

Machine Learning Enabled Discoveries of High Flame Retardancy, Strength and Biodegradable Cellulose Nanofiber (CNF)/Polyvinyl Alcohol (PVA) Films

Yan Feng^{a, 1}, Fengqing Chen^{a, b, c, 1, *}, Yiping Wen^a, Shuoshuo Xue^a, Teng Huang^a, Qiao Kang^a, Jiawei Ye^a, Dan Xie^{a, b, c}, Baiyu Jiang^{a, b, c}, Peng Ding^{d, *}, Jinfeng Dai^{a, b, c, *}, Pingan Song^{e, f}

^a School of Chemistry and Materials Engineering, Zhejiang A&F University, Hangzhou 311300, China.

^b Key Laboratory of Wood Science and Technology of Zhejiang Province, Zhejiang A&F University, Hangzhou 311300, China.

^c Zhejiang Key Laboratory of Green and Low-Carbon Utilization Technology of Agricultural and Forestry Biomass, Hangzhou 311300, China.

^d School of Materials Science and Engineering and Research Center of Nanoscience and Nanotechnology, Shanghai University, Shanghai 200444, China.

^e Centre for Future Materials, University of Southern Queensland, Springfield QLD 4300, Australia

^f School of Agriculture and Environmental Science, University of Southern Queensland, Toowoomba, QLD 4305, Australia.

¹ Contributed equally.

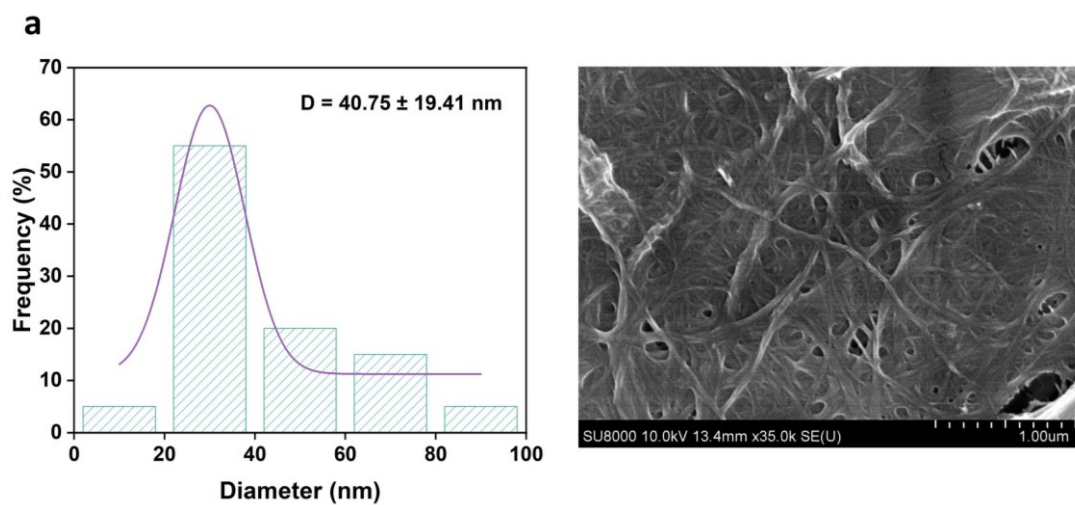
Corresponding author:

E-mail: Fengqingchen@zafu.edu.cn (F.Q. Chen);

dingpeng@shu.edu.cn (P. Ding);

jinfengdai0601@zafu.edu.cn & jinfengdai0601@gmail.com (J.F. Dai);

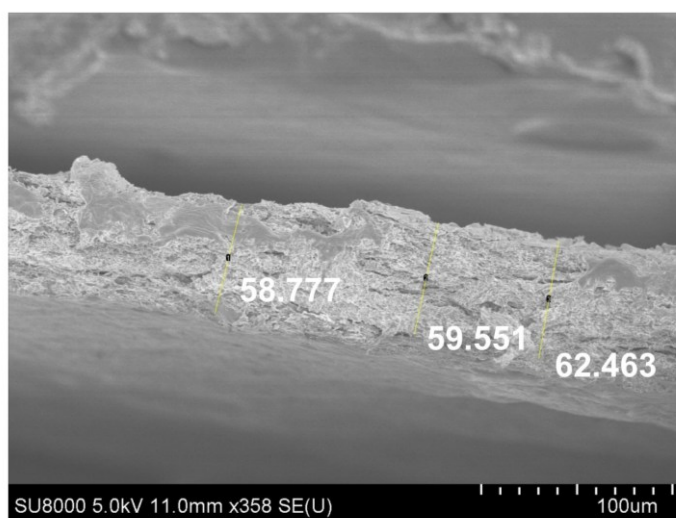
26



27

28 **Fig. S1. The distribution of the diameters of CNF collected from the scanning electron**
29 **microscope.**

30



31

32 **Fig. S2. CNF/PVA composite film cross-sectional scanning electron microscope.**

33

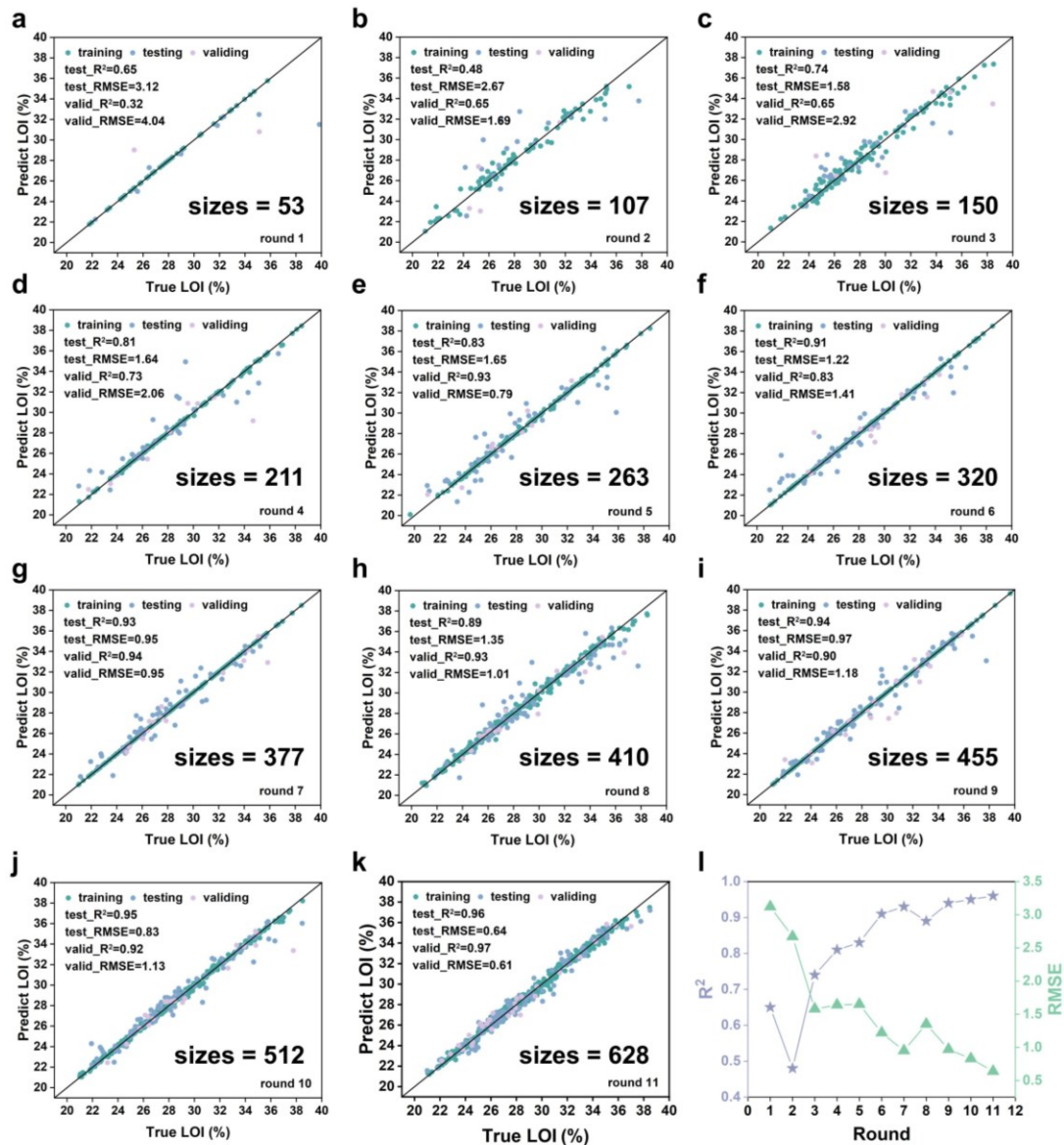


Fig. S3. Moled performance evolutions of expert-guided from active learning loop: (a-k) Changes in performance after iterations based on the XGBoost model, (l) Changes in RMSE and R^2 of the model after each round of iteration

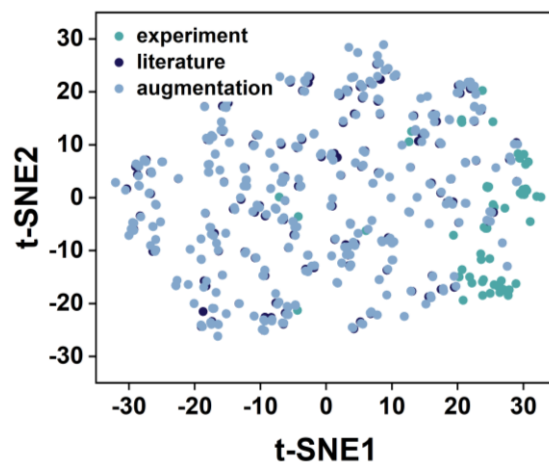


Fig. S4. The t-SNE plot of dataset.

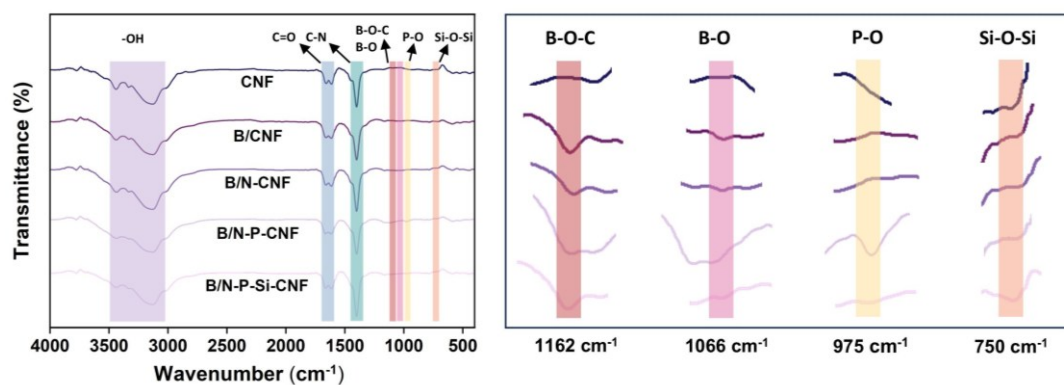


Fig. S5. FT-IR spectra of modified CNF.

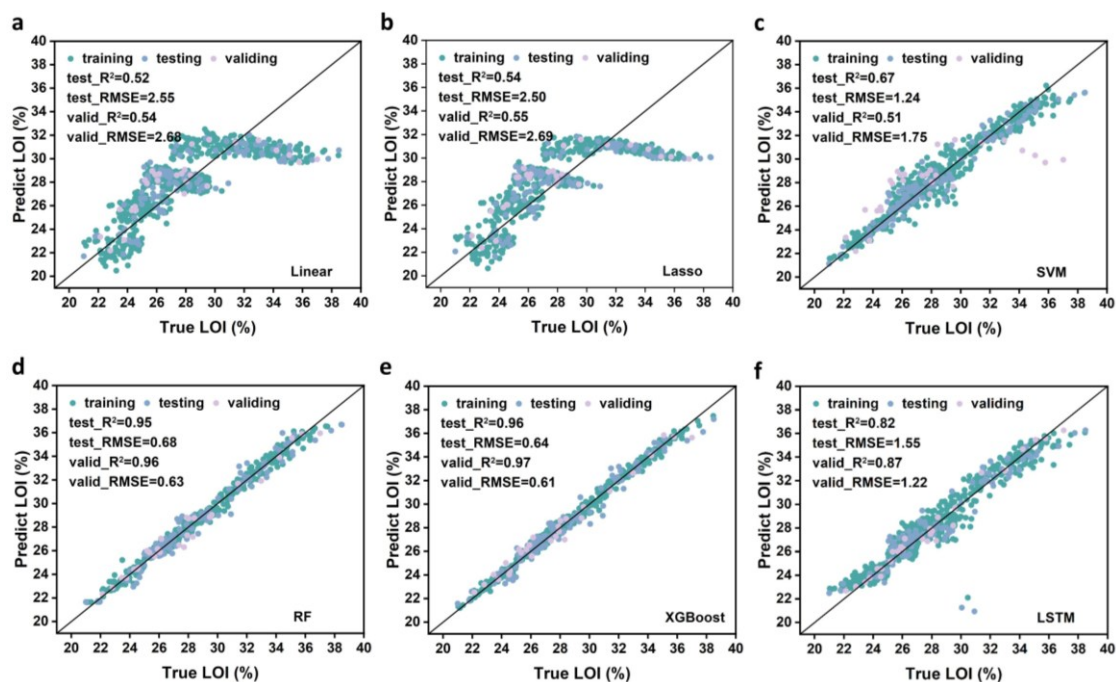


Fig. S6. Six machine learning models established in the experiment. (a) Simple linear

regression model, (b) Lasso regression model, (c) Support vector regression model, (d)
Random forest model, (e) XGBoost model, (f) LSTM model.

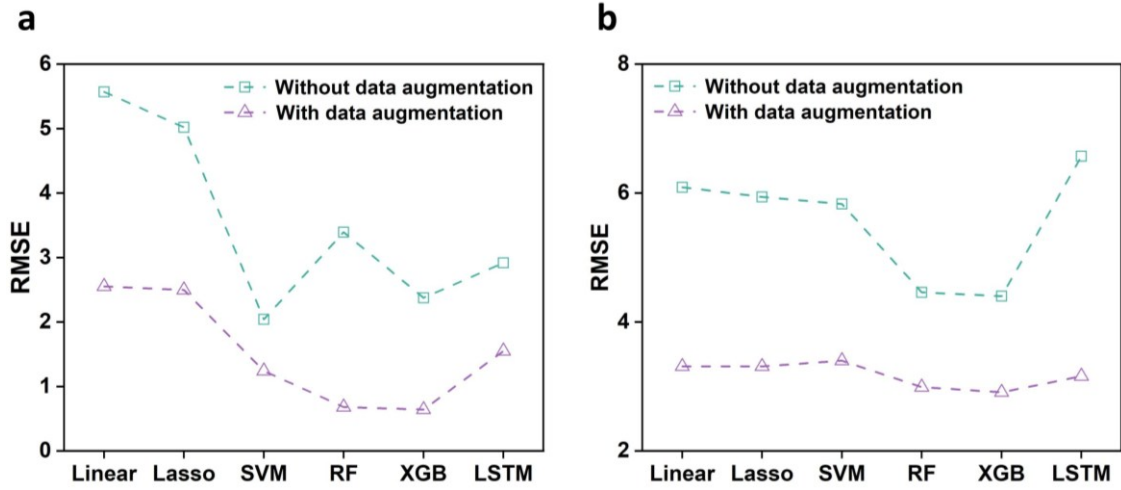


Fig. S7. RMSE values of different prediction model after applying data augmentation strategies (a) LOI model, (b) Tensile strength model.

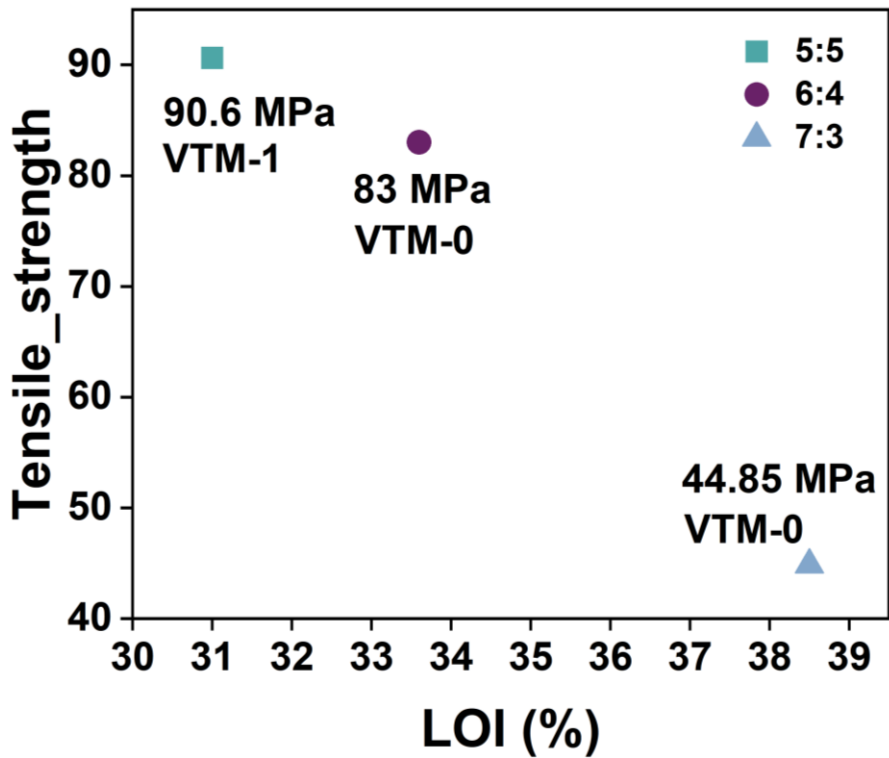


Fig. S8. The screening results under different weight schemes.

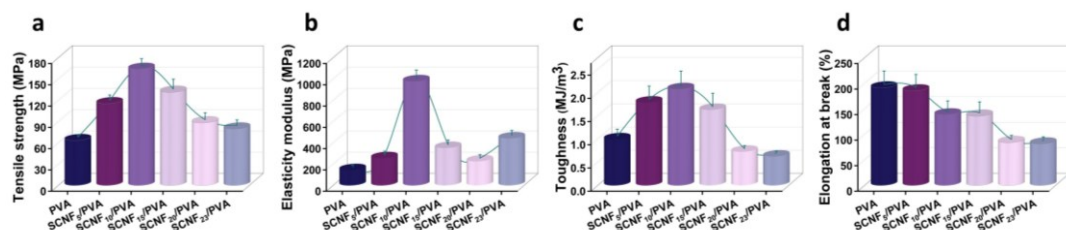


Fig. S9. Mechanical properties analysis of CNF/PVA films: (a) Tensile strength, (b) Elongation at break, (c) Toughness, (d) Elasticity modulus.

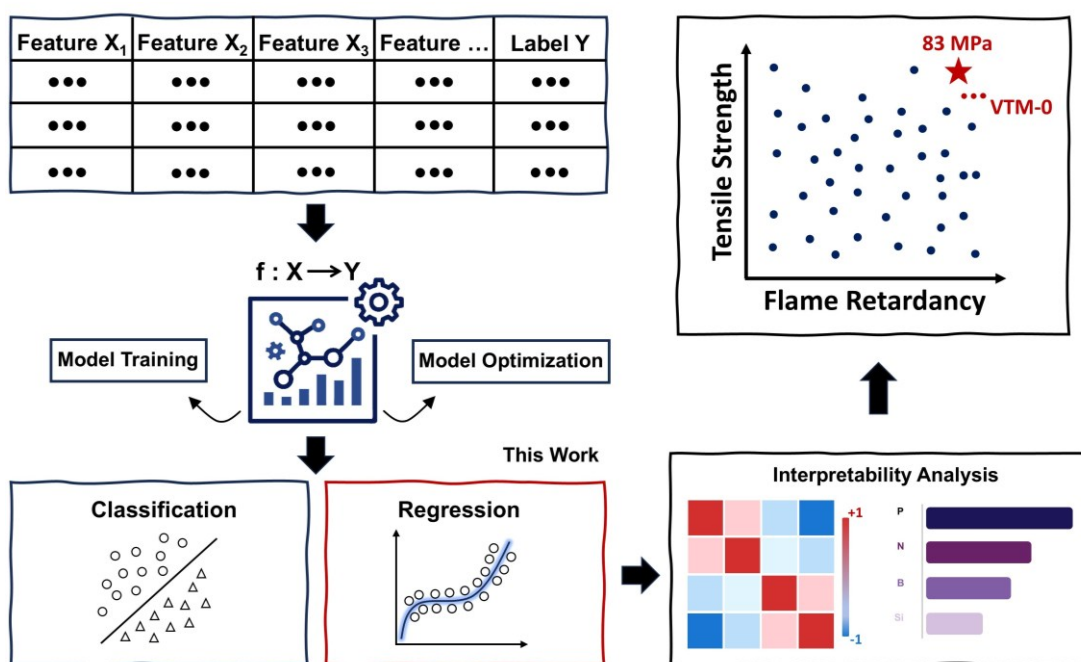


Fig. S10. A general flowchart of machine learning in materials.

Table S1. Changes in PVA concentration under different CNF contents.

CNF/PVA	0 wt %	5 wt %	10 wt %	15 wt %	20 wt %	23 wt %
PVA (g)	4.2	3.99	3.78	3.57	3.36	3.234
CNF (g)	0	0.21	0.42	0.63	0.84	0.966
Deionized water (g)	60	84	105	140	210	210
Concentration (wt %)	7	5	4	3	2	2

Table S2. The training set of the database that was obtained from the experiment.

Num	B	N	P	Si	Filler	T_d	Tensile strength	LOI
1	0.6	1	1	1	1	1.5	83.93	20.5
2	1	0	4	4	50	1.5	61.27	25.8
3	0.6	0	0	1	1	1.5	87.91	20.5
4	0.6	0	1	1	1	1.5	80.9	20.5
5	1.2	2	2	2	2	1.5	88.02	21.7
6	0.6	0	0	1	1	1.5	88.23	20.5
7	0	20	6	0	20	1.5	70	29
8	2	2	2	2	5	2	132.23	26.3
9	1	1.8	5	2.6	5	1.5	133.74	28.4
10	1.2	0	2	2	2	1.5	92.39	21.7
11	0	0	0	0	0	1.5	67	19
12	2.6	3	5.18	1.03	3	1.5	103.93	21.4
13	5	1	3	2	7	1.5	73.01	24.3
14	5	1	6	2	7	1.5	147.92	24.3
15	2.6	3	5.18	1.03	5	2	77.14	22.9
16	3.3	2	0	2	21	1.5	92.35	27.6
17	1.2	2	2	2	2	1.5	114.88	21.7
18	2.4	0	3	3	3	2	97.94	21.9
19	5	1	3	2	7	1.5	136.03	24.3
20	2.4	3	3	3	3	2	103.71	21.9
21	2.6	3	5.18	1.03	2	1.5	92.17	21.3
22	0	0	0	0	0	0	62.07	19
23	0.6	0	0	0	0.5	2	94.86	23.1
24	3.3	1	4	0	27	1.5	58.25	30.6
25	2.4	2	6	3	5	2	115.3	22
26	0.2	2.27	19.18	3.97	13.23	1.74	92.25	28.1
27	1	1.8	5	2.6	20	1.5	138.69	27.3

28	2.6	3	5.18	1.03	1	1.5	77.21	20.2
29	0.6	1	1	1	1	1.5	83.21	20.5
30	0.6	0	0	2	0.5	2	113.82	23.1
31	0.6	0	1	1	1	1.5	107.71	20.5
32	1.2	0	2	2	2	1.5	103.28	21.7
33	0.6	0	1	1	1	1.5	89.42	20.5
34	2.4	0	0	3	3	2	103.93	21.9
35	0	0	5	5	15	2	45	33.5
36	1	1.8	10	2.6	30	1.5	145.02	29.1
37	1.2	0	0	2	2	1.5	87.09	21.7
38	0.6	2	2	2	0.5	2	169.8	23
39	5	1	6	2	7	1.5	147.62	24.3
40	2.4	2	6	3	5	2	127.93	22
41	2.4	3	3	3	3	2	99.8	21.9
42	2.6	3	5.18	1.03	10	2	147.62	25.1
43	4	2	3	0	13	2	72.37	26
44	2.6	3	5.18	1.03	3	1.5	98.06	22.7
45	3.36	0.94	2.15	4.58	31.27	2.65	64.57	32.3
46	2.4	1	6	3	5	2	123.14	22
47	2.5	2	2	1	58	1.5	65.08	25.3
48	1.2	2	2	2	2	1.5	92.4	21.7
49	2.4	0	3	3	3	2	114.03	21.9
50	0	20	6	0	10	1.5	90	25.2
51	1.2	0	0	2	2	1.5	93.93	21.7
52	1.2	0	0	2	2	1.5	92.17	21.7
53	3.3	2	1	3	19	1.5	89.89	27.2
54	2.6	3	5.18	1.03	1	1.5	88.23	20.9
55	2.6	3	5.18	1.03	5	2	120.39	22.2
56	0.6	0	0	1	1	1.5	77.21	20.5

57	1	0	3	3	2	2	82.94	22.6
58	3.3	2	1	3	30	2	67.38	31.3
59	5	1	3	2	7	1.5	117.10	24.3
60	2.6	3	5.18	1.03	2	1.5	87.09	20.7
61	2.4	0	0	3	3	2	70.28	21.9
62	1.25	3.08	5.21	2.3	11	2	109.41	28.4
63	2.4	1	6	3	5	2	138.40	22
64	2.4	0	0	3	3	2	98.06	21.9
65	0.6	0	2	2	0.5	2	127.71	23.5
66	0	5	5	0	10	2	53	27.5
67	0	20	6	0	15	1.5	85	27.1
68	2.4	3	3	3	5	2	120.39	22
69	2.4	1	6	3	5	2	77.14	22
70	0	20	6	0	5	1.5	110	24.2
71	0.6	1	1	1	1	1.5	88.84	20.5
72	5	1	6	2	7	1.5	144.93	24.3
73	0	5	5	0	5	2	60	23.6
74	2.4	0	3	3	3	2	107.24	21.9
75	1.2	0	2	2	2	1.5	97.6	21.7
76	2	0	2	2	5	2	123.08	26.3
77	2.4	2	6	3	5	2	131.7	22
78	3.36	0.94	2.15	4.58	31.27	2.65	64.57	32.3

70
71
72
73
74
75
76
77
78
79
80

81 **Table S3. The part of the training set of the database that was obtained from the literature.**

Num	DOI	Category
1	10.1016/j.foodchem.2024.138506	PVA
2	10.1021/acsanm.4c01785	PVA
3	10.1016/j.ijbiomac.2020.05.044	PVA
4	10.1021/acsami.0c19093	PVA
5	10.1007/s10570-019-02298-z	PVA
6	10.1002/adfm.201905968	PVA
7	10.1016/j.reactfunctpolym.2024.105877	PVA
8	10.1016/j.carbpol.2025.123483	CNF
9	10.1021/acsapm.3c00938	CNF
10	10.1021/acsnano.3c11335	CNF
11	10.1016/j.fbio.2025.107608	CNF
12	10.1021/acs.langmuir.4c01528	CNF
13	10.1016/j.carbpol.2021.118654	CNF
14	10.1002/asia.202300842	CNF
15	10.1016/j.cej.2021.132780	CNF
16	10.1007/s10570-021-04159-0	CNF
17	10.1021/acs.langmuir.4c01528	CNF
18	10.1016/j.carbpol.2022.120068	CNF
19	10.1002/app.56151	CNF
20	10.1016/j.ijbiomac.2021.12.136	CNF

82
83
84
85
86
87
88
89
90
91

92 **Table S4. Comparison of R², RMSE and MSE for six machine learning algorithm models**
 93 **about LOI.**

Algorithm	R ²	RMSE	MSE
Linear	0.52	2.55	6.51
Lasso	0.54	2.50	6.24
SVM	0.66	1.24	1.54
RF	0.95	0.68	0.47
XGBoost	0.96	0.64	0.41
LSTM	0.93	1.55	2.40

94
 95 **Table S5. The comparison of R², RMSE and MSE for six machine learning algorithm models**
 96 **for tensile strength.**

Algorithm	R ²	RMSE	MSE
Linear	0.83	3.31	10.96
Lasso	0.83	3.31	10.96
SVM	0.81	3.40	11.54
RF	0.89	2.99	8.92
XGBoost	0.90	2.91	8.48
LSTM	0.86	3.16	10.01

97
 98
 99
 100
 101
 102
 103
 104
 105
 106
 107

Table S6. The top ten samples with the highest comprehensive scores obtained through model screening.

Num	B	N	P	Si	Filler	T_d	Tensile_strength	LOI	Score
1	2.64	3.014	5.18	1.03	22.67	2.37	83.76	34.27	0.705
2	4.91	3.6	4.04	2.12	21.537	1.11	89.69	33.2679	0.701
3	3.86	4.34	4.85	2.92	21.807	2.4	88.46	33.29385	0.696
4	4.81	1.4	4.67	2.17	22.1895	2.36	91.39	32.6624	0.689
5	1.21	3.37	3.06	0.68	22.473	1.93	88.99	32.01365	0.654
6	2.07	3.15	4.41	2.98	20.691	2.24	85.77	32.4375	0.651
7	2.25	1.16	3.24	0.76	22.4775	2.3	94	31.11405	0.649
8	1.61	4.07	4.5	2.03	21.1815	2.46	84.4	32.351	0.641
9	1	2.47	1.92	1.02	22.428	1.74	88.19	31.71955	0.639
10	0.67	4.45	1.64	2.89	21.8835	2.65	89.84	31.42545	0.638

Table S7. Degradation rate of PVA and SCNF₂₃/PVA composite film.

	Original	Day1	Day3	Day7	Day10
PVA	0.0277±0.0005 g	0.0264 g	0.0253 g	0.0245 g	0.0243 g
SCNF ₂₃ /PVA	0.0320±0.0012 g	0.0290 g	0.0230 g	0.0067 g	0 g

Note S1. Characterization.

The morphology and microstructure of CNF, B/CNF, B/N-CNF, B/N-P-CNF, and B/N-P-Si-CNF were analyzed by Scanning electron microscopy (SEM, TM3030, Hitachi High-Tech Corporation of Japan), combined with Energy dispersive spectrometer (EDS) and Fourier transform infrared spectrometer (FTIR, S20071649, Nicolet, Thermo of the United States). The morphology and thickness of the CNF/PVA composite film were also observed by scanning electron microscopy. The thermal gravimetric analysis (TGA) was conducted using a TG 209 F3 Tarsus (Nanjing Company, Netzsch) thermogravimetric analyzer in a nitrogen environment at a heating rate of 20 °C min⁻¹. The LOI test was performed according to the ASTM D2863 standard on an A-type oxygen index tester (produced by FTT Company in the UK), with the test sample size being 140×52×0.5 mm³. TG-IR testing was conducted using a thermogravimetric-infrared coupling analyzer

(PerkinElmer, TGA4000+SP1, PerkinElmer Instruments Co., Ltd.) under a nitrogen atmosphere. The temperature range spanned from 30°C to 800°C, with a heating rate of 20°C/min and a gas flow rate of 20 mL/min. The mechanical properties of the CNF/PVA composite dispersion, including tensile strength, modulus, and toughness, were tested by a universal mechanical testing machine (CMT 6104, Meister Industrial Four-way Co., Ltd.) according to the ASTM D882 standard.

Note S2. The diameter of CNF.

The diameters of CNF are mostly concentrated in the range of 0.05 - 0.1 micrometers. Through the microscopic morphology analysis using scanning electron microscopy and the elemental distribution analysis with energy dispersive spectrometer. Samples were selected to characterize with slightly larger diameters from numerous cellulose fibers to present more clearly the changes in the surface morphology of cellulose and the distribution of elements.

Note S3. The thickness of CNF/PVA composite film.

The thickness of the CNF/PVA composite film can be determined by scanning electron microscopy, and the thickness range is approximately $60 \pm 2 \mu\text{m}$.

Note S4. Change in PVA solution concentration.

As the cellulose content increases, the concentration of the PVA solution needs to be reduced accordingly, otherwise, severe agglomeration will occur during the dissolution process.

Note S5. Methodological detail of data augmentation.

To expand the design space and mitigate potential overfitting, we employed the Gaussian noise injection technique, a highly established data augmentation strategy in materials informatics. This method essentially simulates the inherent instrumental uncertainties and experimental variances universally present in material synthesis and characterization. Specifically, random noise sampled from a Gaussian distribution ($N(0, \sigma^2)$) was added to the continuous features of the training set. The standard deviation (σ) was strictly constrained to a small percentage (e.g., 5%) of the original feature variance to ensure local perturbation. The sufficiency of this method also has been validated in other publications.¹⁻³

Note S6. Machine learning models.

All machine learning models in this study were written in Python using the Pycharm compiler. The Pearson correlation matrix and SHAP analysis were plotted using the matplotlib package. Furthermore, the hyperparameter optimization and the calculation of R^2 and RMSE were assisted by the s-learn library and the database reading work was performed using the pandas library. Finally, the machine learning models established in this experiment have been packaged and uploaded to Github ([J13feng/CNF-PVA-composite-film: Machine learning model for the research of CNF/PVA composite films](https://github.com/J13feng/CNF-PVA-composite-film: Machine learning model for the research of CNF/PVA composite films)).

Note S7. Active learning loop.

The model's ability to generalize improves with each training cycle. The interpretative results and high-throughput screening outcomes from the previous iteration of the model serve as guidance for the next round of material design. At the same time, incorporating new experimental and literature data improves the model's accuracy through iterative refinement. This process yields more precise interpretive data and screening results, guiding users to design materials with superior performance and broader parameter ranges. These outcomes collectively form the basis for the subsequent iteration. After 11 iterations, the final model for this experiment was obtained.

Note S8. t-distributed stochastic neighbor embedding (t-SNE).

The distribution and rationality of the augmented dataset were validated using t-distributed Stochastic Neighbor Embedding (t-SNE). This non-linear dimensionality reduction method projected the high-dimensional descriptors into a 2-dimensional space to visualize the relative positions of different data sources. The resulting t-SNE plot (Fig. S4) demonstrates that the augmented data points are embedded within the same feature space as the original experimental and literature samples. No distinct clustering or significant deviations were observed for the synthetic data, confirming that the augmentation strategy effectively preserved the intrinsic correlations and physical constraints of the material system. This consistency ensures that the predictive models trained on the expanded dataset remain robust and generalizable to real-world material formulations.

References

- 1 Geng Y, Ji Y, Wang D, Zhang H, Lu Z, Xing A, Gao M, Chen M (2025) Strength prediction of recycled concrete using hybrid artificial intelligence models with Gaussian noise addition. *Eng Appl Artif Intell* 149, 110566. DOI: <https://doi.org/10.1016/j.engappai.2025.110566>.
- 2 Hueffel J A, Bindschaedler Q P, Sala F, Schoenebeck F (2025) Empowering Reactivity Predictions through Noise-Based Data Augmentation. *J Am Chem Soc* 147 (36), 33232-33239. DOI: 10.1021/jacs.5c11632
- 3 Yang H, Wang M, Wang Q, Yu Z, Jin G, Zhou C, Zhou Y (2024) Non-informative noise-enhanced stochastic neural networks for improving adversarial robustness. *Inf Fusions* 108, 102397. DOI: <https://doi.org/10.1016/j.inffus.2024.102397>.