Methods to improve the performance of topological fingerprints in machine learning and molecular similarity calculations

Ruifeng Liu[1,2]*, Zhen Xu[1,2], Carol Rios Rocha[1,2], Mohamed Diwan M. AbdulHameed[1,2], Himanshu Goel[1,2], Valmik Desai[1,2], and Anders Wallqvist[1]*

[1]*Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Development Command, Fort Detrick, MD, USA*

[2]*The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., Bethesda, MD, USA*

*Corresponding authors
Ruifeng Liu, Ph.D.
E-mail: rliu@bhsai.org

Anders Wallqvist, Ph.D.
Deputy Director, DoD Biotechnology HPC Software Applications Institute
Telemedicine and Advanced Technology Research Center
U.S. Army Medical Research and Development Command
ATTN: FCMR-TT, 504 Scott Street, Fort Detrick, MD 21702-5012
Phone: +1 301 619 1989
Fax: +1 301 619 1983
E-mail: sven.a.wallqvist.civ@health.mil

Author's ORCID
RL: 0000-0001-7582-9217
AW: 0000-0002-9775-7469

Abstract

Topological bond path and circular fingerprints are the most common molecular representations used in similarity searches, and in recent years, they have been increasingly used as inputs for chemistry-focused machine learning. To encode all possible atom compositions and chemical bond configurations, fingerprints must be very long, making them resource demanding. To circumvent the length issue, fingerprints are customarily folded to a reduced and more manageable size before use. However, there is no systematic assessment of the information loss associated with folding a fingerprint, especially when folded fingerprints are used as inputs for machine learning. Methods to minimize information loss due to folding exist in proprietary software, but details of the methods and their implementation are not available. As a result, similar methods are lacking in any of the increasingly used open-source software packages. Here, we describe our implementation of such methods and our assessment of their performance on RDKit Morgan fingerprints. Our results demonstrate that by re-indexing the fingerprint bit features, we can significantly reduce the information loss due to folding and can develop robust machine-learning models with a relatively short fingerprint length. We also demonstrate that Tanimoto-similarity calculations using unfolded fingerprints, which avoid information loss completely, are not only feasible, but significantly faster than Tanimoto-similarity calculations using folded fingerprints. We are making our codes to implement these methods freely available, and we encourage the community to use these methods in a broad range of applications.

Keywords: Molecular fingerprint, Bit clashing, Tanimoto similarity, Machine learning

Introduction

Molecular fingerprints are an essential component of cheminformatics [1]. They enable molecular similarity searches, which are key for exploiting the similar structure – similar activity principle in modern drug discovery [2]. Molecular fingerprints are also increasingly used as input features in machine learning [3], eliminating the need for expert handcrafted molecular descriptors specifically designed as inputs for machine learning [4]. Broadly speaking, there are two types of molecular fingerprints. The first type are substructure key-based fingerprints, examples of which include Molecular ACCess Systems keys (MACCS) [5] and PubChem fingerprints [6]. The substructure keys are pre-defined molecular substructures considered important for a broad range of molecular chemical and biological activities. Because the available chemistry space is so large and the number of chemicals tested in all chemical and biological assays only covers a tiny fraction of all possible chemicals, a caveat of substructure key-based fingerprints is that many important substructures may not have been recognized and included in the list of substructure keys. The second type of molecular fingerprints are topological path- and circular atom environment-based fingerprints, examples of which include Daylight fingerprints[7] and Morgan fingerprints [8]. These fingerprints encode all possible chemical bond connectivity paths or atom environments within a user-selected limit of bond-path or atom-environment size (diameter or radius) in the form of a bit-string. Because of the large number of possible bond paths and atom environments, a caveat of this type of fingerprint is that the length of a fingerprint bit-string must be very long to encode all possibilities, making it computationally resource demanding in practical applications. Because an individual molecule always has a limited number of unique structural features, the bit-string of a molecular fingerprint is sparsely populated by 1's (called on-bits), representing unique structural features

present in the molecule, and mostly populated by 0's (called off-bits), representing the absence of structural features encoded by the bit positions. A common approach to reduce the computing resource requirements of this type of fingerprint is to fold the bit-string to a manageable length using the logical OR operation. For example, Daylight fingerprints are typically folded to a fixed length of 1,024 bits, with the folded fingerprints performing satisfactorily in most molecular similarity searches [9]. Following the pioneering work of Daylight Chemical Information Systems, all other fingerprints of the same type adopted the fingerprint-folding approach in their applications with the fingerprint bit-length commonly set to 1,024 [10].

It is well-known that folding a fingerprint leads to information loss due to bit clashing, i.e., two bit positions that represent different structural features are merged into one, leading to structural ambiguity represented by the folded bit position. While fingerprints folded to 1,024 bits show acceptable performance in similarity searches, the impact of fingerprint folding on other applications, e.g., as inputs for machine learning, has not been systematically investigated. For example, when assessing the performance of fingerprints vs. conventional expert handcrafted molecular descriptors for machine learning, some studies only employed fingerprints folded to a fixed length, without considering the impact of folding or exploring ways to minimize information loss due to folding. Conversely, one of the software packages currently available, Pipeline Pilot, implemented a method to minimize folding-led information loss [11]. However, we have not seen studies assessing the effectiveness of the method in improving performance of folded fingerprints. Additionally, Pipeline Pilot implemented a method to calculate Tanimoto similarity using unfolded fingerprints, which does not incur information loss due to folding. Because there is no need to fold fingerprints before a Tanimoto-similarity calculation, this method is also potentially faster than using folded fingerprints.

Here, we report the results of our investigation into the effectiveness of methods to minimize and avoid information loss due to fingerprint folding. We implemented the methods in Python and Java and evaluated their performance using the Morgan fingerprints generated by RDKit [12]. We chose the Morgan 2D fingerprint because it is one of the most popular fingerprints [13]. However, the same methods should work equally well with any folded fingerprints.

Methods

Re-indexing fingerprint bit features to reduce information loss due to folding

In the bit-string of an unfolded fingerprint [Fig. 1(a)], the on-bits appear randomly distributed. Folding the fingerprint merges bits symmetrical with respect to the mid-point of the string into one [Fig. 1(b)], i.e., the first and the last bits are merged into the first bit of the folded fingerprint. Structural ambiguity arises because an on-bit of a folded fingerprint represents either one of two unique structural features present or both structural features present in a molecule. This is referred to as bit clashing. As we successively fold a fingerprint, bit clashing becomes increasingly more severe and is accompanied by increasing information loss.

To minimize folding-induced information loss, one can re-index the bit positions of a fingerprint based on the frequencies of the structural features they encode in a large number of molecules. For example, when we re-index fingerprint bits by ordering from the most frequent to the least frequent structural features [Fig. 1(c), the "Pack Most Significant First" option in Pipeline Pilot], most of the on-bits will appear at the left and most of the off-bits at the right of a molecular fingerprint. Because the bits at the far right encode rarely occurring structural features, folding such a fingerprint leads to less information loss [Fig. 1(d)]. To implement this method, we first generated unfolded binary Morgan_2 fingerprints (equivalent to the ECFP_4 fingerprint of

Pipeline Pilot) for ~800,000 randomly selected compounds from PubChem using RDKit. We counted the number of times each bit was populated by 1's in the ~800,000 fingerprints. We then re-indexed the fingerprint bits based on the counts from high to low. We implemented this process in Python and made the code freely available via download from GitHub (*address to be added here*).

Calculating Tanimoto similarity using unfolded fingerprints

Following the practice of Daylight Chemical Information Systems, the customary procedure to calculate Tanimoto similarity is first folding the fingerprints into a fixed length. However, this is not necessary. By definition, Tanimoto similarity is the ratio between the count of common on-bits between two molecules and the count of unique on-bits of both molecules. Even though the bit-string of a fingerprint can be extremely long, the number of on-bits in a fingerprint is comparatively small, i.e., up to a couple hundred for most molecules. Thus, to reduce resource usage in Tanimoto-similarity calculations, one can ignore the off-bits and store the indices of on-bits in one-dimensional arrays only. From the resulting arrays, one can easily count the common on-bits and total unique on-bits between two molecules to calculate the Tanimoto similarity between the two molecules. We implemented this approach in Java for Konstanz Information Miner (KNIME) [14], and the code is freely available from GitHub (*address to be added here*).

Molecular datasets and evaluation methods

To evaluate the effectiveness of the method for minimizing information loss due to folding, we used four relatively large datasets of molecular physicochemical properties and bioactivities. These datasets consisted of *1*) a dataset of the logarithm of partition coefficients between n-

octanol and water (logP dataset) [15], *2*) an aqueous solubility dataset given in logarithmic form (logS dataset) [16], *3*) a publicly available median lethal dose rat oral toxicity dataset (logLD50) [17], and *4*) a 50% growth inhibition dataset of the human ovarian cancer cell line OVCAR-8 (logGI50) downloaded from the NCI-60 Human Tumor Cell Lines Screen project [18]. After we removed the inorganic and organometallic compounds for which RDKit was not able to generate Morgan_2 fingerprints, 10,154, 8,486, 6,360, and 12,403 compounds remained in the four datasets, respectively.

For each dataset, we first generated Morgan_2 count fingerprints (count of the times each bit feature is present in a molecule) folded to fixed lengths of 2,048, 1,024, 512, and 256 bits for all the molecules. We then generated Morgan_2 count fingerprints folded to the same fixed lengths with re-indexed bit features as described in **Methods**. We used the RDKit and the modified folded fingerprints as inputs to train a two-hidden layer deep neural network (DNN) model. The architecture of the DNN model was ***n***:1,000:500:1, where ***n*** represents the number of input features (i.e., the folded fingerprint length 2,048, 1,024, 512, or 256); 1,000 and 500 denote the number of hidden neurons in the first and second hidden layers, respectively; and 1 represents the numerical output of the DNN model. We applied drop-out regularization with a constant drop-out rate of 20% on all layers. We randomly split each dataset into a 60% training set and a 40% validation set, and used the training sets to train the model parameters by minimizing the mean squared error (MSE) loss function. At each epoch of training, we calculated the MSEs of both the training- and the validation-set compounds. The difference between the validation-set MSEs of the models trained with the two types of folded fingerprints gives an indication of the effectiveness of the method in reducing information loss.

To evaluate the impact of fingerprint folding on Tanimoto similarity, we retrieved 200 compounds with the highest binding affinities to the μ-opioid receptor from the BindingDB database [19] and randomly separated the compounds into two sets of 100 each. We then calculated Tanimoto similarities between compounds in the two sets using RDKit Morgan_2 fingerprints folded to fixed lengths of 4,096, 2,048, 1,024, 512, 256, 128, and 64. We also calculated Tanimoto similarities between the same two sets of compounds using unfolded RDKit Morgan_2 fingerprints as described above.

Results

Impact of fingerprint folding on machine learning

Figure 2 shows the MSE plots of the logP models as model training progressed. We trained the models using the RDKit Morgan_2 count fingerprints and our modified Morgan_2 count fingerprints as input features. As expected, the results showed that with increasing epochs, the training MSEs with both fingerprints as inputs decreased monotonically. At the same time, the validation MSEs with both fingerprints quickly reached their minima, then plateaued or ticked up slightly with increasing training epochs. Surprisingly, the MSEs of the modified fingerprints remained lower than those of the corresponding RDKit fingerprints across all fingerprint lengths. In addition, the difference in MSE was the largest with the shortest fingerprint (256) and smaller with increasing fingerprint length, suggesting that the difference was a result of information loss due to fingerprint folding. The difference in performance between the RDKit and our modified fingerprints was more obvious when the validation MSE traces were plotted side by side on the same scale. Figure 3 shows that not only were the MSEs calculated from the modified fingerprints lower than those of the corresponding RDKit fingerprints, but the point-to-point

MSE variations of the modified fingerprints were also significantly smaller. In addition, the difference between the MSE plots of the RDKit fingerprints folded to different lengths was significantly larger than the difference between the MSE plots of the modified fingerprints. With the RDKit fingerprints as inputs, the shorter the fingerprints were, the higher the validation MSEs and, therefore, the worse the DNN model. With the modified fingerprints as input features, the MSEs of different fingerprint lengths were significantly closer to each other, even though those of the shorter fingerprints still appeared to be higher. Thus, the results indicated that re-indexing the fingerprints reduced information loss to a significant degree, but it did not eliminate all information loss.

Figure 4 shows the results obtained using the other three datasets, which are very similar to those obtained with the logP dataset. Together, the results show that the performance enhancement of the DNN models trained with the modified fingerprints was not restricted to a particular dataset, but was observed across the different datasets. Thus, when using folded fingerprints as inputs for machine learning, re-indexing the fingerprints before folding can significantly reduce information loss and result in more robust models.

Impact of fingerprint folding on Tanimoto similarity

To examine the impact of fingerprint folding on Tanimoto similarity, we calculated Tanimoto similarities between molecules in the two sets of 100 SMILES using RDKit Morgan_2 fingerprints folded to different lengths. We also calculated Tanimoto similarities using RDKit Morgan_2 fingerprints without folding, as described in **Methods**. We performed all the calculations using KNIME on a PC equipped with an Intel 4-core i7-7000 CPU and 16 GB of RAM. Table 1 summarizes the time to completion for the 10,000 Tanimoto-similarity

calculations with and without fingerprint folding to different lengths and shows that calculations using the unfolded fingerprints were the fastest, requiring 4.5 seconds to complete 10,000 calculations. The same similarity calculations using the folded fingerprints took longer to complete, and the shorter the length to which the fingerprints were folded, the longer it took to complete the calculations. For instance, it took 47% longer (6.6 vs. 4.5 seconds) to complete the computations using RDKit fingerprints folded to 1,024 bits–the commonly used fingerprint length. Thus, folding the fingerprints consumed a significant amount of time relative to Tanimoto-similarity calculations.

To assess the impact of fingerprint folding on the calculated Tanimoto-similarity values, we calculated the differences between Tanimoto similarities from folded and unfolded fingerprints. We then grouped the differences in bins and plotted the results in Fig. 5. The similarity values calculated from fingerprints folded to 4,096 bits were the closest to those calculated from the unfolded fingerprints, with 9,973 (99.7%) of the 10,000 similarity values differing by no more than 0.05. The values calculated from fingerprints folded to 2,048 and 1,024 were also relatively close to those calculated from unfolded fingerprints, with 93.4% and 87.4% within 0.05 of the values calculated from the unfolded fingerprints, respectively. However, when the fingerprints were folded to lengths shorter than 1,024, a significant number of similarity values calculated from the folded fingerprints were higher than those calculated from the unfolded fingerprints, and the shorter length to which the fingerprints were folded, the greater number of values that were significantly higher than those calculated from the unfolded fingerprints.

Discussion

A molecular fingerprint is one of the most concise representations of a molecule that encode the chemical identity of a compound in terms of its atomic composition and chemical bond configuration. It reduces the dimensionality of a molecular structure into a one-dimensional array suitable for machine processing and is the basis of molecular similarity searches. In recent years, fingerprints have been increasingly used as input features for machine learning. To avoid ambiguity in its molecular representation, a fingerprint must be very long to encode all possible atom compositions and chemical bonding configurations. However, using long fingerprints was not practical in the early days of cheminformatics due to the increased demand on computing resources. Daylight Chemical Information Systems demonstrated that fingerprints folded to a fixed length of 1,024 could be used to calculate Tanimoto similarity with reasonably good performance. Following this pioneering work, virtually all cheminformatics software packages adopted the approach of folding fingerprints to a fixed length for similarity calculations. Results of this study show that Tanimoto similarities calculated from fingerprints folded to a fixed length of 1,024 or longer are reasonably close to those calculated using the unfolded fingerprints. That is, information loss due to bit clashing does not impact Tanimoto similarity significantly if the fingerprints are folded to 1,024 or longer. However, if fingerprints are folded to a length shorter than 1,024, information loss has a much greater impact and generally results in artificially higher similarity values. Considering that the method we adopted from Pipeline Pilot enables Tanimoto-similarity calculations to be performed faster than calculations using folded fingerprints, and without the information loss seen when using folded fingerprints, we encourage the use of this method for similarity calculations.

Even though we determined that bit clashing has a relatively small impact on Tanimoto similarity as long as the fingerprints are folded to 1,024 or longer, its impact on using folded

fingerprints as inputs for machine learning is much more significant, indicating that the robustness of a DNN model is very sensitive to bit identity, which is compromised by folding. Our results showed that even when the fingerprints were folded to as long as 2,048 bits, the validation MSEs of the DNN models based on the folded RDKit fingerprints as inputs were still significantly higher than those of the models based on our modified folded fingerprints. We demonstrated that modifying the fingerprints before folding was effective in minimizing information loss due to bit clashing, as the validation MSEs of the models based on the modified fingerprints folded to different lengths were very close to each other compared to those of models based on folded RDKit fingerprints. Thus, when using folded fingerprints as input features for machine learning, the fingerprints should be re-indexed before folding to minimize information loss and derive more robust machine-learning models.

Abbreviations

MACCS: Molecular ACCess Systems; KNIME: Konstanz Information Miner; DNN: deep neural network; MSE: mean squared error; FP: fingerprint; SMILES: Simplified Molecular Input Line Entry System.

Acknowledgements

The authors gratefully acknowledge the assistance of Ms. Maria Kuhrmann in editing the manuscript.

Author Contributions

RL and AW designed the study. Programming and computations were performed by ZX and CRR. Data analysis was performed by RL. The first draft of the manuscript was written by RL.

All authors commented on previous versions of the manuscript and approved the final manuscript.

Funding

This research was funded by the U.S. Army Medical Research and Development Command under Contract No. W81XWH20C0031 and by Defense Threat Reduction Agency Grant CBCall14-CBS-05-2-0007.

Availability of data and materials

The datasets (.csv files) used in this study and the Python and Java codes developed in this study are freely available at https://github.com/.... [to be added before submission]

Declarations

Competing interests

The authors declare that they have no competing interests.

Consent for publication

All authors have given consent for publication of the article. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the United States (U.S.) Army, the U.S. Department of Defense, or The Henry M. Jackson Foundation (HJF) for the Advancement of Military Medicine, Inc. This paper has been approved for public release with unlimited distribution.

References

1. Bender A, Brown N: **Special Issue: Cheminformatics in Drug Discovery.** *ChemMedChem* 2018, **13:**467-469.
2. Muegge I, Mukherjee P: **An overview of molecular fingerprint similarity search in virtual screening.** *Expert Opin Drug Discov* 2016, **11:**137-148.
3. Liu R, Zhou D: **Using molecular fingerprint as descriptors in the QSPR study of lipophilicity.** *J Chem Inf Model* 2008, **48:**542-549.
4. Carracedo-Reboredo P, Linares-Blanco J, Rodriguez-Fernandez N, Cedron F, Novoa FJ, Carballal A, Maojo V, Pazos A, Fernandez-Lozano C: **A review on machine learning approaches and trends in drug discovery.** *Comput Struct Biotechnol J* 2021, **19:**4538-4558.
5. Durant JL, Leland BA, Henry DR, Nourse JG: **Reoptimization of MDL keys for use in drug discovery.** *J Chem Inf Comput Sci* 2002, **42:**1273-1280.
6. **PubChem Substructure Fingerprint. (2/20/2021). Available at https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf. Accessed September 13, 2022.**
7. **https://www.daylight.com/dayhtml/doc/theory/theory.finger.html. Accessed September 13, 2022.**
8. Morgan HL: **The Generation of a Unique Machine Description for Chemical Structures - A Technique Developed at Chemical Abstracts Service. .** *J Chem Doc* 1965**:**107-112.
9. Cereto-Massague A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallve S, Pujadas G: **Molecular fingerprint similarity search in virtual screening.** *Methods* 2015, **71:**58-63.

10. Boyle NM, Sayle RA: **Comparing structural fingerprints using a literature-based similarity benchmark.** *J Cheminform* 2016, **8**.

11. **Pipeline Pilot, BIOVIA, Dassault Systèmes. https://www.3ds.com/products-services/biovia/products/data-science/pipeline-pilot/.**

12. **RDKit: Open-source cheminformatics. https://www.rdkit.org.**

13. Hu Y, Lounkine E, Bajorath J: **Improving the search performance of extended connectivity fingerprints through activity-oriented feature filtering and application of a bit-density-dependent similarity function.** *ChemMedChem* 2009, **4:**540-548.

14. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wisedel B: **KNIME: The Konstanz Information Miner.** In *Data Analysis, Machine Learning and Applications Studies in Classification, Data Analysis, and Knowledge Organization.* Edited by Preisach C, Burkhardt, H., Schmidt-Thieme, L., Decker, R. . Berlin, Heidelberg.: Springer; 2008: 319–326

15. **The Pomona LogPstar dataset as an example dataset in Pipeline Pilot.**

16. Sorkun MC, Khetan A, Er S: **AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds.** *Sci Data* 2019, **6:**143.

17. **https://ntp.niehs.nih.gov/whatwestudy/niceatm/test-method-evaluations/acute-systemic-tox/models/index.html.**

18. **https://dtp.cancer.gov/discovery_development/nci-60/.**

19. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J: **BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology.** *Nucleic Acids Res* 2016, **44:**D1045-1053.

Table 1 Time to completion of 10,000 Tanimoto-similarity calculations using unfolded fingerprints and fingerprints folded to different lengths

FP length	**Unfolded**	**4,096**	**2,048**	**1,024**	**512**	**256**	**128**	**64**
Time (ms)	4,510	5,141	5,310	6,630	7,006	7,209	7,423	8,077

The computations were performed using KNIME on a PC equipped with an Intel 4-core i7-7000 CPU and 16 GB of RAM. FP, fingerprint.

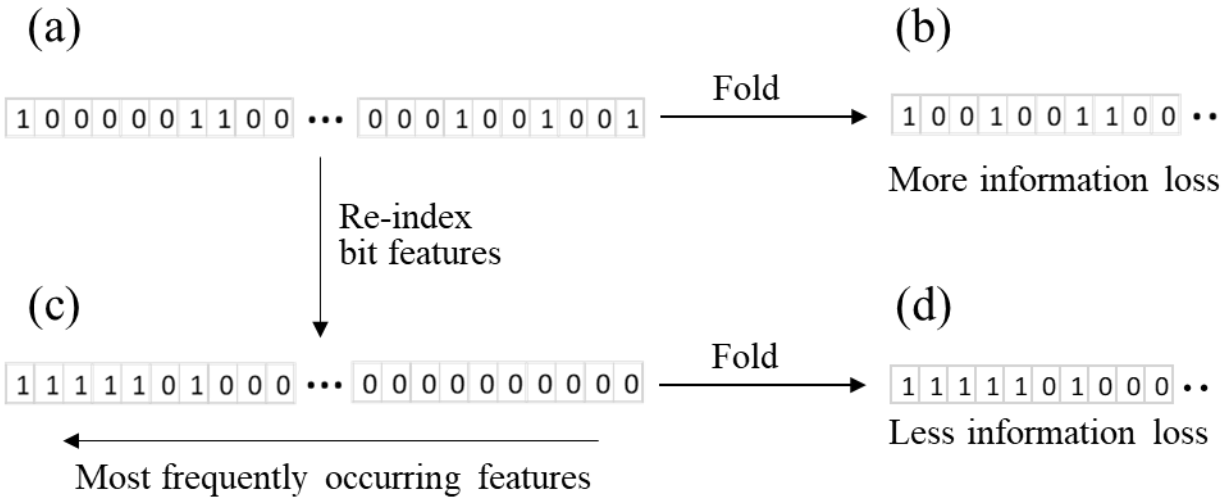

Fig. 1 Chart illustrating fingerprint folding with and without re-indexing bit features.

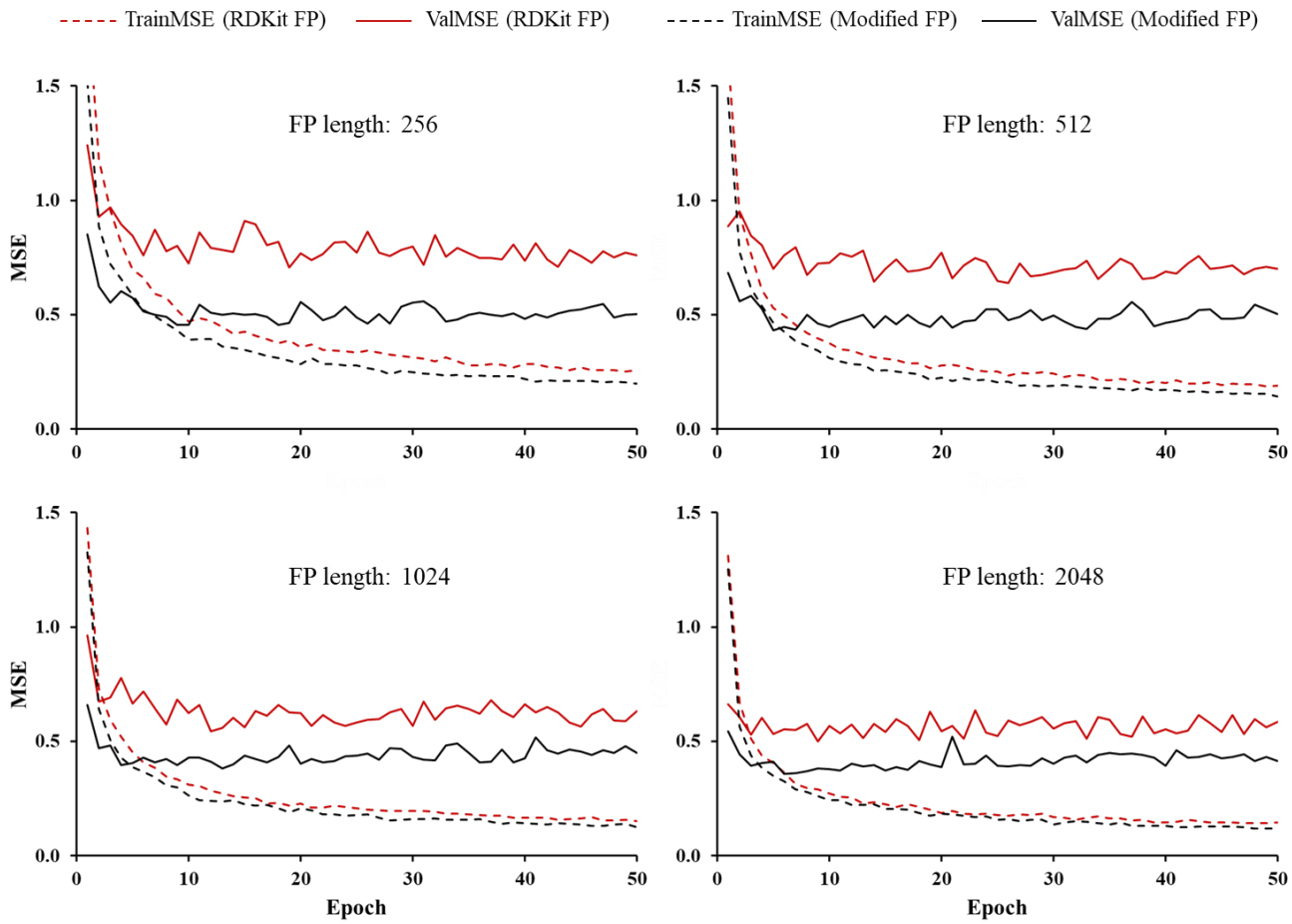

Fig. 2 Mean squared error (MSE) plots of the training (Train, dashed curves) and validation (Val, solid curves) sets of logP models using folded RDKit Morgan_2 count fingerprints as inputs features (black) and using our modified folded RDKit Morgan_2 count fingerprints as input features (red). FP, fingerprint.

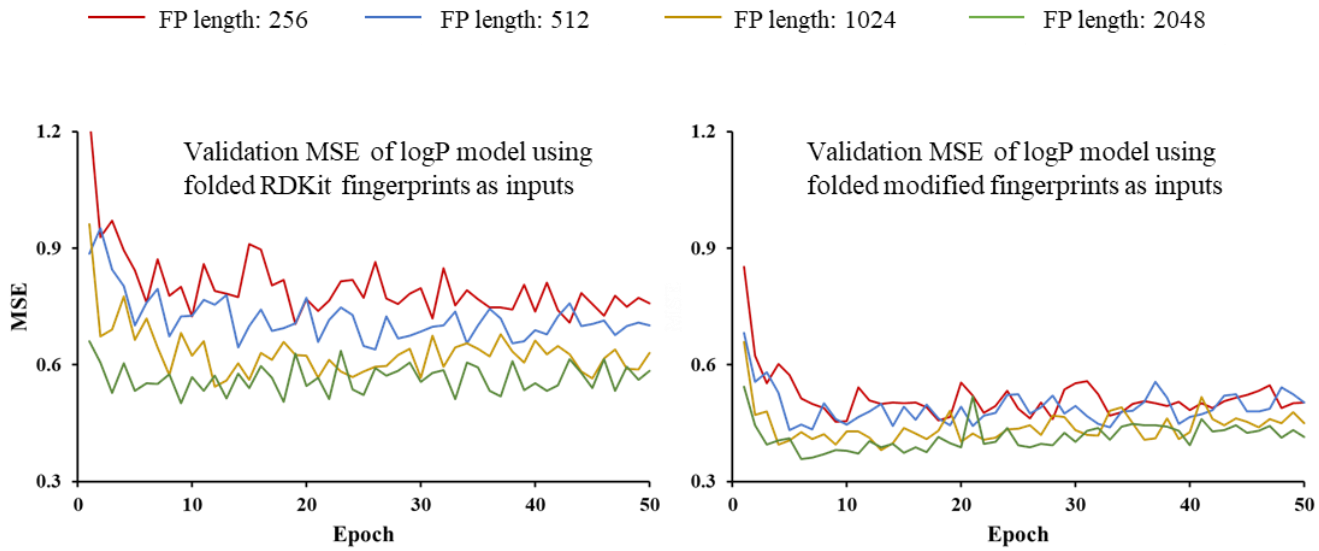

Fig. 3 Validation mean squared error (MSE) plots of the logP models using folded RDKit Morgan_2 count fingerprints as inputs (left) and the models using modified folded RDKit Morgan_2 count fingerprints as inputs (right). FP, fingerprint.

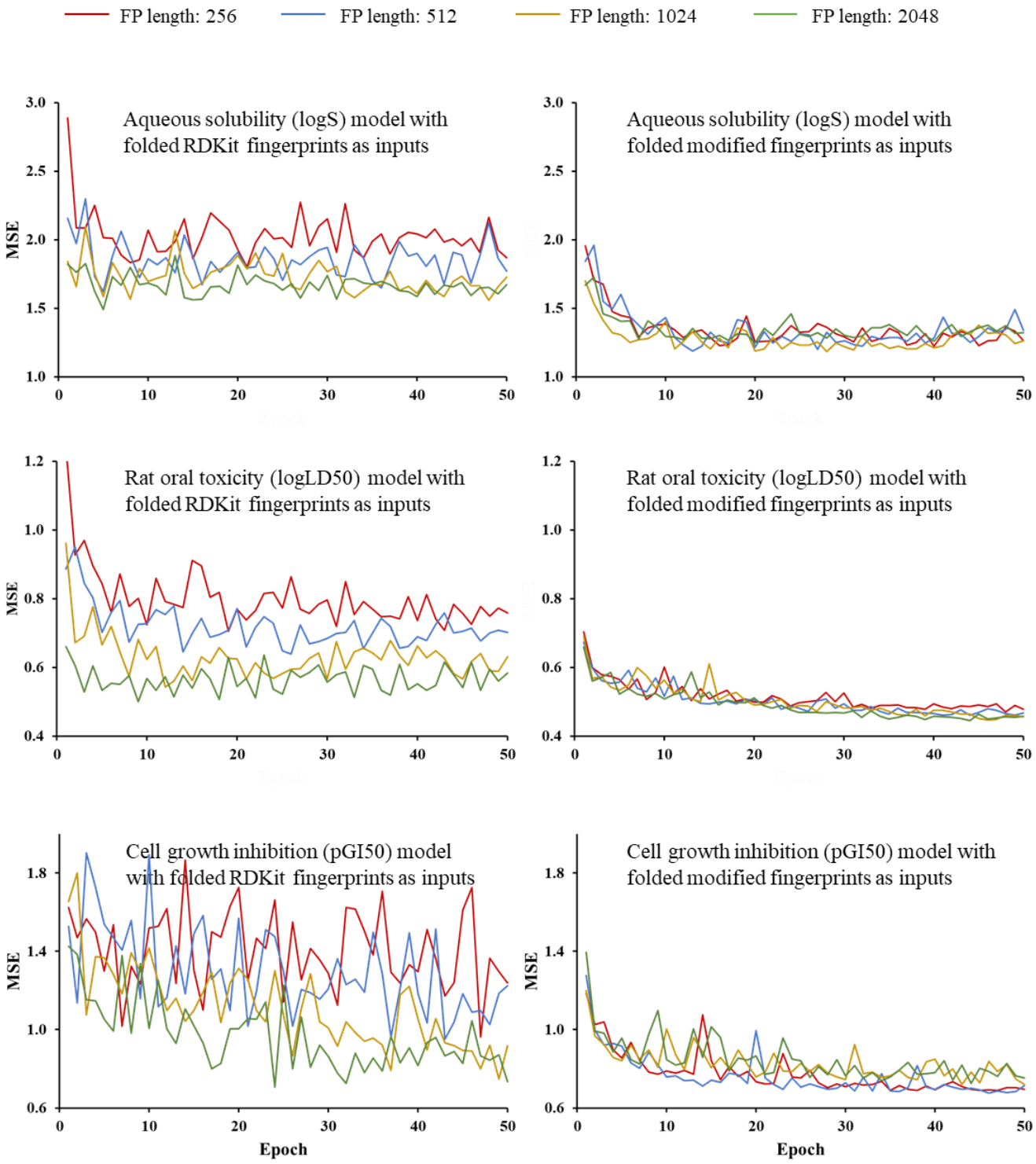

Fig. 4 Validation mean squared error (MSE) plots of the models using RDKit Morgan_2 count fingerprints as inputs (left) and the models using our modified RDKit Morgan_2 count fingerprints as inputs (right). FP, fingerprint.

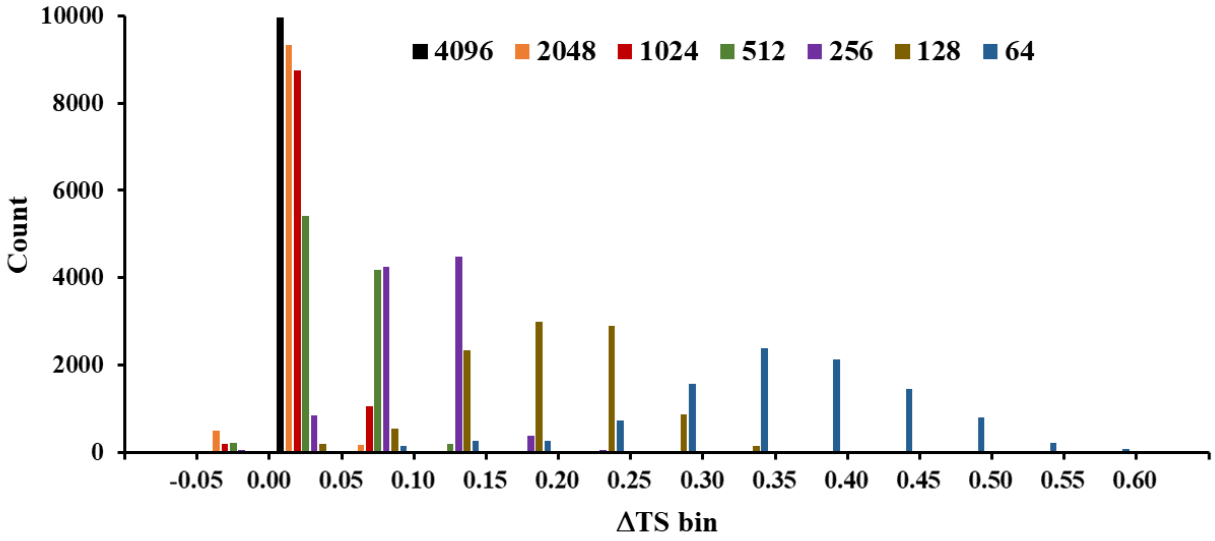

Fig. 5 Histogram of the difference between Tanimoto-similarity values (ΔTS) calculated from RDKit Morgan_2 fingerprints folded to a fixed length of 4,096 (black), 2,048 (orange), 1,024 (red), 512 (green), 256 (purple), 128 (brown), or 64 (blue) and Tanimoto-similarity values calculated from unfolded RDKit Morgan_2 fingerprints. The bin size was 0.05.