

Supporting Information for

Data Extraction from Polymer Literature using Large Language Models

Sonakshi Gupta,^{1,*} Akhlak Mahmood,^{2,*} Pranav Shetty,¹ Aishat Adeboye,³ and Rampi Ramprasad^{2,†}

¹*School of Computational Science and Engineering,
Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta 30332, GA, USA.*

²*School of Materials Science and Engineering, Georgia Institute of Technology,
771 Ferst Drive NW, Atlanta 30332, GA, USA.*

³*School of Chemical and Biomolecular Engineering,
Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta 30332, GA, USA.*

I. F_1 SCORE CALCULATION METHOD

In order to validate the accuracy of the extracted data, a manual curation of the material name, corresponding property name, and property value (inclusive of unit) was conducted for 630 abstracts. Data extraction for a target property was performed on the curated abstract. A data point extracted via a pipeline was labeled as a True Positive (TP) if it corresponded accurately with all four elements of the curated items. Conversely, if the extracted data did not exist within the manually curated set, it was classified as a False Positive (FP). Furthermore, any data from the manual set that was absent in the extracted set was labeled as False Negative (FN). It was observed that in numerous instances, only one of the material names or the property values correctly matched. Such discrepancy contributed to a reduction in final F_1 score.

To further illustrate the F_1 score calculation method, we consider an example scenario, where we extract bandgap data from the abstract of Ref. [1] for which we have manually curated data available. After processing through the GPT-3.5 and MaterialsBERT pipelines, we obtain the following set of data points, shown in **Table S1**. The accuracy of the extracted data determined using the ground truth values is given in the ‘Result’ column.

TABLE S1. Datapoints extracted from Ref. [1] for bandgap by the GPT-3.5 and MaterialsBERT pipelines. The star (*) symbol indicates a material not extracted by the pipeline.

No.	Pipeline	Material	Property	Value	Result
1	GPT-3.5	P-diF2T	bandgap	1.83 eV	TP
2	GPT-3.5	P-TT	bandgap	1.85 eV	TP
3	GPT-3.5	P-BDT	bandgap	1.90 eV	TP
4	GPT-3.5	P-NDT	bandgap	2.10 eV	FP
5	MaterialsBERT	P-diF2 T	optical bandgap	1.83 eV	TP

* These authors contributed equally.

† rampi.ramprasad@mse.gatech.edu

No.	Pipeline	Material	Property	Value	Result
6	MaterialsBERT	TT	optical bandgap	1.85 eV	TP
7	MaterialsBERT	P-diF2 T	optical bandgap	1.90 eV	FP
8	MaterialsBERT	P-diF2 T	optical bandgap	2.04 eV	FP
9	MaterialsBERT	* P-BDT	-	-	FN
10	MaterialsBERT	* P-NDT	-	-	FN

As discussed in the manuscript, there are differences among the extracted material names and the property names by the two pipelines (e.g., note the presence of a space in ‘P-diF2 T’). We normalize such names to identify the true positive cases. Additionally, we treat datapoint 4 as false positive (FP) because the material name was correctly extracted but the value was hallucinated by GPT-3.5. Proceeding to the F_1 score calculation, we obtain the following numbers shown in **Table S2**.

TABLE S2. Metrics obtained for the datapoints extracted in Table S1.

Pipeline	TP	FP	FN	precision	recall	F_1 score	Unique materials
GPT-3.5	3	1	0	0.75	1	0.86	4
MaterialsBERT	2	2	2	0.5	0.5	0.50	2

The Venn diagrams shown in Fig. 4 of the main text, which highlight the differences in the material names extracted by the pipelines, were drawn using the number of unique paragraphs and materials obtained from the Tg and bandgap data points extracted from all papers.

The metrics of precision and recall were computed using the counts of True Positives, False Positives, and False Negatives as illustrated below.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Finally, the F_1 score is calculated as,

$$F_1 = 2 \times \frac{precision * recall}{precision + recall}$$

II. MATERIAL NAMES EXTRACTED FOR TG AND BANDGAP

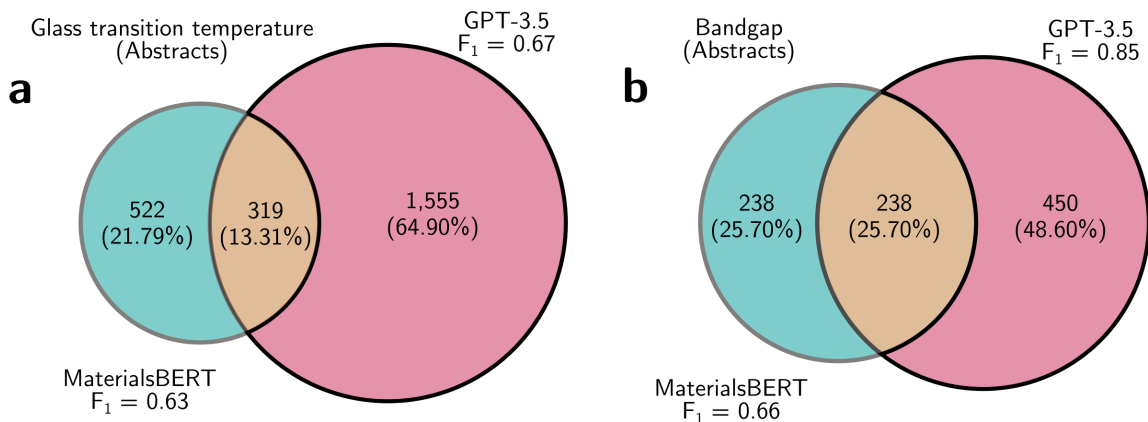


FIG. S1. Number of unique material names extracted for (a) Tg and (b) bandgap using the Materials and GPT-3.5 pipelines from the abstracts of 630 selected papers.

III. SELECTION OF SUBSET OF THE FULL CORPUS

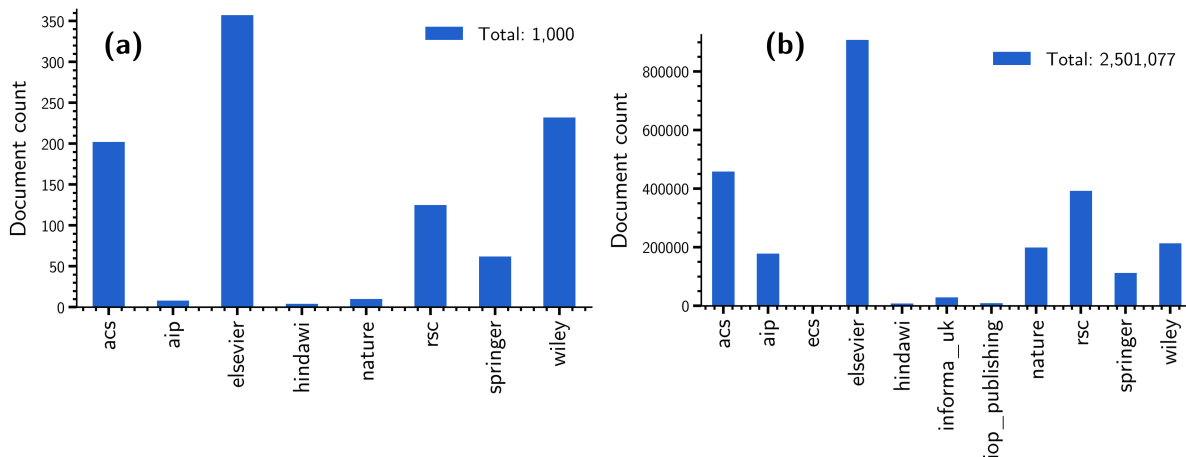


FIG. S2. (a) Number of documents randomly selected from different publishers for performance benchmarking using a subset of the full corpus. (b) Total number of documents available in the corpus from different publishers. The subset represents journal articles from the eight most common publishing houses.

IV. PARITY PLOTS FOR THE Tg AND BANDGAP DATA

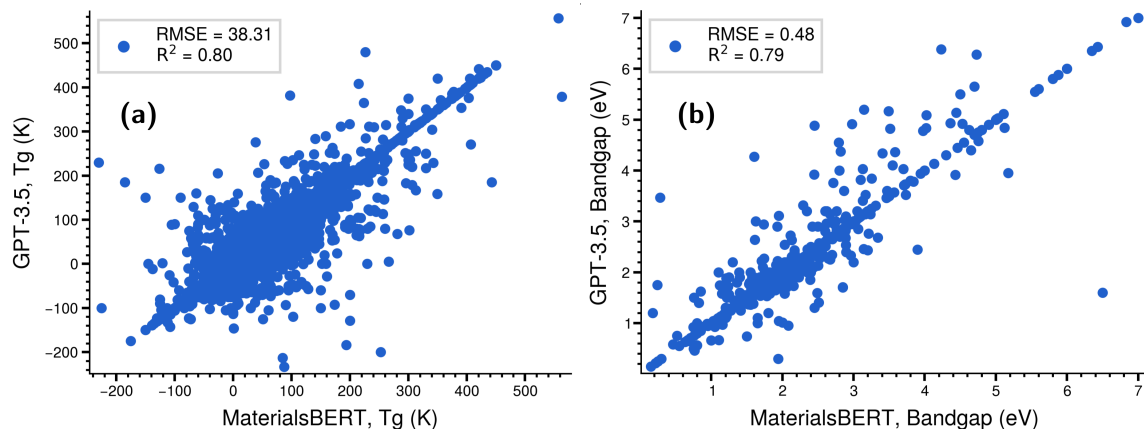


FIG. S3. Parity plot for (a) Tg and (b) bandgap data extracted by the GPT-3.5 and the MaterialsBERT pipelines from the full texts of all polymer related articles. The plots show data where both pipelines extracted the same polymer names and are from the same papers.

The extraction efficacy of bandgap data by the GPT-3.5 pipeline outperformed that of Tg data. **Fig. S4** provides a visual representation of the False Positives versus True Positives and False Negatives versus True Positives. It is noteworthy that the extracted bandgap data demonstrated a minimal amount of both False Positives and False Negatives. In contrast, the Tg data extraction exhibited a higher incidence of False Positives. This discrepancy may be attributed to the frequent occurrence of similar material names and numbers in the text discussing Tg values in many journal articles. Consequently, the model struggles to accurately establish the relationships between the material names and property values.

V. COMPARISON OF PERFORMANCE BETWEEN TG AND BANDGAP

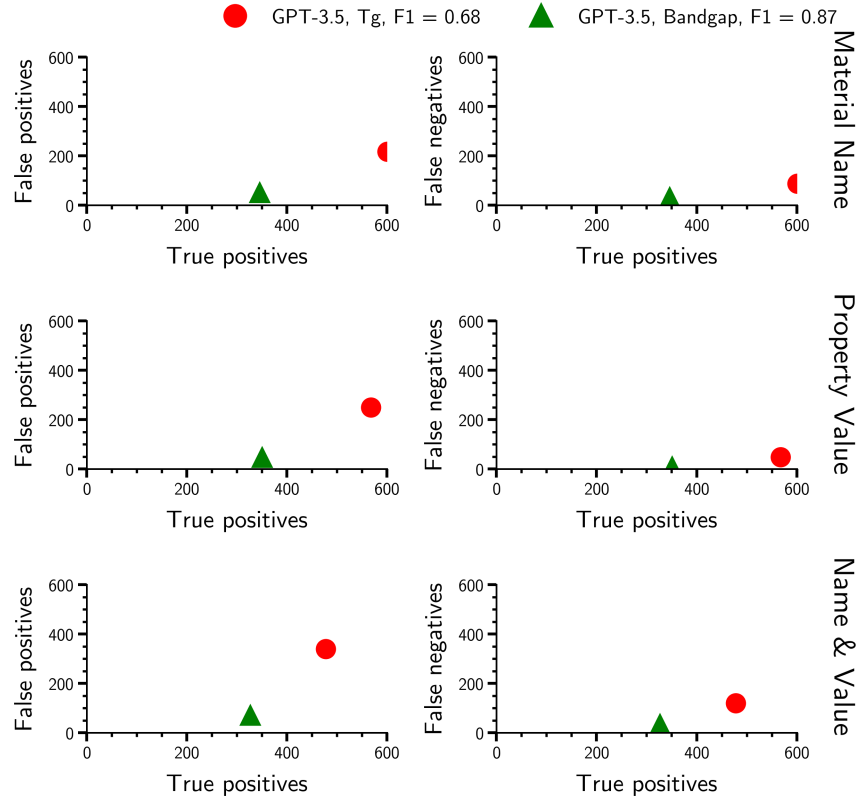


FIG. S4. Number of FP and FN data points for Tg and bandgap values extracted from the curated abstracts by GPT-3.5.

TABLE S3. Detailed metrics for Tg and bandgap data extracted by the GPT-3.5 and MaterialsBERT pipelines from 630 abstracts.

Pipeline	Entity Class	Property	TP	FP	FN	Precision	Recall	F_1 score
GPT-3.5	Material Name	Tg	601	217	88	0.73	0.87	0.79
GPT-3.5	Property Value	Tg	568	250	48	0.69	0.92	0.79
GPT-3.5	Name and Value	Tg	478	340	120	0.58	0.79	0.68
GPT-3.5	Material Name	Bandgap	346	50	30	0.87	0.92	0.89
GPT-3.5	Property Value	Bandgap	351	45	7	0.88	0.98	0.93
GPT-3.5	Name and Value	Bandgap	327	69	33	0.82	0.91	0.87
MaterialsBERT	Material Name	Tg	718	83	171	0.89	0.81	0.85
MaterialsBERT	Property Value	Tg	573	228	14	0.71	0.97	0.83
MaterialsBERT	Name and Value	Tg	449	352	181	0.56	0.71	0.63
MaterialsBERT	Material Name	Bandgap	348	21	150	0.94	0.69	0.81
MaterialsBERT	Property Value	Bandgap	332	37	7	0.89	0.97	0.93
MaterialsBERT	Name and Value	Bandgap	242	127	153	0.65	0.61	0.63

VI. LIMITATIONS OF MATERIALSBERT

Our analysis of the extracted full-text data indicates that MaterialsBERT fails to identify proper relationships among the data when the source text contains many numbers in it. For example, the following paragraph is taken from Ref. [2], which reports water contact angle (WCA) measurements on a blend of poly(3,4-ethylenedioxythiophene) and poly(styrenesulfonate) (PEDOT:PSS) calculated using molecular dynamics simulations method.

“Similarly, the contact angle measurements of formamide, water, and glycerol on PEDOT:PSS were also performed. The snapshots of droplets at 2, 3, 4, 5, and 6 ns are shown in Figures S3-S5. The fitted circles to the points are shown in Figures S6-S8. The contact angles at different times are shown in Table 1. The contact angles of formamide at 2, 3, 4, 5, and 6 ns were 40.9°, 35.0°, 32.8°, 27.9°, and 28.8°, respectively. Contact angle values decreased considerably from 40.9° to 27.9° between 2 and 5 ns and increased to 28.8° at 6 ns, suggesting that contact angle only stabilizes after 5 ns of simulation. Similarly, the water contact angles at 2, 3, 4, 5, and 6 ns were 45.9°, 38.0°, 39.2°, 38.2°, and 40.7°, respectively. Here, the fluctuations in contact angle diminished most after 3 ns of simulation time. The glycerol contact angles at 2, 3, 4, 5, and 6 ns were 73.6°, 66.0°, 67.3°, 64.6°, and 65.1°, respectively. Again, the fluctuations in the contact angle diminished after 3 ns.”

TABLE S4. Water contact angle data extracted from Ref. [2] using MaterialsBERT and GPT-3.5 pipelines.

Method	Material	Property	Value	Unit	Neat Polymer
GPT-3.5	PEDOT:PSS	wca	45.9	°	False
GPT-3.5	PEDOT:PSS	wca	38.0	°	False
GPT-3.5	PEDOT:PSS	wca	39.2	°	False
GPT-3.5	PEDOT:PSS	wca	38.2	°	False
GPT-3.5	PEDOT:PSS	wca	40.7	°	False
MaterialsBERT	PEDOT	wca	32.8	°	True
MaterialsBERT	PEDOT	wca	27.9	°	True
MaterialsBERT	PEDOT	wca	28.8	°	True
MaterialsBERT	PEDOT	wca	40.9	°	True
MaterialsBERT	PEDOT	wca	27.9	°	True
MaterialsBERT	PEDOT	wca	28.8	°	True
MaterialsBERT	PEDOT	wca	40.7	°	True
MaterialsBERT	PEDOT	wca	73.6	°	True
MaterialsBERT	PEDOT	wca	66.0	°	True
MaterialsBERT	PEDOT	wca	67.3	°	True
MaterialsBERT	PEDOT	wca	64.6	°	True
MaterialsBERT	PEDOT	wca	65.1	°	True
MaterialsBERT	PSS	wca	32.8	°	True
MaterialsBERT	PSS	wca	27.9	°	True
MaterialsBERT	PSS	wca	28.8	°	True
MaterialsBERT	PSS	wca	40.9	°	True
MaterialsBERT	PSS	wca	27.9	°	True
MaterialsBERT	PSS	wca	28.8	°	True
MaterialsBERT	PSS	wca	40.7	°	True
MaterialsBERT	PSS	wca	73.6	°	True
MaterialsBERT	PSS	wca	66.0	°	True
MaterialsBERT	PSS	wca	67.3	°	True
MaterialsBERT	PSS	wca	64.6	°	True
MaterialsBERT	PSS	wca	65.1	°	True

In the context of **Table S4**, it is evident that the GPT-3.5 pipeline correctly extracts WCA measurements regarding the PEDOT:PSS blend, effectively filtering out extraneous data related to formamide and glycerol. In contrast, MaterialsBERT exhibits shortcomings in correctly identifying the inter-relationship between entities, leading to the extraction of WCA data points in a disjointed manner for both PEDOT and PSS. This deficiency in MaterialsBERT’s performance in the extraction process gives rise to an increased number of inaccurately extracted data. Furthermore, the number of neat polymers incorrectly extracted by MaterialsBERT is significantly higher. Generally, GPT-3.5 exhibited satisfactory performance when dealing with extensive numerical data and values.

VII. ACCURACY OF DATA IN PAIRWISE PLOT

TABLE S5. Accuracy of the marked data points in **Fig. 4f** of the main text. When multiple values are mentioned in the article, the average is taken.

No.	Source	Pipeline	Extracted Material	True Material	Extracted Bandgap	True Bandgap	Extracted Tg	True Tg
1	Ref. [3] pg. 364, 366	MaterialsBERT	PMMA	PMMA	5.85	5.85	112	112
2	Ref. [4] pg. 1307	GPT-3.5	Polyolefin	Polyolefin	5	5	244	244
3	Ref. [5] pg. 183	MaterialsBERT	PAE	PAE	4.5	4.5	300	300
4	Ref. [6] pg. 28515, 28522	GPT-3.5	PAE	PAE	4.24	4.24	400	N/A in text
5	Ref. [7] pg. 1220, 1226	GPT-3.5	PVA	PVA/ZnO composite	3.82	3.82	480	480
6	Ref. [8] pg. 830, 836	GPT-3.5	PBZT	PBZT modified with PPV	2.205	1.93	680	680

TABLE S6. List of all Tg and bandgap data points extracted from the source articles of the marked points in **Fig. 4f** of the main text. This shows the ability of GPT-3.5 to capture detailed compositions of polymer composites and blends.

Point	Source	Pipeline	Material	Property	Value	Unit
1	Ref. [3]	GPT-3.5	PMMA	Tg	112.0	°C
1	Ref. [3]	GPT-3.5	non-oriented PMMA layer	Bandgap	5.75	eV
1	Ref. [3]	GPT-3.5	oriented PMMA layer	Bandgap	6.07	eV
1	Ref. [3]	MaterialsBERT	PMMA	Tg	112.0	°C
1	Ref. [3]	MaterialsBERT	PMMA	Bandgap	5.75	eV
1	Ref. [3]	MaterialsBERT	PMMA	Bandgap	6.07	eV
1	Ref. [3]	MaterialsBERT	PMMA	Bandgap	5.75	eV
2	Ref. [4]	GPT-3.5	polyolefins	Tg	244.0	°C
2	Ref. [4]	GPT-3.5	o-POFNB	Tg	244.0	°C
2	Ref. [4]	GPT-3.5	o-POFNB	Tg	244.0	°C
2	Ref. [4]	GPT-3.5	POFNB	Tg	186.0	°C
2	Ref. [4]	GPT-3.5	PEEK	Tg	149.0	°C
2	Ref. [4]	GPT-3.5	polyolefins	Bandgap	5.0	eV
2	Ref. [4]	GPT-3.5	PBBTS	Bandgap	1.55	eV
2	Ref. [4]	GPT-3.5	PBBTA	Bandgap	2.25	eV

Point	Source	Pipeline	Material	Property	Value	Unit
2	Ref. [4]	GPT-3.5	o-POFNB	Bandgap	5.0	eV
2	Ref. [4]	MaterialsBERT	o-POFNB	Tg	100.0	°C
3	Ref. [5]	GPT-3.5	PAEs	Tg	300.0	°C
3	Ref. [5]	GPT-3.5	PAEs	Tg	300.0	°C
3	Ref. [5]	GPT-3.5	PAEs	Tg	300.0	°C
3	Ref. [5]	MaterialsBERT	PAEs	Tg	300.0	°C
3	Ref. [5]	MaterialsBERT	PAE	Tg	300.0	°C
3	Ref. [5]	MaterialsBERT	PAEs	Tg	300.0	°C
3	Ref. [5]	MaterialsBERT	PAE	Bandgap	4.5	eV
4	Ref. [6]	GPT-3.5	PEAs	Tg	400.0	°C
4	Ref. [6]	GPT-3.5	PEAs	Bandgap	3.91	eV
4	Ref. [6]	GPT-3.5	PEAs	Bandgap	4.57	eV
5	Ref. [7]	GPT-3.5	PMMA	Tg	92.39	°C
5	Ref. [7]	GPT-3.5	PMMA/ZnO nanocomposites	Tg	118.54	°C
5	Ref. [7]	GPT-3.5	PVC	Tg	66.51	°C
5	Ref. [7]	GPT-3.5	PVC/ZnO	Tg	73.63	°C
5	Ref. [7]	GPT-3.5	PVA	Tg	480.0	°C
5	Ref. [7]	GPT-3.5	PVA/PEO blend	Tg	48.0	°C
5	Ref. [7]	GPT-3.5	PVA/ZnO- Ce_2O_3 nanocomposites	Tg	98.0	°C
5	Ref. [7]	GPT-3.5	PVA/ZnO- Ce_2O_3 nanocomposites	Tg	102.0	°C
5	Ref. [7]	GPT-3.5	PS	Tg	99.0	°C
5	Ref. [7]	GPT-3.5	PS/ZnO nanocomposites	Tg	101.5	°C
5	Ref. [7]	GPT-3.5	ZnO	Bandgap	3.37	eV
5	Ref. [7]	GPT-3.5	AOZO1	Bandgap	3.12	eV
5	Ref. [7]	GPT-3.5	AOZO2	Bandgap	3.17	eV
5	Ref. [7]	GPT-3.5	AOZO3	Bandgap	3.23	eV
5	Ref. [7]	GPT-3.5	ZnO	Bandgap	3.35	eV
5	Ref. [7]	GPT-3.5	0.0 wt % Al-doped ZnO	Bandgap	3.276	eV
5	Ref. [7]	GPT-3.5	0.2 wt % Al-doped ZnO	Bandgap	3.277	eV
5	Ref. [7]	GPT-3.5	0.4 wt % Al-doped ZnO	Bandgap	3.28	eV
5	Ref. [7]	GPT-3.5	0.6 wt % Al-doped ZnO	Bandgap	3.282	eV
5	Ref. [7]	GPT-3.5	0.8 wt % Al-doped ZnO	Bandgap	3.283	eV
5	Ref. [7]	GPT-3.5	1.0 wt % Al-doped ZnO	Bandgap	3.284	eV
5	Ref. [7]	GPT-3.5	2.0 wt % Al-doped ZnO	Bandgap	3.288	eV
5	Ref. [7]	GPT-3.5	3.0 wt % Al-doped ZnO	Bandgap	3.294	eV
5	Ref. [7]	GPT-3.5	AZO thin films	Bandgap	3.38	eV
5	Ref. [7]	GPT-3.5	AZO thin films	Bandgap	3.39	eV

Point	Source	Pipeline	Material	Property	Value	Unit
5	Ref. [7]	GPT-3.5	ZnO	Bandgap	3.2	eV
5	Ref. [7]	GPT-3.5	$Zn_{1-x}Ni_xO$	Bandgap	3.21	eV
5	Ref. [7]	GPT-3.5	$Zn_{1-x}Ni_xO$	Bandgap	3.17	eV
5	Ref. [7]	GPT-3.5	ZnO;B	Bandgap	3.24	eV
5	Ref. [7]	GPT-3.5	ZnO;B	Bandgap	3.35	eV
5	Ref. [7]	GPT-3.5	copper doped ZnO	Bandgap	2.92	eV
5	Ref. [7]	GPT-3.5	PVA	Bandgap	3.82	eV
5	Ref. [7]	GPT-3.5	3 wt % ZnO/PVA nanocomposite	Bandgap	2.78	eV
5	Ref. [7]	GPT-3.5	ZnO/PEO composite (0 wt %)	Bandgap	2.01	eV
5	Ref. [7]	GPT-3.5	ZnO/PEO composite (2 wt %)	Bandgap	1.96	eV
5	Ref. [7]	GPT-3.5	ZnO/PEO composite (6 wt %)	Bandgap	1.9	eV
5	Ref. [7]	GPT-3.5	ZnO/PEO composite (10 wt %)	Bandgap	1.88	eV
5	Ref. [7]	GPT-3.5	ZnO/PEO composite (14 wt %)	Bandgap	1.84	eV
5	Ref. [7]	MaterialsBERT	polyvinylchloride	Tg	66.51	°C
5	Ref. [7]	MaterialsBERT	ZnO	Tg	63.0	°C
5	Ref. [7]	MaterialsBERT	PS	Tg	101.5	°C
5	Ref. [7]	MaterialsBERT	PVA	Tg	102.0	°C
5	Ref. [7]	MaterialsBERT	PVA	Tg	480.0	°C
5	Ref. [7]	MaterialsBERT	PS	Tg	99.0	°C
5	Ref. [7]	MaterialsBERT	PVA	Tg	98.0	°C
5	Ref. [7]	MaterialsBERT	polyvinylchloride	Tg	73.63	°C
5	Ref. [7]	MaterialsBERT	ZnO	Bandgap	3.207	eV
5	Ref. [7]	MaterialsBERT	ZnO	Bandgap	3.17	eV
5	Ref. [7]	MaterialsBERT	ZnO	Bandgap	3.23	eV
5	Ref. [7]	MaterialsBERT	PVA	Bandgap	2.756	eV
5	Ref. [7]	MaterialsBERT	ZnO	Bandgap	3.35	eV
5	Ref. [7]	MaterialsBERT	PVA	Bandgap	3.12	eV
5	Ref. [7]	MaterialsBERT	aluminum	Bandgap	3.294	eV
5	Ref. [7]	MaterialsBERT	PVA	Bandgap	2.78	eV
5	Ref. [7]	MaterialsBERT	ZnO	Bandgap	3.241	eV
5	Ref. [7]	MaterialsBERT	ZnO	Bandgap	3.229	eV
5	Ref. [7]	MaterialsBERT	PVA	Bandgap	3.4	eV
5	Ref. [7]	MaterialsBERT	PVA	Bandgap	3.82	eV
5	Ref. [7]	MaterialsBERT	PVA	Bandgap	2.72	eV
6	Ref. [8]	GPT-3.5	PBZT	Tg	680.0	°C
6	Ref. [8]	GPT-3.5	PBO	Tg	680.0	°C
6	Ref. [8]	GPT-3.5	PBZT	Bandgap	2.205	eV

Point	Source	Pipeline	Material	Property	Value	Unit
6	Ref. [8]	MaterialsBERT	PBZT	Bandgap	0.41	eV
6	Ref. [8]	MaterialsBERT	PBZT	Bandgap	2.205	eV

SUPPLEMENTARY REFERENCES

- [1] M. He, W. Li, H. Tian, H. Tong, J. Zhang, J. Liu, Z. Xie, Y. Geng, and F. Wang, Wide bandgap donor-acceptor conjugated polymers with alkylthiophene as side chains for high-performance non-fullerene polymer solar cells, [Organic Electronics](#) **65**, 31 (2019).
- [2] R. Bhowmik, R. J. Berry, M. F. Durstock, and B. J. Leever, Prediction of the Wetting Behavior of Active and Hole-Transport Layers for Printed Flexible Electronic Devices Using Molecular Dynamics Simulations, [ACS Appl. Mater. Interfaces](#) **9**, 19269 (2017).
- [3] V. Švorčík, O. Lyutakov, and I. Huttel, Thickness dependence of refractive index and optical gap of PMMA layers prepared under electrical field, [J Mater Sci: Mater Electron](#) **19**, 363 (2008).
- [4] A. A. Deshmukh, C. Wu, O. Yassin, A. Mishra, L. Chen, A. Alamri, Z. Li, J. Zhou, Z. Mutlu, M. Sotzing, P. Rajak, S. Shukla, J. Vellek, M. A. Baferani, M. Cakmak, P. Vashishta, R. Ramprasad, Y. Cao, and G. Sotzing, Flexible polyolefin dielectric by strategic design of organic modules for harsh condition electrification, [Energy Environ. Sci.](#) **15**, 1307 (2022).
- [5] C.-C. Lee and W.-Y. Huang, The influence of aryene ether positions and phenyl substituents on the physical properties of functional fluorinated polymers, [Polym J](#) **43**, 180 (2011).
- [6] C. A. Terraza, L. H. Tagle, A. Tundidor-Camba, C. M. González-Henríquez, D. Coll, and M. M. Sarabia, Silarylene-containing oligo(ether-amide)s based on bis(4-(4-amino phenoxy)phenyl)dimethylsilane. Effect of the dicarboxylic acid structure on some properties, [RSC Adv.](#) **5**, 28515 (2015).
- [7] D. Ponnammam, J.-J. Cabibihan, M. Rajan, S. S. Pethaiah, K. Deshmukh, J. P. Gogoi, S. K. K. Pasha, M. B. Ahamed, J. Krishnegowda, B. N. Chandrashekar, A. R. Polu, and C. Cheng, Synthesis, optimization and applications of ZnO/polymer nanocomposites, [Materials Science and Engineering: C](#) **98**, 1210 (2019).
- [8] X.-D. Hu, S. E. Jenkins, B. G. Min, M. B. Polk, and S. Kumar, Rigid-Rod Polymers: Synthesis, Processing, Simulation, Structure, and Properties, [Macromolecular Materials and Engineering](#) **288**, 823 (2003).