Learning Machine Reasoning for Bioactivity Prediction of Chemicals

Suman K. Chakravarti

e-mail: chakravarti@multicase.com

MultiCASE Inc., 23811 Chagrin Blvd., #305, Beachwood, OH 44122, USA

Abstract

We describe a method for learning higher-level vector representations of interactions between molecular features and biology. We named the representations as the *reason vectors*. In contrast to the high-dimensional chemical fingerprints, reason vectors are much simpler with only about 5 dimensions. They allow abstract reasoning for bioactivity of chemicals or absence thereof, uncover causal factors in interactions between chemical features and generalize beyond specific chemical classes or bioactivity. These qualities enable us to perform powerful similarity searches that are vague and conceptual in nature. The methodology can handle novel combinations of features in query molecules and can evaluate chemical classes that are entirely absent in training data. The method consists of similarity-based near neighbor search on a reference database of biologically tested chemicals by a series of substructures obtained from stepwise reconstruction of the test molecule. A data-driven continuous representation of molecular fragments was used for molecular similarity computations. The technique was inspired by the ability of humans to learn and generalize complex concepts by interacting with the physical world. We also show that activity prediction of chemicals using the abstract reason vectors is very easy and straightforward, as compared to modeling in the raw chemistry space, and can be applied to both binary and continuous activity outcomes. Except for utilizing an unsupervised training to construct continuous molecular fingerprints, the methodology is devoid of gradient optimization or statistical fitting.

Introduction

Since its introduction almost 60 years ago,[1] quantitative structure-activity relationship (QSAR) modeling was primarily intended for correlating molecular structures and their properties to predict activity of new molecules. However, QSARs are largely driven by statistical learning and correlations,[2] elements of reasoning and causality are absent. This shortcoming was not too problematic at the beginning, as models were built with small, focused and carefully planned congeneric sets of chemicals with the aid of hand-picked molecular descriptors. They primarily reflected the reasonings postulated by the model builder. However, as increasingly more QSARs

are built with large diverse training sets, this is not the case anymore.[3,4] Large training sets only allow models to be built directly using raw molecular structural data, e.g. molecular fragments, binary fingerprints or deep learning-based continuous representations.[5,6] Unfortunately, such modeling methods pick up superficial patterns in the raw data and fail when models encounter unseen or unusual combinations of features during tests. These issues are well known in the machine learning field and has been found to be the cause of some serious problems.[7]

Reasoning can be defined as the ability to evaluate implicit relationships that may not be explicitly present in the training data.[8] Correlations sometimes give a false sense of reasoning, mostly due to patterns well represented in the training data. Also, QSARs (built using raw structural data) fail to produce meaningfully different results when we make subtle variations in the query structure to ask *what if* questions. Even for basic generalizations, prohibitively large number of examples are needed to cover possible structural variations of model parameters (curse of dimensionality). Models usually contain contributions of individual structural features towards activity and not discerning enough for relative positions or novel combinations in the query. This is why QSAR results frequently need to be reviewed by human experts, especially in safety assessment of chemicals towards human health.[9]

The ability to perceive causality is much more than just capturing patterns in the raw data.[10-13] Traditional correlation-based models give unsatisfactory answers because they only account for individual features' ability to increase or decrease activity (e.g. regression coefficients), but there is no statistical parameter that can represent causality, e.g. group A inhibits toxicity of group B. It is also very easy to falsely assume correlations as causations. As with reasoning, human experts often inject causality after reviewing prediction results. An important benefit of perception of causality would be the ability to compute effects of interventions with much better accuracy, e.g. toxicity reduction or efficacy improvements in drug candidates.

Some of these shortcomings are due to the fact that usually all the learning happens during the model building phase. Then they are applied in tests without adjustments. Humans, on the other hand, continuously form new combinations of existing knowledge and dynamically adjust their importance to solve daily problems. A large part of our intelligence comes from our interaction with the world and assisted by our stored knowledge. Also, humans have a remarkable ability to envision imaginary situations and take actions based on them. It is worth noting about the two

different systems of human decision-making process proposed by Daniel Kahneman[14]: *System 1*, which is automatic, quick, involuntary and requires less effort; and *System 2*, which is slow, sequential, requires effort and can be expressed with languages. Current QSAR and machine learning methods are better at system 1 processes, but we would like them to achieve system 2 abilities. Also, the seminal work of Judea Pearl on causality[13,15] cannot be overlooked. He argues that current AI systems are at level 1 (association), and the higher level 2 (intervention) and level 3 (counterfactuals) cannot be achieved without incorporating causality.

Currently, these problems are focus of intense research in the field of artificial intelligence and machine learning. Concerted effort is being spent to unify principles of classic symbolic AI and modern deep learning techniques to account for causality and reasoning. A general solution seems to be distant at this point, but few clues have started to appear. One such direction is to learn higher levels of abstractions from raw input data.[16-18] Researchers in the field argue that such abstractions allow disentanglement of underlying factors and helps in generalization and transfer. Disentanglement is the phenomenon when the raw data is transformed to the right higher space and the underlying factors become separated.[19] Another approach includes agents interacting with the environment and observing the outcomes to uncover causal factors.[20]

In this study, we developed a simple method to learn reasoning and causality for the purpose of bioactivity prediction, using clues from recent developments in the field of artificial intelligence. It involves learning higher abstract representation to uncover relevant underlying factors by systematically constructing molecules and observing their biological effects. Following are the highlights of the work:

i. Reason vectors are generated by a stepwise reconstruction of the query molecule starting from a single atom to the whole structure, performing near-neighbor predictions with the reference database at every step. This is analogous to an agent interacting with world and observing the outcomes. We consider near-neighbor predictions to be a *System 1* process; fast, works with stored knowledge and gives instant activity predictions. Whereas, reason vector construction is analogous to a *System 2* process; sequential, slower and requires additional computation.
ii. We have described the chemical world using distributed, continuous fingerprints. The building blocks of these fingerprints were learnt using a separate unsupervised learning from

~17 million unlabeled chemical structures. The purpose is to enable generalizations to new combinations of learned chemical features not seen during training and to have rich dense representation of small parts of molecules down to single atoms.

iii. The reason vectors represent general concepts, e.g. "*deactivation of a bioactive functionality by another*", a useful abstraction higher than the raw relationships between specific chemical features and specific biological outcomes. Also, the reason vectors uncover causality in the interaction between the chemical features.

iv. It is straightforward to use the concepts encoded in the reason vectors to get activity predictions, for example, if majority of the vectors of a query molecule indicate activity, then the molecule is active.

v. The methodology enables, to a certain extent, assessment of chemical classes for which no examples are present in the training set. This is possible by the combined benefits of the distributed representations and the sequential activity predictions encoded in the reason vectors.

Methods

Data. We tested our approach on four datasets:

1) Ames mutagenicity (AMES)
2) Aryl hydrocarbon receptor activators (AHR)
3) Skin sensitization (SKIN_SENS)
4) Rat acute oral toxicity (LD50)

These data sets cover both binary (Ames, AHR, skin sensitization) and continuous activity outcomes (LD50); in vivo (LD50 and skin sensitization) and in vitro (Ames and AHR) bioassays; cell based (Ames), whole animal (skin sensitization and LD50) and receptor binding high throughput screening (AHR) assays. The data sets range in size from 3122 to 23070 compounds. They were subdivided in train and test sets via random split (Table 1). Two of the datasets, Ames[27,28] and LD50,[21] are suitable for comparing prediction performances as they appear in previous QSAR based publications. We kept the train and test sets same.

Except for the Ames set, which in part contains proprietary data, the data was collected from publicly available sources. The LD50 data is from the inventory of National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods

(NICEATM) and the U.S. EPA National Center for Computational Toxicology (NCCT), who collected and curated the data from several sources.[21] For skin sensitization, we have collected guinea pig maximization test, Buehler and human patch test data from several publications[22-24] and public data sources including European Chemical Agency (ECHA) dossiers via eChemPortal.[25] AHR dataset is from a PubChem bioassay AID 2796.[26] We have published the details of the mutagenicity data set in previous articles.[27,28] In short, it contains data from a number of public and proprietary sources. The proprietary data is mainly from the Ames/QSAR International Challenge conducted by the Division of Genetics and Mutagenesis, National Institute of Health Sciences, Japan.[29]

The datasets were subjected to some common preprocessing steps including aromaticity perception, elimination of stereochemistry, neutralizing charges on certain atoms and removal of alkali metal salt parts, only one chemical with the highest activity was retained in case of duplicates. The AHR data set initially contained 324858 compounds with an overrepresentation of the negative class, therefore, we randomly excluded a majority of the negatives to make it reasonably balanced with about one third positives.

Table 1. Size of train and test splits of the datasets used in this study. First three sets have binary activity labels, whereas the LD50 dataset have continuous activity labels.

dataset	train	test	total (positive fraction)
Ames mutagenicity (AMES)	17005	1942	18947 (38%)
Skin sensitization (SKIN_SENS)	2810	312	3122 (30%)
Aryl hydrocarbon receptor activators (AHR)	20763	2307	23070 (33%)
Rat acute oral toxicity (LD50) *	6279	2134	8413

*activity units: log LD_{50} (mmol/kg body weight); higher values indicate lower potency.

Molecular fingerprints. A continuous, distributed representation of molecular fragments was used, which we and others reported recently.[30,31] The publications provide complete description of these fingerprints. In brief, these fingerprints were built via unsupervised training on a text corpus

comprised of atom centered fragments from approximately 17 million unlabeled PubChem chemicals. The Word2Vec algorithm[32] was used for the unsupervised training. As a result, we got 167, 21395 and 340412 vectors for fragments of depth zero, one and two respectively. Summation of the vectors of these small fragments produce fingerprints for bigger molecular fragments. The size of the fingerprints was kept at 600 for this study. It is worth noting that the fingerprints can be generated for fragments of any arbitrary size, and consequently similarity between any two fragments can be calculated, even between a single atom and a full molecule. These qualities are particularly suitable for this work.

Traditional fragment-based 1024-bits binary hashed fingerprints were also used in some parts of this study. They were built using linear molecular fragments of 2-10 path length.

Similarity computation. Cosine and Tanimoto similarity measures were used for computing similarity within continuous and binary fingerprints respectively. Euclidean distances were used for the reason vectors.

Generating reason vectors. A query compound and a reference dataset of molecules tested experimentally in the bioassay of interest are needed for computing these vectors. The process is shown in Figure 1, and consists of stepwise reconstruction of the query molecule, starting from each of its atoms and performing *k*-nearest neighborhood (*k-nn*) similarity search using the fragments from each step. Following is the process to generate one reason vector:

i. Compute fingerprints for the molecules of the reference dataset.
ii. Select an arbitrary atom *m* on the query chemical (a single atom fragment) and compute its fingerprint.
iii. Compute similarity between the above fragment and all the molecules of the reference dataset, average the bioactivity for the *k* (5, 7 or 9) most similar reference structures. This is the first element of the reason vector.
iv. Expand the fragment by adding neighboring bonded atoms to *m*, compute its fingerprint, perform step *iii* and add the average activity of the neighbors as the second element to the reason vector.
v. Repeat steps *iii* and *iv* until the whole molecule is covered, producing the full reason vector.

It is important to note that each element of the vector is a result of the similarity measurement between a fragment and all the molecules of the reference database, not a substructure hit search. Also, the above methodology can be repeated for every atom to produce all reasons vectors of the query molecule, as shown in Figure 2a for the example query molecule with reference to the Ames mutagenicity dataset.

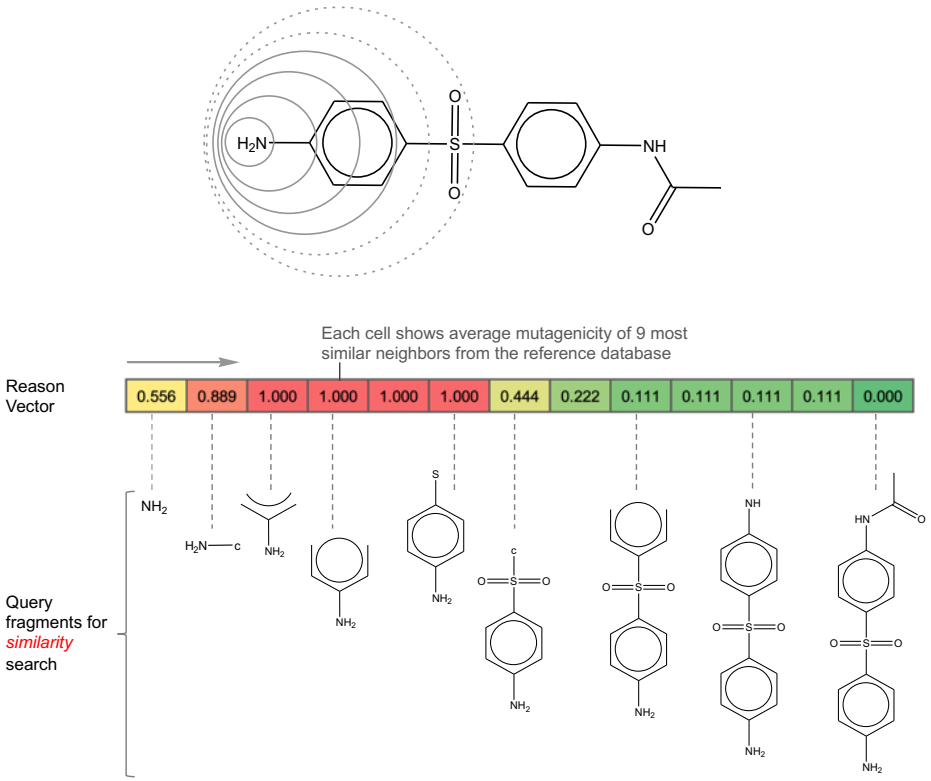
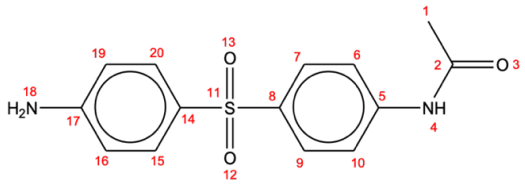


Figure 1: A reason vector being generated from an atom of a query molecule, using the Ames mutagenicity dataset as the reference. The vector starts from the nitrogen of the aromatic amine. It is important to note that each element of the vector is a result of similarity calculations between a fragment and all the molecules of the reference database, not a substructure hit search.



Atom #													
1	0.111	0.333	0.111	0.333	0.444	0.444	0.333	0.333	0.000	0.000	0.000	0.000	0.000
2	0.444	0.111	0.333	0.444	0.444	0.333	0.333	0.000	0.000	0.000	0.000	0.000	
3	0.556	0.444	0.111	0.333	0.444	0.444	0.333	0.333	0.000	0.000	0.000	0.000	0.000
4	0.556	0.333	0.444	0.444	0.333	0.333	0.000	0.000	0.000	0.000	0.000		
5	0.111	0.444	0.667	0.333	0.333	0.000	0.000	0.000	0.000	0.000			
6	0.111	0.222	0.444	0.444	0.000	0.000	0.000	0.000	0.000				
7	0.111	0.222	0.333	0.000	0.000	0.000	0.000	0.000					
8	0.111	0.111	0.000	0.000	0.000	0.000	0.000						
9	0.111	0.222	0.333	0.000	0.000	0.000	0.000	0.000					
10	0.111	0.222	0.444	0.444	0.000	0.000	0.000	0.000	0.000				
11	0.444	0.000	0.000	0.000	0.000	0.111	0.111	0.000					
12	0.556	0.222	0.000	0.000	0.000	0.000	0.111	0.111	0.000				
13	0.556	0.222	0.000	0.000	0.000	0.000	0.111	0.111	0.000				
14	0.111	0.111	0.000	0.000	0.111	0.111	0.111	0.111	0.000				
15	0.111	0.222	0.333	0.444	0.222	0.111	0.111	0.111	0.111	0.000			
16	0.111	0.222	1.000	1.000	0.444	0.222	0.111	0.111	0.111	0.111	0.000		
17	0.111	1.000	1.000	1.000	1.000	0.444	0.222	0.111	0.111	0.111	0.111	0.000	
18	0.556	0.889	1.000	1.000	1.000	1.000	0.444	0.222	0.111	0.111	0.111	0.111	0.000
19	0.111	0.222	1.000	1.000	0.444	0.222	0.111	0.111	0.111	0.111	0.000		
20	0.111	0.222	0.333	0.444	0.222	0.111	0.111	0.111	0.111	0.000			

a. Vectors from each atom

Atom #					
1	0.185	0.407	0.222	0.000	0.000
2	0.278	0.389	0.389	0.167	0.000
3	0.370	0.407	0.222	0.000	0.000
4	0.444	0.444	0.333	0.000	0.000
5	0.278	0.500	0.167	0.000	0.000
6	0.167	0.444	0.000	0.000	0.000
7	0.167	0.167	0.000	0.000	0.000
8	0.111	0.111	0.000	0.000	0.000
9	0.167	0.167	0.000	0.000	0.000
10	0.167	0.444	0.000	0.000	0.000
11	0.222	0.000	0.056	0.056	0.056
12	0.389	0.000	0.000	0.111	0.000
13	0.389	0.000	0.000	0.111	0.000
14	0.111	0.000	0.111	0.111	0.000
15	0.167	0.389	0.167	0.111	0.056
16	0.167	1.000	0.333	0.111	0.074
17	0.556	1.000	0.722	0.167	0.083
18	0.815	1.000	0.259	0.111	0.000
19	0.167	1.000	0.333	0.111	0.074
20	0.167	0.389	0.167	0.111	0.056

b. Transfomed vectors

Figure 2. Full list of reason vectors of a query molecule by using the Ames mutagenicity dataset as the reference. a) List of full-length raw vectors, note varying lengths for different atoms, b) vectors after transformation to a uniform length of 5.

As shown in Figure 2a, raw reason vectors from a molecule can have different lengths depending on the location of the start atom and number of steps required to cover the whole structure. This creates a difficulty in their effective use. Therefore, we transform them to a fixed uniform length (e.g. 5 or 7) as shown in Figure 2b, by simple compression or expansion. This conversion preserves the overall character of the vectors. Mainly, it allows for similarity calculations using distances in Euclidean space. We have used the transformed vectors for all computations in this paper.

Predicting activity of a reason vector: As shown in Figure 3a, the activity of a reason vector can be predicted by finding similar vectors from the reference chemicals (whose bioactivities are known). First, the reference vectors are annotated with the bioactivity of their parent chemicals. Second, a search is made to find a few (usually 10) vectors that are similar to the query vector.

The biological activities of the found vectors are averaged and assigned as the activity of the query vector.

Predicting activity of a query molecule: Once the activity of the reason vectors of a query molecule have been predicted, its own activity can be calculated based on the distribution of predicted activities of its vectors. This is shown in Figure 3b. First, activity range of the training data is divided in equally spaced bins. Second, the vectors are placed in corresponding bins based on their predicted activity. A probability value is then computed for each bin based on the count of vectors in it. The predicted activity is simply the average of the mean activity of the bins weighted by the probability values. We found this procedure to be equally effective for both binary and continuous activity outcomes.

Figure 3. Schematics of activity prediction method for (a) a reason vector and (b) a molecule. First, the activities of individual reason vectors of the molecule are predicted using the method shown in 'a'. Second, the activity distribution of the vectors is used in prediction of the query molecule's activity, shown in 'b'.

Additional modeling techniques for comparison: We have used three standard modeling methods for comparisons:

1. *k*-Nearest neighborhood using distributed fingerprints (knn_DISTR_FP): Activity of a molecule is predicted from the activity labels of *k* most similar molecules from the reference dataset. Distributed fingerprints are used for similarity computations.
2. *k*-Nearest neighborhood using binary fingerprints (knn_BINARY_FP): Same as above, but binary fragment-based hashed fingerprints were used for similarity computations.[33]
3. ECFP fragment-based regression modeling (LOGIST_REGR_ECFP): ECFP type atom-centered fragment descriptors were used to build regression models. Logistic and ordinary regression was used for binary and continuous activity outcomes respectively. L1 regularization was used to pick relevant fragments.[27]

Validation methodology: In addition to the external test sets, we used multiple train-test subsets using random split from the primary training sets. A series of training sets of size of 100, 200, …, 800, 6400 etc. were formed. The size of the test sets was kept at 2000, 2000, 281 and 1000 for mutagenicity, AHR, skin sensitization and LD50 respectively. Every combination of the train-test set was repeated multiple times to obtain stable estimates. Specifically, *k-nn* methods were repeated 50 times because they are computationally inexpensive as opposed to the reason vector methodology which we could repeat only 5 times for a few combinations. ROC-area under curve (ROC-AUC) was used for the binary datasets whereas root mean square error (RMSE) was used for LD50 prediction.

Software: Python package *Gensim*[34] was used for accessing the *Word2Vec* algorithm. The R package *Rtsne*[35] was used for generating t-Distributed Stochastic Neighbor Embedding (t-SNE) plots. An in-house cheminformatics software library was used for handling chemical structures, fragmenting chemicals, computing fingerprints, reason vector methodology, traditional QSAR analysis and all other operations described in this paper.

Results and Discussion

Perception of causality: By means of a series of substructures following a systematic, stepwise procedure, the reason vectors mimic the sequential response of a biological system. Interactions between structural features of molecules can be deduced from these sequences, e.g. deactivation

or boosting of biological effects of a functional group by another. Moreover, cause and effect can be separated. In Figure 1 for example, the reason vector generated from the amine nitrogen (atom #18) suggests that aromatic amines can cause mutagenicity as shown by the elevated mutagenicity (red cells) in the first half of the vector. But, as soon as the sulfonyl group gets added to the growing substructure, mutagenicity decreases significantly (green cells). Whereas, when a vector is generated from the sulfonyl group (atom #11) of the same molecule (Figure 4), the vector is devoid of any activity, implying a causality supported reasoning that sulfonyl group suppresses mutagenic potential of the aromatic amine. Human experts with domain knowledge routinely perform such causal deductions, whereas, we are computationally generating such reasoning for more effective use.

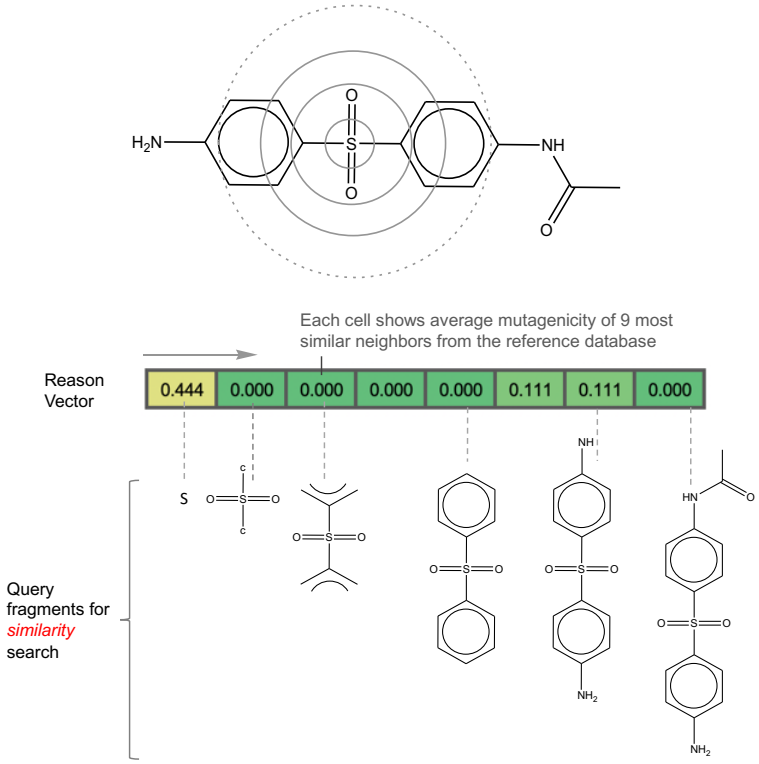
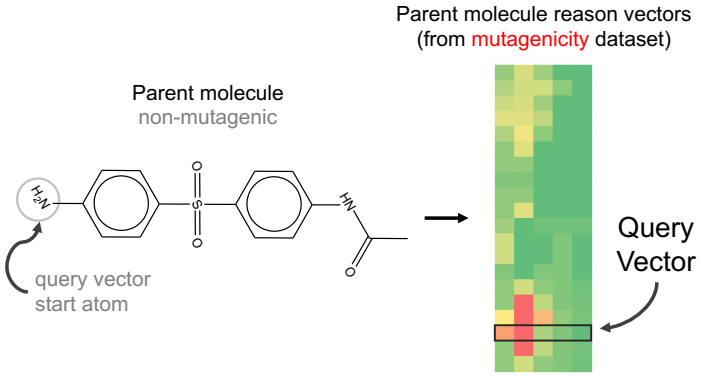


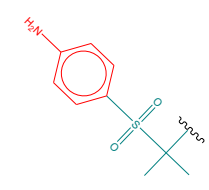
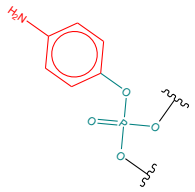
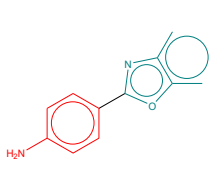
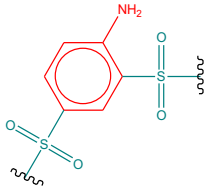
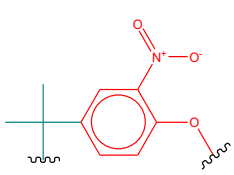
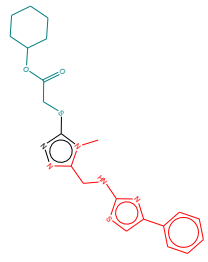
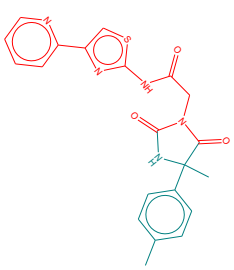
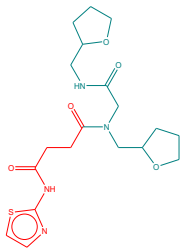
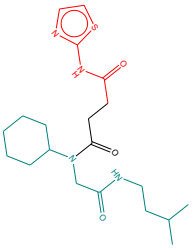
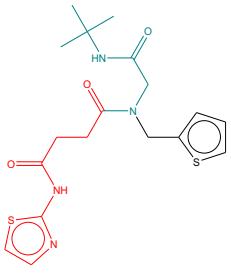
Figure 4. Using the same molecule from Figure 1, a mutagenicity reason vector is being generated from the sulfur atom of the sulfonyl group. The Ames mutagenicity dataset was used as the reference. Note the absence of mutagenic activity in the resulting vector.

Ability to generalize: We present evidence in this section to show that the vector shown in Figure 1 represents "*deactivation of a bioactive functionality by another*" rather than "*deactivation of mutagenic potential of aromatic amino group by the sulfonyl group*".

Two nearby vectors in Euclidean space represent similar reasoning for bioactivity or absence thereof. For instance, vectors that are similar to the one in Figure 1 depict some type of deactivation mechanism and their parent molecules are almost all non-mutagenic. The search results are shown in Table 2. Specifically, the aromatic amine functionality is deactivated by the sulfonyl group in the query molecule, however, the associated functional groups are not always the same in the search results. For example, the mutagenicity hits in Table 2 include an aromatic nitro group deactivated by a bulky tertiary-butyl group. Most interestingly, hits from other activity domains (e.g. aryl hydrocarbon or skin sensitization) also represent deactivation of a bioactive functionality by another and their parent molecule are all inactive (i.e. AHR non-activators or skin non-sensitizers). We got similar results using an 'active' query vector from the AHR space as shown in Table 3. The hits are active compounds from skin sensitization and mutagenicity datasets. We believe this to be an important finding, i.e. a reason vector from one activity domain can be used for searching in a completely different bioactivity domain. This shows that rather than being directly anchored to any specific molecular feature or bioactivity, reason vectors are abstractions of activity mechanisms. Therefore, facilitate useful generalizations and wider applicability. They are much simpler than the chemical fingerprints (5 elements vs. 600) yet capture more than just patterns in the raw data and bring us closer to the underlying causality. The results shown in Tables 2 and 3 also represents a different type of similarity search, where the query is not a molecule or a substructure, but an imprecise concept.

Table 2. Search hits from different bioactivity domains, using a reason vector from **mutagenicity** domain as a query representing deactivation of a bioactive functionality by another. All hits are inactive and contains a variety of active functionality (red) deactivated by another (green) from different activity domains.



Mutagenicity hits				
				
Non-Mutagenic	Non-Mutagenic	Non-Mutagenic	Non-Mutagenic	Non-Mutagenic
				
AHR hits				
				
AHR inactive	AHR inactive	AHR inactive	AHR inactive	AHR inactive
				
Skin sensitization hits				

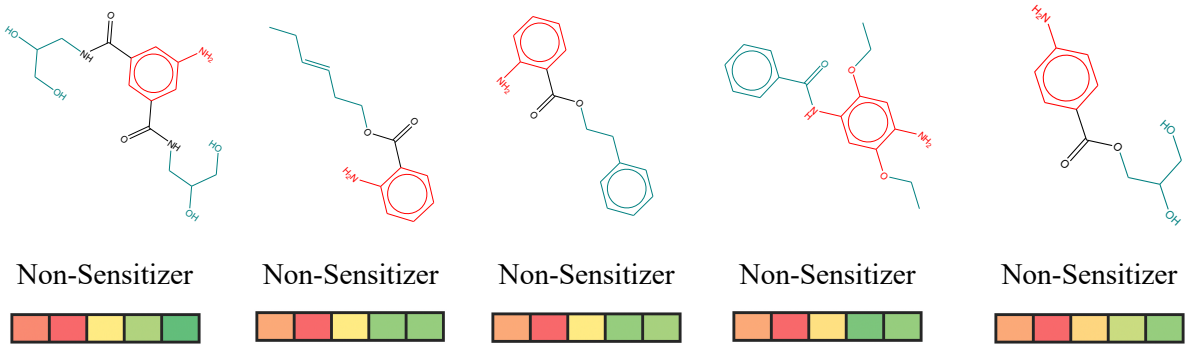
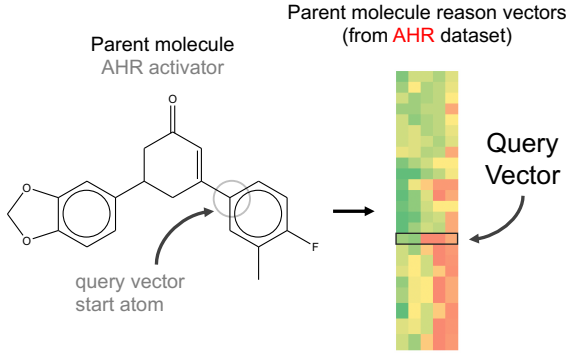
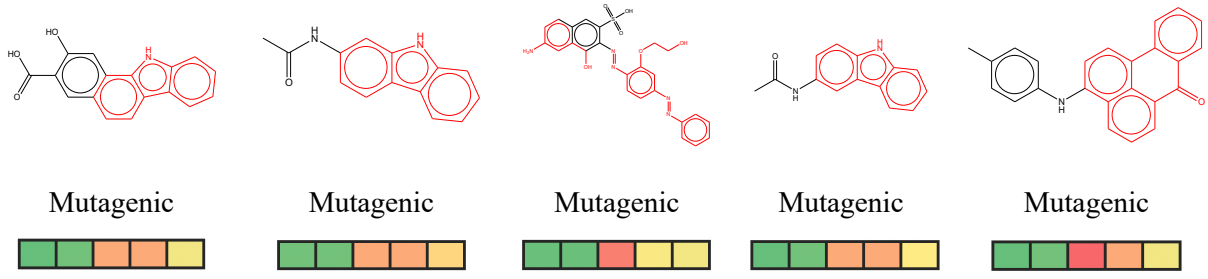


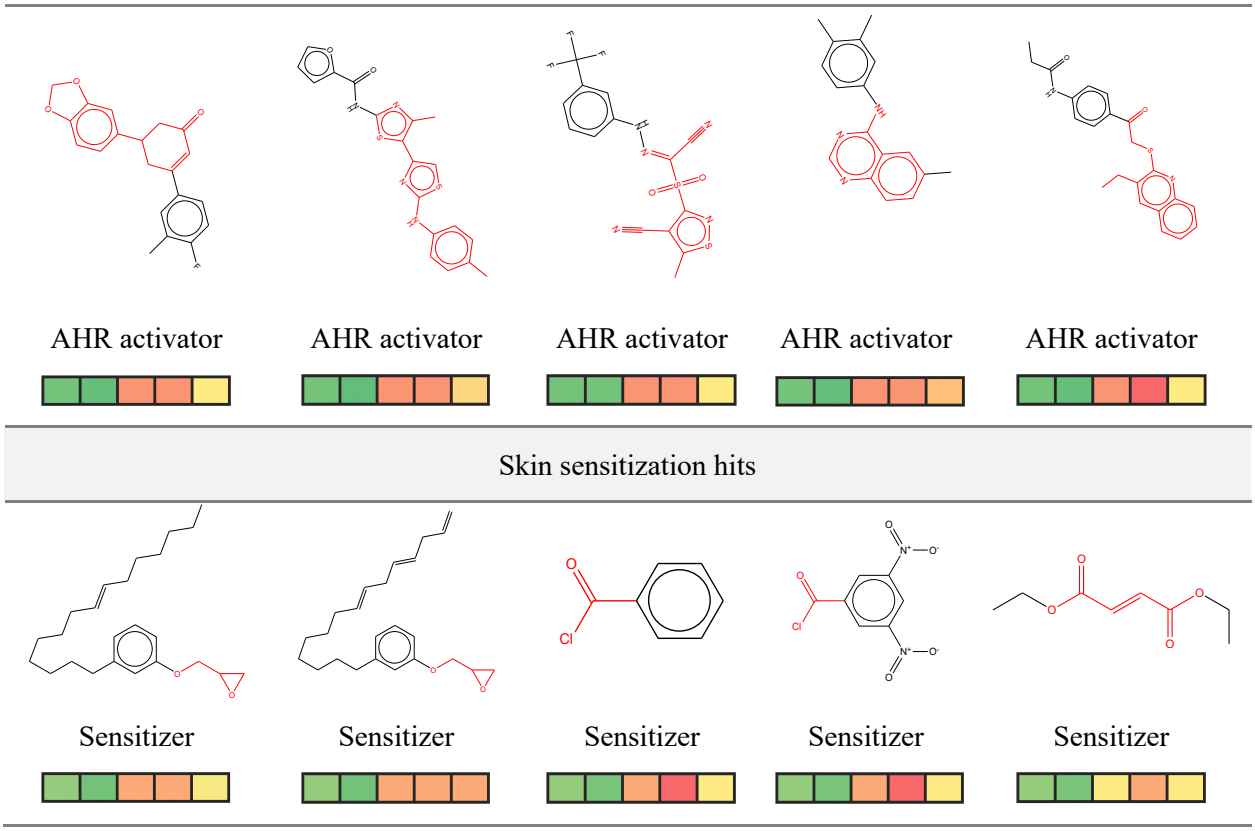
Table 3. Hits from searching different bioactivity domains, using a reason vector from **aryl hydrocarbon** domain as a query representing a reason for activity. All hits are active and contains a variety of active functionalities (red) from different activity domains.



Mutagenicity hits



AHR hits



Next, a more detailed experiment was performed to confirm the generalizing ability of the reason vectors. A thousand 'active' vectors were randomly selected from each activity domain, then a cross-search was performed in different domains to obtain a few (e.g. 9 or 11) closest vectors. Same process was repeated for sets of 'inactive' reason vectors. The percentage of active and inactive vectors are counted in the results. Mutagenicity, AHR and skin sensitization datasets (binary outcome sets) were used for this purpose. The results are shown in Table 4. We found that active query vectors from mutagenicity and AHR produce active hits from other domains. For example, Ames positive vectors returned 81% and 84% active vectors from AHR and skin sets respectively. Active vectors from skin sensitization were an exception, returning a mix of active and inactive vectors from other datasets, possibly due to a relatively small dataset size. Inactive query vectors from any set, on the other hand, returned mostly inactive hits across all three activity domains.

Table 4. Results from searching active and inactive reason vectors across different activity domains. Each query set (first column) is comprised of a thousand vectors from a particular domain.

query vector *	activity distribution of vectors in search result					
	AHR+	Skin+	Ames+	AHR–	Skin–	Ames–
Ames+	81%	84%	-	19%	16%	-
Ames–	5%	10%	-	95%	90%	-
AHR+	-	79%	70%	-	21%	30%
AHR–	-	17%	12%	-	83%	88%
Skin+	42%	-	45%	58%	-	55%
Skin–	7%	-	8%	93%	-	92%

* '+' or '–' indicates active or inactive query vectors, e.g. AHR+ stands for active vectors from the aryl hydrocarbon dataset.

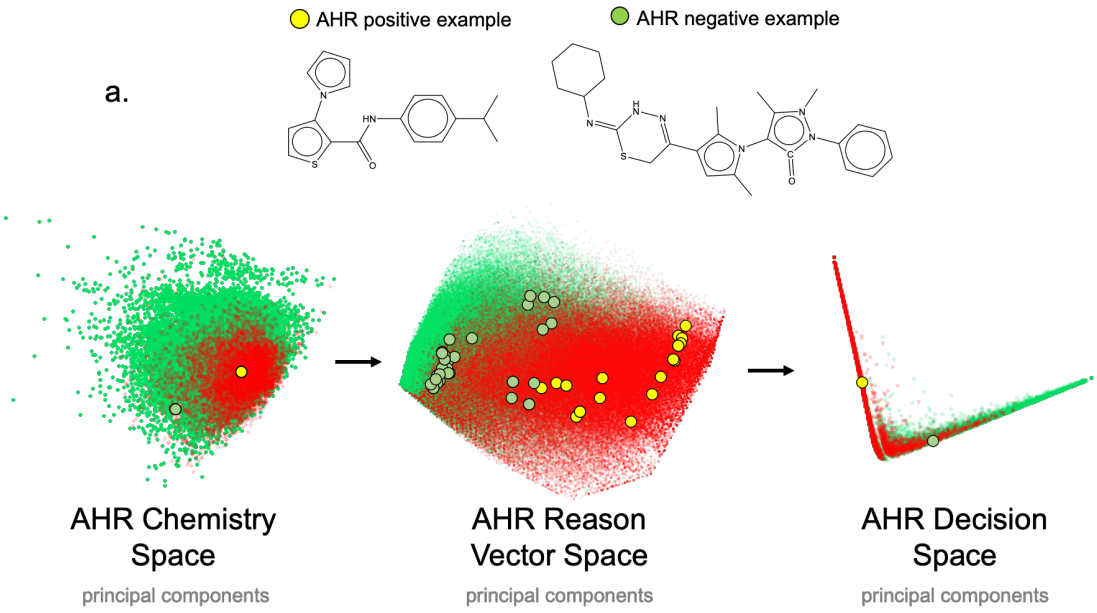
Analysis of different vector spaces: From a broader perspective, the process of activity prediction is equivalent to placing the query molecule in three consecutive vector spaces, with progressive simplification from one to the next:

1. *Chemistry space:* Represented by the high dimensional (600D) molecular fingerprints from the input data. Each molecule is represented by one point in this space.
2. *Reason vector space:* Consists of reason vectors of low dimensions (5D or 7D). Each chemical is represented by multiple points (depending on the number of reason vectors from the molecule). The vectors contain only a few factors relevant to bioactivity.
3. *Decision space:* Consists of 10-20D vectors representing the distribution of predicted activities of reason vectors in the query molecule. This space is used for activity prediction. Every molecule is represented by one point.

These three spaces are actually a confirmation of the manifold hypothesis[36,37] that high dimensional data usually lie close to a low dimensional manifold and real data of interest lives in a space of low dimensions. This is illustrated in Figure 5a with two molecules' (AHR active and inactive respectively) in the corresponding vector spaces of AHR activity domain. We have used principal components to help the display. It can be seen that the reasons for activity and inactivity are separated with less overlap in the reason vector space as compared to the chemistry

space. The reason vector space has considerably more data points as compared to the chemistry space, however, the vectors only contain a few key features. The two example molecules are represented by multiple points in the reason vector space and their activity distributions are well separated as shown in Figure 5b. The decision space for the AHR dataset practically has only two dimensions, one rich in actives while the other in inactive molecules, in line with the binary nature of this dataset.

Figure 6 shows the vector spaces for the LD50 dataset. We have used a t-SNE plot to show the chemistry space because PCA was not able to provide any visual separation of molecules of varying toxicities. The stepwise simplification and distillation of the reasons of activity results in the final heart shaped decision space, which contains the most toxic chemicals at left side. Toxicity of the molecules reduces smoothly towards right.



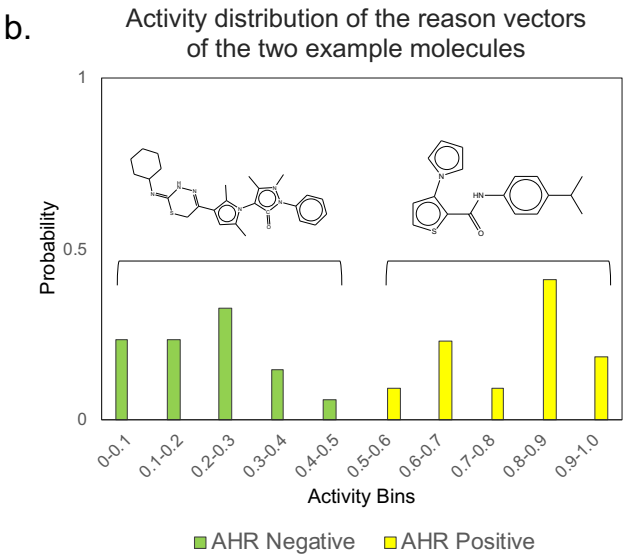


Figure 5. (a) Depiction of chemistry, reason vector and decision vector spaces for the aryl hydrocarbon receptor dataset. Active and inactive vectors are shown with red and green color respectively. Two example molecules are placed in the vector spaces, yellow dots were used for the active and green for the inactive molecule respectively. (b) Predicted activity distribution of the reason vectors for the two example molecules.

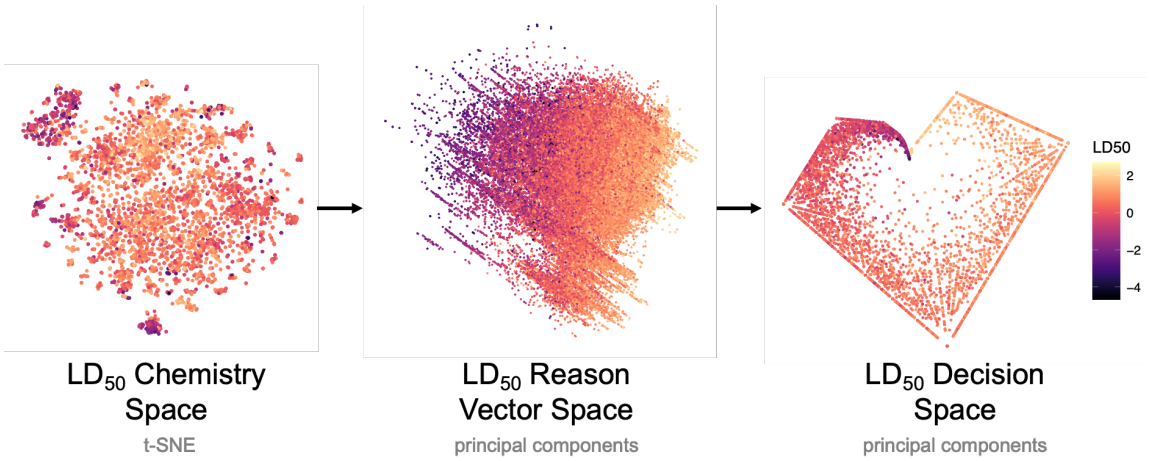


Figure 6. Depiction of chemistry, reason vector and decision spaces for the LD50 activity domain. Higher toxicity (lower LD50 values) is depicted using darker shade of color.

Using reason vectors to identify biologically relevant substructures: Although the reason vectors are higher-level abstract representations, they can be mapped back to the structural features of the query molecules, allowing identification of biologically relevant substructures or

'biophores'. Also, this can be done during test time as opposed to traditional techniques in which the biophores have to be extracted during the model building phase. This allows identification of novel biophores that do not exist in the training set. As shown in Figure 7, the mapping method consists of annotating the atoms of the test molecule with the corresponding activity values at a particular depth of reason vectors. The chosen depth can be varied to observe the change in the biophores, presenting a dynamic picture of the underlying mechanism in terms of relevant substructures.

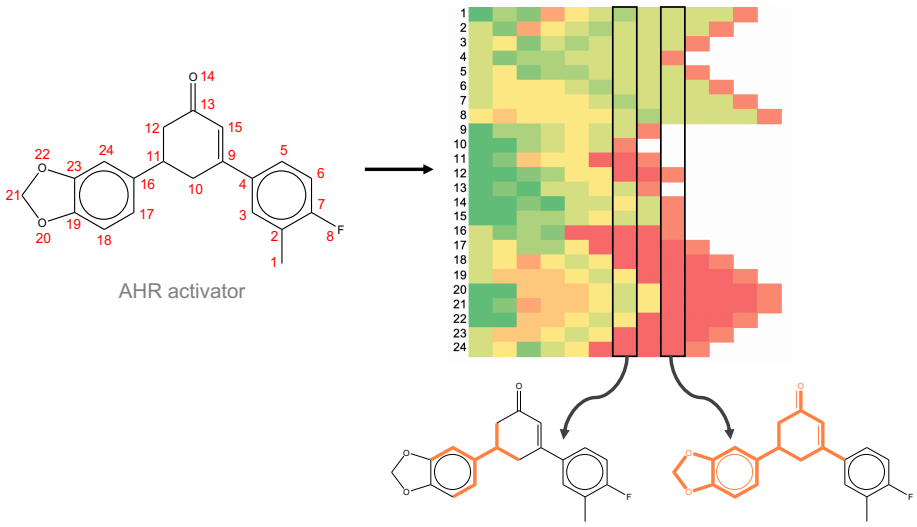


Figure 7. Mapping of biologically relevant substructures (biophores) in an example AHR activator. The biophores are highlighted on the molecule. Note that as the chosen depth is increased, the biophore also expands to cover larger part of the query molecule.

Reason vectors account for structural environments and interactions between different chemical features dynamically in the test molecule, as a result more sensitive to subtle changes in the query structure. This is illustrated in Figure 8 with the aid of two hypothetical molecules subjected to mutagenicity prediction. The aromatic amino moiety in molecule #1 is flanked by two bulky t-butyl groups, blocking its mutagenic potential. However, when the bulky substituents are removed from the vicinity of the amino group, mutagenic potency should increase. This change is reflected nicely in the reason vectors of the two molecules, as a sizable high activity red patch appeared in the reason vectors in Molecule #2. It is worth noting that the two molecules showed negligible difference when evaluated using near neighbors in the chemistry space, both were predicted inactive.

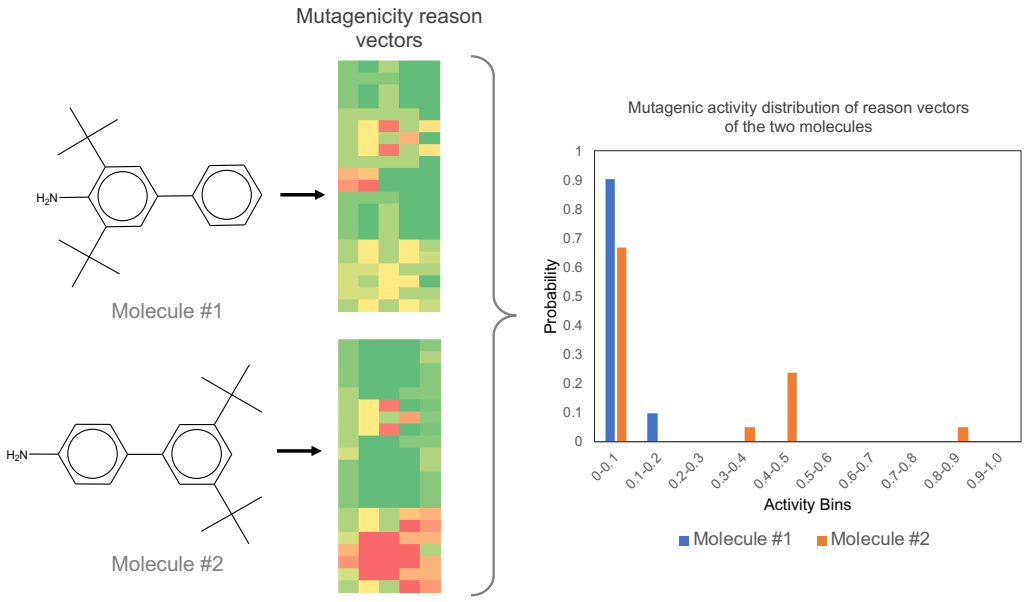


Figure 8. Change in the activity distribution of the reason vectors as a result of change of relative position of functional groups in the two hypothetical query molecules predicted for mutagenicity.

Bioactivity prediction performance of reason vectors: We envision reasoning and causality perception to be the main function of the reason vectors. Consequently, primary objective of this paper is to develop the concept of the reason vectors and not to focus entirely on prediction metrics. Nevertheless, it is important to check if they have acceptable ability to predict biological activity of new chemicals, else their practical applications will certainly be limited. As mentioned in the methods, we used a few standard methods for comparison, i.e. *k-nn* using binary and distributed fingerprints and ECFP fragment-based regression. The results are presented in Figure 9 and the external test metrics are given in Table 5 and 6 separately.

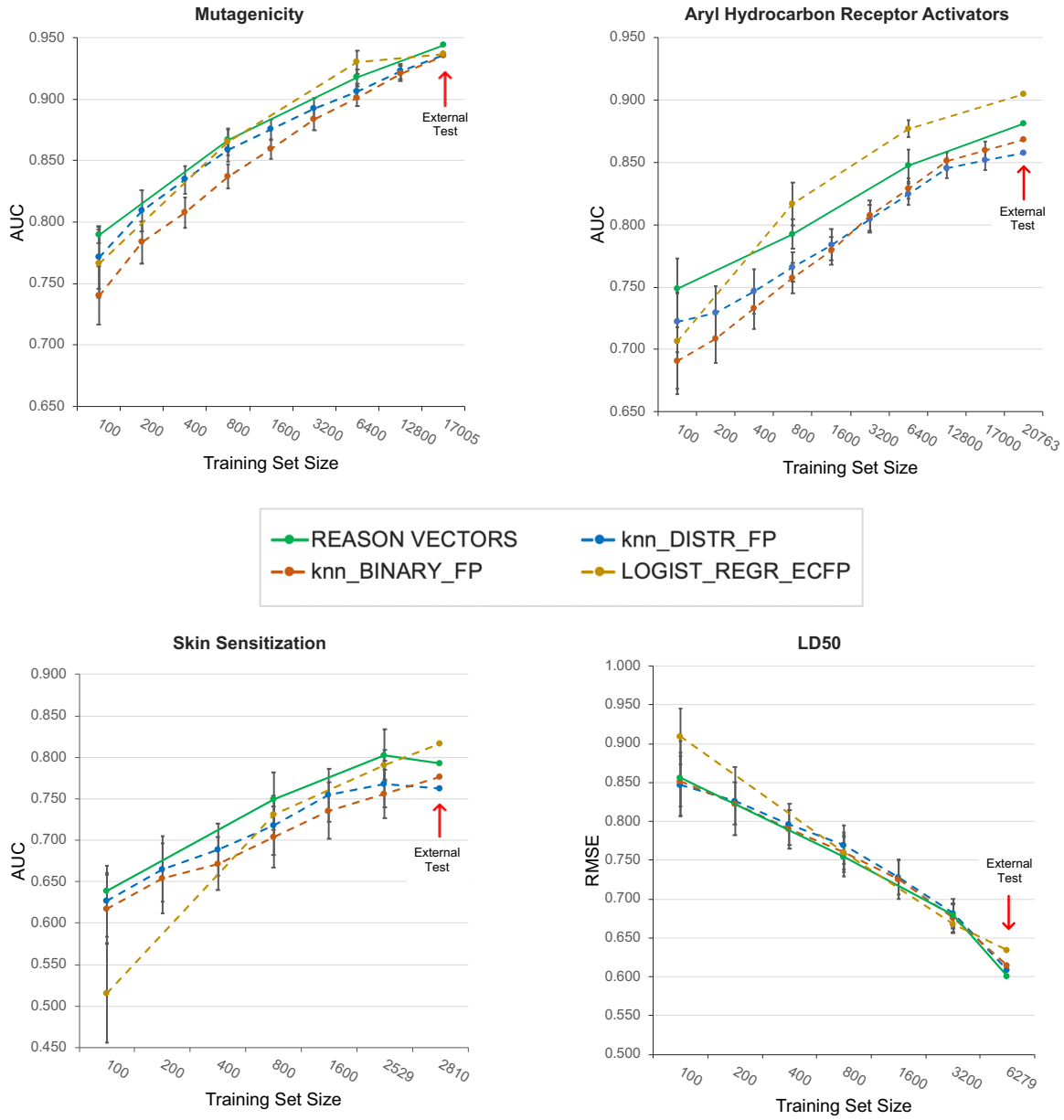


Figure 9. Prediction performance plots for cross-validations and external tests. Note that the external test was performed only once (pointed by the red arrow on right end of every plot), while the cross-validations were repeated multiple times for every training set size. The error bars indicate the standard deviation of trials for different training sets. Also note that the REASON_VECTORS and the LOGIST_REGR_ECFP validations' training set sizes are not as many as the knn methods, for being computationally expensive. For the cross-validations, the test set size was kept at 2000, 2000, 281 and 1000 for mutagenicity, AHR, skin sensitization and LD50 respectively.

The reason vectors performed quite well in cross-validations as well as in the external tests. In the cross-validations, it consistently outperformed the *k-nn* methods in all the datasets and almost for all training set sizes. Performance increased consistently with the training data set size for all methods. The LOGIST_REGR_ECFP performed best in the AHR dataset while the reason vectors were the top performer in the skin sensitization cross-validations. In the external tests, reason vectors gave the best performance for mutagenicity and LD50 and second best in AHR and skin sensitization (Table 5 and 6). We do not think that the external tests are the best indicators of performance, mainly because they were performed only once. On the other hand, cross-validations were repeated several times with multiple combinations of train-test sets.

We have recently published prediction results using a LSTM deep learning model for this mutagenicity dataset and an AUC of 0.938 was achieved for the same external set. In comparison, the reason vectors produced a slightly better AUC of 0.944.

The prediction performance of this LD50 external set was also reported by others using a variety of modeling techniques. For example, Gadaleta *et al* reported r^2 and RMSE of 0.590 and 0.585 respectively using random forests. In comparison, the reason vectors produced a slightly lower r^2 and RMSE of 0.554 and 0.601 respectively. However, it should be noted that Gadaleta *et al*'s results include enforcement of applicability domain, resulting in a coverage of about 91% of the test chemicals. Whereas, the reason vector methodology includes 100% of the test chemicals and therefore, a slight decrease in performance is expected.

In summary, these results support the notion that the reason vectors are not deficient in terms of prediction performance and works well for both binary and continuous activity outcomes.

Table 5. External set prediction results in terms of ROC-AUC and RMSE.

methodology	AMES	AHR	SKIN_SENS	LD50
	ROC-AUC *higher is better*			RMSE *lower is better*
REASON_VECTORS	**0.944**	0.881	0.793	**0.601**
knn_DISTR_FP	0.936	0.857	0.763	0.608
knn_BINARY_FP	0.935	0.868	0.777	0.615
LOGIST_REGR_ECFP	0.937	**0.905**	**0.816**	0.634

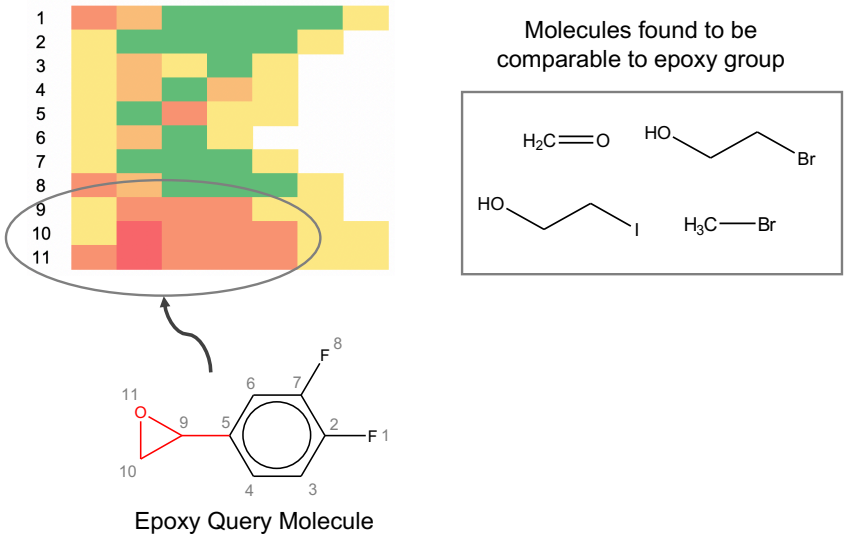
Table 6. External set prediction results in terms of sensitivity, specificity and r^2 for REASON_VECTORS and the LOGIST_REGR_ECFP.

dataset	reason vectors			ECFP logistic regression		
	sensitivity	specificity	accuracy	sensitivity	specificity	accuracy
AMES	85.25	89.09	**87.17**	81.97	91.32	86.75
AHR	83.25	77.50	80.37	85.09	81.86	**83.48**
SKIN_SENS	72.97	72.14	72.56	72.97	77.61	**75.29**
LD50	r^2 = **0.554**			r^2 = 0.506		

Prediction of chemical classes that are absent in training data: The validation experiments described above were based on test sets by random splitting of the data. As a convention, the models are only expected to successfully test chemicals that are well represented in the training set. Present study offers a possibility to overcome this limitation due to two reasons: i. the use of a distributed, continuous representation of molecular chemistry and ii. the reason vectors encode biological activity of parts of the molecules. The first point makes it possible to identify chemicals in the training data that have 'similar' chemistry in spite of not of being identical to that of the query chemical. The latter enables activity prediction of small chunks of query molecules, reducing chances of a complete failure if the whole query chemical is not represented well. Therefore, we checked if we can correctly identify biologically relevant core features of classes of query chemicals for which no examples are present in the training set.

The mutagenicity dataset was utilized to explore this idea because it contains many well-known chemical classes with known mechanism of actions, e.g. nitroso, aromatic nitro, alkyl halides, aromatic amines etc. First, a training set was created by removing all instances of compounds of the chemical class in question. Second, a query chemical of the excluded class was tested using the reason vector methodology in an attempt to identify biophores. We evaluated two mutagenic classes: aromatic nitro and reactive three membered ring compounds (e.g. epoxides, aziridines). In the results presented in Figure 10a and 10b, it can be seen that the appropriate mutagenic functionality was identified in both instances, shown as elevated activity zones in the reason vectors. Next we examined the underlying training compounds that were utilized in the absence of matching examples to ensure if results actually agree with our chemical intuition. We found that for epoxides, some alkyl halides were identified as similar chemicals. Both epoxides and alkyl halides cause mutagenicity by alkylation, therefore the results are not wrong. For the aromatic nitro class, hydroxylamines, azo compounds, aromatic amines and azoxy compounds were identified as replacements. Aromatic nitro group causes mutagenicity via metabolic reduction to hydroxylamine and amines. Similarly, azo compounds are metabolically reduced to amines and hydroxylamines. The results indicate that the distributed fingerprints indeed encode notions of chemistry and the reason vectors are suitable candidates for identifying causes of activity in query molecules that are not well represented in the training data.

a.



b.

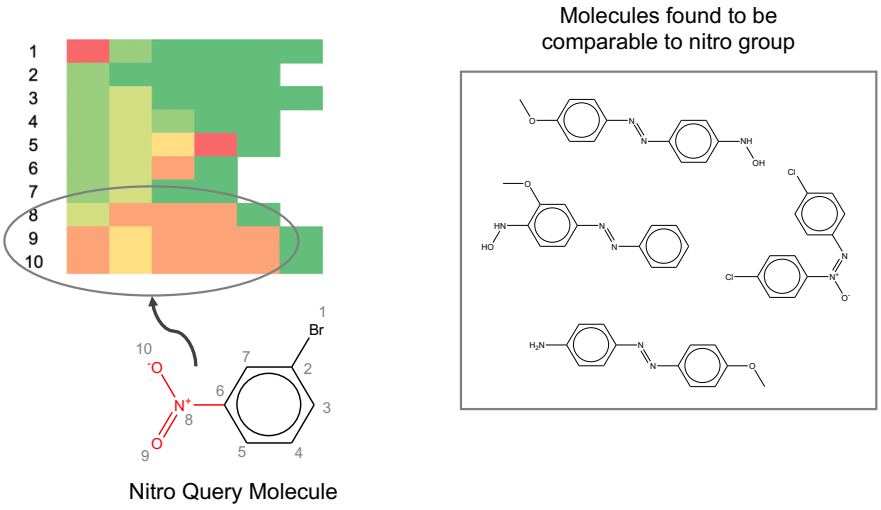


Figure 10. In the absence of any matching training examples, mutagenicity assessment of (a) an epoxy and (b) an aromatic nitro molecule using reason vectors. The epoxy and aromatic nitro 'biophores' were correctly identified by the reason vectors. The molecules shown in the boxes were utilized while forming the reason vectors as similar to epoxy or aromatic nitro functionality.

Conclusions

In this paper we describe the *reason vectors* which are high-level abstract representations of interaction between chemical features and a biological system. These vectors representations are produced by a series of near neighbor searches using a list of sequentially grown atom-centered substructures. They are much simpler than raw chemical fingerprints and are closer to the underlying causality. Evidence was presented demonstrating the reason vectors' powerful generalizing ability, i.e. a vector obtained from a particular bioactivity domain can be used for finding chemicals with similar behavior in a different domain. Although initially produced from raw input data, they represent general concepts independent of any particular bioactivity domain or chemical space. They are able to handle novel combination of features in the query molecule that are not explicitly represented in training data. We also showed that they are able to evaluate classes of chemicals never seen by the training set. They perform very well in bioactivity prediction of molecules and able to predict both binary and continuous activity outcomes. We

believe that this work is a step forward for towards making computational reasoning and causality as an integral part of the QSAR modeling.

Acknowledgments

The author is thankful to his colleagues Dr. Roustem D. Saiakhov, Gianna Cioffi, Mounika Girireddy and Sai Radha Mani Alla for reading the manuscript and offering useful suggestions for improvement.

References

1. Hansch, C., and Fujita, T. p-σ-π analysis. A method for the correlation of biological activity and chemical structure. J. Am. Chem. Soc. 1964, 86, 1616–1626.
2. Hansch, C., and Leo. A. Substituent Constants for Correlation Analysis in Chemistry and Biology, 1979, New York, John Wiley & Sons.
3. A. Lusci, G. Pollastri and P. Baldi, Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules, J. Chem. Inf. Model., 2013, 53, 1563–1575.
4. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2017, 9(2): 513-530.
5. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M, Palmer A, Settels V, Jaakkola T, Jensen K, Barzilay R. Analyzing Learned Molecular Representations for Property Prediction. J Chem Inf Model. 2019, 59(8):3370-3388.
6. Winter R, Montanari F, Noé F, Clevert DA. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem Sci. 2018, 10(6):1692-1701.
7. Jason Jo, Yoshua Bengio, Measuring the tendency of CNNs to Learn Surface Statistical Regularities, 2017, arXiv:1711.11561.
8. Leon Bottou, From Machine Learning to Machine Reasoning, 2011, arXiv:1102.1808.
9. Barber C, Amberg A, Custer L, Dobo KL, Glowienke S, Van Gompel J, Gutsell S, Harvey J, Honma M, Kenyon MO, Kruhlak N, Muster W, Stavitskaya L, Teasdale A, Vessey J, Wichard J. Establishing best practise in the application of expert review of mutagenicity under ICH M7. Regul Toxicol Pharmacol. 2015, 73(1):367-77.
10. J. Peters, D. Janzing, and B. Schölkopf. Elements of Causal Inference - Foundations and Learning Algorithms. 2017, MIT Press, Cambridge, MA, USA.
11. Bernhard Schölkopf, Causality for Machine Learning, 2019, arXiv:1911.10500.

12. J. Pearl. Causality: Models, Reasoning, and Inference, 2nd. 2009, Cambridge University Press, New York, NY.
13. Pearl, Judea and Mackenzie, Dana, The Book of Why: The New Science of Cause and Effect, 2018, Basic Books, Inc., ISBN:978-0-465-09760-9
14. Kahneman, Daniel. Thinking, fast and slow. 2011, New York: Farrar, Straus And Giroux.
15. J. Pearl, The new science of cause and effect, with reflections on data science and artificial intelligence, 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, p-4.
16. Yoshua Bengio, Deep Learning of Representations: Looking Forward, 2 May 2013, arXiv:1305.0445.
17. Hinton GE. Learning multiple layers of representation. Trends Cogn Sci. 2007, 11(10):428-34. Review.
18. Yoshua Bengio, Aaron Courville, Pascal Vincent, Representation Learning: A Review and New Perspectives, 2012, arXiv:1206.5538.
19. Yoshua Bengio, The Consciousness Prior, 2017, arXiv:1709.08568.
20. Valentin Thomas and Jules Pondard and Emmanuel Bengio and Marc Sarfati and Philippe Beaudoin and Marie-Jean Meurs and Joelle Pineau and Doina Precup and Yoshua Bengio, Independently Controllable Factors, 2017, arXiv:1708.01289.
21. Gadaleta, D., Vuković, K., Toma, C. et al. SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data. J Cheminform. 2019, 11, 58.
22. Alves VM, Capuzzi SJ, Muratov E, Braga RC, Thornton T, Fourches D, Strickland J, Kleinstreuer N, Andrade CH, Tropsha A. QSAR models of human data can enrich or replace LLNA testing for human skin sensitization. Green Chem. 2016,18(24):6501-6515.
23. Basketter DA, Selbie E, Scholes EW, Lees D, Kimber I, Botham PA. Food Chem Toxicol. Results with OECD recommended positive control sensitizers in the maximization, buehler and local lymph node assays, 1993; 31(1):63-7.
24. Cronin M. T., Basketter D. A, Multivariate QSAR analysis of a skin sensitization database, SAR QSAR Environ. Res. 1994, 2(3): 159-179.
25. https://www.echemportal.org/echemportal/ and https://echa.europa.eu/cs/information-on-chemicals/registered-substances.
26. National Center for Biotechnology Information. PubChem Database. Source=The Scripps Research Institute Molecular Screening Center, AID=2796, https://pubchem.ncbi.nlm.nih.gov/bioassay/2796 (accessed on May 7, 2020).
27. Suman K. Chakravarti and Sai Radha Mani Alla; Descriptor Free QSAR Modeling Using Deep Learning with Long Short-Term Memory Neural Networks, Frontiers, August 22nd, 2019 DOI: 10.3389/frai.2019.00017.
28. Computing similarity between structural environments of mutagenicity alerts, Chakravarti, S.K., Saiakhov, R. D.; Mutagenesis, 2018, DOI: https://doi.org/10.1093/mutage/gey032.

29. Honma, M., Kitazawa, A., Cayley, A., Williams, R.V., Barber, C., Hanser, T., et al. Improvement of quantitative structure-activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project. Mutagenesis. 2019, 34(1): 3-16.

30. Chakravarti S.K.; Distributed Representation of Chemical Fragments; ACS Omega. 2018, 3(3): 2825-2836.

31. Jaeger S, Fulle S, Turk S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. J Chem Inf Model. 2018, 58(1):27-35.

32. Mikolov, T., Chen, K. Corrado, G., Dean, J. Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR. 2013.

33. Zheng, W. and Tropsha, A. Novel variable selection quantitative structure-property relationship approach based on the k- nearest-neighbour principle. J. Chem. Inf. Comput. Sci. 2000, 40, 185-194.

34. Řehuřek, R.; Sojka, P. Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks; ELRA: Valletta, Malta, 2010, 45–50.

35. Krijthe, J. H. Rtsne: T-distributed stochastic neighbor embedding using a Barnes-Hut implementation, 2015, URL: https://github.com/jkrijthe/Rtsne.

36. Charles Fefferman and Sanjoy Mitter and Hariharan Narayanan, Testing the Manifold Hypothesis, 2013, arXiv:1310.0425.

37. Vikas Verma and Alex Lamb and Christopher Beckham and Amir Najafi and Ioannis Mitliagkas and Aaron Courville and David Lopez-Paz and Yoshua Bengio, Manifold Mixup: Better Representations by Interpolating Hidden States, 2018, arXiv:1806.05236.