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

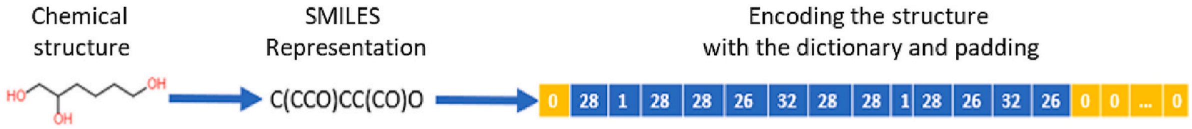

Fig. 1. Encoding the SMILES with cardinal encoding. We added a 0 at the beginning of each string and completed them with a padding of 0 s to have the same length for all the instances.

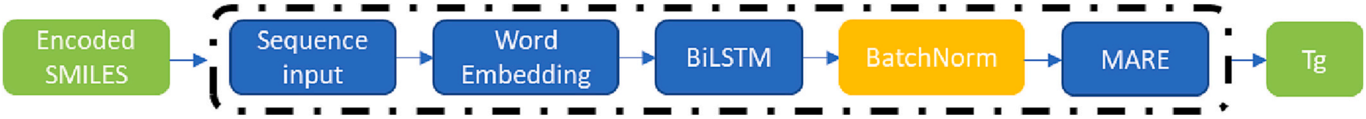

Fig. 2. The ANN architecture comprises a Sequence Input layer, a Word Embedding layer, a Bidirectional LSTM layer, a Batch Normalisation layer, and a mean absolute relative error output layer. It takes as an input the encoded SMILES and outputs the T_g of the molecule(s).

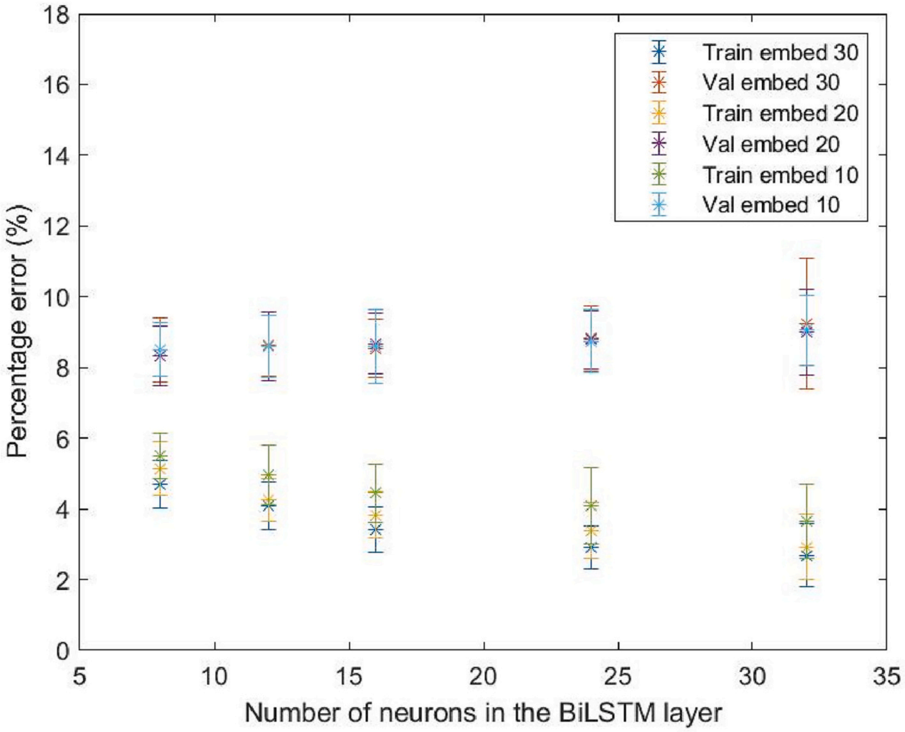

Fig. 3. Architecture test: we tested different values of neurons in the BiLSTM layer (from 8 to 32 nodes) and several values of the word embedding dimension (10, 20, 30).

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

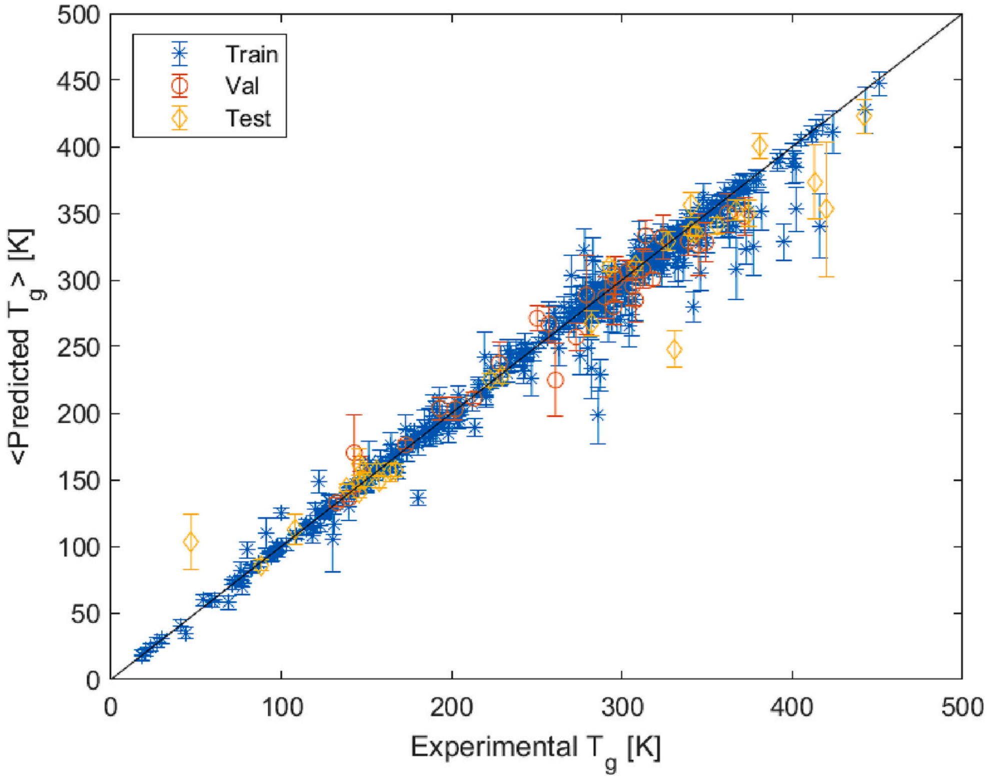

Fig. 4. Average prediction of the glass transition temperature for the training (blue), validation(orange) and test(yellow) set. The mean standard deviation for the training, validation and test set are 7 K, 13 K, and 9 K, respectively. (For interpretation of the references to colour in this figure legend,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

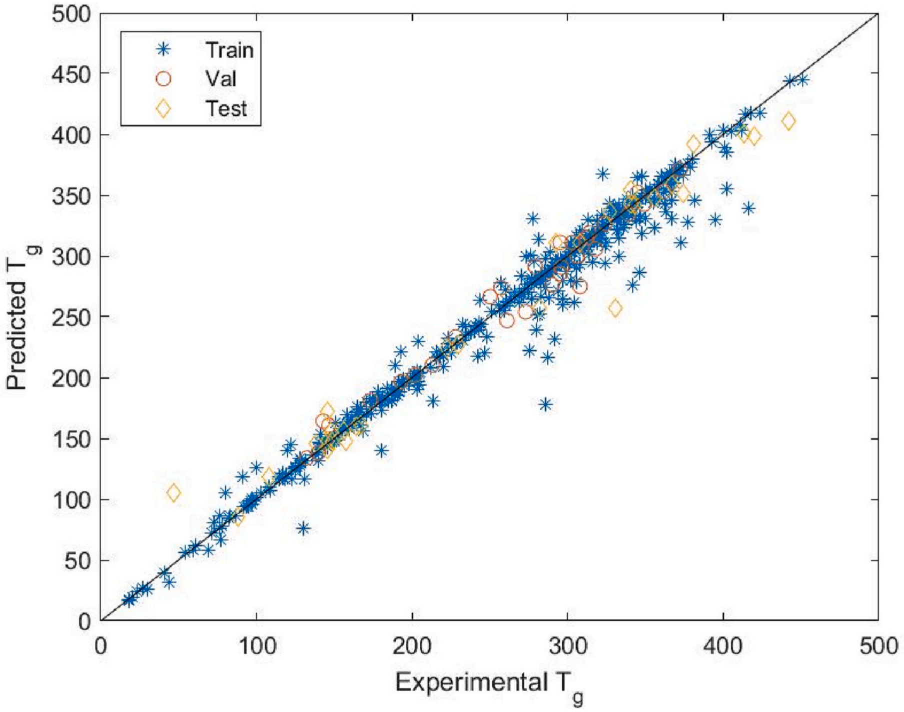

Fig. 5. Performance of the chosen neural network for the training (blue), validation (orange) and test (yellow) sets. The data points lay almost perfectly on the bisector axis, indicating an excellent agreement between experimental and predicted T_g. The MAPE values obtained are 3.4%, 3.8% and 8.7% for 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.)

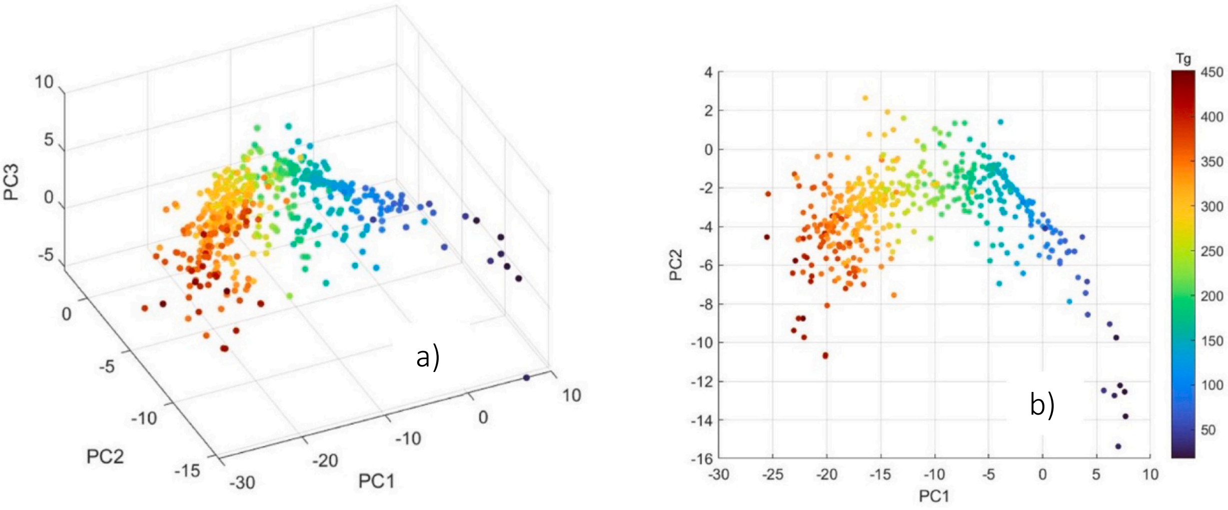

Fig. 6. PCA projection of the batch normalisation layer activations. We use a colour map to enhance the trend of the glass transition temperature, which goes from blue colours (low T_g) to red colours (high T_g). (For interpretation of the references to colour in this figure legend, the reader is referred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

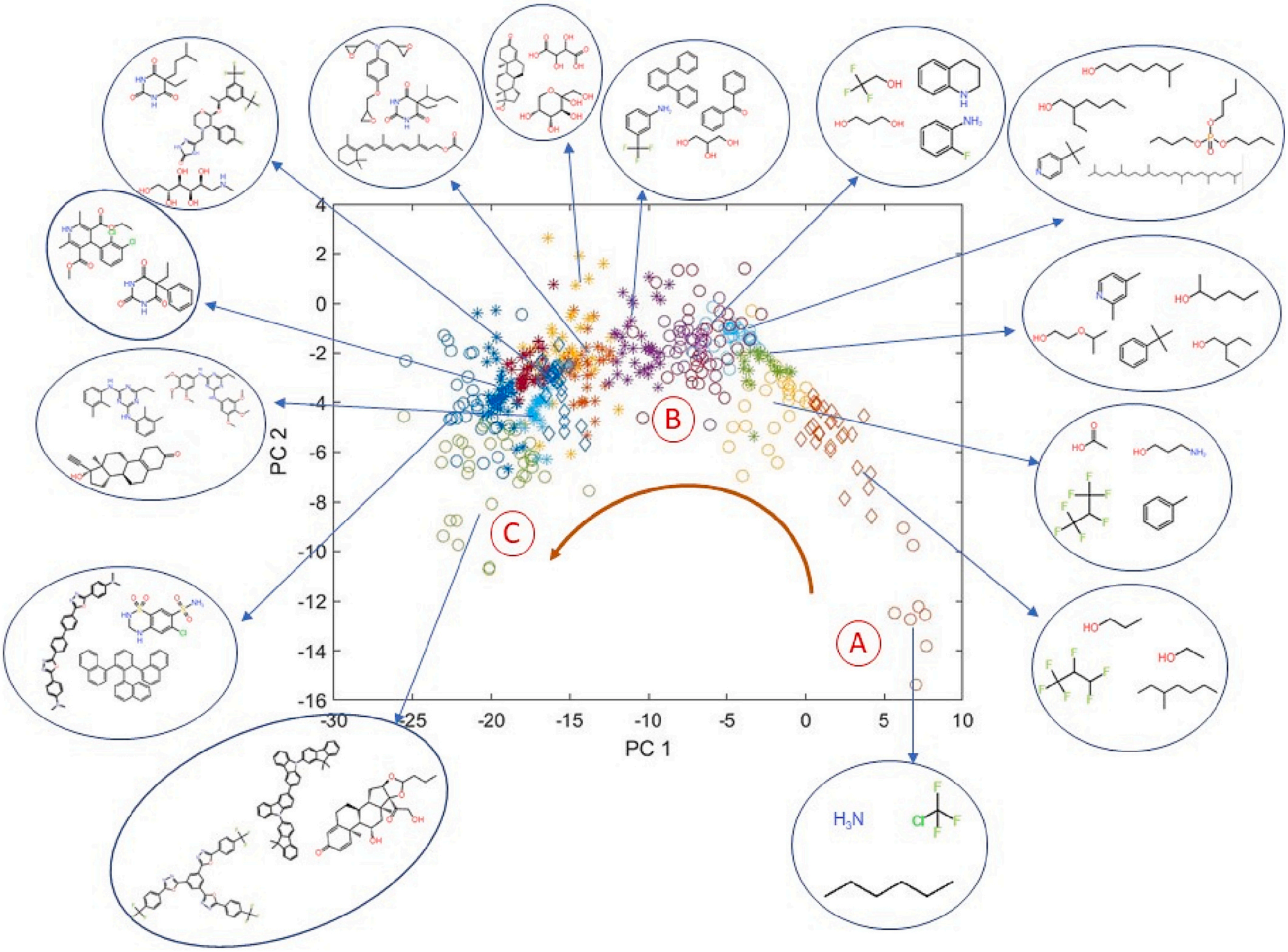

Fig. 7. We clustered the chemical structures within the m-dimensional space using the fuzzy C algorithm. In this way, we can observe the structural changes along the trajectory of the PCA, going from low-molecular-weight, linear chains and weak intermolecular forces (A) to high-molecular-weight, a higher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

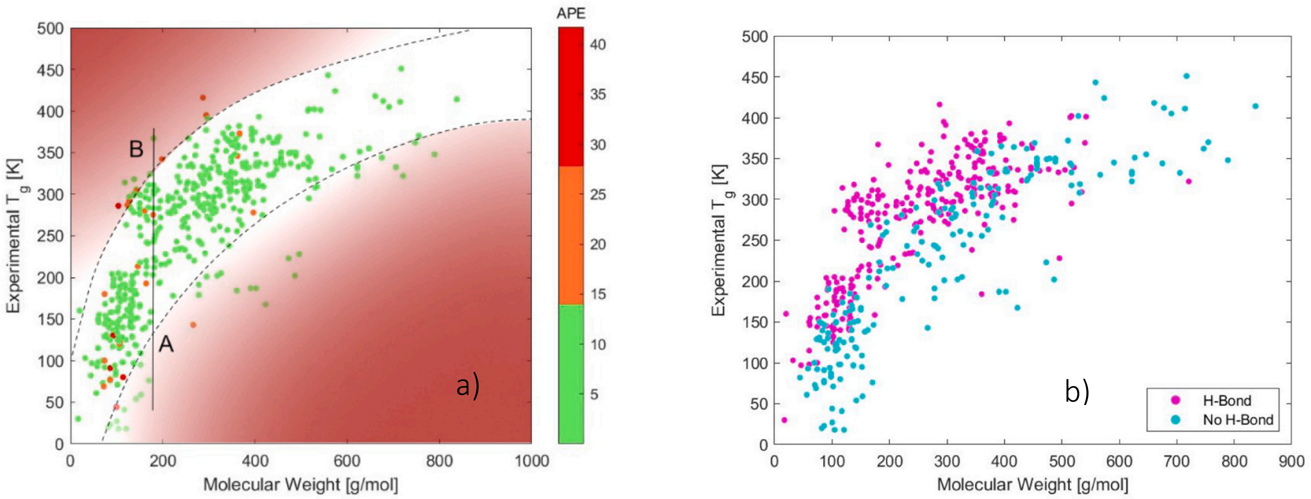

Fig. 8. Molecular weight dependence of the glass transition temperature for the training set molecules. The colour map in Fig. 8a represents the absolute percentage error (APE) when predicting the T_g of the compound. The area between the dashed lines represents the confidence interval of the neural 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

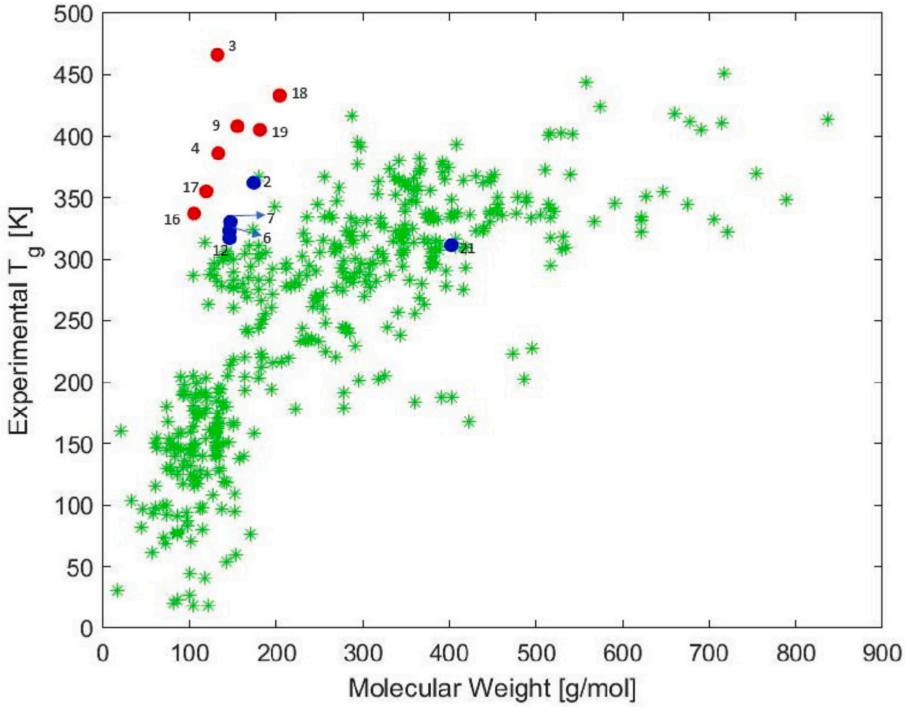

Fig. 9. Distribution of the amino acids and 3-Lys within the training set chemical space (green). The red dots represent molecules that lay outside of the confidence area of the neural network, while the blue dots are well represented by the dataset and lay inside the confidence area of the neural network. (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).