Enhancing Deep Learning Predictive Models with HAPPY (Hierarchically Abstracted rePeat unit of PolYmers) Representation

Jihun Ahn[1], Gabriella Pasya Irianti[1], Yeojin Choe[1], Su-Mi Hur[1,2*]

[1] Department of Polymer Engineering, Graduate School, Chonnam National University, Gwangju 61186, Korea

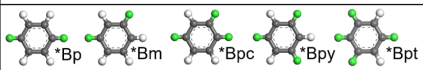
[2] School of Polymer Science and Engineering, Chonnam National University, Gwangju 61186, Republic of Korea

Supplementary Information

Supplementary Information 1. Lists of Subgroups

As noted in the main text, subgroups are defined for sub-structures consisting of ring structures, functional groups, and frequently occurring structures. These frequently occurring structures encapsulate small substituents structures, like ethyl, propyl, and butyl structures. Supplementary Table 1 below compiles the chemical structures of subgroups defined in this study, along with their assigned characters in HAPPY representation. The SMILES notation for these subgroups is also provided for comparison in Supplementary Information 1. For cases where the types and location of bonds need to be differentiated, we introduced topo-HAPPY. All extended topological subgroups and their characters in topo-HAPPY are also listed in Supplementary Table 1.

Supplementary Table 1. Compilation of subgroups defined from the experimental dataset used in this study, along with their string representations derived from HAPPY (topo-HAPPY where applicable) and SMILES.

No.	Structure	HAPPY	SMILES	Topo_HAPPY
1		*B	c1ccccc1	 *Bp *Bm *Bpc *Bpy *Bpt
2		*Bn	c2ccc1[nH]cnc1c2	
3		*Bc	c2ccc1ccccc1c2	
4		*Bh	O=c1[nH]c(=O)c2ccccc12	
5		*Nbb	c1ccc3c(c1)[nH]c2ccccc23	
6		*Ma	O=C1CCC(=O)N1	 *Mat
7		*H	C1CCCCC1	
8		*Co	CC(=O)O	
9		*Ct	CC	
10		*Cf	CCCC	
11		*Cq	CC(C)C	

Supplementary Information 2. Experimental Dataset

Our experimental datasets consisted of 56 repeat unit structures, each with its experimental dielectric constant values. Dielectric constant values can be affected by structure, composition and morphology of polymer, the presence of impurities and different temperature and frequency of measurement. There is still a lack in polymers databases with standardized and well-informed details about experimental conditions and method of measurement. Such inconsistency in the dataset might affect to the network performance. To mitigate the problems. We limit the source of our database and most of data is retrieved from J. Bicerano, "Prediction of Polymer Properties" which informed the empirical Quantitative Structure-Property Relationship (QSPR) scheme successfully implemented in Synthia module of commercial Materials Studio 2022 software. Supplementary Table 2 display the full list of experimental data that used in this study, and corresponding HAPPY and SMILES representation for their repeat unit structures.

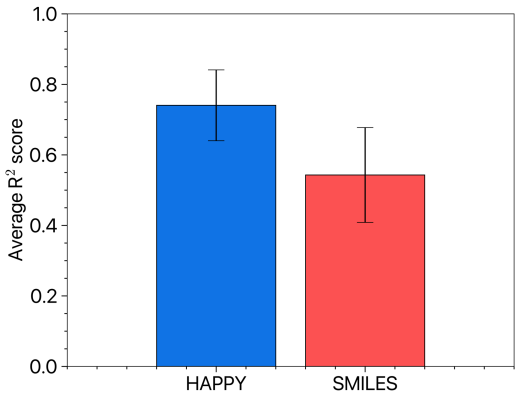
Supplementary Table 2. Full list of experimental dataset featuring polymers with various repeat unit structures, their dielectric constant values, and corresponding HAPPY and SMILES representations.

Polymer	SMILES	HAPPY	Dielectric constant
Polytetrafluoroethylene	FC(=C(F)F)F	HC#FFC#FFT	2.1
Poly(4-methyl-1-pentene)	CC(CC=C)C	HCC@*CqT	2.13
Polypropylene	C=CC	HCC@CT	2.2
Polyisobutylene	CC(C)=C	HCC#CCT	2.23
Poly(vinyl cyclohexane)	C(=C)C1CCCCC1	HCC@*HT	2.25
Poly(1-butene)	C=CCC	HCC@*CtT	2.27
Polyethylene	C=C	HCCT	2.3
Poly (a,a,a',a'-Tetrafluoro-p-xylene)	FC(C1=CC=C(C=C1)C(F)F)F	HC#FF*BC#FFT	2.35
Polyisoprene	C=CC(C)=C	HCCC@CCT	2.37
Poly(o-methyl styrene)	CC1=C(C=C)C=CC=C1	HC*BCCT	2.49
Poly(1,4-butadiene)	C=CC=C	HCCCCT	2.51
Poly(beta-vinyl naphthalene)	C(=C)C1=CC2=CC=CC=C2C=C1	HCC@*BcT	2.51
Polystyrene	C=CC1=CC=CC=C1	HCC@*BT	2.55
Poly(a-methyl styrene)	CC(=C)C1=CC=CC=C1	HC@*BC@CT	2.57
Poly(cyclohexyl methacrylate)	C(C(=C)C)(=O)OC1CCCCC1	HC@CC@*Oc@*HT	2.58
Polychlorotrifluoroethylene	ClC(=C(F)F)F	HC#FFC#ClFT	2.6
Poly(alpha-vinyl naphthalene)	C(=C)C1=CC=CC2=CC=CC=C12	HCC@*BcCC@*BcT	2.6
Poly[oxy(2,6-dimethyl-1,4-phenylene)]	[O]C1=C(C=[C]C=C1C)C	HO*B#CCT	2.6
Poly[1,1-cyclohexane-bis(4-phenyl) carbonate]	O=C(Oc3ccc(C2(c1ccc(O[*])cc1)CCCCC2)cc3)[*]	HC@OO*B*Ch*BOT	2.6
Poly(p-xylylene)	C1(=CC=C(C=C1)[CH2])[CH2]	H*BT	2.65
Poly(isobutyl methacrylate)	C(C(=C)C)(=O)OCC(C)C	HCC#C*Oc@*CqT	2.7
Poly(dimethyl siloxane)	C[Si]([O])C	HOSi#CCT	2.75
Poly[oxy(2,6-diphenyl-1,4-phenylene)]	[O]C1=C(C=[C]C=C1C1=CC=CC=C1)C1=CC=CC=C1	HC*B#*B*BT	2.8
Poly(m-chloro styrene)	ClC=1C=C(C=C)C=CC1	HCC@*B@ClT	2.8
Poly(n-butyl methacrylate)	C(C(=C)C)(=O)OCCCC	HC@CC@*Oc@*CfT	2.82
Poly(vinylidene chloride)	C(=C)(Cl)Cl	HCC#ClClT	2.85
Bisphenol-A polycarbonate	CC(C)(C1=CC=C(C=C1)O)C2=CC=C(C=C2)O.C(=O)(O)O	HC*BC#CC*BOC@OT	2.9
Poly(N-vinyl carbazole)	C(=C)N1C2=CC=CC=C2C=2C=CC=CC12	HCC@*NbbT	2.9
Poly[1,1-ethane bis(4-phenyl)carbonate]	CC(c1ccc(OC(=O)[*])cc1)c2ccc([*])cc2	HC@OO*BC@C*BOT	2.9
Poly(3,4-dichlorostyrene)	ClC=1C=C(C=C)C=CC1Cl	HCC@*B#ClClT	2.94
Poly(chloro-p-xylylene)	ClC1=C(C=CC(=C1)[CH2])[CH2]	HC*B@ClCT	2.95
Poly(vinyl chloride)	C(=C)Cl	HCC@ClT	2.95
Poly(ethyl methacrylate)	C(C(=C)C)(=O)OCC	HCC#C*Oc@*CtT	3
Poly(oxy-2,2-dichloromethyltrimethylene)	[O]CC([CH2])(CCl)CCl	HCCC#C@ClC@ClCT	3
Poly(p-methoxy-o-chloro styrene)	COC1=CC(=C(C=C)C=C1)Cl	HCC@*B#ClO@CT	3.08
Poly(methyl methacrylate)	C(C(=C)C)(=O)OC	HCC#C*Oc@CT	3.1
Poly[thio(p-phenylene)]	[S]C1=CC=[C]C=C1	HS*BT	3.1
Polyoxymethylene	[O][CH2]	HOCT	3.1
Poly(tetramethylene terephthalate)	C1(C2=CC=C(C(=O)OCCCCO1)C=C2)=O	HC@O*BC@OOCCCOT	3.1
Poly(ethyl alpha-chloroacrylate)	ClC(C(=O)OCC)=C	HCC#Cl*Oc@*CtT	3.1
Ultem 1000	Cc1ccc(cc1)C(C)(C)c7ccc(Oc3ccc2C(=O)N(C(=O)c2c3)c4cccc(c4)N6C(=O)c5ccc(C)cc5C6=O)cc7	H*BhO*BC#CC*BO*Bh*BT	3.15
Poly(ether ether ketone)	[Na]Oc1ccc(cc1)O[Na].O=C(c1ccc(cc1)Cl)c1ccc(cc1)Cl	HO*BC@O*BO*BT	3.2
Poly(hexamethylene sebacamide) (measured dry)	C(CCCCCCCCC(=O)N)(=O)N1CCCCCC1	HNCCCCCCNC@OCCCCCCCC@OT	3.2
Poly(vinyl acetate)	C(C)(=O)OC=C	HCC@*Oc@CT	3.25
Poly(ethylene terephthalate)	C1(C2=CC=C(C(=O)OCCO1)C=C2)=O	HC@O*BC@OOCCOT	3.25
Poly(p-hydroxybenzoate)	OC1=CC=C(C(=O)[O-])C=C1	HO*BC@OT	3.28
Poly[2,2'-(m-phenylene)-5,5'-bibenzimidazole]	Cc5cccc(c4nc3cc(c2ccc1[nH]c(C)nc1c2)ccc3[nH]4)c5	H*Bn*Bn*BT	3.3
Poly(methyl alpha-chloroacrylate)	ClC(C(=O)OC)=C	HC#Cl*Oc@CT	3.4
Poly(epsilon-Caprolactam) (measured dry)	C1(CCCCCN1)=O	HNCCCCCC@OT	3.5
Poly(hexamethylene adipamide) (measured dry)	C(CCCCC(=O)N)(=O)N1CCCCCC1	HOCCCCCCNC@OCCCCC@OT	3.5
Torlon 2000 (basic unfilled grade, measured dry)	CNc4ccc(Cc3ccc(n2c(=O)c1ccc(C(C)=O)cc1c2=O)cc3)cc4	HC@O*Bh*BC*BNT	3.7
Polyacrylonitrile	C(C=C)#N	HC@C@NCT	4
4,4-dipheny methane Bismaleimide	O=C1CCC(=O)N1c4ccc(Cc3ccc(N2C(=O)CCC2=O)cc3)cc4	H*Ma*BC*B*MaT	3.45
Phenyl methane maleimide	CCc2cc(C)cc(N1C(=O)CCC1=O)c2	HC*B@*MaT	3.42

Bisphenol A diphenylether Bismaleimide	CC(C)(c3ccc(Oc2ccc(N1C(=O)CCC1=O)cc2)cc3)c6ccc(Oc5ccc(N4C(=O)CCC4=O)cc5)cc6	H*Ma*BO*BC#CC*BO*B*MaT	3.07
3,3-dimethyl-5,5-diethyl-4,4-diphenylmethane Bismaleimide	CCc3cc(Cc2cc(C)c(N1C(=O)CCC1=O)c(CC)c2)cc(C)c3N4C(=O)CCC4=O	H*Ma*B#CC@CC*B#CC@C*MaT	2.87

Supplementary Information 3. Network Performance with k-fold cross-validation

To further ensure the reliability of our representation HAPPY, we implemented a 5-fold cross-validation for the dielectric constant dataset in the same condition as Figure 2, which has provided us with a more rigorous assessment of model performance. As expected, we remarked a decline in the R^2 score values as we averaged over different 5-fold cross-validation datasets. As shown in Supplementary Figure 1, the average accuracy in terms of the R^2 score for HAPPY is 0.7404. However, it still significantly outperforms network trained with SMILES which has an average R^2 score value of 0.5430.



Supplementary Figure 1. Network performance averaged over 5-fold cross-validation result. While there is a decrease in R^2score, HAPPY (0.76) till outperforms SMILES (0.58)

Supplementary Information 4. Performance Comparison of HAPPY, Topo-HAPPY, PolyBERT, and Polymer Genome

To strengthen the validation of HAPPY representation, we conducted a performance benchmark against other established representations and predictors in the field. For this purpose, we augmented our experimental dataset for glass transition temperature (Tg) with an additional 312 values from various repeat unit structures, as detailed in Jozef Bicerano's book[1]. This expansion increased our Tg dataset to 527 data points, and correspondingly, the number of subgroups within HAPPY and Topo-HAPPY representations grew to 42 and 73, respectively.

We perform Tg prediction tasks on five different testing datasets, each comprising 100 randomly selected data points from the expanded dataset. The training datasets were incrementally varied, ranging from 100 to 400 to assess the robustness of the models when faced with limited data. In Supplementary Figure 2 below, we present averaged R^2 scores over five different testing dataset, comparing the performance of SMILES, HAPPY, Topo-HAPPY, PolyBERT[2], and Polymer Genome[3].

As the size of the training dataset increases, the prediction accuracy of all models improves. Notably, HAPPY consistently surpass over PolyBERT starting from dataset size 200, which is attributed to PolyBERT's expansive latent space requiring more comprehensive and extensive data for effective training. This necessity is reflected in the improved accuracy of PolyBERT when it is trained with larger dataset.

Our comparison with Polymer Genome, a renowned open-source property prediction tool trained on millions of molecules, showcases HAPPY and Topo-HAPPY's competitive edge; its prediction performance differs only by less than 5%, while the size of data set is decreased by more than thousand times. Despite Polymer Genome's high accuracy (R^2 score= 0.948), the accuracies of SMILES, HAPPY and Topo-HAPPY are commendable given their model simplicity and the constrained size of the dataset. This suggests that our coarse-grained representations offer a viable, alternative for polymer structure representation, especially in scenarios with limited data.

SMILES initially appears to have higher accuracy than HAPPY and Topo-HAPPY with smaller training datasets. This indicates the expansion of subgroups might degrade the performance of network for lower amount of training data. However, as the training data increases, the performance of network increases and becomes comparable to that of SMILES, demonstrating HAPPY's adaptability as a coarse-grained representation for polymer. Interestingly, when we eliminated data points containing subgroups absent from the training dataset, we observed a significant increase in the accuracy of a refined version of Topo-HAPPY. This refinement led to an accuracy surpassing that of SMILES, which underscores the importance of selecting the right level of coarse-graining based on the available data and the properties in question.

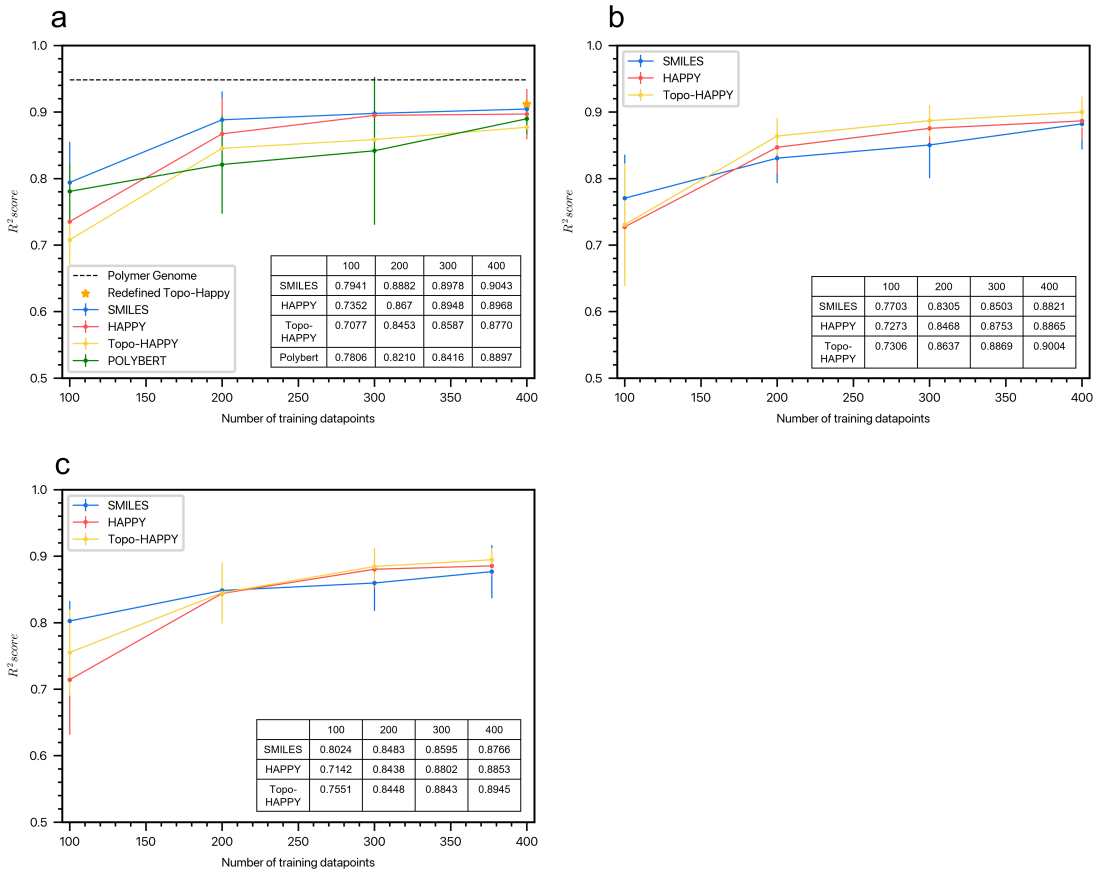
To solidify our findings and substantiate the adaptability of HAPPY and Topo-HAPPY , we conducted two additional evaluations, utilizing our entirely experimental dataset of 527 Tg datapoints.

The first approach involved redefining certain subgroups to reduce the total number from 42 to 37. Specifically, we limited carbon chain subgroups to a maximum of three C atoms, allowing longer chain to be represented through combinations of these subgroups. To make a further statistically confident conclusion, we average the evaluation result of 25 different testing dataset of 100 datapoints each randomly picked from the overall dataset. Similar to Supplementary Figure 2 (a), we vary the number of training datasets from 100, 200, 300, and 400. This simple modification led to enhanced performance of HAPPY and Topo-HAPPY, surpassing that of SMILES, as seen in Supplementary Figure 2 (b), with Topo-HAPPY achieving the highest R2 score of 0.9.

In the second approach, we focused on data points containing more frequently occurring subgroups. By removing subgroups that appeared in less than five instances, along with their associated data points, we reduced the dataset to 477 data points with 20 subgroups. The evaluations, averaged over 25 different testing sets of 100 data points each, showed that both HAPPY and Topo-HAPPY performed better than SMILES, particularly in datasets with 300 to 400 data points (Supplementary Figure 2 (c)).

These comprehensive tests underscore the importance of balancing subgroup representation and coarse-graining in relation to the available data and targeted properties. By tailoring the HAPPY representation to specific datasets, users can achieve optimal performance, demonstrating the versatility and potential of our approach.

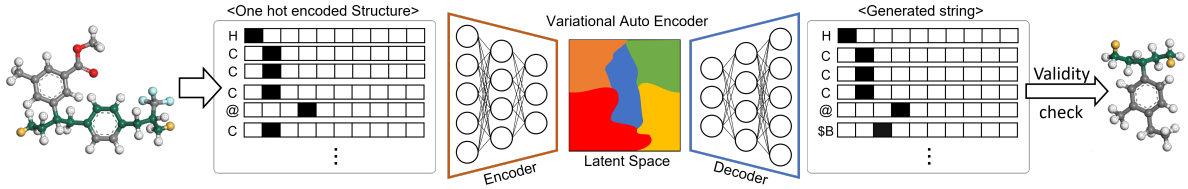
In our comprehensive network performance assessment, we also focused on training and evaluation times. To obtain realistic time measurements, we significantly increased the dataset size by replicating it 100 times, creating datasets with dimensions of 100*100 for both training and testing. Our findings indicate a marked difference in processing times between the networks. The SMILES-based network recorded a training time of 67.8ms per step. In contrast, Topo-HAPPY and HAPPY required significantly less time, with Topo-HAPPY taking 19.5ms per step and HAPPY only 18ms per step for each epoch iteration. Regarding evaluation times, the SMILES-based network took 4.31 seconds, whereas Topo-HAPPY and HAPPY were faster, recording times of 4.10 seconds and 3.64 seconds, respectively. These measurements were conducted using a batch size of 32/10000 on an Nvidia-A100 GPU, without TensorRT optimization.



Supplementary Figure 2. (a)Average R^2score over 5 different training and testing dataset for HAPPY, Topo-HAPPY, SMILES, PolyBERT, and Polymer Genome. The x axis indicates the number of training dataset that is used to trained the network which vary from 100, 200, 300, to 400 datapoints. (b) Average R^2score over 25 different training and testing dataset for HAPPY, Topo-HAPPY, and SMILES trained and evaluated with redefined carbon chain subgroups. (c) Average R^2score over 25 different training and testing dataset for HAPPY, Topo-HAPPY, and SMILES trained and evaluate with dataset disregarding datapoints with less frequently occurring subroups. The x axis indicates the number of training dataset that is used to trained the network which vary from 100, 200, 300, to 377 datapoints.

Supplementary Information 5. The Employment of HAPPY in Generative Model

One of the potential applications of machine learning in polymer informatics is the ability to sample new polymer structures with desired properties. As a first step, we tested the robustness of HAPPY representation in terms of generating valid strings. The generative model was implemented using Python3 with PyTorch library. In order to model the input strings of characters which represent repeat unit structures in the dataset obtained from either HAPPY (Topo-HAPPY) or SMILES, the strings were transformed into binary matrix representations using one-hot encoding. The generative model consisted of a Variational Autoencoder (VAE) with an encoder and decoder architecture. The encoder included three 1-dimensional convolutional layers with 9, 9, and 10 filters, respectively, and kernel sizes of 5, 5, and 7. This was followed by two fully connected layers with 256 units and a batch normalization layer. The encoder within the VAE produces mean and log variance parameters for the Gaussian distribution in the latent space, each with a dimensionality of 16. Sampling was done using the reparameterization trick, where an epsilon vector was sampled from a normal distribution and scaled by the exponential of half the log variance, then added to the mean. This sampled vector z was used as input to the decoder. The decoder contained an initial linear layer followed by a Gated Recurrent Unit (GRU) with three hidden layers of 384 units each. The output from the encoder was then fed into a linear layer equipped with N units, where N corresponds to the length of the longest string-encoded repeat unit structure within the dataset. The output of this layer was subsequently passed through a softmax activation function. The model was trained using an Adam optimizer with a learning rate of 1e-3. The loss function was a combination of Binary Cross Entropy (BCE) to measure the reconstruction loss and the KL divergence for the latent loss. The model was trained over 4000 epochs, with a batch size of 256.



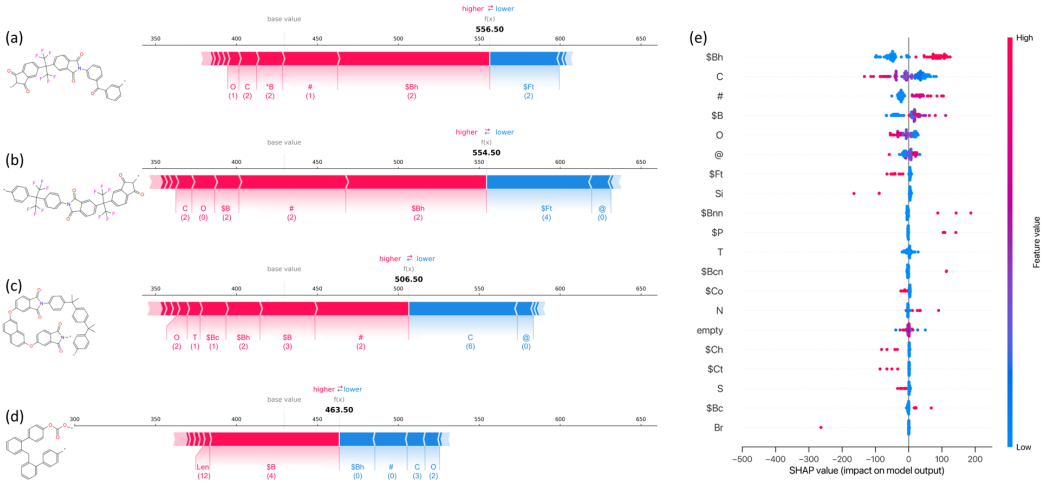
Supplementary Figure 3. Outline of the utilization HAPPY as input to Variational Autoencoder to sample new polymer repeat unit structures.

Supplementary Information 6. SHAP((SHapley Additive exPlanations) Analysis on Glass Transition Temperature Predictions

We performed SHAP analysis[4,5] on our glass transition temperature prediction model which was trained with the expanded dataset used in Supplementary Information 3, satisfying SHAP's requirement of 100 test data points. SHAP analysis elucidates the influence of each constituent by examining all possible combinations to determine their individual impact on the Tg values. For instance, as depicted in Supplementary Figure 4, SHAP analysis for structure (a) illustrates how each constituent contributes to counterbalance the Tg-reducing effect of subgroup $Ft is counterbalanced by Tg-increasing contributions from subgroups $Bh and O.

The compiled SHAP values in Supplementary Figure 4 (e) summarize the impact of each feature across 100 test predictions. The scatter plot visualizes the SHAP values for each feature, depicting the individual contribution of each feature to the overall prediction. The color coding represents the frequency of each feature's occurrence in the repeat units. In the summary plot, we observed consistent trend: the presence of the phthalimide subgroup represented by $Bh leads to a rise in Tg, while a decrease or absence of $Bh contributes to a decline in Tg. This trend of increasing Tg with the addition of phthalimide in a chemical structure was also observed in other experimental research[6].

The analysis of the SHAP values enables us to look into the impact of each subgroup on our data point to another advantage of HAPPY as a coarse-grained string representation. This level of insight is challenging to achieve with all-atomistic representations. This analysis also can be the initial guiding for a generative model based on HAPPY string representation.



Supplementary Figure 4. (a)-(d) SHAP explanation force plots for four repeat unit structures in the Glass Transition Temperature (Tg) dataset. (e) SHAP summary plot on Tg predictions. The features are ordered according to their importance. According to the behavior of the model, the existence of subgroup "$Bh" which denotes the chemical structure of phthalimide will increase Tg.

References

1. Bicerano, J. *Prediction of Polymer Properties*. *CRC Press* (Boca Raton, 2002). doi:10.1201/9780203910115.

2. Kuenneth, C. & Ramprasad, R. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nat. Commun.* **14**, (2023).

3. Kim, C., Chandrasekaran, A., Huan, T. D., Das, D. & Ramprasad, R. Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *J. Phys. Chem. C* **122**, 17575–17585 (2018).

4. Lundberg, S. M. *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).

5. Lundberg, S. M., Allen, P. G. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process Syst.* **30**, (2017).

6. Kim, T., Ju, C., Park, C. & Kang, H. Enhanced, parallel liquid crystal alignment based on polystyrene substituted with phthalimidoyl groups. *RSC Adv.* **9**, 9755–9761 (2019).