: A summary of the employed datasets.

Property	Unit	Source	Data range	# datapoints
Thermophysical				
Density ρ	g/cm^3	PolyInfo	[0.770, 2.868]	1032
Thermal conductivity κ	$W/(m \cdot K)$	PolyInfo	[0.013, 23]	110
Specific heat c_p	$cal/(g \cdot C)$	PolyInfo	[0.085, 2.520]	243
Pyrolysis kinetic				
Heat release capacity η_c	$J/(g \cdot K)$	Literature	[20.291, 1527.251]	88
Heat of combustion h_c	kJ/g	Literature	[3.823, 46.528]	88
Temperature range of pyrolysis dT	K	Literature	[46.093, 333.964]	88
Temperature at peak mass loss rate T_p	K	Literature	[386.285, 765.922]	88
Experimental cone calorimeter				
Time to ignition t_{ig}	s	Literature	[1, 538]	45
Peak heat release rate pHRR	kW/m^2	Literature	[19, 1761]	45
Smoke extinction area SEA	m^2/kg	Literature	[33, 1300]	38
Synthetic cone calorimeter				
Time to ignition t_{ig}	s	FDS	[20.2, 600]	3237
Peak heat release rate pHRR	kW/m^2	FDS	[0.117, 1864.828]	3237

Quantifying uncertainty of synthetic data

A key question is how reliable the generated synthetic data is. Both GC and RFR introduce unavoidable errors, which induce uncertainty in the FDS inputs. These uncertainties further propagate through the course of FDS simulation, resulting in posterior uncertainty in cone calorimeter predictions. To determine the trust regions for the synthetic t_{ig} and pHRR of a particular polymer, we must quantify the uncertainty of FDS simulation from its inputs. In this regard, we use an additive independent and identically distributed (i.i.d.) noise to model the errors of FDS inputs. Specifically, we follow the practice in scientific ML [40–42] and assume that the distributions of all inputs follow Gaussian distributions, $\xi \sim \mathcal{N}(E_\xi, \sigma_\xi^2)$. The only exception is char fraction μ, where a log-normal distribution is assumed: $\ln(\mu) \sim \mathcal{N}(E_{\ln(\mu)}, \sigma_{\ln(\mu)}^2)$, to guarantee the physical consistency of nonnegativity. For the kinetic parameters μ, h_c, and η_c obtained from GC, we approximate their means and standard deviations as the test results and the averaged test errors of GC. To estimate the distributions of dT and T_p, we employ the probabilistic collocation method (PCM) [43–47] on RFR models to determine the means and standard deviations. For the thermophysical inputs ρ, K, and C_p generated via MoLFormers, we use MoLFormer's predictions and their averaged test errors to estimate their

(a) Uncertainty quantification workflow

(b) Computed probability distributions of physics-based variables: (mean, std) assuming normal distributions

(c) Trust region vs. experimental measurements

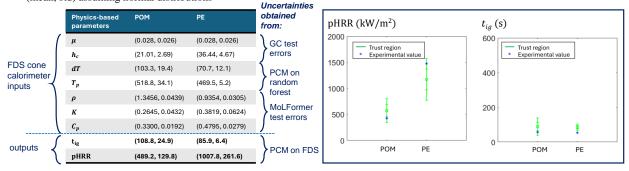| Physics-based parameters	POM	PE	Uncertainties obtained from:
μ	(0.028, 0.026)	(0.028, 0.026)	GC test errors
h_c	(21.01, 2.69)	(36.44, 4.67)	
dT	(103.3, 19.4)	(70.7, 12.1)	PCM on random forest
T_p	(518.8, 34.1)	(469.5, 5.2)	
ρ	(1.3456, 0.0439)	(0.9354, 0.0305)	
K	(0.2645, 0.0432)	(0.3819, 0.0624)	MoLFormer test errors
C_p	(0.3300, 0.0192)	(0.4795, 0.0279)	
t_{ig}	(108.8, 24.9)	(85.9, 6.4)	PCM on FDS
pHRR	(489.2, 129.8)	(1007.8, 261.6)	

FDS cone calorimeter inputs — outputs

Figure 4: Quantifying uncertainties on the generated synthetic data: (a) general workflow, (b) the computed probability distributions of physics-based variables, and (c) trust region vs. experimental measurement plots.

distributions. With the standard deviations for all seven FDS inputs obtained, we again apply PCM to the FDS model and find the distributions for t_{ig} and pHRR. To verify this process, we consider two polymers from the synthetic dataset that correspond to known polymers with available experimental measurements, namely Polyethylene (PE) and Polyoxymethylene (POM). Figure 4b lists the means and standard deviations for the FDS inputs and outputs. Figure 4c displays the trust regions for t_{ig} and pHRR predicted by the data generation method versus experimental measurements for PE and POM. The green segments span four standard deviations corresponding to 95% of the possible values. The inclusion of the measured values within or near the trust regions demonstrates the reliability and trust-worthiness of the synthetic data generation approach.

Phase-1 supervised pretraining

A significant advantage of the proposed trinity framework is the supervised pretraining of LLMs that aims to align the model to a physically consistent initial state prior to finetuning. In the circumstance of cone calorimeter test, we pretrain the LLM with labeled synthetic data in terms of time to ignition t_{ig} and peak heat release rate per unit area pHRR. In particular, we first categorize the synthetic data into two classes, one with ignitable polymers (indicated by $t_{ig} \leq 600s$) and the other with un-ignitable polymers. We slightly modify the MoLFormer architecture by adding a sigmoid function to the decoder output together with the binary cross-entropy loss as the loss function to train a classifier that identifies if a given polymer is
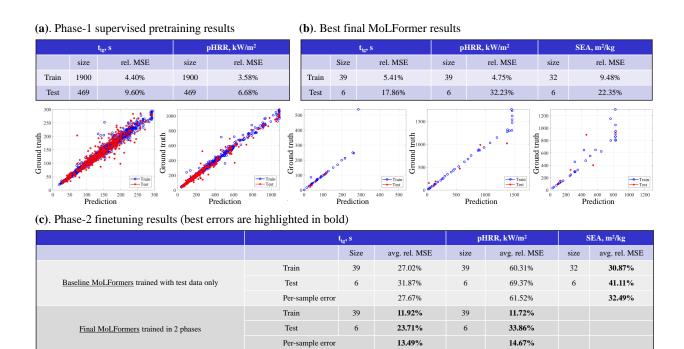
(a). Phase-1 supervised pretraining results

	t_{ig}, s		pHRR, kW/m²	
	size	rel. MSE	size	rel. MSE
Train	1900	4.40%	1900	3.58%
Test	469	9.60%	469	6.68%

(b). Best final MoLFormer results

	t_{ig}, s		pHRR, kW/m²		SEA, m²/kg	
	Size	rel. MSE	size	rel. MSE	size	rel. MSE
Train	39	5.41%	39	4.75%	32	9.48%
Test	6	17.86%	6	32.23%	6	22.35%

(c). Phase-2 finetuning results (best errors are highlighted in bold)

		t_{ig}, s		pHRR, kW/m²		SEA, m²/kg	
		Size	avg. rel. MSE	size	avg. rel. MSE	size	avg. rel. MSE
Baseline MoLFormers trained with test data only	Train	39	27.02%	39	60.31%	32	**30.87%**
	Test	6	31.87%	6	69.37%	6	**41.11%**
	Per-sample error		27.67%		61.52%		**32.49%**
Final MoLFormers trained in 2 phases	Train	39	**11.92%**	39	**11.72%**		
	Test	6	**23.71%**	6	**33.86%**		
	Per-sample error		**13.49%**		**14.67%**		

Figure 5: MoLFormer training results for cone calorimeter metrics prediction.

ignitable. We then use the ignitable synthetic polymers for supervised pretraining. This pre-processing step is critical because un-ignitable polymers are characterized by a constant t_{ig} of 600 s and very small pHRRs, which could otherwise introduce unbalanced outputs and deteriorate prediction accuracy. By implementing a pre-classification step, we address this data distribution challenge by focusing on the ignitable polymers of interest, thereby enhancing the model's predictability. Next, we randomly split the 2369 ignitable polymers into 1900 for training and 469 for testing. The supervised pretraining results can be found in Figure 5a, where the best trained MoLFormers achieve training and test errors of 4.40%/9.60% and 3.58%/6.68% for t_{ig} and pHRR, respectively. With these superb training results, the two MoLFormers are anticipated to inherit good physical priors from the physics-based synthetic data and qualify as initial models for phase-2 finetuning.

Phase-2 finetuning

In phase 2, we finetune the obtained phase-1 models with good physics-based priors. Due to the limited size of the experimental cone calorimeter datasets (cf. Table 1), we randomly split each dataset 5 times and train 5 MoLFormers for both t_{ig} and pHRR to mitigate the effect of bias in splitting data. We then report the averaged relative MSE in Figure 5c, together with the results from the best MoLFormers (out of the 5 trained) in Figure 5b. In general, we observe a good agreement between the ground truth data and the prediction, as indicated by their alignment along the $y = x$ line. We note that during training, the loss function is defined based on relative MSE. As such, the denominator becomes smaller for polymers with smaller pHRR and SEA values, which is equivalent to placing a larger weight on these polymers. Consequently, the model is anticipated to provide better predictions for polymers with lower flammability.

This loss function is chosen because optimal polymer structures with lowest flammability are often more favorable [48]. The effect of the relative MSE loss can also be observed in our results shown in Figure 5, where the model generally presents larger errors in the higher pHRR and SEA regimes, underscoring the importance of an application-oriented loss function. We also point out that with a proper selection of suitable loss functions or simply the augmentation of more data of large pHRRs and SEAs, the trained MoLFormers can be further finetuned to reach higher accuracy. This is, however, out of the scope of the current study.

Enhanced predictability of the learned models

To demonstrate the enhanced predictability of the final MoLFormers utilizing the 2-phase training approach, we compare their accuracy with the baseline MoLFormers trained with test data only in Figure 5c. Note that, due to the inability of FDS to generate synthetic SEA data, we do not compare the performance in terms of SEA. However, synthetic SEA data can be generated using other physics-based methods such as the smoke calculation method in [49], which is left as a future research direction. With this caveat in mind, by comparing qualitatively across the prediction plots in Figure 5b, a much wider dispersion from the $y = x$ line together with an early-stage deviation is observed in the SEA plot, emphasizing the indispensable need of the phase-1 supervised pretraining. By comparing the performance between the baseline single-phase MoLFormers (i.e., trained with test data only) and our final physics-guided MoLFormers in Figure 5c, the final MoLFormers manifest significantly higher accuracy in the prediction of both t_{ig} and pHRR, where the test errors are improved by 25.6% and 51.2%, respectively. Note that, since the best models are chosen based on the best test errors, the training errors in the best models are also effectively reduced. By comparing the per-sample errors between the baseline and final MoLFormers, a remarkable improvement of 51.2% and 76.2% is observed in the prediction of t_{ig} and pHRR, respectively. These results highlight the effectiveness of the proposed 2-phase training framework in imposing physics-based priors and tackling the pathology of data scarcity.

3. Discussion

We present a physics-guided ML pipeline tailored for finetuning LLMs with limited data availability, in which a 2-phase training strategy is designed to blend physics knowledge with pretrained LLMs and finite experimental measurements. Our training strategy begins with a supervised pretraining phase (phase 1), wherein physics-based synthetic data is leveraged to lead the LLM to a physically consistent initial state that establishes the prior physical knowledge for subsequent finetuning. In phase 2, the obtained phase-1 model, having acquired a condensed representation of polymers and a deep physical understanding of the downstream problem, undergoes finetuning that makes the best use of limited experimental data, thereby achieving the final LLM with enhanced precision.

The realization of the proposed 2-phase training framework relies on a suite of physics-based numerical models for efficient synthetic data generation. On one hand, group contribution plays a critical role, not only in generating a vast array of hypothetical polymers for LLM pretraining, but also in correlating each

generated polymer with a diverse range of fundamental physical properties that materialize the hypothetical polymer with actual physical significance. On the other hand, the proposed physics-based modeling framework of the physical process enables the production of a multitude of synthetic data for supervised pretraining. Another outstanding feature of the 2-phase training strategy lies in its relaxation of the stringent accuracy requirement on synthetic data, as the phase-2 finetuning aims to utilize the most accurate experimental data to correct the phase-1 prediction and mitigate the impact of numerical errors and simplified assumptions intrinsically inherited from the numerical simulation model, thus maximizing prediction accuracy.

Interlocking the three pillars of scientific machine learning—LLMs, physics-based modeling and experimental measurements—we demonstrate the proposed trinity framework via the lens of learning polymer flammability metrics, where cone calorimeter data is limited (to only a few dozen measurements). We show that, through supervised pretraining of the LLM with synthetic data, the model's per-sample accuracy can be effectively enhanced by at least 50%. This extraordinary capability in making accurate predictions while relaxing the requirement of an excessive amount of expensive experimental data enables the exploration of the gigantic polymer design universe and the discovery of optimal polymer structures for downstream applications in a precise and efficient manner.

4. Methods

SMILES canonicalization

SMILES (Simplified Molecular Input Line Entry Systems) is developed in the 1980s [17] and can be viewed as a chemical language for describing molecular structures. While there can be more than one SMILES expressions for a given structure, canonical SMILES are unique. Canonicalization is an algorithm that maps any valid SMILES for a given structure to the unique (canonical) version for that molecule [50].

Since the chemical structures of polymers can be described by their repeating units, researchers have adopted SMILES notations to represent polymers (e.g., [51] and [52]). In this work, we employ a similar notation as in [52], where the polymerization point is denoted by an asterisk (i.e., "*"). For copolymers, the corresponding SMILES consists of SMILES of the constituent monomers, connected by dots (i.e., ".").

Group contribution

The group contribution (GC) method is a low-fidelity methodology for estimating various properties of chemical compounds by dividing them into smaller constituents (i.e., sub-molecular structures), sometimes referred to as molar groups. A compound's property is then assumed to be computable according to the *contribution* of the constituent molar groups [53]. For any particular property, and given a dataset containing this property for a number of compounds, one needs to gather a finite set of molar groups that describe all the given compounds and any other compounds for which one aims to obtain the property. Then, a weighted linear equation is assumed that relates the property to the contribution of molar groups, and a regression

problem is solved to obtain the group contributions. Once the contributions are found, they can be used to estimate the properties for compounds outside of the original dataset.

The selection of molar groups and the equation that relates properties to the group contributions are not unique and vary across applications. In this work, we use the GC method by Lyon [29] that provides estimates for heat release capacity η_c, specific heat of combustion h_c, and char fraction μ for polymer compounds based on a set of 38 molar groups, using:

$$\eta_c = \sum_{i=1}^{n_g} \frac{N_i \Psi_i}{N_i M_i}, \quad h_c = \sum_{i=1}^{n_g} \frac{N_i \Omega_i}{N_i M_i}, \quad \mu = \sum_{i=1}^{n_g} \frac{N_i X_i}{N_i M_i}, \tag{4.1}$$

where n_g is the number of distinct groups in the compound, N_i is the number of occurrence of group i in the compound, M_i is the molar mass of the group i, and Ψ, Ω_i, X_i are contributions of group i to the η_c, h_c, μ properties, respectively.

Numerical modeling of the physical process

Two essential ingredients of a synthetic data generator are the numerical accuracy and computational efficiency of the simulation model. The model needs to be numerically validated against experimental measurements to produce trustworthy data. This is often achieved via constructing a high-fidelity model that simulates the complete process of the physical asset. Concurrently, the model should also be computationally efficient to be able to generate a large amount of data in a relatively short amount of time. This is often accomplished via reduced-order modeling of the high-fidelity process.

We instantiate the modeling framework of the physical process in the form of cone calorimeter experiments [54] that are used to study fire behaviors in small samples of condensed-phase materials. In this case, Fire Dynamics Simulator (FDS) [37] is adopted as the computational framework. FDS is a computational fluid dynamics (CFD) model based on large-eddy simulation developed by the National Institute of Standards and Technology, which predicts the thermo-chemo-physical response of the surrounding solid-phase materials and their potential contribution to the fire through the generation of pyrolysis gases that create fuel for the fire. In this context, we construct a high-fidelity FDS model to predict the fire behavior of polymers exposed to a constant heat flux from the cone heater. A high-fidelity simulation model is constructed that performs a single simulation to predict the temperature rise and the decomposition of polymers over time in the solid phase as well as the combustion of pyrolysis gases in the gas phase. This model, albeit being numerically accurate against experimental data, is computationally slow in that it takes 1-4 hours to run one simulation. To alleviate the computational burden, we proceed to build a reduced-order model (ROM) by only considering the solid-phase mode with an additional simulation that includes a step increase in the heat flux of 25 kW/m^2 after ignition to account for the heat flux from the flames back down onto the polymer sample. This ROM produces comparable results to the high-fidelity model and can be completed in approximately one second, thus enabling rapid generation of a large synthetic dataset for supervised pretraining.

Probabilistic collocation for uncertainty quantification

The Probabilistic Collocation Method (PCM) [43–47, 55–57] is a popular numerical method for stochastic models. It retains the ease of implementation found in Monte Carlo methods, as only solutions at the sample points are needed. This makes PCM particularly suitable in cases where the deterministic model operates as a black box or pre-compiled/pre-trained software. Additionally, when combined with sparse grids such as the Smolyak formulation [58], PCM reduces the number of sample points needed to achieve a given numerical accuracy, especially for problems with relatively small (≤ 50) random dimensions [47] and sufficient solution smoothness in the parameter space [56]. These attributes make PCM an ideal candidate for our application.

We elaborate on PCM for a given stochastic model, $f(\boldsymbol{x}, \boldsymbol{\xi})$, where $\boldsymbol{x} \in \mathbb{R}^{d_x}$ denotes physical parameters and $\boldsymbol{\xi} \in \Omega_\xi$ represents random variables of the random field. Here, Ω_ξ is the random space. It is often assumed that all elements of $\boldsymbol{\xi}$ are i.i.d. random variables with a probability density $\rho : \Omega_\xi \to \mathbb{R}^+$. The goal of PCM is to construct an approximate solution manifold of $f(\boldsymbol{x}, \boldsymbol{\xi})$ based on Lagrange interpolation in the random space. Specifically, let $\Theta_N = \{\boldsymbol{\xi}_k\}_{k=1}^K \subset \Omega_\xi$ be a set of prescribed nodes such that the Lagrange interpolation can be performed in the random space Ω_ξ, where N is the dimension of the parametric space. Then, the function $f(\boldsymbol{x}, \cdot) : \Omega_\xi \to \mathbb{R}$ can be approximated using the Lagrange interpolation polynomial:

$$\mathcal{J}[f](\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{k=1}^K f(\boldsymbol{x}, \boldsymbol{\xi}_k) J_k(\boldsymbol{\xi}) \ ,$$

where $J_k(\boldsymbol{\xi})$ is the Lagrange polynomial satisfying $J_k(\boldsymbol{\xi}_j) = \delta_{kj}$. Then, the statistical moments of the stochastic model f can be evaluated as follows:

$$\mathbb{E}[f](\boldsymbol{x}) \approx \int_\Gamma \sum_{k=1}^K f(\boldsymbol{x}, \boldsymbol{\xi}_k) J_k(\boldsymbol{\xi}) \rho(\boldsymbol{\xi}) d\boldsymbol{\xi} \ , \ \ \sigma[f](\boldsymbol{x}) \approx \sqrt{\int_\Gamma \left[\sum_{k=1}^K f(\boldsymbol{x}, \boldsymbol{\xi}_k) J_k(\boldsymbol{\xi}) \right]^2 \rho(\boldsymbol{\xi}) d\boldsymbol{\xi} - [\mathbb{E}[f](\boldsymbol{x})]^2} \ , \quad (4.2)$$

and so on. In our application, we take f as the RFR model (when estimating the uncertainty of dT and T_p) or the FDS model (when estimating the uncertainty in cone calorimeter tests). We evaluate the RFR and FDS models at each collocation point, $\boldsymbol{\xi}_k$, producing a set of system responses corresponding to different input values. Using these responses, we can then estimate the statistical moments of the output distribution, such as the mean, variance, and higher-order moments, following (4.2).

Large chemical language models for polymer property prediction

Without loss of generality, we employ the pretrained large chemical language model, MolLFormer [14], as our LLM backbone. MolLFormer is an efficient transformer encoder model with linear attention mechanism and rotary positional embeddings. Although the proposed trinity framework is applicable to any large chemical language model (i.e., it is LLM-agnostic), we choose MoLFormer for demonstration because it is pre-trained on a relatively large database of 1.1 billion unlabeled SMILES representations of compounds from the PubChem and ZINC datasets. This pretraining database covers a diverse range of real-world compounds and positions MoLFormer as an arguably more general-purpose chemical linguist. It has been shown that MolLFormer can capture an adequate amount of structural and chemical information for accurate predictions

of a diverse array of downstream applications [14].

Two-phase training strategy

The two-phase prediction-correction training strategy consists of a supervised pretraining phase with a myriad of synthetic data, and a finetuning phase using limited experimental data. In the first phase, we aim to instil in the encoder and decoder the physical knowledge governing the underlying problem, via a data-driven process using a multitude of synthetic data. The rationale behind this phase closely resembles the unsupervised pretraining of transformer encoders in most LLMs that aims to pretrain the encoder to understand the syntactical and grammatical rules of the chemical language. Another significant advantage of the proposed strategy is that it relaxes the stringent requirement on the accuracy of the synthetic dataset. In fact, the phase-2 finetuning leverages the most accurate experimental measurements, treated as ground truth data, to correct the phase-1 predictions and mitigate the effects of numerical errors and simplified assumptions intrinsically inherited from the numerical simulation model. In this way, our two-phase training strategy enhances prediction accuracy by seamlessly blending prior physical knowledge with experimental measurements.

References

[1] P. Tran, Q. T. Nguyen, K. Lau, Fire performance of polymer-based composites for maritime infrastructure, Composites Part B: Engineering 155 (2018) 31–48.

[2] C. Silvestre, D. Duraccio, S. Cimmino, Food packaging based on polymer nanomaterials, Progress in polymer science 36 (12) (2011) 1766–1782.

[3] C. Fu, Z. Wang, Y. Gao, J. Zhao, Y. Liu, X. Zhou, R. Qin, Y. Pang, B. Hu, Y. Zhang, et al., Sustainable polymer coating for stainproof fabrics, Nature Sustainability 6 (8) (2023) 984–994.

[4] G. Li, R. Zhu, Y. Yang, Polymer solar cells, Nature photonics 6 (3) (2012) 153–161.

[5] H. Doan Tran, C. Kim, L. Chen, A. Chandrasekaran, R. Batra, S. Venkatram, D. Kamal, J. P. Lightstone, R. Gurnani, P. Shetty, et al., Machine-learning predictions of polymer properties with polymer genome, Journal of Applied Physics 128 (17) (2020).

[6] E. Kazemi-Khasragh, J. P. F. Blázquez, D. G. Gómez, C. González, M. Haranczyk, Facilitating polymer property prediction with machine learning and group interaction modelling methods, International Journal of Solids and Structures 286 (2024) 112547.

[7] C. Kim, R. Batra, L. Chen, H. Tran, R. Ramprasad, Polymer design using genetic algorithm and machine learning, Computational Materials Science 186 (2021) 110067.

[8] L. Tao, V. Varshney, Y. Li, Benchmarking machine learning models for polymer informatics: an example of glass transition temperature, Journal of Chemical Information and Modeling 61 (11) (2021) 5395–5413.

[9] W. E. McMullen, K. F. Freed, A density functional theory of polymer phase transitions and interfaces, The Journal of chemical physics 92 (2) (1990) 1413–1426.

[10] J. Wu, Density functional theory for chemical engineering: From capillarity to soft materials, AIChE journal 52 (3) (2006) 1169–1193.

[11] J. Park, Y. Shim, F. Lee, A. Rammohan, S. Goyal, M. Shim, C. Jeong, D. S. Kim, Prediction and interpretation of polymer properties using the graph convolutional network, ACS Polymers Au 2 (4) (2022) 213–222.

[12] R. Gurnani, C. Kuenneth, A. Toland, R. Ramprasad, Polymer informatics at scale with multitask graph neural networks, Chemistry of Materials 35 (4) (2023) 1560–1567.

[13] N. Liu, Y. Yu, H. You, N. Tatikola, Ino: Invariant neural operators for learning complex physical systems with momentum conservation, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2023, pp. 6822–6838.

[14] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, P. Das, Large-scale chemical language representations capture molecular structure and properties, Nature Machine Intelligence 4 (12) (2022) 1256–1264.

[15] C. Kuenneth, R. Ramprasad, polybert: a chemical language model to enable fully machine-driven ultrafast polymer informatics, Nature Communications 14 (1) (2023) 4099.

[16] C. Xu, Y. Wang, A. Barati Farimani, Transpolymer: a transformer-based language model for polymer property predictions, npj Computational Materials 9 (1) (2023) 64.

[17] D. Weininger, Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules, Journal of chemical information and computer sciences 28 (1) (1988) 31–36.

[18] G. B. Goh, N. O. Hodas, C. Siegel, A. Vishnu, Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties, arXiv preprint arXiv:1712.02034 (2017).

[19] H. Öztürk, A. Özgür, E. Ozkirimli, Deepdta: deep drug–target binding affinity prediction, Bioinformatics 34 (17) (2018) i821–i829.

[20] A. Paul, D. Jha, R. Al-Bahrani, W.-k. Liao, A. Choudhary, A. Agrawal, Chemixnet: Mixed dnn architectures for predicting chemical properties using multiple molecular representations, arXiv preprint arXiv:1811.08283 (2018).

[21] B. Shin, S. Park, K. Kang, J. C. Ho, Self-attention based molecule representation for predicting drug-target interaction, in: Machine learning for healthcare conference, PMLR, 2019, pp. 230–248.

[22] F. Wu, D. Radev, S. Z. Li, Molformer: Motif-based transformer on 3d heterogeneous molecular graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 5312–5320.

[23] L. Maziarka, D. Majchrowski, T. Danel, P. Gainski, J. Tabor, I. Podolak, P. Morkisz, S. Jastrzebski, Relative molecule self-attention transformer, Journal of Cheminformatics 16 (1) (2024) 3.

[24] Y. Wang, S. Li, T. Wang, B. Shao, N. Zheng, T.-Y. Liu, Geometric transformer with interatomic positional encoding, Advances in Neural Information Processing Systems 36 (2024).

[25] J. Ross, B. Belgodere, S. C. Hoffman, V. Chenthamarakshan, Y. Mroueh, P. Das, Gp-molformer: A foundation model for molecular generation, arXiv preprint arXiv:2405.04912 (2024).

[26] B. Belgodere, V. Chenthamarakshan, P. Das, P. Dognin, T. Kurien, I. Melnyk, Y. Mroueh, I. Padhi, M. Rigotti, J. Ross, et al., Cloud-based real-time molecular screening platform with molformer, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2022, pp. 641–644.

[27] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, V. Pande, Moleculenet: a benchmark for molecular machine learning, Chemical science 9 (2) (2018) 513–530.

[28] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, M. Rarey, On the art of compiling and using'drug-like'chemical fragment spaces, ChemMedChem 3 (10) (2008) 1503.

[29] R. E. Lyon, M. T. Takemori, N. Safronava, S. I. Stoliarov, R. N. Walters, A molecular basis for polymer flammability, Polymer 50 (12) (2009) 2608–2617.

[30] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al., Chatgpt for good? on opportunities and challenges of large language models for education, Learning and individual differences 103 (2023) 102274.

[31] R. E. Lyon, M. L. Janssens, et al., Polymer flammability, Tech. rep., United States. Federal Aviation Administration. Office of Aviation Research (2005).

[32] S. M. Kraft, B. Y. Lattimer, C. B. Williams, Flammability of 3-d printed polymers–composition and geometry factors (2016).

[33] K. Carpenter, M. Janssens, Using heat release rate to assess combustibility of building products in the cone calorimeter, Fire technology 41 (2005) 79–92.

[34] V. Babrauskas, The development and evolution of the cone calorimeter: a review of 12 years of research and standardization, Fire Standards in the International Marketplace (1995).

[35] R. N. Walters, R. E. Lyon, Molar group contributions to polymer flammability, Journal of Applied Polymer Science 87 (3) (2003) 548–563.

[36] A. S. Hukkerikar, B. Sarup, A. Ten Kate, J. Abildskov, G. Sin, R. Gani, Group-contribution+ (gc+) based estimation of properties of pure components: Improved property estimation and uncertainty analysis, Fluid Phase Equilibria 321 (2012) 25–43.

[37] K. McGrattan, S. Hostikka, R. McDermott, J. Floyd, C. Weinschenk, K. Overholt, Fire dynamics simulator user's guide, NIST special publication 1019 (6) (2013) 1–339.

[38] S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu, M. Yamazaki, Polyinfo: Polymer database for polymeric materials design, in: 2011 International Conference on Emerging Intelligent Data and Web Technologies, IEEE, 2011, pp. 22–29.

[39] J. L. Hodges, B. Y. Lattimer, A. Kapahi, J. E. Floyd, An engineering model for the pyrolysis of materials, Fire safety journal 141 (2023) 103980.

[40] C. Moya, A. Mollaali, Z. Zhang, L. Lu, G. Lin, Conformalized-deeponet: A distribution-free framework for uncertainty quantification in deep operator networks, arXiv preprint arXiv:2402.15406 (2024).

[41] Z. Zou, X. Meng, G. E. Karniadakis, Uncertainty quantification for noisy inputs-outputs in physics-informed neural networks and neural operators, arXiv preprint arXiv:2311.11262 (2023).

[42] A. Ghosh, S. V. Kalinin, M. A. Ziatdinov, Discovery of structure–property relations for molecules via hypothesis-driven active learning over the chemical space, APL Machine Learning 1 (4) (2023).

[43] D. Xiu, J. S. Hesthaven, High-order collocation methods for differential equations with random inputs, SIAM Journal on Scientific Computing 27 (3) (2005) 1118–1139.

[44] F. Nobile, R. Tempone, C. G. Webster, An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data, SIAM Journal on Numerical Analysis 46 (5) (2008) 2411–2442.

[45] X. Ma, N. Zabaras, An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations, Journal of Computational Physics 228 (8) (2009) 3084–3113.

[46] G. Zhang, M. Gunzburger, Error analysis of a stochastic collocation method for parabolic partial differential equations with random input data, SIAM Journal on Numerical Analysis 50 (4) (2012) 1922–1940.

[47] G. Lin, A. M. Tartakovsky, An efficient, high-order probabilistic collocation method on sparse grids for three-dimensional flow and solute transport in randomly heterogeneous porous media, Advances in Water Resources 32 (5) (2009) 712–722.

[48] S. Bourbigot, S. Duquesne, C. Jama, Polymer nanocomposites: how to reach low flammability?, in: Macromolecular Symposia, Vol. 233, Wiley Online Library, 2006, pp. 180–190.

[49] A. Tewarson, A. Tewarson, Smoke point height and fire properties of materials, US Department of Commerce, National Institute of Standards and Technology, 1988.

[50] D. Weininger, A. Weininger, J. L. Weininger, Smiles. 2. algorithm for generation of unique smiles notation, Journal of chemical information and computer sciences 29 (2) (1989) 97–101.

[51] Bigsmiles, `https://olsenlabmit.github.io/BigSMILES/docs/line_notation.html#the-bigsmiles-line-notation`.

[52] Psmiles, `https://psmiles.readthedocs.io/en/latest/`.

[53] K. G. Joback, R. C. Reid, Estimation of pure-component properties from group-contributions, Chemical Engineering Communications 57 (1-6) (1987) 233–243.

[54] V. Babrauskas, The cone calorimeter, SFPE handbook of fire protection engineering (2016) 952–980.

[55] D. Xiu, G. E. Karniadakis, The wiener–askey polynomial chaos for stochastic differential equations, SIAM journal on scientific computing 24 (2) (2002) 619–644.

[56] Y. Fan, X. Tian, X. Yang, X. Li, C. Webster, Y. Yu, An asymptotically compatible probabilistic collocation method for randomly heterogeneous nonlocal problems, Journal of Computational Physics 465 (2022) 111376.

[57] Y. Fan, H. You, X. Tian, X. Yang, X. Li, N. Prakash, Y. Yu, A meshfree peridynamic model for brittle fracture in randomly heterogeneous materials, Computer Methods in Applied Mechanics and Engineering 399 (2022) 115340. `doi:https://doi.org/10.1016/j.cma.2022.115340`.

[58] S. A. Smolyak, Quadrature and interpolation formulas for tensor products of certain classes of functions, in: Doklady Akademii Nauk, Vol. 148, Russian Academy of Sciences, 1963, pp. 1042–1045.