University of Amsterdam

Master Artificial Intelligence

The effect of added chemical attributes on the classification of graph based molecular datasets

Author:
Jeroen Baars
(206873049)

Lecturer:
dr. Pascal Mettes

July 29, 2020

Universiteit van Amsterdam

Rijksinstituut voor Volksgezondheid en Milieu
Ministerie van Volksgezondheid, Welzijn en Sport

Abstract

This work studies the influence of altered conventional graph representations on the classification of toxicity data-sets. Predicting physico-chemical properties is an important instrument in chemo-informatics used for different applications and screening tools. A core challenge in this field is to exploit the role of the spatial structure in graphs and the correct use of chemical attributes. To that end, we propose three new graph representations options with added chemically based spatial or node attributes. The attributes have a known correlation with toxicity. For the first representation, these are molecular bonds. For the second, hydrogen bonds, hydrophobicity, and aromatic rings. For the third representation, these are reactivity sites for metabolism. To research the influence we make use of the dutch substances of very high concern list as a use case, two different verified toxic endpoints are tested. Experimental evaluation shows a potential of data-set specific representations to improve the recall and balanced accuracy in contrast to the conventional representation, complementary models show an improvement in precision but fail to improve the recall. We conclude that these additions show considerable potential and are confident that further research can adhere to better screening tools.

Contents

1 Introduction

In the field of chemo-informatics, there is an ongoing stream of newly produced chemicals. The production or introduction of these chemicals requires the classification, labeling, and activity prediction of these chemicals and therefore testing. These tests can be done in vitro, in vivo, or in silico; meaning in living organisms, in petri dishes, or in (computer) modeling [1]. In vitro or in vivo testing is as time and cost consuming for the companies or organizations creating or using the chemicals as for the regulatory instances monitoring these chemicals. For example, the testing of nine traditional hazard classifications. The assays used for the testing of these hazards take up to 57 percent of all animal testing in Europe, which accounts for about 600.000 animals per year [2]. The costs of bringing for example a pesticide into production, which is an extreme case, is 20 million dollars and requires about 10.000 animals before it can be brought to the market. A less extreme case, such as industrial chemicals, can still require 5 million dollars of testing before it is allowed[2].

To reduce the need and cost of in vitro or in vivo testing or at least to narrow down the set of chemicals on which these tests are required the chemicals can already be screened in silico. The molecular structure is one of the properties that lend itself for these tests, modeling the Structure Activity Relationship (SAR)[3] can give insights into the probabilities of a molecule belonging to a certain toxicity class or having particular characteristics. This relationship states that molecules with a similar structure have a high likelihood to share the same characteristics. New molecules that have a similar structure to a group of toxic molecules already classified as such are thus more likely, because of this relationship, to also have the same toxic property. Modeling the SAR can be used to prioritize chemical screening and to assert the viability of newly introduced chemicals. The SAR modeling was introduced fifty years ago as Quantitative Structure Activity Relationship (QSAR)[4] modeling which can be used for the classification of chemical labels or the regression of chemical characteristics.

In these fifty years, QSAR modeling has improved within the field of chemo-informatics and recently also combined with the field of artificial intelligence [5]. Together with the growth of open-source chemical datasets, the field has multiple machine learning and deep learning applications introduced. From modeling the chemical fingerprints of a molecule with a support vector machine to letting the deep learning algorithms learn the complex relations from the molecular structure itself. In combination with deep learning the creation, testing, and analysis of chemicals is shifted to computational tools before the lab testing is involved. For example, generative adversarial network developing chemicals [6], molecular algorithms predicting physico-chemical properties like solubility of a molecule [7] and different toxicity screening and prediction tools [5].

Together with these advancements, questions arise on how to tackle and solve emerging problems in the chemo-informatics field. For Graph Convolutional Neural Networks (GCNN), a molecular graph-based approach, is stated that ".. helped us reason about the role of structure, as we found that structure-agnostic

baselines outperform GNNs on some chemical data-sets, thus suggesting that structural properties have not been exploited yet."[8] Also, the strength of the GCNN's is disputed, by removing the spatial based approach and creating a simpler network, the same results can be achieved as with GCNN's [9]. Further, the usage of added attributes is questioned as is stated that "It will be an interesting question to see how to incorporate attributes in a more effective yet still simple manner in our graph representation"[10]. On the other hand, for larger non chemo-informatic data-sets the GCNN's are found to outperform the baseline [9] [8]. Thus, GCNN's methods are reviewed as not fully exploiting the structural data within chemo-informatic datasets, and possibilities are seen on incorporating attributes.

To take a step towards solving these outstanding questions in the field we propose and research new representations; the representations have chemical properties added as supplementary information to the conventional molecular graph structure used in graph convolutional neural networks. We will research the performance influence of this supplementary information on the prediction of toxic endpoints within chemo-informatics data-sets. To do so, we will research the difference in performance between the conventional representation with atoms and bonds as edges and nodes and 3 modified representations, the three representations have supplementary chemical information incorporated all of which have a known correlation with toxic properties.

The use case for these new representations will be the classification of two small toxicity datasets. These data-sets are gathered from the dutch Substances of Very High Concern (nSVHC) lists of the RijksInstituut van Volksgezondheid en Milleu (RIVM) in the Netherlands, which contain verified toxic molecules, and are complemented with a set of verified non-toxic molecules.

1.1 Research Question

Can the use of added chemical based attributes contribute to the classification of toxicity within chemo-informatics data-sets with the use of graph convolution neural networks?

The next items are sub-goals intended to research the influence of added chemical attributes.

- Implement three new representations with chemical attributes added and research the influence of these new representations measured against the conventional representation.

- Introduce the new data-sets based on the nSVHC list and investigate the possible influence of the different characteristics within these toxicities in comparison to the new representations.

- Investigate the use of the nSVHC datasets against a baseline to compare the performance of the GCNN's on these chemo-informatic data-sets.

1.2 Contribution

This thesis is positioned in two fields. On one side the emerging field of graph convolutional networks and the other side is the field of chemo-informatics in relation to toxicity prediction. In the field of graph convolutional networks, the benchmark detests are usually a set of chemo-informatic, social data-sets, and citation networks. The chemo-informatics datasets are found to be not well documented by our own experience. To adjust the data-sets to the new representations the molecule has to be known, for some of the benchmark data-sets, this was not possible as the molecules could not be traced back. Thus, one contribution will be the introduction of new data-sets with re-traceable and known molecules for the best explainable possibilities.

Furthermore, the graph convolutional networks field uses the same conventional representation for molecules with sometimes extra information added to the atoms as a feature vector. We would like to introduce and research the influence of adding chemical knowledge to the graph representation, as a feature vector or as part of the spatial structure.

2 Related work

First, we will explore chemical-based modeling and several contributions to the chemo-informatics field. Then, artificial intelligence based applications are researched followed, by a more in-depth analysis of GCNN's and the different approaches and (sub) categories in these GCNN's.

2.1 Chemical based modeling

The field of chemo-informatics has a widespread growing selection of applications, one of these applications is for example a screening tool to find nSVHC chemicals [11]. To model the toxicity of nSVHC chemicals, the similarity between chemicals can be compared with a combination of similarity coefficients and chemical fingerprints [11]. Chemical fingerprints are binary bit-strings formed with chemical-based knowledge to describe the structural information of a molecule [12], these fingerprints can then be used to compare the structural similarity of the molecules and model the structure-activity relationship. Several options exist in these fingerprints describing the 2D or 3D information with dictionary-based, path-based, circular-based, and pharmacophore-based fingerprints [11][13], these fingerprints can describe the molecular structure and also particular chemical properties as for example binding sites or hydrophobic properties [14]. The nSVHC similarity model can then be used as a screening tool for new chemicals[11], which shows the development that can be obtained with modeling the structure-activity relationship. The fingerprints can also have a downside, modeling the deemed important properties or structural components to 166 bits or 1024 bits representation can cause information loss. This information loss was accepted as the fingerprints resulted in computational efficiency and homogeneous representations.

Modeling relationships especially lends itself to the use of machine learning algorithms. The use of shallow and deep machine learning algorithms was therefore introduced as a solution to model the structure-activity relationship[15]. The combination of representation and the chosen algorithm is a crucial and greatly explored choice. Shallow algorithms can already acquire state of the art results but feature engineering is needed for the representations, while deep learning algorithms model more complex relationships and usually require fewer feature engineering[15]. The use of simple features and shallow algorithms for toxicity prediction already proves itself, Karim et al [16] describe their work as a "step toward model simplicity and less compute intensiveness while still maintaining similar or higher accuracy to the DNN (deep neural network) models for moderate size toxicity data sets."

Although shallow modeling with simple features can reach the state of the art results, the progression in deep learning methods still moves forward. For example, different representations options for molecules are explored to apply to molecular modeling; representing the molecular structure as a graph-based image, a graph with nodes and edges, or with a simple SMILES format [17]. With SMILES [18] the molecules are represented in a string format using ASCII sym-

bols, all the atoms of a molecule are represented with their respective letter(s) and molecular bonds, aromatic rings and other properties are represented with chosen symbols. This format allows for the conversion back and forth between a string representation into a 2- or 3d- model of the molecule. With constant advancements in the different fields of artificial intelligence, these advancements cross over in other fields very often. In combination with the representations emerging, the progress of convolutional neural networks or recurrent neural networks in particular deep learning applications gives new opportunities to implement them elsewhere [19]. The created Chemception algorithm, based on the google inception algorithm, uses rendered 2D images of the molecules to match the state of the art in several applications, in predicting toxicity, activity, and solubility properties [20]. Whereas the Deepscreen algorithm make use of 2d images of molecules to predict drug interaction mechanisms [21]. Also, with the use of a Long Short Term Memory model (LSTM) and the SMILES representation of a molecule, different toxicity endpoints can be classified by feeding the recurrent network a character at a time[22].

These advancements made with the images or SMILES representations [21] [22] [20] are not the only representation options emerging in the chemo-informatics field, other advancements are also made in the field of Grap Convolutional Neural Network (GCNN) [23] and Graph Kernels (GK) [24][25]. Here, the molecules are represented as a graph where the atoms are nodes and the edges are bonds. The graph kernels are used to project the graphs to another dimensional plane and in this plane, methods as support vector machines are used to classify the data. This research focuses on the GCNN's, as the new representations are an attempt to solve outstanding questions in this field. Therefore, the next section explores these networks in more depth.

2.2 Graph convolutional neural networks

To illustrate the difference between graph convolutions and image convolutions filters, the different neighborhood filtering is shown in figure 1. The red marked nodes are the center nodes, the gray edges are connected to the neighboring nodes and the larger blue circle marks the neighborhood. The difference with image convolutions is that with graphs, the amount of neighbors is variable and the set of neighbors is unordered which is fitting for molecules. Within these filters, the GCNN's use the spatial or spectral information of the graphs [23] together with the convolutional architecture to create unified representations of the graphs. Because of these unified representations the networks can handle different sized graphs which is required for particular data-sets of molecules and our use case. Spectral GCNN's are based on the use of the adjacency matrix of a graph and their respective eigenvalues while spatial GCNN's are based on aggregating neighboring feature information. In the latest years, the spatial GCNN's have gained the most attention, these models are preferred over spectral models because of the efficiency, generality, and flexibility issues of spectral networks[23]. Therefore, the spatial models are preferred for this research and will be further explored.

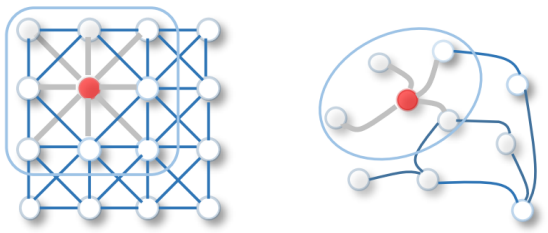

Figure 1: On the left the convolution operation of an image, on the right the convolution of a graph. Image is taken from other source [23].

Dwivedi et al. [26] have implemented a framework to test and compare different spatial graph convolutional neural networks, in which these networks are divided into two subcategories. Both categories work with graphs that consist of nodes and edges between the nodes, both also work with the spatial structure defined by the bonds. The division is made on the anisotropic or isotropic character of the algorithms, the isotropic networks consider each neighbor and their edge directions equal. The anisotropic networks are defined by an attention or gating mechanism: meaning that not every edge direction and thus neighbor is treated equally. Networks in this group use for example gaussian mixture models, gating mechanisms, or sparse attention mechanisms to model the edge weights. An addition to these mechanisms is to use the corresponding edge label eq the bond type. GatedGCN for example, which uses gating mechanisms, can have the edge weights initialized with the corresponding edge representations [26] and thus take into account the molecular bond types. In the benchmark of Dwivedi et al. [26] the anisotropic networks are described as consistent performers which improve over isotropic GNNs in 5 out of 7 benchmarked datasets.

In conclusion, these spatial networks can work with different graph representations and the anisotropic networks have better properties for molecules because of the working with the edges and possibly bond based labels. But in contrast to the chemical fingerprints, the different chemical characteristics such as hydrophobicity and aromatic rings are not exploited in graph representations. Therefore, we propose different representations with additions based on chemical fingerprints patterns and chemo-informatic expert knowledge.

3 Methodology

First, a general mapping of graph convolutional networks is explained to grasp the mapping of the graphs to a unified representation. Then, the data-sets and the main characteristics of the use case data-sets are explained to later on reason about the case-specific difference. Also, altered representations in another application are explained followed by the introduction of chemical additions to the graph representations, Then, we will introduce the representations and the transformation needed from the conventional representation.

3.1 Graph data classification

To research the difference between the different representations we will use one algorithm of each aforementioned category; meaning an isotropic and anisotropic model. For the anisotropic model, the GatedGCN model with the addition that the edge weights are initialized with the corresponding edge label is chosen. Because the edge initialization we can compare against the isotropic networks which work without edge labels. For the isotropic model, GIN, which stands for graph isomorphism network, is chosen. Both will be compared with a baseline Multi Layer Perceptron (MLP) for the comparison with a simple network and if the GCNN's can improve over a baseline on all use case data-sets. Although two algorithms are picked out, the graph representations are considered uniform and useable by other GCNN's or graph applications. In figure 2, a high-level illustration of graph neural networks can be seen. The representations fit in this pipeline and can be used by spatial, spectral, isotropic, and/or anisotropic models and outside this pipeline, by graph kernels or other graph-based applications.

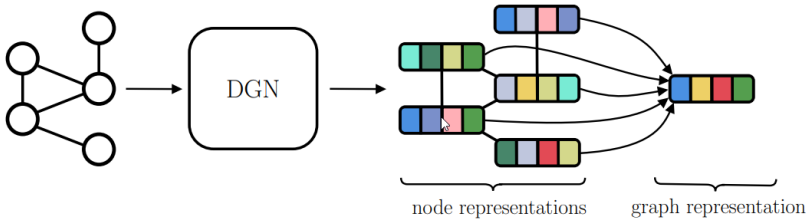

Figure 2: High level illustration of graph neural networks. The image is from another source [27]

All networks will work with a graph as input, we will discuss the generic mapping of the graph data from one layer to the next layer for a graph convolutional network as described by Dwivedi et al. [26]. Also, the adjustments for the MLP to create a graph agnostic baseline are formalized. And last, the mapping from the graph to a graph classification layer together with the mapping of two combined networks to a classification outcome is put into a formula.

The graphs exist of node features α_i for each node i and edge features $\beta_i j$ for each edge connecting node i and node j. The input features are embedded to d-dimensional hidden features $h_i^{l=0}$ and $e_i^{l=0}$ via a simple linear projection before passing them to a graph neural network. Each graph neural network layer computes d-dimensional representations through neighborhood aggregation, where each graph node gathers features from its neighbors to represent local graph structure [26]. The most generic version of a feature vector h_i^{l+1} can be seen in equation 1. The $j- > i$ denotes the set of neighboring nodes j connected to node i, which can also be denoted as $\{j \in \mathcal{N}_i\}$. The choice of this mapping is what defines a network.

$$h^{(l+1)} = f(h_i^l, h_j^l : j- > i) \tag{1}$$

The baseline from Dwivedi et al. [26] is used; this applies an MLP on each node feature vector independently of other nodes. The update is seen in equation 2, the node feature vector of a next layer is only dependent on its feature vector.

$$h_i^{l+1} = ReLU(U^l h_i^l) \tag{2}$$

The final layer of each network is for prediction, the outputs will be fed to a loss function to train the network parameters. To perform graph classification, first, a d-dimensional graph-level vector representation y_g is build by averaging over all node features in the final GNN layer as seen in equation 3. This causes a graph size agnostic representation which can then be used for classification. These graph features are then passed to an MLP, this outputs un-normalized scores y_{pred} for each class, the formula can be seen in equation 4. The predicted class is then the maximal value of either the toxic or not toxic class. $P \in \mathcal{R}^{dxC}$, $Q \in \mathcal{R}^{dxd}$, and C is the number of classes.

$$y_g = \frac{1}{V} \sum_{i=0}^{V} h_i^l \tag{3}$$

$$y_{pred} = PReLU(Qy_g) \tag{4}$$

To compare different complementary properties of different trained networks, we decided to also combine algorithms. For this, we have chosen to first normalize the y_{pred} of both respective algorithms with a sigmoid function. After that, the average of the sum of y_{pred}^a of an algorithm a and y_{pred}^b of an algorithm b is taken as seen in equation 5. Then, the predicted class becomes also the maximal value of either the toxic or not toxic class.

$$y_{pred} = (sigmoid(y_{pred}^a) + sigmoid(y_{pred}^b))/2 \tag{5}$$

3.2 Chemical background

To dive into the chemical knowledge applicable to this data, we first investigate the chemical background of the data. After that particular relations of chemical characteristics with toxicity and thus the molecular data is discussed. Following, in the section chemical representations, our proposed representations.

The chemicals within the use cases are deemed substances of very high concern and collected from the nSVHC list, which is the dutch SVHC chemical list. Different from the European Chemical Agency (ECHA) SVHC list, the nSVHC[28] list also contains chemicals from other lists as OSPAR and KRW. In these other lists, slightly different thresholds or other criteria for the nSVHC chemicals are set. This causes that all the SVHC chemicals of the ECHA are present together with other SVHC deemed chemicals by different lists. The OSPAR list, for example, is a list that highlights chemicals that are considered to have a risk for the marine environment and don't get enough attention under the Registration, Evaluation, Authorization, and restriction of Chemicals (REACH) initiative of the ECHA, where SVHC chemicals fall under. Thus, the lists where the classifications originated from can cause differences in criteria.

The use case data consists of different chemicals where molecular structures are used. The chemicals are labeled with PBT/vPvB, CMR, or nontoxic labels. The PBT label stands for persistent, bio-accumulative, and toxic[28], meaning that the molecules are persistent and therefore not degradable, bio-accumulative which means that molecules are accumulating inside living creatures (plants or animals) and toxic. Also, together with PBT, the vPvB chemicals are taken into this dataset which stands for very persistent and very bio-accumulative; so a category higher persistent and accumulating than PBT. These molecules are under concern because they persist for a long time in the environment, can also accumulate inside living creatures and are deemed toxic. A chemical is classified as PBT if the chemical is persistent, bio-accumulative, and toxic so if all of the three endpoints are true, or vPvB if both very persistent and very bio-accumulative are true. If one of the one endpoints is not true, the chemical is not classified in PBT/vPvB and only deemed persistent or bio-accumulative. Because of this all or nothing principle of PBT it is possible to classify chemicals that are labeled as not persistent or not bioaccumulative as not PBT although the other endpoint is not tested at this point.

The CMR stands for carcinogenic, mutagenic, or reprotoxic [28]. Meaning toxicity concerning cancer, alterations in the genetic material, or problems with reproduction. A chemical is classified as CMR if one of the CMR endpoints is true, so if it is carcinogenic, mutagenic or reprotoxic but it can also be classified to more than one endpoint. A classification not carcinogenic, not mutagenic or not reprotoxic can mean that chemical is still considered one of these labels, but not in the category required to be deemed CMR. There are categories for the labels, which are 1A, 1B, and 2: in category 1A or 1B the evidence is based on data from humans or enough animals, while in category 2 the evidence is based on limited data of animals. A 1A or 1B label is considered to be the criteria for classification as CMR [29].

These two data-sets originate from the nSVHC datasets and aforementioned classifications, both deemed small data-sets with an unbalanced character, the non-CMR or non-PBT molecules are the larger part of the datasets. An approach to exploit a small data-set better is the use of expert knowledge to inject the information provided in the data representation which then can be used for better classification [30] [31] [32]. Some data-sets also contain single points of knowledge (SPOK), as described in [11]. These SPOK's are molecules which are not comparable to any other chemicals in this group, the addition of extra knowledge can also adhere to these SPOK's. The extra information can contain comparable characteristics to other chemicals for the classification of these SPOK's. Chemical knowledge that can contribute to the classification of graph-based data-sets is for example other structural information. Exploiting structural data in the structure-activity relationship doesn't have to be based on only the atoms and their connections in terms of bonds. The bond type and particular characteristics of the molecular substructures can also be viewed as structural data. With a convolutional architecture on 2D images[33], structural components are highlighted with the use of six different filters, the different filters can be seen in figure 3. Highlighting these features and feeding all this information to a convolutional network has let the algorithm exploit more structural information and therefore the available chemical knowledge.

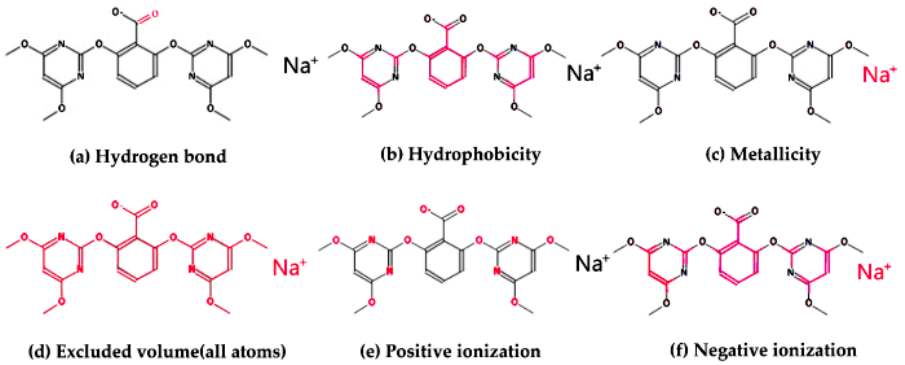

Figure 3: Molecular descriptors used in [33] and their highlights in images

The hydrophobicity used in these filters is also deemed an important characteristic in general toxicity prediction [34][35], this possible structural information could, therefore, be regarded as extra expert information regarding toxicity. Hydrogen bonds are also an interesting part, these can be split up in are the hydrogen bond acceptors and donors, which can cause a hydrophilic character. Hydrophobicity can be a characterization of bioaccumulation which is the B in PBT, the opposite of this (hydrophilic) could indicate which are not deemed so. Next to the role of these characteristics, aromatic compounds can play a role in persistent traits as these are seen as "among the most prevalent and persistent pollutants in the environment" [36]. Aromatic rings can thus correlate

with the persistent classification of the P class in PBT. Besides, the aromatic rings of a molecule have a relation with the reactivity, which has a relation with carcinogenicity. Further, the biodegradability properties of a chemical can also indicate properties regarding the metabolism, which can then have an impact on the level of bio-accumulation. A molecule that is not processed in the metabolism can accumulate in the organism. The metabolism can also indicate reactive properties of CMR chemicals, where for some chemicals the toxic properties become active after metabolic processes. In line with this, sites of metabolism (SOM) in molecules, are found to have a relation with toxicity. The SOMs of the protein glutathione for example[37], the most abundant peptide in the human body, is shown to correlate with toxic characteristics. The SOMs for the family of enzymes known as cytochromes P450 (CYPs) are deemed "the most important class of drug metabolizing enzymes" [38] and "CYPs are also the cause of the majority of drugdrug interactions and metabolism-dependent toxicity issue"[39] showing the importance of SOMs regarding toxicity. Next to these substructures, the bond strengths, and thus the bonds labels can also indicate reactivity and biodegradability. All these substructures and labels can be regarded as structural information based on chemical expert knowledge.

3.3 Proposed graph representations

To exploit this expert knowledge we introduce three new graph representations for molecules. To research the influence of these representations in contrast with the conventional configuration, we will also use the conventional representation with atoms as nodes and bonds as edges. We introduce all representations, an explanation of what made us use these representations, and the high-level overview of how the representations are created.

The first representation we propose, the **bonds addition representation** (bonds), is where the chemical bonds are added as extra nodes. This representation follows from the isotropic networks mentioned where the bonds labels are not taken into the network setup. Further, from the relation of the bond types, thus the bond labels, with metabolism and reactivity. The new representation is a graph, where nodes are atoms and also bonds. To label the edges we used the bond label. For the isotropic networks this edge label makes no difference but for the anisotropic network must be noted this can be regarded as redundant information together with the bond nodes. To form this representation we processed the molecules into a graph representation and removed the edges. For every edge that we removed, we converted the edge feature (bond label) β_{ij} into a new node feature, which has the edge feature converted into a newly created node feature α_k different from the edge feature. Then, we created new edges from node i to node k and node j to node k, the transformation of the old to the new situation can be seen in figure 4. Nodes will hereafter have an atom or bond label and edges have a bond label.

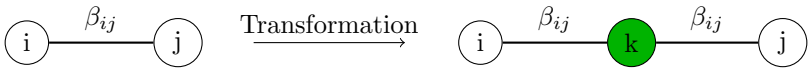

Figure 4: Bonds addition representation. The transformation for the chemical addition of bonds as nodes, the green node is added and the new edges are added with the same label as the old edge.

The second representation, **the hydro/aromatic addition representation** (hydro/arom), will have particular substructures added: the hydrophobicity, hydrogen bond acceptors, hydrogen bond donors, and aromatic rings will be highlighted as spatial information. These substructures can be possible indicators and have a known relation with toxicity, therefore for both the isotropic and anisotropic networks, these are deemed as exploitable expert knowledge. To add these substructures to the graph representations, first, the substructures are located in the molecules. These substructures are located with SMILES Arbitrary Target Specification (SMARTS), an extension of SMILES, which is a substructure searcher that can identify a particular subpattern in a molecular graph. For the hydrophobicity, the SMARTS patterns from RDKit [40] are used, a python based molecular modeling library. For the hydrogen-bond acceptors, hydrogen-bond donors, and aromatic rings, the SMARTS from a chemical fingerprint(morgan) [41] are used. These SMARTS can be found in the appendix. For every atom in the substructure, a new node is connected with a new edge. The feature vector of the node will have the label of the new substructure α_{sub} and the new edge feature vector will also have the label of the new substructure, to separate the edges from the bonds we decided to keep this feature vectors different. If for example a graph has nodes i,j,k, and edge ij, ik and the SMARTS find a substructure of interest, present as nodes j and k, then we add nodes l and m together with edges jl and km to highlight the structural presence of this substructure. The transformation is visualized in figure 5. Nodes will hereafter have a substructure or atom label and edges a substructure or bond label.

Figure 5: Hydro/aromatic additions representation. The transformation show the highlighted yellow nodes as found substructure, added to these yellow nodes is the substructure labeled nodes to highlight this substructure.

The third representation, **the SOM addition representation** (SOM), will add information about Sites of Metabolism of the CYPs to the feature vectors of the nodes. For the CYPs, 71 SOM are formalized in Olsen et al.[38] as SMART

patterns. Because of the amount of 71 patterns and the more general character regarding the reactivity of the whole molecule and the role in the metabolism, the choice here is to add the SOM to the feature vector of every atom. The 71 SMARTS patterns were searched on every molecular graph representation, if one was present, this presence was encoded into a boolean vector. After the search of the entire molecule, this boolean vector was concatenated with every atom feature vector creating a multi hot encoded feature vector for every atom containing information about the atom label and the possible presence of SOM for CYPs in the molecule.

In figure 6 the different representations are visualized. The most left representation is the conventional representation. The first proposed representation, in the middle, has the bonds added between the nodes as extra nodes. The second representation, on the right, has hydro/aromatic additions, the green nodes are the extra additions. The third representation isn't visualized since it only contains an altered node feature vector.

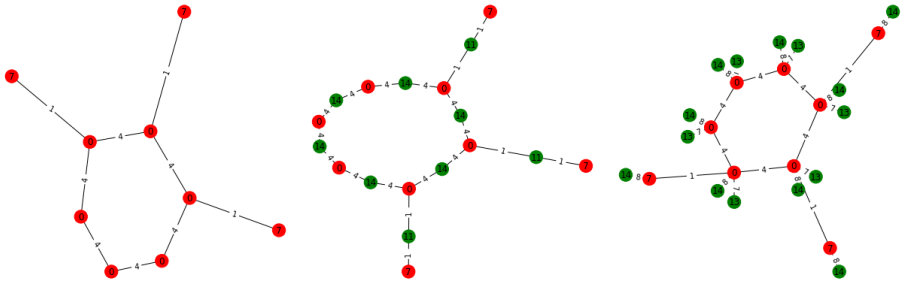

Figure 6: The three different representations in graph form, from left to right it is the conventional representation, the hydro/arom representation and the bonds representation. Added information is visualized with green nodes.

4 Experimental setup

4.1 nSVHC data-sets

We have taken the data-sets from Wassenaar et al. [11], the chemicals are collected from the dutch substances of very high concern (nSVHC) list which overlaps but also has a broader selection of chemicals than the REACH data. The data was extracted on 01-03-2018 by Wassenaar et al. [11] and contains known CMR or PBT/vPvB toxicants, further a set of chemicals that are considered inherently safe and known not CMR or PBT/vPvB chemicals is added. This non-list is considered as suited for the purpose as the chemicals have been tested experimentally and are negative for these endpoints according to the SVHC criteria. The PBT/vPvB datasets will be named the PBT dataset from now, the CMR dataset contains the CMR chemicals and we have created a PBT+ dataset which is an extended version of the PBT dataset containing chemicals that are either classified as not P or not B and therefore can not be classified as PBT.

The data consists of molecules written in SMILES code which are converted to a graph structure with the use of RDKIT-python[41]. Every atom is a node and every bond is an edge in the normal conventional representations, in the new representation the bonds and other features are also represented as nodes. In table 1 is described how many toxic and non-toxic molecules there are and for each, what the mean number of edges and nodes are and the maximal and minimal number of nodes in a molecule. In table 2 and table 3 these statistics are shown for the bonds and hydro/arom representation, to depict the difference with the altered representations and the conventional representation. In table 4 is described for the nontoxic and the toxic molecules which of the chemical features are present in which percentage of the total molecules, this can mean that the molecule has 1 or more of the specific features. In figure 9 and figure 10 in the appendix can be seen how these features are distributed over the different molecules, this is visualized in different pie charts for each possible feature and toxicity combination.

	CMR		PBT		PBT+	
class	0	1	0	1	0	1
molecules	381	271	381	113	859	113
mean # edges	21.4	16.5	21.4	20.9	15.9	20.9
mean # nodes	20.6	15.8	20.6	19.5	15.7	19.5
max # nodes	65	73	73	43	73	43
min # nodes	2	2	2	9	2	9

Table 1: Table containing the graph characteristics for each data-set for the conventional representation

	CMR		PBT		PBT+	
class	0	1	0	1	0	1
molecules	381	271	381	113	859	113
mean # edges	42.8	32.9	42.8	41.7	31.8	41.7
mean # nodes	42.0	32.3	42.0	40.3	31.6	40.3
max # nodes	151	136	151	85	151	85
min # nodes	3	3	3	18	3	18

Table 2: Table containing the graph characteristics for each data-set for the bond addition representation

	CMR		PBT		PBT+	
class	0	1	0	1	0	1
molecules	381	271	381	113	859	113
mean # edges	48.6	27.9	48.6	49.1	36.4	49.1
mean # nodes	47.8	27.5	47.8	47.8	36.2	47.8
max # nodes	168	167	168	85	168	85
min # nodes	3	2	3	20	3	20

Table 3: Table containing the graph characteristics for each data-set for the hydro/aromatic addition representation

	CMR		PBT		PBT+	
class	0	1	0	1	0	1
HB acceptor	96.85%	84.13%	96.85%	50.44%	91.04%	50.44%
HB donor	60.37%	44.65%	60.37%	18.58%	54.60%	18.58%
Aromatic bond	72.97%	57.93%	72.97%	65.49%	60.30%	65.49%
Hydrophobe	98.95%	95.57%	98.95%	100.00%	99.30%	100.00%

Table 4: Table containing the graph characteristics for each data-set showing in which percentage of the molecules a particular substructure is present.

4.2 Representations

To research the effect of the different representations on the case study, each representation will be fitted to two different algorithms and the MLP baseline resulting in three different accuracies per algorithm per dataset. These accuracies will be the mean balanced accuracy of 10-fold cross-validation performed on the dataset. The difference with the baseline representation, which is the conventional representation, will be examined to see the performance impact of the different representations. Also, the difference with the baseline MLP will

be examined to see if the graph convolutional networks exploit their architecture. After that, the different representation/algorithm combinations will be examined on their respective precision and recall to also compare the difference between false positive's and false negative's. Also, some molecules that are miss-classified by the combinations will be examined case for case together with a chemical expert. This will give insight into not only the performance difference but also the possible different outcomes the combinations and representations induce.

In short, the next four results are produced:

- For 2 data-sets, 3 algorithms and 4 different representations an accuracy for every possible combination; resulting in 24 different mean 10-fold balanced accuracies
- For those 24 different outcomes, the recall and precision are taken into account to see the trade-off between false positive's and false negative's.
- The amount of miss-classified molecules categorized by the amount that they are present in different settings. So, in how many of the four different representation-algorithms combinations the molecules are miss-classified.

To further investigate the different representations and the possible different aspects that are learned in the training process the different trained models per representation are combined to a complementary model where the score per graph is the average of the two models combined. To investigate further if the models learn different aspects, the resulting accuracies of the different combined networks are analyzed to see differences with the seperate network accuracies. Also, the model's outcomes are further analyzed with their precision and recall scores to see the impact of the combination in more depth.

In short, the next two results are produced:

- The 4 representation models per data-set and algorithm combination are combined into a new model, for every combination the 10 fold mean balanced accuracy is calculated. The combination is always the conventional and then with 1, 2 or 3 other representation algorithms.
- For all complementary models, the precision and recall are investigated.

All the representations are converted with python and the use of jupyter notebook. The framework of Dwivedi et al [26] can load datasets where each graph consists of a vector with the node type, an adjacency matrix where the edge type is the matrix component, the total numbers of atoms, and the toxic label of the graph. All the representations are converted with the help of RDKIT[41], the SMILES[18] are converted into molecules and then converted into the needed dataset. For 2 representations the SMARTS, which is substructure search language for SMILES, is used to find and locate the chemical components added to the representation. For representation 2 the SMARTS for the hydrogen bond acceptor, hydrogen bond donor, and the hydrophobicity

is added, the SMARTS can be found in the appendix. The smarts for the hydrophobicity is used from the RDKIT [41]. For representation 4 the 71 SMARTS for the SOMs of a family of enzymes CYPs are added to the feature vector of the atom if the atom is part of the binding site, the SMARTS can be found in Olsen et al. [38].

4.3 Evaluation

The approach described by Dwivedi et al [26] and the need expressed for a reproducible and bench-marked machine learning approach motivates us to use stratified 10-fold cross-validation. Also, the seed which defines which randomness is used is set to 79, meaning that every performed 10 fold cross-validation is one the same train, validation, and test split so that the results are comparable. The framework used in Dwivedi et al [26] will be implemented and followed, the engineered representations can be implemented and used in this framework. This allows verifying and securing the reasoning about significant differences. The work is compared with 2 GCNN's, namely GIN and GatedGCN which are graph isomorphism network and gated graph convolutional network[26].

For every representation and dataset combination possible there is a hyperparameter search done, the max amount of epochs is set to 1000 as used in Dwivedi et al.[26]. The different hyper-parameters settings can be found in the appendix. The experiments are run on a bio-informatics grid of the RIVM which is an IBM based grid. On every setup, batch normalization and graph normalization following the benchmark is applied [26]. To evaluate the network, balanced accuracy is used. This metrics is the sensitivity and the specificity together, divided by 2. This metric is used because of the unbalanced character of the dataset. Further, the recall and precision per dataset are compared to reason about the balanced accuracy.

4.4 Models

For every model, a hyperparameter search is done on the possible hyperparameters. The models are retrieved from the Github of Dwivedi et al [42]. Different hyperparameters were searched for different models. For the MLP, the in dimension, the hidden dimension, the out dimension, the layers, and the readout aggregation function were searched for the best setting. For the GIN model the in dimension, the hidden dimension, the layers, number of MLP layers for GIN, the neighborhood aggregation function, and the readout aggregation function were searched for best settings. For the GatedGCN the in dimension, the hidden dimension, the out dimension, the layers, and the readout aggregation function were searched for the best settings. For the MLP, GIN and GatedGCN the best hyperparameter settings are shown in table 12,13,14 of the appendix.

5 Experiments

For the experiments, first, we will discuss the mean balanced accuracy per data-set. The PBT data-set is found to be not sufficient enough in terms of data and complexity to be further discussed. For the CMR and PBT+ data-sets, particular additions are found to be of added value to the performance. To investigate and compare the added value, the recall and precision of the CMR and PBT+ data-sets are discussed, which shows the added value of the addition of the bonds for the PBT+ data-set. For the CMR data-set, the addition of the SOMs and the hydro/arom shows improvement over the conventional representation. Further, we investigate the miss classifications for patterns and with a separate case study to find that the borderline cases are strongly represented. To research the complementary effect of different additions, we combine the addition representations with the conventional in different set-ups and compare these in contrast to the conventional representation based networks. The complementary networks show a steady performance increase over the conventional set-up but these are not able to improve over the best separate representation based networks, both in recall, precision, or in balanced accuracy.

5.1 Chemical additions

5.1.1 Balanced accuracy

For the CMR dataset, the balanced accuracy's gathered from the different setups are depicted in table 5. The first thing that pops out is the influence of the representations on the MLP baseline. The bonds additions and the hydro/aromatic additions show an increase for the conventional representations and the largest increase is seen in the SOM addition representations. Following from the metabolism and toxicity correlations with reactivity sites, these indicators have a strong positive effect on the prediction of the CMR data-set for a simple MLP. Further, the conventional representation shows a significant difference of almost 19 percent with the GCNN's conventional accuracies, which indicates that the GCNN architectures can exploit the spatial structure of the graphs for these representations. The difference is decreased when looking at the difference between the SOM addition for the MLP and the two GCNN's, nonetheless, the GIN performs better. The difference for the other two representations shows that GCNNs can also exploit their architecture on these representations, having a significant difference with their baseline accuracies. From the combination of the isotropic GIN algorithm and the CMR dataset can be seen that the best performing combination is with the SOM additions representation, scoring an accuracy of 82.3 percent. Further, in contrast to the MLP baseline, the addition of the bonds shows a decrease regarding the conventional representation for both the isotropic and anisotropic networks. Although the addition of the bonds seems to have a positive influence on a simple baseline, the addition influences the GCNN's negatively, especially in the GatedGCN. Thus, the redundant aspect of adding the bonds to the anisotropic is supported with a

declining accuracy, a more interesting aspect is that the addition of bond types has a negative influence on the isotropic network. Inversely, the hydro/arom additions show an increase of accuracy for both the networks on the conventional representations, showing a positive influence of these additions. Not in line with these findings is the influence of the SOMs addition, whereas the GIN network improves over the conventional representation, the GatedGCN decreases over the conventional representation.

	CMR MLP	GIN	Gated GCN
Conventional	61.2 ± 5.1	79.9 ± 4.1	81.3 ± 5.5
Bonds	67.6 ± 4.7	78.3 ± 5.1	76.8 ± 5.6
Hydro/Arom	66.7 ± 6.7	81.5 ± 3.9	**83.2 ± 3.7**
SOM	**77.9 ± 6.0**	**82.3 ± 3.8**	80.8 ± 3.3

Table 5: Mean balanced accuracy in percentage of a 10 fold cross validation for each representation and network combination. The bold marked are the best in their column and the red marked are the best for the whole dataset. The highlights are the best performing score of 83.2 for the hydro/arom presentation with GatedGCN and the best scoring of 82.3 for the SOM representation with GIN. Further, the table shows an improvement in scores for the new representations with the baseline. Also, a significant difference of the baseline with GCNN scores showing the added value of the architecture.

Switching from the CMR data-set to the PBT data-set, we will describe the differences in patterns regarding the CMR accuracy's next to the outcomes. The table with the accuracies can be found in the appendix, in table 15. The first difference regarding the CMR data-set is the balanced accuracy of the MLP baseline with the GCNN's. The MLP baseline shows for all the 4 representations an approximately equal performance ranging between 88.1 for the conventional representation to 89.8 for the SOM additions representation. This indicates that for this PBT dataset the GCNN architecture doesn't exploit extra information or to a minimal extent. The additions in the representations have no beneficial effect on the baseline MLP. Also, the difference with the GCNN's to the baseline is significantly lower than with the CMR, the scores of the GIN and GatedGCN range from 90.2 to 93.7. Further, the best performing representation for GIN, the representation with bonds additions, has an accuracy of 93.7, where the conventional representation has just a slightly lower accuracy of 93.5. Thus, with the PBT data-set, the GCNN architectures have no additional influence and the additions have no added value for any network.

In contract with the PBT data-set, the PBT+ data-set shows a similar pattern to the CMR data-set for the MLP baseline. The scores of the PBT+ data-set can be seen in 6. The same differences between representations and the same highest-scoring representation come out on in the MLP baseline. The MLP baseline increases his performance with the new additions and the SOM

additions representation is the highest-scoring in the MLP baseline. Further, the difference between an MLP or GCNN and a representation shows the same pattern in an increase in performance as with the CMR data-set. Thus, the addition of the more borderline cases of non P or non B adhere to the complexity of this data-set and therefore the GCNN architectures have an additional value. The best performing combination here is the bonds addition representation, opposite to the CMR data-set which has the lowest performance with this representation. The other additions have no significant higher performance compared with the conventional representation, which indicates a lesser added value than with the CMR data-set. The bonds addition shows a positive correlation with the classification of PBT+ chemicals.

	PBT+ MLP	GIN	Gated GCN
Conventional	77.6 ± 6.4	92.1 ± 4.3	89.5 ± 4.7
Bonds	84.4 ± 4.8	**94.9 ± 3.4**	90.1 ± 5.3
Hydro/arom	83.1 ± 2.9	92.9 ± 6.4	89.4 ± 5.4
SOMs	**85.4 ± 4.4**	91.8 ± 6.4	**90.9 ± 5.3**

Table 6: Mean balanced accuracy after 10 fold cross validation for each representation and network combination. The bold marked are the best in their column and the red marked are the best for the whole dataset. The highlights are the best performing score of 94.9 for the bonds presentation with GIN. Further, the table shows an improvement in scores for the new representations with the baseline. Also, a significant difference of the baseline with GCNN scores showing the added value of the architecture.

The accuracies show a positive impact of particular additions for both data-sets, the difference in for which additions this is, shows the effect of the different toxic characteristics of the data-set. With CMR, the isotropic network with the SOMs addition increases the performance. With both networks, the hydro/aromatic addition increases the performance over the conventional representation. For the PBT+ data-set, the addition of the bonds shows a performance increase in contrast to the other additions. To research the difference in more depth, we will consider the precision and recall to analyze the performance in classifying more nontoxic or toxic chemicals. An important note on this, because these are nSVHC chemicals and a miss classified toxic chemical is worse than a miss classified nontoxic chemical, the recall metric is important. Also, because of the small differences in performance within the PBT data-set, we will focus on the PBT+ and CMR data-set.

5.1.2 Precision & recall

In table 7 the precision and recall for the different algorithms with the CMR data-set show differences not visible from only the accuracy. The MLP shows

an increase in recall for the additions together with a decrease in precision. Although classifying more CMR chemicals correct it also increases the false positives. Switching to the GCNN's, the recall with the SOMs additions for GIN shows the highest performance and also the highest precision, this in line with also having the highest balanced accuracy of 83.2%. In the accuracies, the addition of the bond shows a decrease in performance for both networks, while the precision of the isotropic network is the highest, showing a decrease in the false positives. Both the highest recalls of 74.2% and 78.2% are seen in the highest performing accuracies, in the isotropic with SOMs addition with an accuracy of 82.3% and in the anisotropic network with the hydro/arom addition with an accuracy of 83.2%. The anisotropic network shows the highest recall overall together with the hydro/arom addition, recalling 78.2% of all CMR chemicals. Together with the highest accuracy, this shows the added value of the hydro/arom representation for the CMR data-set in the increase in performance on both points.

		CMR MLP			GIN			Gated GCN		
Conventional	P - R	**82.2**	-	26.5	81.5	-	67.9	83.2	-	73.4
Bonds	P - R	76.0	-	46.8	**85.7**	-	66.0	79.5	-	66.0
Hydro/Arom	P - R	67.5	-	50.9	85.3	-	73.1	83.2	-	**78.2**
SOMs	P - R	79.4	-	**69.0**	**85.7**	-	**74.2**	**85.3**	-	70.8

Table 7: Precision - Recall in percentage per representation and algorithm combination for the three different algorithms for the CMR data-set. The best performing precision and recall per column are depicted in bold, the best performing precision and recall for the whole data-set are shown in red. Together with the best performing accuracy in the column, the SOMs representation shows the best precision and recall scores in the GIN column. Further, the overall best recall score is seen in the hydro/arom representation with the GatedGCN.

For the PBT+, the precision and recall in table 8 for the MLP shows a different trade-off. The recall improves for the bond and the hydro/arom representation but the precision increases slightly. A greater precision increase is reached in the hydro/arom addition together with a slight increase in recall. The GCNN networks show better recall and precision scores, the best performing recall of 94.6% with the addition of the bond can thus recall 94.6% of the PBT/vPvB chemicals. The highest performance in precision is the conventional representation with the anisotropic network, both showing the performance increase with the bonds addition or bonds label initialization.

		PBT+ MLP			GIN			Gated GCN		
Conventional	P - R	57.7	-	61.7	72.9	-	85.7	**97.2**	-	79.3
Bond	P - R	59.8	-	76.0	75.9	-	**94.6**	80.7	-	83.0
Hydro/Arom	P - R	**73.8**	-	69.7	**88.3**	-	86.7	95.2	-	79.4
SOMs	P - R	61.2	-	**77.7**	77.6	-	87.4	79.9	-	**84.9**

Table 8: Precision - Recall per representation and algorithm combination for the three different algorithms for the PBT+ data-set. The best performing precision and recall per column are depicted in bold, the best performing precision and recall for the whole data-set are shown in red. The best performing recall is seen in the bonds representation, the same as with the best performing accuracy for this dataset. Further, the best precision score is seen with the conventional representation in the GatedGCN, although also scoring the lowest recall score for the GCNNs.

The precision-recall scores also show a positive impact of particular additions for both data-sets. In PBT+ the best recall score and the best accuracy is with the isotropic network with the addition of the bonds, the best precision is with the anisotropic network with the bonds initialization showing the added value of the bond labels. In the CMR dataset, the best recall and accuracy is with the hydro/arom additions. The best precision is with the bonds and SOMs addition, whereas the best recall for the isotropic is also with the SOMs addition. For both datasets, this shows the improvements that can be made with the additions in representations.

To research the particular miss classified chemicals of both data-sets, we will dive into the miss classified chemicals by all network and representation combinations, investigating which chemicals are miss classified and by how many representations. This shows if some chemicals are persistently miss classified and gives a reason to investigate these case by case.

5.1.3 Miss classifications

In table 9 the number of miss classifications per data-set for the network and representations combinations are shown. The eleven chemicals of the PBT+ data-set that are persistently miss classified by the GatedGCN were analyzed to find possible causes, all these chemicals were falsely labeled non-PBT. For this set, one chemical is classified as PBT/vPvB by the SVHC list of the ECHA whilst the other 10 chemicals originated from the OSPAR list. The OSPAR list can be seen as a list with less strict criteria and therefore these cases can be seen as borderline cases which can be may not be considered PBT/vPvB by other criteria. Thus, these can also be classified correctly under different criteria. The 13 miss classified PBT+ chemicals of the GIN network contain a set of 6 falsely labeled PBT chemicals and a set of 7 falsely labeled non-PBT chemicals. The PBT chemicals contain mostly the same as in the GatedGCN set, chemicals

originating from the OSPAR list. Further, a chemical containing the atom tin (Sn) is the only chemical containing this atom and can be considered a SPOK. The non-PBT chemicals have another pattern, this set of chemicals contain mostly chemicals that are either labeled non P or non B, and thus can still be one of the others but not PBT, these can also be considered borderline cases and hard to classify. Further, because the CMR miss classifications contain a large set of chemicals, we analyzed the ones with the highest difference in prediction score and the lowest difference in prediction score from only GatedGCN to investigate possible patterns. The lowest difference shows no mention-worthy patterns, only in the highest-scoring differences patterns arise. There are 2 labeled CMR whilst the true label is non-CMR, but these chemicals have a reprotoxic category 2 label and a carcinogenic category 2 label, these chemicals do have CMR effects, but just not to a high enough extent to be classified as CMR (cat 1A/B). These cases can also be seen as borderline cases where the model finds characteristics considered to be CMR properties. These borderlines cases in both data-sets show the challenge within the data-sets and also reason to emphasize the performance of the models. The miss classified cases are borderline and the models perform well under these circumstances.

Amount of missclassifications	CMR		PBT+	
	GIN	GatedGCN	GIN	GatedGCN
Single	69	77	51	35
Double	42	55	25	20
Triple	39	26	9	7
Quadruple	51	52	13	11

Table 9: The amount of miss classifications of a the false labeled chemicals, single if one representation - algorithm labels the chemical false to quadruple if all combinations label the chemical false. The quadruple, or four times misclassified molecules, show a similar value within the datasets which can indicate a similar set of molecules that is miss classified.

Further, to dive into a case study, we compared the predicted scores of the best CMR accuracy in contrast to the conventional representation. The best CMR is the GatedGCN network in combination with the hydro/aromatic additions. Particular chemicals showed an improvement in scores in contrast to the scores with the conventional representation. The chemical in figure 7 where the difference between the two representations is shown, was converted from a false labeled borderline prediction score in the conventional representation to a certain prediction of the CMR label in the hydro/aromatic representation. For the chemical in figure 8 this was reversed, where the borderline false labeled score is done by the hydro/aromatic representation in combination with GatedGCN, the conventional representation has a certain prediction for the true labeled non CMR class. The green nodes are different added substructures which are labeled with an 13,14,15,16 in the nodes. These numbers are the substructural labels of hydrogen bond acceptors, donors, aromaticity, and hydrophobicity. Both chem-

icals in both figures contain the labels 13 and 16 for hydrogen bond acceptors and hydrophobicity, these substructures parts indicate parts of the representations the network trained could hit on and where the miss classification comes from.

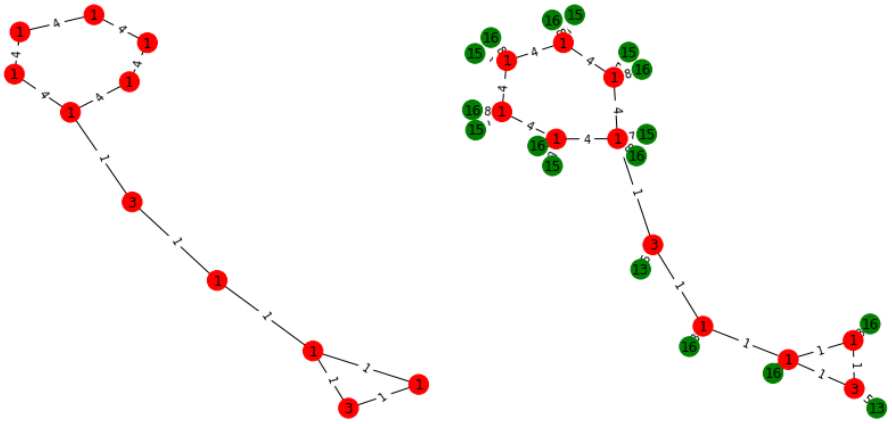

Figure 7: On the left, the conventional representation of a CMR chemical, on the right the hydro/aromatic representation of the same chemical. This chemical went from a borderline case in the conventional representation to a confident classification in the hydro/aromatic representation.

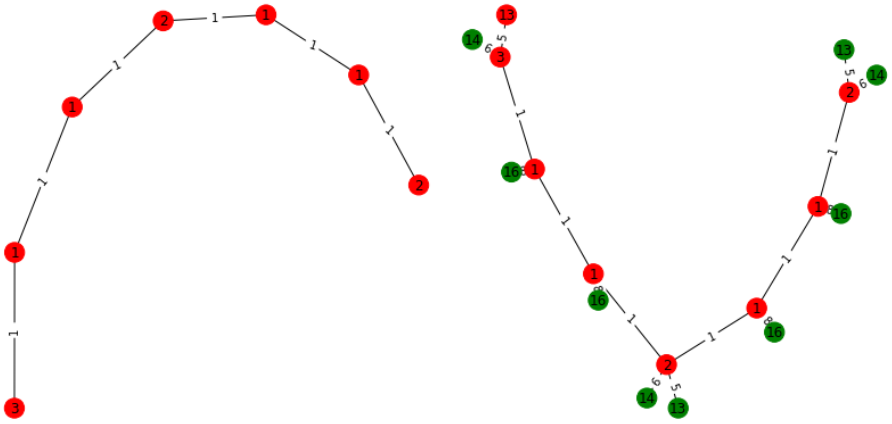

Figure 8: On the left, the conventional representation of a CMR chemical, on the right the hydro/aromatic representation of the same chemical. This chemical went from a borderline case in the hydro/arp, representation to a confident classification in the conventional representation.

5.2 Combining chemical additions

Following the number of miss classifications, the different representations also have a large list of chemicals that are only miss classified by one or two representation networks. If the networks learn different patterns, added together they could enhance each other's performance. To research the complementary effect in contrast to the conventional representation network, we will compare the effect of the conventional representation combined with other representations. The first pattern that arises is that the best performing combination for almost all data-sets is everything combined. Only in the GatedGCN with the hydro/arom combination the accuracy is better, although with a very small margin. For the PBT+ data-set with the isotropic network, the combinations with the best score of 91.9%, don't reach the best performing accuracy of 94.9% from the separate networks. For the anisotropic network, the performance of 91.3% is an improvement with a small margin over the score of 90.9% of the separate network. For the CMR data-set, the best performing combination in GatedGCN improves over the best performing separate representation accuracy, a combination of the representations has a positive effect on the accuracy. We will highlight the best performing recall and precision scores of the combined network, all the scores can be found in 19 in the appendix. For the GIN-CMR, the best performing recall score is 69.0% which is improved by the best performing recall in the combinations, all the representations combined achieve a recall score of 72.3%. For all the others, the recall score performs worse than with the best separate representation. The best performing recall score of GIN-PBT+ of 94.6%, of GatedGCN-PBT+ of 84.9%, and GIN-CMR of 74.2% are all better than the scores of GIN-PBT+ 87.5%, of GatedGCN-PBT+ of 84.7%, and GIN-CMR of 72.3%. The recall score of GIN for PBT+ with the bonds representation shows the highest difference with the combinations. The precisions score of the CMR networks shows a similar performance, for the GIN it is 85.7% for separate to 86.6% for combined, and for GatedGCN this is 85.3% to 86.2%. The precision scores of the PBT networks shows 82.6% to 88.6% and 97.5% to 92.7% for these. The improvement in accuracy for the CMR combination with all representations is not found in the recall scores and shows no persistent performance.

				CMR		PBT+	
1	2	3	4	GIN	GatedGCN	GIN	GatedGCN
x				77.9 ± 3.9	81.3±5.5	90.2±4.3	89.5±4.7
x	x			78.8 ± 4.7	79.5±5.7	91.2±5.4	91.2±4.2
x		x		80.7 ± 4.8	83.4±4.1	91.0±5.3	**91.3±4.2**
x			x	79.6 ± 3.7	83.4±3.8	91.7±5.6	90.6±5.2
x	x	x		80.9 ± 4.4	81.6±5.4	91.3±5.7	89.5±4.5
x	x		x	80.9 ± 4.4	81.6±5.4	91.3±5.7	89.5±4.5
x	x	x	x	**82.0 ± 4.9**	**84.1±4.2**	**91.9±5.6**	91.2±4.9

Table 10: Balanced accuracy scores in percentage of a combination of representation networks, where the 1 is the conventional network, the 2 is the bonds representation, the 3 is the hydro/arom representation and the 4 is the protein representation. The best balanced accuracy scores per dataset are shown, the best per column are marked in bold. The combination of all the representation has three times the best score and only in GatedGCN for PBT+ the best score is with the conventional and the hydro/arom combination.

5.3 Sub goals

The next items are sub-goals intended to research the influence of added chemical attributes. To summarize the experiments we will answer the sub-items, intended to research the research question.

Implement three new representations with chemical attributes added and research the influence of these new representation measured against the conventional representation. The particular additions have added value for different data-sets. The bonds representation improves the performance of the isotropic network over the conventional representation for the PBT+ data-set. For the CMR data-set, the SOMs add value for the isotropic network. For both networks, the hydro/aromatic addition increases the performance. Further, the complementary models show improvement over the conventional representation. A representation that wraps all the additions in one representation could improve the performance even more. The influence of these representations, therefore, seems data-set dependent. The influence is present and can contribute to significant improvement in accuracy, recall, and precision.

Introduce the new data-sets based on the nSVHC list and research the possible influence of the different characteristics within these toxicities in comparison to the new representations. For the GCNN's, the PBT+ data-set shows the best performance with the addition of the bonds, or with the bonds initialization. The bond type has shown to have the best influence on PBT classified chemicals. The hydro/arom and the SOMs show the highest performance increase with the CMR data-set, the hydro/arom shows the best influence on the CMR data-set. Thus, showing the influence of the different additions for the representations on a particular toxic class.

Research the use of the nSVHC lists data-sets against a baseline to compare the performance of the GCNN on these chemo-informatic

data-sets. The CMR and PBT+ data-sets show a significant difference between the baseline performance and their respective GCNN performances. The PBT data-set shows a small difference between the baseline and the respective GCNN performances and is therefore deemed less valuable as a GCNN data-set.

6 Conclusion

The answer of the research question, "Can the use of added chemical-based attributes contribute to the classification of toxicity within small unbalanced chemo-informatics data-sets with the use of graph convolution neural network", can be answered with the conclusion that we see a contribution of particular additions to the performance increase.

The performance in the baseline shows the added value of the chemical-based attributes outside the GCNN scope. The performance increase within the GCNN and representation combinations show that toxic class-specific improvements can be made with particular representations. The performance increase shows results with combinations of algorithms and representations and is data-set specific. These increases can make a difference within the classification of nSVHC chemicals and the use within screening tools for the aforementioned chemicals. The higher recall shows that the representations can add value to the increase in finding more nSVHC chemicals and decrease the chance of classifying toxic chemicals as nontoxic.

We conclude that these additions show considerable potential and are confident that further research can adhere to better screening tools.

Research into the influence of the representations over larger data-sets could indicate the performance increase these representations can give in general. Further, research into other representations options or more altered features vectors could help the classification. The substructures and bonds can be exploited in newly developed architectures or other data-set specific representations. The influence of an representation including all or a more sophisticated version of the representations is also a new angle that can shed light on the influence of these attributes. These results differ in the different toxic classes thus power can also be seen in toxic specific representations enhanced for the different toxic classes. To get insights into the workings of the new representations, an research can be considered into the performance of the representations together with explainability methods [43][44] for graphs that could give insights into the importance of the additions and for which class they contribute in which manner to the predictions. Case-specific research could not only help the performance increase but also give toxicologists insights in what the networks learn from the representations. New relations could arise from explainability methods and expand the field of both artificial intelligence and toxicology.

References

[1] E. Fröhlich and S. Salar-Behzadi, "Toxicological assessment of inhaled nanoparticles: Role of in vivo, ex vivo, in vitro, and in silico studies," *International Journal of Molecular Sciences*, vol. 15, p. 4795–4822, Mar 2014.

[2] T. Hartung, "Predicting toxicity of chemicals: software beats animal testing," *EFSA Journal*, vol. 17, 07 2019.

[3] R. Guha, "On exploring structure–activity relationships," *Methods in molecular biology (Clifton, N.J.)*, vol. 993, pp. 81–94, 04 2013.

[4] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, *et al.*, "Qsar modeling: where have you been? where are you going to?," *Journal of medicinal chemistry*, vol. 57, no. 12, pp. 4977–5010, 2014.

[5] A. O. Basile, A. Yahi, and N. P. Tatonetti, "Artificial intelligence for drug toxicity and safety," *Trends in pharmacological sciences*, vol. 40, no. 9, pp. 624–635, 2019.

[6] N. De Cao and T. Kipf, "Molgan: An implicit generative model for small molecular graphs," *arXiv preprint arXiv:1805.11973*, 2018.

[7] F. Montanari, L. Kuhnke, A. Ter Laak, and D.-A. Clevert, "Modeling physico-chemical admet endpoints with multitask graph convolutional networks," *Molecules*, vol. 25, p. 44, Dec 2019.

[8] F. Errica, M. Podda, D. Bacciu, and A. Micheli, "A fair comparison of graph neural networks for graph classification," in *International Conference on Learning Representations*, 2020.

[9] T. Chen, S. Bian, and Y. Sun, "Are powerful graph neural nets necessary? A dissection on graph classification," *CoRR*, vol. abs/1905.04579, 2019.

[10] C. Cai and Y. Wang, "A simple yet effective baseline for non-attribute graph classification," *CoRR*, vol. abs/1811.03508, 2018.

[11] P. Wassenaar, E. Rorije, N. Janssen, W. Peijnenburg, and M. Vijver, "Chemical similarity to identify potential substances of very high concern – an effective screening method," *Computational Toxicology*, vol. 12, p. 100110, 09 2019.

[12] M. Wójcikowski, M. Kukiełka, M. M. Stepniewska-Dziubinska, and P. Siedlecki, "Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions," *Bioinformatics*, vol. 35, pp. 1334–1341, 09 2018.

[13] J. Duan, M. Garikapati, S. Dixon, J. Lowrie, and W. Sherman, "Analysis and comparison of 2d fingerprints: Insights into database screening performance using eight fingerprint methods," *Journal of Cheminformatics*, vol. 3, pp. 1–1, 04 2011.

[14] M. Radifar, N. Yuniarti, and E. Istyastono, "Pyplif: Python-based protein-ligand interaction fingerprinting," *Bioinformation*, vol. 9, pp. 325–8, 03 2013.

[15] G. Idakwo, J. Luttrell, M. Chen, H. Hong, Z. Zhou, P. Gong, and C. Zhang, "A review on machine learning methods for in silico toxicity prediction," *Journal of Environmental Science and Health, Part C*, vol. 36, pp. 1–23, 01 2019.

[16] A. Karim, A. Mishra, M. A. H. Newton, and A. Sattar, "Efficient toxicity prediction via simple features using shallow neural networks and decision trees," *CoRR*, vol. abs/1901.09240, 2019.

[17] D. C. Elton, Z. Boukouvalas, M. D. Fuge, and P. W. Chung, "Deep learning for molecular generation and optimization - a review of the state of the art," *CoRR*, vol. abs/1903.04388, 2019.

[18] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[19] Y. Wu and G. Wang, "Machine learning based toxicity prediction: From chemical structural description to transcriptome analysis," *International Journal of Molecular Sciences*, vol. 19, p. 2358, 08 2018.

[20] G. B. Goh, C. Siegel, A. Vishnu, N. O. Hodas, and N. Baker, "Chemception: A deep neural network with minimal chemistry knowledge matches the performance of expert-developed qsar/qspr models," *ArXiv*, vol. abs/1706.06689, 2017.

[21] A. S. Rifaioglu, V. Atalay, M. J. Martin, R. Cetin-Atalay, and T. Doğan, "Deepscreen: High performance drug-target interaction prediction with convolutional neural networks using 2-d structural compound representations," *bioRxiv*, 2018.

[22] S. Chakravarti and S. Alla, "Descriptor free qsar modeling using deep learning with long short-term memory neural networks," *Frontiers in Artificial Intelligence*, vol. 2, 09 2019.

[23] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *CoRR*, vol. abs/1901.00596, 2019.

[24] L. Jia, B. Gaüzère, and P. Honeine, "Graph Kernels Based on Linear Patterns: Theoretical and Experimental Comparisons." working paper or preprint, Mar. 2019.

[25] N. Kriege, F. Johansson, and C. Morris, "A survey on graph kernels," *Applied Network Science*, vol. 5, 12 2020.

[26] V. P. Dwivedi, C. K. Joshi, T. Laurent, Y. Bengio, and X. Bresson, "Benchmarking graph neural networks," *ArXiv*, vol. abs/2003.00982, 2020.

[27] D. Bacciu, F. Errica, A. Micheli, and M. Podda, "A gentle introduction to deep learning for graphs," *Neural Networks*, vol. 129, p. 203–221, Sep 2020.

[28] RIVM, "nsvhc." https://rvs.rivm.nl/stoffenlijsten/Zeer-Zorgwekkende-Stoffen, 2020.

[29] RIVM, "Cmr." https://rvs.rivm.nl/gevaarsindeling/CMR, 2020.

[30] V. Mirčevska, M. Luštrek, and M. Gams, "Combining machine learning and expert knowledge for classifying human posture," in *Proceedings of the 18th International Electrotechnical and Computer Science Conference. B*, pp. 183–186, 2009.

[31] A. Zappone, M. Di Renzo, M. Debbah, T. T. Lam, and X. Qian, "Model-aided wireless artificial intelligence: Embedding expert knowledge in deep neural networks for wireless system optimization," *IEEE Vehicular Technology Magazine*, vol. 14, pp. 60–69, Sep. 2019.

[32] C. Baumgartner, C. Böhm, and D. Baumgartner, "Modelling of classification rules on metabolic patterns including machine learning and expert knowledge," *Journal of biomedical informatics*, vol. 38, pp. 89–98, 04 2005.

[33] C. Yuan, Wei, G. Wanbing, J. Jiang, Z. Wang, M. Zhang, and M. Li, "Toxicity prediction method based on multi-channel convolutional neural network," *Molecules*, vol. 24, p. 3383, 09 2019.

[34] M. Cronin, "The role of hydrophobicity in toxicity prediction," *Current Computer - Aided Drug Design*, vol. 2, pp. 405–413, 12 2006.

[35] A. Debnath, A. Shusterman, R. Compadre, and C. Hasch, "The importance of the hydrophobic interaction in the mutagenicity of organic compounds," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 305, pp. 63–72, 02 1994.

[36] J.-S. Seo, Y.-S. Keum, and Q. Li, "Bacterial degradation of aromatic compounds," *International journal of environmental research and public health*, vol. 6, pp. 278–309, 02 2009.

[37] T. Hughes, G. Miller, and S. J. Swamidass, "Site of reactivity models predict molecular reactivity of diverse chemicals with glutathione," *Chemical research in toxicology*, vol. 28, 03 2015.

[38] L. Olsen, M. Montefiori, K. P. Tran, and F. S. Jørgensen, "Smartcyp 3.0: enhanced cytochrome p450 site-of-metabolism prediction server," *Bioinformatics*, 2019.

[39] L. Afzelius, C. Hasselgren Arnby, A. Broo, L. Carlsson, C. Isaksson, U. Jurva, B. Kjellander, K. Kolmodin, K. Nilsson, F. Raubacher, *et al.*, "State-of-the-art tools for computational site of metabolism predictions: comparative analysis, mechanistical insights, and future applications," *Drug metabolism reviews*, vol. 39, no. 1, pp. 61–86, 2007.

[40] "RDKit: Open-source cheminformatics." http://www.rdkit.org. [Online; accessed 11-April-2013].

[41] A. Gobbi and D. Poppinger, "Genetic optimization of combinatorial libraries," *Biotechnology and Bioengineering*, vol. 61, no. 1, pp. 47–54, 1998.

[42] V. P. Dwivedi, C. K. Joshi, T. Laurent, Y. Bengio, and X. Bresson, "Benchmarking graph neural networks." https://github.com/graphdeeplearning/benchmarking-gnns/tree/arXivV1, 2020.

[43] F. Baldassarre and H. Azizpour, "Explainability techniques for graph convolutional networks," *CoRR*, vol. abs/1905.13686, 2019.

[44] P. Pope, S. Kolouri, M. Rostami, C. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 06 2019.

7 Appendix

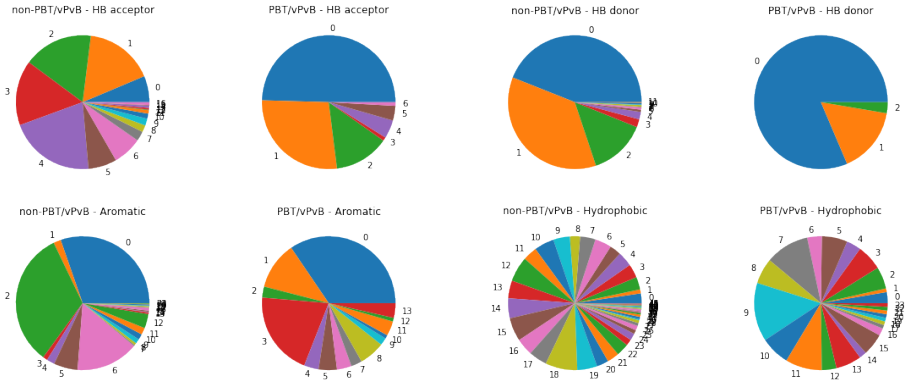

Figure 9: Distribution of chemical features dependent on toxicity endpoint CMR

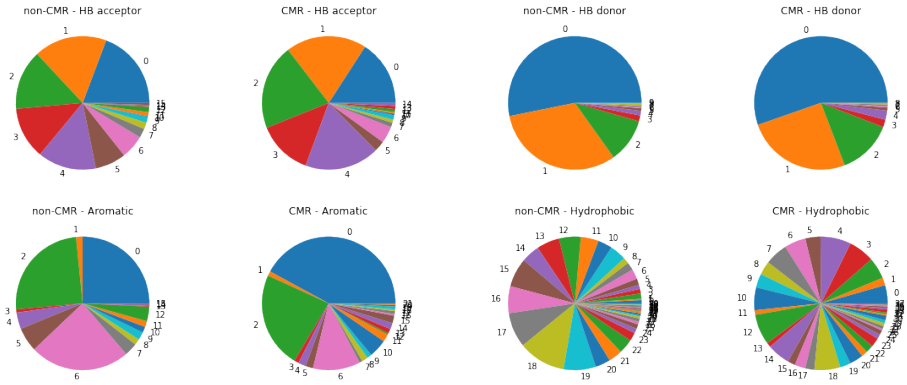

Figure 10: Distribution of chemical features dependent on toxicity endpoint PBT/vPvB

Substructural type	SMARTS pattern
Hydrogen bond Acceptor	{[}([$O, S; H1; v2$; !(*-*={[}O,N,P,S{]}){]})
	$({[}O,S;H0;v2{]}),$({[}O,S;-{]})
	$({[}N;v3;!$(N-*={[}O,N,P,S{]}){]})
	\textbackslash{}n&H0&+0
	$({[}o,s;+0;!([$o, s$] : n); !({[}o,s{]}:c:n){]}){]}
Hydrogen bond Donor	{[}$({[}N;!H0;v3,v4&+1{]})
	$({[}O,S;H1;+0{]})
	{}n&H1&+0{]}
Aromatic	{[}a{]}

Table 11: The SMARTS patterns for the hydrogen bond acceptors, hydrogen bond donors and the aromaticity

	Representation	I	H	O	C	L	IFD	D	R	GN
CMR	Conventional	17	200	128	2	2	0.5	0.5	mean	True
	Bonds	25	50	128	2	2	0.5	0.5	mean	True
	Hydro/arom	21	50	128	2	2	0.5	0.5	mean	True
	SOM	88	200	128	2	2	0.5	0.5	mean	True
PBT	Conventional	15	50	128	2	2	0.5	0.5	mean	True
	Bonds	23	25	128	2	4	0.5	0.5	mean	True
	Hydro/arom	19	50	128	2	2	0.5	0.5	mean	True
	SOM	86	100	128	2	2	0.5	0.5	mean	True
PBT+	Conventional	15	200	128	2	6	0.5	0.5	mean	True
	Bonds	23	150	128	2	2	0.5	0.5	mean	True
	Hydro/arom	19	200	128	2	6	0.5	0.5	mean	True
	SOM	86	200	128	2	4	0.5	0.5	mean	True

Table 12: The hyperparameters for the different MLP and representations combinations. The I is the in dimension, the H is the hidden dimension, the O is the out dimension, the C are the classes, the L is the layers, the IFD is the in feat dropout, the D is the dropout, the R is readout aggregation function and the GN is the graph normalization.

	Representation	I	H	C	L	NM	D	NE	R	GN	BN
CMR	Conventional	17	128	2	2	2	0.5	sum	sum	True	True
	Bonds	25	128	2	2	2	0.5	sum	sum	True	True
	Hydro/arom	21	96	2	2	2	0.5	sum	sum	True	True
	SOM	88	48	2	8	2	0.5	sum	sum	True	True
PBT	Conventional	15	96	2	8	2	0.5	sum	sum	True	True
	Bonds	23	128	2	8	2	0.5	sum	sum	True	True
	Hydro/arom	19	96	2	8	2	0.5	sum	sum	True	True
	SOM	86	48	2	8	2	0.5	sum	sum	True	True
PBT+	Conventional	15	128	2	6	2	0.5	sum	sum	True	True
	Bonds	23	128	2	6	2	0.5	sum	sum	True	True
	Hydro/arom	19	128	2	4	2	0.5	sum	sum	True	True
	SOM	86	96	2	4	2	0.5	sum	sum	True	True

Table 13: The hyperparameters for the different GIN and representations combinations. The I is the in dimension, the H is the hidden dimension, the C are the classes, the L is the layers, the NM is the number of MLP layers for GIN, the D is the dropout, the NE is the neighborhood aggregation function, the R is readout aggregation function, the GN is the graph normalization, and the BN is the batch normalization.

	Representation	I	H	O	C	L	D	R	GN	BN
CMR	Conventional	17	96	96	2	4	0.5	sum	True	True
	Bonds	25	96	96	2	6	0.5	sum	True	True
	Hydro/arom	21	96	96	2	6	0.5	sum	True	True
	SOM	88	96	96	2	6	0.5	sum	True	True
PBT	Conventional	15	96	96	2	6	0.5	sum	True	True
	Bonds	23	128	64	2	4	0.5	sum	True	True
	Hydro/arom	19	128	64	2	4	0.5	sum	True	True
	SOM	86	96	96	2	6	0.5	sum	True	True
PBT+	Conventional	15	196	196	2	2	0.5	sum	True	True
	Bonds	23	96	96	2	4	0.5	sum	True	True
	Hydro/arom	19	16	64	2	2	0.5	sum	True	True
	SOM	86	96	96	2	6	0.5	sum	True	True

Table 14: The hyperparameters for the different GatedGCN and representations combinations. The I is the in dimension, the H is the hidden dimension, the O is the out dimension, the C are the classes, the L is the layers, the D is the dropout, the R is readout aggregation function, the GN is the graph normalization, and the BN is the batch normalization.

	PBT MLP	GIN	Gated GCN
Conventional	88.1 ± 8.4	**93.7 ± 4.6**	92.2 ± 4.7
Bond add.	89.3 ± 7.0	93.0 ± 3.9	**92.4 ± 4.8**
Hydro/Arom add.	88.4 ± 7.5	93.0 ± 3.4	90.2 ± 4.8
SOM add.	**89.8 ± 5.9**	92.4 ± 5.8	91.5 ± 3.8

Table 15: Mean balanced accuracy after 10 fold cross validation for each representation and network combination. The bold marked are the best in their column and the red marked are the best for the whole dataset.

		PBT MLP		GIN		Gated GCN	
Conventional	P - R	0.9043	0.7856	**0.9137**	0.9023	0.8523	**0.8939**
Hydro/Arom prior	P - R	0.7847	**0.8394**	0.8501	0.9121	**0.9049**	0.8773
Bond prior	P - R	**0.9261**	0.8053	0.8509	**0.9114**	0.7940	0.8750
Protein prior	P - R	0.9097	0.8227	0.8401	**0.9114**	0.8242	0.8924

Table 16: Precision - Recall per representation and algorithm combination for the three different algorithms for the PBT+ data-set. The best performing precision and recall per column are depicted in bold, the best performing precision and recall for the whole data-set are shown in red.

				PBT	
1	2	3	4	GIN	GatedGCN
x				93.67±4.63	92.21±4.71
x	**x**			**93.80±4.53**	**94.38±4.26**
x		**x**		93.18±4.43	93.99±4.24
x			**x**	92.93±4.54	93.21±3.89
x	**x**	**x**		93.31±4.35	93.80±4.20
x	**x**		**x**	93.31±4.35	93.80±4.20
x	**x**	**x**	**x**	93.59±4.48	94.26±4.37

Table 17: Balanced accuracy scores of a combination of representation networks,where the 1 is the conventional network, the 2 is the bonds representation, the 3 is the hydro/arom representation and the 4 is the protein representation. The best balanced accuracy scores per dataset are shown, the best per column are marked in bold.

1	2	3	4		PBT GIN		GatedGCN	
x				p-r	0.68	0.80	0.85	0.89
x	**x**			p-r	0.72	**0.84**	**0.92**	**0.91**
x		**x**		p-r	0.68	0.82	0.90	**0.91**
x			**x**	p-r	0.69	**0.84**	0.86	**0.91**
x	**x**	**x**		p-r	**0.85**	0.71	**0.92**	0.90
x	**x**		**x**	p-r	**0.85**	0.71	**0.92**	0.90
x	**x**	**x**	**x**	p-r	**0.87**	0.72	**0.92**	**0.91**

Table 18: Precision and recall scores for the PBT dataset of a combination of representation net-works, where the 1 is the conventional network, the 2 is the bonds representation,the 3 is the hydro/arom representation and the 4 is the protein representation.

1	2	3	4	CMR GIN			GatedGCN			PBT+ GIN			GatedGCN		
x				80.5	-	67.9	83.2	-	73.4	72.9	-	85.7	97.2	-	79.3
x	x			82.1	-	68.2	81.7	-	70.4	76.0	-	86.6	**97.5**	-	83.0
x		x		83.8	-	71.6	85.0	-	76.7	76.7	-	86.6	96.5	-	82.9
x			x	83.6	-	69.3	85.8	-	76.0	79.8	-	86.6	92.9	-	82.0
x	x	x		84.7	-	71.2	85.1	-	72.6	77.9	-	86.6	97.3	-	79.4
x	x		x	84.7	-	70.1	84.9	-	76.0	80.4	-	**87.5**	94.0	-	**84.7**
x	x	x	x	**86.6**	-	**72.3**	**86.2**	-	**77.1**	**82.6**	-	86.6	95.8	-	83.0

Table 19: Precision and recall scores for the CMR and PBT+ dataset of a combination of representation networks, where the 1 is the conventional network, the 2 is the bonds representation, the 3 is the hydro/arom representation and the 4 is the protein representation.

	CMR Isotropic Conv.	Isotropic SOM	Anisotropic Hydro/arom
Accuracy	79.9	82.3	83.2
True negative	335	345	336
False positive	46	36	45
True positive	184	201	212
False negative	87	70	59

	PBT+ Isotropic Conv.	Isotropic Bonds
Accuracy	92.1	94.9
True negative	814	823
False positive	45	36
True positive	96	106
False negative	16	6

	Accuracy
CMR - Screening tool	79.9 %
CMR - Best GCNN	83.2 %
PBT - Screening tool	91.1 %
PBT - Best GCNN	94.9 %
PBT+ - Best GCNN	93.7 %