Supporting Information:

Predicting Glass Transition Temperature and Melting Point of Organic Compounds via Machine Learning and Molecular Embeddings

Tommaso Galeazzo1,* and Manabu Shiraiwa1,*

1. Department of Chemistry, University of California, Irvine, CA92697, USA

* Correspondence to: tommaso.galeazzo@gmail.com; m.shiraiwa@uci.edu
1. **tgBoost model**

1.1 Comparison with extrapolated $T_g$ from viscosity measurements

Rothfuss and Petters (1) derived $T_g$ for atmospherically-relevant organic compounds by viscosity measurements followed by extrapolation using the Vogel-Fulcher-Tammann (VFT) equation. $T_g$ of these compounds are also available in the dataset in Koop et al. (2). Figure S1 shows the correlation plots between: a) $T_g$ experimental measurements (2) and predictions from the tgBoost model, b) $T_g$ extrapolated from viscosity measurements (1) and predictions from the tgBoost model, and c) $T_g$ experimental measurements (2) and $T_g$ extrapolated from viscosity measurement (1). As shown in Fig. S1a, the tgBoost model reproduces experimental $T_g$ very well, while there are some deviations for extrapolated $T_g$ from viscosity measurements (Fig. S1b), reflecting differences between extrapolated and measured values (Fig. S1c). It is worth noting that $T_g$ has been extrapolated from viscosity measurements where the experimental measurements were very low in viscosity value (i.e., $\eta < 10$ Pa s, which is very far from the glassy state with $\eta = 10^{12}$ Pa s) for a number of species. In this case, it is difficult to accurately estimate $T_g$ by extrapolation of viscosity measurements using the VFT equation.

1.2 Comparison of tgBoost predictions to $T_g$ estimations from O:C and $C^0$

Li et al. (3) developed $T_g$ parameterizations based on molecular O:C ratio and $C^0$. They compared their $T_g$ predictions with $T_g$ estimated through the Boyer-Kauzman rule on $T_m$ from EPI Suite for compounds in the dataset from Shiraiwa et al. (4). Their estimated absolute mean percentage relative error (MPE) (i.e., defined as AAVRE in their study) is 6% with $R = 0.96$. We have compared the $T_g$ predicted by the tgBoost model with $T_g$ estimated through the Boyer-Kauzman rule (2) on $T_m$ from EPI Suite for compounds in the dataset from Shiraiwa et al. (4). Our MPE is 10.6% with $R = 0.8$. Note that, the $C^0$ used by Li et al. (3) was evaluated using the EVAPORATION model by Compernolle et al. (5), and that the $T_g$ values estimated from $T_m$ evaluated by EPI Suite were based on the MPBPWIN model. Both MPBPWIN and EVAPORATION are QSAR models developed on a mix of experimental measurements and model predictions and the models use chemical species boiling points to build their QSAR. Both models use a combination of different methods, but they are both using derivations of the Antoine equation. As a result, both MPBPWIN and EVAPORATION have predictions strongly correlated to boiling point values. These approaches might introduce a correlation bias based on the similar estimation methods and linked to the same variable implicitly used for the prediction. As a result, even if $C^0$ is suited to predict $T_g$, there might a correlation bias to account
when comparing the estimated MPE of the two methods in relation to the $T_g$ estimated from $T_m$ evaluated by EPI Suite.

2. **$T_m$ regression models with additional datasets**

   We have developed two additional $T_m$ regressors using separate datasets to compare the performances of molecular embeddings in the $T_m$ regression. The first dataset is the “Bradley good melting point dataset” (i.e. $T_m$-Bradley) which is a highly curated dataset of experimental melting points of drug-like compounds (6). The second dataset has been generated using the $T_m$ of environmentally relevant compounds by Wei et al. (7) and evaluated using MPBPWIN by the EPI Suite Software (8) (i.e. $T_m$-EPI). The $T_m$-Bradley dataset contains 3041 entries, which is reduced to 3025 compounds after cleaning. The $T_m$-EPI dataset contains 29488 entries and 29487 compounds after cleaning. A summary of the specific datasets used in this study and their properties is reported in Table S1. We have developed a Deep Neural Network regressor (DNN) for the $T_m$-Bradley dataset and tested three models (i.e., Random Forest, RF; Extreme Gradient Boosting, XGBoost; Deep Neural Network, DNN) for the regression task of the $T_m$-EPI dataset. The best performances of the $T_m$ developed models are reported in Table S2.

   The DNN model trained on the $T_m$-Bradley dataset has a MAE of 32.3 K, slightly above the results from the gold standard models for $T_m$ prediction. The model has a positive correlation of $R = 0.76$ and a variance of $R^2_{CV} = 0.89$. It is important to note that the state-of-the-art models for $T_m$ regression are built on top of very complex architectures such as convolutional neural networks (CNN) (9), a combination of a neural network and an associative ensemble step (ASNN) (10), and a Gaussian Process with dataset specific embeddings (11). No significant difference was observed between the performance of the DNN model developed on the $T_m$-Bradley dataset and the DNN developed on top of the $T_m$-Tetko dataset with MAE = 31.0 K. Remarkably, Tetko et al. measured a similar Root Mean Squared Error (RMSE) for ASNN models developed on both the Bradley good melting point dataset and their curated $T_m$-Tetko dataset (10,12). These similar performances of the ASNN architecture on the two datasets suggest that more complex model architectures are needed to predict with higher accuracy the trends from experimental $T_m$. The $T_m$ regression DNN model developed on the Tm-EPI dataset performs really well with an MAE of 12.3 K, a positive correlation of $R = 0.94$ and a variance of $R^2_{CV} = 0.97$. This result demonstrates a good performance of molecular embeddings in reproducing the algorithm of MPBPWIN, but its predictions remain strongly linked to the limitations of the prediction module of the EPI Suite.
Figure S1: Correlation plots between a) $T_g$ experimental measurements and predictions from the $T_g$ tgBoost model, a) $T_g$ extrapolated from viscosity measurements (Rothfuss and Petters, 2016) and predictions from the tgBoost model, and b) $T_g$ experimental measurements and $T_g$ extrapolated from viscosity measurement.

Figure S2: SOA compounds from the dataset of Shiraiwa et al. (4) with the 5 highest deviations between the $T_g$ predicted by the tgBoost model and the $T_g$ predicted by the MPBPWIN module from the EPI Suite.

Figure S3: SOA compounds from the dataset of Shiraiwa et al. (4) with the 5 highest deviations between the $T_g$ predicted by the compositional parametrization and the $T_g$ predicted by the MPBPWIN module from the EPI Suite.
Figure S4: Estimated $T_g$ of alkanes and esters as a function of the number of carbon atoms within the molecule. The functional groups in ethers are positioned at the end of the alkyl chain.
Table S1: A summary of the datasets used to develop the additional $T_m$ models for comparison of molecular embeddings results.

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Author</th>
<th>Data</th>
<th>Initial entries</th>
<th>Final entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_m$-Bradley</td>
<td>Bradley et al., (2014) (6)</td>
<td>$T_m$, experimental</td>
<td>3041</td>
<td>3025</td>
</tr>
<tr>
<td>$T_m$-EPI</td>
<td>Wei et al., (2012) (7)</td>
<td>$T_m$, EPI Suite estimated</td>
<td>29488</td>
<td>29487</td>
</tr>
</tbody>
</table>

Table S2: Comparison of the performances in the regression tasks of developed models on the additional $T_m$ datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Algorithm*</th>
<th>MAE (K)</th>
<th>RMSE</th>
<th>$R^2_{CV}$</th>
<th>$R$</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_m$-EPI</td>
<td>RF</td>
<td>15.9</td>
<td>25.7</td>
<td>0.90</td>
<td>0.95</td>
<td>This work</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>20.5</td>
<td>30.4</td>
<td>0.86</td>
<td>0.93</td>
<td>This work</td>
</tr>
<tr>
<td></td>
<td>DNN</td>
<td>12.3</td>
<td>19.2</td>
<td>0.94</td>
<td>0.97</td>
<td>This work</td>
</tr>
<tr>
<td>$T_m$-Bradley</td>
<td>DNN</td>
<td>31.3</td>
<td>41.9</td>
<td>0.89</td>
<td>0.76</td>
<td>This work</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>26.2</td>
<td>35.5</td>
<td></td>
<td></td>
<td>(9) †</td>
</tr>
<tr>
<td></td>
<td>ASNN</td>
<td>32.0</td>
<td></td>
<td></td>
<td></td>
<td>(10) †</td>
</tr>
<tr>
<td></td>
<td>GPR</td>
<td>28.85</td>
<td>0.78</td>
<td></td>
<td></td>
<td>(11) †</td>
</tr>
</tbody>
</table>

*CNN = Convolutional Neural Network, GPR = Gaussian Process Regression, ASNN = Adversarial Neural Network. † The datasets used in these studies are all different variations of the “Bradley Good Melting Points Dataset” from Bradley at al. (6).
References


