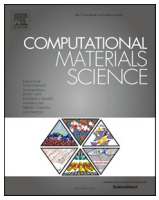
Computational Materials Science 236 (2024) 112863

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci



Full Length Article

A glimpse inside materials: Polymer structure – Glass transition temperature relationship as observed by a trained artificial intelligence

Luis A. Miccio [a,b,c,*], Claudia Borredon [a], Gustavo A. Schwartz [a,b,*]

[a] Centro de Física de Materiales (CSIC-UPV/EHU) – Materials Physics Center (MPC), P. M. de Lardizábal 5, 20018 San Sebastián, Spain
[b] Donostia International Physics Center, P. M. de Lardizábal 4, 20018 San Sebastián, Spain
[c] Institute of Materials Science and Technology (INTEMA), National Research Council (CONICET), Colón 10850, 7600 Mar del Plata, Buenos Aires, Argentina

ARTICLE INFO

Keywords:
QSPR
Polymer properties
Convolutional neural networks
Grad-CAM

ABSTRACT

Artificial neural networks (ANNs), a subset of Quantitative Structure-Property Relationship (QSPR) methods, offer a promising avenue for addressing challenges in materials science. In particular, ANNs can learn intricated patterns within the experimental data, enabling them to predict properties and recognize complex relationships with remarkable accuracy. However, the opacity of ANNs, normally acting as black boxes, raises concerns about their reliability and interpretability. To enhance their transparency and to uncover the underlying relationships between chemical features and material properties, we propose a novel approach that employs Gradient-weighted Class Activation Mapping (Grad-CAM) applied to Convolutional Neural Networks (CNNs). By analyzing these attention maps, we identify the crucial chemical features influencing the prediction of a polymer property, specifically the glass transition temperature (Tg). Our methodology is validated using a dataset of atactic acrylates, allowing us to not only predict Tg values for a control group of polymers but also to quantitatively assess the impact of individual monomer structural elements on these predictions. This work proposes a step towards transparent models in materials science, contributing to a deeper understanding of the intricate relationship between chemical structures and material properties.

1. Introduction

In the last decades, the design and development of new materials has gained impulse by the progressive introduction of many techniques, both experimental and numerical. Among the later, the Quantitative Structure Property Relationship models (QSPR) have been successfully employed for estimating several material properties, especially those too experimentally expensive to be measured. Among the many available tools that QSPR methods offer, it is particularly interesting the case of artificial neural networks (ANNs) applied to materials science. To date, several interesting approaches have been proposed for solving problems in a wide number of challenging fields [1–7], like visual pattern recognition of experimental data [8], generation of complex configurations in proteins [9–11] or the estimation of unknown properties in polymers [12–16].The process begins by collecting data from either experimental results or literature (or directly from an existing database when available) and using ANNs for learning the hidden patterns within. When successful, this approach leads to a trained ANN, able to perform the previously mentioned tasks (property prediction, pattern recognition) on new data. In this pursuit, several ANNs architectures and codifications have been proposed, and it has been shown that ANNs can be trained by using solely a representation of the chemical structure as input and the property of interest as output [14–17]. From the architectures that have proven to be successful in the task, there are some very interesting examples of good performance (with varying degrees of complexity and accuracy), being to date those based on convolutional (CNN), graph convolutional (GCN) and recurrent neural networks (RNN) the most general and accurate [18–22].

Notwithstanding these fascinating improvements, a common issue with all these methods is their complexity - interpretability trade off i.e.: the ANNs act as black boxes where codified information is fed into the network and a given output (either a numeric value or class prediction) is obtained [23]. No information on which elements of the input data are responsible of the output results is obtained from this process. Therefore, in order to establish intelligent systems as fully reliable tools and move towards a better integration of them into our everyday tasks, it is

* Corresponding author at: Donostia International Physics Center, P. M. de Lardizábal 4, 20018 San Sebastián, Spain (L.A. Miccio); Centro de Física de Materiales (CSIC-UPV/EHU) – Materials Physics Center (MPC), P. M. de Lardizábal 5, 20018 San Sebastián, Spain (G.A. Schwartz).
E-mail addresses: luisalejandro.miccios@ehu.eus (L.A. Miccio), gustavo.schwartz@csic.es (G.A. Schwartz).

https://doi.org/10.1016/j.commatsci.2024.112863
Received 24 October 2023; Received in revised form 27 January 2024; Accepted 7 February 2024
Available online 17 February 2024

mandatory to develop 'transparent' models (i.e.: models or training procedures that can provide an explanation of how and why they yield a particular output [24]). In an effort to obtain more transparent models, akin to the steps taken in bias removal processes, it is necessary to use tools for studying potential biases and deviations that transcend conventional benchmarking performance metrics, such as absolute and relative errors. The utilization of graph attention networks (GAN) [25] or a combination of saliency maps and GCN [26] stand as exemplary approaches, possessing the capability to discern and reveal latent knowledge embedded within atoms and substructures. Furthermore, the information acquired can contribute with valuable insights to ongoing scientific discussions such as the interplay between chemical structure and the collective relaxation process in glass transition. In this sense, ANN training can be thought as a two-fold process: a) the mining and curing of (related) critical information hidden in the multidimensional training data (chemical features) and b) the correlation of these features to the experimental property value for constructing the prediction. These chemical features, if accessible, might be employed as complementary information, like a numerically directed focus for studying the underlying physical mechanisms involved in the emergence of the material property under study.

In this work we use attention maps constructed with Gradient-weighted Class Activation Mapping (Grad-CAM) [27] on the layers of convolutional neural networks (CNNs) to find and show the relationship between the chemical features that are more actively influencing the predicted polymer property. In this way, we can evaluate where the focus of the CNN is directed, therefore finding and measuring the contributions of the different chemical features to the glass transition temperature. We perform a comparative analysis between the attention maps generated from two distinct chemical structure representations: one hot encoded binary matrices and 2D graphical representations. By employing Grad-CAM on both types of inputs, we were able to study the networks' attention patterns and the subsequent impact on interpretability and performance. Although the graphical representations introduce an additional layer of complexity, due to their larger dimensions and the inclusion of more spatial information, it is noteworthy that these representations remain entirely comprehensible to humans. This comparison thus sheds light on how different input representations influence the networks' ability to discern crucial chemical features governing Tg predictions. To obtain a proof of concept of the feasibility of these methods, and to avoid including other chemical and experimental effects (different degrees of tacticity, varying backbone architecture and flexibility, among others), we used a dataset composed by more than 200 chemically similar polymeric compounds (atactic acrylates). After training the ANN with this dataset, we can find, for a control group of polymers (not included in the training or validation datasets), not only the value of the glass transition temperature but also the relative contribution of each part of the monomer structure to the corresponding T_g.

2. Methods

In this section we describe the details of the different numerical methods employed along this work. We focus on the dataset, the encoding methods used for converting the chemical structures into ANN compatible data (we used and compared two different approaches), the ANN's architectures and their corresponding tuning process.

Dataset. We used a curated dataset composed by the chemical structures of about 200 atactic polyacrylates and their corresponding T_g values [28–30] (see Table SI 1). The monomer units of these acrylates were selected as the minimum information block for ANN training, i.e.: the smallest amount of polymer info that contains enough relevant information to allow the learning process.

For comparison purposes, we used two different encoding processes commonly employed for training an ANN. On the one hand, we transformed the chemical structure data into linear strings by using a Simplified Molecular Input Line Entry System (SMILES) [31–35] and converted it into binary matrices by using a one hot encoding algorithm [36] and a *dictionary* (see Figure SI 1. Despite any limitations, like for instance the sparsity of the resulting matrices, the benefit of this method lies in the fact that each position of the input directly represents a chemical species of the monomer. Furthermore, since all monomers belong to the same family, backbone structure and interactions are implicitly learned by the models. Therefore, it allows a direct exploration of the ANN attention degree to the structure - T_g relationship along the training process.

On the other hand, we constructed a second set of inputs, this time as graphic representations of the monomer's molecular structure (showing how the atoms are arranged in 2D-space and how they are chemically bonded). Unlike other approaches, which have a more limited number of symbols (and therefore less representational capabilities), graphical inputs provide additional geometric information of the molecular structure (see Figure SI 2). As a result, an even more direct human interpretability of the structure is accomplished. This second dataset was in turn augmented by following the transformations typically employed in artificial vision (the original structures were vertically, horizontally, and vertically + horizontally rotated, therefore quadruplicating dataset size). It is worth noticing that this input images have 512 x 256 pixels (elements), far more than the 66 x 16 of the one hot encoded approach. Therefore, although they are interpretable representations of the chemical structure, they also are more sparse matrices in comparison.

ANNs architecture and optimization. We employed two convolutional neural networks (CNN-a, fed with one hot encoded matrices, and CNN-b, fed with the graphic representations of the monomers) and the dataset of corresponding glass transition temperatures to be learned and predicted. We call features those parts of the molecule that strongly contribute to the structure- T_g relationship, and exhibit a notably heightened significance compared to other elements within the input after the training process. Fig. 1 shows a schematic view of architecture employed for CNN-a (see Figure SI 3 for the schematic of the architecture employed for CNN-b). As shown, the monomer structure is codified (through SMILES) into a 2D binary matrix which is then fed to the convolutional layers (C_1 to C_5) that in turn extract the relevant features. The result is flattened into a 1D vector ($X \in R^n$); this vector feeds a fully connected layer (FC_1) with ReLU activations. To achieve the best possible performance for the ANNs before extracting features, different values of the network hyperparameters were explored. In this way, several networks with varying parameter values were trained and compared. We based the comparison on the raw performance (minimum relative error) achieved on the dataset, and then also on the attention map (see attention maps section) of each layer. In addition, a dropout [37] algorithm was used, with dropping probabilities ranging from 0 to 0,3. Finally, the last hidden layer (FC_1) was connected to a single neuron with a linear activation function responsible of providing the glass transition temperature value.

As done in our previous works, to ensure equal weighting of low and high T_g data values during training, the loss function was defined as the average relative error between the actual (A_i) and forecasted (F_i) glass transition temperatures.

$$Loss = \frac{100}{m_x} \cdot \sum_{i=1}^{m_x} \left(\frac{A_i - F_i}{A_i} \right)$$

An ADAM optimizer [38] with a learning rate (lr) ranging from 0.0001 to 0.00001 was employed for speeding up the convergence during training (beta 1 and beta 2 were 0.99 and 0.999, respectively). The calculations were performed by using mini batches of 5 to 20 samples.

As usual, the data were divided into train and validation subsets for the training process (as in previous works [14–16], a grid search was conducted by using the parameters shown in Table 1). In addition, an external control group (independent from the previous subsets and

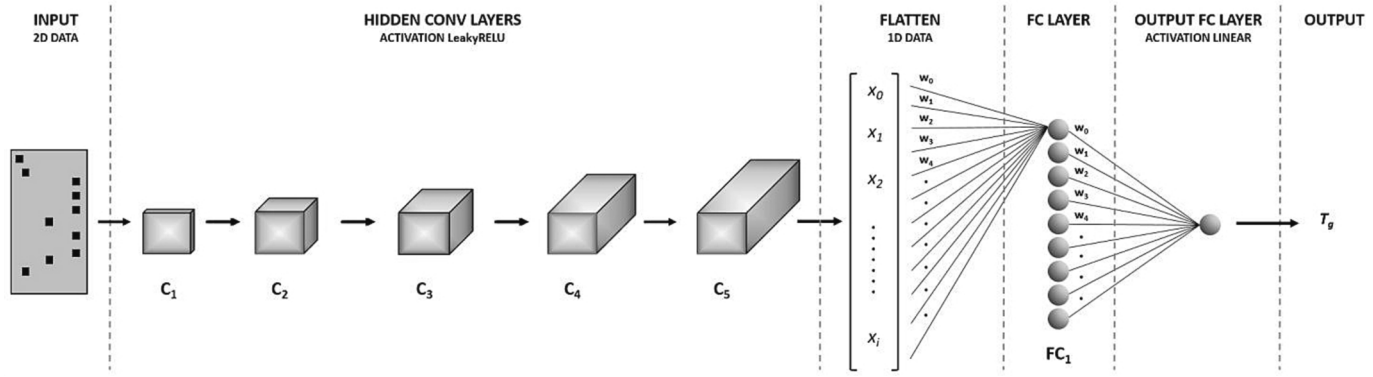


Fig. 1. Schematic picture of the artificial neural network employed for studying the attention maps on the one hot encoded matrices.

Table 1
ANN Hyperparameters.

Item	Value
Data split ratio (train/test)	80/20
Dropout probability	0 to 0,3
Mini batch size	5 to 20
Learning rate	0,0001 to 0,00001
Beta1 (Beta2)	0,99 (0,999)
# Hidden neurons in FC1	50

representative of the main chemical features) was included for studying the CNNs attention maps.

Attention maps. According to literature, there typically exists a trade-off between accuracy and interpretability of the ANNs' output [23,27]. Some systems are highly interpretable but at the same time not accurate or robust enough [39]. On the other hand, these pipelines in which each stage has been designed separately can be thought as more interpretable. This is because, to a certain extent, each constituent element of the process attains an intuitive explanation. In this regard, several authors have worked in this interpretability issue. For instance, Zhou et al. [40] proposed a technique called Class Activation Mapping (CAM) for identifying discriminative regions used by a restricted class of image classification CNNs (by directly trading off complexity and performance for more transparency into the inner working of the model). More recently, Selvaraju et al. [27] proposed a generalization of the previous method, Gradient-weighted Class Activation Mapping or Grad-CAM, which can be applied to a significantly broader range of CNN models without requiring any architectural changes or re-training.

It has been shown that in a CNN, the deeper layers are able to capture higher-level visual details of the input [41,42]. Furthermore, convolutional layers can naturally retain spatial information which is lost in fully-connected layers. In this way, it is expected the last convolutional layers to have the best compromise between high-level features and detailed spatial information. The neurons in these layers look for specific information in the image (i.e.: object parts), and Grad-CAM uses the gradient information flowing into this last convolutional layer of the CNN to assign importance values to each neuron for a given decision. However, while Grad-CAM is class-discriminative and localizes relevant image regions, it lacks the ability to highlight fine-grained details like other techniques, namely Guided Backpropagation [43] or Deconvolution [44]. Guided Backpropagation visualizes gradients with respect to the image where negative gradients are suppressed when back-propagating through ReLU layers. Intuitively, this aims to capture pixels detected by neurons, not the ones that these same neurons are suppressing. In order to combine the best aspects of both methods, Guided Backpropagation and Grad-CAM visualizations were fused via element-wise multiplication, therefore yielding a visualization that is both high-resolution and class-discriminative [44]. Therefore, we employed this guided Grad-Cam algorithm in the hidden layers of our CNN trained with the chemical structure and the glass transition temperature of atactic acrylates. In this way, we constructed and analysed the attention maps of both approaches: "CNN-a" (trained with one hot encoded inputs) and "CNN-b" (in turn trained with images of the chemical structures as inputs).

3. Results and Discussion

Fig. 2 summarizes the predicted vs experimental glass transition temperature values obtained from the trained CNN-a on the one hot encoded monomers for the external control group (see also the results obtained on the training and validation sets, together with their chemical structures-relative deviation analysis in Figure SI 4). As shown, the convolutional neural network does capture the relationship between the chemical structure and the T_g of the polyacrylates all along the studied temperature range. Particularly in the external control group, and except for poly (octyl methacrylate), the individual relative deviations are all well within a 10 % deviation, in close agreement with the observed values for the validation sets. It is worth noticing that poly (octyl methacrylate) can undergo nanophase separations (like other long n-alkyl acrylates) and therefore show different behaviour in

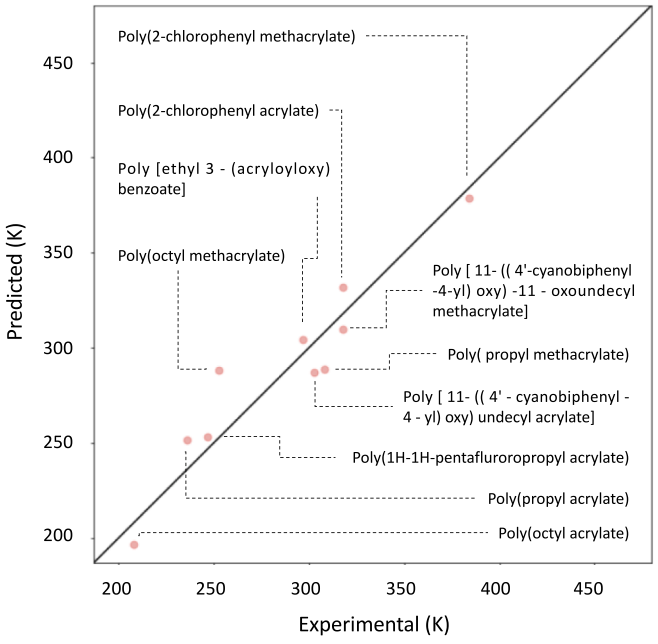


Fig. 2. Predicted vs experimental glass transition values obtained from trained CNN-a on the external control group of acrylates.

relationship with the glass transition and overall dynamics[45,46]).

Beyond the relatively low errors during the numerical validations (in the range of 10 %), our aim is to study which substructures are more relevant to the network during training and prediction processes. In that sense, both the presence of outliers and the chemical structures that are not well generalized, i.e.: not well learned during training due to other underlying physical processes like the above-mentioned nanophase separations, sample preparation protocols, or simply due to underrepresented chemical features in the training set, could be quantitatively analysed and compared beyond average error metrics. While error metrics provide a quantifiable measure of performance, the complex interplay between chemical structures makes it crucial to explore the models' interpretability and their ability to generalize across different scenarios. By analysing the attention maps and the inputs, we can gain a richer understanding of the models' comprehension of chemical features' impact on glass transition temperature.

Fig. 3 shows the one hot encoded matrices and the corresponding attention maps of the last hidden convolutional layer (C_5 in Fig. 1) for the monomers in the control group. As shown, the coloured heatmaps confirm that the network pays attention exclusively to the chemical structure.

It is noteworthy that no influence of the rest of the input matrix is detected, all zeros in the original binary matrix are mapped to the lowest attention scale (i.e.: dark blue), meaning therefore that in this scale the sparsity of the matrix does not influence the T_g prediction. It can be also observed that the attention degree to different substructures is not constant throughout the monomer. For example, there appears to be a tendency to include phenyl groups as a very relevant feature (i.e.: higher in the attention colour scale) in most of these cases where they are present in the structure (like in cases a, b, i, and j of the control group). A similar behaviour has also been observed in our previous approaches on fully connected networks [14] (in these cases, since the method is insensitive to changes in the spatial information of the input, it yields only approximate results compared to the current approach). It is noteworthy that the current approach focuses on the degree of relevance of the features, and not in the specific way they affect the glass transition (i.e. whether they tend to increase or reduce it).

From these observations, it can be argued that the attention maps of the trained network are in close agreement with reported glass transition - chemical structure behavior [47]. For example, CNN-a is sensitive to the change in the backbone stiffness of the polymer [48–50] due to species adjacent to the monomer's functional group (i.e.: the presence of a methyl group in methacrylates increases T_g). This effect can cause T_g differences of about 100 K, like these observed between Poly (2 - chlorophenyl methacrylate) and Poly (2 - chlorophenyl acrylate) (h and I, respectively, in Fig. 3).

The scheme in Fig. 4 shows in more detail the areas of the one hot encoded binary matrices of these monomers, and the degree of attention on the methacrylate position. As shown, the last convolutional layer of CNN-a evaluates the presence/absence of side-groups in both Poly (2-chlorophenyl methacrylate) and Poly (2-chlorophenyl acrylate). On the other hand, the trained network also shows sensitivity to the chemical characteristics of the monomer's "tail", like the 2-chlorophenyl group in Poly (2 - chlorophenyl methacrylate). For example, it has been reported that the presence of alkyl chains can cause changes in T_g by decreasing the resistance to movement (lubrication effect[51,52]). As a simplification, it can be assumed that the longer the alkyl chain the lower the observed glass transition temperature [51,52]. However, in the case of long linear alkyl chains, other effects have been also reported, like the above-mentioned nanophase separation for chains longer than 4 or 6C atoms [45,46]. As shown for Poly (propyl methacrylate), Poly (propyl acrylate), Poly (octyl methacrylate) and Poly (octyl acrylate) (d to g in Fig. 3), the CNN's degree of attention to the linear chain is high in comparison with the rest of the structure. It has been also widely reported the effect of bulkier phenyl groups, whose presence in the tail of a given polymer will dramatically affect the observed glass transition

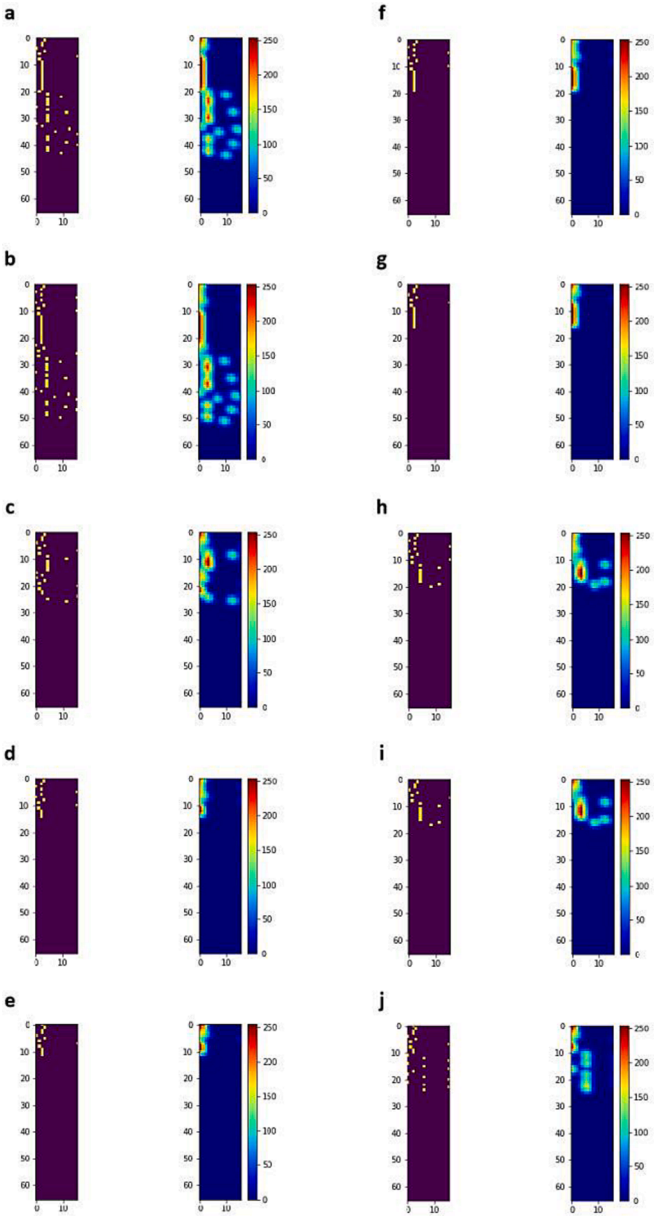


Fig. 3. One hot encoded matrices and attention maps (arbitrary units) of the monomers in the control group. The labels a-j correspond to: a) Poly [11-((4′-cyanobiphenyl-4-yl)oxy)undecyl acrylate], b) Poly [11-((4′-cyanobiphenyl-4-yl) oxy)-11-oxoundecyl methacrylate], c) Poly[ethyl 3-(acryloyloxy)benzoate], d) Poly(propyl methacrylate), e) Poly(propyl acrylate), f) Poly(octyl methacrylate), g) Poly(octyl acrylate), h) Poly(2-chlorophenyl methacrylate), i) Poly(2-chlorophenyl acrylate), and j) Poly(1H,1H-pentafluoropropyl acrylate).

temperature [53,54]. Both effects are also shown to be taken into account by trained CNN-a, as evidenced by the high attention degree (dark red in the colorscale) observed for the linear chain and phenyl groups of Poly [11 - ((4' - cyanobiphenyl −4- yl) oxy) −11- oxoundecyl methacrylate] in Fig. 4.

In order to take a step further in the analysis of the neural networks' training process, and to also shed some more light on the sensitivity of these architectures to the performance - interpretability trade-off, we compared the above discussed results with the outputs of a network trained on images of the monomer's structures.

As mentioned in the methods section, CNN-b was trained with a graphical representation of each monomer structure, over the same dataset as before. We used a similar architecture for both neural

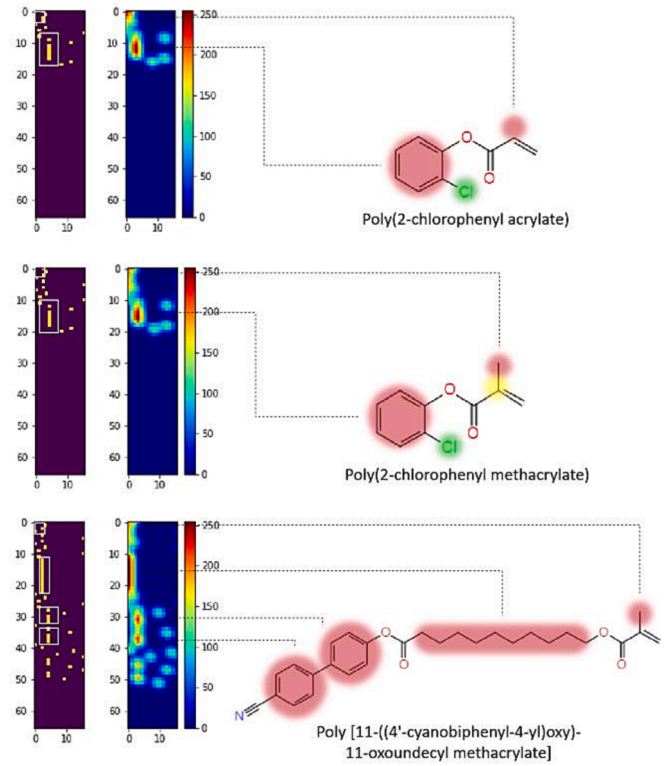


Fig. 4. One hot encoded matrices and attention maps (arbitrary units) of Poly (2 - chlorophenyl methacrylate), Poly (2 - chlorophenyl acrylate), and Poly [11-((4′-cyanobiphenyl-4-yl) oxy) −11- oxoundecyl methacrylate]. In order to facilitate the interpretation, some parts of the one hot encoded matrix have been boxed and the corresponding color-scale has been overlapped to the chemical structure.

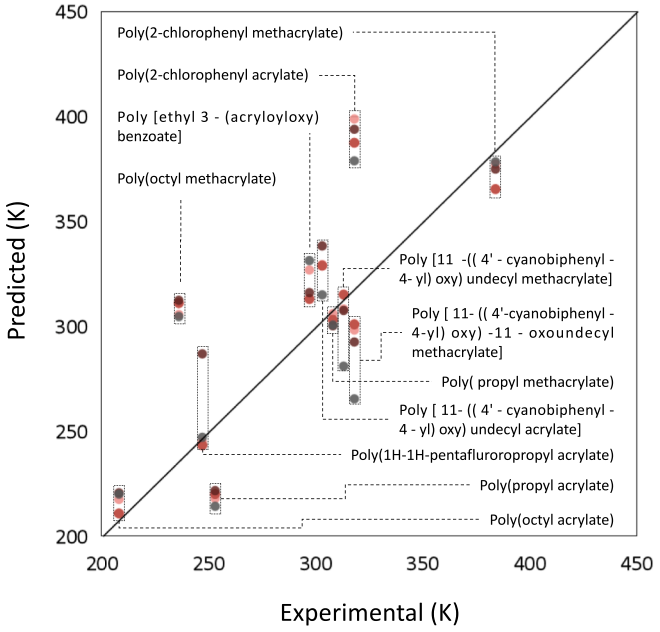


Fig. 5. Predicted vs experimental glass transition values obtained from the trained CNN-b on the external control group of acrylates. Each dot color represent a different transformation of the data: red = original, pink = vertical, darkgrey = horizontal, and grey = both.

networks, in order to obtain a direct comparison among the attention along the different convolutional layers. The details of CNN-b architecture are presented in the Supplementary Information (see Figure SI 3). Fig. 5 shows the experimental vs predicted glass transition values obtained on the control group with CNN-b. As shown, the generalization of the network is worse than that of CNN-a, yielding less accurate predictions, like in the case of Poly (2 - chlorophenyl acrylate, whose predicted T_g clearly deviates from the reported experimental value. Additionally, due to the augmentation of the dataset (now there are 4 inputs per polymer instead of only one) it is possible to see that the precision of the prediction (i.e.: how close the same instances are predicted to each other) in some cases is also low, most noticeably in Poly (propyl methacrylate) and Poly (2- chlorophenyl methacrylate). In addition, the experimental vs predicted glass transition temperatures obtained from the augmented internal training and validation sets of acrylates (see Figure SI 5) reveal a tendency of trained CNN-b to underestimate the T_g of the input values at high temperatures.

No obvious predominance of any applied transformation of the input has been noticed to present systematically larger generalization problems. The results suggest that some chemical structures are just harder to generalize rather than a transformation-related problem. There are some differences, however, related to the sensitivity to particular chemical structures, like fluorinated tails. These facts can be better explained by comparing the attention maps of the last convolutional layer in Figure SI 6, which show that regardless the applied transformation the focus of CNN-b is directed towards the same chemical substructures.

The attention maps along all the hidden convolutional layers of CNN-b for Poly [11- ((4′- cyanobiphenyl −4- yl) oxy) −11- oxoundecyl methacrylate] (see Fig. 6) also show that the first layers (C_1 and C_2 in Figure SI 3) primarily separate the chemical structure from the background, which appears higher in the colorscale. Then, it can be noticed that the attention progressively shifts to the chemical structure, that becomes clearly dominant already in C_3. Finally, the last layers tend to focus only on the chemical structure and mainly in some substructures like the phenyl groups and species like oxygen atoms. This effect is especially strong in C_5 and C_6, where the background is completely neglected and only the chemical structure appears to be relevant to the network (see Figure SI 7 for layer by layer details on the the one hot

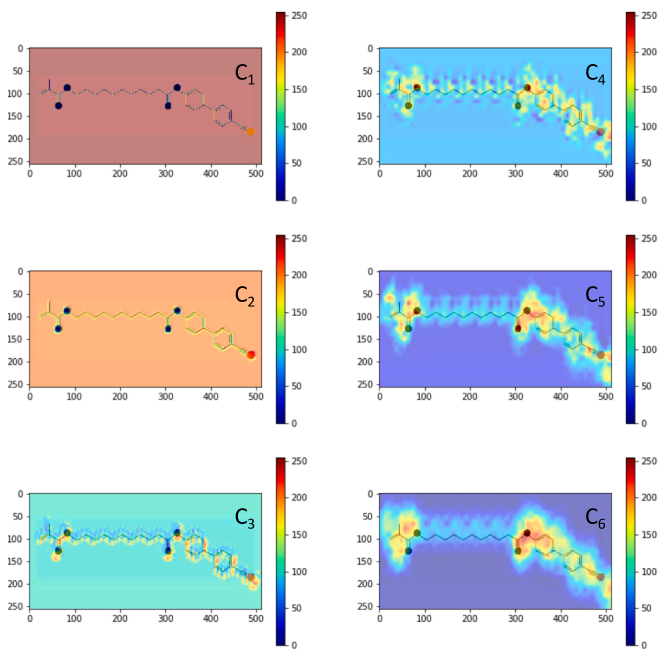


Fig. 6. Attention maps (arbitrary units) along the hidden convolutional layers of CNN_b (C_1, C_2, C_3, C_4, C_5 and C_6) of Poly [11- ((4 '- cyanobiphenyl −4- yl) oxy) −11- oxoundecyl methacrylate].

encoded approach).

It is noteworthy that some of the key chemical features mentioned during CNN-a analysis are now slightly overlooked, like the linear chains in Poly [11- ((4′- cyanobiphenyl-4-yl) oxy) −11- oxoundecyl methacrylate]. Moreover, the relative importance of phenyl groups is also reduced in comparison to the results of CNN-a on one hot encoding binary matrices (see for example the attention map for the binary input of Poly [11- ((4′-cyanobiphenyl-4-yl) oxy) −11- oxoundecyl methacrylate] in Fig. 4). As a result, and in spite of CNN-b training with an augmented dataset, CNN-a shows that the generalization directly from images of the one hot encoded structures by using a similar architecture yields better results in comparison with the learning provided by the direct images.

One way of rationalizing these results is to focus on the fact that when inputting the data as one hot encoded binary matrices, the relative position of each group gets fixed to a particular pattern (that depends on the dictionary employed in the codification).

Therefore, the number of variables that an ANN must learn and later generalize to new data is also reduced in comparison to the images fed to CNN-b. For example, the presence/absence of a side methyl group (i.e.: methacrylate/acrylate) always appear in the same region of the binary matrix (and with the same relative shape), while in the image input case it can vary both in size and in position (therefore having several more degrees of freedom in comparison). In the same way, other important chemical features have an equally well defined relative position and shape, like the phenyl groups or the linear chains, as they can only appear in fixed columns of the input matrix (defined according to their position in the dictionary). Finally, the number of matrix elements to process by the ANN is also different, being much smaller in the case of the one hot encoded matrices where, in addition, the ratio of useful information to background is also larger.

Our results suggest that the one hot encoding of the inputs is, while less explicit than the images, easier to manage for training with these network's architectures. As shown by the comparison between CNN-a and CNN-b, the degree of attention to important chemical features is more efficient in the case of the one hot encoded matrices. We noticed that the increased interpretability of the image inputs of CNN-b led to a comparative decrease in the performance, not only as shown through standard metrics like the relative average deviation, but also observed when analysing the obtained attention maps. On the other hand, our work shows that an excellent compromise between interpretability and performance can be obtained with the rather simple and computationally economic approach employed in CNN-a.

4. Conclusions

We employed convolutional neural networks and gradient-weighted class activation mapping to study the connection between certain chemical structures and the emerging material properties, with a focus on the glass transition temperature of atactic acrylates. Our approach helps to understand the main chemical features influencing Tg predictions and has demonstrated robust performance when validated against an external control group. The attention maps generated by the CNNs highlighted their ability to prioritize chemical components over matrix sparsity and depicted varying importance of substructures, aligning with experimentally known glass transition-chemical structure relationships. To address the interpretability-performance trade-off, we compared this approach with a CNN trained on graphical monomer representations. Although both trained models captured the chemical features, in this case the results suggest a balance between interpretability and performance. In summary, we propose a transparent approach for understanding the interplay between chemical structures and material properties. This method offers insights bridging conventional models and machine learning, with implications for informed material design and development.

Associated Content

Additional information on the glass transition temperatures dataset, one hot encoded inputs and dictionary, CNN-b architecture, predicted vs experimental results and attention maps can be found in the Supplementary Information file (SI).

Data Availability Statements

The data that supports the findings of this study are available within the article and in the Supporting Information file (SI).

CRediT authorship contribution statement

Luis A. Miccio: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Claudia Borredon:** Validation, Writing – review & editing. **Gustavo A. Schwartz:** Funding acquisition, Project administration, Resources, Formal Analysis, Investigation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We gratefully acknowledge the financial support from the Spanish Government Ministerio de Ciencia e Innovación Research Project PID2019-104650GB-C21 / PID2022-138070NB-C22 MCIN/ AEI /10.13039/501100011033 and the Basque Government (IT-1566-22). We also acknowledge the support of NVIDIA Corporation with the donation of two GPUs used for this research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.commatsci.2024.112863.

References

[1] G.A. Schwartz, Prediction of rheometric properties of compounds by using artificial neural networks, Rubber Chem. Technol. 74 (2001) 116–123.
[2] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1798–1828.
[3] D. Silver, et al., Mastering the game of go with deep neural networks and tree search, Nature 529 (2016) 484–489.
[4] Y. Kim, Y. Jernite, D. Sontag, A.M. Rush, Character-Aware neural language models, in: 30th AAAI Conf Artif. Intell. AAAI 2016, 2016, pp. 2741–2749.
[5] G. Chen, Z. Shen, Y. Li, A machine-learning-assisted study of the permeability of small drug-like molecules across lipid membranes, Phys. Chem. Chem. Phys. 22 (2020) 19687–19696.
[6] A. Lavecchia, Machine-learning approaches in drug discovery: Methods and applications, Drug Discov. Today 20 (2015) 318–331.
[7] Y. Liu, T. Zhao, W. Ju, S. Shi, Materials discovery and design using machine learning, J. Mater. 3 (2017) 159–177.
[8] L. Chen, et al., Polymer informatics: Current status and critical next steps, Mater. Sci. Eng. R Reports 144 (2021).
[9] H. Frauenfelder, B. McMahon, Dynamics and function of proteins: The search for general concepts, Proc. Natl Acad. Sci. USA 95 (1998) 4795–4797.
[10] T. Nguyen, N. Le, Journal of Molecular Graphics and Modelling Prediction of ATP-binding sites in membrane proteins using a two- dimensional convolutional neural network, J. Mol. Graph. Model. 92 (2019) 86–93.
[11] G. Gupta, A. Liu, J. Ghosh, Automated hierarchical density shaving: A robust automated clustering and visualization framework for large biological data sets, IEEE/ACM Trans. Comput. Biol. Bioinforma. 7 (2010) 223–237.
[12] Y. Zhang, X. Xu, Machine learning glass transition temperature of polymers, Heliyon 6 (2020).
[13] A.L. Nazarova, et al., Dielectric Polymer Property Prediction Using Recurrent Neural Networks with Optimizations, J. Chem. Inf. Model. 61 (2021) 2175–2186.

[14] L.A. Miccio, G.A. Schwartz, Localizing and quantifying the intra-monomer contributions to the glass transition temperature using artificial neural networks, Polymer (guildf). 203 (2020) 122786.
[15] L.A. Miccio, G.A. Schwartz, Mapping Chemical Structure-Glass Transition Temperature Relationship through Artificial Intelligence, Macromolecules 54 (2021) 1811–1817.
[16] L.A. Miccio, G.A. Schwartz, From chemical structure to quantitative polymer properties prediction through convolutional neural networks, Polymer (guildf). 193 (2020) 122341.
[17] B.E. Mattioni, P.C. Jurs, Prediction of Glass Transition Temperatures from Monomer and Repeat Unit Structure Using Computational Neural Networks, J. Chem. Inf. Comput. Sci. 42 (2002) 232–240.
[18] G. Chen, L. Tao, Y. Li, Predicting Polymers' Glass Transition Temperature by a Chemical Language Processing Model. Polym. 2021, Vol. 13, Page 1898 13, 1898 (2021).
[19] C. Borredon, L.A. Miccio, S. Cerveny, G.A. Schwartz, Characterising the glass transition temperature-structure relationship through a recurrent neural network, J. Non-Crystalline Solids X 18 (2023) 100185.
[20] I.V. Volgin, et al., Machine Learning with Enormous 'synthetic' Data Sets: Predicting Glass Transition Temperature of Polyimides Using Graph Convolutional Neural Networks, ACS Omega 7 (2022) 43678–43691.
[21] P. Reiser, et al., Graph neural networks for materials science and chemistry, Commun. Mater. 3 (2022).
[22] C. Li, et al., Correlated RNN Framework to Quickly Generate Molecules with Desired Properties for Energetic Materials in the Low Data Regime, J. Chem. Inf. Model. 62 (2022) 4873–4887.
[23] Z.C. Lipton, The Mythos of Model Interpretability. (2016).
[24] M.T. Ribeiro, S. Singh, C. Guestrin, 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 13-17-August-2016, 1135–1144 (2016).
[25] T. Nguyen, M. Bavarian, A Machine Learning Framework for Predicting the Glass Transition Temperature of Homopolymers, Ind. Eng. Chem. Res. (2022), https://doi.org/10.1021/ACS.IECR.2C01302/SUPPL_FILE/IE2C01302_SI_001.PDF.
[26] J. Hu, Z. Li, J. Lin, L. Zhang, Prediction and Interpretability of Glass Transition Temperature of Homopolymers by Data-Augmented Graph Convolutional Neural Networks, ACS Appl. Mater. Interfaces 15 (2023) 54006–54017.
[27] R.R. Selvaraju, et al., Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, Int. J. Comput. vis. 128 (2020) 336–359.
[28] D.J. Plazek, K.L. Ngai, The Glass Temperature. in Physical Properties of Polymers Handbook (ed. Mark, J. E.) 187–215 (Springer New York, 2007).
[29] C. Bertinetto, et al. Prediction of the glass transition temperature of (meth) acrylic polymers containing phenyl groups by recursive neural network. **48**, 7121–7129 (2007).
[30] G. Wypych, Handbook of Polymers, Elsevier, 2016.
[31] M. Hirohara, Y. Saito, Y. Koda, K. Sato, Y. Sakakibara, Convolutional neural network based on SMILES representation of compounds for detecting chemical motif, BMC Bioinformatics 19 (2018).
[32] S. Zheng, X. Yan, Y. Yang, J. Xu, Identifying Structure-Property Relationships through SMILES Syntax Analysis with Self-Attention Mechanism, J. Chem. Inf. Model. 59 (2019) 914–923.
[33] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28, 31–36 (1988).
[34] M. Hirohara, Y. Saito, Y. Koda, K. Sato, Y. Sakakibara, Convolutional neural network based on SMILES representation of compounds for detecting chemical motif, BMC Bioinformatics 19 (2018) 526.
[35] G.A. Pinheiro et al., Machine Learning Prediction of Nine Molecular Properties Based on the SMILES Representation of the QM9 Quantum-Chemistry Dataset. (2020) doi:10.1021/acs.jpca.0c05969.
[36] H. Alkharusi, Categorical variables in regression analysis: a comparison of dummy and effect coding, Int. J. Educ. 4 (2012) 202–210.
[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[38] D.P. Kingma, J.A. Ba, A Method for Stochastic Optimization. arXiv1412.6980 [cs] (2014).
[39] P. Jackson, Introduction to expert systems. (1998).
[40] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Object detectors emerge in deep scene CNNs. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. (2015).
[41] Y. Bengio, A. Courville, P. Vincent, Representation Learning: A Review and New Perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2012) 1798–1828.
[42] A. Mahendran, A. Vedaldi, Visualizing Deep Convolutional Neural Networks Using Natural Pre-Images, Int. J. Comput. vis. 120 (2015) 233–255.
[43] J.T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Work. Track Proc. (2015).
[44] R.R. Selvaraju, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int. J. Comput. Vis. 128, 336–359 (2020).
[45] M. Beiner, K. Schröter, E. Hempel, S. Reissig, E. Donth, Multiple glass transition and nanophase separation in poly(n-alkyl methacrylate) homopolymers, Macromolecules 32 (1999) 6278–6282.
[46] M. Beiner, H. Huth, Nanophase separation and hindered glass transition in side-chain polymers, Nat. Mater. 29 (2) (2003) 595–599.
[47] L.A. Miccio, C. Borredon, U. Casado, A.D. Phan, G.A. Schwartz, Approaching Polymer Dynamics Combining Artificial Neural Networks and Elastically Collective Nonlinear Langevin Equation, Polymers (basel) 14 (2022) 1573.
[48] J.H. Gibbs, E.A. DiMarzio, Nature of the glass transition and the glassy state, J. Chem. Phys. 28 (2004) 373.
[49] T.G. Fox, P.J. Flory, Second-Order Transition Temperatures and Related Properties of Polystyrene. I. Influence of Molecular Weight, J. Appl. Phys. 21 (2004) 581.
[50] J.H. Gibbs, Nature of the Glass Transition in Polymers, J. Chem. Phys. 25 (2004) 185.
[51] C.G. Overberger, L.H. Arond, R.H. Wiley, R.R. Garrett, Monomers containing large alkyl groups. IV. Polymerization and properties of the polymers of 2-alkyl-1,3-butadienes, J. Polym. Sci. 7 (1951) 431–441.
[52] M.L. Dannis, Thermal expansion measurements and transition temperatures, first and second order, J. Appl. Polym. Sci. 1 (1959) 121–126.
[53] J.M.G. Cowie, I.J. McEwen, Glass and sub-glass transitions in methylphenyl and chlorophenyl polyitaconic acid esters, Eur. Polym. J. 18 (1982) 555–558.
[54] A. Gallardo, J.S. Román, Effect of large polar side groups on the glass transition temperature of acrylic copolymers, Macromolecules 26 (2002) 3681–3686.