eman ta zabal zazu

Universidad del País Vasco Euskal Herriko Unibertsitatea

On the dynamics of polymers and biomolecules through the use of machine learning algorithms

Claudia Borredon

SUPERVISED BY

Dr. Gustavo Ariel Schwartz

Dr. Luis Alejandro Miccio

PH.D. THESIS DONOSTIA-SAN SEBASTIÁN, 2023

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Centro de Física de Materiales
Materials Physics Center
Materialen Fisika Zentroa

On the dynamics of polymers and biomolecules through the use of machine learning algorithms

Claudia Borredon

SUPERVISED BY

Dr. Gustavo Ariel Schwartz

Dr. Luis Alejandro Miccio

PH.D. THESIS

DONOSTIA-SAN SEBASTIÁN, 2023

Index

Acknowledgements

Completing this Ph.D. journey has been an incredibly challenging and rewarding experience, and I couldn't have reached this point without the support, encouragement, and guidance of many individuals.

Firstly, I am profoundly grateful to my advisors, Dr. Gustavo Schwartz and Dr. Alejandro Miccio, for their unwavering dedication, insightful feedback, and constant belief in my abilities. Their mentorship has been instrumental in shaping the direction of my research and fostering my growth as a scholar.

Then I acknowledge the University of the Basque Country (UPV/EHU) and the Material Physics Center (MPC, CFM) for the financial support but, moreover, for providing a stimulating environment where to develop my research.

Finally, I extend my gratitude to my friends and family.

I want to thank my friends, both those who have been with me from the beginning and those who have moved on, for sharing memorable times and offering their presence during challenging moments that have unfolded over these three years.

To my family, I want to express my deep appreciation for the unwavering support they have provided since my first year at university, a support that has brought me to this significant juncture.

Thank you all,

Claudia

Resumen

El diseño y desarrollo de nuevos materiales con propiedades mecánicas, químicas o fisicoquímicas específicas a menudo involucra procesos que demandan una considerable inversión de tiempo y dinero, entre otros recursos. Por otra parte, la caracterización de estos materiales se enfrenta a su vez a dificultades relacionadas con el proceso de síntesis, la preparación de muestras o con las condiciones experimentales requeridas para realizar las mediciones. A la hora de superar estos desafíos, el aprendizaje automático y las redes neuronales son herramientas de gran potencial predictivo, especialmente en el estudio de la relación entre la estructura y las propiedades (QSPR, por sus siglas en inglés). En este trabajo se aplican métodos QSPR a la predicción y al estudio de propiedades características de moléculas y polímeros, con foco en la precisión y la interpretabilidad de los resultados.

El aprendizaje automático se refiere a la habilidad de un algoritmo computacional para aprender a partir de un conjunto de datos. Según la conceptualización de Mitchell, un programa de ordenador aprende de la experiencia E en relación con una tarea T y medidas de rendimiento P si su ejecución en las tareas de T, medida por P, mejora con la experiencia E. Aquí T se define como la manera en que el programa debe procesar una entrada dada. La métrica de rendimiento P se establece específicamente para evaluar la precisión del modelo en cumplir la tarea T, generalmente a través de la medición de un error. La experiencia E consiste en los datos proporcionados al algoritmo durante la fase de entrenamiento. En el marco de la presente Tesis, la experiencia está dada por una base de datos de estructuras químicas, la tarea es la predicción de la temperatura de transición vidria y la métrica de rendimiento es representada por el error porcentual sobre esa predicción. A la hora de entrenar, los datos se dividen principalmente en dos conjuntos: entrenamiento y prueba. El conjunto de entrenamiento es empleado en la fase de aprendizaje, y es quien permite que el algoritmo adquiera la información sobre las características de los datos y genere un modelo. Por otra parte, el conjunto de prueba consiste en ejemplos desconocidos para dicho modelo que se utilizan para evaluar la capacidad de generalización del mismo, es decir, su habilidad a la hora de procesar entradas inéditas. De esta manera, el aprendizaje automático implica una tarea de optimización en la que se busca que el rendimiento no solo sea óptimo en el conjunto de entrenamiento, sino también en el conjunto de prueba. El rendimiento del modelo también se ve influido por la cantidad de parámetros con los que se construye, si es supervisado o no supervisado, entre otros. Los algoritmos no supervisados procesan una base de datos cuyas características más relevantes son inferidas de

la estructura del conjunto, mientras que los algoritmos supervisados trabajan con una base de datos acompañada de etiquetas para cada ejemplo. Ejemplos de algoritmos no supervisados son los de agrupamiento o “clustering”, mientras que la regresión lineal es un caso de algoritmo supervisado. En este trabajo, se emplean tanto algoritmos supervisados como no supervisados para desarrollar y validar modelos que permitan estudiar y predecir la temperatura de transición vítrea de un material y para indagar en las propiedades dinámicas de su estructura.

Los algoritmos de aprendizaje automático, y en especial las redes neuronales (ANN, por sus siglas en inglés), han revolucionado el campo de la física de materiales al posibilitar la creación de modelos capaces de capturar relaciones complejas entre las estructuras moleculares y las propiedades físicas. Las ANNs son modelos informáticos inspirados en la estructura y el funcionamiento del cerebro humano. En términos sencillos, constan de capas interconectadas de neuronas artificiales que procesan y transforman los datos de entrada para generar salidas que se ajustan al mapeo no lineal de propiedades complejas. De esta manera, las redes aprenden a extraer características relevantes de dichas representaciones moleculares y a relacionarlas con las propiedades objetivo. Una de las ventajas primordiales de las ANN radica en su habilidad para lidiar con datos multidimensionales y no lineales, lo cual las convierte en herramientas idóneas para nuestro objetivo de capturar relaciones entre estructura y propiedades. Estos modelos tienen la capacidad de procesar y analizar con eficacia volúmenes considerables de información química, acelerando notablemente la concepción y el hallazgo de nuevos materiales con características específicas. Asimismo, el desarrollo orientado a la interpretabilidad de las ANN posibilita la adquisición de conocimiento directo sobre los factores químicos subyacentes que más influyen en las propiedades de interés. En este trabajo, por ejemplo, utilizo metodologías de agrupamiento y de análisis de componentes principales precisamente para entender como el algoritmo está procesando la información de la estructura química para enlazarla con el valor de la temperatura de transición vítrea. Es importante remarcar que la comprensión del algoritmo puede ser utilizada como guía para el diseño de nuevos compuestos con atributos específicos a cada tarea. No obstante, el rendimiento de los modelos QSPR basados en ANN depende de la calidad y representatividad de los datos de entrenamiento. La selección y depuración meticulosa del conjunto de datos son críticas para asegurar pronósticos precisos y fiables. Además, la interpretación de las ANN en el contexto de QSPR sigue siendo objeto de investigación, dado que descifrar las características moleculares específicas y las interacciones que aportan a los pronósticos de propiedades continúa siendo un desafío.

En el ámbito de las propiedades macroscópicas de los materiales, la temperatura de transición vítrea (T_g) aparece como una de las más relevantes, tanto en el ámbito académico como en el industrial. La T_{g} denota la temperatura a la cual un material amorfo transita de un estado fluido a uno rígido y vítreo, motivo por el cual desempeña un papel crucial en la definición de las características de procesamiento de polímeros y otros materiales formadores de vidrio. No obstante, la comprensión de los mecanismos físicos subyacentes al fenómeno de la transición vítrea sigue hoy en día siendo un desafío, dado que este proceso está influido por una variedad de factores como la estructura molecular, la movilidad de las cadenas y las interacciones intermoleculares. Los enfoques tradicionales para estudiar la T_{g} de los materiales a menudo se basan en técnicas experimentales (que dependiendo de la preparación de la muestra y la técnica experimental pueden ser muy laboriosas) o en simulaciones computacionalmente intensivas basadas en primeros principios. En este contexto, los modelos QSPR representan una alternativa eficiente y complementaria en términos de tiempo y costo a los enfoques experimentales, agilizando la detección y el hallazgo de nuevos materiales con un comportamiento en particular. La T_{g} guarda además una estrecha relación con la dinámica de los materiales. En el caso de los polímeros, por ejemplo, su dinámica puede ser explorada a través de técnicas como la espectroscopía dieléctrica de banda ancha (BDS), la reología o los ensayos mecánico dinámicos (DMA). Estos métodos experimentales proporcionan información valiosa sobre el movimiento molecular y los procesos de relajación de distintas porciones de las estructuras moleculares en función de la temperatura. En este sentido, las ANN ofrecen una oportunidad única para capturar las relaciones entre las características estructurales y la T_{g}, brindando de esta manera una perspectiva indirecta de la dinámica de los materiales sin medirla explícitamente. Los descriptores moleculares empleados en los modelos QSPR contienen información implícita del movimiento molecular y la relajación, tales como la existencia de segmentos flexibles, el empaquetamiento molecular o las interacciones intermoleculares. En consecuencia, los valores de T_{g} pronosticados por los modelos QSPR pueden ser empleados como una estimación de la dinámica de los materiales. Por este motivo, en este trabajo se propone la utilización de modelos híbridos que fusionan ANNs con marcos teóricos como la ecuación de Langevin elástica colectiva no lineal (ECNLE, por sus siglas en inglés) para modelar la dinámica de los materiales. De esta manera, se utilizan ANNs para identificar correlaciones complejas y no lineales entre los descriptores moleculares y la T_{g}, la cual a su vez se emplea como valor de entrada para la teoría ECNLE, que estima la dinámica de relajación del material. Al incorporar la teoría ECNLE en el marco del modelado, el modelo híbrido puede brindar predicciones no solo para la T_{g}, sino también para la dinámica del material, como escalas de tiempo de relajación o viscosidad. En este trabajo se emplea el “Simplified Molecular Input-Line Entry System” (SMILES) para representar la estructura molecular de un compuesto mediante una cadena alfanumérica de caracteres. A través de esta representación de la estructura química como entrada, se investiga la modelización de la T_{g} mediante ANN de diferentes arquitecturas, así como también se profundiza en la interpretación de los resultados, la codificación en espacios multidimensionales y su agrupamiento. El trabajo se presenta a través de 3 publicaciones científicas en revistas indexadas internacionalmente. Es importante destacar que se trata de un estudio transversal, que involucra diferentes arquitecturas, propiedades, datasets y herramientas de optimización e interpretabilidad de los resultados, por lo que debe ser considerado en su conjunto antes que como desarrollos independientes.

El primer artículo aborda el desarrollo de un modelo híbrido mediante el uso de una ANN para predecir la T_{g} y un modelo teórico para capturar la dinámica de formadores de vidrio moleculares. En particular, se aplica esta metodología para estimar la dinámica de la relajación α de los compuestos a través de la T_{g} y la teoría ECNLE. Para ello se emplea una arquitectura de red neuronal “fully connected” y una codificación estilo “one hot encoding” de las cadenas alfanuméricas obtenidas a través del SMILES, que permiten predicciones de la T_{g} con errores porcentuales promedio inferior al 8%. Este resultado es especialmente destacable dado que en muchos casos la naturaleza cinética de la medida experimental de T_{g} no permite establecer un único valor en la literatura, sino más bien un rango de temperaturas que depende de la velocidad de calentamiento-enfriamiento (entre otros factores), y por ende las incertezas no pueden reducirse con facilidad más allá de ese punto. La concordancia entre las predicciones y los resultados experimentales es notable y demuestra la validez de este enfoque híbrido para realizar inferencias sobre los materiales a partir de solo una representación de su estructura química. Además, este enfoque se puede emplear para comprender cómo las variaciones en la estructura molecular inducen cambios en la estimación de la T_{g}.

El segundo artículo, es similar en términos de la predicción de la dinámica a partir de la T_{g}, pero se enfoca en la implementación de redes neuronales convolucionales aplicada a una familia de polímeros: los poliacrilatos atácticos. Haciendo uso de la capacidad de las CNN para detectar patrones en las estructuras químicas, se obtienen estimaciones de T_{g} que después se emplean como entradas para el modelo ECNLE. Con el fin de entrenar el modelo de CNN, se codifican los monómeros como matrices derivadas del SMILES. Es importante mencionar que a pesar de trabajar sobre cadenas de polímeros (y no sobre moléculas individuales), esta metodología produce errores porcentuales promedio en las predicciones de menos del 9%, lo que constituye un logro considerable ya que la red es entrenada únicamente a partir de la estructura del monómero, sin añadir ningún tipo de información física adicional. Posteriormente, se integran estos resultados con la teoría ECNLE para obtener estimaciones sobre la dinámica de los polímeros. Esta modalidad híbrida que aprovecha las CNN podría abrir nuevos caminos en la creación de materiales poliméricos, permitiendo una aproximación significativa a la dinámica de los compuestos exclusivamente a partir de la estructura del monómero.

En el tercer artículo, se estudia el proceso por el cual las redes neuronales recurrentes pueden modelar la física detrás del proceso de transición vítrea. En esta instancia, el SMILES se codifica con una codificación cardinal y se emplean neuronas bidireccionales de memoria a largo plazo (BiLSTM por sus siglas en ingles). Estas neuronas son especialmente ventajosas ya que analizan la secuencia proporcionada tanto de izquierda a derecha como de derecha a izquierda, facilitando la identificación de patrones significativos en la misma. El error porcentual promedio en este caso es inferior al 9%. Luego, se demuestra mediante el Análisis de Componentes Principales (PCA) que la red es capaz de reconocer y seguir características en la estructura química que influyen en el valor de la T_g. Se aplica el algoritmo de clusterización Fuzzy-C a la última capa oculta de la red para evaluar su capacidad de distinguir entre diversas estructuras químicas. Finalmente, se emplea la red neuronal para predecir los valores de T_g de aminoácidos esenciales y un péptido corto (3-lisina), la mayoría de gran dificultad para su medida. En el caso de aquellos con valores experimentales, se constata que los aminoácidos que se encuentran más cercanos al intervalo de confianza de la red, son efectivamente predichos con mayor precisión que aquellos que están más alejados del intervalo. De esta manera, se concluye que es viable emplear las ANN como un laboratorio virtual para explorar el impacto de la estructura molecular en la T_g.

Section 1

1.1 Introduction

The design and the development of new materials with desired physico-chemical properties often involves time-consuming and resource-intensive processes. In addition, the characterization of these materials has to deal with the complications involved in the synthesis process. To overcome these difficulties, machine learning and neural networks have emerged as powerful tools within the framework of quantitative structure-property relationship (QSPR) models [1–3]. Machine learning algorithms, particularly artificial neural networks (ANNs), have revolutionized the field by enabling the development of data-driven models that can capture complex relationships between molecular structures and properties[4–11].

ANNs are computational models inspired by the structure and functioning of the human brain[12]. In simple terms, ANNs consist of interconnected layers of artificial neurons that process and transform input data to produce the desired outputs. In the context of QSPR, ANNs are trained on large datasets of molecular structures and their corresponding properties[4,13,14]. The network learns how to extract relevant features from the molecular descriptors and map them to the target property. The advantage of ANNs lies in their ability to handle high-dimensional and non-linear data, making them suitable for capturing intricate structure-property relationships. By utilizing hidden layers with nonlinear activation functions, ANNs can model complex dependencies that may exist between molecular features and properties. Through an iterative process called training, the network adjusts its internal parameters to minimize the difference between predicted and actual property values. This optimization enables the network to generalize its learning and make accurate predictions on new, unseen data. Furthermore, the integration of machine learning techniques, such as ANNs, has facilitated the exploration and analysis of large chemical databases. These models can efficiently process and analyse vast amounts of chemical data, significantly accelerating the design and discovery of new materials with specific properties. Additionally, the interpretability of ANNs allows researchers to gain insights into the underlying factors influencing the properties of interest. This understanding can guide the rational design of new compounds with desired properties. However, it is important to note that the performance of ANN-based QSPR models is dependent on the quality and representativeness of the training data[15,16]. Careful selection and curation of the dataset are crucial to ensure accurate and reliable predictions. Moreover, the interpretability of ANNs in the context of QSPR is an ongoing research area, as understanding the specific molecular features and interactions that contribute to property predictions remains a challenge.

Among the numerous macroscopic properties of materials, the glass transition temperature (T_g) holds significant academic and industrial relevance[17–21]. T_g represents the temperature at which an amorphous material changes from a fluid state into a rigid, glassy state. It plays a crucial role in determining the mechanical, thermal, and processing characteristics of polymers and other glass forming materials. Despite its importance, understanding the physical mechanisms underlying the glass transition phenomenon remains challenging. It is a complex and dynamic process influenced by various factors, including molecular structure, chain mobility and intermolecular interactions [22,23]. Traditional approaches to comprehend T_g often rely on laborious experimental techniques[24,25] or computationally demanding simulations based on first principles[26,27]. In this regard, QSPR models offer a valuable and efficient approach to explore the associations between the molecular structure of a compound and its corresponding T_g value. By using machine learning algorithms and ANNs, these models can uncover complex relationships between the structural features of a molecule and its glass transition dynamics[6,7,10,28,29]. These models have the potential to guide the design and development of materials with tailored T_g values, enabling the optimization of desired properties for specific applications. Furthermore, QSPR models offer a more time and cost-efficient alternative to experimental approaches, accelerating the screening and discovery of new materials with targeted glass transition behaviour.

In addition to its relevance for material properties, the T_g is closely tied to the dynamics of materials. The dynamics of polymers, for instance, can be probed through techniques such as broadband dielectric spectroscopy (BDS) or differential scanning calorimetry (DSC). These experimental methods provide valuable insights into the molecular motion and relaxation processes as a function of the temperature. ANN models offer a unique opportunity to bridge the gap between molecular structure, T_g and material dynamics by capturing the relationships between structural features and T_g, therefore providing indirect information about the dynamics of materials without explicitly measuring them. The molecular descriptors used in QSPR models can capture important aspects of molecular motion and relaxation, such as the presence of flexible segments, molecular packing, or intermolecular interactions. As a result, the predicted T_g values from QSPR models can serve as approximation for the dynamics of materials. In this sense, there has been a growing interest in developing hybrid models that combine ANNs with theoretical frameworks such as the Elastically Collective Non-linear Langevin Equation (ECNLE) theory [30] to model the dynamics of materials. In such hybrid models, ANNs are employed to capture the complex and non-linear relationships between molecular descriptors and the glass transition temperature. The T_g, in turn, serves as an input to the ECNLE theory, which calculates the relaxation dynamics of the material. By incorporating the ECNLE theory into the modeling framework, the hybrid model can provide predictions not only for T_g but also for material dynamics, such as relaxation timescales or viscosity.

In summary, the combination of machine learning, ANNs, and experimental techniques for material dynamics allows for a comprehensive understanding of the relationship between molecular structure, glass transition temperature, and the dynamic behaviour of materials. This integrated approach holds great potential for accelerating materials design and optimization by providing valuable insights into the complex interplay between molecular structure, T_g, and material dynamics.

The studies presented in this work employ ANNs to predict the T_g of molecular glass-formers and polymers. The Simplified Molecular Input Line Entry System (SMILES) molecular descriptor is used, representing the molecular structure of a compound through an alphanumeric string of characters. By using this chemical structure representation as input, I investigate the modelling of T_g behavior using three different ANN architectures. On the one hand, I explore the dynamics of the alpha relaxation process in molecular glass formers and polymers, by using a hybrid combination of fully connected and convolutional neural networks with disordered systems theory (ECNLE). These hybrid models successfully predict the dynamics associated with the alpha relaxation process. In addition, a recurrent neural network is employed to predict the glass transition of molecular glass-formers. While this architecture is well-suited for capturing temporal dependencies and sequential patterns, providing an effective tool for modelling the

T_g behaviour in these systems, I apply machine learning techniques to ensure transparency and interpretability of the resulting model. These techniques facilitate the understanding of how the ANNs arrive at their predictions, enabling researchers to gain insights into the factors driving the glass transition phenomenon.

1.2 Methods

This section describes the main theoretical background behind this Thesis. First, a definition of what is machine learning and the different types of artificial neural network is provided, then a general description of the glass transition process and the different experimental techniques which measure the glass transition temperature is presented.

1.2.1 Machine learning

Machine learning is the field which deals with the capability of a computational algorithm to learn from a set of data. A conceptualization of what learning is for a machine can be found in Mitchell[31] : "A computer program is said to learn from experience E with respect to some class of tasks T and performance measures P, if its performance at tasks in T, as measured by P, improves with experience E". A task T is defined by how a program should process a given input. For example, for a regression task the algorithm should output a function $f: R^n \to R$. The performance P consists in a metric defined specifically to evaluate the accuracy of the model in accomplishing the task T, like the measuring of an error. The experience E lies in the variety and amount of data that the algorithm is provided with during the training phase. The data is collected in a dataset from which I define mainly two sets: the training and the test set. The first set is the one which is used in the training phase of the model, as to say the one from which the algorithm learns the features of the data, while the second one is a set of unseen examples with which we can measure the so-called generalization power of the algorithm, as to say the capability of the algorithm to process an unknown input. We assume that the examples in each set are independent and identically distributed, and that the probability distribution with which the training and test set are generated coincide. In this way, it is possible to make assumptions on the relationship between the training set error and the test set error. The main aim of machine learning is to correctly perform a task T on the test set of examples, evaluating it with a performance P. So, machine learning is a special type of optimization task because we do not want only the performance on the training set to be the best possible, but, also, we want it to be good on the test set. The performance of an algorithm depends also on the number of parameters it is built with. According to this number there could be two types of limit behaviour of the algorithm: underfitting or overfitting. It is underfitting when the parameters of the model are not sufficient to grasp the features of the dataset, while it is overfitting when the model is able to predict the examples of the training set but is incapable of generalization (the performance on the test set is scarce). Another factor which describes an algorithm is whether it is supervised or unsupervised. An unsupervised algorithm typically is fed with a dataset of features and the algorithm infers information from the structure of the dataset; on the other hand, a supervised algorithm is fed with a dataset provided with a label for each example in it. Clusterization is a typical example of unsupervised algorithm, while linear regression is an example of supervised algorithm. In this work both unsupervised and supervised algorithms are used in order to find a model which predicts the glass transition temperature of a given material and to investigate what is happening inside the network.

1.2.2 Artificial neural networks

Artificial neural networks (ANNs) are a field of machine learning consisting in a non-linear mapping of information between an input and a given output. ANNs find a lot of applications, from pattern or face recognition, to data processing and natural language processing. In physics, ANNs can help understanding the complicated correlations of cause-effect of a given phenomenon without the use of first principles. It is called "neural" network because the parts which make it are thought to resemble the behaviour of a biological neuron. The biological neuron functions can be modelled by considering the dendrites (which represent the input) and the axons (which represent the output). When an electric signal is perceived by the dendrites, the axon responds with a signal according to a given threshold.

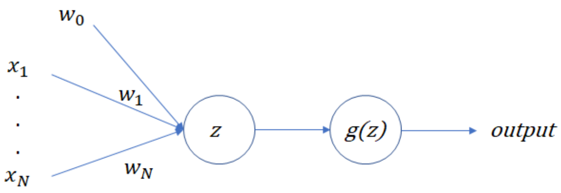


Figure 1.2.1 A representation of the McCulloch-Pitts neuron. Adapted from [12].

A simple model of a single neuron was introduced by McCulloch and Pitts in 1943[32]. We can use a computational graph to have a scheme of the model (Figure 1.2.1). In these graphs, every node is a variable (scalar, vector, tensor …) and every edge is a function operating on the variable to give an output. Given an n-dimensional array *x*, it is multiplied by a set of parameters *w* called weights as shown in Equation 1.

$$z = \sum_{i=1}^{n} x_i w_i + w_0 \tag{1}$$

The element w_0 is called bias and is an offset parameter which simulates the threshold of the neuron. The output of the neuron o is given by applying a function g over the results of z. This function g is called activation function and can be linear or non-linear. For the McCulloch Pitts model the step function was used as activation function. The most used activation function is the sigmoidal, but hyperbolic tangent and linear activation functions are used as well.

An ANN has to be provided with a dataset of examples {*x*}, which is typically divided in three: a training set, an internal validation set and a test set. The training set is the set of examples from which the ANN learns the features of the input, the validation set is an internal set with which it is possible to monitor the generalization power of the network during the training phase, and the test set is a set of examples which were not used during the training phase and is used to test the generalization power of the network once it is trained. The training typically consists in the minimization of a loss function (see Equation 3), and this usually is achieved by minimizing the gradient of such function.

1.2.2.1 Backpropagation & gradient descent

When training an ANN, the information of the input flows from the first layers of the network to the last one generating an output. This part of the algorithm is usually addressed as feedforward propagation. The output $o(w)$, is then compared with the expected value y by means of a loss function, which calculates the error of the prediction of the network with respect to the expected value. The aim of the algorithm is to lower the value of such error by adjusting its free parameters (the weights) to minimize the loss function. To do so, the algorithm must calculate the gradient of the loss function taking into account all the parameters of the model. The backpropagation algorithm is a tool with which is possible to calculate the gradient of the loss function of the ANN. It is based to the chain rule of calculus, which states for a function y = g(x) and a function z = f(g(x)) = f(y):

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx} \tag{2}$$

So, given the loss function L:

$$L = L(o(w), y) \tag{3}$$

we can imagine it as a m-dimensional surface characterized by a certain number of minima (like in Figure 1.2.2), where m is the number of parameters of the model. It depends on the parameters w by the quantity expressed in Equation 1, to which is applied the activation function $g(z)$, so that the derivative of L with respect to a parameter w_i is

$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial g(z)} * \frac{\partial g(z)}{\partial z} * \frac{\partial z}{\partial w_i} \tag{4}$$

This derivative can be now used to update and optimize the weights w of the model. One of the simplest approaches to update the weights is the gradient descent algorithm, for which

$$w_i := w_i - \alpha \frac{\partial L}{\partial w_i} \tag{5}$$

where α is called learning rate. This operation is made calculating the gradient over all the set of examples x, but it can be also calculated by dividing the set in minibatches and updating the weights one batch at a time.

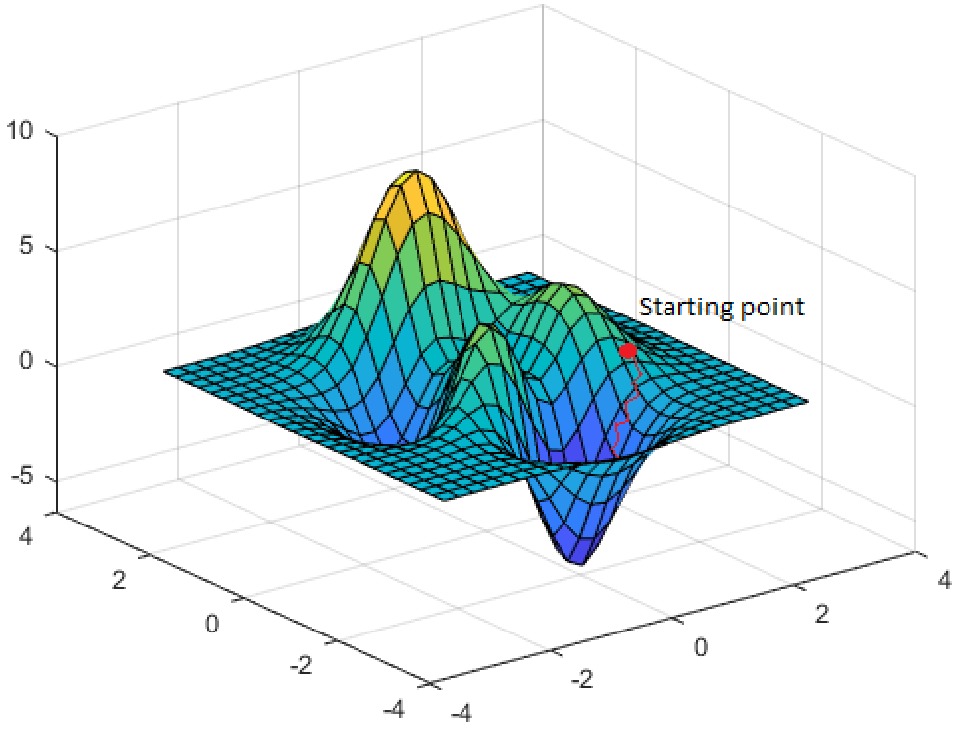


Figure 1.2.2 A scheme representing the process of the gradient descent. The point follows the gradient of the m-dimensional surface according to the direction pointing at the minimum.

This is usually more efficient and less time-consuming than the gradient descent. If the batch is made of 1 example, then it is called stochastic gradient descent, as the algorithm moves stochastically on the surface of the loss function trying to minimize the gradient. Actually, optimization algorithms have been developed in order to enhance the search of the minimum. Among these, we used the adaptive moment estimation (Adam) optimizer, a standard optimization method which combines the advantages of two other extensions of gradient descent: the momentum and the root mean square propagation (RMSProp). The momentum algorithm consists in considering the parameters w_i as points moving on a surface with a given velocity v_i so that the update to the weights is:

$$v_i \coloneqq \beta v_i + (1-\beta) dw_i \tag{6.1}$$

$$w_i \coloneqq w_i - \alpha v_i \tag{6.2}$$

where $dw_i = \frac{\partial L}{\partial w_i}$. The RMSProp update is given by:

$$s_i \coloneqq \beta s_i + (1-\beta) dw_i^2 \tag{7.1}$$

$$w_i \coloneqq w_i - \alpha \frac{dw_i}{\sqrt{s_i}} \tag{7.2}$$

and, finally, the Adam optimization algorithm is given by:

$$v_i \coloneqq \beta_1 v_i + (1 - \beta_1) dw_i \tag{8.1}$$

$$s_i \coloneqq \beta_2 s_i + (1 - \beta_2) dw_i^2 \tag{8.2}$$

$$v_i^{corr} \coloneqq \frac{v_i}{1 - (\beta_1)^t} \tag{8.3}$$

$$s_i^{corr} \coloneqq \frac{s_i}{1 - (\beta_2)^t} \tag{8.4}$$

$$w_i \coloneqq w_i - \alpha \frac{v_i^{corr}}{\sqrt{s_i^{corr}} + \varepsilon} \tag{8.5}$$

where v_i^{corr} and s_i^{corr} are bias corrections to v_i and s_i. This algorithm is one of the most used nowadays because it is a fast, computational efficient method to search the minimum of complex functions as the loss function.

1.2.2.2 Fully connected

The fully connected neural network (or multilayer perceptron) is the simplest architecture. It consists in a sequence of layers made of neurons in which each neuron from the previous layer is connected to each neuron in the following one.

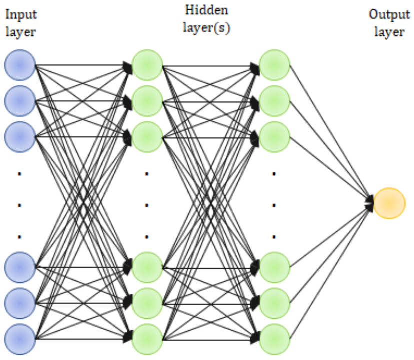


Figure 1.2.3 Fully connected architecture. It is composed by a first input layer (blue), two hidden layers (green) and one output layer (yellow). Adapted from [33].

They are also called "feedforward" networks as the information flows in only one direction. The computational graph of the fully connected neural network is shown in Figure 1.2.3. This Figure shows the architecture of a multilayer perceptron, with one input layer (blue), followed by 2 hidden layers (green) and ending in a single neuron output layer (yellow).

1.2.2.3 Convolutional neural network

Convolutional neural networks (CNNs) are networks which use convolution in place of general matrix multiplication in at least one of their layers. They are designed to work mainly with grid-like data. Unlike the fully connected neural network, the convolutional is characterised by sparse interaction, meaning that not every output unit interacts with the input units. This happens because the convolutional layer is based on the multiplication of the input with a filter (or kernel), which is smaller than the input and spans different areas of the input, generally top to bottom and left to right.

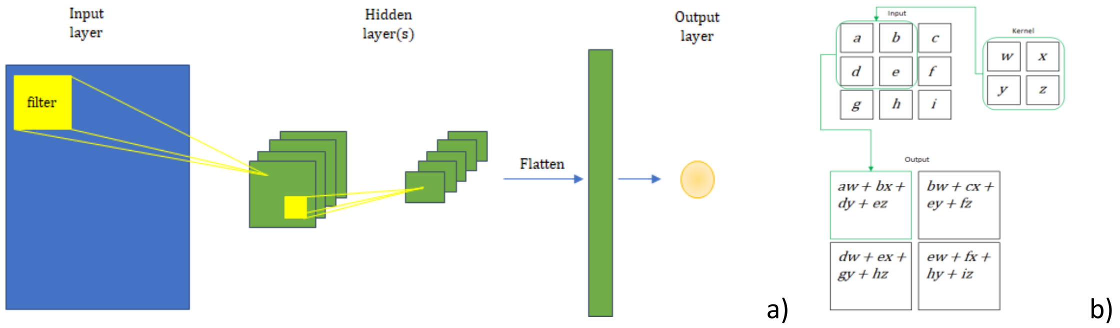


Figure 1.2.4 a) scheme of a CNN. It is composed by an image-input layer, to which a series of filters are applied; in the end, the output is flattened and the result is fed to an output layer. Adapted from[6]; b) example of filter application, adapted from[33].

The filters applied can be more than one, and this influences the depth of the convolutional layer. One of the key advantages of CNNs is the parameter sharing, meaning that the algorithm learns one local set of parameters from the filter which are able to find important features spanning the whole input. This property enables the network to capture important features that are spatially invariant. For example, if a certain filter detects a vertical edge, it can recognize that pattern regardless of its position in the image. Parameter sharing significantly reduces the number of parameters required compared to fully connected neural networks, where each parameter corresponds to a connection between individual neurons.

In a typical CNN architecture, the convolutional layers are often followed by other types of layers, such as pooling layers and fully connected layers. Pooling layers reduce the spatial dimensions of the feature maps, reducing computational complexity and providing some degree of translational invariance. Fully connected layers are responsible for making predictions based on the extracted features and are commonly found at the end of the network. These layers combine the learned features from the previous layers and map them to the desired output, such as class probabilities in image classification. An example of a convolutional neural network is shown in Figure 1.2.4a, where an image (input layer, blue) is spanned by 4 filters and then by 5 filters (hidden layers, green), then the result is flattened and sent to the output layer (yellow). Figure 1.2.4b shows an example of the application of the convolutional filter to the input.

1.2.2.4 Recurrent Neural network

The Recurrent Neural Network (RNN) is an architecture designed to deal with data expressed as sequences like time series, sentences or speech. It is characterized by the fact that it can learn patterns from parts of the sequence by keeping memory of it in a hidden state s which is updated according to the learnt parameters. Similarly to the convolutional neural network, it is characterized by the property of parameter sharing, as to say that the weights locally learnt during the training are used on every time step of the sequence.

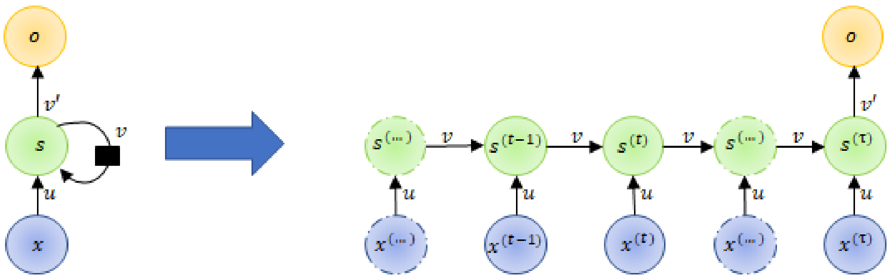


Figure 1.2.5 Recurrent neural network. On the right part of the arrow, the unfolded version shows how the input sequence (blue) is fed to the recurrent hidden layer (green) and then outputs a single value (yellow). Adapted from [33].

At each time step, the current input is transformed by using a set of learnable weights and biases. This transformation produces an intermediate representation, often referred to as the hidden state or hidden activation. The hidden state captures information about the current input and incorporates information from previous time steps through the recurrent connections. The process can be expressed in terms of the following function:

$$s^{<t>} = f(s^{<t-1>}, x^{<t>};\ W) \tag{9}$$

where $s^{<t>}$ is the state of the system at timestep t, $x^{<t>}$ is the input at timestep t and W are the learnt parameters. In this case, W represents both the weights which connect a state to the other (v), the weights which connect the state to the input $x^{(t)}$ (u) and the weights that connect the final state to the output (v'). The ";" inside the function indicates that the terms on the left are timestep dependent, while the term on the right is shared along all the timesteps. It is possible to unfold the recurrence of the expression in Equation (9) for a finite number of timesteps τ in order to get an expression which does not involve recurrence. For example, for $\tau\ =\ 3$ we obtain:

$$s^{<3>} = f(s^{<2>}, x^{<3>};\ W)\ = \tag{10.1}$$
$$= f(f(s^{<1>}, x^{<2>}; W), x^{<3>};\ W) \tag{10.2}$$

With this expression it is now possible to build the computational graph of the recurrent neural network in an acyclic way (Figure 1.2.5). In this graph the sequence input (blue), the hidden states (green) and the output state (yellow) are depicted. The output can be produced at each time step or only at the final time step, depending on the specific task (in our case we focus on the final step). One of the most used units of recurrent neural network is the long short-term memory (LSTM) unit. The state s of this unit accounts for its activation a and a cell flag c which is the kernel of the memory process. The value of the activation is influenced by the value of the cell flag, and the value of the cell flag is decided according to the value of the so-called gates, as shown in Figure 1.2.6. The LSTM has 3 types of gates (forget, update and output gate) which are defined as follows:

$$\Gamma_f = \sigma(w_f[a^{<t-1>},x^{<t>}] + b_f) \tag{11.1}$$

$$\Gamma_u = \sigma(w_u[a^{<t-1>},x^{<t>}] + b_u) \tag{11.2}$$

$$\Gamma_o = \sigma(w_o[a^{<t-1>},x^{<t>}] + b_o) \tag{11.3}$$

where σ is the sigmoid activation function, w and b are respectively the weights and the bias vectors, a is the activation and x is the input value. On each iteration a candidate for the cell value $\hat{c}$ is calculated as follows:

$$\hat{c}^{<t>} = tanh(w_c[a^{<t-1>},x^{<t>}] + b_c) \tag{12}$$

and combined with the update gate and the forget gate and the value of the cell at time *t-1* to compute the new value of the cell as follows:

$$c^{<t>} = \Gamma_f c^{<t-1>} + \Gamma_u \hat{c}^{<t>} \tag{13}$$

Finally, the new value for the activation is calculated using the value of the cell at time *t* and the value of the output gate as follows:

$$a^{<t>} = \Gamma_0 tanh(c^{<t>}) \tag{14}$$

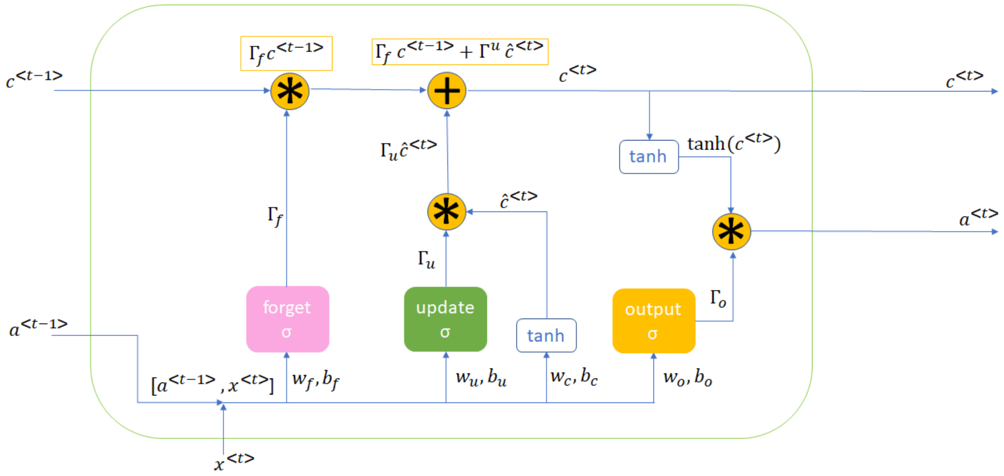


Figure 1.2.6 Scheme of the Long Short-Term Memory unit

This peculiar state is what allows the network to register and recognize patterns along a sequence. In the third paper, I used a Bidirectional-LSTM approach, which reads the SMILES sequences of different molecular structures both from the beginning to the end and from the end to the beginning.

1.2.3 Principal Component Analysis

The information in ANNs is codified in high dimensionality spaces which can be manipulated with mathematical tools, but then there is the necessity to visualize the results in a lower dimensional space. The Principal Component Analysis (PCA)[34] is a dimensionality reduction technique. The concept behind this is that the high dimensionality of the data might be redundant and it can be explained by an intrinsic lower-dimensional structure emerging from linear combinations of the dimensionalities of the original space.

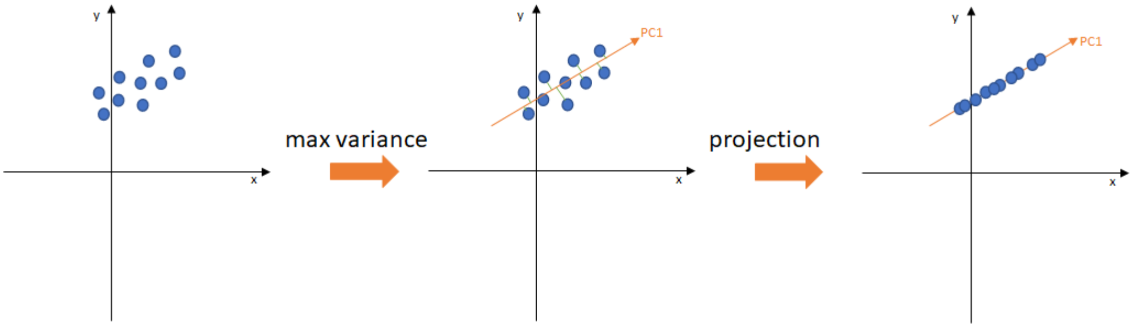


Figure 1.2.7 Principal Component Analysis. The data are projected on the direction of maximum variance, called Principal Component 1. Adapted from [34]

The PCA is an orthonormal transformation which projects the data of the sample in a space in which the variance of the sample is maximised (see Figure 1.2.7). In this way it is possible to analyse and manipulate the high dimensionality data in the original space (for example with a clustering) and then easily interpret it by embedding it in a low-dimensionality human-readable space. I used the PCA in order to interpret the output of the intermediate layers of the ANN, to verify that the network would in fact get the physics of the glass transition process. To do so, I extracted the activations of the last hidden layer of the ANN, as it is supposed to contain information connecting the molecular structure to the T_g. In this way it is possible to embed the molecular structure of the glass formers (as encoded by this layer) into an m-dimensional T_g-oriented space, where m corresponds to the number of neurons of the last hidden layer (in this case $m = 16$). Having this mathematical representation of the molecular structure allows to easily make operations on it. However, the high-dimensionality of the space does not allow an immediate human-readable output. For this reason, the PCA is used to reduce the dimensionality of the space from 16 to 3 or 2 dimensions, allowing a clear reading of the mathematical representation of the data.

1.2.4 Clustering

Clustering is a technique with which a set of N data is divided into C groups called "clusters"[34]. Given a numerical representation of the data, like molecular descriptors, the division among the clusters is due to similarity between points of the dataset, so that molecules which share some similarity are grouped together. Classical clustering (like for example the K-means[35], see Figure 1.2.8) works by defining a number of centroids on the data and calculating the distance of each point from every centroid.

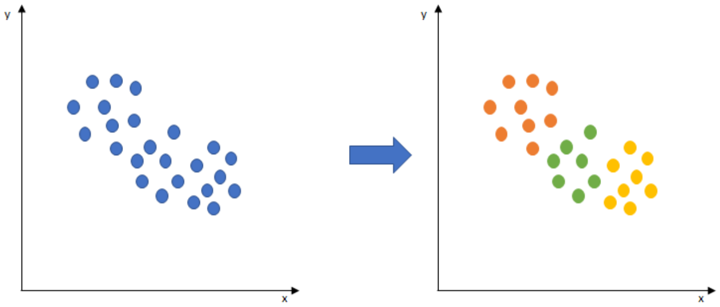
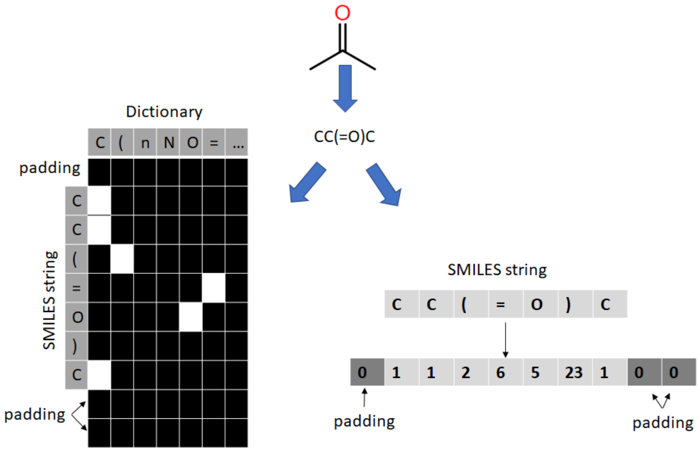


Figure 1.2.8 A schematic example of classical clustering.

A variant of the classical clustering is the fuzzy clustering (fuzzy C)[36], in which the points of the dataset are assigned to a cluster with a probability to belong to it. This makes the clustering more flexible, as the point is not forced to belong to a single cluster, but it can be shared among two or more clusters.

1.2.5 Chemical structure of molecules and polymers

The chemical structure representation of a given molecule reflects the arrangement and connectivity of atoms within it, providing valuable insights into its composition and reactivity. At the core of this representation there are the atoms, which are the building blocks of molecules. Each atom is denoted by its elemental symbol (e.g., C for carbon, O for oxygen) and the arrangement of atoms in a molecule is determined by the bonds that connect them. The most common type of bond is the covalent bond, where atoms share electrons to achieve a stable electron configuration. Covalent bonds can be single, double, or triple bonds, indicating the number of electron pairs shared between atoms. In addition to covalent bonds, other types of bonds, such as ionic bonds and metallic bonds, may exist in certain molecular systems. The connectivity of atoms in a molecule is often depicted using graphical representations, such as molecular diagrams or line structures. In molecular diagrams, atoms are represented by their elemental symbols, and bonds are depicted as lines connecting the atoms. This visual representation provides a clear understanding of the spatial arrangement and connectivity of atoms within the molecule. In the case of polymers, the chemical structure involves the repetition of molecular units known as monomers. Since the chemical structure of molecules and polymers encodes relevant information about their properties, the main idea is that the representation of such structures can be used as an input for an ANN to predict a given property. The non-linearity mappings of the ANN can indeed grasp the correlation between the molecular structure and the property to predict. The simplified molecular input line entry system (SMILES)[37] is a way to codify the chemical structure of a molecule with alpha-numeric strings of characters.



a) One-hot encoding

b) Cardinal encoding

Figure 1.2.9 SMILES strings and encodings. a) the one hot encoding represents the sequence as a sparse matrix of 0s and 1s; b) the cardinal encoding represents the sequence as a string of numbers. In both cases 0-padding was added so that all the sequences had the same length.

In order to use this chemical information in the model, it is necessary to convert the alpha-numerical string in a full numerical input, and this can be done with a variety of strategies. In particular, in this Thesis I opted for the one-hot and cardinal encoding. In both cases, I defined a dictionary of characters contained in the strings of the dataset. Then, the dictionary is used in the one-hot encoding to define a matrix of 0s and 1s, where the 1s correspond to the position of a letter inside a string of characters. In the cardinal encoding I assigned a number to each character according to its position in the dictionary to then convert the strings of characters into strings of numbers. Also, padding of 0s were added in order to have the same length for all the sequences. These two procedures are shown in Figure 1.2.9.

1.2.6 Glass transition temperature and relaxation dynamics

1.2.6.1 Glass transition temperature

In a liquid, the relative position of a molecule with respect to its neighbours changes according to a relaxation time determined by the influence of the intermolecular forces on the random thermal motion of the molecules[38]. Usually this behaviour is modelled as a molecule which jumps out of a cage of neighbours with a characteristic time τ. This time can be approximate knowing that the molecule vibrates in the cage with a given frequency ν and has to overcome an energy barrier ε in order to escape the cage. The probability of such event is given by the Boltzmann distribution, so that the estimation of the characteristic escape time is:

$$\tau^{-1} \approx \nu \; exp(-\varepsilon/kT) \qquad (15)$$

For a simple liquid the characteristic time is about $(10^{-12} - 10^{-10})s$. Another key parameter describing the behaviour of a fluid is the viscosity η, which depends on the relaxation time in the form:

$$\eta = G_0 * \tau \qquad (16)$$

where G_0 is an instantaneous modulus which characterises the elastic response of the material at times shorter than the relaxation time. By putting the previous estimation of τ we obtain the Arrhenius behaviour

$$\eta = (G_0/\nu) * exp(\varepsilon/k_B T) \qquad (17)$$

which is the typical behaviour of viscous liquids at high temperature. When lowering the temperature, assuming that the liquid does not crystallise, the material undergoes glass transition. The glass transition is a kinetic phenomenon concerning the solidification of a super-cooled liquid into an amorphous solid. It is not considered as a phase transition as its features depend on the thermal history of the material, characterized, for example, by the cooling rate of the process and, on the other hand, it does not show any discontinuity in the first derivative of the free energy. The temperature dependence of the viscosity is given by the empirical formula of Vogel-Fulcher-Tammann:

$$\eta = \eta_0 \ * \ exp(B/\ (T - T_0)) \tag{18}$$

The glass transition temperature is defined by convention when the viscosity of the liquid reaches $10^{12}\ Pa * s$, or when its relaxation time reaches $10^2 s$, as to say that the relaxation time becomes comparable to that of the experiment. A way to visualize the glass transition is through the behaviour of the volume of the substance or through its heat capacity at constant pressure. From the volume point of view, the cooling of a liquid would result in a change of slope in the graph volume vs temperature, given that the cooling rate is fast enough to avoid the crystallization of the sample. This process is displayed in Figure 1.2.10, where two T_g are shown, corresponding to a faster ($T_{g,1}$) and a slower ($T_{g,2}$) cooling rate, along with the crystallization temperature T_c.

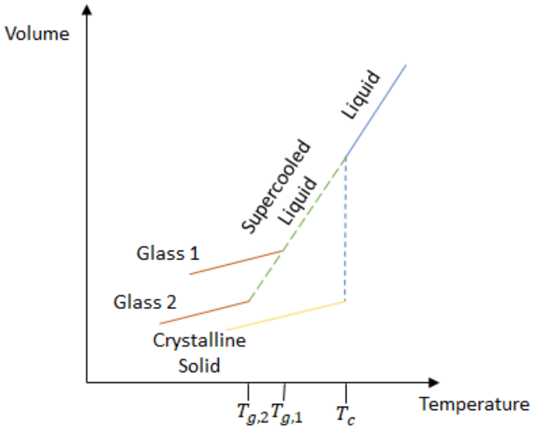


Figure 1.2.10 Volume behaviour in the case of glass transition and crystallisation. The $T_{g,1}$ corresponds to a faster cooling rate, while $T_{g,2}$ corresponds to a slower cooling rate. T_c is the crystallisation temperature.

Experimentally, the glass transition temperature is usually measured with calorimetry (see section 1.2.7.1), by following the changes in heat capacity at constant pressure. The knowledge of the glass transition temperature is of most interest as it is involved in relevant processes like food processing, pharmaceutical stocking and polymer characterization. In the developing phase of a new material, given the difficulties that might arise in the synthesis processes, it is interesting to study new approaches to have at least an estimation of the T_g starting from the chemical structure of the compound. This can also give an insight on how the structure of a molecule can influence the value of the T_g, suggesting new routes to material property tuning.

1.2.6.2 Material Dynamics

The understanding of relaxation dynamics plays a crucial role in the behaviour and performance of materials. For instance, in the pharmaceutical industry it is important to obtain materials of high purity and whose making process is reproducible in terms of physical, chemical and biological properties. In this framework, amorphous materials play a crucial role as this character is common in polymers used as excipients, peptides and proteins and small organic and inorganic molecules. The relaxation processes are key to understanding the dynamics and properties of these materials.

When a material is subjected to an external perturbation, it undergoes a response that reflects its molecular-chain dynamics. Under the assumption of linear response theory, the observed response is directly proportional to the perturbation. This theory allows to apply the fluctuation-dissipation theorem, which states that the response of a system at thermodynamic equilibrium to a small applied disturbance is equivalent to its response to a spontaneous fluctuation. Therefore, by examining the response over a range of frequencies, the equivalence in time-temperature concept can be employed. This concept enables the exploration of a wide range of molecular dynamics by mapping the effects of temperature changes onto changes in the time scale.

Material dynamics are typically characterized by a specific relaxation time and its dependence on temperature. The broad range of relaxation times observed experimentally, spanning several decades, necessitates the use of multiple techniques such as broadband dielectric spectroscopy (BDS), dynamic mechanical analysis (DMA) or any other frequency sensitive technique. The knowledge of T_g in the biological and pharmaceutical framework is critical to anticipate the spontaneous changes in the properties of the solid during storage and/or handling of the material. The development phase of a new material can be both costly and time-consuming, hence the aim of this Thesis is to propose different strategies to estimate the value of the T_g starting from the knowledge of the molecular structure of the glass former.

1.2.6.3 Elastically Collective Non-Linear Langevin Equation

The Elastically Collective Non-linear Langevin Equation (ECNLE) theory [30] describes the molecular dynamics of glass-formers with a hard-sphere based model. It considers: 1) the interactions of a tagged particle with its neighbours and 2) the cooperative motions of particles beyond the first neighbours shell. The motion of a single particle in its cage of first neighbours is described by a dynamic free energy $F_{dyn}(r) = F_{ideal}(r) \ - F_{caging}(r)$, where $F_{ideal}(r)$ is the ideal fluid state, while $F_{caging}(r)$ characterises the localised state of the particle. The dynamical constraint of the cage is given by the emergence of a barrier in $F_{dyn}(r)$ and, defining r_L as the localization length and r_B as the barrier position, we get $F_B = F_{dyn}(r_L) - F_{dyn}(r_B)$, which considers the energy to overcome the cage barrier. Then, the collective rearrangement of the outer molecules of the material is accounted by a displacement field $u(r)$ to which is associated the elastic free energy:

$$F_e \ = \int_{r_{cage}}^{\infty} 4\pi\rho r^2 g(r) K_0 u^2(r)/2dr \tag{19}$$

where ρ is the number of particles per volume, g(r) is the radial distribution function and $K_0 = \frac{\partial F_{dyn}(r)}{\partial r}|_{r\,=\,r_L}$. Using Kramer's theory we obtain:

$$\frac{\tau_\alpha}{\tau_s} = 1 + \frac{2\pi}{\sqrt{K_B K_0}} \frac{k_B T}{d^2} \, exp(\frac{F_B + aF_e}{k_B T}) \tag{20}$$

where $K_B = \frac{\partial F_{dyn}(r)}{\partial r}|_{r = r_B}$, τ_s is the experimental time, d is the diameter of the sphere and a is an adjustable parameter which improves the modelling of the external elastic field. The glass transition temperature enters in this model due to a density-to-temperature conversion, as the equation above gives the τ_α as a function of the volume fraction $\Phi = \rho\pi d^3/6$, where ρ is the number of particles per volume. The conversion is based on a thermal expansion process and it gives the thermal mapping $T = T_g + (\Phi_g - \Phi)/(\beta\Phi_0)$, where Φ_g is defined as $\tau_\alpha(\Phi_g)$= 100s, Φ_0 is a characteristic volume fraction, and β is the effective thermal expansion coefficient. This model provides a description of the α-relaxation dynamics of a molecular glass former, and it is used in the papers to model the α-relaxation dynamics given the T_g predicted by the network.

1.2.7 Experimental techniques

1.2.7.1 Differential Scanning Calorimetry

The differential scanning calorimetry (DSC) is a thermoanalyitical technique which measures the heat necessary to raise the temperature of a sample as a function of temperature. It registers changes in the heat flow, which is related to the heat capacity at constant pressure by the equation:

$$HF = c_p \frac{dT}{dt} \tag{21}$$

where c_p is the heat capacity and $\frac{dT}{dt}$ is the cooling or heating rate of the experiment. The setup is made by a reference and the sample which are heated or cooled at the same time. Changes in the heat flow correspond, among others, to changes in the phase of the sample, as they correspond to endothermic or exothermic transformations. In this way, it is possible to follow the transformation of a material and identify the different state passages. The glass transition temperature is identified by a step function behaviour of the heat flow as shown in Figure 1.2.11.

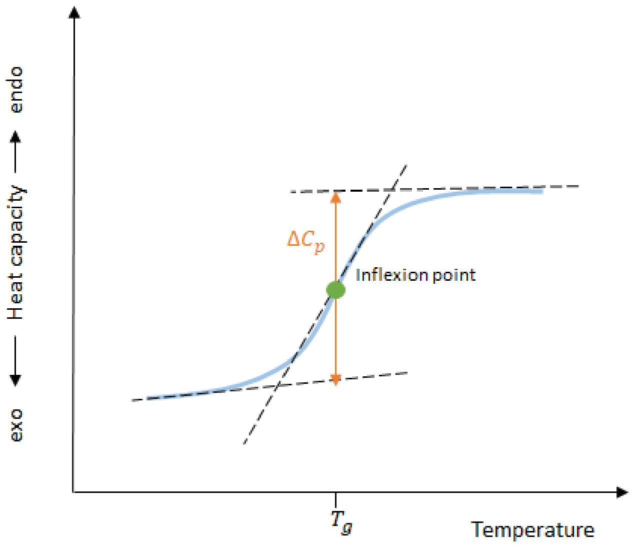


Figure 1.2.11 DSC measurement. In this case, the glass transition temperature corresponds to the inflection point of the heat capacity.

1.2.7.2 Broadband Dielectric Spectroscopy

Broadband dielectric spectroscopy (BDS) is an experimental technique based on the study of the response of a material to a given electric field through its dielectric permittivity, which is a measure of how the dipoles of the system are oriented with respect to an external electric field. The dielectric permittivity can be written in its complex form as

$$\varepsilon^{*}(\omega) = \varepsilon'(\omega) \; - \; i\varepsilon''(\omega) \tag{22}$$

where ε' is the real part and ε'' is the imaginary part. The relaxation processes are identified by a peak in the imaginary part of the permittivity and a step in its real part, as shown in Figure 1.2.12.

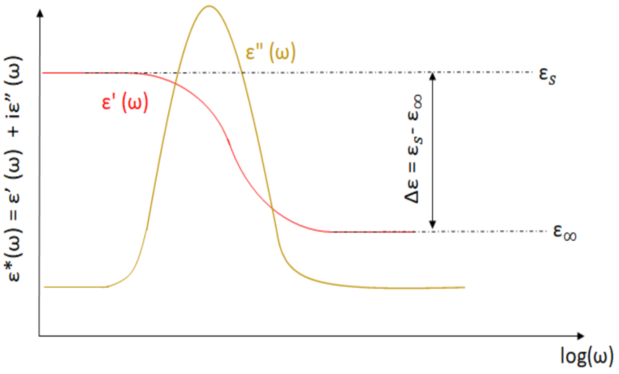

ε*(ω) = ε' (ω) + iε'' (ω)
ε" (ω)
ε' (ω)
ε_S
$\Delta\varepsilon = \varepsilon_S - \varepsilon_\infty$
ε_∞
log(ω)

Figure 1.2.12 Dielectric permittivity behaviour as a function of the frequency.

In BDS experiments the sample is put in between two gold-plated electrodes, creating a capacitor in which the sample acts as insulator. Then, an electric field E(ω) is applied to the sample and the permittivity is measured. According to Debye's equation, for an ideal case of non-interacting dipoles with a single time constant τ it is possible to write:

$$\varepsilon^*(\omega) = \varepsilon_\infty + \frac{\varepsilon_S - \varepsilon_\infty}{1+i\omega\tau} = \varepsilon_\infty + \frac{\Delta\varepsilon}{1+i\omega\tau} \tag{23}$$

where i is the imaginary unit, $\Delta\varepsilon = \varepsilon_S - \varepsilon_\infty$ is the dielectric strength, ε_S is the low frequency permittivity, ε_∞in the high frequency permittivity and τ is the Debye relaxation time. From Equation (23) it is possible to obtain an expression for the real and imaginary part of the permittivity:

$$\varepsilon'(\omega) = \varepsilon_\infty + \frac{\Delta\varepsilon}{1+(\omega\tau)^2} \tag{24.1}$$

$$\varepsilon''(\omega) = \frac{\Delta\varepsilon \, * \, \omega\tau}{1+(\omega\tau)^2} \tag{24.2}$$

This model, though, is only valid in few rare cases, and fails describing the permittivity behaviour of many materials, which usually show a broader peak in their imaginary part. For this reason, experimentally it is preferred to use empirical models like the Cole-Cole or the Havriliak-Negami equations (respectively Equation (25.1) and Equation (25.2)):

$$\varepsilon^*(\omega) = \varepsilon_\infty + \frac{\Delta\varepsilon}{1+(i\omega\tau_{CC})^\alpha} \tag{25.1}$$

$$\varepsilon^*(\omega) = \varepsilon_\infty + \frac{\Delta\varepsilon}{[1+(i\omega\tau_{HN})^\alpha]^\beta} \tag{25.2}$$

where τ_{CC} and τ_{HN} are characteristic times and α and β are tuneable parameters. As shown in Figure 1.2.13a, the position of the peak is temperature dependent, and following its path by changing the temperature allows to map the relaxation dynamics of the sample as a function of the temperature. The main process is called α-relaxation and it corresponds to a cooperative motion of the molecules related to the structural relaxation of the material, coinciding with the glass transition process when $\tau = 100s$. It is displayed in Figure 1.2.13b.

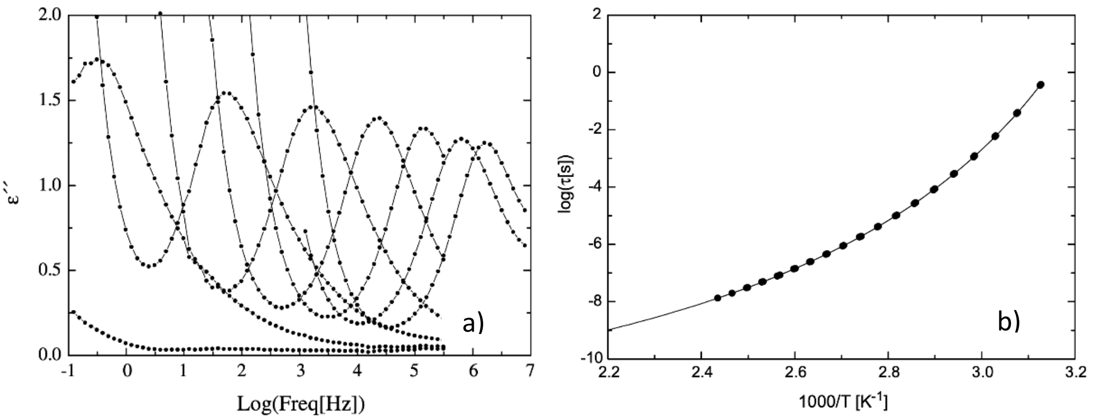


Figure 1.2.13 Typical glass former behaviour. Figure 1.2.13a reports the BDS measurements of the imaginary part of the permittivity, with the temperature increasing from left to right. Figure 1.2.13b shows the relaxation map given by the position of the peaks as a function of the inverse of the temperature.

Figure 1.2.13a shows the frequency dependence of the imaginary part of the complex dielectric permittivity at different temperatures for a typical glass former. Figure 1.2.13b shows the corresponding relaxation map where the logarithm of the maximum relaxation time is plotted as a function of inverse temperature.

1.3 Hypothesis and objectives

1.3.1 Hypothesis

The central hypothesis of this study is that the use of machine learning techniques, specifically ANNs, can effectively capture the non-linear relationship between the chemical structure of materials and their physico-chemical properties. By developing quantitative structure-property relationship models, it is possible to accurately predict and understand the properties of materials based on a representation of their chemical structure. The combination of different ANN architectures and the integration of the machine learning output with theoretical models extends the applicability of the predictions. The underlying assumption is that the chemical structure of materials holds crucial information that determines some of their properties. By leveraging the capabilities of ANNs to handle complex, non-linear relationships, it becomes feasible to establish correlations between the structure and property of materials. This hypothesis also acknowledges the potential of machine learning output as an input for theoretical models, implying that the combination of numerical and theoretical approaches can further improve the understanding of material properties.

1.3.2 Objectives

1.3.2.1 General Objective

To develop and apply machine learning techniques to establish quantitative structure-property relationships and estimate the glass transition temperature and relaxation dynamics of molecular glass formers and polymers based on their chemical structure.

1.3.2.2 Specific Objectives

- Contribute to the scientific understanding of structure-properties relationships in amorphous materials and accelerate the development of new materials.
- Assess the feasibility and accuracy of estimating the dynamics and T_g of molecular glass formers and polymers before their synthesis, thereby saving time and resources in the material development process.
- Validate the developed models and methodologies using experimental data and compare the predicted results with actual measurements.
- Contribute to model interpretability by performing a chemical embedding into an m-dimensional T_g-oriented space
- Explore different ANN models to establish a correlation between the chemical structure and T_g of molecular glass formers.
- Test the models on relevant compounds like essential amino acids and peptides.
- Combine ANNs and the elastically collective nonlinear Langevin equation (ECNLE) to estimate the temperature dependence of the main structural relaxation time of polymers and molecular glass formers using only the knowledge of their chemical structure.

1.4 Overview and discussion of the results

In this Thesis I present three works related with the theme of ANNs applied to the prediction of the glass transition temperature and relaxation dynamics of molecular and polymeric compounds. These strategies are based on the idea that the molecular structures contain enough information to get an estimation of the general properties of a compound, in this case the T_g. Thanks to the numerical/theoretical prediction of the property it is possible to save time and resources in the development of new compounds. Three different datasets were used, spanning biological and pharmaceutical compounds and polymers. The study was grounded in a combination of variables: the level of complexity of the network (starting with a fully connected network, then a convolutional neural network and finishing with a recurrent neural network), the complexity of the chemical structures, and by the nature of the output (dynamic properties).

1.4.1 First paper: Molecular glass formers and ECNLE

The first paper deals with the possibility to make hybrid models by using an ANN to predict the desired property and a theoretical model to show the physics of the dynamics. In particular, I do this applied to molecular glass formers and predicting their glass transition temperature, which is then used along with the ECNLE theory to estimate the α-relaxation dynamics of the compounds. In this article, I cured a dataset of around 200 molecular glass formers like biological molecules, pharmaceutical drugs, additives or sugars. The T_g of the compounds is in a range between 200 K and 400 K and the structures are expressed in SMILES strings (see Section 1.2.5). Figure 1.4.1a reports the T_g histogram of the dataset, and in Figure 1.4.1b shows the trend of the T_g with respect to the SMILES length of the molecular glass formers. Note that the value of the T_g increases with increasing SMILES's length, which is proportional to the molecular weight of the compounds.

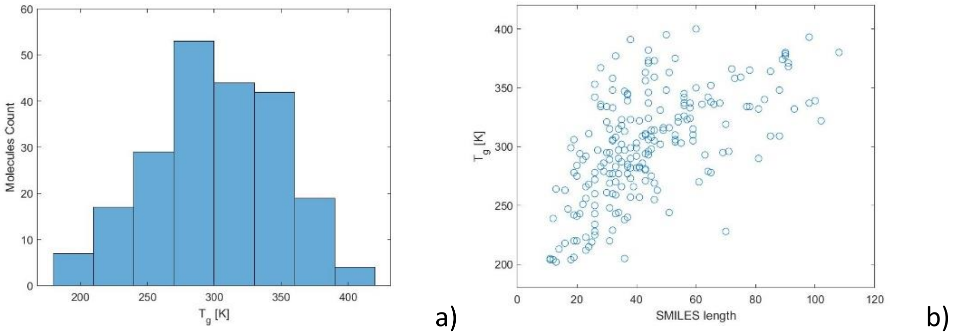


Figure 1.4.1 Characterisation of the dataset. a) shows the dataset distribution of the glass transition temperature, while b) shows the trend of the glass transition temperature with respect to the SMILES length of the compounds.

The pre-processing of the data consists in converting the SMILES strings into 2d matrices with the one-hot matrix encoding (see Section 1.2.5). This means that the i-th row of the matrix is filled with zeros except that for the dictionary position where there is the i-th symbol of the string. Then the matrix is flattened and fed to the fully connected neural network (see Figure 1.4.2), obtaining a prediction of the glass transition with average percentage errors below 8% (see Figure 1.4.3). This result is particularly good as in most cases the kinetic nature of the glass transition does not allow to find in literature a single value for the T_g, but rather a range of temperatures. The network was tested with the dropout layer before each hidden layer. This special layer, which is turned off when the network has been trained, is needed to avoid overfitting, as it puts to zero a fraction of the neurons of a layer so that the learning of a given feature is not linked to a restricted number of neurons, but is more evenly shared along all the neurons of the layer. I tested different architectures, changing the number of neurons for each layer and some of the hyperparameters, like the value of the dropout probability or the learning rate. Finally, I chose as best architecture two hidden layers of 40 neurons each and eLU activation function, with 40% probability of dropout and a starting learning rate of 0.01.

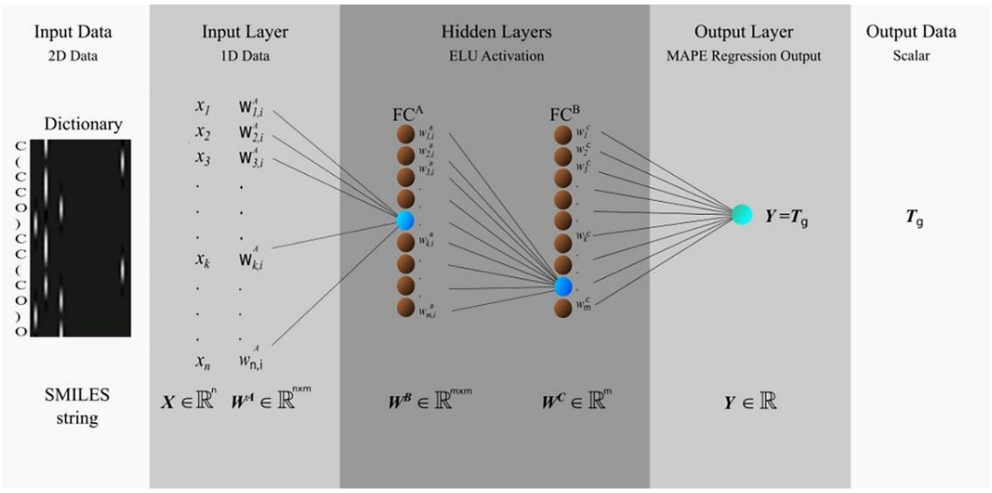


Figure 1.4.2 Fully connected architecture. The network, fed with a flattened version of the SMILES one-hot encoded, is composed by two hidden layers of 40 neurons each and eLU activation function and outputs the T_g.

The results of the predicted T_g are shown in Figure 1.4.3. The black line represents the bisector of the first quadrant of the cartesian axes, so this means that, according to the position of the dots in the graph, there is good accordance between the prediction and the actual value of the T_g.

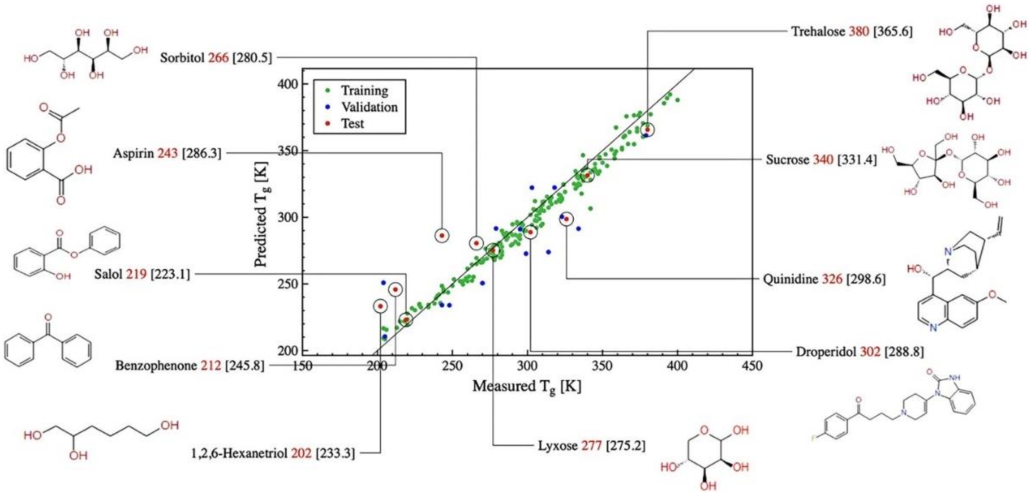


Figure 1.4.3 T_g predicted vs T_g experimental. The average percentage error on the predicted T_g is below 8%.

Then I used the predicted T_g as an input for the ECNLE theory (see Section 1.2.6.3), obtaining the molecular dynamics shown in Figure 1.4.4. This particular graph was obtained by using the correction a in the ECNLE theory so that the contribution of the elastic force F_e would be taken into account more precisely.

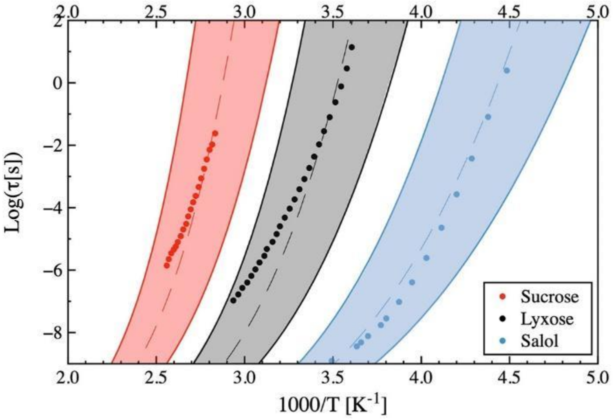


Figure 1.4.4 ECNLE results. The coloured bands represent the uncertainty due to the error on the prediction of the T_g*, while the dots represent the BDS experimental data of the molecular dynamics.*

Figure 1.4.4 compares the theoretical results (coloured bands and dashed lines) with the experimental data obtained by BDS measurements (dots). The concordance between theory and experiment is noteworthy and it shows that this hybrid approach is indeed valid to make assumptions on materials starting from a representation of their chemical structure. Moreover, this method can be used to understand how changes in the molecular structure lead to changes in the estimation of the T_g. As shown in Figure 1.4.5, it is possible to use the ANN as if it was a "virtual laboratory", where the relative positions of the atoms of the molecules change and reflect their effect on the estimation of the T_g.

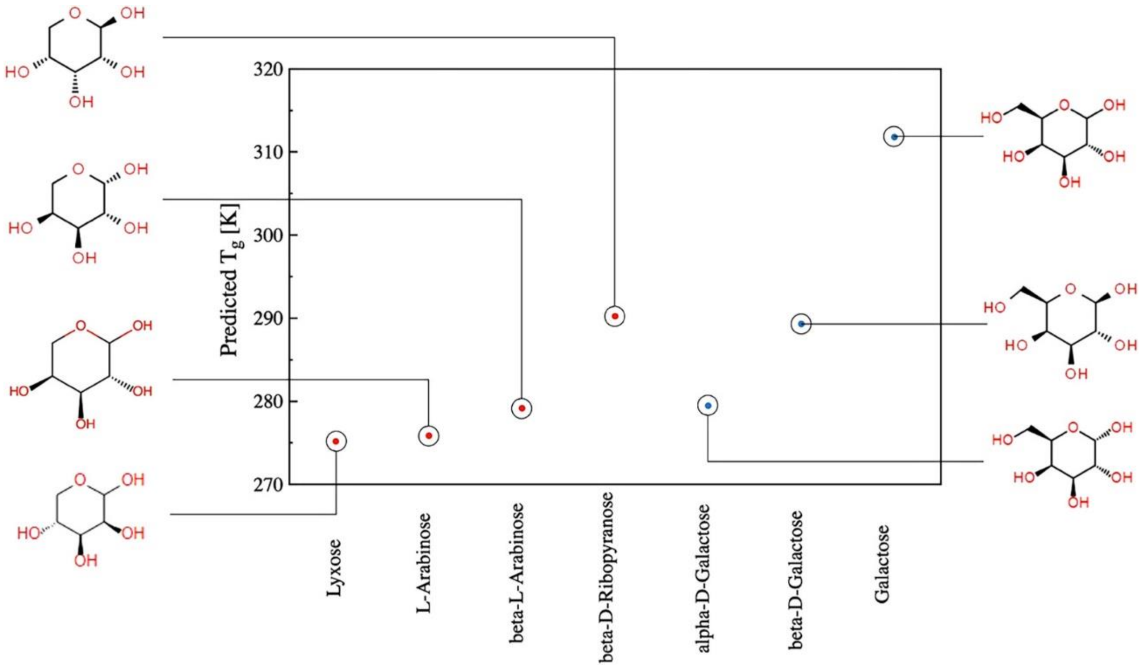


Figure 1.4.5 ANN as virtual laboratory. Small changes in the relative position of the atoms lead to changes in the estimation of the T_g *of the compound.*

1.4.2 Second paper: Acrylates and ECNLE

In the second paper, I focus on the application of convolutional neural networks to the prediction of the glass transition temperature for a family of polymers: the atactic polyacrylates. By taking advantage of the power of CNNs to detect patterns in the chemical structures, I obtain estimations of $T_g$which are subsequently utilized as inputs for the ECNLE model (see Section 1.2.6.3). To begin, I present the characterization of the dataset, consisting of approximately 200 monomers, as depicted in Figure 1.4.6. Figure 1.4.6a showcases the distribution of T_g values among the monomers, while Figure 1.4.6b illustrates the relationship between the SMILES length and the glass transition temperature. Since this approach relies on only the differential chemical structure (i.e.: all polymers share the same backbone structure, so we employ only the monomer structure as input, thus focusing the network to the study of the pending chains), the observed trend differs from that obtained in previous studies, where individual molecules were studied.

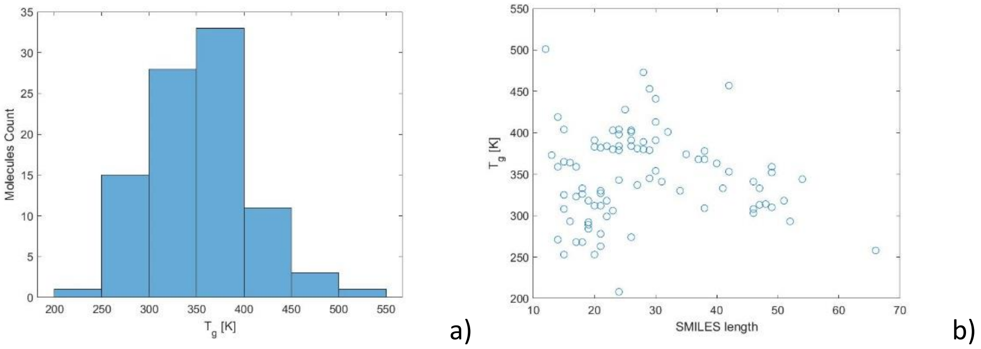


Figure 1.4.6 Characterisation of the dataset. a) shows the dataset distribution of the glass transition temperature, while b) shows the trend of the glass transition temperature with respect to the SMILES length of the monomers

To train the CNN model, the monomers are encoded as one-hot matrices derived from their SMILES strings, as explained in Section 1.2.5. Remarkably, this approach yields average percentage errors in the predictions of less than 9%, representing a significant outcome given that the network solely relies on monomer structure without any additional physical information.

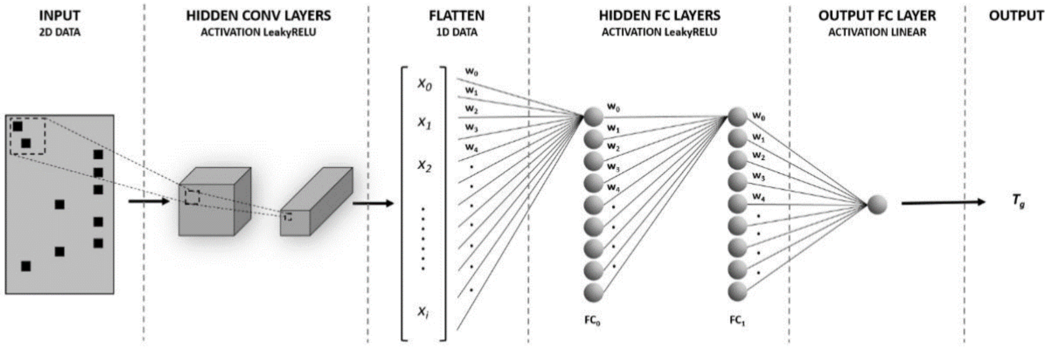


Figure 1.4.7 Convolutional neural network architecture. The 2D matrix encoding the SMILES is fed to the CNN, then the filtered image is flattened and passed to a fully connected architecture.

Subsequently, I feed these prediction results into the ECNLE theory, in order to obtain insights into the molecular dynamics of this specific class of polymers, as demonstrated in Figure 1.4.8. Figure 1.4.8 compares the main relaxation dynamics obtained by BDS measurement (blue line) with the ECNLE prediction obtained with the predicted T_g (red and dashed lines). The model was calculated by taking into account the value of α in the ECNLE, to have a more precise estimation of the contribution of the external elastic force F_e .

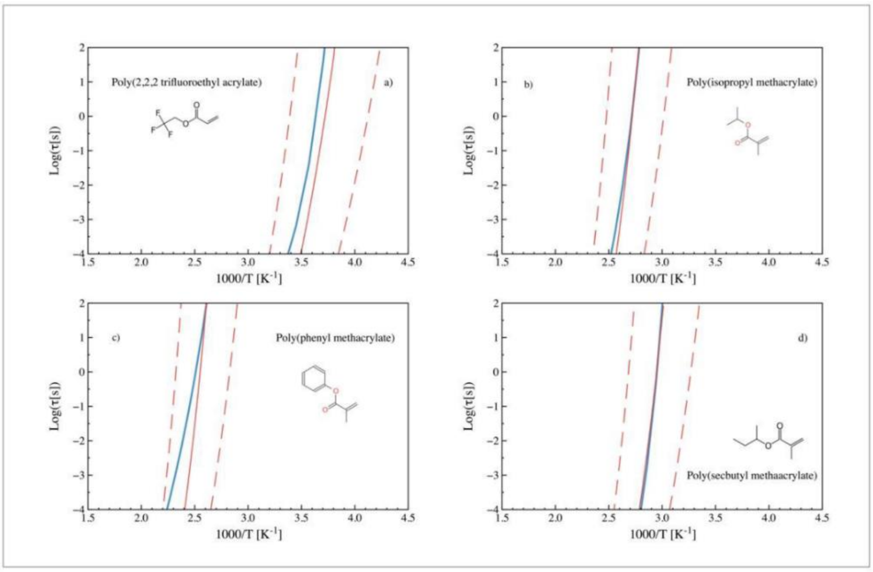


Figure 1.4.8 ECNLE results for the atactic poliacrylates. The blue line corresponds to VFT fit obtained by BDS measurements, while the red and dashed lines correspond to the range of the corresponding dynamics estimated by the ECNLE.

This hybrid approach of utilizing CNNs can open new routes in the design process of new polymeric materials, as it allows to have a good approximation of the dynamics of the compounds starting solely from the monomer.

1.4.3 Third paper: RNN and glass transition

In the third paper I show that the recurrent neural network is able to group and recognize the physics behind the glass transition process. In this case, the collected dataset counted about 500 molecular glass formers, and its characterisation is shown in Figure 1.4.9. The temperature range of the T_g spans between 18K and 450K. As in the first paper, we can see in Figure 4.9b that there is a strong correlation between the T_g and the SMILES's length, which increases with increasing molecular weight.

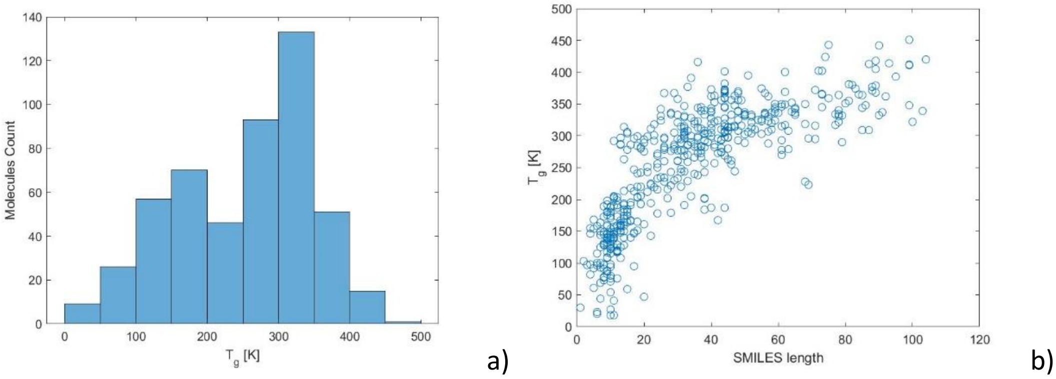


Figure 1.4.9 Characterisation of the dataset. a) shows the dataset distribution of the glass transition temperature, while b) shows the trend of the glass transition temperature with respect to the SMILES length of the compounds

Figure 1.4.10 shows a scheme of the RNN architecture used in the paper. In this case, the SMILES was encoded with the cardinal encoding (see Section 1.2.5) and bidirectional long short-term memory neurons were used. These particular nodes are advantageous because they analyse the given sequence both from left to right and from right to left, making it simpler to find significative patterns inside the sequence. The batch-normalization layer is essential for the performance of the network, as it is intrinsically a deep network and the batch-normalization helps to make the output of the BiLSTM less prone to saturation. After testing various architectures, I opted for a 8 nodes BiLSTM architecture, which become 16 as the network reads the sequence on both directions.

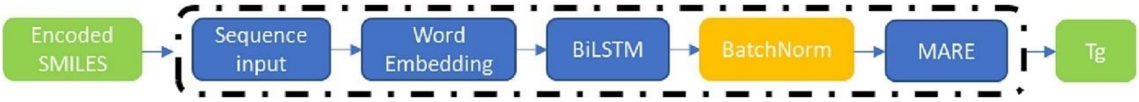


Figure 1.4.10 RNN architecture. The network is fed with the SMILES sequences expressed in cardinal encoding. Its core is the BiLSTM layer, which examines the sequences in both directions and is able to recognize significative patterns.

The graph in Figure 1.4.11 shows that the value of the predicted T_g as a function of the experimental T_g lies on the bisector of the cartesian plane, meaning that there is accordance between the prediction and the real value of the T_g. The average percentage error in this case is lower than 9%.

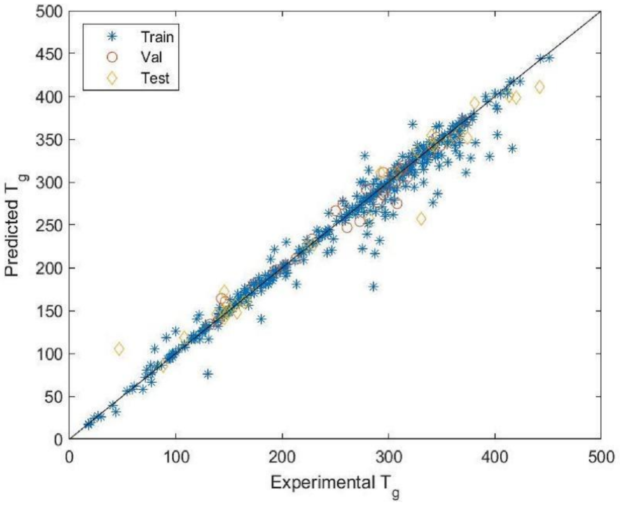


Figure 1.4.11 T_g predicted vs T_g experimental. The predicted T_g have an average percentage error below 9%

I showed with the Principal Component Analysis that the network is able to recognize and follow features in the chemical structure which influence the value of the glass transition temperature. I applied the Fuzzy-C clustering algorithm to the last hidden layer of the network (the batch-normalization layer, as it is supposed to be the most informative being it the one right before the output layer) to assess that the network is able to distinguish among the different structures. Then, I applied the PCA to reduce the dimensionality from 16 (number of neurons of the layer) to 2. The results of this analysis are shown in Figure 1.4.12.

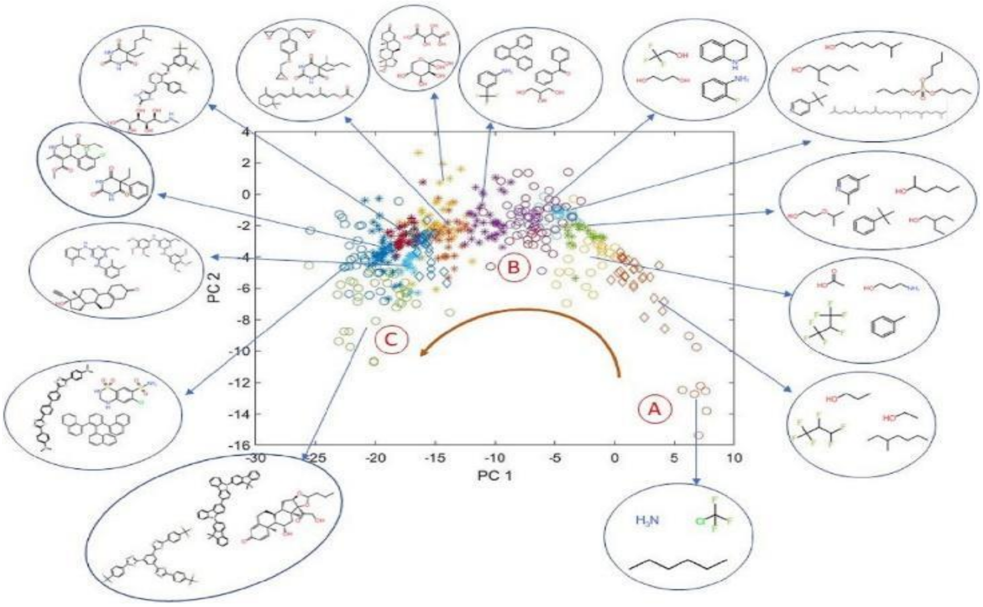


Figure 1.4.12 PCA and clustering. The PCA projects in 2 dimensions the results of the clustering applied in 16 dimensions. From point A to point C the network groups the molecules based on their T_g and structure relationship.

It is also possible to define a confidence interval based on the distribution of the T_g with respect to the molecular weight of the compounds, as shown in Figure 1.4.13. Such interval is a visualisation of the network working area, as to say of the features that it learnt from the dataset. The line from point A to point B shows that, at same molecular weight, the T_g increases according to an interplay of other molecular forces, demonstrating that the network was indeed able to grasp complex interactions laying in the process.

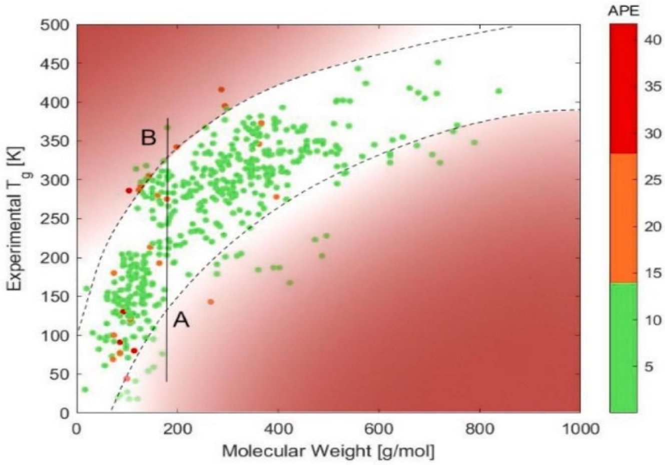


Figure 1.4.13 Confidence interval. This graph shows the chemical area which the neural network learned based on the molecular weight and the glass transition temperature

Then I used the recurrent neural network to predict the values of the glass transition temperatures of the 20 essential amino acids and a short peptide (3-lys). When possible, I compared the experimental T_g with the predicted one and found that the amino acids which were closer to the confidence interval were indeed predicted better than those which were farther from it, as shown in Figure 1.4.14 with blue dots for the more accurate predictions and red dots for the worse ones.

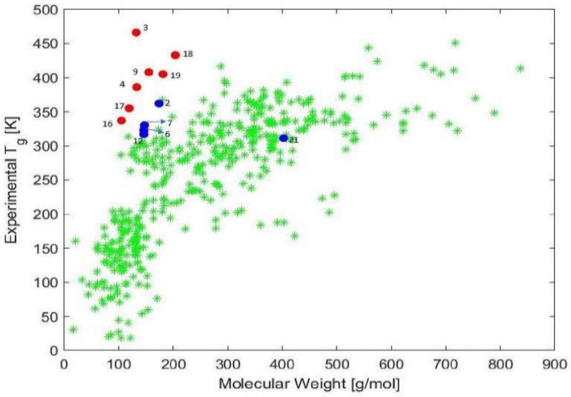


Figure 1.4.14 Essential amino acid prediction. The amino acids which are closer to the confidence interval (green dots) of the network (blue dots) are better predicted than those farther (red dots)

Also in this case, I demonstrated how it is possible to use the ANNs as a virtual laboratory where to test the effect of the molecular structure on the T_g. In particular, when dealing with biomolecules there are several issues which might occur, like for example the degradation of the sample upon heating or its crystallisation upon cooling. The use of numerical strategies to estimate the properties of such molecules is a way to overcome these difficulties and open new routes in the understanding of the relationship between the property and the molecular structure in Nature.

Bibliography

[1] A.R. Katritzky, V.S. Lobanov, M. Karelson, QSPR: the correlation and quantitative prediction of chemical and physical properties from structure, Chem. Soc. Rev. 24 (1995) 279–287. https://doi.org/10.1039/CS9952400279.

[2] A.R. Katritzky, M. Karelson, V.S. Lobanov, QSPR as a means of predicting and understanding chemical and physical properties in terms of structure, Pure and Applied Chemistry. 69 (1997) 245–248. https://doi.org/10.1351/pac199769020245.

[3] T. Le, V.C. Epa, F.R. Burden, D.A. Winkler, Quantitative Structure–Property Relationship Modeling of Diverse Materials Properties, Chem. Rev. 112 (2012) 2889–2919. https://doi.org/10.1021/cr200066h.

[4] D.R. Cassar, A.C.P.L.F. de Carvalho, E.D. Zanotto, Predicting glass transition temperatures using neural networks, Acta Materialia. 159 (2018) 249–256. https://doi.org/10.1016/j.actamat.2018.08.022.

[5] Machine Learning Prediction of Nine Molecular Properties Based on the SMILES Representation of the QM9 Quantum-Chemistry Dataset | The Journal of Physical Chemistry A, (n.d.). https://pubs.acs.org/doi/full/10.1021/acs.jpca.0c05969 (accessed June 6, 2023).

[6] L.A. Miccio, G.A. Schwartz, From chemical structure to quantitative polymer properties prediction through convolutional neural networks, Polymer. 193 (2020) 122341. https://doi.org/10.1016/j.polymer.2020.122341.

[7] L.A. Miccio, G.A. Schwartz, Localizing and quantifying the intra-monomer contributions to the glass transition temperature using artificial neural networks, Polymer. 203 (2020) 122786. https://doi.org/10.1016/j.polymer.2020.122786.

[8] W.X. Shen, X. Zeng, F. Zhu, Y. li Wang, C. Qin, Y. Tan, Y.Y. Jiang, Y.Z. Chen, Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations, Nat Mach Intell. 3 (2021) 334–343. https://doi.org/10.1038/s42256-021-00301-6.

[9] T. Hasebe, Knowledge-Embedded Message-Passing Neural Networks: Improving Molecular Property Prediction with Human Knowledge, ACS Omega. 6 (2021) 27955–27967. https://doi.org/10.1021/acsomega.1c03839.

[10] L.A. Miccio, G.A. Schwartz, Mapping Chemical Structure–Glass Transition Temperature Relationship through Artificial Intelligence, Macromolecules. 54 (2021) 1811–1817. https://doi.org/10.1021/acs.macromol.0c02594.

[11] A. Karthikeyan, U.D. Priyakumar, Artificial intelligence: machine learning for chemical sciences, J Chem Sci (Bangalore). 134 (2022) 2. https://doi.org/10.1007/s12039-021-01995-2.

[12] C.M. Bishop, Neural networks and their applications, Review of Scientific Instruments. 65 (1994) 1803–1832. https://doi.org/10.1063/1.1144830.

[13] L. Tao, V. Varshney, Y. Li, Benchmarking Machine Learning Models for Polymer Informatics: An Example of Glass Transition Temperature, J. Chem. Inf. Model. 61 (2021) 5395–5413. https://doi.org/10.1021/acs.jcim.1c01031.

[14] Z. Tan, Y. Li, W. Shi, S. Yang, A Multitask Approach to Learn Molecular Properties, J. Chem. Inf. Model. 61 (2021) 3824–3834. https://doi.org/10.1021/acs.jcim.1c00646.

[15] A. Tropsha, P. Gramatica, V.K. Gombar, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, QSAR & Combinatorial Science. 22 (2003) 69–77. https://doi.org/10.1002/qsar.200390007.

[16] W. Sha, K.L. Edwards, The use of artificial neural networks in materials science based research, Materials & Design. 28 (2007) 1747–1752. https://doi.org/10.1016/j.matdes.2007.02.009.

[17] D. Champion, M. Le Meste, D. Simatos, Towards an improved understanding of glass transition and relaxations in foods: molecular mobility in the glass transition range, Trends in Food Science & Technology. 11 (2000) 41–55. https://doi.org/10.1016/S0924-2244(00)00047-9.

[18] M.G. Abiad, M.T. Carvajal, O.H. Campanella, A Review on Methods and Theories to Describe the Glass Transition Phenomenon: Applications in Food and Pharmaceutical Products, Food Eng. Rev. 1 (2009) 105–132. https://doi.org/10.1007/s12393-009-9009-1.

[19] N.R. Jadhav, V.L. Gaikwad, K.J. Nair, H.M. Kadam, Glass transition temperature: Basics and application in pharmaceutical sector, Asian Journal of Pharmaceutics (AJP): Free Full Text Articles from Asian J Pharm. 3 (2014). https://doi.org/10.22377/ajp.v3i2.246.

[20] S.A. Umoren, M.M. Solomon, Protective polymeric films for industrial substrates: A critical review on past and recent applications with conducting polymers and polymer composites/nanocomposites, Progress in Materials Science. 104 (2019) 380–450. https://doi.org/10.1016/j.pmatsci.2019.04.002.

[21] Vidya, L. Mandal, B. Verma, P.K. Patel, Review on polymer nanocomposite for ballistic & aerospace applications, Materials Today: Proceedings. 26 (2020) 3161–3166. https://doi.org/10.1016/j.matpr.2020.02.652.

[22] J. Bicerano, Prediction of Polymer Properties, CRC Press, 2002.

[23] D.W. van Krevelen, K. te Nijenhuis, Properties of Polymers: Their Correlation with Chemical Structure; their Numerical Estimation and Prediction from Additive Group Contributions, Elsevier, 2009.

[24] S.Z.D. Cheng, Handbook of Thermal Analysis and Calorimetry: Applications to Polymers and Plastics, Elsevier, 2002.

[25] Y. Dong, Y. Ruan, H. Wang, Y. Zhao, D. Bi, Studies on glass transition temperature of chitosan with four techniques, Journal of Applied Polymer Science. 93 (2004) 1553–1558. https://doi.org/10.1002/app.20630.

[26] J.A. Torres, P.F. Nealey, J.J. de Pablo, Molecular Simulation of Ultrathin Polymeric Films near the Glass Transition, Phys. Rev. Lett. 85 (2000) 3221–3224. https://doi.org/10.1103/PhysRevLett.85.3221.

[27] N.C. Ekdawi-Sever, P.B. Conrad, J.J. de Pablo, Molecular Simulation of Sucrose Solutions near the Glass Transition Temperature, J. Phys. Chem. A. 105 (2001) 734–742. https://doi.org/10.1021/jp002722i.

[28] G.B. Goh, N. Hodas, C. Siegel, A. Vishnu, SMILES2vec: Predicting Chemical Properties from Text Representations, (2018). https://openreview.net/forum?id=B1NTHukPf (accessed November 2, 2022).

[29] G. Chen, L. Tao, Y. Li, Predicting Polymers' Glass Transition Temperature by a Chemical Language Processing Model, Polymers. 13 (2021) 1898. https://doi.org/10.3390/polym13111898.

[30] A.D. Phan, K.S. Schweizer, Elastically Collective Nonlinear Langevin Equation Theory of Glass-Forming Liquids: Transient Localization, Thermodynamic Mapping, and Cooperativity, J. Phys. Chem. B. 122 (2018) 8451–8461. https://doi.org/10.1021/acs.jpcb.8b04975.

[31] T.M. Mitchell, Machine Learning, McGraw-Hill, New York, 1997.

[32] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bulletin of Mathematical Biophysics. 5 (1943) 115–133. https://doi.org/10.1007/BF02478259.

[33] Ian Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning, MIT Press, 2016. http://www.deeplearningbook.org.

[34] Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong, Mathematics for Machine Learning, Cambridge University Press, 2020. https://mml-book.com.

[35] H.-H. Bock, Clustering Methods: A History of k-Means Algorithms, in: P. Brito, G. Cucumel, P. Bertrand, F. de Carvalho (Eds.), Selected Contributions in Data Analysis and Classification, Springer, Berlin, Heidelberg, 2007: pp. 161–172. https://doi.org/10.1007/978-3-540-73560-1_15.

[36] E.H. Ruspini, J.C. Bezdek, J.M. Keller, Fuzzy Clustering: A Historical Perspective, IEEE Computational Intelligence Magazine. 14 (2019) 45–55. https://doi.org/10.1109/MCI.2018.2881643.

[37] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36. https://doi.org/10.1021/ci00057a005.

[38] Richard A. L. Jones, Soft Condensed Matter, Oxford University Press, 2002.

Section 2

Conclusions

The QSPR is a field which is growing to find in silico alternatives to the experimental paths in the development of new materials. The main aim of these models is to generate accurate predictions with the only information of the molecular structure of the compounds. One of the most powerful tools to achieve such results is the use of ANNs, as they are able to identify non-linear correlations between their inputs and outputs. The papers focus on the use of ANNs as a way to predict a given property (in this case the glass transition temperature) starting from a representation of the molecular structure. The most powerful item shared among the papers is that the representation of the molecular structure is given by an alpha-numerical string of characters, which does not contain the actual position of the atoms in the space, but whose pattern is used to generate a graphical drawing of the molecular structure. I demonstrate that this pattern contains enough information for the network to get the physics behind the process by finding the non-linear correlation between structure and property. Moreover, the possibility to access the layers of the network allows to have a mathematical representation of the structure which also contains information about its relationship with the predicted property. The activations of these layers can in fact be thought as numerical vectors and eventually be used as inputs for other neural networks (transfer learning) or used to apply mathematical operations on them (like clustering). This feature provides new tools to study the structure-property relationship of non-synthesised new materials and to make hypothesis on experimentally forbidden areas, like for example for biomolecules which often undergo degradation in the process of characterization. Also, the output of the network can be used as an input for theoretical models, avoiding the costly process of synthesis involved in the characterisation of new materials. For example, once obtained a certain degree of accuracy in the prediction of the T_g, it was possible to obtain relaxation maps very similar to the experimental results by using the ECNLE theory. On the one hand, in this Thesis I analyse the theme of the molecular glass-formers, which are often characterized by complex interactions which influence their glass transition temperature and, though, are well predicted by the networks (first and third paper). On the other hand, I study a family of polymers, the atactic polyacrylates, showing how it is possible to obtain a good approximation of their molecular dynamics starting from their monomer structure (second paper). Altogether, in this Thesis I propose three different methods to estimate the value of the T_g starting from a representation of the molecular structure of the glass former, showing that it is possible to obtain reliable results which are comparable with their experimental counterparts, opening new routes in the developing of new materials and in the study of experimentally forbidden regions.

Section 3

Annex

The following articles were published in the frame of this Ph.D. Thesis. Below the complete reference to each paper is detailed.

- *Estimating glass transition temperature and related dynamics of molecular glass formers combining artificial neural networks and disordered systems theory*, Claudia Borredon, Luis A. Miccio, Anh D. Phan, Gustavo A. Schwartz, *Journal of Non-Crystalline Solids: X*, **Volume 15**, 100106, (2022).

  This journal was Q2 in *Materials Chemistry* and had an impact factor of 2.21 at the time of the publication.

- *Approaching Polymer Dynamics Combining Artificial Neural Networks and Elastically Collective Nonlinear Langevin Equation*, Luis A. Miccio, Claudia Borredon, Ulises Casado, Anh D. Phan, Gustavo A. Schwartz, *Polymers* **2022**, *14*, 1573.

  This journal was Q1 in *Polymer Science* and had an impact factor of 5 at the time of the publication.

- *Characterising the glass transition temperature-structure relationship through recurrent neural network*, Claudia Borredon, Luis A. Miccio, Silvina Cerveny, Gustavo A. Schwartz, *Journal of Non-Crystalline Solids: X*, **Volume 18**, 100185, (2023).

  This journal was Q2 in *Materials Chemistry* and had an impact factor of 2.21 at the time of the publication.

Contents lists available at ScienceDirect

Journal of Non-Crystalline Solids: X

journal homepage: www.sciencedirect.com/journal/journal-of-non-crystalline-solids-x

Estimating glass transition temperature and related dynamics of molecular glass formers combining artificial neural networks and disordered systems theory

Claudia Borredon [a], Luis A. Miccio [a,b,c], Anh D. Phan [d,e], 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
[d] Faculty of Materials Science and Engineering, Phenikaa University, Hanoi 12116, Viet Nam
[e] Phenikaa Institute for Advanced Study (PIAS), Phenikaa University, Hanoi 12116, Viet Nam

ARTICLE INFO

Keywords:
QSPR
Properties prediction
Artificial neural networks
ECNLE

ABSTRACT

Glass transition temperature and related dynamics play an essential role in amorphous materials research since many of their properties and functionalities depend on molecular mobility. However, the temperature dependence of the structural relaxation time for a given glass former is only experimentally accessible after synthesizing it, implying a time-consuming and costly process. In this work, we propose combining artificial neural networks and disordered systems theory to estimate the glass transition temperature and the temperature dependence of the main relaxation time based on the knowledge of the molecule's chemical structure. This approach provides a way to assess the dynamics of molecular glass formers, with reasonable accuracy, even before synthesizing them. We expect this methodology to boost industrial development, save time and resources, and accelerate the scientific understanding of structure-properties relationships.

1. Introduction

Quantitative structure-property relationships (QSPR) models can boost both materials design and scientific understanding of molecular glass formers. They can correlate the molecular structure with important properties like glass transition temperature and its related dynamics, which are among the most significant issues associated with the behaviour of glass formers. Many challenging problems, especially in the pharmaceutical industry, like the tendency to recrystallize, water solubility and dissolution rate, or the long-term stability [1–4], are related to the structural relaxation dynamics and the glass transition temperature. This molecular relaxation process is usually described by a characteristic relaxation time and its temperature dependence, which can be experimentally measured using broadband dielectric spectroscopy (BDS), dynamic light scattering (DLS), or dynamic mechanical analysis (DMA), among other techniques. However, when designing new molecular glass formers, we do not know their dynamics before synthesizing and characterizing them, which are costly and time-consuming processes. In this sense, QSPR models are able to estimate the desired properties based only on the chemical structure of the molecules.

Some theoretical approaches can help in these challenging tasks. For instance, the elastically collective nonlinear Langevin equation (ECNLE) theory has been recently used to successfully describe the temperature dependence of the relaxation times of different amorphous materials [5]. However, this approach requires the knowledge of the glass transition temperature (T_g) to estimate the molecular dynamics. Although this information is not available for new glass formers until synthesized, recent developments based on artificial neural networks (ANN) allow estimating their glass transition temperature based only on their chemical structure [6–8], without involving any experimental measurements or complex synthesis.

This work proposes a joint theoretical and numerical approach to estimate the glass transition temperature and the temperature dependence of the structural relaxation time for several molecular glass formers (including amorphous drugs and biomolecules), based only on their chemical structure. A neural network approach is firstly used to

[*] Correspondence to: G. A. Schwartz, Centro de Física de Materiales (CSIC-UPV/EHU), Materials Physics Center (MPC), P. M. de Lardizábal 5, 20018 San Sebastián, Spain.
E-mail address: gustavo.schwartz@csic.es (G.A. Schwartz).

https://doi.org/10.1016/j.nocx.2022.100106
Received 17 March 2022; Received in revised form 1 June 2022; Accepted 11 June 2022
Available online 17 June 2022

estimate the glass transition temperature of “new” compounds from their chemical structure codified into a Simplified Molecular Input Line Entry System (SMILES) representation; then, this information is used in ECNLE theory to estimate the temperature dependence of the relaxation time for the structural relaxation process. In this way, we can estimate the dynamics of molecular glass formers, even before synthesizing them, only knowing their chemical structure.

2. Theoretical background

ECNLE theory describes glass-forming liquids using a hard-sphere fluid [1,5,9–15]. Key characteristics of the fluid are the particle size, d, and the number of particles per volume, ρ, from where the volume fraction is estimated as $\Phi = \rho\pi d^3/6$. Two main factors affecting the mobility of a tagged particle are 1) interactions with its nearest neighbours and 2) cooperative motions of particles beyond the first shell. The local dynamics (or the motion of the tagged particle within a particle cage) is quantified by the dynamic free energy [1,5,9–15], $F_{dyn}(r) = F_{ideal}(r) + F_{caging}(r)$, where r is the displacement. $F_{ideal}(r)$ corresponds to the delocalized or ideal fluid state and $F_{caging}(r)$ characterizes the localized state of the particle via caging forces, which strongly depends on the density and structure of systems.

Fig. 1 shows an example for calculations of $F_{dyn}(r)$ at $\Phi = 0.58$ and indicates physical quantities of the local dynamics. In a sufficiently dense fluid, a reduction of the free volume dynamically restricts the motion of particles and forms a particle cage surrounding a tagged particle. The dynamical constraint is characterized by the emergence of a barrier in $F_{dyn}(r)$. A particle cage radius, r_{cage}, is roughly estimated by the first minimum of the radial distribution function, $g(r)$. Other important length scales of the local dynamics are a localization length, r_L, a barrier position, r_B, a jump distance, $\Delta r = r_B - r_L$, and a local barrier, $F_B = F_{dyn}(r_B) - F_{dyn}(r_L)$. From these, we can calculate $K_0 = \frac{\partial^2 F_{dyn}(r)}{\partial r^2}\rceil_{r=r_L}$ and $K_B = \frac{\partial^2 F_{dyn}(r)}{\partial r^2}\rceil_{r=r_B}$ corresponding to harmonic curvatures at r_L and r_B, respectively. K_0 can be interpreted as a spring constant at the localization length.

Escaping of a particle from its cage requires reorganization of both the nearest neighbours and all particles outside the cage to generate the extra space. Thus, the collective motions are strongly coupled to the local dynamics within the cage. The cooperative rearrangement creates a displacement field, $u(r)$, from the surface of the particle cage, that triggers particles beyond the first coordination shell to vibrate as oscillators and radially propagates through the rest medium. By employing Lifshitz's linear continuum mechanics [16], one can analytically calculate the displacement field for $r \geq r_{cage}$ as $u(r) = \frac{\Delta r_{eff} r_{cage}^2}{r^2}$, where Δr_{eff} is the amplitude of the field, whose mathematical expression was reported elsewhere [14,15]. Since $u(r)$ is small, the oscillation of each particle is approximately harmonic and, thus, the elastic energy of an oscillator is $K_0\frac{u^2(r)}{2}$. From this, we quantify collective motion effects on the relaxation process by summing the harmonic elastic energy of particles outside the cage to obtain the collective elastic barrier, which is $F_e = 4\pi\rho\int_{r_{cage}}^{\infty} drr^2 g(r)K_0\frac{u^2(r)}{2}$.

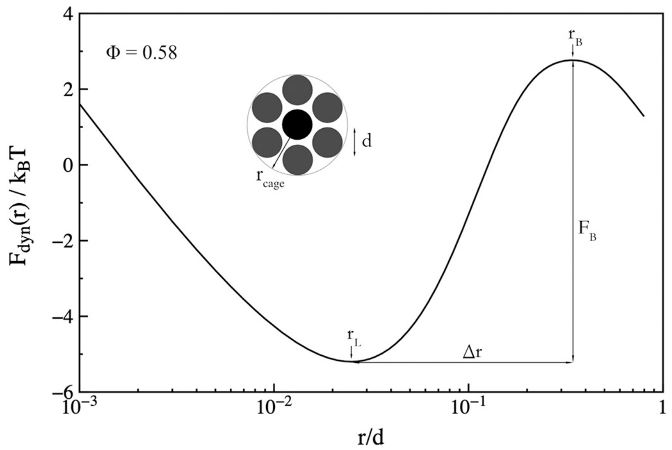


Fig. 1. The free energy profile normalized by k_BT for a hard-sphere fluid with $\Phi = 0.58$, where k_B is the Boltzmann constant and T is the temperature. Characteristic length and energy scales for the local dynamics are defined.

Inserting the local and elastic components into Kramer's theory gives us the structural relaxation time

$$\frac{\tau_\alpha}{\tau_s} = 1 + \frac{2\pi}{\sqrt{K_0 K_B}} \frac{k_B T}{d^2} exp\left(\frac{F_B + F_e}{k_B T}\right) \tag{1}$$

where τ_s is a short time scale and its analytical form was previously reported [1,5,9–15]. Note that the above calculations provide $\tau_\alpha(\Phi)$. Direct comparisons between theory and experiments need a density-to-temperature conversion. In prior works [1,9–12], based on a thermal expansion process, it was proposed a simple thermal mapping $T = T_g + (\Phi_g - \Phi)/\beta\Phi_0$, where T_g is the dynamic glass transition temperature defined by $\tau_\alpha(T_g) = 100$s, Φ_g is the volume fraction when $\tau_\alpha(\Phi_g \approx 0.6157) = 100$s, $\Phi_0 \approx 0.50$ is a characteristic volume fraction, and $\beta \approx 12 \times 10^{-4}\text{K}^{-1}$ is an effective thermal expansion coefficient considered constant for all amorphous materials [1,9–12].

Fig. 2 shows the theoretical and experimental temperature dependence of structural relaxation time for griseofulvin, nordazepam, celecoxib, tetrazepam, and ibuprofen. Overall, theoretical results quantitatively agree with experimental data over a wide temperature range or timescale without any fitting parameter.

The slight deviation observed for ibuprofen (solid line) is related to the fact that we assume that local and collective dynamics correlate to each other in a universal manner for all materials. F_B and F_e in Eq. (1) are summed with the ratio of prefactor equal to 1. In prior works [12,17], ECNLE calculations were improved by multiplying the elastic barrier with an adjustable parameter a to change the relative importance of collective dynamics in the glass transition. The new adjusted elastic barrier is $F_e \rightarrow a^2 F_e$ and it modifies the structural relaxation time in Eq. (1) as

$$\frac{\tau_\alpha}{\tau_s} = 1 + \frac{2\pi}{\sqrt{K_0 K_B}} \frac{k_B T}{d^2} exp\left(\frac{F_B + a^2 F_e}{k_B T}\right) \tag{2}$$

The value of Φ_g and the thermal mapping are strongly dependent on

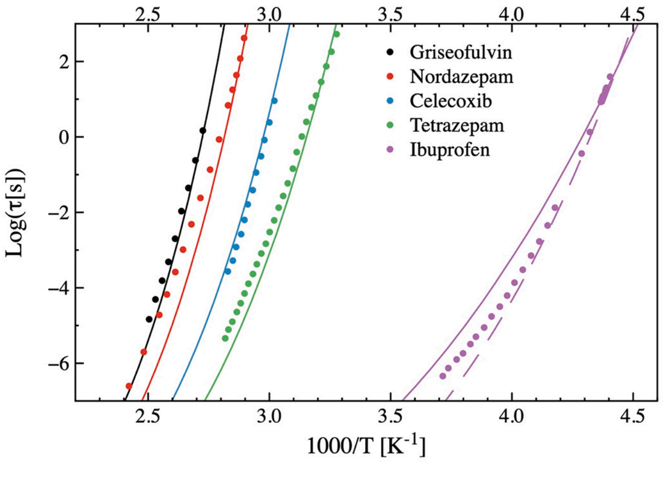


Fig. 2. Temperature dependence of the structural relaxation time calculated using the ECNLE theory (solid lines) and the corresponding experimental values (dots) for several molecular glass formers. The dashed line represents the ECNLE prediction for ibuprofen with the adjustable parameter $a = 2.5$ taken from Fig. 3.

the parameter a, which accounts for the non-universal effects of biological, conformational, and chemical complexities on the collective motions of molecules. It was empirically observed that the scaling parameter (a) typically increases with increasing fragility and depends on the glass transition temperature, as shown in Fig. 3. Although the correlation is not strong, there is a clear trend indicating an increment of the parameter a upon decreasing glass transition temperature. The estimated value of the parameter a for ibuprofen is 2.5 (being the experimentally observed value 2.4). The dashed line in Fig. 2 shows the ECNLE prediction for ibuprofen using this value for the scaling parameter.

3. Numerical background

This section describes the dataset, the encoding of the chemical structures, and the corresponding T_g prediction using ANN.

3.1. Dataset

The dataset for this work is composed of 216 molecules, including pharmaceutical drugs (like benzocaine or ibuprofen), biological molecules (like sucrose or ribose), and typical additives used in the pharmaceutical industry (like benzophenone), with T_g in the range 200 K – 400 K. We accounted for the chemical and spatial structure of the compounds by using the Simplified Molecular Input Line Entry System (SMILES) [18,19], which codifies the molecules into a string of characters. Table 1 in the Supporting Information (SI) shows the name of the compounds, their corresponding SMILES code, and the experimental and estimated T_g values.

3.2. Data treatment (encoding)

Following the same approach reported in previous works [6–8], we used a one-hot encoding method and an appropriate dictionary with all the existing characters in the SMILES code to convert the SMILES strings (1D) into binary matrices (2D). Thus, row i-th of the matrix is filled with zeros except for the position of the dictionary that coincides with the same character on the i-th position in the SMILES code. A one is placed in this case. Therefore, the number of characters in the dictionary (nd) and the length of each SMILES code (npos) define the columns and the rows of the matrix (as shown in Fig. 1 in the SI).

3.3. ANN's architecture

Based on previous results [6], we used a fully connected neural network for this work. We tried different architectures varying the number of hidden layers, the number of neurons, the dropout probability, and the activation functions to improve the performance of the ANN. Fig. 4 shows a scheme of the optimal network: the inputs to the ANN are the flattened versions of the 2D SMILES matrices; we then pass the input to two fully connected hidden layers, containing 40 neurons each, and a single output regression layer. We used the ELU activation function in the hidden layers, a variation of the most common ReLU activation function, characterized by an exponential contribution [20,21]. In addition, we imposed a 40% dropout probability on each hidden layer in the training phase, whereas the output regression activation function was linear, and its loss function was the mean average percentage error (MAPE), defined by:

$$Loss = \frac{1}{m_x}\sum_{i=1}^{m_x}\frac{|T_i - T_i'|}{T_i} \quad (3)$$

where m_x represents the number of elements in the x-th mini-batch, T_i represents the experimental T_g value collected for the i-th compound in the mini-batch and T_i' the calculated value using the ANN for the same compound.

3.4. ANN's optimization

The network was trained using the Adam optimization algorithm [22] provided by MATLAB with the default parameters for beta_1, beta_2, and epsilon (0.9, 0.999 and 10^{-8}, respectively) and applying the mini-batch strategy (mini-batch size = 16) to estimate the gradient of the loss function. In addition to the Adam algorithm, we imposed an external drop of the initial learning rate (lr), starting from $lr = 0.01$ and multiplying it by 0.25 every 500 epochs. We found that initializing the hidden layers and the output layers with a bias vector and a weight tensor whose elements are all ones, significantly improved the network's performance, compared to other initialization methods [23], most likely due to the inclusion of the dropout layers during training, which randomize the data transfer from one layer to the next one [24].

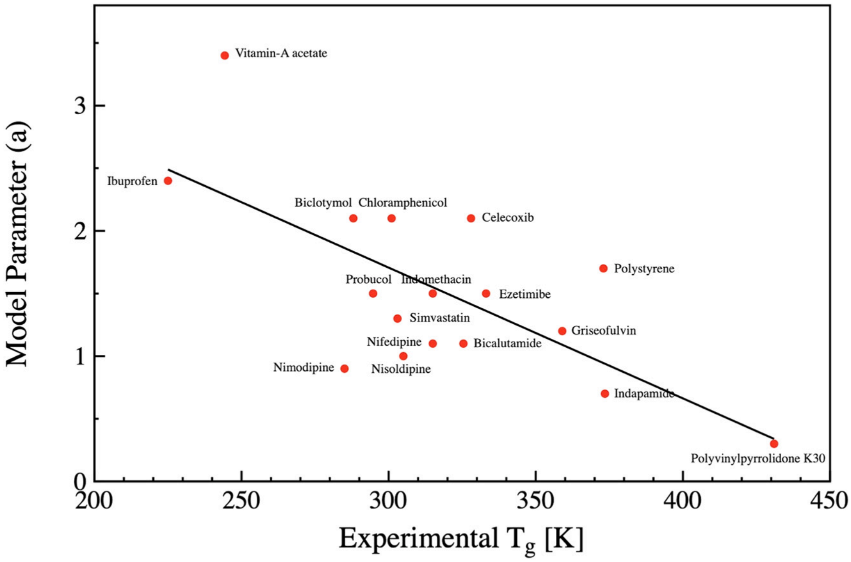


Fig. 3. Glass transition temperature dependence of the model adjustable parameter a for several glass formers. The solid line represents the linear fit of the experimental data. Data were taken from reference 12.

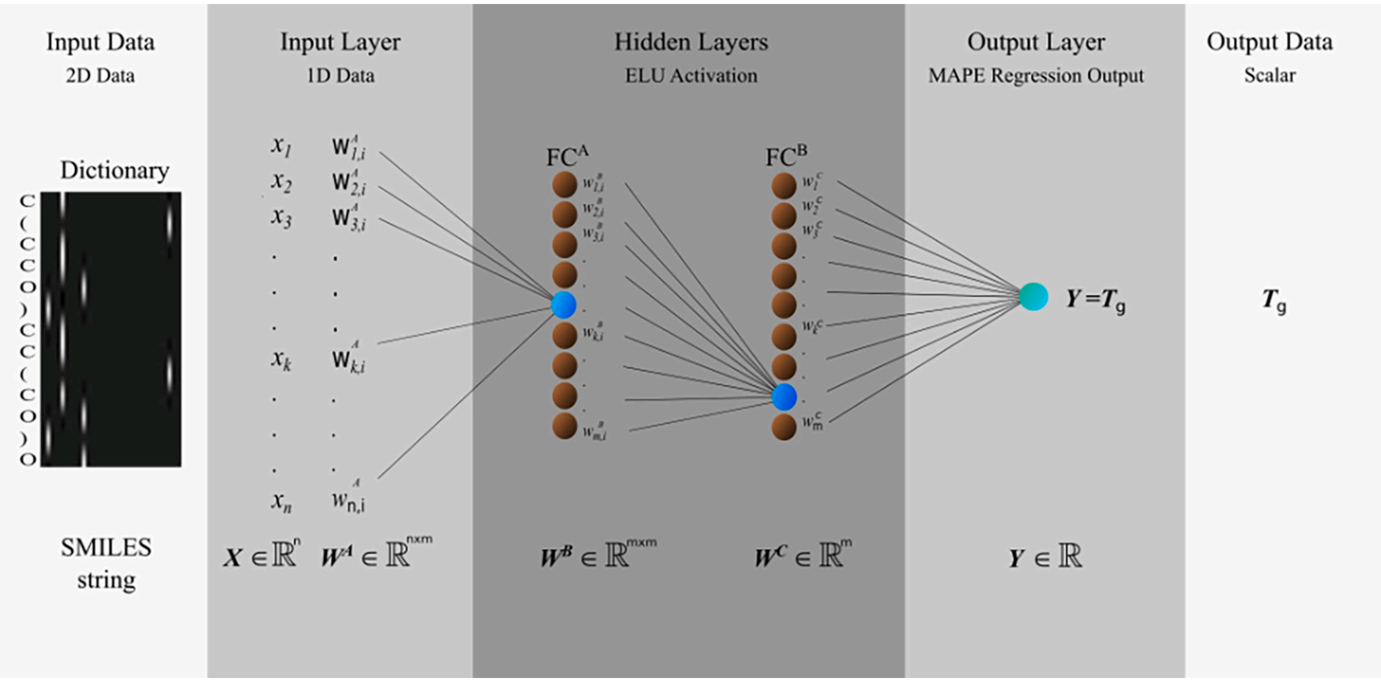


Fig. 4. Schematic picture of the artificial neural network used to predict the glass transition temperature.

3.5. ANN training

The first step consists of training the neural network with known pairs of SMILES strings – glass transition temperatures. This data set is called the *training set*. The examples in the training set are fed into the network, which compares the predicted value with the corresponding experimental T_g. Then, the network adjusts its weights and bias values, using an appropriate learning algorithm, to minimize the average relative error between predicted and known T_g values. In parallel to the training process, the average relative error of the *validation set* is also calculated after each epoch (1 epoch = number of iterations over the mini-batches so that the whole training set is spanned). Since the molecules in the validation set do not participate in the training process, the prediction of their T_g values gives an estimation of the generalization power of the neural network. Once the network can generalize, it is fed with the molecules of the *test set*. This data set corresponds to molecules that were never fed to the network during the training phase, and it is in this sense that we say they are “unknown” (or “new”) compounds. Since these molecules do not belong to training or validation sets, we can say that the neural network *predicts* their glass transition temperature.

We divided the data set into 90% training set, 6% validation set, and 4% test set for two main reasons: on the one hand, due to the structural complexity and variability of the molecules in our data set, the network needs to learn many different features and therefore it needs to have as many examples (molecules) as possible in the training set; on the other hand, we selected for test set those molecules for which we could find published experimental measurements of the alpha-relaxation dynamics, in order to have physical feedback to compare our method with. In addition to this, we also wanted to ensure that the T_gs of the test set span over the whole temperature range (200 K–400 K). Once we extracted the test set molecules from the data set, we tried different partitions between training and validation sets, looking for a good representation of the chemical features (in the training set) that minimizes the average percentage error (in the validation set).

4. Results and discussion

4.1. Estimation of the glass transition temperature

Fig. 5 shows the predicted vs. measured glass transition temperature for all the molecules in the data set. The green, blue and red points correspond to training, validation, and test sets, respectively. In addition, the chemical structure of the test set compounds is also shown. We got average relative errors of 7.26% and 7.63% for validation and test sets, respectively. These errors are comparable to similar previous ANN published results [25–27] and are close to half the error obtained with linear regression models. Thus, the ANN does capture the relationship between the chemical structure and the glass transition temperature of molecular glass formers. The observed differences can be rationalized by analysing the chemical features in the training and the test sets. As shown in Table 1 in the SI, the training set contains several examples of molecules with strong intermolecular forces (with several hydrogen bond acceptors and donors in specific molecules), aliphatic cycles, and stereochemistry close to sucrose, lyxose, and trehalose. A somewhat similar situation is observed for sorbitol, where the structure and presence of OH groups are also well represented in the training set (especially in xylitol and meglumine). As a result, the ANN can correctly learn the structure-glass transition temperature relationship of these compounds from sucrose benzoate, galactose, fructose, salicin, xylose, halothane, lactose, meglumine, and ribose (see Fig. 2 in SI).

4.2. Estimation of the temperature dependence of the relaxation times

Figs. 6 and 7 show the temperature dependence of the relaxation times for sucrose, lyxose, salol, trehalose, and sorbitol. Dots represent the experimental values, as measured by BDS and reported elsewhere [28–31], while shaded bands represent the range of relaxation times obtained by ECNLE theory (from ANN’s predicted T_g values, including error bands for $T_g \pm 8\%$ corresponding to the average percentage error on the validation set). As shown, the experimental observations are in these cases inside (or very close to) the predicted relaxation region having an excellent agreement for sucrose, lyxose, salol, sorbitol, and trehalose. It is worth reminding here that the only input to the joint

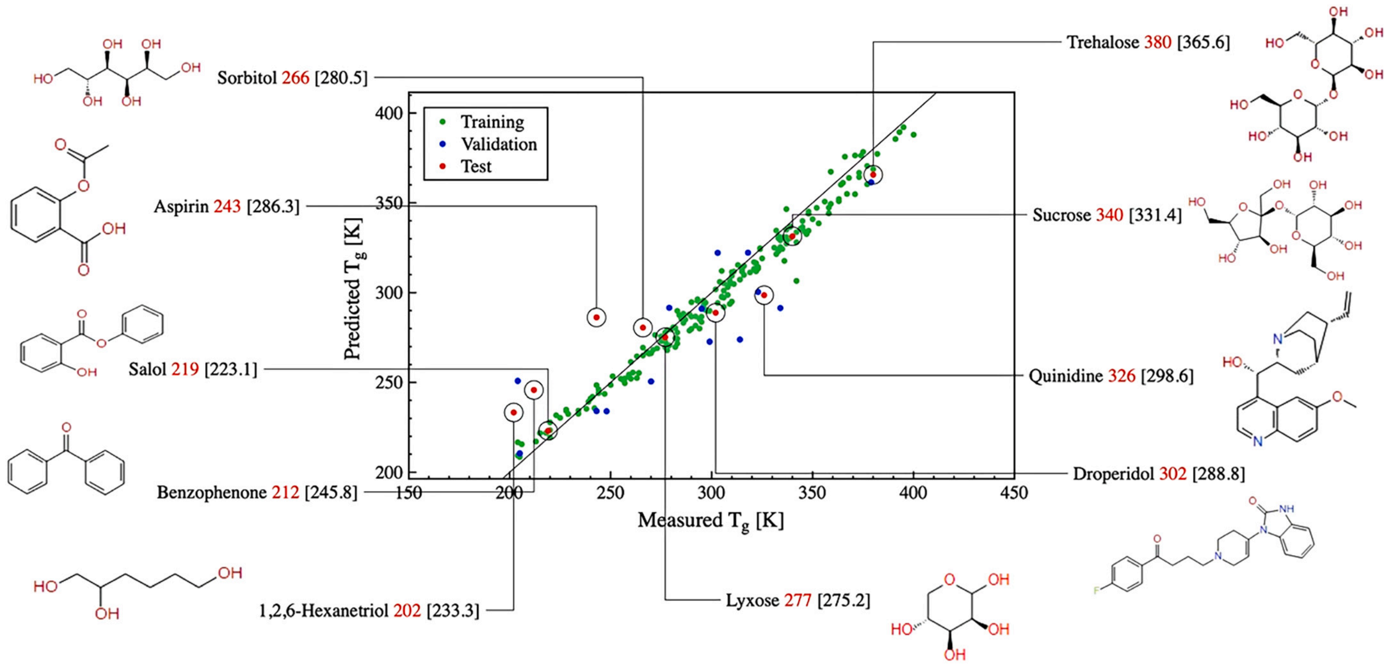


Fig. 5. Predicted vs. measured glass transition temperature for all the molecules in the data set. The green, blue and red points correspond to training, validation, and test sets, respectively. Experimental (in red) and predicted (in brackets) glass transition temperatures are indicated (in kelvin) for all the molecular glass formers in the test set. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

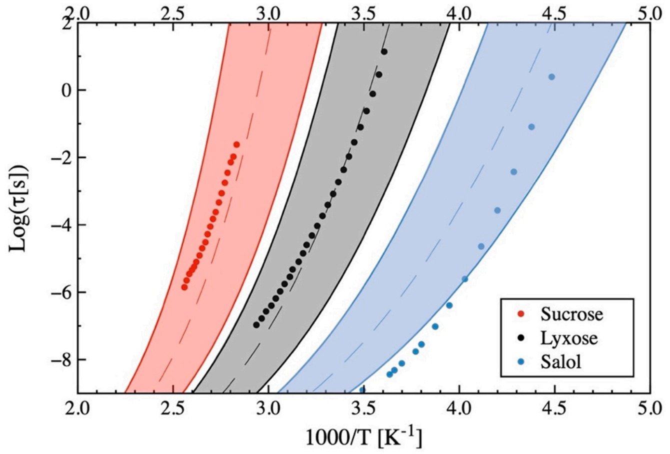


Fig. 6. Relaxation map for sucrose, lyxose and salol. Dots represent experimental values measured by BDS taken from references 28–30. Dashed lines represent the ECNLE prediction for the temperature dependence of the relaxation times based on the glass transition temperature estimated from the ANN. Shaded bands indicate the range of relaxation times as predicted by ECNLE theory based on the prediction error of the ANN (T_g ± 8%).

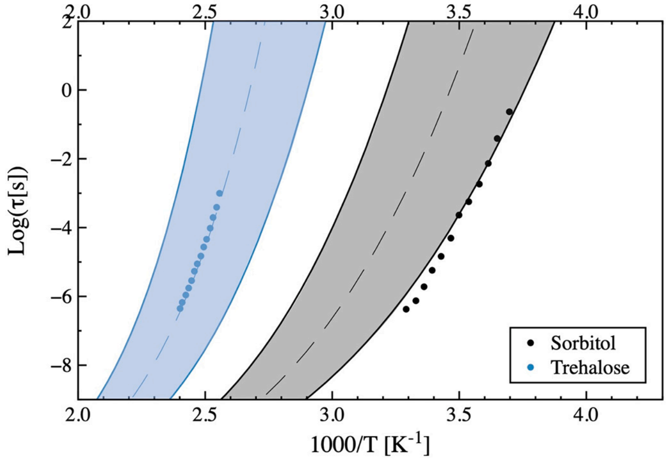


Fig. 7. Relaxation map for sorbitol and trehalose. Dots represent experimental values measured by BDS taken from references 31 and 29. Dashed lines represent the ECNLE prediction for the temperature dependence of the relaxation times based on the glass transition temperature estimated from the ANN. Shaded bands indicate the range of relaxation times as predicted by ECNLE theory based on the prediction error of the ANN (T_g ± 8%).

numerical-theoretical approach we propose in this work is the chemical structure represented as a SMILES code. Even if the molecule has not been synthesized yet, we can still have a good estimation of its glass transition temperature and the temperature dependence of its main relaxation time.

For specific compounds, like hexanetriol (see Fig. 3 in SI) or benzophenone, some deviations between estimated and experimental dynamics are observed. These differences may arise from two sources. On the one hand, due to the complexity and the variability of the molecular structures of the data set, it becomes difficult to obtain an excellent generalization from the ANN: we have to consider that mapping the chemical features during the ANN training determines the chemical structure-glass transition temperature relationship for each compound. Intuitively, we can see that chemical features better represented in the training set are more likely to be accurately mapped (see Fig. 2 in SI), and therefore, the corresponding T_g is better predicted. On the contrary, molecules with non-common features will be underrepresented, so their estimated glass transition temperature will likely present higher uncertainties. Therefore, it is expected that the ANN precision will further improve when more examples (that appropriately represent the chemical features observed in the test set) are added to the training set.

On the other hand, in some cases (see salol in Fig. 6), the glass transition temperature is well predicted by the ANN, but the estimated dynamics (the temperature dependence of the relaxation time) deviates from the experimental values. This behaviour is most likely related to the assumption that local and collective dynamics (in the ECNLE model) correlate to each other in the same way for all materials. As shown in Fig. 3, the coupling between the two dynamics (parameter *a*) depends on T_g, and therefore, predicting the dynamics for those compounds with lower values of T_g is less accurate. These deviations can be corrected by considering non-universal effects of chemical and biological complexities (as previously discussed).

In this sense, it is noteworthy that trehalose, with a relatively high T_g (380 K), shows excellent agreement between the experimental and predicted values for the temperature dependence of the relaxation times. For intermediate T_g values, like lyxose (277 K), the predicted temperature dependence of the relaxation times slightly deviates from the experimental values (provided the T_g is well predicted). On the low T_g side, we have the case of salol (219 K), for which the T_g is correctly estimated, but the dynamics departs from the experimental points upon increasing temperature. These differences in the temperature dependence of the relaxation times are expected due to different molecular structures, glass transition temperature, and the different number of H bond donors (OHs in the structure) and acceptors (oxygen atoms in the structure) in the studied compounds [32].

Fig. 8 shows some corrected dynamics calculated, including non-universal effects (a $\neq$ 1) of chemical and biological complexities in molecules like sucrose, lyxose, and salol. According to Fig. 3, we took *a* = 1.37, 1.96, and 2.5 for sucrose, lyxose, and salol, respectively. Theoretical curves (with the corresponding values of the adjustable parameter) are now closer to the experimental data. We expect that a better understanding of the dependence of this parameter on chemical structure or glass transition temperature further improves the theoretical predictions of the temperature dependence of the relaxation times.

It is important to discuss here some limitations of the proposed approach. For new materials or those without experimental data of $\tau_\alpha(T)$, the parameter *a* cannot be determined. To zeroth-order approximation, we use a linear function to empirically describe the *a*-T_g relation as shown in Fig. 3. Thus, combining the T_g value predicted from the chemical structure and ANN network with the empirical *a*-T_g relation allows us to determine (through ECNLE) $\tau_\alpha(T)$ without any adjustable/fit parameter. However, the linear *a*-T_g fit means that a given T_g leads to one value of *a* or dynamic fragility. As a result, our approach deduces that glass-forming materials having the same T_g have the same fragility, and this is not necessarily the case as shown in previous publications [33,34]. A good option to overcome this limitation is to use the ANN not only to predict the T_g, but also the fragility. We tried this approach, but unfortunately, there is a lack of data for the fragility in the literature, that makes it highly inaccurate. We expect that the available amount of data will increase in the next years, allowing the use of ANNs for predicting fragility.

Besides estimating the glass transition temperature, the ANN can also provide a new understanding of molecular glass formers. For instance, Fig. 9 shows the predicted T_g for isomers of lyxose and galactose (except for these two, we could not find the corresponding experimental T_g values for the rest of the molecules in the scientific literature). It is interesting to note that for L-Arabinose and beta-L-Arabinose, which only differ on the position of the upper right OH group, the T_g only changes three degrees. However, the same structural change between galactose and alpha-D-Galactose gives a T_g difference of 33 K. In this case, the presence of the upper left group induces a higher sensitivity of the dynamics to minor structural changes. It is worth mentioning that although the average relative error of the ANN's prediction is about 8%, in the case of lyxose and galactose, the corresponding errors are below 1% (see Table 1 in SI), making sense of the observed differences in Fig. 9. The same analysis can be performed on molecules not even synthesized, boosting the development of new materials with tuned properties.

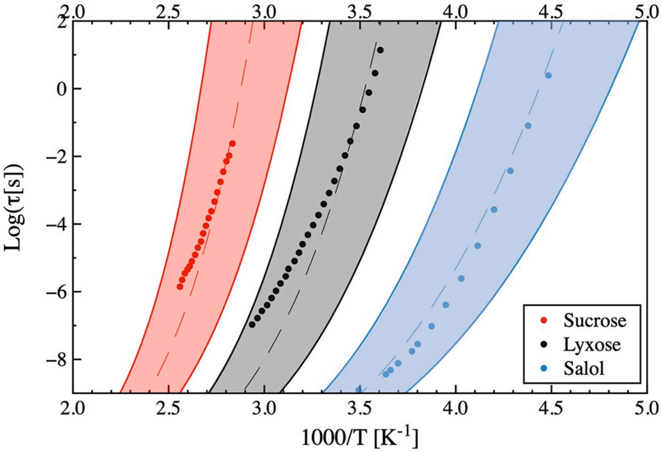


Fig. 8. Relaxation map for sucrose, lyxose and salol. Dots represent experimental values measured by BDS taken from references 28–30, and, respectively. Dashed lines represent the ECNLE prediction for the temperature dependence of the relaxation times based on the glass transition temperature estimated from the ANN and corrected with the adjustable parameter *a* as taken from Fig. 3. Shaded bands indicate the range of relaxation times predicted by the joint numerical/theoretical approach proposed here.

5. Conclusions

We have presented in this work a new approach that combines numerical methods with theory to estimate the temperature dependence of the structural relaxation time for molecular glass formers. Firstly, we built, optimized and trained an artificial neural network to assess the glass transition temperature of molecular glass formers only based on their chemical structure. Then, we used a theoretical approach based on the elastically collective nonlinear Langevin equation to estimate the full relaxation map. Although there is still some room to improve accuracy and overcome limitations, this first joint theoretical and numerical approach constitutes a suitable tool for giving a reasonable estimation of the dynamics of unknown molecular glass formers based on their chemical structure. This approach will boost materials and drug development by designing molecular glass formers with desired properties and will also increase the understanding of the physical mechanisms related to molecular dynamics.

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 author statement

Author 1: Claudia Borredon.

Collected the data; Contributed data or analysis tools; Performed the analysis; Wrote the paper.

Author 2: Luis A. Miccio.

Conceived and designed the analysis; Contributed data or analysis tools; Wrote the paper.

Author 3: Anh D. Phan.

Contributed data or analysis tools; Wrote the paper.

Author 4: Gustavo A. Schwartz.

Conceived and designed the analysis; Contributed data or analysis tools; Wrote the paper; other contributions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

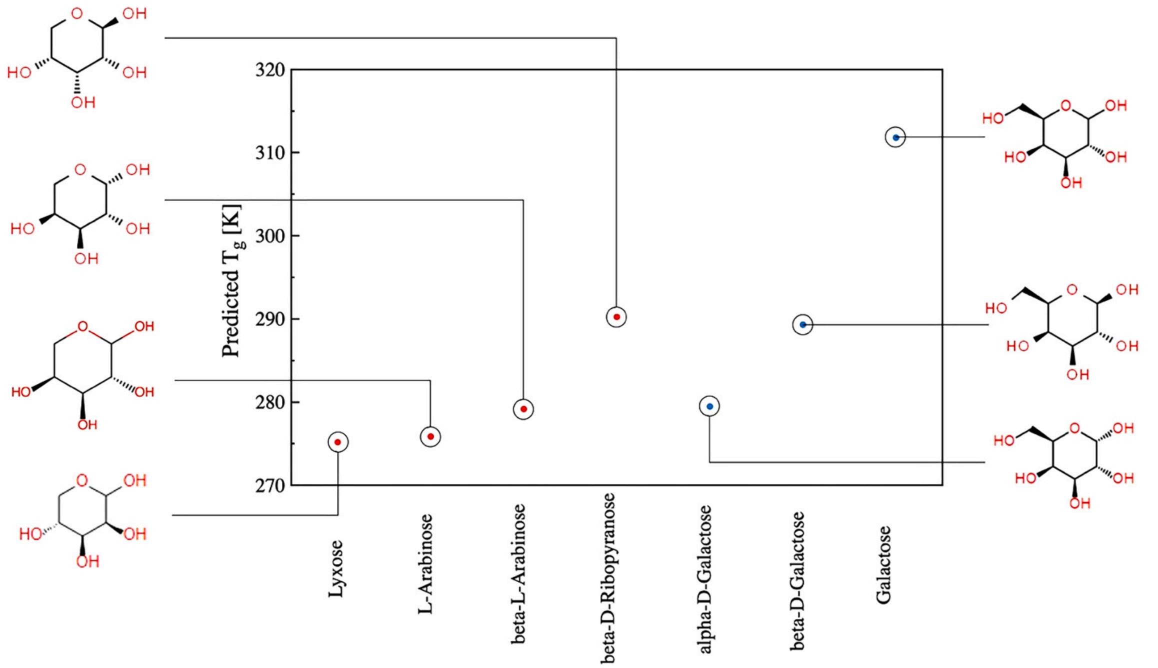


Fig. 9. Predicted glass transition temperature for isomers of lyxose (red) and galactose (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the work reported in this paper.

Acknowledgments

We gratefully acknowledge the financial support from the Spanish Government "Ministerio de Ciencia e Innovación" (PID2019-104650GB-C21) and the Basque Government (IT-1566-22). We also acknowledge support of the publication fee by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URICI). This research was also funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 103.01-2019.318. We also acknowledge the support of NVIDIA Corporation with the donation of two GPUs used for this research.

Appendix A. Supplementary data

Additional information on the glass formers, their glass transition temperatures, chemical structures, and SMILES code can be found in the Supporting Information file (SI). Supplementary data to this article can be found online at https://doi.org/10.1016/j.nocx.2022.100106.

References

[1] A.D. Phan, J. Knapik-Kowalczuk, M. Paluch, T.X. Hoang, K. Wakabayashi, Theoretical model for the structural relaxation time in coamorphous drugs, Mol. Pharm. 16 (2019) 2992–2998, https://doi.org/10.1021/acs.molpharmaceut.9b00230.

[2] W. Tu, J. Knapik-Kowalczuk, K.C. Molecular, Glass transition dynamics and physical stability of amorphous Griseofulvin in binary mixtures with low-Tg excipients, Chem. Rev. (2019), https://doi.org/10.1021/acs.molpharmaceut.9b00476.

[3] K. Grzybowska, A.G. Molecular, Molecular dynamics and physical stability of ibuprofen in binary mixtures with an acetylated derivative of maltose, Chem. Rev. (2020), https://doi.org/10.1021/acs.molpharmaceut.0c00517.

[4] A.D. Phan, Determination of Young's Modulus of Active Pharmaceutical Ingredients by Relaxation Dynamics at Elevated Pressures, 2020, pp. 1–7, https://doi.org/10.1021/acs.jpcb.0c05523.

[5] A.D. Phan, K.S. Schweizer, Elastically collective nonlinear Langevin equation theory of glass-forming liquids: transient localization, thermodynamic mapping, and cooperativity, J. Phys. Chem. B 122 (2018) 8451–8461, https://doi.org/10.1021/acs.jpcb.8b04975.

[6] L.A. Miccio, G.A. Schwartz, Localizing and quantifying the intra-monomer contributions to the glass transition temperature using artificial neural networks, Polymer 203 (2020), 122786, https://doi.org/10.1016/j.polymer.2020.122786.

[7] L.A. Miccio, G.A. Schwartz, Mapping chemical structure–glass transition temperature relationship through artificial intelligence, Macromolecules 54 (2021) 1811–1817, https://doi.org/10.1021/acs.macromol.0c02594.

[8] L.A. Miccio, G.A. Schwartz, From chemical structure to quantitative polymer properties prediction through convolutional neural networks, Polymer 193 (2020), 122341, https://doi.org/10.1016/j.polymer.2020.122341.

[9] A.D. Phan, K. Wakabayashi, Theory of structural and secondary relaxation in amorphous drugs under compression, Pharm 12 (2020) 177, https://doi.org/10.3390/pharmaceutics12020177.

[10] A.D. Phan, T.T.T. Thuy, N.T.K. An, J. Knapik-Kowalczuk, M. Paluch, K. Wakabayashi, Molecular relaxations in Supercooled liquid and glassy states of amorphous Gambogic acid: dielectric spectroscopy, calorimetry, and theoretical approach, AIP Adv. 10 (2020), 025128, https://doi.org/10.1063/1.5139101.

[11] A.D. Phan, A. Jedrzejowska, M. Paluch, K. Wakabayashi, Theoretical and experimental study of compression effects on structural relaxation of glass-forming liquids, Acs Omega 5 (2020) 11035–11042, https://doi.org/10.1021/acsomega.0c00860.

[12] A.D. Phan, K. Wakabayashi, M. Paluch, V.D. Lam, Effects of cooling rate on structural relaxation in amorphous drugs: elastically collective nonlinear Langevin equation theory and machine learning study, RSC Adv. 9 (2019) 40214–40221, https://doi.org/10.1039/c9ra08441j.

[13] S. Mirigian, K.S. Schweizer, Elastically cooperative activated barrier hopping theory of relaxation in viscous fluids. II. Thermal liquids, J. Chem. Phys. 140 (2014), 194507, https://doi.org/10.1063/1.4874843.

[14] S. Mirigian, K.S. Schweizer, Unified theory of activated relaxation in liquids over 14 decades in time, J. Phys. Chem. Lett. 4 (2013) 3648–3653, https://doi.org/10.1021/jz4018943.

[15] S. Mirigian, K.S. Schweizer, Elastically cooperative activated barrier hopping theory of relaxation in viscous fluids. I. General formulation and application to hard sphere fluids, J. Chem. Phys. 140 (2014), 194506, https://doi.org/10.1063/1.4874842.

[16] L.D. Landau, E.M. Lifshitz, Theory of Elasticity, Permagon Press, London, 1975. ISBN 9780080570693.

[17] S.-J. Xie, K.S. Schweizer, Nonuniversal coupling of cage scale hopping and collective elastic distortion as the origin of dynamic fragility diversity in glass-forming polymer liquids, Macromolecules 49 (2016) 9655–9664, https://doi.org/10.1021/acs.macromol.6b02272.

[18] N.M. O'Boyle, Towards a universal SMILES representation - a standard method to generate canonical SMILES based on the InChI, Aust. J. Chem. 4 (2012) 22, https://doi.org/10.1186/1758-2946-4-22.

[19] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Model. 28 (1988) 31–36, https://doi.org/10.1021/ci00057a005.

[20] C. Nwankpa, W. Ijomah, A. Gachagan, S. Marshall, Activation functions: comparison of trends in practice and research for deep learning, 2nd International Conference on Computational Sciences and Technology, Jamshoro, Pakistan. (2021) 124–133.

[21] D.-A. Clevert, T. Unterthiner, S. Hochreiter, Fast and accurate deep network learning by exponential linear units (ELUs), 4th International Conference on Learning Representations (2016) 1–14.

[22] D.P. Kingma, Jimmy Ba, Adam: a method for stochastic optimization, 3rd International Conference on Learning Representations (2015) 1–14.

[23] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics *9*, **2010**, pp. 249–256.

[24] D. Hendrycks, K. Gimpel, Adjusting for dropout variance in batch normalization and weight initialization, Workshop track - 5th International Conference on Learning Representations (2017) 1–10.

[25] W. Liu, 2010 prediction of glass transition temperatures of aromatic heterocyclic polyimides using an ANN model, Wiley Online Library 50 (2010) 1547–1557, https://doi.org/10.1002/pen.21670.

[26] L. Ning, Artificial neural network prediction of glass transition temperature of fluorine-containing polybenzoxazoles, J. Mater. Sci. 44 (2009) 3156–3164, https://doi.org/10.1007/s10853-009-3420-0.

[27] W. Liu, C. Cao, Artificial neural network prediction of glass transition temperature of polymers, Colloid Polym. Sci. 287 (2009) 811–818, https://doi.org/10.1007/s00396-009-2035-y.

[28] P.S. Lokendra, A. Alegría, J. Colmenero, Broadband dielectric spectroscopy and calorimetric investigations of D-Lyxose, Carbohydr. Res. 346 (2011) 2165, https://doi.org/10.1016/j.carres.2011.06.029.

[29] K. Kaminski, K. Adrjanowicz, D.Z. Molecular, Dielectric studies on molecular dynamics of two important disaccharides: sucrose and trehalose, Chem. Rev. (2012), https://doi.org/10.1021/mp2004498.

[30] J. Bartoš, M. Iskrová, M. Köhler, R. Wehn, O. Šauša, P. Lunkenheimer, J. Krištiak, A. Loidl, Positron annihilation response and broadband dielectric spectroscopy: salol, European Phys J E 34 (2011) 104, https://doi.org/10.1140/epje/i2011-11104-x.

[31] A. Faivre, G. Niquet, M. Maglione, J. Fornazero, J.F. Jal, L. David, Dynamics of sorbitol and maltitol over a wide time-temperature range, The Eur. Phys. J. B - Condensed Matter Complex Syst. 10 (1999) 277–286, https://doi.org/10.1007/s100510050856.

[32] Q. Qian, Gregory B. McKenna, Gregory correlation between dynamic fragility and glass transition temperature for different classes of glass forming liquids, J. Non-Cryst. Solids 352 (2006) 2977, https://doi.org/10.1016/j.jnoncrysol.2006.04.014.

[33] K. Kunal, C.G. Robertson, S. Pawlus, S.F. Hahn, A.P. Sokolov, Role of chemical structure in fragility of polymers: a qualitative picture, Macromolecules 41 (2008) 7232–7238, https://doi.org/10.1021/ma801155c.

[34] W.-S. Xu, K.F. Freed, Influence of cohesive energy and chain stiffness on polymer glass formation, Macromolecules 47 (2014) 6990–6997, https://doi.org/10.1021/ma501581u.

polymers

MDPI

Article

Approaching Polymer Dynamics Combining Artificial Neural Networks and Elastically Collective Nonlinear Langevin Equation

Luis A. Miccio [1,2,3,*], Claudia Borredon [1], Ulises Casado [3], Anh D. Phan [4,5] and Gustavo A. Schwartz [1,2,*]

[1] Centro de Física de Materiales (CSIC-UPV/EHU)—Materials Physics Center (MPC), P. M. de Lardizabal 5, 20018 San Sebastian, Spain; borredon.claudia@gmail.com
[2] Donostia International Physics Center, P. M. de Lardizábal 4, 20018 San Sebastian, Spain
[3] Institute of Materials Science and Technology (INTEMA), National Research Council (CONICET), Colon 10850, Mar del Plata 7600, Argentina; ulisescasado@fi.mdp.edu.ar
[4] Faculty of Materials Science and Engineering, Phenikaa University, Hanoi 12116, Vietnam; anh.phanduc@phenikaa-uni.edu.vn
[5] Phenikaa Institute for Advanced Study (PIAS), Phenikaa University, Hanoi 12116, Vietnam
* Correspondence: lamiccio@gmail.com (L.A.M.); gustavo.schwartz@csic.es (G.A.S.)

Abstract: The analysis of structural relaxation dynamics of polymers gives an insight into their mechanical properties, whose characterization is used to qualify a given material for its practical scope. The dynamics are usually expressed in terms of the temperature dependence of the relaxation time, which is only available through time-consuming experimental processes following polymer synthesis. However, it would be advantageous to estimate their dynamics before synthesizing them when designing new materials. In this work, we propose a combined approach of artificial neural networks and the elastically collective nonlinear Langevin equation (ECNLE) to estimate the temperature dependence of the main structural relaxation time of polymers based only on the knowledge of the chemical structure of the corresponding monomer.

Keywords: QSPR; dynamics prediction; polymers; artificial neural networks; smart design

check for updates

Citation: Miccio, L.A.; Borredon, C.; Casado, U.; Phan, A.D.; Schwartz, G.A. Approaching Polymer Dynamics Combining Artificial Neural Networks and Elastically Collective Nonlinear Langevin Equation. *Polymers* **2022**, *14*, 1573. https://doi.org/10.3390/polym14081573

Academic Editor: Sasaki Takashi

Received: 18 March 2022
Accepted: 11 April 2022
Published: 12 April 2022

1. Introduction

The mechanical behavior of polymeric materials is key to several industries such as aerospace, transport, energy, and construction, among many others [1–7]. Since the mechanical properties, together with the overall service life performance of these materials, are directly related to their dynamics, the knowledge of the latter becomes highly relevant. For instance, in transport and aerospace industries, some materials are expected to be able to perform well through wide ranges in terms of frequency, presenting a low rolling resistance and at the same time a large dissipation of energy during a braking period (processes that correspond to approximately 10^{-2} Hz and 10^4–10^7 Hz, respectively) [8–11]. Therefore, for obtaining the required on-service behavior, adequate polymer selection is combined with the fine-tuning of several other properties such as processability, durability, and energetic efficiency. Molecular dynamics determines such mechanical properties of the compound, and it is usually described in terms of a characteristic relaxation time and its temperature dependence. The experimental window of these relaxations (that can extend over several decades) imposes the necessity of a combination of techniques (such as broadband dielectric spectroscopy (BDS), dynamic light scattering (DLS), or dynamic mechanical analysis (DMA)), in turn converting this practice into a costly and time-consuming process that could increase development costs.

Nevertheless, some theoretical approaches can help when designing and developing new materials since there is no prior information about their dynamics before synthesizing and characterizing them. Among these approaches, the elastically collective nonlinear

Langevin equation (ECNLE) [12–14] theory was developed and successfully applied to describe the molecular dynamics of various amorphous materials. This model solely relies on the knowledge of the glass transition temperature (T_g), which requires a non-negligible amount of time and resources to be determined when unknown. However, recent advances in the field of artificial neural networks (ANN) [15–17] enable the estimation of the glass transition temperature of polymers based only on the monomer's chemical structure.

In this work, we combine theoretical and numerical approaches to estimate, from a representation of the chemical structure of amorphous acrylates, their glass transition temperature and the temperature dependence of the structural relaxation time. Firstly, we codify the chemical structure of the compounds using the Simplified Molecular Input Line Entry System (SMILES) [18,19] representation and employ it as an input for a neural network algorithm that would output an estimation of the polymer's T_g; then, we exploit this information as an input for the ECNLE to theoretically compute the trajectory of the molecular dynamics of the structural relaxation process, expressed as the temperature dependence of its relaxation time. We propose this approach as a tool to speed up research and development in the field of polymeric materials.

2. Methods and Theoretical Background

In this section, we explain the characteristics of the dataset, the process that the data undergo, the ANN's architecture, and how it is tuned. In addition, we include a description of how ECNLE theory is applied to the estimation of the acrylates' dynamics.

2.1. Dataset

We employed a cured dataset composed of about 200 atactic polyacrylates and their corresponding T_g values above chain length saturation [20–23] (see Table S1). These acrylates' monomer units were codified using a Simplified Molecular Input Line Entry System (SMILES) [18,19] and converted into binary matrices, which are then fed to the ANN.

The external control set was composed of those polymers for which the experimental dynamics was published. These data are essential since we want to compare the predicted dynamics against the experimental dynamics. Table SI2 shows the parameters of the Vogel–Fulcher–Tammann (VFT) equation that fits the corresponding observed dynamics together with the references the data were taken from.

2.2. Chemical Structure Encoding

As we proposed in recent works [15–17], to consider the structure and composition of the monomeric units, we transformed the chemical structures into linear strings using SMILES [18,19]. Then, we converted these strings into binary matrices using a one hot encoding algorithm [24] and a dictionary (composed by the following list of symbols: ‘(’, ‘O’, ‘C’, ‘=‘, ‘c’, ‘S’, ‘F’, ‘N’, ‘X’, ‘2’, ‘d’, ‘1’, ‘#’, ‘]’, ‘/’, ‘)’). Section S3 in Supplementary Materials provides a brief explanation of this encoding process.

2.3. ANN's Architecture and Optimization

We used convolutional neural networks fed with the polyacrylates' monomeric structures (codified into binary matrices) and the corresponding glass transition temperatures. Figure 1 shows a schematic view of the ANN's architecture: the monomer structure is codified (through a one-hot encoding process applied on its SMILES string) into a 2D matrix which is then fed to convolutional layers to extract relevant chemo-structural features; the result is flattened into a 1D vector ($X \in R^n$) feeding two fully connected layers (FC_0 and FC_1) with LeakyReLU activations. Section S4 in Supplementary Materials provides more details about the neural network architecture. We compared several combinations of hyperparameters to achieve the best possible performance for the ANNs. Such comparison among ANNs was based on the raw performance (minimum relative error) obtained on the dataset. A dropout [25] 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 for providing the glass transition temperature value.

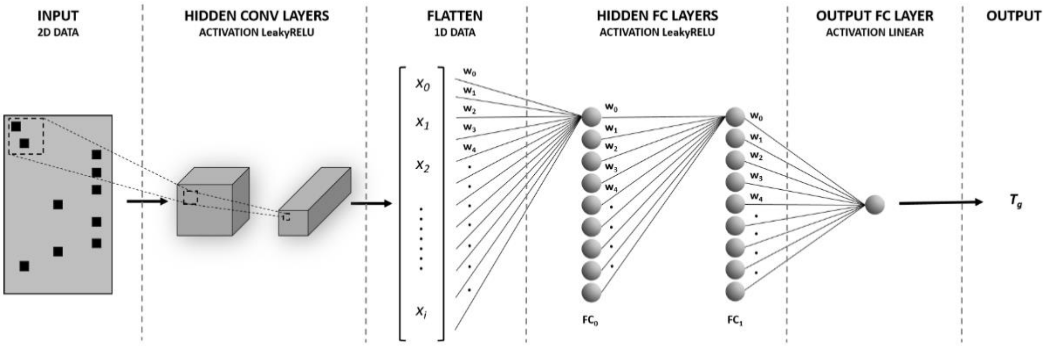


Figure 1. Schematic picture of the artificial neural network employed for predicting the glass transition temperatures of acrylates.

As done in previous works [15–17], we implement the mean absolute relative error as a loss function in the training process to ensure equal weighting of low and high T_g data values. Given E_i (experimental T_g), F_i (forecasted T_g), and the number of acrylates in one mini-batch m_x, we define the mean absolute percentage error as

$$Loss = \frac{100}{m_x} \cdot \sum_{i=1}^{m_x} \left| \frac{E_i - F_i}{E_i} \right| \tag{1}$$

We adopt a mini-batch gradient descent technique to minimize the loss function, using an Adam optimizer [26] with a learning rate (lr) of 0.0001 for speeding up the convergence and mini-batches of 20 acrylates each.

As usual, the data were randomly divided into test and train subsets during the training process, and no enforcement of any preference in the way the data are split was applied. In addition, an external control group (independent from the previous subsets) was formed for studying polymer dynamics through ECNLE theory. ANN details are summarized in Table 1 and Figure 1 (more details are provided in Section S4 of the SI).

Table 1. ANN hyperparameters.

Item	Value
Data split ratio (train/test)	80/20
Dropout probability	0 to 0.3
Mini batch size	20
Learning rate	0.0001
Beta1 (Beta2)	0.99 (0.999)
# Hidden neurons (FC0–FC1)	30–20

2.4. Nonlinear Langevin Equation

ECNLE theory describes glass-forming liquids using a hard-sphere fluid [12–14] of volume fraction $\Phi = \rho\pi d^3/6$, where d is the particle size and ρ is the number of particles per volume. The local dynamics takes account of a tagged particle considering: (1) interactions with its nearest neighbors, and (2) cooperative motions of particles beyond the first shell. The dynamics is quantified by the dynamic free energy [12–14], $F_{dyn}(r) = F_{ideal}(r) + F_{caging}(r)$, where r is the displacement, $F_{ideal}(r)$ represents the ideal fluid dynamics and $F_{caging}(r)$ characterizes the local state of a particle subject to caging forces conditioned by the structural features of the system. When the fluid has a sufficiently large density ($\Phi \geq 0.432$) or is in a low enough temperature, the motion of particles is restricted within a particle cage of radius r_{cage} and a barrier in $F_{dyn}(r)$ emerges with a barrier height given by

$F_B = F_{dyn}(r_B) - F_{dyn}(r_L)$, where r_L is the localization length of the particle and r_B is the barrier position. The escaping of a particle from its cage produces a collective elastic long-range rearrangement of the molecules in the fluid, whose energy contribution is given by a sum over harmonic oscillators which is described in Section S5 of the SI. Once the local and elastic dynamics are defined and the harmonic curvatures at r_B and r_L (respectively K_0 and K_B, see SI) is estimated, we calculate the structural relaxation time using Kramer's theory

$$\frac{\tau}{\tau_s} = 1 + \frac{2\pi}{\sqrt{K_0 K_B}} \frac{k_B T}{d^2} exp\left(\frac{F_B + F_e}{k_B T}\right) \tag{2}$$

where τ_s is a short time scale [12–14]. As the above calculations provide $\tau(\Phi)$, we use a simple thermal mapping $T = T_g + \left(\Phi_g - \Phi\right)/\beta\Phi_0$, where T_g is the dynamic glass transition temperature defined by $\tau\left(T_g\right) = 100$ s, Φ_g is the volume fraction when $\tau\left(\Phi_g \approx 0.6157\right) = 100$ s, $\Phi_0 \approx 0.5$ is a characteristic volume fraction, and $\beta \approx 12 \times 10^{-4}$ K^{-1} is an effective thermal expansion coefficient considered constant for all amorphous materials. Further details to derive the theory are given in the Supplementary Materials and elsewhere [12–14].

3. Discussion

Figure 2 shows predicted vs. experimental values of the glass transition temperature for the external control set of polyacrylates, as obtained with our trained ANN (see also Figure S2 for the training and internal test sets). We obtained mean absolute percentage errors of 4.3% (training set), 8.5% (validation set), and 4.5% (control set). In comparison with other neural network approaches that we have used in the past [15], the relative number of parameters (and, therefore, calculations) is reduced thanks to a convolutional approach (due to the stride convolution operation that tosses out parts of the input image). It is worth remembering here that we are feeding the ANN only with the monomer chemical structure without any other physical or chemical input data (neither measured nor calculated).

As shown, the ANN does capture the relationship between the chemical structure and the glass transition temperature of the polyacrylates all along the 200–400 K range (see also Figure S2). The individual relative deviations in the external control group are within (or close to) a 10% margin (see Table S1), in agreement with the observed values for the internal test. More details on the obtained relative deviation for the different chemical structures are depicted in Figure S3. Aside from the obtained low errors, our aim is not only to predict the T_g, but also to obtain some insight into the dynamics of the polymers under study. For this purpose, the predicted glass transition temperatures are used as input for ECNLE theory, thus creating a hybrid ANN-theory approach for yielding a possible relaxation area (in terms of log (τ) vs. 1000/T).

Hence, Figure 3 shows the temperature dependence of the alpha relaxation times for (a) Poly (propyl methacrylate), (b) Poly (phenyl methacrylate), (c) Poly (butyl methacrylate), and (d) Poly (isopropyl methacrylate). Blue lines represent the experimental values, reported elsewhere [27–31], while dashed lines represent the range of relaxation times obtained by ECNLE theory (from ANN's predicted T_g values), including error bands for T_g ± 10% (corresponding to the maximum relative error on the external control set). As shown, the predicted relaxation region is very close to the experimental observations, having, therefore, an acceptable agreement (especially considering that only the chemical structure of the monomer is used as input).

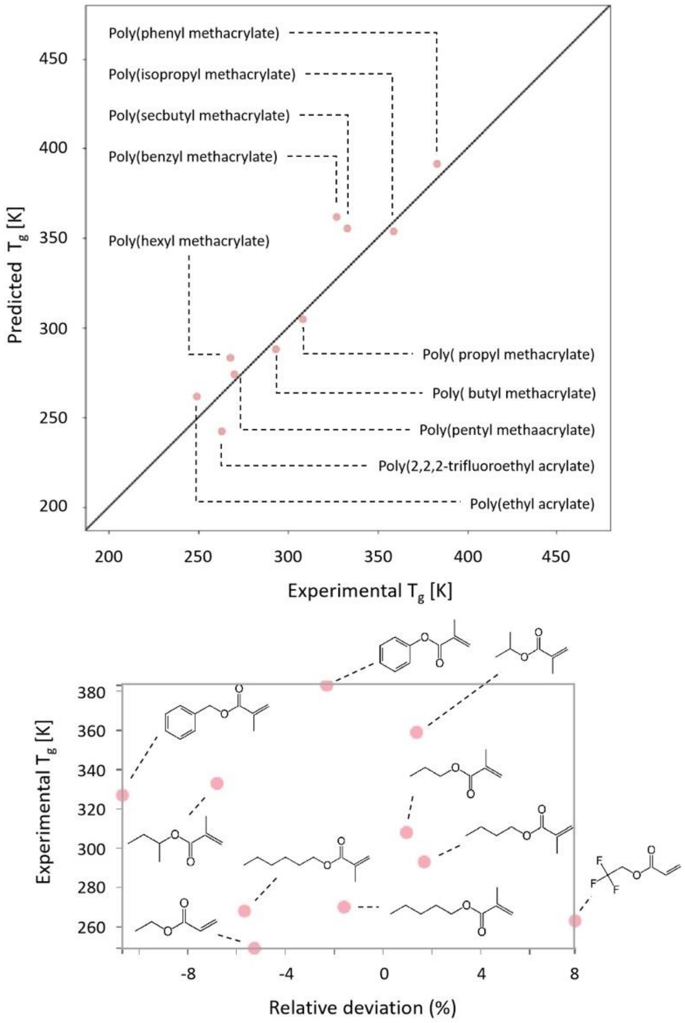


Figure 2. Predicted vs. experimental glass transition values obtained from the trained ANN on the external control group of acrylates. Relative deviations are shown below with the corresponding monomeric chemical structures.

In some cases, as for poly (phenyl methacrylate) or poly (isopropyl methacrylate) (see Figure 3b,d), the glass transition temperature is well predicted, but the curvature of the estimated dynamics deviates from the experimental values. In some other cases, as in Figure 3a,c, the deviations are even more pronounced. Therefore, despite being inside the proposed confidence interval, the curvature obtained from ECNLE theory does not follow the experimental dynamics. This behavior is most likely related to the assumption that local and collective dynamics correlate to each other for all materials in the same way (which is an excellent approach in terms of not needing any other inputs to obtain an approximated relaxation map but tends to oversimplify the behavior of the materials). In particular, local and collective dynamics in Equation (2) are summed with equal weights (i.e., the ratio of prefactor equal to 1). It has been shown [32,33] that ECNLE calculations gain accuracy by weighting the collective elastic contribution with a parameter $a \neq 1$, to change its relative importance in the glass transition process. The new adjusted elastic barrier is $F_e \rightarrow a^2 F_e$ and it modifies the structural relaxation time in Equation (2) as

$$\frac{\tau}{\tau_s} = 1 + \frac{2\pi}{\sqrt{K_0 K_B}} \frac{k_B T}{d^2} exp\left(\frac{F_B + a^2 F_e}{k_B T}\right) \tag{3}$$

The parameter a strongly influences the structural features of the model (value of Φ_g and the thermal mapping), as it accounts for the non-universal effects on the collective motions of molecules due to conformational and chemical complexities. It was empirically observed that the T_g is typically inversely proportional to the scaling parameter a [13]. Figure S4 shows the glass transition temperature dependence of the model adjustable parameter a for several polymers and glass formers. Although the correlation is not strong, there is a clear trend indicating an increment of the parameter a upon decreasing glass transition temperature. Thus, we can estimate the scaling parameter a based on the T_g.

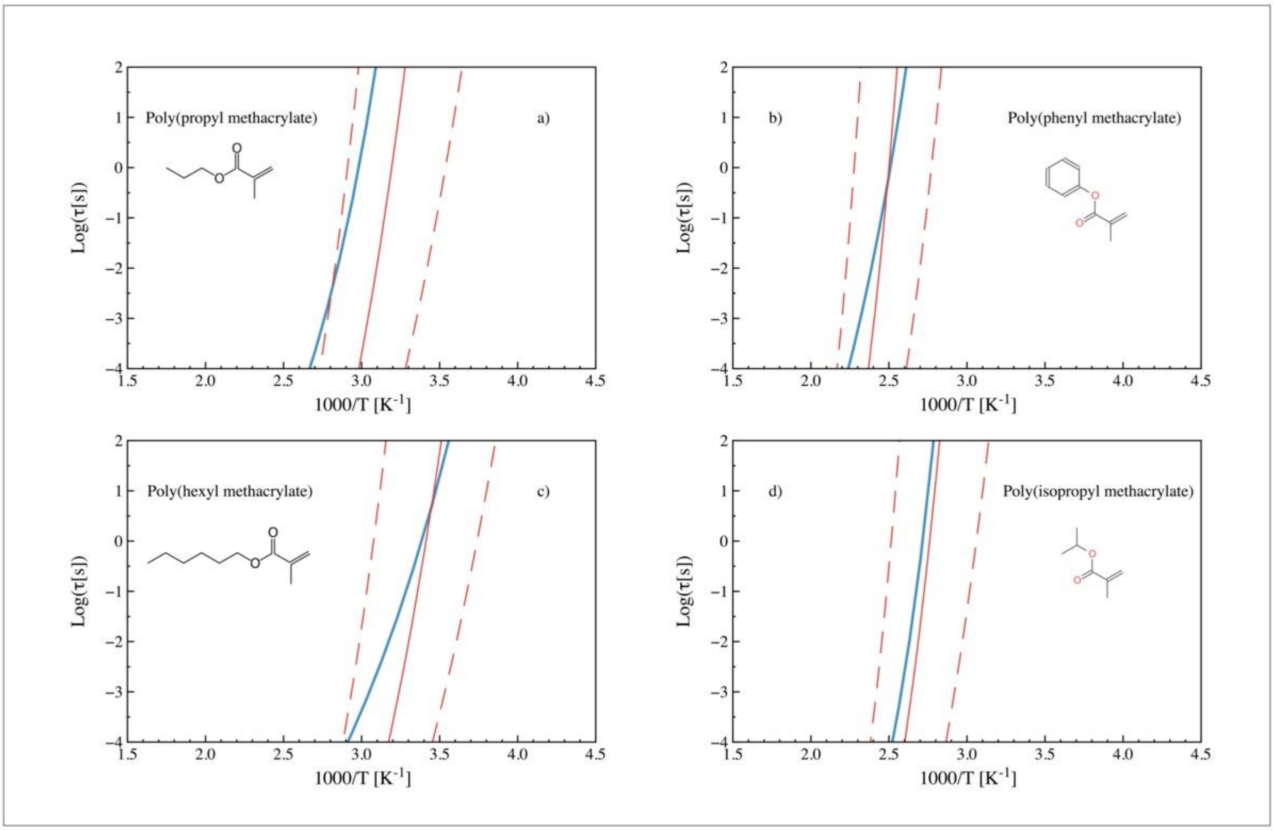


Figure 3. Experimental (blue) and predicted (red) relaxation times (obtained from ECNLE theory) vs. 1000/T. Dashed lines stand for the confidence interval corresponding to the typical deviation in the ANN prediction (10% relative error): (**a**) Poly (propyl methacrylate) [28], (**b**) poly (phenyl methacrylate) [30], (**c**) poly (hexyl methacrylate) [27], and (**d**) poly (isopropyl methacrylate) [30].

Figure 4 shows the temperature dependence of alpha relaxation times for the same polymers as Figure 3 after introducing the scaling parameter (a). The predicted relaxation times change their curvature, displaying a better agreement (for cases b and d) with the experimental observations. In the case of poly (propyl methacrylate), no further improvement is perceived. It is also observed that, in the case of polymers with linear alkane tails, the experimental-predicted agreement appears to decrease as the length of the tail increases. As shown in Figure 5 (b) poly (propyl methacrylate) and (c) poly (butyl methacrylate) already reflect this trend, which intensifies for (d) poly (pentyl methacrylate) and (e) poly (hexyl methacrylate), while it is much smaller for (a) poly (ethyl acrylate).

Fragilities and dynamics data of members of the polyacrylates family have been obtained from mechanical and dielectric data by several authors [33–42]. From this experimental point of view, the increase in the length of the alkyl chain causes a strengthening effect. The variation of fragility (m) with the length of the alkyl chain appears to have three ranges: for less than three atoms, m is nearly constant; between three and five atoms, it drastically decreases; and, for more than five atoms, m slowly decreases. Moreover, Balabin studied the enthalpy difference between conformations of normal alkanes and showed that n-alkyl chains are more and more flexible as the chain length increases [43]. In addition, some local order structure gradually develops as the carbon number in the side chain increases due to a self-assembly process that forms supramolecular systems such as “hairy rods” [44,45].

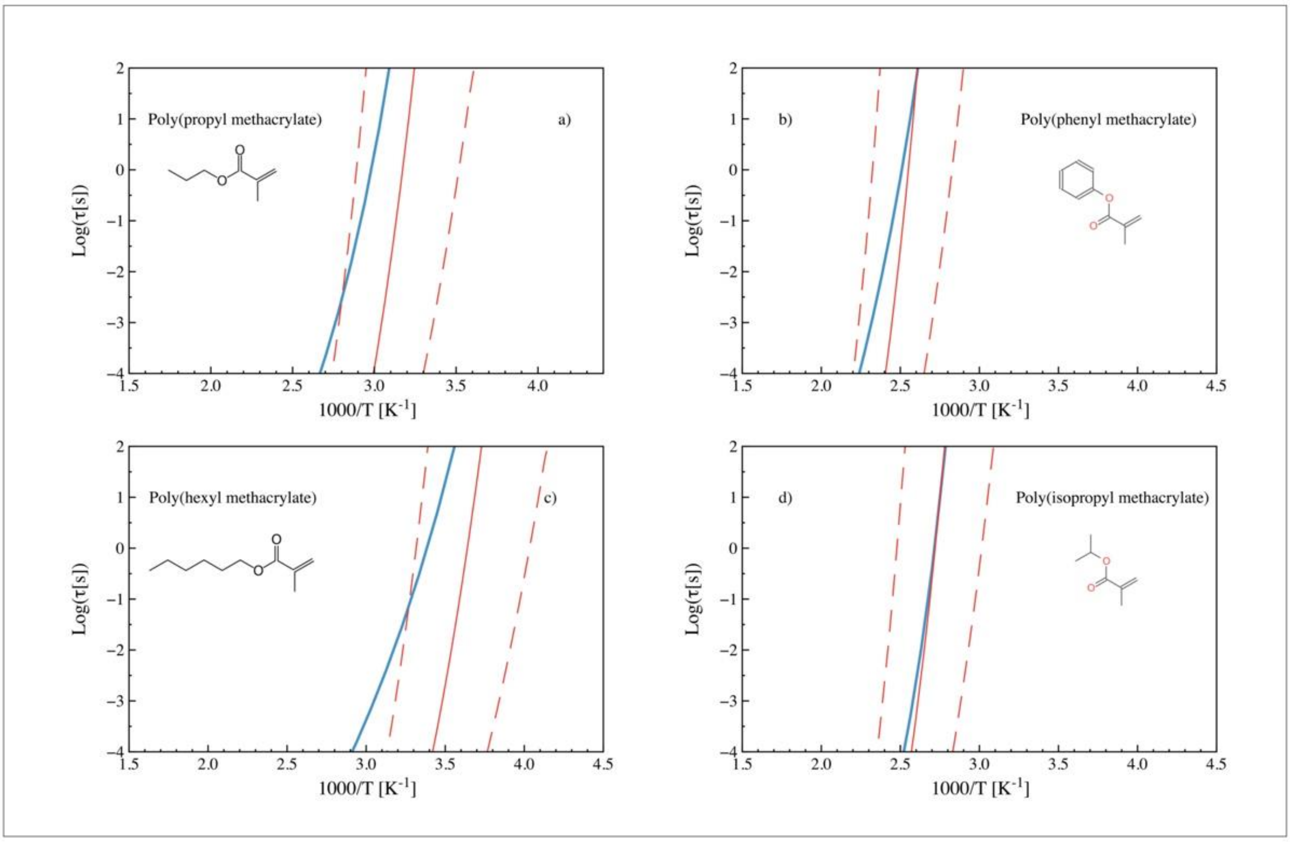


Figure 4. Experimental (blue) and predicted (red) relaxation times (obtained from ECNLE theory) vs. 1000/T after introducing the scaling parameter (*a*). Dashed lines stand for the confidence interval corresponding to the typical deviation in the ANN prediction (10% relative error). (**a**) Poly (propyl methacrylate), (**b**) poly (phenyl methacrylate), (**c**) poly (hexyl methacrylate) and (**d**) poly (isopropyl methacrylate).

Finally, it has also been reported that nanophase separation of incompatible main and side-chain parts occurs in amorphous side-chain polymers with long alkyl groups (for polymers with 4 or more C atoms in the side chain) [46–49]. Considering that the cooperative dynamics changes if the confinement size becomes comparable to the size of cooperatively rearranging regions (CRRs), these crystalline regions could affect the relaxation, thus creating a hindered glass transition [48]. Published results indicate that the CRR size for alkyl sequences is in the range of one nanometer [50–52].

A more detailed view of this effect on the prediction differences with the experimental data can be observed in Figure 5, where the relaxation maps of a series of alkyl-acrylates are presented. As shown, the predictions progressively deviate from the experimental curves as the side-chain length increases. Deviations in polymers with two or three atoms in the tail are almost exclusively related to deviations in the T_g predicted by the ANN, while for longer chains, a difference in the predicted curvature is additionally noticed. It can be argued that the proposed approach yields acceptable predictions up to four or five atoms in the linear chain.

These predicted and experimental results can be reconciled by considering the ECNLE theory assumptions, which predicts the material dynamics in terms of a fluid composed of hard spheres and does not consider other processes (such as packing density, induced crystallization or nanophase separation). Therefore some deviations are expected from the experimental observations in these polymers where other processes occur. These deviations are related to the typical relaxation length of the alpha relaxation, which is in the nanometer range for these materials.

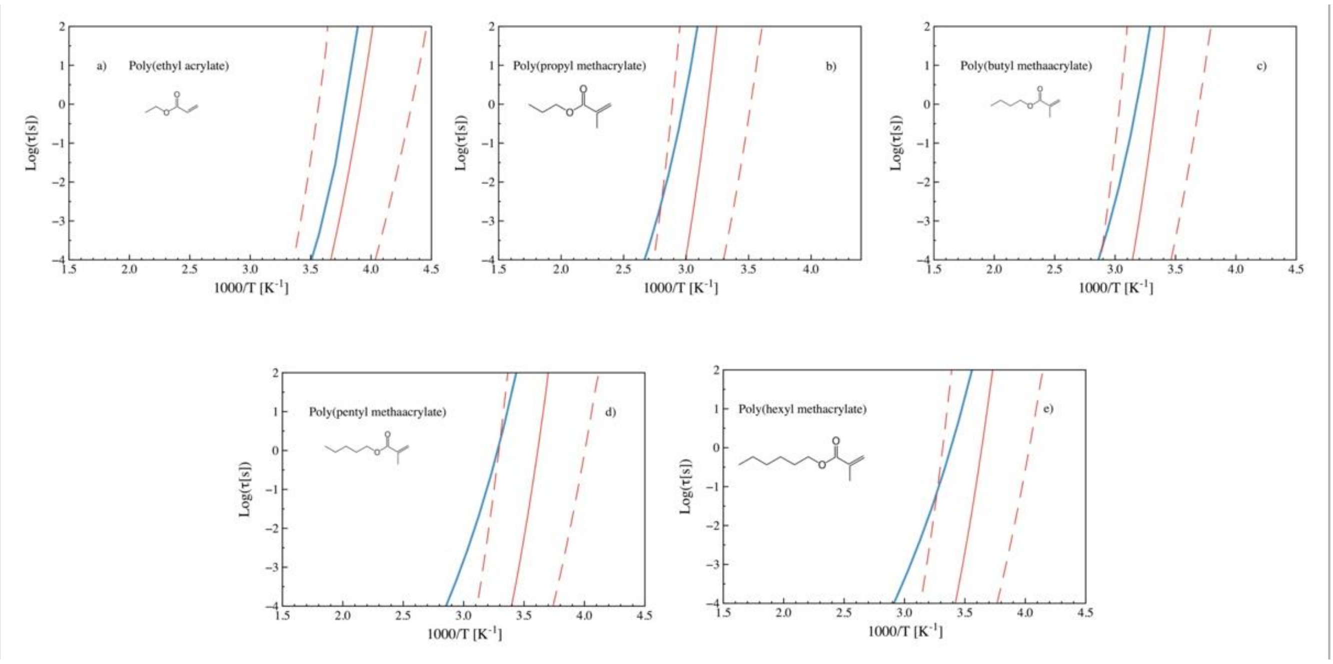


Figure 5. Experimental (blue) and predicted (red) relaxation times (obtained from ECNLE theory) vs. 1000/T (n-alkyl acrylates, with n ranging from 2 to 6). The corresponding monomeric chemical structures are also shown. (**a**) Poly (ethyl acrylate) [27], (**b**) poly (propyl methacrylate), (**c**) poly (butyl methacrylate), (**d**) poly (pentyl methacrylate) [27], and (**e**) poly (hexyl methacrylate) [27]. The plots correspond to predictions after introducing the scaling parameter (*a*) for linear tailed polymers.

We can move further by analyzing the experimental-predicted dynamics relationship for polymers where the side-chain length effects are not present. In that sense, Figure 6 shows experimental (blue) and predicted (red) relaxation times obtained from ECNLE after introducing the scaling parameter *a* for nonlinear tailed polymers. Poly (2, 2, 2 trifluoroethyl acrylate) (a), poly (isopropyl methacrylate) (b), poly (phenyl methacrylate) (c), and poly (secbutyl methacrylate) (d) present a much better agreement than the long linear tailed polymers (such as pentyl or hexyl methacrylates).

For this joint theoretical/numerical approach, we have two sources of uncertainty: on the one hand, the prediction of the T_g by the ANN; on the other hand, the accuracy of the ECNLE model to follow the temperature dependence of the relaxation times (i.e., fragility). Although the errors in both cases are not significant, there is still some room for improvement. The accuracy of the ANN can be improved by increasing the size of the training set; especially if we include polymers with chemical features similar to those we want to predict. In the case of the ECNLE model, a better understanding of the dependence of the parameter '*a*' with the chemical structure or the glass transition temperature would improve the predicted fragility.

In summary, and from a chemical structure point of view, many different factors have been reported to affect the glass transition and the polymer dynamics, thus increasing the difficulties in obtaining simple but realistic model approximations. The presence of bulky groups (as phenyl) can be 'diluted' by the presence of long alkyl chains in the same structure, whereas the lubricating effect of long alkyl chains can be hidden by very stiff backbones or by nanophase separations. The hybrid approach proposed can recognize these chemical features and quantify their relevance for estimating an alpha relaxation map area. It is important to highlight here that this knowledge is self-learned by the network, based only on the monomer chemical structure and the corresponding T_g value, and that ECNLE theory converts this output into a relaxation map. This approach could substantially help gain both qualitative and quantitative insights into the behavior of polymeric materials, especially for properties that are difficult and/or expensive to measure.

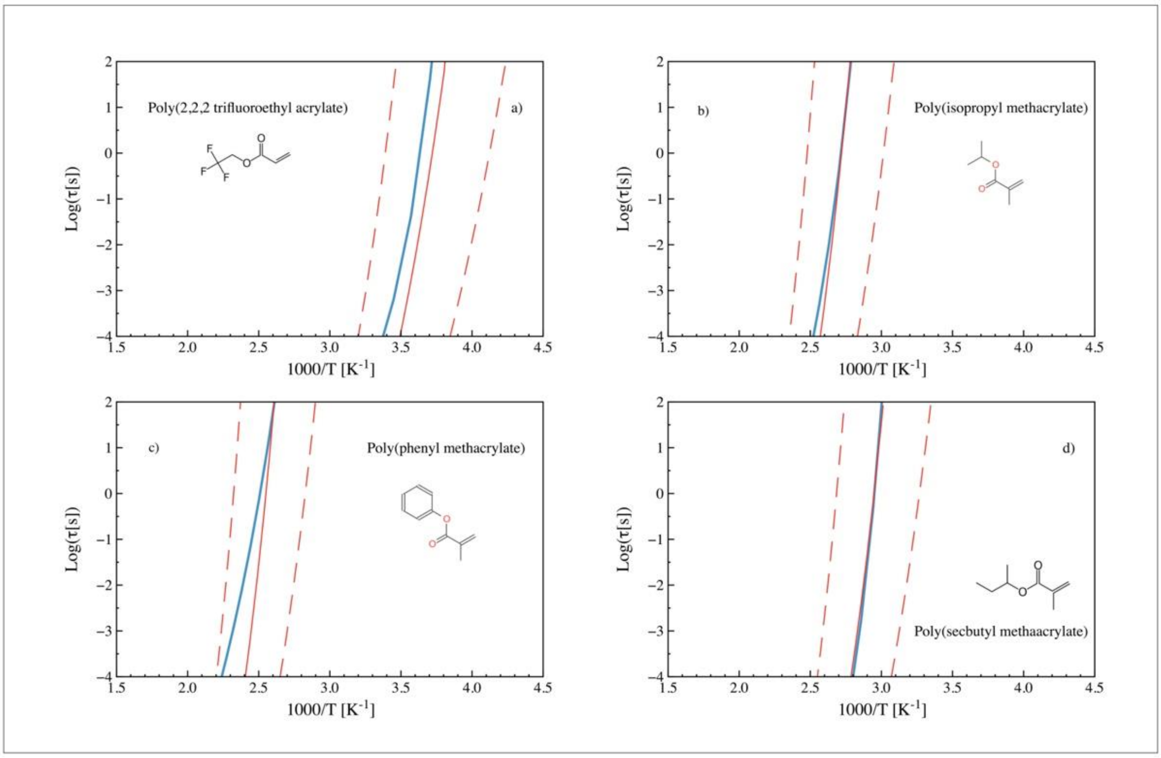


Figure 6. Experimental (blue) and predicted (red) relaxation times (obtained from ECNLE theory) vs. 1000/T after introducing the scaling parameter (*a*) for nonlinear tailed polymers. The corresponding monomeric chemical structure are also shown. (**a**) Poly (2, 2, 2 trifluoroethyl acrylate) [31], (**b**) poly (isopropyl methacrylate) [29], (**c**) poly (phenyl methacrylate), and (**d**) poly (secbutyl methacrylate) [30].

4. Conclusions

The feasibility of joining artificial neural networks and theory into a hybrid system to provide an estimation of the temperature dependence of the polymer alpha relaxation, based only on the knowledge of the chemical structure of the monomer, has been demonstrated. The proposed method has been tested on a set of polyacrylates providing, for short side-chain polymers, an excellent agreement between the predicted and experimental temperature dependence of the relaxation times. This approach relies only on the knowledge of the monomeric chemical formula and does not require any kind of experimental measurements or calculations as input, and constitutes a valuable tool for boosting the scientific understanding of structure–property relationships.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/polym14081573/s1, Table S1. The list of polyacrylates used in this work; Table S2. Parameters of the Vogel-Fulcher-Tammann (VFT) equation for the external control group; Figure S1. Schematic picture of the encoding process; Table S3. The number of filters and window sizes in the convolutional layers and the number of neurons in the fully connected layers; Figure S2. Shows the predicted vs experimental T_g values for the internal subset of polyacrylates after finishing the training process; Table S4. ECNLE caltulations; Figure S3. Relative deviations (Experimental – Predicted) / Experimental (in %) histogram for the training and internal test sets (a). The chemical structures for those molecules with more significant relative deviations are shown in (b). Figure S4. Glass transition temperature dependence of the model adjustable parameter *a* for several polymers and glass formers.

Author Contributions: Conceptualization, L.A.M. and G.A.S.; methodology, L.A.M., A.D.P. and G.A.S.; validation, C.B., U.C., G.A.S., L.A.M. and A.D.P.; formal analysis, C.B., G.A.S. and L.A.M.; investigation, G.A.S. and L.A.M.; resources, G.A.S.; data curation, L.A.M. and U.C.; writing—original draft preparation, L.A.M.; writing—review and editing, G.A.S., L.A.M. and A.D.P.; visualization, C.B., L.A.M. and U.C.; supervision, G.A.S.; project administration, G.A.S.; funding acquisition, G.A.S. All authors have read and agreed to the published version of the manuscript.

Funding: Spanish government, Ministerio de Ciencia e Innovación: PID2019-104650GB-C21; Basque Government: IT1566-22; Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 103.01-2019.318.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available within the article and in the Supporting Information (SI) file.

Acknowledgments: We gratefully acknowledge the financial support from the Spanish Government "Ministerio de Ciencia e Innovación" (PID2019-104650GB-C21) and the Basque Government (IT1566-22). This research was also funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 103.01-2019.318. We also acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nakajima, H.; Dijkstra, P.; Loos, K. The Recent Developments in Biobased Polymers toward General and Engineering Applications: Polymers That Are Upgraded from Biodegradable Polymers, Analogous to Petroleum-Derived Polymers, and Newly Developed. *Polymers* **2017**, *9*, 523. [CrossRef] [PubMed]
2. Umoren, S.A.; Solomon, M.M. Protective Polymeric Films for Industrial Substrates: A Critical Review on Past and Recent Applications with Conducting Polymers and Polymer Composites/Nanocomposites. *Prog. Mater. Sci.* **2019**, *104*, 380–450. [CrossRef]
3. de Leon, A.C.C.; da Silva, Í.G.M.; Pangilinan, K.D.; Chen, Q.; Caldona, E.B.; Advincula, R.C. High Performance Polymers for Oil and Gas Applications. *React. Funct. Polym.* **2021**, *162*, 104878. [CrossRef]
4. Wu, X.; Chen, X.; Zhang, Q.M.; Tan, D.Q. Advanced Dielectric Polymers for Energy Storage. *Energy Storage Mater.* **2022**, *44*, 29–47. [CrossRef]
5. Wang, Y.; Ghanem, B.S.; Han, Y.; Pinnau, I. State of the Art Polymers of Intrinsic Microporosity for High-Performance Gas Separation Membranes. *Curr. Opin. Chem. Eng.* **2022**, *35*, 100755. [CrossRef]
6. Devaraju, S.; Alagar, M. Polymer Matrix Composite Materials for Aerospace Applications. *Encycl. Mater. Compos.* **2021**, *1*, 947–969. [CrossRef]
7. Vidya; Mandal, L.; Verma, B.; Patel, P.K. Review on Polymer Nanocomposite for Ballistic & Aerospace Applications. *Mater. Today Proc.* **2020**, *26*, 3161–3166. [CrossRef]
8. Gambino, T.; Alegría, A.; Arbe, A.; Colmenero, J.; Malicki, N.; Dronet, S. Modeling the High Frequency Mechanical Relaxation of Simplified Industrial Polymer Mixtures Using Dielectric Relaxation Results. *Polymer* **2020**, *187*, 122051. [CrossRef]
9. Menard, K.P.; Menard, N.R. Dynamic Mechanical Analysis in the Analysis of Polymers and Rubbers. *Encycl. Polym. Sci. Technol.* **2015**, 1–33. [CrossRef]
10. Capiel, G.; Miccio, L.A.; Montemartini, P.E.; Schwartz, G.A. Water Diffusion and Hydrolysis Effect on the Structure and Dynamics of Epoxy-Anhydride Networks. *Polym. Degrad. Stab.* **2017**, *143*, 57–63. [CrossRef]
11. Otegui, J.; Miccio, L.A.; Arbe, A.; Schwartz, G.A.; Meyer, M.; Westermann, S. Determination of Filler Structure in Silica-Filled SBR Compounds by Means of SAXS and AFM. *Rubber Chem. Technol.* **2015**, *88*, 690–710. [CrossRef]
12. Phan, A.D.; Schweizer, K.S. Elastically Collective Nonlinear Langevin Equation Theory of Glass-Forming Liquids: Transient Localization, Thermodynamic Mapping, and Cooperativity. *J. Phys. Chem. B* **2018**, *122*, 8451–8461. [CrossRef] [PubMed]
13. Phan, A.D.; Wakabayashi, K. Effects of Cooling Rate on Structural Relaxation in Amorphous Drugs: Elastically Collective Nonlinear Langevin Equation Theory and Machine Learning Study. *RSC Adv.* **2019**, *9*, 40214–40221. [CrossRef]
14. Phan, A.D.; Knapik-Kowalczuk, J.; Paluch, M.; Hoang, T.X.; Wakabayashi, K. Theoretical Model for the Structural Relaxation Time in Coamorphous Drugs. *Mol. Pharm.* **2019**, *16*, 2992–2998. [CrossRef]
15. Miccio, L.A.; Schwartz, G.A. From Chemical Structure to Quantitative Polymer Properties Prediction through Convolutional Neural Networks. *Polymer.* **2020**, *193*, 122341. [CrossRef]
16. Miccio, L.A.; Schwartz, G.A. Localizing and Quantifying the Intra-Monomer Contributions to the Glass Transition Temperature Using Artificial Neural Networks. *Polymer.* **2020**, *203*, 122786. [CrossRef]

17. Miccio, L.A.; Schwartz, G.A. Mapping Chemical Structure-Glass Transition Temperature Relationship through Artificial Intelligence. *Macromolecules* **2021**, *54*, 1811–1817. [CrossRef]
18. Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. [CrossRef]
19. O'Boyle, N.M. Towards a Universal SMILES Representation-A Standard Method to Generate Canonical SMILES Based on the InChI. *J. Cheminform.* **2012**, *4*, 22. [CrossRef]
20. Plazek, D.J.; Ngai, K.L. The Glass Temperature. In *Physical Properties of Polymers Handbook*; Mark, J.E., Ed.; Springer: New York, NY, USA, 2007; pp. 187–215. ISBN 978-0-387-69002-5.
21. Plastic Library, Chemical Retrieval on the Web, Crow. Available online: https://polymerdatabase.com (accessed on 19 April 2019).
22. Bertinetto, C.; Duce, C.; Micheli, A.; Solaro, R.; Starita, A.; Tine, M.R. Prediction of the Glass Transition Temperature of (Meth)Acrylic Polymers Containing Phenyl Groups by Recursive Neural Network. *Polymer* **2007**, *48*, 7121–7129. [CrossRef]
23. Wypych, G. *Handbook of Polymers*, 2nd ed.; ChemTec Publishing: Toronto, ON, Canada, 2016; ISBN 978-1-895198-92-8.
24. Alkharusi, H. Categorical Variables in Regression Analysis: A Comparison of Dummy and Effect Coding. *Int. J. Educ.* **2012**, *4*, 202–210. [CrossRef]
25. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
26. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
27. He, X.; Wu, J.; Huang, G.; Wang, X. Effect of Alkyl Side Chain Length on Relaxation Behaviors in Poly(n-Alkyl Acrylates) and Poly(n-Alkyl Methacrylates). *J. Macromol. Sci. Part B Phys.* **2010**, *50*, 188–200. [CrossRef]
28. Qin, Q.; McKenna, G.B. Correlation between Dynamic Fragility and Glass Transition Temperature for Different Classes of Glass Forming Liquids. *J. Non-Cryst. Solids* **2006**, *352*, 2977–2985. [CrossRef]
29. Menissez, C.; Sixou, B.; David, L.; Vigier, G. Dielectric and Mechanical Relaxation Behavior in Poly(Butyl Methacrylate) Isomers. *J. Non-Cryst. Solids* **2005**, *351*, 595–603. [CrossRef]
30. Sato, A.; Sasaki, T. Cooperativity of Dynamics in Supercooled Polymeric Materials and Its Temperature Dependence Predicted from a Surface Controlled Model. *Eur. Polym. J.* **2018**, *99*, 485–494. [CrossRef]
31. Merino, E.G.; Atlas, S.; Raihane, M.; Belfkira, A.; Lahcini, M.; Hult, A.; Dionísio, M.; Correia, N.T. Molecular Dynamics of Poly(ATRIF) Homopolymer and Poly(AN-Co-ATRIF) Copolymer Investigated by Dielectric Relaxation Spectroscopy. *Eur. Polym. J.* **2011**, *47*, 1429–1446. [CrossRef]
32. Xie, S.J.; Schweizer, K.S. Nonuniversal Coupling of Cage Scale Hopping and Collective Elastic Distortion as the Origin of Dynamic Fragility Diversity in Glass-Forming Polymer Liquids. *Macromolecules* **2016**, *49*, 9655–9664. [CrossRef]
33. Godard, M.; Saiter, J. Fragility and Non-Linearity in Polymethyl (n-Alkyl) Acrylates. *J. Non-Cryst. Solids* **1998**, *237*, 635–639. [CrossRef]
34. Calleja, R.D.; Jaime, C.; Sanchis, M.J.; Romcin, J.S.; Gargallo, L. Dynamic Mechanical and Dielectric Relaxations in Poly(Pentachloropheny1 Methacrylate). *Macromol. Chem. Phys.* **1998**, *581*, 575–581. [CrossRef]
35. Calleja, R.D.; Jaime's, C.; Sanchis-Sanchez, M.J.; Martinez-Piña, F.; Gargallo, L.; Radic, D. Mechanical and Dielectric Properties of Bulky Side Chain Poly(Methacry1ates). Analysis of the Low Frequency Phenomena. 1: Poly(5-Lndanyl Methacrylate). *Polym. Eng. Sci.* **1997**, *37*, 882–887. [CrossRef]
36. Fredrickson, G.H.; Bates, F.S. Dynamics of Block Copolymers: Theory and Experiment. *Annu. Rev. Mater. Sci.* **1996**, *26*, 501–550. [CrossRef]
37. Böhmer, R.; Ngai, K.L.; Angell, C.A.; Plazek, D.J. Nonexponential Relaxations in Strong and Fragile Glass Formers. *J. Chem. Phys.* **1993**, *99*, 4201–4209. [CrossRef]
38. Diaz-Calleja, R. Dielectric Relaxation Studies on Phenyl and Chlorophenyl Esters of Poly(Acry1ic Acid). *Macromolecules* **1991**, *24*, 264–269. [CrossRef]
39. Floudas, G.; Štepánek, P. Structure and Dynamics of Poly(n-Decyl Methacrylate) below and above the Glass Transition. *Macromolecules* **1998**, *31*, 6951–6957. [CrossRef]
40. Garci, A.; Di, R.; Guzma, J. Relaxation Behavior of Acrylate and Methacrylate Polymers Containing Dioxacyclopentane Rings in the Side Chains. *J. Polym. Sci. Part B Polym. Phys.* **2000**, *39*, 286–299. [CrossRef]
41. Sanchis, M.J.; Saiz, E.; Marti, F. Dynamic Mechanical and Dielectric Relaxations of Poly (Difluorobenzyl Methacrylates). *J. Polym. Sci. Part B Polym. Phys.* **2000**, *38*, 2179–2188.
42. García, N.; Compañ, V.; Díaz-Calleja, R.; Guzmán, J.; Riande, E. Comparative Study of the Relaxation Behaviour of Acrylic Polymers with Flexible Cyclic Groups in Their Structure. *Polymer* **2000**, *41*, 6603–6611. [CrossRef]
43. Balabin, R.M. Enthalpy Difference between Conformations of Normal Alkanes: Raman Spectroscopy Study of n-Pentane and n-Butane. *J. Phys. Chem. A* **2009**, *113*, 1012–1019. [CrossRef]
44. Wind, M.; Graf, R.; Renker, S.; Spiess, H.W.; Steffen, W. Structure of Amorphous Poly-(Ethylmethacrylate): A Wide-Angle x-Ray Scattering Study. *J. Chem. Phys.* **2004**, *122*, 014906. [CrossRef] [PubMed]
45. Wind, M.; Graf, R.; Renker, S.; Spiess, H.W. Structural Reasons for Restricted Backbone Motion in Poly(n-Alkyl Methacrylates): Degree of Polymerization, Tacticity and Side-Chain Length. *Macromol. Chem. Phys.* **2005**, *206*, 142–156. [CrossRef]
46. Beiner, M.; Schröter, K.; Hempel, E.; Reissig, S.; Donth, E. Multiple Glass Transition and Nanophase Separation in Poly(n-Alkyl Methacrylate) Homopolymers. *Macromolecules* **1999**, *32*, 6278–6282. [CrossRef]

47. Beiner, M.; Kabisch, O.; Reichl, S.; Huth, H. Structural and Dynamic Nanoheterogeneities in Higher Poly (Alkyl Methacrylate). *J. Non-Cryst. Solids* **2002**, *310*, 658–666. [CrossRef]
48. Beiner, M.; Huth, H. Nanophase Separation and Hindered Glass Transition in Side-Chain Polymers. *Nat. Mater.* **2003**, *2*, 595–599. [CrossRef] [PubMed]
49. Arbe, A.; Genix, A.; Arrese-Igor, S.; Colmenero, J.; Richter, D. Dynamics in Poly (n-Alkyl Methacrylates): A Neutron Scattering, Calorimetric, and Dielectric Study. *Macromolecules* **2010**, *43*, 3107–3119. [CrossRef]
50. Qazvini, N.T.; Mohammadi, N. Segmental Dynamics in Net-Poly(Methyl Methacrylate)-Co-Poly(n-Butyl Acrylate) Copolymer Networks. *J. Macromol. Sci. Part B Phys.* **2008**, *47*, 1161–1175. [CrossRef]
51. Cangialosi, D.; Schwartz, G.A.; Alegría, A.; Colmenero, J. Combining Configurational Entropy and Self-Concentration to Describe the Component Dynamics in Miscible Polymer Blends. *J. Chem. Phys.* **2005**, *123*, 144908. [CrossRef]
52. Schwartz, G.A.; Alegría, Á.; Colmenero, J. Adam-Gibbs Based Model to Describe the Single Component Dynamics in Miscible Polymer Blends under Hydrostatic Pressure. *J. Chem. Phys.* **2007**, *127*, 154907. [CrossRef]

Journal of Non-Crystalline Solids: X 18 (2023) 100185

Contents lists available at ScienceDirect

Journal of Non-Crystalline Solids: X

journal homepage: www.sciencedirect.com/journal/journal-of-non-crystalline-solids-x

Characterising the glass transition temperature-structure relationship through a recurrent neural network

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

[a] Centro de Física de Materiales (CSIC-UPV/EHU)-Material Physics Centre (MPC), P. M. de Lardizábal 5, San Sebastián 20018, Spain
[b] Donostia International Physics Center, P. M. de Lardizábal 4, San Sebastián 20018, 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
Machine learning
Molecular glass former
Amino acid
RNN

ABSTRACT

Quantitative structure-property relationship (QSPR) is a powerful analytical method to find correlations between the structure of a molecule and its physicochemical properties. The glass transition temperature (T_g) is one of the most reported properties, and its characterisation is critical for tuning the physical properties of materials. In this work, we explore the use of machine learning in the field of QSPR by developing a recurrent neural network (RNN) that relates the chemical structure and the glass transition temperature of molecular glass formers. In addition, we performed a chemical embedding from the last hidden layer of the RNN architecture into an *m*-dimensional T_g-oriented space. Then, we test the model to predict the glass transition temperature of essential amino acids and peptides. The results are very promising and they can open the door for exploring and designing new materials.

1. Introduction

In the field of quantitative structure-property relationship (QSPR) [1–6], machine learning (ML) methods open new routes to investigate and explore the physico-chemical properties of materials [6–11]. ML methods typically use molecular descriptors or a representation of molecular structures to predict several material properties. Among the most relevant material properties, the glass transition temperature (T_g) stands out since it is used in quality control of food and pharmaceutical drugs, defining the polymer production process parameters or tuning the mechanical properties of compounds [12–14], among many others. The T_g of numerous glass formers has been measured using different experimental techniques like differential scanning calorimetry [15,16], broadband dielectric spectroscopy [17–19], or rheology [20] and is widely reported in the literature. Several theories also model the glass transition mechanism [12,21–23], usually involving phenomenological parameters that account for still not fully understood processes.

Among the first attempts to estimate the T_g of glass formers based on their chemical structure, we can mention a method developed in the polymers field by Weyland et al. [21], which consists of considering the glass transition temperature as a sum of weighted group contribution of the atoms of the polymer. However, there is no specific way to choose these weights. More recent studies use artificial neural networks (ANN) and physico-chemical features to predict the T_g of materials but neglect the molecular structure and the interaction between atoms [24,25]. Also, whereas there are several studies dealing with the glass transition temperature of polymers [26–30] and inorganic glasses [25], we have found a lack of studies in the literature concerning the use of neural networks for predicting the T_g of organic molecular glass formers. These are very complex systems presenting a variety of intermolecular interactions that makes necessary a different and innovative approach that overcomes the limitations of the standard ANN.

In this work, we present a recurrent neural network (RNN) capable of predicting the glass transition temperature of several molecular glass formers (including biomolecules, pharmaceutical molecules, and additives typically used in the pharmaceutical industry). In particular, we show that by using a dataset of individual organic molecules structures and a bidirectional long short-term memory (Bi-LSTM) architecture, it is possible to achieve a prediction of the T_g with average deviations lower than 9%. Furthermore, we show that these networks also capture physically meaningful variables underneath the glass transition process in molecular glass formers, like the influence of intermolecular forces

* Corresponding authors at: Centro de Física de Materiales (CSIC-UPV/EHU), Materials Physics Center (MPC), P. M. de Lardizábal 5, San Sebastían 20018, Spain.
** Corresponding author at: Donostia International Physics Center, P. M. de Lardizábal 4, San Sebastián 20018, Spain
E-mail addresses: lamiccio@gmail.com (L.A. Miccio), silvina.cerveny@ehu.es (S. Cerveny), gustavo.schwartz@csic.es (G.A. Schwartz).

https://doi.org/10.1016/j.nocx.2023.100185
Received 8 March 2023; Received in revised form 20 April 2023; Accepted 29 April 2023
Available online 2 May 2023

and molecular weight. Finally, we apply our model to predict the T_g of the 20 biologically relevant amino acids and compare the results with the experimental measurements of a group of amino acids and peptides.

2. Materials and methods

In this section, we define the dataset, the data treatment, and the characteristics of the neural network, including the architecture and the training options.

2.1. Dataset

We have collected a dataset of 501 organic molecules whose experimental glass transition temperature was reported in the literature. The dataset includes alcohols, hydrocarbons, sugars, aromatic compounds, and pharmaceutical products, spanning a T_g range from 18 K to 450 K. A detailed dataset description can be found in Section 1 of the Supplementary Information file (SI).

2.2. Data treatment

We identified each molecular structure with their simplified molecular-input line-entry system (SMILES) [31] string using the open-source cheminformatics software RDkit [32] to get a unique representation of the molecular structures. We then numerically encoded each string using the following dictionary:

$$\{(, c, 4, F, =, \#, n, S, @, 3, I, o, s, 6, N, H, X, 7, +, Y, 2, d, 5, 1, P, O,], C, -, /, [,)\}$$

We assigned a number to each symbol according to its position in the dictionary (cardinal encoding), obtaining a 1-dimensional numerical array for each structure to feed the neural network. We padded the SMILES strings by adding a 0 at the beginning of each sequence and completing them with 0 s (only one final 0 for the longest string) so that all instances have the same length. The scheme in Fig. 1 shows an example of the encoding process.

2.3. RNN's architecture

We employed a long short-term memory neural network architecture [33,34], constructed using MATLAB. In Fig. 2, we show a schematic picture of the whole network, starting with a sequence input (which takes as input the SMILES encoded as expressed in the previous section), a word embedding layer, which feeds a bidirectional long short-term memory (BiLSTM) layer, a batch normalisation layer and finally a mean absolute relative error (MARE) regression that outputs the T_g.

We tested different values of neurons in the BiLSTM layer (from 8 to 32 nodes) and several values of the word embedding dimension (10,20,30). We chose the network architecture for which the value of the mean absolute percentage error (MAPE) of the validation set was minimum, as shown in Fig. 3. Thus, we finally have 8 neurons in the BiLSTM and a word embedding dimension of 20. Note that, as we use a Bidirectional LSTM, the number of neurons doubles to 16 as the network reads the sequences in both directions. We selected this set of hyperparameters by keeping fixed the training-validation division and running the learning algorithm for each architecture 100 times.

2.4. RNN training and optimisation

We extracted a test set of 30 elements from the dataset, trying to represent its variety of chemical composition as closely as possible. Then, we randomly shuffled 100 times the remaining dataset, splitting it into a training set of 441 molecules and a validation set of 30 molecules. This results in ~90%, 5%, and 5% partition for training, validation, and test set, respectively. For each split, we ran the learning algorithm of the RNN 100 times, to investigate the sensitivity of the architecture concerning the initial conditions. We used the gradient descent method and the Adam optimisation protocol during the training procedure. We employed a learning rate of 0.01 and trained each network for 1000 epochs. We selected a network that satisfied the following requirements:

- MARE Train <0.06;
- $\frac{\textit{MARE Train}}{\textit{MARE Val}} > 0.8$;
- min(MARE Val).

By fulfilling these requirements, the performance of the RNN on the validation set should be similar to that of the training set. Therefore the value of the MAPE of the validation set is below 9% (i.e., the performance of the selected network can be defined as validation set oriented). In Fig. 4, we show the average T_g predicted for 100 runs versus the corresponding experimental values, also reporting the mean standard deviation of each set.

3. Results and discussion

In this section, we show that the network is sensitive to the physically meaningful variables of the glass transition process by embedding the last activation layer and performing non-supervised clustering analysis and dimensionality reduction techniques. In addition, we explore the possibility of employing the proposed dataset and architecture to estimate the glass transition temperature of amino acids and short peptides.

3.1. Characterisation of the network

Based on the modality described in “RNN training and optimisation”, we select a network for which the average error of the validation set is similar to that of the training set. In Fig. 5, we show the predicted glass transition temperatures as a function of the corresponding experimental counterpart. The data lay almost perfectly on the bisector of the Cartesian plane, implying a concordance between the experimental and predicted T_g values for the molecules in the training (blue), validation (orange), and test (yellow) sets. The observed deviations are below 9%.

Once the RNN has been trained (and optimised) to predict the T_g value, we can assume that the activation of the neurons, particularly those of the last layer, codify enough chemical information to embed molecular structures into a T_g-oriented m-dimensional space. This procedure allows performing associations among molecules in the dataset by applying clusterization algorithms without using molecular fingerprints or other descriptors. By embedding the molecular structures in such high-dimensional space, it is possible to lead mathematical operations with these representations of the chemical structures. We then plotted the activation vectors in 3 dimensions using the principal component analysis (PCA) [35]. This dimensionality reduction is needed to ensure human-readability since each activation vector contains 16

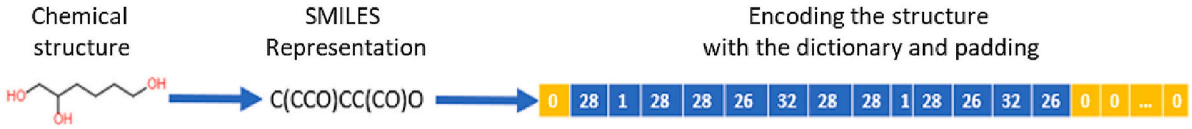


Fig. 1. Encoding the SMILES with cardinal encoding. We added a 0 at the beginning of each string and completed them with a padding of 0 s to have the same length for all the instances.

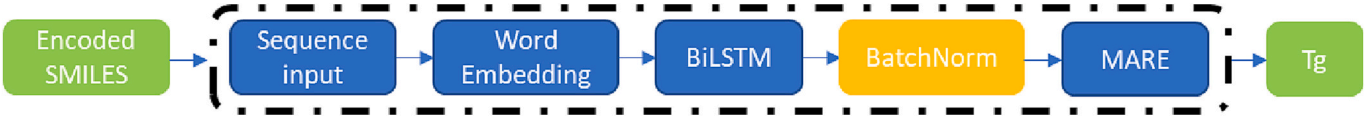


Fig. 2. The ANN architecture comprises a Sequence Input layer, a Word Embedding layer, a Bidirectional LSTM layer, a Batch Normalisation layer, and a mean absolute relative error output layer. It takes as an input the encoded SMILES and outputs the T_g of the molecule(s).

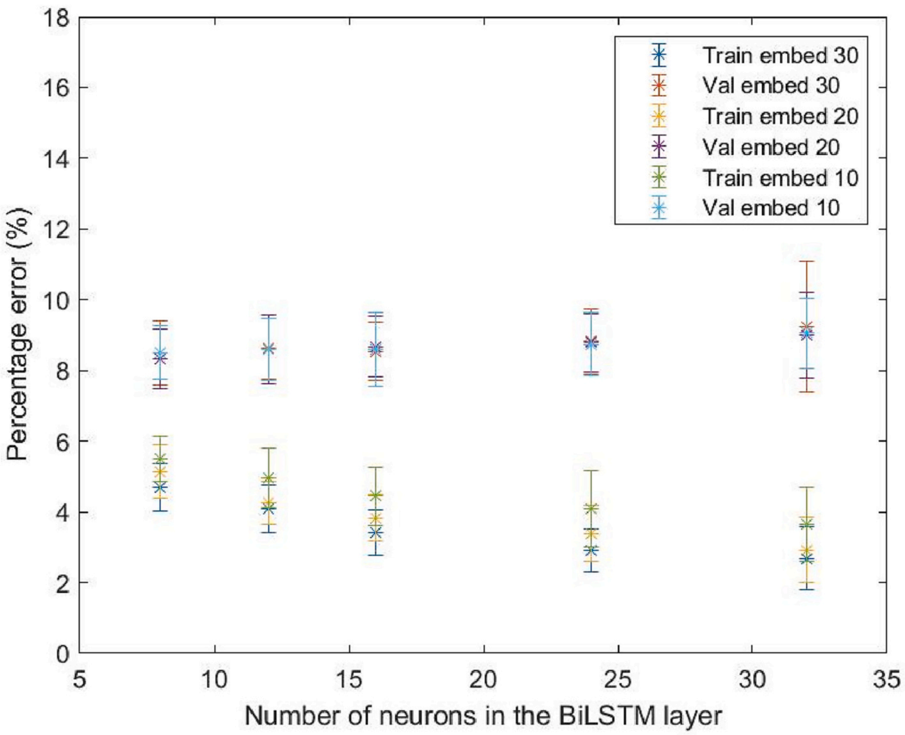


Fig. 3. Architecture test: we tested different values of neurons in the BiLSTM layer (from 8 to 32 nodes) and several values of the word embedding dimension (10, 20, 30).

values (16 dimensions, each corresponding to a neuron's activation). We observed that most of the variance, and therefore most of the chemical information (~88%), is contained in the first three components of the PCA: PC1 = 75.03%, PC2 = 6.94%, and PC3 = 6.02%. Fig. 6 shows a 3-dimensional colour map of the obtained components and the corresponding 2D projection on the main axes (PC1 and PC2), where the colours represent the experimental T_g of each compound. The data follow a gradient from blue to red colours (i.e., from lower to higher glass transition temperatures).

We performed a non-supervised analysis of the data by clustering using the fuzzy-c algorithm [36] on the batch normalisation layer. We show the obtained results in Fig. 7 (for the components PC1 and PC2). Fuzzy-c algorithm allows knowing the probability with which each molecule belongs to a given cluster (i.e., each molecule can participate in more than one cluster with a certain probability). Since this algorithm requires predefining the number of clusters, we employed the Elbow method to determine the optimum parameter (n = 16, we show the details in the SI). Clustering can help identify patterns and relationships between the molecular structure and the T_g by grouping similar molecules together. This process also helps reveal how the network deals with the variables affecting the glass transition temperature, such as the molecular weight, intermolecular forces and other chemical composition-related factors. Furthermore, clustering can also be used to identify potential outliers in the employed datasets, which can be further studied to gain insights into the underlying mechanism of the glass transition phenomenon and the neural network training processes.

In Fig. 7, we present chemical structures within different clusters and the trajectory these compounds follow on the map. The clusters on the bottom right (A) mainly consist of low-molecular-weight, flexible linear carbonated chains and weak intermolecular forces. Conversely, the left side of the representation (C) is composed of molecules with high-molecular weight, more rigid phenyl groups, and strong intermolecular forces. Notably, in the middle section (B), we observe a change in the intermolecular forces and the structural composition of the molecules as they progressively become more branched and incorporate bulkier molecular groups into their structure. These results show that the network can recognise and classify complex features linked to the glass transition temperature by learning from the SMILES representation of the chemical structure of the molecular glass formers.

3.2. T_g vs. molecular weight

The previous analysis can be complemented using known experimental variables such as glass transition temperature and molecular weight, which show a well-established trend, as seen in Fig. 8a. It is worth noting that the network was only provided with the molecular structure expressed as a SMILES string, and no other chemical information was given. Therefore, the RNN implicitly learned the general trend between T_g and the molecular weight from the chemical structures encoded as SMILES strings.

Fig. 8a can also be interpreted as an indicator of the trained neural network's confidence area for predicting the T_g of new molecular glass formers (i.e., where new chemical structures would be well represented by the elements in the dataset). Thus, the region enclosed by the dashed

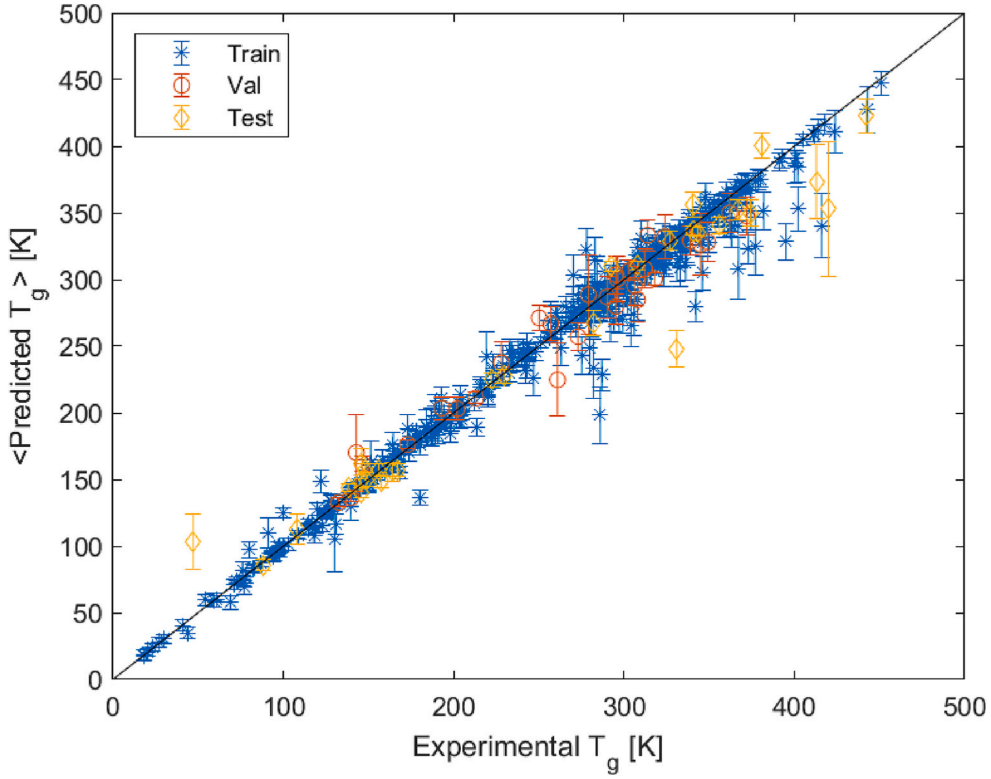


Fig. 4. Average prediction of the glass transition temperature for the training (blue), validation(orange) and test(yellow) set. The mean standard deviation for the training, validation and test set are 7 K, 13 K, and 9 K, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

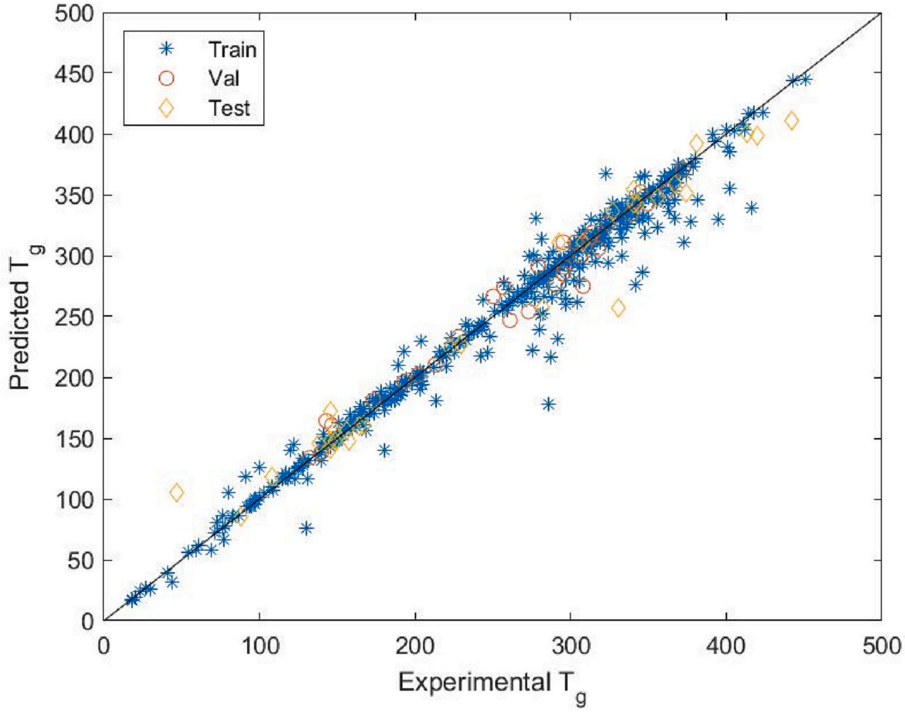


Fig. 5. Performance of the chosen neural network for the training (blue), validation (orange) and test (yellow) sets. The data points lay almost perfectly on the bisector axis, indicating an excellent agreement between experimental and predicted T_g. The MAPE values obtained are 3.4%, 3.8% and 8.7% for the training, validation and test sets, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

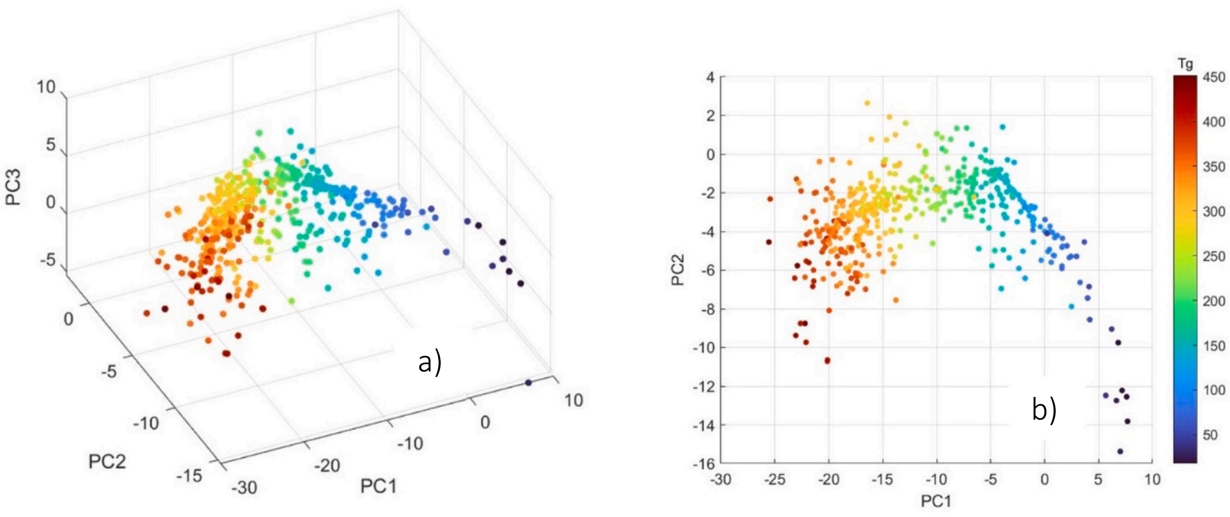


Fig. 6. PCA projection of the batch normalisation layer activations. We use a colour map to enhance the trend of the glass transition temperature, which goes from blue colours (low T_g) to red colours (high T_g). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

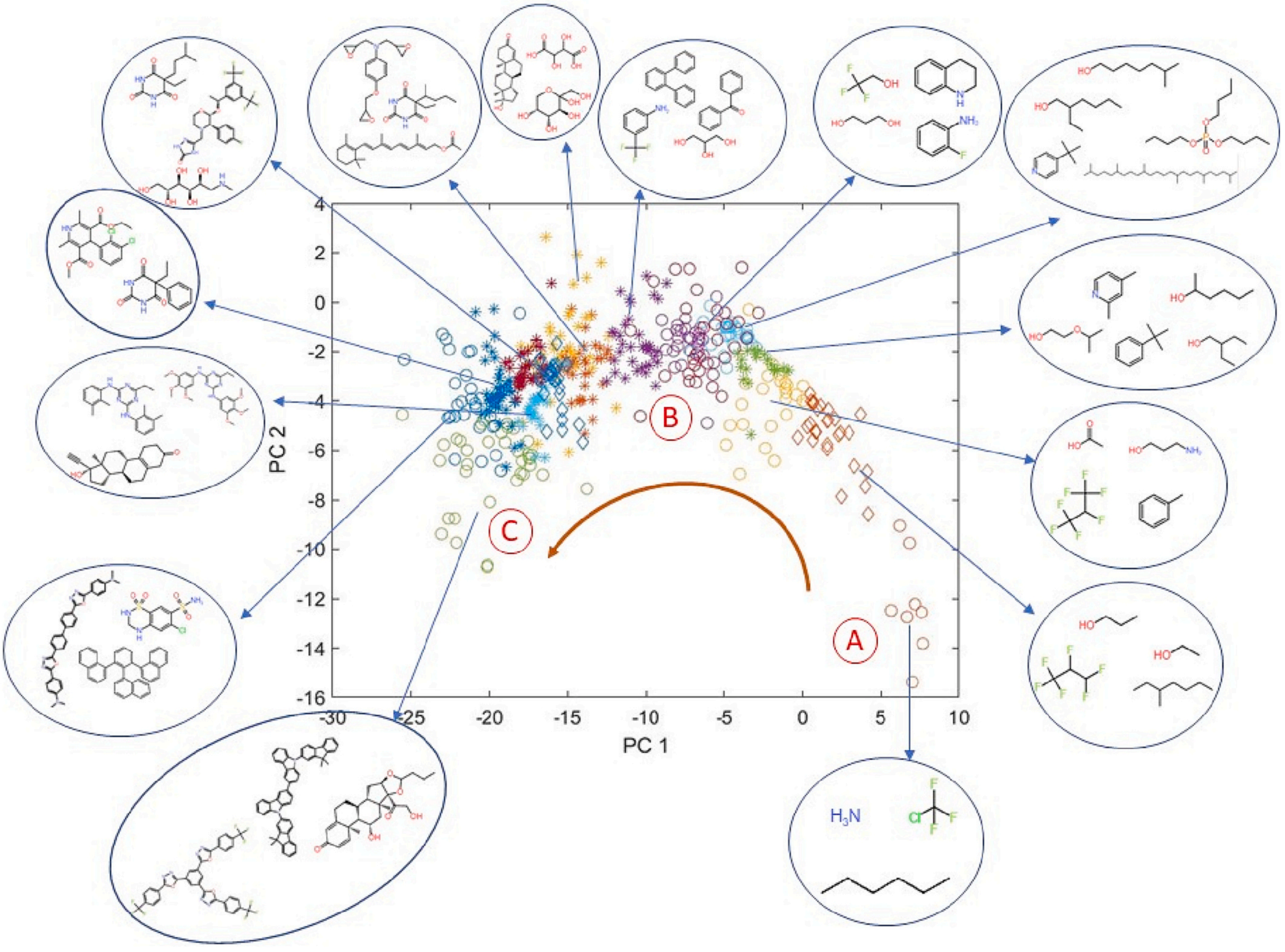


Fig. 7. We clustered the chemical structures within the m-dimensional space using the fuzzy C algorithm. In this way, we can observe the structural changes along the trajectory of the PCA, going from low-molecular-weight, linear chains and weak intermolecular forces (A) to high-molecular-weight, a higher concentration of more rigid groups and strong intermolecular forces (C).

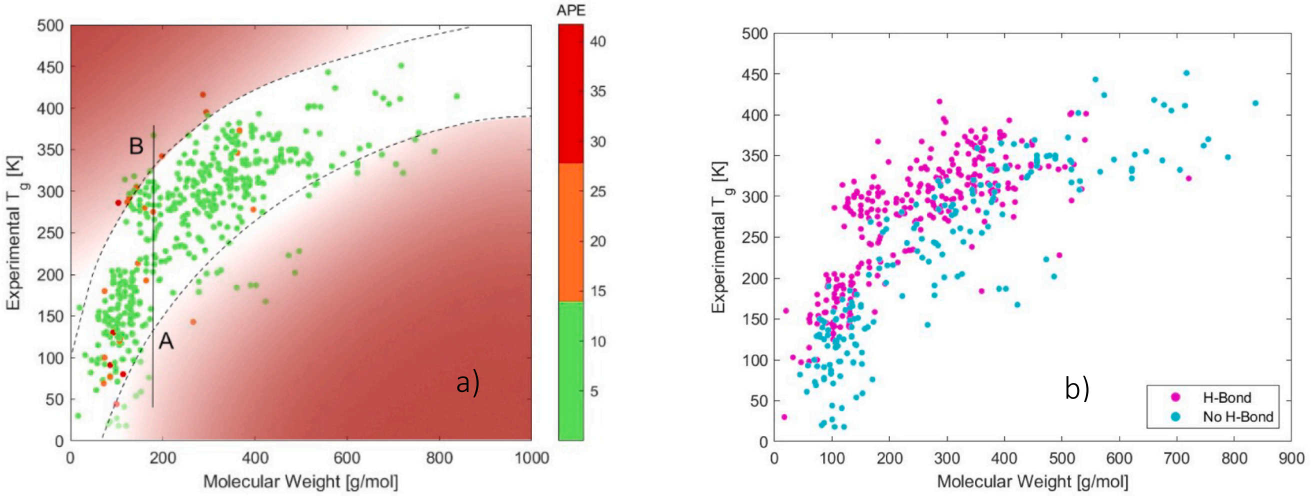


Fig. 8. Molecular weight dependence of the glass transition temperature for the training set molecules. The colour map in Fig. 8a represents the absolute percentage error (APE) when predicting the T_g of the compound. The area between the dashed lines represents the confidence interval of the neural network. Also, the vertical line from point A to point B (fixed molecular weight) indicates the raising of the T_g due to the contribution of intermolecular forces. Fig. 8b shows the hydrogen bond distribution over the molecular weight trend. Lines are just a guide for the eyes and indicate approximate regions of low and large intermolecular forces.

lines represents the chemical space from which the network learned the underlying features of the glass transition process. In this plot, at fixed molecular weight (going, for example, vertically from point A to point B) variations in T_g are due to changes in the molecular structure (at constant molecular weight) most likely because of the increasing of intermolecular forces (see Fig. 8b and the next paragraph for more details). The colour map on the plot represents the errors of the RNN in predicting the glass transition temperature of the training set. Therefore, the observed homogeneous distribution of red and orange dots indicates no bias due to molecular weight or intermolecular forces. For those elements located on the upper side of the general trend, the neural network must consider the effect of molecular weight, the flexibility of the different groups, and the impact of intermolecular forces.

In Fig. 8b, we show the same molecular weight dependence of the glass transition temperature dividing the molecules into those able to form (pink) or not (light blue) H-bond networks (only considering the existence of H-bonds donors and acceptors, disregarding the amount and location in the molecule). The molecules which lack donors or acceptors of hydrogen bond fall into the "no H-bond" category and occupy the lower part of the graph. In contrast, the molecules with potential hydrogen bonding properties fill the upper part of the plot.

These results, along with the clusterisation ones, agree with traditional experimental observations of glass transition temperature trends for several glass formers, indicating that the network has effectively learnt some features of the underlying physics of the glass transition phenomena.

3.3. Application to biological molecules

The study of the properties of amino acids is a hot topic in many fields, such as biophysics, food, and pharmaceutical industries. Overall, measuring the glass transition temperature of amino acids can be complex and challenging due to many factors affecting the measurement, including the presence of absorbed moisture, the sensitivity to measurement conditions, and their degradation temperatures. In addition, many biomolecules are not "good" glass formers because partial or complete crystallization may occur during cooling, or the sample might degrade when melting. For these reasons, it is interesting to explore numerical routes to estimate the physical properties of biomolecules. Therefore, we used our model to predict the glass transition temperature of the 20 essential amino acids and a short peptide. Table 1 shows the predicted values for the T_g of the essential amino acids [37,38] and the corresponding experimental values (for some of them).

Table 1
Results of predicting the glass transition temperature of the 20 amino acids and oligomer 3-Lys.*Own DSC measurements of 3-Lys samples (see section 3 in SI).

Molecule number	Name	Predicted T_g [K]	T_g [K]	APE [%]
1	Alanine	284		
2	Arginine	339	362 [37]	6.2
3	Asparagine	330	466 [37]	29.2
4	Aspartic acid	312	386 [37]	19.1
5	Cysteine	314		
6	Glutamine	323	323 [37]	0.1
7	Glutamic acid	310	330 [37]	6.1
8	Glycine	229		
9	Histidine	318	408 [37]	22.2
10	Isoleucine	273		
11	Leucine	278		
12	Lysine	311	317 [38]	1.9
13	Methionine	281		
14	Phenylalanine	307		
15	Proline	195		
16	Serine	301	337 [37]	10.7
17	Threonine	275	355 [37]	22.4
18	Tryptophan	330	433 [37]	23.8
19	Tyrosine	327	405 [37]	19.3
20	Valine	284		
21	3-Lys	311	312.5*	0.3

In Fig. 9, we plot the amino acids in the previously analysed T_g versus molecular weight map. The red dots represent amino acids for which the absolute percentage error on the prediction of the T_g is higher than 10%. Noticeably, these compounds are all located outside the model's predictive region. On the other hand, blue dots represent the amino acids and the peptides for which the prediction error is lower than 7%. These molecules, which are closer to the chemical space covered by the training set (green dots), have more accurate predictions for T_g. These results clearly show that the glass transition temperature of amino acids (at least those within the prediction confident area) can be predicted by our RNN trained on different chemical families. As a particular test, we also included the 3-lysine (3-Lys) data, which has a more complex chemical structure but still falls within the model's confidence area. In this case, the agreement between the predicted and the measured value of the glass transition temperature is excellent. These findings open the

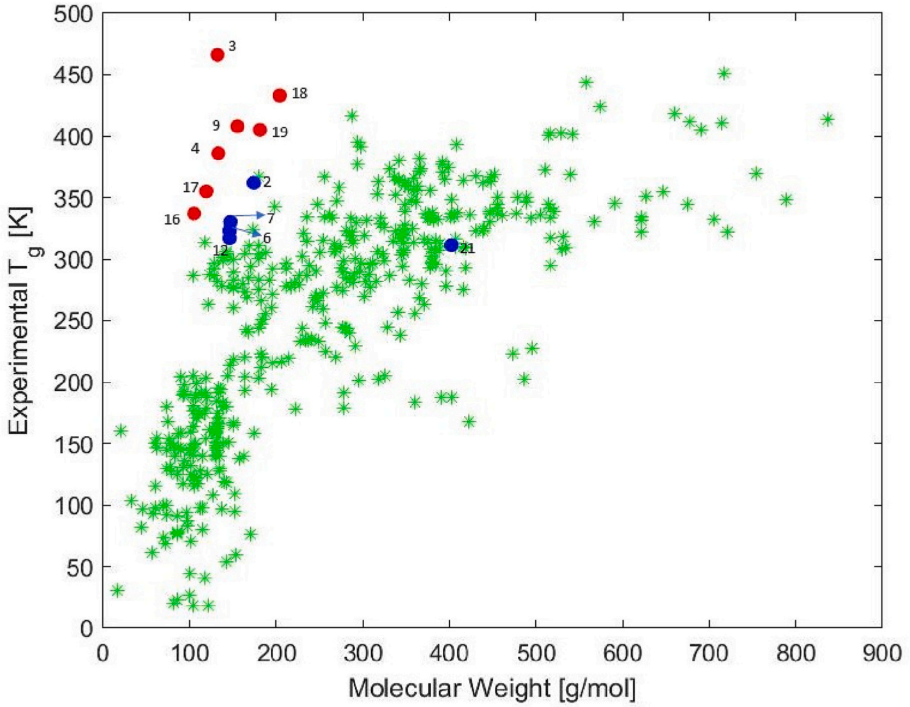


Fig. 9. Distribution of the amino acids and 3-Lys within the training set chemical space (green). The red dots represent molecules that lay outside of the confidence area of the neural network, while the blue dots are well represented by the dataset and lay inside the confidence area of the neural network. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

door to using numerical approaches to estimate the glass transition temperature of complex molecular glass formers, especially when its experimental determination is difficult or even before synthesizing them.

4. Conclusions

We have presented in this work a dataset of organic molecular glass formers with their T_g,which has been used to train an RNN with a Bi-LSTM architecture. We have shown that the network can detect patterns from SMILES strings and correlate them with the corresponding molecule's physical property, in this case, the T_g. We have observed the result of such learning by embedding the activations of the neurons of the last layer into a T_g-oriented m-dimensional space and analysing them by clusterization and PCA. We further have shown that it is possible to predict the T_g of other complex molecules and that such predictions are accurate when the molecules lay in the confidence area of the model. In particular, we have led this analysis on the group of 20 essential amino acids and a short peptide (3-Lys). Finally, we have shown that this kind of architecture is a powerful tool for exploring and designing new materials and correlating macroscopic physical properties to the corresponding molecular structure.

CRediT authorship contribution statement

Claudia Borredon: Data curation, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Luis A. Miccio:** Methodology, Conceptualization, Writing – original draft, Writing – review & editing. **Silvina Cerveny:** Validation, Writing – review & editing, Project administration, Funding acquisition. **Gustavo A. Schwartz:** Conceptualization, Methodology, Writing – review & editing, Project administration, Funding acquisition.

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

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

Acknowledgements

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 the publication fee by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URICI). We also acknowledge the support of NVIDIA Corporation with the donation of two GPUs used for this research.

Appendix A. Supplementary data

The dataset and a description of its composition in terms of T_g

- the description of the elbow method

- the experimental procedure with which the T_g of 3-lys was measured. Supplementary data to this article can be found online at [https://doi.org/10.1016/j.nocx.2023.100185].

References

[1] A.R. Katritzky, V.S. Lobanov, M. Karelson, QSPR: the correlation and quantitative prediction of Chemical and physical properties from structure, Chem. Soc. Rev. 24 (4) (1995) 279–287, https://doi.org/10.1039/CS9952400279.

[2] M. Karelson, V.S. Lobanov, A.R. Katritzky, Quantum-Chemical descriptors in QSAR/QSPR studies, Chem. Rev. 96 (3) (1996) 1027–1044, https://doi.org/10.1021/cr950202r.

[3] A.R. Katritzky, M. Karelson, V.S. Lobanov, QSPR as a means of predicting and understanding Chemical and physical properties in terms of structure, Pure Appl. Chem. 69 (2) (1997) 245–248, https://doi.org/10.1351/pac199769020245.

[4] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being Earnest: validation is the absolute essential for successful application and interpretation of QSPR models, QSAR & Combinator. Sci. 22 (1) (2003) 69–77, https://doi.org/10.1002/qsar.200390007.

[5] T. Le, V.C. Epa, F.R. Burden, D.A. Winkler, Quantitative structure–property relationship modeling of diverse materials properties, Chem. Rev. 112 (5) (2012) 2889–2919, https://doi.org/10.1021/cr200066h.

[6] H. Liu, Z. Fu, K. Yang, X. Xu, M. Bauchy, Machine learning for glass science and engineering: a review, J. Non-Cryst. Solids 557 (2021), 119419, https://doi.org/10.1016/j.jnoncrysol.2019.04.039.

[7] E. Alcobaça, S.M. Mastelini, T. Botari, B.A. Pimentel, D.R. Cassar, A.C.P.L.F. de Carvalho, E.D. de Zanotto, Explainable machine learning algorithms for predicting glass transition temperatures, Acta Mater. 188 (2020) 92–100, https://doi.org/10.1016/j.actamat.2020.01.047.

[8] J.A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, A. Tkatchenko, Combining machine learning and computational chemistry for predictive insights into Chemical systems, Chem. Rev. 121 (16) (2021) 9816–9872, https://doi.org/10.1021/acs.chemrev.1c00107.

[9] W.X. Shen, X. Zeng, F. Zhu, Y. li Wang, C. Qin, Y. Tan, Y.Y. Jiang, Y.Z. Chen, Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations, Nat. Mach. Intell. 3 (4) (2021) 334–343, https://doi.org/10.1038/s42256-021-00301-6.

[10] Z. Tan, Y. Li, W. Shi, S. Yang, A multitask approach to learn molecular properties, J. Chem. Inf. Model. 61 (8) (2021) 3824–3834, https://doi.org/10.1021/acs.jcim.1c00646.

[11] J. Deng, Z. Yang, I. Ojima, D. Samaras, F. Wang, Artificial intelligence in drug discovery: applications and techniques, Brief. Bioinform. 23 (1) (2022) bbab430, https://doi.org/10.1093/bib/bbab430.

[12] M.G. Abiad, M.T. Carvajal, O.H. Campanella, A review on methods and theories to describe the glass transition phenomenon: applications in food and pharmaceutical products, Food Eng. Rev. 1 (2) (2009) 105–132, https://doi.org/10.1007/s12393-009-9009-1.

[13] D. Champion, M. Le Meste, D. Simatos, Towards an improved understanding of glass transition and relaxations in foods: molecular mobility in the glass transition range, Trends Food Sci. Technol. 11 (2) (2000) 41–55, https://doi.org/10.1016/S0924-2244(00)00047-9.

[14] N.R. Jadhav, V.L. Gaikwad, K.J. Nair, H.M. Kadam, Glass transition temperature: basics and application in pharmaceutical sector, Asian J. Pharmaceut. (AJP): Free Full Text Articles From Asian J. Pharm. 3 (2) (2014), https://doi.org/10.22377/ajp.v3i2.246.

[15] L.-P. Blanchard, J. Hesse, S.L. Malhotra, Effect of molecular weight on glass transition by differential scanning calorimetry, Can. J. Chem. 52 (18) (1974) 3170–3175, https://doi.org/10.1139/v74-465.

[16] M.J. Richardson, N.G. Savill, Derivation of accurate glass transition temperatures by differential scanning calorimetry, Polymer 16 (10) (1975) 753–757, https://doi.org/10.1016/0032-3861(75)90194-9.

[17] U. Schneider, P. Lunkenheimer, A. Pimenov, R. Brand, A. Loidl, Wide range dielectric spectroscopy on glass-forming materials: an experimental overview, Ferroelectrics 249 (1) (2001) 89–98, https://doi.org/10.1080/00150190108214970.

[18] F. Kremer, Dielectric spectroscopy – yesterday, today and tomorrow, J. Non-Cryst. Solids 305 (1) (2002) 1–9, https://doi.org/10.1016/S0022-3093(02)01083-9.

[19] Y. Zhang, S. Katira, A. Lee, A.T. Lambe, T.B. Onasch, W. Xu, W.A. Brooks, M. R. Canagaratna, A. Freedman, J.T. Jayne, D.R. Worsnop, P. Davidovits, D. Chandler, C.E. Kolb, Kinetically controlled glass transition measurement of organic aerosol thin films using broadband dielectric spectroscopy, Atmos. Measurem. Techniq. 11 (6) (2018) 3479–3490, https://doi.org/10.5194/amt-11-3479-2018.

[20] C.B. Holmes, M.E. Cates, M. Fuchs, P. Sollich, Glass transitions and shear thickening suspension rheology, J. Rheol. 49 (1) (2005) 237–269, https://doi.org/10.1122/1.1814114.

[21] H.G. Weyland, P.J. Hoftyzer, D.W. Van Krevelen, Prediction of the glass transition temperature of polymers, Polymer 11 (2) (1970) 79–87, https://doi.org/10.1016/0032-3861(70)90028-5.

[22] E. Donth, Characteristic length of glass transition, J. Non-Cryst. Solids 131–133 (1991) 204–206, https://doi.org/10.1016/0022-3093(91)90300-U.

[23] C.A. Angell, Formation of glasses from liquids and biopolymers, Science 267 (5206) (1995) 1924–1935, https://doi.org/10.1126/science.267.5206.1924.

[24] W. Liu, C. Cao, Artificial neural network prediction of glass transition temperature of polymers, Colloid Polym. Sci. 287 (7) (2009) 811–818, https://doi.org/10.1007/s00396-009-2035-y.

[25] D.R. Cassar, A.C.P.L.F. de Carvalho, E.D. Zanotto, Predicting glass transition temperatures using neural networks, Acta Mater. 159 (2018) 249–256, https://doi.org/10.1016/j.actamat.2018.08.022.

[26] A. Jha, A. Chandrasekaran, C. Kim, R. Ramprasad, Impact of dataset uncertainties on machine learning model predictions: the example of polymer glass transition temperatures, Model. Simul. Mater. Sci. Eng. 27 (2) (2019), 024002, https://doi.org/10.1088/1361-651X/aaf8ca.

[27] L.A. Miccio, G.A. Schwartz, From Chemical structure to quantitative polymer properties prediction through convolutional neural networks, Polymer 193 (2020), 122341, https://doi.org/10.1016/j.polymer.2020.122341.

[28] L.A. Miccio, G.A. Schwartz, Localizing and quantifying the intra-monomer contributions to the glass transition temperature using artificial neural networks, Polymer 203 (2020), 122786, https://doi.org/10.1016/j.polymer.2020.122786.

[29] L. Tao, V. Varshney, Y. Li, Benchmarking machine learning models for polymer informatics: an example of glass transition temperature, J. Chem. Inf. Model. 61 (11) (2021) 5395–5413, https://doi.org/10.1021/acs.jcim.1c01031.

[30] L.A. Miccio, G.A. Schwartz, Mapping chemical structure–glass transition temperature relationship through artificial intelligence, Macromolecules 54 (4) (2021) 1811–1817, https://doi.org/10.1021/acs.macromol.0c02594.

[31] D. Weininger, SMILES, a Chemical Language and Information System. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1) (1988) 31–36, https://doi.org/10.1021/ci00057a005.

[32] G. Landrum, RDKit Documentation (2019).

[33] G.B. Goh, N. Hodas, C. Siegel, A. Vishnu, SMILES2vec: Predicting Chemical Properties from Text Representations, 2018.

[34] G. Chen, L. Tao, Y. Li, Predicting Polymers' glass transition temperature by a Chemical Language processing model, Polymers 13 (11) (2021) 1898, https://doi.org/10.3390/polym13111898.

[35] H. Abdi, L.J. Williams, Principal component analysis, WIREs Computat. Statist. 2 (4) (2010) 433–459, https://doi.org/10.1002/wics.101.

[36] E.H. Ruspini, J.C. Bezdek, J.M. Keller, Fuzzy clustering: a historical perspective, IEEE Comput. Intell. Mag. 14 (1) (2019) 45–55, https://doi.org/10.1109/MCI.2018.2881643.

[37] H. Tam Do, Y. Zen Chua, A. Kumar, D. Pabsch, M. Hallermann, D. Zaitsau, C. Schick, C. Held, Melting properties of amino acids and their solubility in water, RSC Adv. 10 (72) (2020) 44205–44215, https://doi.org/10.1039/D0RA08947H.

[38] Private Communication (2023).