 doi:10.20944/preprints202202.0220.v1

Article

A Parametric Approach to Molecular Encodings of Carbon-based Multilevel Atomic Neighborhoods

Georges Hattab [1]*, Nils Neumann[1], Aleksandar Anžel [1] and Dominik Heider [1]

[1] Department of Mathematics and Computer Science, Philipps-Universität Marburg, Marburg, 35032, Germany.
* Correspondence: georges.hattab@uni-marburg.de

Abstract: Exploring new ways to represent and discover organic molecules is critical for developing novel therapies. With recent advances in bioinformatics, virtual screening of databases is possible. However, biochemical data must be encoded using computer algorithms to make them machine-readable, taking into account distance and similarity measures to support tasks such as similarity searching. Motivated by the ubiquity of the carbon element and the structured patterns that emerge, we propose a parametric approach to molecular encodings of carbon-based multilevel atomic neighborhoods. It implements a walk along the carbon chain of an organic molecule to compute different representations of its feature encoding in the form of a binary or numerical array that can be exported later into an image. Resulting encodings are reproducible and readily formatted for various domain tasks including machine learning tasks. This approach was evaluated using a 10-fold stratified cross validation for binary classification with eight data sets and six different encodings (384 models) in the domain knowledge of cell-penetrating peptides. The parametric approach is built on open-source software and is implemented as a Python package (cmangoes). Source code and documentation are available at https://github.com/ghattab/cmangoes.

Keywords: parametric; encoding; fingerprinting; machine learning; classification; transporter; cell-penetrating peptide.

1. Introduction

Computational approaches for molecular analysis support a variety of biologically oriented applications. From identifying the interactions between drugs and target proteins [1,2], to revealing quantitative relationships between structural properties of chemical compounds and biological activities [3], to screening a handful of membrane proteins for drug delivery [4], computational approaches are facilitated by the similar property principle [5]. It asserts that similar molecules will also tend to exhibit similar properties. For example, the virtual screening method is primarily used in drug discovery and allows researchers to find candidate treatments for Alzheimer's disease or HIV [6–8]. It is carried out by calculating similarity measures of compounds in a database to a reference compound. Using a similarity search, compounds are ranked in descending order and manual screening is performed on the highest ranked compounds [9]. Yet to support the growing number of machine-related tasks, the structure of a molecule must be encoded to a machine-readable format.

For instance, certain structural information may be represented as a numeric feature by means of mapping a large data item to a much shorter bit string. Indeed, different molecular fingerprint types have been proposed: Substructure key-based (*e.g.*, MACCS [2]), topological (*e.g.*, FP2 OpenBabel [10]), circular (*e.g.*, MNA [11]), pharmacophore, and hybrid. This process leads to a molecular fingerprint, which uniquely identifies each molecule through data encoding.

Given such a fingerprint, we can abstract task-specific information at different levels, from the atom, to the neighborhood of an atom, to the amino acid of a protein or even to the base of a DNA molecule. Thanks to this process of abstraction, various biological and chemical aspects may be characterized, similarities and differences may be noted. In the similarity searching example, distances such as Tanimoto or Dice coefficients are

 doi:10.20944/preprints202202.0220.v1

calculated between the fingerprint of a certain molecule and its reference during the search [4,12]. Although researchers have identified numerous distance measures and studied their limitations (*e.g.*, Manhattan, Soergel) [13], it has been shown that it is important to also uncover the underlying factors of a molecule that determine its similarity value or prediction by a machine learning (ML) model [14].

Given various bioinformatics tools implementing different molecular fingerprint types and fingerprinting algorithms [15,16], there is an immediate need for parametric molecular approaches that can be adapted to different tasks and specific user needs while respecting domain standards. This work addresses the specific problem of binary classification for cell-penetrating peptide data. Used in research and medicine, cell-penetrating peptides are also known as protein transduction domains and carry a variety of cargoes across the cellular membranes in an intact and functional form [17].

In considering a circular fingerprinting algorithm, the main idea is to be able to successfully classify peptides with certain features that characterize cell-penetrating peptides. To our knowledge and although various approaches exist, the concept of relying on the hierarchy of the neighborhoods have not been considered for this task. To this extent, we rely on the element carbon (C) to create different encodings. As the centerpiece of organic life, C is ubiquitous and very good at forming large and stable chains of various organic molecules. Inspired by its central role, we introduce a parametric approach to molecular encodings of carbon-based multilevel atomic neighborhoods as an open source standalone executable and a GitHub source repository; namely cmangoes. It takes as input a molecular sequence, user parameters such as the levels to be considered and whether an image representation is needed. This parametric approach allows the creation of user-defined molecular encodings. This paves the way for further efforts to tailor molecular encodings to specific user requirements while taking into account the parameter space of fingerprinting algorithms. In the following, we introduce the methodology of the proposed parametric approach and demonstrate its usefulness on eight data sets.

2. Materials and Methods

The presented work takes into account the ubiquity of the carbon element and its central role in holding together the structure of organic molecules and organizing their neighborhoods. The parametric approach encodes the neighborhoods around the carbon chain of a molecule in multiple levels. Various design considerations are followed to meet established domain standards and create compatible encodings for common similarity measures and distances. This section describes the parametric approach, design considerations of the underlying algorithm to fit the domain specificity, the data sets used, and the evaluation of the machine learning.

2.1. Parametric approach

The parametric approach handles the input data, generates intermediate data representations as a graph, traverses it to record the relevant neighborhoods according to user-specified parameters, transforms the recorded features to their final output format, and generates the corresponding image representations. The walk along the carbon chain iteratively lists the neighboring atoms of each visited atom. The hierarchies are multiple levels of an atom's neighborhood and are defined hierarchically based on their proximity to the carbon chain. By incorporating hierarchies in the encoding, molecules of varying lengths containing different substructures can be represented appropriately. An implementation of this approach is provided as a Python package for easy reproducibility. The core development of the algorithm was performed using Python programming language, version 3.8.5 [18,19]. The chosen language offers high compatibility with existing computational approaches commonly used in bioinformatics and cheminformatics. All core-related dependencies are listed in Table A1. The package accepts FASTA or SMILES file format specifications and follows a seven-step encoding pipeline. Figure 1 depicts an example workflow diagram for an input molecule as a SMILES string.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2022 doi:10.20944/preprints202202.0220.v1

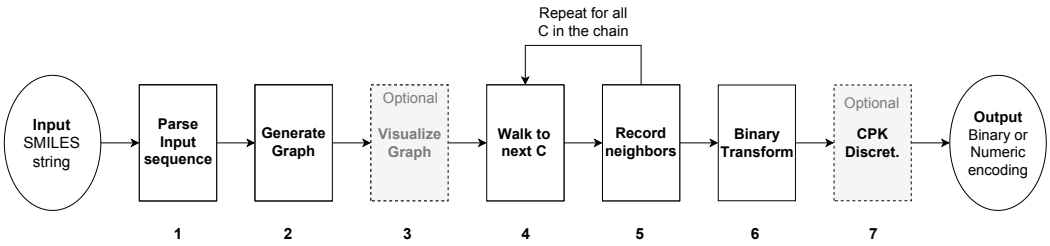


Figure 1. Example workflow of the encoding pipeline for a given molecule. C: Carbon. Corey-Pauling-Koltun Discretization: CPK Discret.

The first step consists of parsing and processing the input data, given in one of two available formats. The file format specification is identified. When a FASTA file is provided, the molecular information is converted into the SMILES format. To ensure all atoms of a given molecule are present for all following steps of the parametric algorithm, hydrogen atoms (H) are added upon data import.

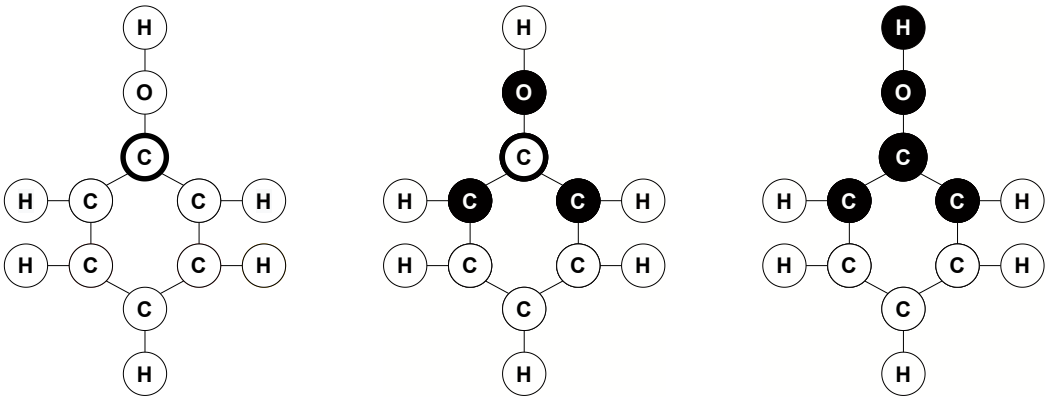
Second, an intermediary molecular graph data structure is employed to efficiently traverse the carbon chain to record the relevant neighborhoods. To generate the molecular graph, the input data is parsed into an adjacency matrix which is then transformed into a graph. To create a robust and deterministic encoding, all atoms of a given molecule are represented by nodes in the molecular graph and are numbered with a unique identifier. Each node in the molecular graph stores the type of element they represent using its element symbol from the periodic table. The element labels are required in subsequent steps to generate the feature vectors. To avoid redundant information not required as part of the encoding, the edges of the molecular graph do not store any additional information aside from the nodes they connect, *i.e.*, an unweighted graph.

Third, to aid the identification of the optimal depth, the molecular graph may be visualized. When a data set is used, users select the molecule of their choice in the data set. The intermediary graph is visualized.

Fourth, the walk along the carbon chain corresponds to an iteration over a numbered list of carbon atoms. This list is created by only retaining the nodes that correspond to the carbon (C) element symbol. Each carbon node is included exactly once. The filtered nodes are then sorted in ascending order by their unique node identifier. The immediate neighbors include all nodes connected directly by an edge to the respective carbon node. Aside from the immediate neighbors, additional hierarchy levels of a neighborhood can be saved. Figure 2 shows an illustrative iteration for the example phenol molecule.

Fifth, neighborhoods along the previously mentioned walk are saved. The additional hierarchy levels are defined as the immediate neighbors of all nodes belonging to the previous level. For instance, the second-level hierarchy includes all nodes with a direct connection to any node from the first-level hierarchy. To avoid redundancy in the encoded information, an additional filter is applied when recording more than one hierarchy. Since neighborhoods are recorded as part of the main iteration, this filter excludes nodes containing carbon atoms. The number of recorded hierarchies can be set using the level parameter. The data structure used for saving the neighborhoods is a dictionary. Only the element symbols belonging to the neighborhood's nodes are saved. The list of element symbols is, by nature of the iteration, automatically sorted according to the unique node identifiers. This ensures that the feature vectors are deterministic across multiple runs of the encoding. To simplify subsequent steps of the algorithm and feature-based operations, such as transformation, the dictionary is transformed to a data frame. Table 2 reports the resulting hierarchies for the example phenol molecule.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2022 doi:10.20944/preprints202202.0220.v1



(a) C at example iteration. **(b)** The first level: C,C,O. **(c)** The second level: C,H.

Figure 2. Visual demonstration of a computation for two-level hierarchies of the phenol molecule (C_6H_6O). Each figure corresponds to one iteration along the carbon chain The algorithm reaches (a) the highlighted carbon atom at an example iteration and records (b) first- and (c) second-level hierarchy. The resulting hierarchy is C,C,O,C,H. The algorithm iterates onto the next carbon atom.

Sixth, feature transformation (binary and discretization) is applied on the hierarchies. Feature transformation enables numeric operations and image generation of the resulting encodings. In the example of the binary encoding, the feature vectors are represented as bit strings, 0 and 1 encode the absence or presence of an atom in the respective neighborhood. The resulting categorical data frame may include missing values depending on the structure of the encoded compound. This occurrence may result when recording more than one hierarchy level as aforementioned when carbon nodes are excluded. Indeed, to preserve the overall data structure integrity and avoid the occurrence of an unequal number of recorded atoms in the neighborhoods along the carbon chain, the data frame is automatically filled with missing values in the relevant positions. Since the value 0 represents the absence of information, this procedure does not distort the resulting feature vector.

Seventh, the numerical encodings in the image space follow either a 1-bit coding or the Corey Pauling Koltun colors (CPK). This optional step exports either binary or numeric/discretized encodings. Table 3 and table 4 show the resulting output after feature transformation for the example phenol molecule. Figure 4 depicts the image representations of the resulting transformations.

2.2. Domain specificity

The created feature vectors are compatible with domain-specific tasks such as database querying and virtual screening [2,4].

In the special case of cyclic molecules, for example aromatic cycles, the unique node identifiers and the sorted filtered list permit the algorithm to avoid cyclical substructures. Figure 2 shows an example iteration for the phenol molecule. Table 2 reports the resulting hierarchies. Figure 4 depicts the image representations of the resulting transformations.

The image representations complement the mathematical feature vectors to provide an accessible way to understand the resulting encodings and enable additional operations in the image space [20]. Figure 4 depicts the image representations of the resulting transformations for the phenol molecule. To avoid discrepancies or dimensional mismatches in the output feature vector and avoid bit collision for different molecule sizes, a padding step is included in the encoding pipeline when applying the encoding to more than one molecule. It includes two padding strategies to either top-left shift or center the encoding by introducing new empty pixels around the edges of an image.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2022 doi:10.20944/preprints202202.0220.v1

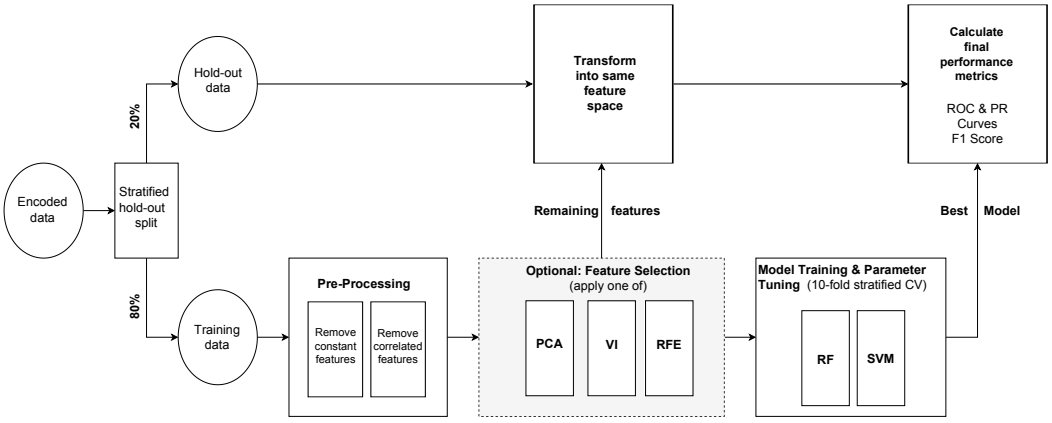


Figure 3. Machine Learning pipeline to evaluate an encoding using the parametric approach. RF: Random Forest. RFE: Recursive Feature Elimination. PCA: Principal Component Analysis. VI: Variable Importance. SVM: Support Vector Machine.

2.3. Data sets

Table 1 lists all the data sets included in this work and reports their class imbalance for the evaluation. Data sets prefixed by *cpp_* contain peptides with the common task of identifying cell-penetrating peptides (CPPs). The properties encoded in the target vectors are represented by ones and zeros, corresponding to the presence or absence of the relevant property, respectively. For six data sets, the property encoded in the target vector is whether or not the peptide is cell-penetrating. The `ace_vaxinpad` data set also contains peptides, but its target vector represents information about stimulating antigen presenting cells. The `hiv_protease` data set consists of data used for the prediction of human immunodeficiency virus HIV-1 protease cleavage sites.

2.4. Evaluation

We rely on a ML pipeline to evaluate the parametric approach. Eight data sets in the domain knowledge of peptide encodings are employed. The pipeline includes dimensionality reduction, model training, validation strategies, and evaluation metrics. The evaluation pipeline is presented as a workflow diagram in Figure 3.

The R Project for Statistical Computing language was used to carry out the evaluation of the resulting encodings. The external packages employed in the implementation support tasks such as computationally efficient machine learning model training and the extraction of performance metrics. An overview of all external packages used for this work can be found in Table A2.

To minimize bias based on the data set choice, 48 different scenarios are evaluated (six encodings per data set). These scenarios result from all possible combinations of a data set, a recorded hierarchy level, and a feature transformation strategy. They are combined with four different dimensionality reduction options (none, principal component analysis, variable importance, and recursive feature elimination) and two classifiers (the random forest and support vector machine), totaling 384 trained models.

To ascertain whether the class-imbalance and the data set size has an effect on the prediction quality, both the class distribution of the respective target vectors and the number of observations contained in each data set vary.

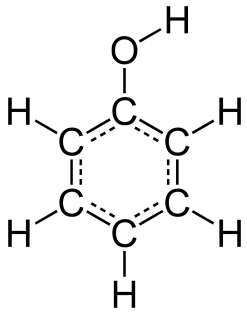
3. Results

The parametric approach provides a simplified set of parameters to adapt the encoding to user-specific needs. It is available as a standalone Linux executable and the source code GitHub repository to create encodings, explore their parameter space, and generate hypotheses and design ML experiments.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2022 doi:10.20944/preprints202202.0220.v1

Data set	Size	Imbalance Ratio	Source
ace_vaxinpad	688	0.79	[21]
cpp_cellppd	1614	1	[22]
cpp_cellppdmod	1462	1	[23]
cpp_cppredfl	924	1	[24]
cpp_kelmcpp	1003	1	[25]
cpp_mlcpp	1903	0.63	[26]
cpp_mlcppue	374	1	[26]
hiv_protease	947	0.19	[27]

Table 1. Data sets used in the evaluation from the specific domain of cell-penetrating peptides.



C_0	C_1	C_2	C_3	C_4	C_5
C	C	C	C	C	C
C	C	C	C	C	C
H	H	H	O	H	H
C	C	C	C	C	C
-	-	-	H	-	-

Table 2. The structural formula of the phenol molecule and its recorded neighborhoods using one- and two-level hierarchies. Phenol or C_6H_6O has the SMILES specification: `C1=CC=C(C=C1)O`. Canonical SMILES: `Oc1ccccc1`. C_0 is located at the bottom of the cycle. C_3 is at the top and is connected to the Oxygen element (O).

C_0C	C_0H	C_1C	C_1H	C_2C	C_2H	C_3C	C_3H	C_3O	C_4C	C_4H	C_5C	C_5H
1	0	1	0	1	0	1	0	0	1	0	1	0
1	0	1	0	1	0	1	0	0	1	0	1	0
0	1	0	1	0	1	0	0	1	0	1	0	1
1	0	1	0	1	0	1	0	0	1	0	1	0
0	0	0	0	0	0	0	1	0	0	0	0	0

Table 3. Recorded neighborhoods of the phenol molecule using one- and two-level hierarchies after binary transformation. To enable distance-based, similarity searching and machine learning tasks, the categorical encoding is transformed using dummy encoding.

C_0C	C_0H	C_1C	C_1H	C_2C	C_2H	C_3C	C_3H	C_3O	C_4C	C_4H	C_5C	C_5H
3	0	3	0	3	0	3	0	0	3	0	3	0
3	0	3	0	3	0	3	0	0	3	0	3	0
0	2	0	2	0	2	0	0	5	0	2	0	2
3	0	3	0	3	0	3	0	0	3	0	3	0
0	0	0	0	0	0	0	2	0	0	0	0	0

Table 4. Recorded neighborhoods for the phenol molecule using one- and two-level hierarchies after CPK-based discretization. The parametric approach transforms the features to integers ranging from 0 to 16 as per the CPK coloring system.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2022 doi:10.20944/preprints202202.0220.v1

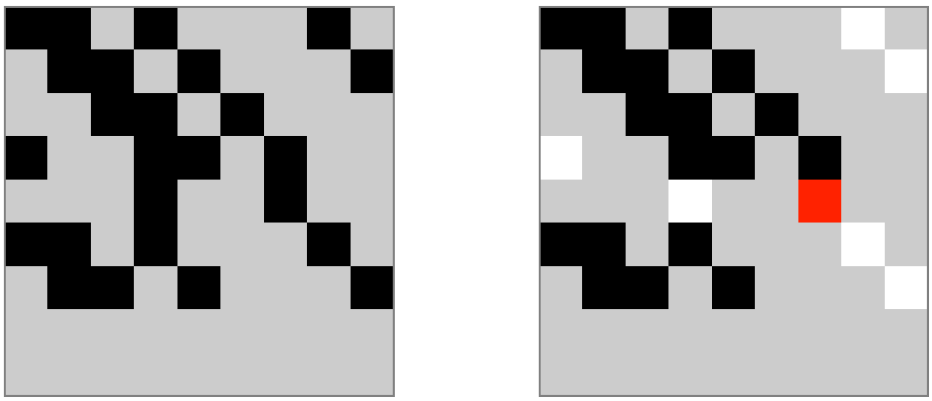


Figure 4. Image representations of the encoding for the phenol molecule. (Left) The binary encoding. (Right) The CPK-colored encoding. The images are generated based on the feature vectors found in Table 3 and Table 4, respectively. Since image generation introduces additional computational overhead, this step is implemented as an optional parameter.

Method	Retained features (%)
Principal Component Analysis	5
Variable importance	26
Recursive Feature Elimination	10

Table 5. Average percentage of retained features after dimensionality reduction across all encoded data sets.

With the pre-processing and feature selection steps, it was possible to reduce the average number of retained features across all encoded data sets by up to 90%. A full overview over the mean reduction in features for the pre-processing step and all three applied feature selection methods can be found in Table 5.

With a p-value of $p \approx 0.003$, the RF model performed significantly better than the linear weighted SVM across all eight data sets. On average and across all data sets, the best performance was achieved with the RF model and RFE feature selection. In this case, only first-level hierarchies were recorded and transformed using the binary transformation strategy. The model's complete performance metrics can be found in Table 6. The corresponding Receiver Operating Characteristic (ROC) and Precision-Recall curves can be found in Figure 5 and Figure 6, respectively. By means of sensitivity analysis, the optimal classification model was obtained for each data set, as shown in Table 7.

4. Discussion

First, further computational improvements can be made for both the parametric approach and the evaluation pipeline. On one hand, very large data sets can be batched encoded and as such parallel processing of the input molecules is relevant. On the other hand, the large number of training iterations makes computational optimizations especially important. Although we rely on two-dimensional multilevel neighborhoods alone, this work provides a proof of concept and the evaluation of other fingerprinting algorithms for binary classification should be considered, as reported in previous work [16]. However, due to the more elaborate ML pipeline, our results cannot be directly compared to the classification results reported in the Peptide Reactor tool. To enable a fair and direct comparison, we are working on integrating this parametric approach into this tool and running the same machine learning models.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2022 doi:10.20944/preprints202202.0220.v1

Data set	ROC AUC	F_1 Score
ace_vaxinpad	0.73	0.64
cpp_cellppd	0.82	0.77
cpp_cellppdmod	0.84	0.74
cpp_cppredfl	0.59	0.64
cpp_kelmcpp	0.71	0.63
cpp_mlcpp	0.73	0.62
cpp_mlcppue	0.61	0.68
hiv_protease	0.74	0.49

Table 6. Performance metrics for the best performing classification model overall. (Hierarchies: 1st level; Feature transformation: Binary; Classifier: Random Forest; Feature selection: RFE)

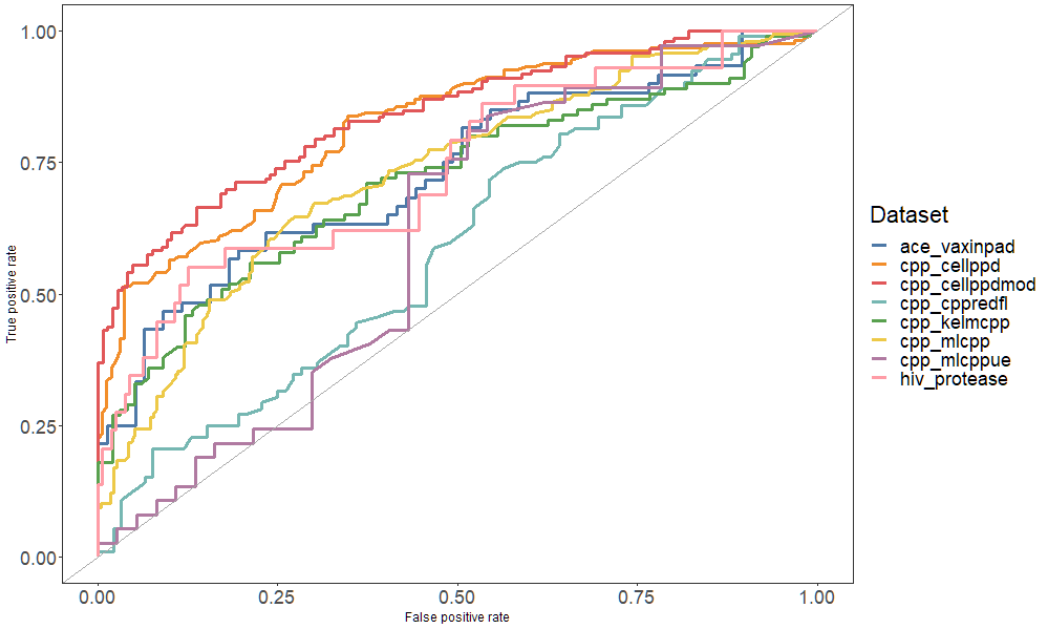


Figure 5. ROC curves for the best performing classification model overall. (First-level hierarchy. Feature transformation: Binary. Classifier: Random forest. Feature selection: RFE).

Data set	Encoding	ML	AUC	F_1 Score
ace_vaxinpad	2-level HC; CPK	RF; All features	0.85	0.79
cpp_cellppd	1-level HC; Binary	RF; RFE	0.82	0.77
cpp_cellppdmod	2-level HC; CPK	RF; RFE	0.89	0.83
cpp_cppredfl	1-level HC; CPK	RF; RFE	0.69	0.72
cpp_kelmcpp	2-level HC; CPK	RF; PCA	0.73	0.75
cpp_mlcpp	1-level HC; Binary	RF; All features	0.73	0.63
cpp_mlcppue	1-level HC; Binary	SVM; PCA	0.64	0.71
hiv_protease	3-level HC; Binary	RF; All features	0.82	0.50

Table 7. Optimal classification model for each data set. H: Hierarchies. CPK: Corey-Pauling-Koltun. RF: Random Forest. RFE: Recursive Feature Elimination. PCA: Principal Component Analysis. SVM: Support Vector Machine.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 February 2022 doi:10.20944/preprints202202.0220.v1

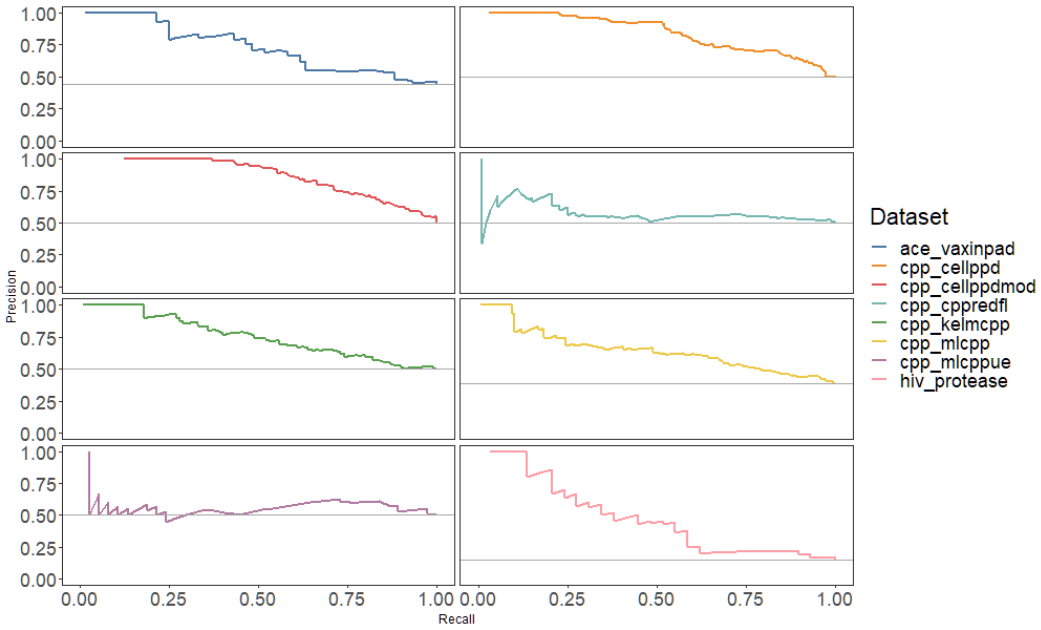


Figure 6. Precision-Recall curves for the best performing classification model overall. (First-level hierarchy. Feature transformation: Binary. Classifier: Random forest. Feature selection: RFE).

Second, compared to the state-of-the-art results reported in the original works that have published the six CPPs data sets (*c.f.*, Table 1), our results were found to have an acceptable to good separation of the two classes, *i.e.*, robustness. It is important to note that in the majority of these works, both the accuracy and the Matthews correlation coefficient performance metrics were used and this is in discordance with good practices for binary classification. Indeed, the AUC provides robustness of the resulting classifier and is more discriminative than Matthews correlation coefficient, while the accuracy is the measure of the closeness to a specific value and the AUC is the measure across all the possible thresholds [28–30]. The empirical evaluation leads us to conclude that our results are better except for the `cpp_cellppdmod` data set. Compared to the related work which achieved a near-perfect classification with an AUC of 0.98, we reached AUC values of 0.89 and 0.84 for the optimal model on this data set and the overall model, respectively. This implies that there is little trade-off between specificity and sensitivity [23].

Third, the molecular complexity field is noteworthy. It provides fundamental concepts that underly current fragment-based lead discovery. It considers the general index of molecular complexity, where features that make a molecule more or less complex are taken into account [31,32]. For example, size, symmetry, branching, rings, multiple bonds and heterogeneity in the atoms. Such concepts have been used in various application domains such as chromatography analysis and synthesis pathways. It would be very useful to rely on such features to improve the proposed approach and introduce further parameters such as symmetry, the presence of a cycle, or even the distances among atoms. Such additions may be made at the second step of the parametric approach, that is the intermediate molecular graph, by addition of metadata to the aforementioned dictionary. In turn, this could lead to an enrichment of the resulting encodings and in extension an increase of the user-settable parameters.

Fourth, although this parametric approach proved useful for cell-penetrating peptides and achieved acceptable classification results, it is important to extend its usage to include larger molecules and more heterogeneous data sets such as membrane proteins [33,34]. Moreover, since this parametric approach is designed to accommodate the calculation of similarity measures and distances, it is relevant to gauge its performance for other tasks such as similarity searching.

 doi:10.20944/preprints202202.0220.v1

Fifth, the image representations of the resulting encodings constitute an interesting research starting point. Images of the twenty essential amino acids encoded as a one- or two-level hierarchy using the CPK discretization are reported in Figure A1 and A2, respectively. Indeed, they open up a new space of representation by using the image domain. For example, convolutional neural networks may be used for the same task of classification yet by relying on the images of the resulting encodings. Since such neural networks convolve learned features with input data, and use two-dimensional convolutional layers, their architecture is suited to processing two-dimensional data, such as images. Indeed, they would eliminate the need for manual feature extraction required to classify the images.

Sixth, the dimensionality reduction results indicate that the resulting encodings are by default very sparse. This is especially the case when the encodings are padded or centered. Hence, it is important to develop specialized methods to address sparsity and evaluate its effects. This relates to the problem of representation and has potential links to data compression. Further considerations are warranted for a more faithful space of representation so to reduce the data and preserve its relevant structure.

Seven, evaluating all models using good practices in machine learning is a standard approach to optimize the prediction performance of the ML models. Since the parametric approach provides a parameter space, researchers can conduct a sensitivity analysis to better fine-tune resulting machine learning models.

Eighth and last, although the parametric approach is focused on organic compounds or molecules, it is possible to adapt the underlying algorithm to create multilevel atomic neighborhoods of molecules that lack C-H bonds. That is to say, considering inorganic polymers with a skeletal structure that does not include carbon atoms.

5. Conclusions

The presented parametric approach is created as an easy-to-use and install solution that includes the necessary operations to create custom multilevel encodings of molecular data. Employing the ML pipeline for the binary classification task produced very promising results when combining random forests and recursive feature elimination. Performance metrics such as the ROC AUC and F1-score reached 0.89 and 0.83, respectively. The evaluation using the ML pipeline indicated that the first two-level hierarchies carry the most meaningful information for the classification task. We foresee that the proposed work will be a valuable tool to complement and enhance current molecular fingerprinting algorithms and offer further insights into the parameters and the use of hierarchies and their potential combination.

Author Contributions: Conceptualization, G.H. and N.N.; methodology, G.H., N.N. and A.A.; software, N.N. and A.A.; validation, G.H. and N.N.; investigation, G.H. and N.N.; resources, A.A.; data curation, G.H. and A.A; writing—original draft preparation, G.H.; writing—review and editing, G.H., A.A. and D.H.; visualization, G.H. and N.N.; supervision, D.H.; project administration, G.H.; funding acquisition, D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Hessian Ministry for Science and the Arts (LOEWE) in the context of the Molecular Storage for Long-Term Archiving (MOSLA) consortium.

Data Availability Statement: The employed data sets employed in this study are available with their respective original source at https://github.com/ghattab/CMANGOES/tree/main/Data

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript (alphabetical order):

 doi:10.20944/preprints202202.0220.v1

AUC	Area under the ROC Curve
C	Carbon
CPK	Corey Pauling Koltun
CMANGOES	Carbon-based multilevel atomic neighborhood encodings
CPPs	Cell-penetrating peptides
FASTA	FAST-All
H	Hydrogen
HC	Hierarchies
HIV	Human immunodeficiency virus
ML	Machine Learning
PCA	Principal Component Analysis
RF	Random Forest
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
SMILES	Simplified Molecular-Input Line-Entry System

Appendix A

Package name	Version number
Biopython	1.7.7
Matplotlib	3.3.0
NetworkX	2.4
NumPy	1.19.1
pandas	1.1.0
RDKit	2020.09.1
pysmiles	1.0.1

Table A1. Python packages used for the implementation of the parametric approach.

Package name	Version number
caret	6.0.86
doParallel	1.0.16
dplyr	1.0.2
e1071	1.7.4
ggplot2	3.3.2
ggpubr	0.4.0
ggthemes	4.2.0
janitor	2.0.1
MLeval	0.3
mlr	2.18.2
ranger	0.12.1
Rfast	2.0.1

Table A2. R packages used for the implementation of the ML pipeline.

References

1. Csermely, P.; Korcsmáros, T.; Kiss, H.J.; London, G.; Nussinov, R. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & therapeutics* **2013**, *138*, 333–408.
2. Cereto-Massagué, A.; Ojeda, M.J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58–63.
3. Neves, B.J.; Braga, R.C.; Melo-Filho, C.C.; Moreira-Filho, J.T.; Muratov, E.N.; Andrade, C.H. QSAR-based virtual screening: advances and applications in drug discovery. *Frontiers in pharmacology* **2018**, *9*, 1275.
4. Bajusz, D.; Rácz, A.; Héberger, K. Chemical data formats, fingerprints, and other molecular descriptions for database analysis and searching. In *In Silico Drug Discovery Tools*; Elsevier Inc., 2017; pp. 329–378.

 doi:10.20944/preprints202202.0220.v1

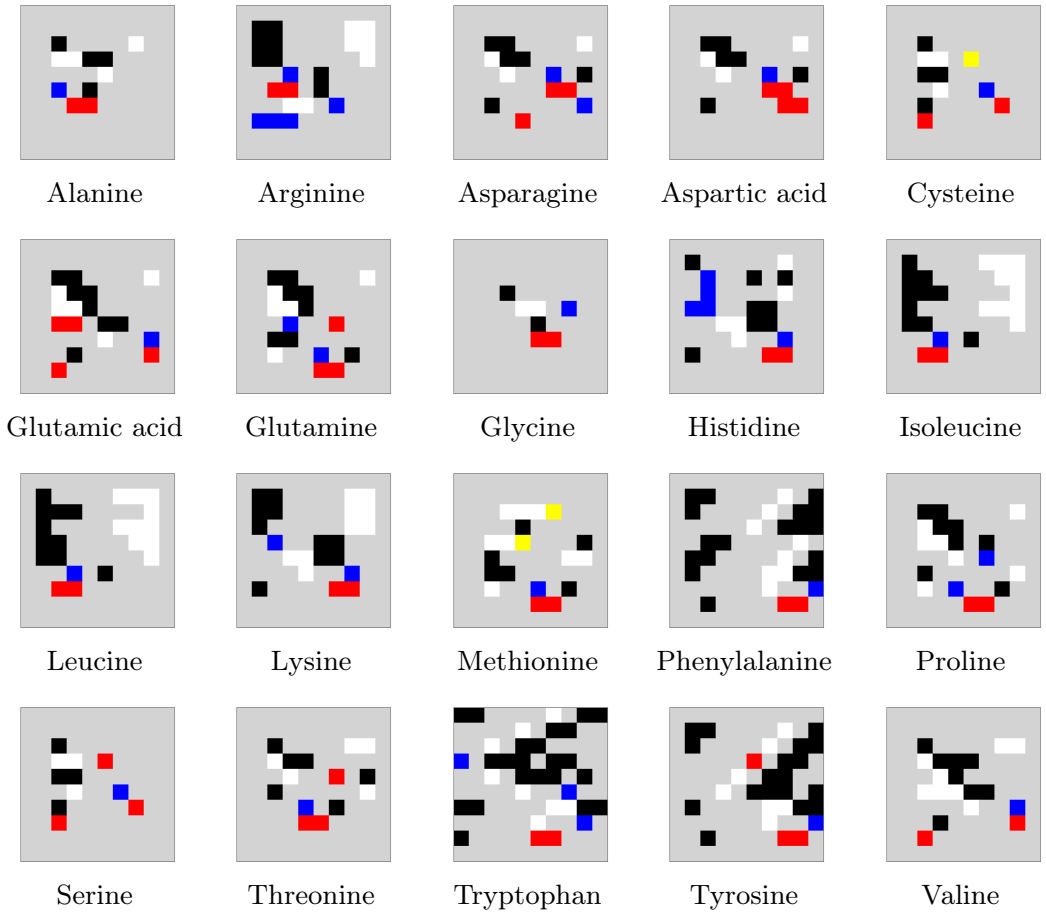


Figure A1. Images of the twenty essential amino acids. First-level hierarchy is encoded and transformed (CPK). The recorded neighborhoods along the carbon chain may be identified. For instance, for the cysteine amino acid, the oxygen atom (red), the nitrogen atom (blue), and the sulfur atom (yellow) can easily be seen. Other aspects like the minor structural difference between leucine and isoleucine are immediately apparent from the arrangement of carbon (black) and hydrogen (white) blocks in the respective images.

 doi:10.20944/preprints202202.0220.v1

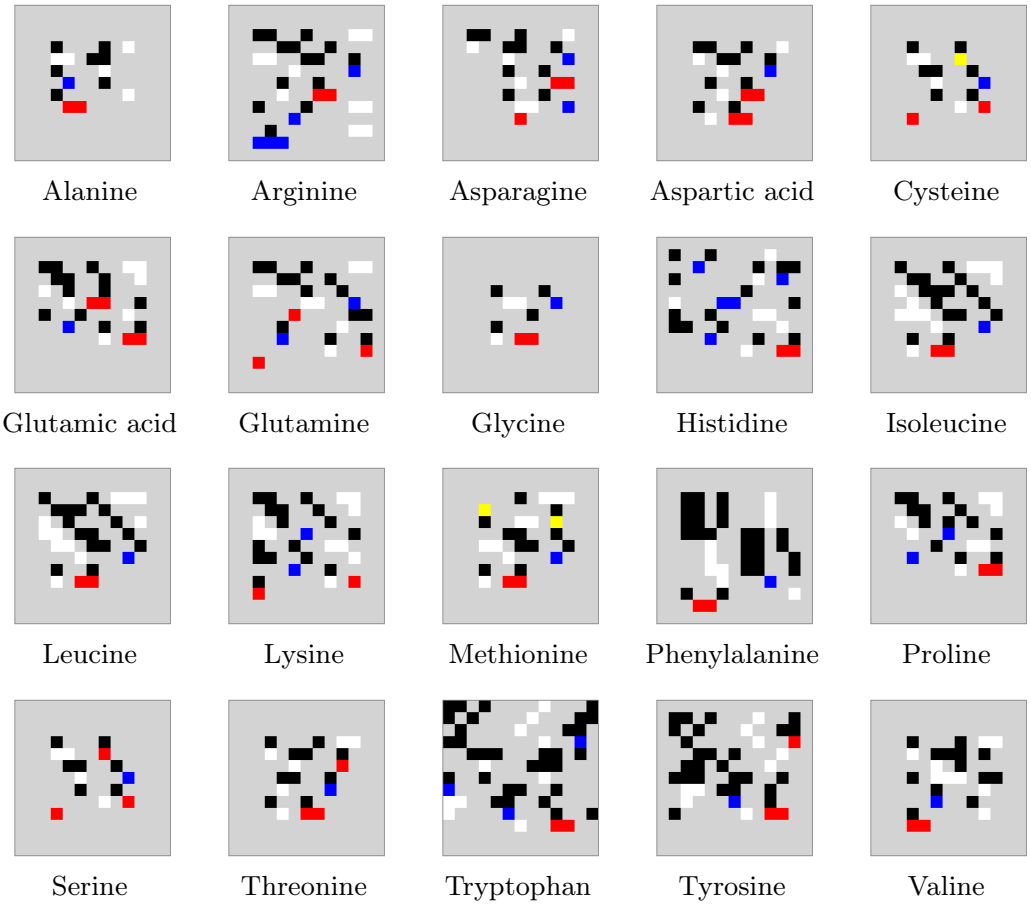


Figure A2. Images of the twenty essential amino acids. Two-level hierarchy encoded and transformed (CPK).

 doi:10.20944/preprints202202.0220.v1

5. Johnson, M.A.; Maggiora, G.M. *Concepts and applications of molecular similarity*; Wiley, 1990.
6. Ponzoni, I.; Sebastián-Pérez, V.; Martínez, M.J.; Roca, C.; De la Cruz Pérez, C.; Cravero, F.; Vazquez, G.E.; Páez, J.A.; Díaz, M.F.; Campillo, N.E. QSAR Classification Models for Predicting the Activity of Inhibitors of Beta-Secretase (BACE1) Associated with Alzheimer's Disease. *Scientific reports* **2019**, *9*, 1–13.
7. Vora, J.; Patel, S.; Sinha, S.; Sharma, S.; Srivastava, A.; Chhabria, M.; Shrivastava, N. Molecular docking, QSAR and ADMET based mining of natural compounds against prime targets of HIV. *Journal of Biomolecular Structure and Dynamics* **2019**, *37*, 131–146.
8. Dybowski, J.N.; Riemenschneider, M.; Hauke, S.; Pyka, M.; Verheyen, J.; Hoffmann, D.; Heider, D. Improved Bevirimat resistance prediction by combination of structural and sequence-based classifiers. *BioData mining* **2011**, *4*, 1–13.
9. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug discovery today* **2006**, *11*, 1046–1053.
10. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An open chemical toolbox. *Journal of cheminformatics* **2011**, *3*, 33.
11. Filimonov, D.; Poroikov, V.; Borodina, Y.; Gloriozova, T. Chemical similarity assessment through multilevel neighborhoods of atoms: definition and comparison with the other descriptors. *Journal of chemical information and computer sciences* **1999**, *39*, 666–670.
12. Deepak, P.; Deshpande, P.M. *Operators for similarity search: Semantics, techniques and usage scenarios*; Springer, 2015.
13. Riniker, S.; Landrum, G.A. Similarity maps-a visualization strategy for molecular fingerprints and machine-learning methods. *Journal of cheminformatics* **2013**, *5*, 43.
14. Godden, J.W.; Xue, L.; Bajorath, J. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *Journal of Chemical Information and Computer Sciences* **2000**, *40*, 163–166.
15. Spänig, S.; Heider, D. Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. *BioData Mining* **2019**, *12*, 7.
16. Spänig, S.; Mohsen, S.; Hattab, G.; Hauschild, A.C.; Heider, D. A large-scale comparative study on peptide encodings for biomedical classification. *NAR genomics and bioinformatics* **2021**, *3*, lqab039.
17. Taylor, R.E.; Zahid, M. Cell penetrating peptides, novel vectors for gene therapy. *Pharmaceutics* **2020**, *12*, 225.
18. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, 2009.
19. Oliphant, T.E. Python for scientific computing. *Computing in Science & Engineering* **2007**, *9*, 10–20.
20. Keim, D.A.; Mansmann, F.; Schneidewind, J.; Ziegler, H. Challenges in visual data analysis. Tenth International Conference on Information Visualisation (IV'06). IEEE, 2006, pp. 9–16.
21. Nagpal, G.; Chaudhary, K.; Agrawal, P.; Raghava, G.P. Computer-aided prediction of antigen presenting cell modulators for designing peptide-based vaccine adjuvants. *Journal of translational medicine* **2018**, *16*, 181.
22. Gautam, A.; Chaudhary, K.; Kumar, R.; Sharma, A.; Kapoor, P.; Tyagi, A.; Raghava, G.P.; et al. In silico approaches for designing highly effective cell penetrating peptides. *Journal of translational medicine* **2013**, *11*, 74.
23. Kumar, V.; Agrawal, P.; Kumar, R.; Bhalla, S.; Usmani, S.S.; Varshney, G.C.; Raghava, G.P. Prediction of cell-penetrating potential of modified peptides containing natural and chemically modified residues. *Frontiers in microbiology* **2018**, *9*, 725.
24. Wei, L.; Tang, J.; Zou, Q. SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. *BMC genomics* **2017**, *18*, 1–11.
25. Pandey, P.; Patel, V.; George, N.V.; Mallajosyula, S.S. KELM-CPPpred: Kernel extreme learning machine based prediction model for cell-penetrating peptides. *Journal of proteome research* **2018**, *17*, 3214–3222.
26. Manavalan, B.; Subramaniyam, S.; Shin, T.H.; Kim, M.O.; Lee, G. Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *Journal of proteome research* **2018**, *17*, 2715–2726.
27. Rögnvaldsson, T.; You, L.; Garwicz, D. State of the art prediction of HIV-1 protease cleavage sites. *Bioinformatics* **2015**, *31*, 1204–1210.
28. Ling, C.X.; Huang, J.; Zhang, H.; et al. AUC: a statistically consistent and more discriminating measure than accuracy. Ijcai, 2003, Vol. 3, pp. 519–524.
29. Calders, T.; Jaroszewicz, S. Efficient AUC optimization for classification. European conference on principles of data mining and knowledge discovery. Springer, 2007, pp. 42–53.
30. Halimu, C.; Kasem, A.; Newaz, S.S. Empirical comparison of area under ROC curve (AUC) and Mathew correlation coefficient (MCC) for evaluating machine learning algorithms on imbalanced datasets for binary classification. Proceedings of the 3rd international conference on machine learning and soft computing, 2019, pp. 1–6.
31. D'Amboise, M.; Bertrand, M.J. General index of molecular complexity and chromatographic retention data. *Journal of Chromatography A* **1986**, *361*, 13–24.
32. Hendrickson, J.B.; Huang, P.; Toczko, A.G. Molecular complexity: a simplified formula adapted to individual atoms. *Journal of Chemical Information and Computer Sciences* **1987**, *27*, 63–67.
33. Chou, K.C.; Elrod, D.W. Prediction of membrane protein types and subcellular locations. *Proteins: Structure, Function, and Bioinformatics* **1999**, *34*, 137–153.
34. Hattab, G.; Warschawski, D.E.; Moncoq, K.; Miroux, B. Escherichia coli as host for membrane protein structure determination: a global analysis. *Scientific reports* **2015**, *5*, 1–10.