ChemComm

COMMUNICATION

View Article Online
View Journal | View Issue

Cite this: *Chem. Commun.*, 2021, **57**, 2633

Received 10th November 2020,
Accepted 11th January 2021

DOI: 10.1039/d0cc07384a

rsc.li/chemcomm

Melting point prediction of organic molecules by deciphering the chemical structure into a natural language†

Weiming Mi, ‡[a] Huijun Chen,‡[b] Donghua (Alan) Zhu, *[c] Tao Zhang*[a] and Feng Qian *[b]

Establishing quantitative structure–property relationships for the rational design of small molecule drugs at the early discovery stage is highly desirable. Using natural language processing (NLP), we proposed a machine learning model to process the line notation of small organic molecules, allowing the prediction of their melting points. The model prediction accuracy benefits from training upon different canonicalized SMILES forms of the same molecules and does not decrease with increasing size, complexity, and structural flexibility. When a combination of two different canonicalized SMILES forms is used to train the model, the prediction accuracy improves. Largely distinguished from the previous fragment-based or descriptor-based models, the prediction accuracy of this NLP-based model does not decrease with increasing size, complexity, and structural flexibility of molecules. By representing the chemical structure as a natural language, this NLP-based model offers a potential tool for quantitative structure–property prediction for drug discovery and development.

The advances of various innovative chemical biology and medicinal chemistry technologies have prompted the identification of new therapeutic targets, as well as vastly expanded chemical libraries for accelerated drug discovery.[1–4] Although the identification of chemical binders with pharmacological targets is the first critical step towards successful drug discovery, the subsequent pharmaceutical "develop-ability" assessment of lead compounds often presents another key hurdle preventing the transformation of discoveries into drug products.[5–7] Using small organic molecules as common examples, their physicochemical properties, including the solubility, hydrophobicity, cell membrane permeability, and stability, could significantly impact their "drug-likeness".[5–7] Therefore, any tools allowing early prediction of such molecular properties or rational design of compound libraries based on the properties mentioned above would be highly desirable.

The melting point (T_m) of small organic molecules specifies the temperature wherein an orderly packed, 3D crystal lattice reaches thermodynamic equilibrium with its corresponding disordered melt, thus converting from the solid to the liquid state. T_m is intimately correlated with some important properties of pharmaceutical compounds, especially the solubility, arguably the most critical physiochemical property of small molecular drugs.[8,9] In this study, we proposed a novel T_m prediction model (More information about the model see Method in ESI†) based on natural language processing (NLP), without any input of chemical descriptors requiring additional computational or experimental efforts, due to the following considerations:

First, despite the ease of the experimental determination of T_m, there is still a practical need for rapid T_m prediction without any synthesized materials to allow the rational profiling of compound libraries with a tremendous amount of new chemical entities and the early assessment of the drug-like properties of target compounds to avoid downstream developability challenges.[8–11]

Second, a variety of existing T_m prediction models can be classified into a descriptor-based or fragment-based approach,[12–17] and both rely on prior quantum and solid-state chemistry theories and contain some hard-to-understand chemical descriptors. In the current NLP approach, 3D chemical structures are topologically represented by line notations of chemical features, using a simplified molecular-input line-entry system (SMILES) as the only input for the model. SMILES is the most widely used line notation in chemical nomenclature, similarity search, and data exchange for its handy readability and writability by both humans and computers. SMILES stores the structural information of atoms, bonds, connectivity,

[a] *Department of Automation, Tsinghua University, Beijing National Research Center for Information Science and Technology, Beijing 100084, P. R. China. E-mail: zhangtao@tsinghua.edu.cn*
[b] *School of Pharmaceutical Sciences, Tsinghua University, Beijing 100084, P. R. China. E-mail: qianfeng@tsinghua.edu.cn*
[c] *Pharmaceutical Product Development & Supply, Chemical Pharmaceutical Development & Supply, Janssen Research & Development, Johnson & Johnson, Shanghai 200233, P. R. China. E-mail: dzhu7@its.jnj.com*

† Electronic supplementary information (ESI) available: Discussion, methods, and some major exported results. See DOI: 10.1039/d0cc07384a

‡ W. M. and H. C. contributed equally to this work.

aromaticity, and stereochemistry in a linguistic construct using ASCII strings. Practically, conversion of 3D chemical structures into SMILES could be easily achieved, while a SMILES string could correspond to many 3D molecular structures.[17]

Last, NLP is a process that transfers natural language to computers, which has been widely used in the fields of automatic translation, speech recognition, chatbots, *etc.*[18] The study of natural language relies on the relationship between words, given the premise that words with similar meanings tend to appear in similar contexts and patterns. The idea is quite similar in establishing quantitative structure–property relationships, which assumes that molecules with similar chemical structures tend to possess similar properties. Similarly, some other properties of molecules, such as chemical stability, toxicity,[19] and nuclear magnetic resonance spectroscopy,[20] have been also analysed by NLP-based models.

Our model is predicting the structure-related properties of compounds solely through artificial intelligence-based linguistic methodologies, which have no apparent linkage with any existing chemistry theories. The feasibility and prediction accuracy of the model was evaluated and compared to previously reported fragment-based or descriptor-based models. Then, the impact of the SMILES forms and molecular size and complexity on the prediction accuracy was analysed. Since the major objective of this work was to demonstrate successful T_m prediction by the NLP approach solely from SMILES strings, the impact of polymorphism on the melting point was not considered. Molecules with very different T_m values (range > 5 °C) in the dataset, possibly due to polymorphs or experimental variabilities, were thus excluded on purpose.

In Fig. 1a, the experimental and predicted T_m values for both the training and test data are shown. A significant correlation was found for a majority of compounds in the dataset, with correlation coefficients (R^2) of 0.8746 and 0.8103 for the training and test data, respectively. The absolute mean errors (MAEs) were 23.87 °C and 30.00 °C for the training and test data, respectively, while the root mean square errors (RMSEs) were 32.34 °C and 39.04 °C, respectively. The chemical structures of several of the worst predicted molecules from the test data set are shown in Fig. 1b.

Residual analysis was performed by plotting the T_m residuals against the predicted T_m to check the appropriateness of the model. The residuals were almost symmetrically distributed along the *y*-axis, and no obvious pattern was found (Fig. 1c), confirming that the regression model was appropriate. Nevertheless, when T_m residuals were plotted against the experimental T_m (Fig. 1d), there was a mild decreasing trend of the residuals with increasing experimental T_m, indicating an overestimation for compounds with a low experimental T_m and underestimation for those with a high experimental T_m.

Since the SMILES string is the only input of the model, the way SMILES is incorporated into the model could play a critical role in the model performance. In general, one SMILES string corresponds to only one specific chemical structure. However, depending on the selection of the starting point, main chain, and open-loop position, multiple SMILES strings exist for a given structure, and canonicalized SMILES strings generated by different toolkits for one given structure might differ significantly[21,22] (Fig. 2a). It is similar to natural language, where one could use different ways of expression to convey the same meaning. For example, we could say 'We knew nothing.' or 'Nothing did we know.', wherein the inversion of the verb and subject does not change the meaning of the sentence. Hence, it is important to know whether our model could not only distinguish different molecules whereby analysing the SMILES strings but also recognize the same molecule from the manifold with an equivalent SMILES expression. Therefore, the impact of two canonicalized SMILES forms (Open Babel and RDKit) on the prediction accuracy was investigated by incorporating into the models in different combinations (Fig. 2b).

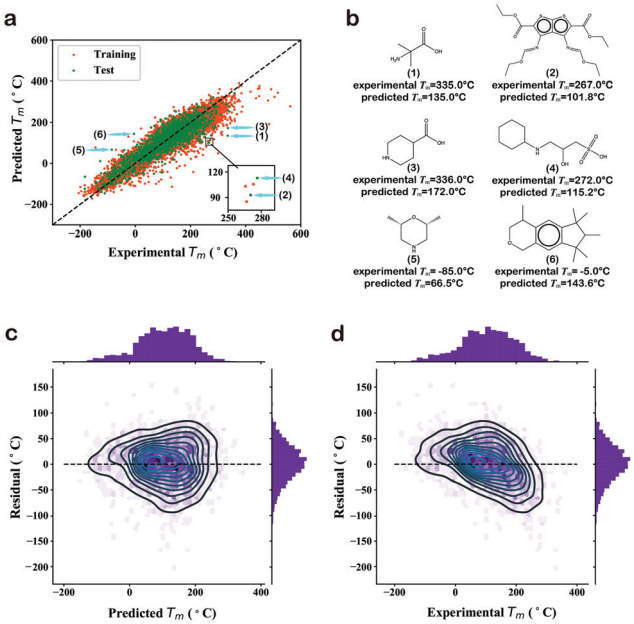

Fig. 1 Model prediction performance. (a) The joint distribution of predicted T_m and experimental T_m values on both the training data and test data. Six of the worst predicted molecules are marked from (1) to (6). (b) Chemical structures, experimental T_m, and predicted T_m of the outliers stated in (a). (c) The joint distribution of predicted T_m and T_m residuals of the molecules in the test data. (d) The joint distribution of experimental T_m and T_m residuals of molecules in the test data. (Note: In (c or d), the high-value area and low-value area shown in the contour plots represent areas with dense and sparse data points, respectively.)

The results (Table 1 and Fig. 2c) of our model showed that when only one SMILES form was imported into the model, the model had difficulty predicting another SMILES form that was not previously imported, resulting in a significant decrease in the prediction accuracy (A1 *vs.* A3 and A2 *vs.* A4). When multiple forms of canonicalized SMILES were imported into the model for training, the prediction accuracy was comparable for all of the form, and the overall performance regarding the RMSE and MAE was improved (B1 *vs.* A1 and B2 *vs.* A2), with a similar or (to some extent) superior effect as the sample size increased (B1 and B2 *vs.* C).

We hypothesized that the input of diverse canonicalized SMILES forms might help the model to find not only the

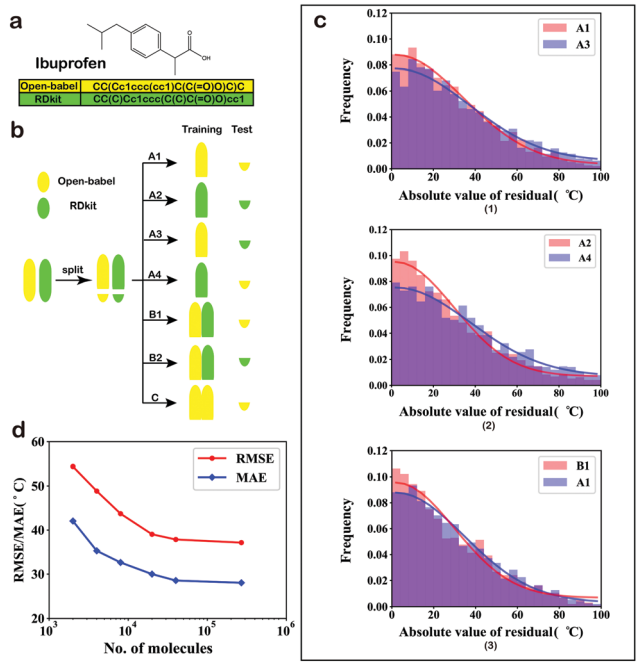

Fig. 2 Impact of different canonicalized SMILES forms and sample sizes on the prediction accuracy. (a) The chemical structure and canonicalized SMILES strings of ibuprofen in the two toolkits: Open Babel and RDKit. (b) The combination of different canonicalized SMILES forms of the input in the training and test data. (c) The frequency distribution difference of the absolute values of the T_m residuals caused by different canonicalized SMILES forms. From 2c(1) to 2c(3), each figure shows the frequency distribution histogram and the frequency Gaussian regression curve of the absolute values of the T_m residuals in the corresponding two experiments. The purple area in each figure is the overlap of the histogram of the corresponding two experiments, and all the bar widths are 5 °C. (d) The relationship between the prediction performance and the size of the training dataset. As the sample size of the dataset increased, the RMSE and MAE gradually decreased, but the improvement was more obvious when the sample size was less than 20 000, and it levelled off when the sample size was greater than 20 000.

Table 1 Prediction performance of the SMILES-based model using different canonicalized SMILES forms[a]

	R^2 [b]	RMSE (°C)	MAE (°C)
A1	0.81(0.812 ± 0.002)	39.04(38.75 ± 0.34)	30.00(29.89 ± 0.31)
A2	0.82(0.815 ± 0.003)	38.52(38.29 ± 0.38)	29.07(28.99 ± 0.21)
A3	0.75(0.749 ± 0.002)	44.90(44.79 ± 0.31)	34.73(34.52 ± 0.26)
A4	0.74(0.737 ± 0.002)	46.00(45.90 ± 0.33)	35.54(35.33 ± 0.30)
B1	0.83(0.825 ± 0.002)	37.35(37.54 ± 0.29)	28.33(28.27 ± 0.11)
B2	0.83(0.830 ± 0.001)	36.88(37.00 ± 0.16)	27.88(27.94 ± 0.25)
C	0.82(0.822 ± 0.001)	37.86(37.92 ± 0.20)	28.54(28.49 ± 0.18)

[a] See Table S1 (ESI) for a table including training data results. [b] Values in parentheses are average values calculated by five repeated experiments using different training and test data.

sequential alignment or the distance of fragments in the SMILES strings but also other strong relationships between fragments, which could be features that are more intrinsic for T_m prediction. When the size of the dataset available for training is limited, increasing the diversity of SMILES forms could provide us with an alternative way to improve the model prediction accuracy. Furthermore, even with a sufficiently large training set, the amount of information that the model extracted from the SMILES strings could fail to increase with an increase in the SMILES sample size after a certain threshold (Fig. 2d and Table S2, ESI†), indicating that the information residing within the SMILES strings could be exhausted. Therefore, the diverse canonicalized SMILES forms could still offer extra merits to improve the prediction accuracy.

Hence, the similarity between SMILES and natural language entitles the application of NLP approaches to explore the relationship between the structure and T_m. Our model also demonstrated that a successful T_m prediction with an accuracy comparable to or even better than the state of the art (Table 2) in terms of the RMSE, MAE, and R^2 could be achieved solely and directly from SMILES by the application of the NLP approach.

Furthermore, it was reported that the size of molecules, the complexity and flexibility of molecular structures, and the intermolecular interaction could either be related to the T_m or have an impact on the prediction accuracy.[12,23] It raised the question of whether the prediction accuracy of our model was consistent for molecules with various properties. To answer this question, we performed residual analyses on these related properties, including the molecular weight (MW), length of SMILES string, number of branches, number of rotatable bonds, number of hydrogen acceptors, and number of hydrogen donors.

When plotted against MW (Fig. 3a), the T_m residuals of the test data were almost evenly distributed on both sides of the x-axis, and the variation in the T_m residuals did not increase with increasing MW. This fact suggests that the model works almost consistently for compounds with distinct MWs within the range of the dataset. For the length of the SMILES string (Fig. 3b), the T_m residuals were distributed evenly along the x-axis, but the variance of the residuals changed with increasing length. For the number of branches (Fig. 3c), heteroscedasticity was also observed, and the variance of the T_m residuals decreased with an increasing number of branches, suggesting that the prediction accuracy of molecules will improve with longer SMILES string lengths or more branches. Regarding the number of rotatable bonds and the number of hydrogen bond acceptors and donors (Fig. 3d–f), likewise, the T_m residuals

Table 2 Comparison of the melting point prediction accuracy with other state of the art methods

	RMSE (°C)	R^2	MAE (°C)
Our method[a]	36.88	0.83	27.88
Karthikeyan *et al.*[12]	49.8	0.57	N/A
Hughes *et al.*[14]	48.1	0.50	33.8
McDonagh *et al.*[9]	41.3	0.75	N/A
Tetko *et al.*[b 17]	36.8 ± 0.3	N/A	N/A

[a] Test data of B2. [b] The reported RMSE of the model developed by Tetko *et al.*[17] was slightly lower than our model, but in that model, the outliers were excluded, while in our study, the outliers in the training set were kept to maintain the diversity of the molecules.

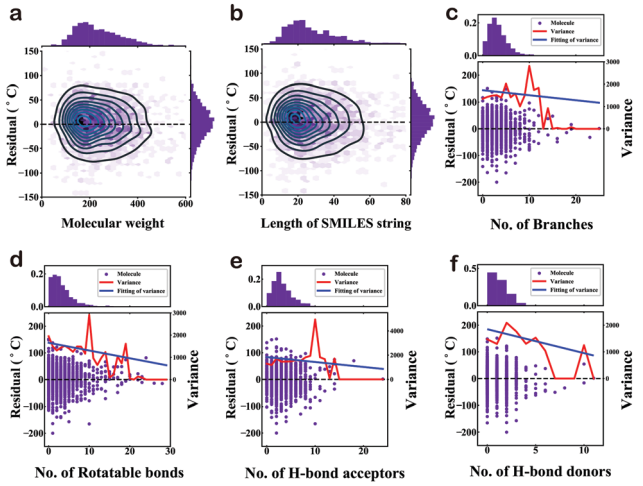

Fig. 3 Impact of the molecular size and complexity on the prediction accuracy. (a–f) The joint distribution of the T_m residual *vs.* the molecular weight, the SMILES string length, the number of branches, the rotatable bond number, the hydrogen bond acceptor number, and the hydrogen bond donor number.

were distributed evenly on the two sides of the *x*-axis, but the variation in the residuals did not increase; instead, they tended to decrease when increasing the number of rotatable bonds or increasing the number of hydrogen bond acceptors or donors. The heteroscedasticity pattern implied that the variance of the T_m residuals slightly decreased with the increasing number of hydrogen bonds or rotatable bonds, which is quite different from previously reported models. These results suggest that our model had uniqueness in the accuracy of its predictions compared to traditional models.

In recent years, although the Transformer[24] has brought unprecedented influence to the field of machine translation, it does not apply to all linguistic problems.[25,26] On the contrary, through the improvement and popularization of many scholars,[27,28] the long short term memory (LSTM)[29] used in our model has been successfully in many language models.[30,31] In fact, in addition to LSTM, we also tested four other structures, including the Transformer, with the similar number of parameters. The LSTM offered the best performances (Table S3, ESI†), indicating that LSTM developed its ability to extract information from line notations of chemical structures.

In conclusion, by solely using SMILES as a structural input without any chemical descriptors, we applied a natural language processing approach to predict the melting point of small organic molecules, and the model demonstrated a surprising accuracy comparable to or even better than state of the art. The model is applicable for molecules with diverse properties, and the prediction accuracy is a function of the sample size, as well as the number of multiple SMILES forms used in the training set. Different from previous fragment-based or descriptor-based models, our model provides better prediction accuracy for molecules with many rotatable bonds and thus more flexible configurations. The model provides a potential tool to establish a quantitative structure–property relationship and facilitate rational molecular design for drug discovery and development.

This research is supported by Johnson & Johnson and the Beijing Advanced Innovation Center for Structural Biology.

Conflicts of interest

There are no conflicts to declare.

Notes and references

1 L. H. Jones and M. E. Bunnage, *Nat. Rev. Drug Discovery*, 2017, **16**, 285–296.
2 D. A. Fidock, *Nature*, 2016, **538**, 323–325.
3 G. Zimmermann and D. Neri, *Drug Discovery Today*, 2016, **21**, 1828–1834.
4 E. J. Martin, J. M. Blaney, M. A. Siani, D. C. Spellmeyer, A. K. Wong and W. H. Moos, *J. Med. Chem.*, 1995, **38**, 1431–1436.
5 C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 1997, **23**, 3–25.
6 C. A. Lipinski, *J. Pharmacol. Toxicol. Methods*, 2000, **44**, 235–249.
7 C. A. Lipinski, *Adv. Drug Delivery Rev.*, 2016, **101**, 34–41.
8 S. H. Yalkowsky and S. C. Valvani, *J. Pharm. Sci.*, 1980, **69**, 912–922.
9 J. L. McDonagh, T. van Mourik and J. B. Mitchell, *Mol. Inf.*, 2015, **34**, 715–724.
10 M. Grover, B. Singh, M. Bakshi and S. Singh, *Pharm. Sci. Technol. Today*, 2000, **3**, 28–35.
11 M. Grover, B. Singh, M. Bakshi and S. Singh, *Pharm. Sci. Technol. Today*, 2000, **3**, 50–57.
12 M. Karthikeyan, R. C. Glen and A. Bender, *J. Chem. Inf. Model.*, 2005, **45**, 581–590.
13 L. Zhao and S. H. Yalkowsky, *Indus. Eng. Chem. Res.*, 1999, **38**, 3581–3584.
14 L. D. Hughes, D. S. Palmer, F. Nigsch and J. B. Mitchell, *J. Chem. Inf. Model.*, 2008, **48**, 220–232.
15 U. P. Preiss, W. Beichel, A. M. Erle, Y. U. Paulechka and I. Krossing, *ChemPhysChem*, 2011, **12**, 2959–2972.
16 C. W. Coley, R. Barzilay, W. H. Green, T. S. Jaakkola and K. F. Jensen, *J. Chem. Inf. Model.*, 2017, **57**, 1757–1772.
17 I. V. Tetko, D. M. Lowe and A. J. Williams, *J. Cheminf.*, 2016, **8**, 2.
18 T. Young, D. Hazarika, S. Poria and E. Cambria, *IEEE Comput. Intelligence Magazine*, 2018, **13**, 55–75.
19 S. Zheng, X. Yan, Y. Yang and J. Xu, *J. Chem. Inf. Model.*, 2019, **59**, 914–923.
20 E. Jonas, presented in part at Advances in neural information processing systems, 2019.
21 D. Weininger, *J. Chem. Inf. Model.*, 1988, **28**, 31–36.
22 D. Weininger, A. Weininger and J. L. Weininger, *J. Chem. Inf. Model.*, 1989, **29**, 97–101.
23 C. A. Bergström, U. Norinder, K. Luthman and P. Artursson, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1177–1185.
24 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, *et al.*, presented in part at Neural Information Processing Systems, 2017.
25 Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le and R. Salakhutdinov, presented in part at Annual Meeting of the Association for Computational Linguistics, 2019.
26 T. Domhan, presented in part at Annual Meeting of the Association for Computational Linguistics, 2018.
27 A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke and J. Schmidhuber, IEEE transactions on pattern analysis machine intelligence, 2008, **31**, 855–868.
28 J. Schmidhuber, *Neural networks*, 2015, **61**, 85–117.
29 S. Hochreiter and J. Schmidhuber, *Neural Comput.*, 1997, **9**, 1735–1780.
30 I. Sutskever, O. Vinyals and Q. V. Le, presented in part at Advances in neural information processing systems, 2014.
31 A. Graves, A.-r. Mohamed and G. Hinton, presented in part at 2013 IEEE international conference on acoustics, speech and signal processing, 2013.

