Chemistry SL Investigation

Research Question:

Which SMILES components contribute most to the prediction of melting points?

Number of pages: 17

Done by: Matvey Ryabov

Candidate number: 000556-0010

School: Winston Churchill High School, Lethbridge Alberta

Contents

1 Research Question:

Which SMILES components contribute most to the prediction of melting points?

2 Introduction:

The inspiration for this investigation initially came about while I was researching some different molecular representation methods. I was interested in applying my knowledge of machine learning to the field of chemistry but wasn't sure how I would be able to feed a computer 3D molecular models. I was looking at the different ways we can input molecular data and get the computer to make then predictions based on the molecular structure. I eventually came across a method called Simplified Molecular Input Line Energy Specification (SMILES), which looked to be quite promising and was being used by several projects. This method represents a valence model of a molecule and allows us to describe the structure of a molecule in one dimension (via a string of characters) while remaining syntactically intuitive and comprehensive enough to be valuable (enough data is still preserved in this format).

Also, I was interested in getting a more intuitive understanding of the factors affecting the melting points of a substance, since this was an area I lost marks in on some of my exams.

Putting these two inspirations together, I wanted to see what part of

SMILES - each SMILES representation was cut into 2-4 character long slices - would contribute most to the prediction of melting points using a machine learning model; I was interested in determining what 2-4 character slices, when removed from the dataset would most affect the performance of a machine learning model aimed at predicting the melting point.

3 Background Research/Information:

We learned in class that as the size of a molecule increases, the melting point goes up. This due to the increase in Van der Waals' forces. We also learned that the presence of a polar group such as the carbonyl group increases the melting point because it creates a dipole-dipole interaction. Molecules that can hydrogen bond, we learned, generally have a higher melting point - and that hydrogen bonding is significantly stronger than Van der Waal's forces. For this reason, alcohols generally have a higher melting/boiling point because they have an OH which can hydrogen bond. Molecules that can pack together more tightly, we learned, have a greater melting point also. For example, symmetrical neopentane molecules have a higher melting point than isopentane, which can't pack together so well.

In my background research I discovered that impurities play a part in a molecule's melting point [3]. However, for this dataset, the chemicals are fairly pure, so I am not using this in my hypothesis.

The Bergstrom study looked at determining whether certain 3D chemical representations can viably/performantly predict the melting point of a solid drug (useful in solubility research). The Bergstrom study, the source I am using my dataset from, has a paper [1] associated with it that came to the conclusion that descriptors of hydrophilicity, polarity, partial atom charge, and molecular rigidity were found to be positively correlated with melting point while non-polar atoms and high flexibility were found to be negatively correlated. Although I am using the same dataset, my experiment focused on using the SMILES method of representation rather than 3D or 2D representations (calculated representation values) the paper uses. This means my machine learning model doesn't have access to the quantitative values associated with molecule rigidity and partial atom charge. However, by looking at the SMILES representations, my model still has access to rough measures of hydrophilicity (it should be able to identify hydrogen bonds), polarity, and molecule size, which actually yielded a correlation coefficient of 0.68 which is pretty good. In this experiment, I aim at finding what parts of SMILES were most useful to the model to make its predictions.

An important thing to note is that cutting SMILES representations into 2-4 character slices doesn't remove the model's ability to see what is in total present in the molecule structure. It just looks at one 2-4 character slice at a time - which could be anything like (=O), which would represent the presence of a carbonyl functional group. This means that using my 'leave one out' methodology, we are able to isolate what 'indicator' (2-4 character slice) - such as the presence of a carbonyl group - contributes most to the performance of the model.

4 Hypothesis:

Polarity, hydrophilicity, and molecule size contribute to melting point. SMILES descriptors of these qualities will be most useful to the model and will influence its performance most when taken out of the dataset.

5 Materials:

- Weka 3.8: Software used to rapidly prototype machine learning models and which was used to configure and run the model for this investigation.
- GNUMeric: Open source equivalent of Microsoft Excel. Was used to clean up the data and get it ready to be inputted into Weka.
- Bergstrom Melting Point Dataset: Obtained from ChemDB and available at: `ftp:ftp.ics.uci.edu/pub/baldig/learning/Bergstrom/`.
- Multilayered Perceptron (Neural Network): The machine learning model used for this investigation (see appendix for configuration). This model is, put simply, a universal function finder and thus is commonly used to find the line either separating groups of data (classification) or predicting datapoints (regression). Model learns ('finds' the correct function) using the negative gradient of a cost function (see appendix).
- A SMILES reference, available at `https://archive.epa.gov/med/med_archive_03/web/html/smiles.html`, was used.

5.1 The Dataset:

The dataset consisted of three columns: one contained the chemical name of the molecule, one contained the SMILES representation of the molecule, and one contained the melting point of the molecule. There were 184 molecules in the dataset.

Here are the first 20 rows of the dataset:

Chemical Name:	SMILES Representation:	M.P. (°C)
Abecarnil	COCc3c(ncc4[nH]c2ccc(OCc1ccccc1)cc2c34)C(=O)OC(C)C	150
Acebutolol	CCCC(=O)Nc1ccc(OCC(O)CNC(C)C)c(c1)C(C)=O	119
Acecarbromal	CCC(Br)(CC)C(=O)NC(=O)NC(C)=O	109
Aceclofenac	OC(=O)COC(=O)Cc1ccccc1Nc2c(Cl)cccc2Cl	149
Acedapsone	CC(=O)Nc1ccc(cc1)S(=O)(=O)c2ccc(NC(C)=O)cc2	289
Acediasulfone	Nc1ccc(cc1)S(=O)(=O)c2ccc(NCC(O)=O)cc2	194
Aceglutamide	CC(=O)NC(CCC(N)=O)C(O)=O	197
Acemetacin	COc3ccc2n(C(=O)c1ccc(Cl)cc1)c(C)c(CC(=O)OCC(O)=O)c2c3	150
Acetaminophen	CC(=O)Nc1ccc(O)cc1	169
Acetaminosalol	CC(=O)Nc2ccc(OC(=O)c1ccccc1O)cc2	187
Acetorphan	CC(=O)SCC(Cc1ccccc1)C(=O)NCC(=O)OCc2ccccc2	89
Acetylpheneturide	CCC(C(=O)NC(=O)NC(C)=O)c1ccccc1	100
AcetylsalicylicAcid	CC(=O)Oc1ccccc1C(O)=O	142.4
Acyclovir	Nc2nc1n(COCCO)cnc1c(=O)[nH]2	255
Acifran	CC1(OC(=CC1=O)C(O)=O)c2ccccc2	176
Acitretin	COc1cc(C)c(C=CC(C)=CC=CC(C)=CC(O)=O)c(C)c1C	228

Continued on next page

Continued from previous page

Chemical Name:	SMILES Representation:	M.P. (°C)
Adrafinil	ONC(=O)CS(=O)C(c1ccccc1)c2ccccc2	159
Ahistan	CN(C)CC(=O)N2c1ccccc1Sc3ccccc23	144
Albendazole	CCCSc2ccc1[nH]c(NC(=O)OC)nc1c2	208
Albutoin	CC(C)CC1NC(=S)N(CC=C)C1=O	210

6 Methodology and Experiment:

6.1 Preprocessing:

Before the data could be fed into the machine learning model (which is a glorified function finder, if you are not aware), I needed to cut the SMILES representations into 2-4 character slices. This was done using a tokenizer method, which used the TF-IDF[1] method to find the top 105 most 'unique' 2-4 character slices. The number of times each slice appeared in each molecule was tallied, and this resulting information was compiled into a dictionary file where each column corresponds to a 2-4 character slice, and each row corresponds to an individual molecule - the element at the cross of the two is the tallied presence of that slice in the molecule. The model was then run on this dictionary file and was predicting the melting point of each molecule based on the slices present in each molecule's structure. Here are is a sample

[1] TF-IDF: short for term frequency - inverse document. A numerical statistic that reflects how relevant a term is in a given document (in this case a molecule).

of selected 'indicators' (slices) and their interpretations:

SMILES Slice:	What it indicates:	TF-IDF Rank:
ccc	presence of ring with # of C >= 5	100
C(C)	presence of methyl branch	54
cccc	presence of ring with # of C >= 6	104
CCCC	presence of carbon chain with length of 4	64
O	presence of oxygen in molecule	75
(=O)	presence of carbonyl group in molecule	5
4	4 breakage points in molecule which relates to presence of >= 4 rings	41
C(O)	presence of hydroxyl group	56

Note that for "ccc" and "cccc" the minimum possible ring length is current length plus 2 since every ring needs an open and close atom. So the smallest possible rings are "c1cccc1" for "ccc" and "c1ccccc1" for "cccc" where 1 indicates the location of the breakage point - the place where the ring closes. This is how SMILES describes rings - using numbers to show where the ring closes (the two matching by number atoms would be the ones joined).

6.2 Experimental Procedure:

- Data was cleaned and arranged into three columns (headings were removed) in GNUMeric.
- The tokenizer was applied to the SMILES column, slicing the SMILES

representation into 2-4 character slices, yielding the dictionary file.

- The model was run on different, non-overlapping, 90-10% training-testing splits (10 fold cross validation) of the dictionary 10 times to make sure the results were unbiased (perhaps one 10% was easier to predict melting points for) and valuable (to make sure the SMILES slices contained enough data - which they did because the correlation coefficient was 0.68[2]).

- The model was run with the same split pattern as above, but one column in the dictionary - relating to a different 2-4 character slice (descriptor) - was left out. The effect of leaving out this slice on the correlation coefficient was recorded, and this was done for each of the 105 different slices.

- The slices were ranked by their effect on the correlation coefficient.

- The top-ranked slices are then interpreted to come to conclusions.

6.3 A Note About Interpretation:

Some SMILES slices are difficult to interpret in isolation from other slices. A computer has access to more than one slice and is capable to then find more complicated relationships than we can with just one. For example, "4" tells us that there are $>=$ 4 rings but not what branches might be attached or what atoms compose the molecule - for example, the presence of a hydroxyl functional group would also have an effect on melting point.

[2] This means there is a function which can predict melting point. A value closer to 1 means the data has a correlation with the found function.

However, even though we don't have the full picture, we can see that "4" relates to a complicated molecular structure (4 rings) and we can make an interpretation that the computer might use this to say that this molecule would have a higher melting point - more Van der Waal's forces means a higher melting point.

7 Data and Analysis:

7.1 Data Table:

Here are the top 20 slices ranked by their effect on model performance when taken out of the dictionary:

Effect on Correlation:	TF-IDF Rank:	Slice:
0.04414	31	1ccc
0.03014	39	3c
0.02938	11	(O)
0.02914	53	C(C
0.029	21	)NC
0.02777	48	=O)c
0.02467	40	3cc
0.02441	41	4
0.02313	17	)C
0.023	24	)c2
0.02215	49	C
0.02087	37	2ccc

Continued on next page

Continued from previous page

Effect on Correlation:	TF-IDF Rank:	Slice:
0.02076	56	C(O)
0.02065	2	(
0.01978	46	=O)C
0.01793	10	(O
0.01763	52	C(=O
0.01674	62	CC(
0.01575	19	)CC
0.01411	1	150

7.2 Analysis of Results:

If you look at the results, the slice "1ccc" has the most effect on the correlation coefficient. This slice indicates that there is a carbon ring with a length >= 5. The '1' in the slice indicates that there are at least 1 ring present in this structure. Both of these 'indications' can be used by the computer to come to conclusions about the melting point of a molecule. For example, rings in general rings have a higher melting point because the molecules in the ring can get closer together and have more rigidity than in straight-chain molecules [2]. Larger rings would tend to have a larger melting point because they have higher LD forces. Having many rings can cause molecular asymmetry [4] thus leading to increased polarity - which increases the melting point. All this is what the model likely understood from the "4" slice. The slice "3c" also had a large effect. The c in "3c" doesn't tell us too much -

it doesn't tell you much about the real length of the ring just that it's >= 3. However, the 3 in "3c" would indicate that there are >= 3 rings, which would be related with a higher melting point.

The next 4 slices are all indicating that there is a branch on the molecule. "(O)" indicates that there is a hydroxyl functional group, "C(C" indicates that there is a >= methyl branch, ")NC" indicates that there is a nitrogen in a non -ring chain plus that there is a branch ")", and "=O)c" indicates both that there is a ring and that there is a carbonyl branch (not necessarily an OH also). "(O)" indicates that there is a hydroxyl group present in the molecule which would be related to a higher boiling point because of hydrogen bonding. More branching can relate to a lower melting point as more branching can result in more spherical shape, which would relate to a lower melting point [5] - thus why "=O)c" might be more useful; aside from its potential to create dipole forces. Different types of atoms present in the chain have a chance at increasing polarity, contributing to a higher melting point - thus ")NC" could prove useful.

The model, since it has access to the frequencies of these slices, is able to weigh their presence and thus predict the melting point of the molecule - to an extent of course.

I believe that 150 is an outlier/error since it doesn't appear in the dataset at all, but does in the labels. I think that somehow the model got access to the label (a melting point of 150 for two molecules) and was using the label to then better its predictions.

8 Evaluation and Limitations:

This experiment isn't perfect because the SMILES representation is still a 'summary' - it isn't as comprehensive as a 3D model. This means that the conclusions arrived by the end of this investigation could be improved if I used a methodology such as the one used in the Bergstrom study - if I used a 3D representation. However, my investigation had a nuance - to find out which 'descriptor' affected the performance most, I needed to run the model and exclude one 'descriptor' at a time. This means that I needed to run the model more than one-hundred times more than for a regular test of whether using SMILES as a representation is viable/possible - Bergstrom study. For this reason, I was in a sense limited in computing resources - I only have a laptop, and it isn't fast enough to run these computationally intensive calculations. Working with a more 'complete' representation of molecular structure - 3D and 2D representations of structure - would be impossible on my current computer (I already had to leave the experiment running for more than 24 hours). Possibly if I had a better computer, I would be able to run a more complex model that would take 3D or 2D structures as input. This would allow the model to take advantage of other different traits that might be important for predicting melting point - such as molecule flex. One potential improvement that could have been done to my methodology, would be if I used a character level convolutional neural network (CNN) model rather than a n-gram[3] neural network model - the former actually has a better sense of spatial relationships [6] (A CNN is commonly used in

[3] An n-gram is simply a sequence of words or characters. It's the more data-scientist term for my 2-4 character slices.

image classification tasks) - the model I used is more frequency orientated (frequency of 'slice').

Overall, this investigation has been able to find out what sort of 'indicators' (slices) matter for predicting the melting point (see the results table) and provide a potential explanation for their significance (see Analysis section). The slices that had the greatest effect and their potential explanations line up quite well with the results of the Bergstrom study - plus, there is the extra explanation about rings. From the results of this investigation we can infer that our model looks at and values most the traits responsible for polarity (types of branches), hydrophilicity (hydrogen bonding), and ring size/number of rings. Size didn't matter as much in this investigation, perhaps since a lot of the molecules were already quite large (drugs tend to be complex molecules). This means that my hypothesis was partially incorrect. I thought size would matter more than it seemingly did. For the other 'traits' - hydrophilicity, polarity - I was correct that this model would value them.

9 Conclusion:

All in all, this investigation was successful. I was able to get a correlation coefficient of 0.68 and was able to determine which 'indicators' have the greatest effect on predicting the melting point. This investigation can have potential applications in the medical field where predicting the melting point of a new drug might prove useful since being able to also understand what factors contribute to melting point, would allow us to potentially develop better drugs with a controlled solubility [1]. This important because having a controlled solubility greatly affects the effective deployment of a drug.

References

[1] Christel AS Bergström, Ulf Norinder, Kristina Luthman, and Per Artursson. Molecular descriptors influencing melting point and their role in classification of solid drugs. *Journal of chemical information and computer sciences*, 43(4):1177–1185, 2003.

[2] Jim Clark. An introduction to alkanes and cycloalkanes. *Chemguide UK*, Jul 2015.

[3] Claire Gillespie. What factors affect melting point? *SCIENCING*, Apr 2018.
`https://sciencing.com/factors-affect-melting-point-8690403.html`
.

[4] Jacob W Martin, Kimberly Bowal, Angiras Menon, Radomir I Slavchov, Jethro Akroyd, Sebastian Mosbach, and Markus Kraft. Polar curved polycyclic aromatic hydrocarbons in soot formation. *Proceedings of the Combustion Institute*, 37(1):1117–1123, 2019.

[5] William Reusch. Boiling melting points - msu chemistry. May 2013. `https://www2.chemistry.msu.edu/faculty/reusch/VirtTxtJml/physprop.htm`.

[6] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

10 Appendix:

10.1 Model Configuration:

All I changed was the layer layout and the learning rate.

Layer Layout: I used 105, 10, and 10 node in the three layers.

Learning Rate: 0.3 and turned on decay.

10.2 Machine Learning:

Machine learning is a field of computer science looking at creating models capable of 'learning' the solution to a problem. In this investigation the model was learning to predict the melting point of a molecule based on the SMILES representation. The model chosen, if explained at a very simple level, is a collection of matrix operations representing layers of 'neurons'. This type of model feeds data through these operations and then adjusts for error to get the right output. A model does this by calculating the negative gradient of a loss function, such as a mean squared error - although this loss function is not advised. The adjustable parts which the gradient is computed for, are the matrices within these operations.