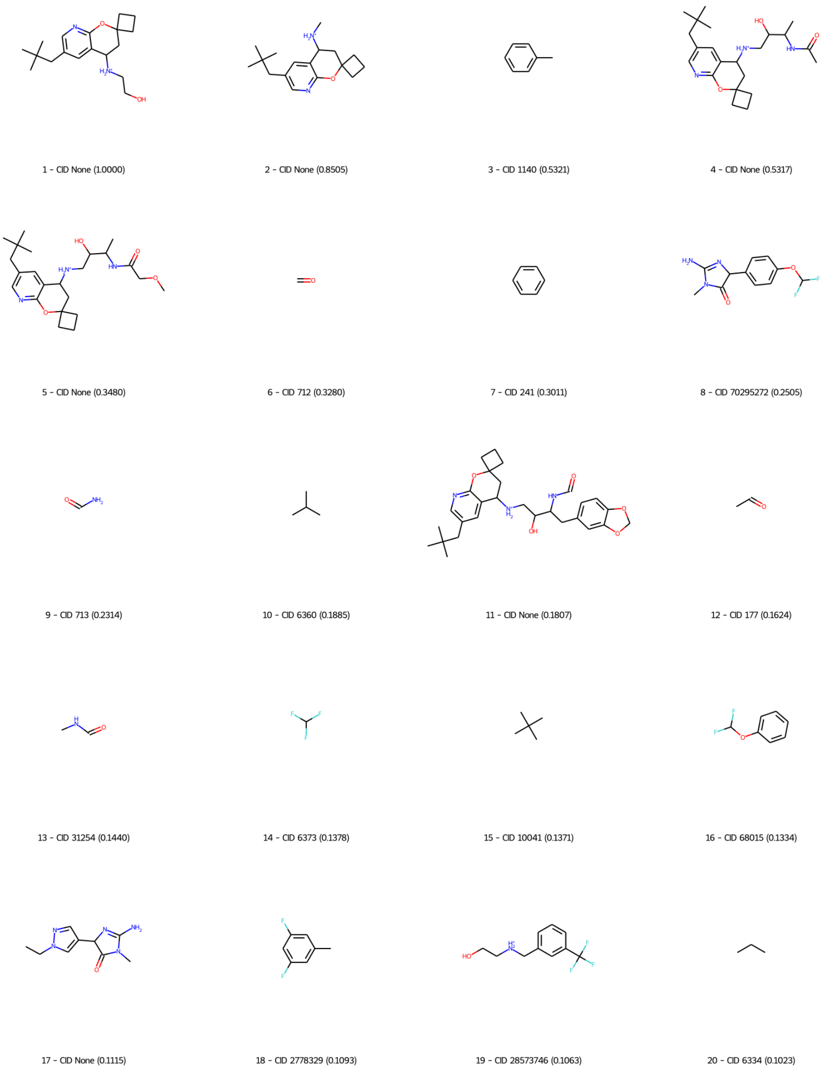
Appendix A Top 20 Fragments



1 - CID None (1.0000)

2 - CID None (0.8505)

3 - CID 1140 (0.5321)

4 - CID None (0.5317)

5 - CID None (0.3480)

6 - CID 712 (0.3280)

7 - CID 241 (0.3011)

8 - CID 70295272 (0.2505)

9 - CID 713 (0.2314)

10 - CID 6360 (0.1885)

11 - CID None (0.1807)

12 - CID 177 (0.1624)

13 - CID 31254 (0.1440)

14 - CID 6373 (0.1378)

15 - CID 10041 (0.1371)

16 - CID 68015 (0.1334)

17 - CID None (0.1115)

18 - CID 2778329 (0.1093)

19 - CID 28573746 (0.1063)

20 - CID 6334 (0.1023)

Fig. A1 Top 20 fragments for the BACE Classification task.

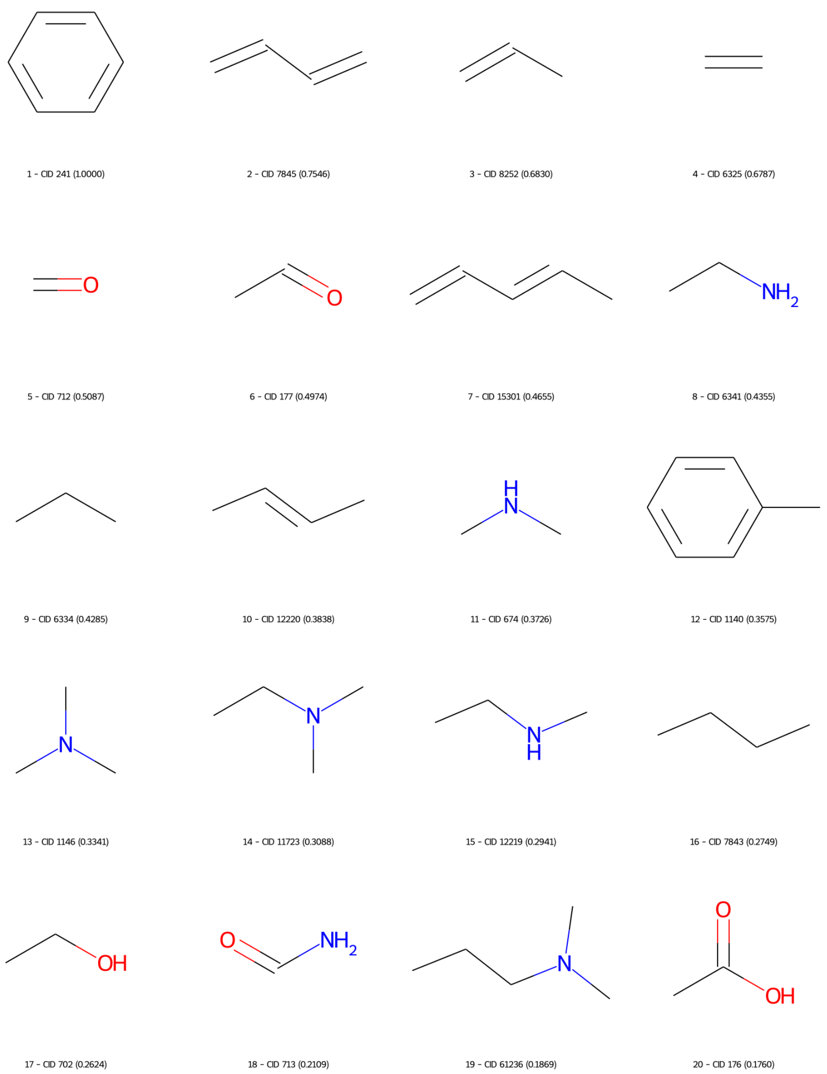
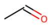
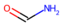


Fig. A2 Top 20 fragments for the BBBP task.

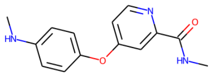


1 - CID 177 (1.0000)

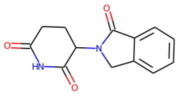
2 - CID 8252 (0.8864)



3 - CID 713 (0.8436)



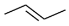
4 - CID 54909138 (0.7412)



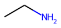
5 - CID 91585 (0.5978)

6 - CID 6325 (0.4544)

7 - CID 6334 (0.3743)

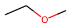


8 - CID 12220 (0.3203)



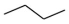
9 - CID 6341 (0.2737)

10 - CID 6360 (0.1937)

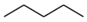


11 - CID 10903 (0.1527)

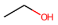
12 - CID 8254 (0.1471)



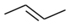
13 - CID 7843 (0.1248)



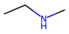
14 - CID 8003 (0.1024)



15 - CID 702 (0.0894)

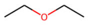


16 - CID 62695 (0.0708)

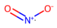


17 - CID 12219 (0.0633)

18 - CID 21585139 (0.0503)

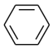


19 - CID 3283 (0.0428)



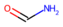
20 - CID 24529 (0.0279)

Fig. A3 Top 20 fragments for the ClinTox task.

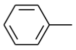


1 - CID 241 (1.0000)

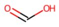
2 - CID 712 (0.7787)



3 - CID 713 (0.6064)



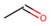
4 - CID 1140 (0.5599)



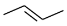
5 - CID 284 (0.5536)

6 - CID 6325 (0.3996)

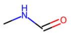
7 - CID 6334 (0.3936)



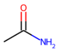
8 - CID 177 (0.3813)



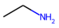
9 - CID 12220 (0.3667)



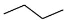
10 - CID 31254 (0.3225)



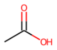
11 - CID 178 (0.3111)



12 - CID 6341 (0.2156)

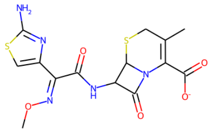


13 - CID 7843 (0.2001)

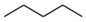


14 - CID 176 (0.1878)

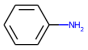
15 - CID 674 (0.1717)



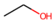
16 - CID 74687419 (0.1552)



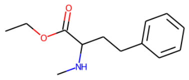
17 - CID 8003 (0.1540)



18 - CID 6115 (0.1511)

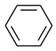


19 - CID 702 (0.1413)

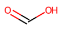


20 - CID 59900860 (0.1378)

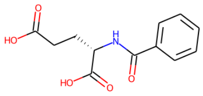
Fig. A4 Top 20 fragments for the SIDER task.



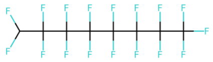
1 - CID 241 (1.0000)



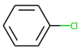
2 - CID 284 (0.3565)



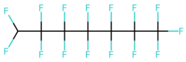
3 - CID 11345724 (0.1913)



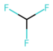
4 - CID 101652 (0.1826)



5 - CID 7964 (0.1652)

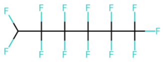


6 - CID 9778 (0.1391)

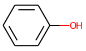


7 - CID 6373 (0.1304)

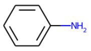
8 - CID 6334 (0.1304)



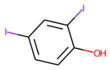
9 - CID 67730 (0.1043)



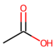
10 - CID 996 (0.1043)



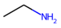
11 - CID 6115 (0.1043)



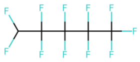
12 - CID 72858 (0.0870)



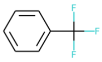
13 - CID 176 (0.0870)



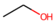
14 - CID 6341 (0.0870)



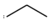
15 - CID 9816874 (0.0783)



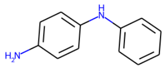
16 - CID 7368 (0.0783)



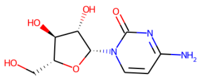
17 - CID 702 (0.0783)



18 - CID 6334 (0.0783)

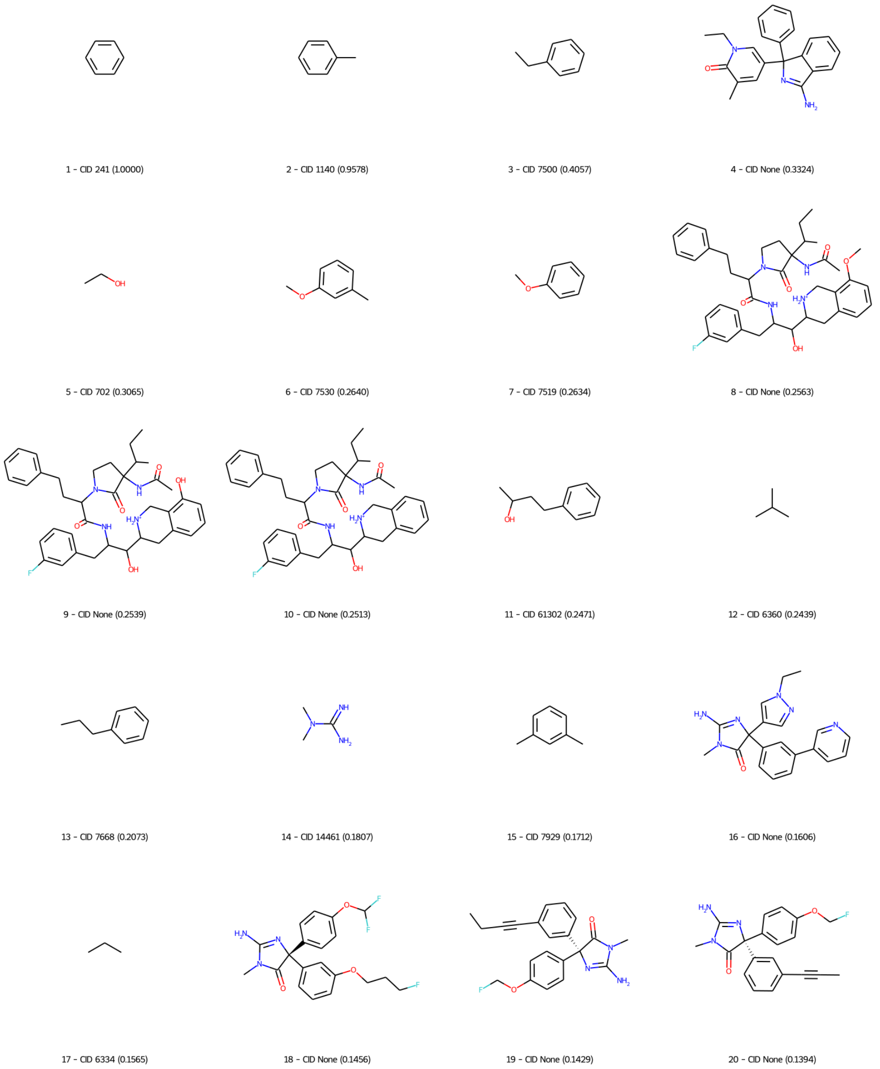


19 - CID 7564 (0.0696)



20 - CID 6253 (0.0696)

Fig. A5 Top 20 fragments for the Tox21 task.



1 - CID 241 (1.0000)

2 - CID 1140 (0.9578)

3 - CID 7500 (0.4057)

4 - CID None (0.3324)

5 - CID 702 (0.3065)

6 - CID 7530 (0.2640)

7 - CID 7519 (0.2634)

8 - CID None (0.2563)

9 - CID None (0.2539)

10 - CID None (0.2513)

11 - CID 61302 (0.2471)

12 - CID 6360 (0.2439)

13 - CID 7668 (0.2073)

14 - CID 14461 (0.1807)

15 - CID 7929 (0.1712)

16 - CID None (0.1606)

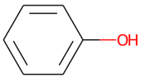
17 - CID 6334 (0.1565)

18 - CID None (0.1456)

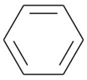
19 - CID None (0.1429)

20 - CID None (0.1394)

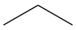
Fig. A6 Top 20 fragments for the BACE Regression task.



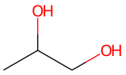
1 - CID 996 (1.0000)



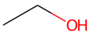
2 - CID 241 (0.9630)



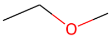
3 - CID 6334 (0.7809)



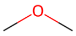
4 - CID 1030 (0.5216)



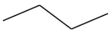
5 - CID 702 (0.4198)



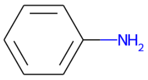
6 - CID 10903 (0.4167)



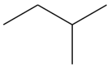
7 - CID 8254 (0.4074)



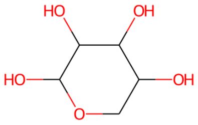
8 - CID 7843 (0.3735)



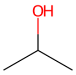
9 - CID 6115 (0.3549)



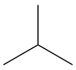
10 - CID 6556 (0.3117)



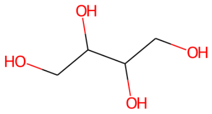
11 - CID 229 (0.3056)



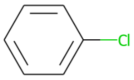
12 - CID 3776 (0.3056)



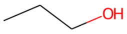
13 - CID 6360 (0.2747)



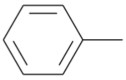
14 - CID 8998 (0.2438)



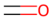
15 - CID 7964 (0.2407)



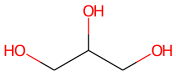
16 - CID 1031 (0.2346)



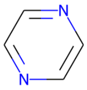
17 - CID 1140 (0.2191)



18 - CID 712 (0.2099)

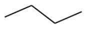


19 - CID 753 (0.2006)

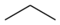


20 - CID 9261 (0.1975)

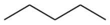
Fig. A7 Top 20 fragments for the ESOL task.



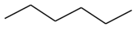
1 – CID 7843 (1.0000)



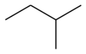
2 – CID 6334 (0.9461)



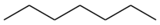
3 – CID 8003 (0.9079)



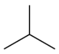
4 – CID 8058 (0.8652)



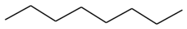
5 – CID 6556 (0.6809)



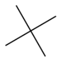
6 – CID 8900 (0.6787)



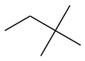
7 – CID 6360 (0.6719)



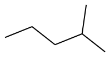
8 – CID 356 (0.4764)



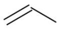
9 – CID 10041 (0.4562)



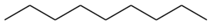
10 – CID 6403 (0.3169)



11 – CID 7892 (0.3079)

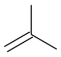


12 – CID 8252 (0.2989)

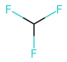


13 – CID 8141 (0.2000)

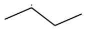
14 – CID 6325 (0.1685)



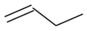
15 – CID 8255 (0.1416)



16 – CID 6373 (0.1191)



17 – CID 7843 (0.0966)



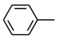
18 – CID 7844 (0.0539)

Fig. A8 Top 20 fragments for the FreeSolv task.

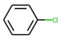
1 - CID 241 (1.0000)

2 - CID 713 (0.5263)

3 - CID 712 (0.5078)

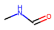


4 - CID 1140 (0.4339)



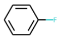
5 - CID 7964 (0.2120)

6 - CID 177 (0.2112)



7 - CID 31254 (0.1782)

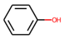
8 - CID 6334 (0.1534)



9 - CID 10008 (0.1452)

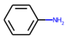
10 - CID 1049 (0.1238)

11 - CID 178 (0.1191)

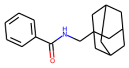


12 - CID 996 (0.1127)

13 - CID 6373 (0.1089)

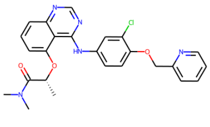


14 - CID 6115 (0.1078)

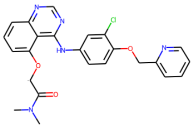


15 - CID 3452818 (0.1019)

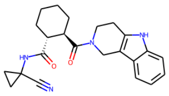
16 - CID 6341 (0.0909)



17 - CID 11569158 (0.0890)



18 - CID None (0.0861)



19 - CID 25142173 (0.0859)

20 - CID 674 (0.0838)

Fig. A9 Top 20 fragments for the Lipophilicity task.

Appendix B Model Details

```
Input: Sentence hidden vectors $S = \{\mathbf{h_1}, \mathbf{h_2}, \ldots, \mathbf{h_N}\}$, beginning index $b$ and ending index $e$
Output: root node $R_{S[b:e]}$ of sequence $S[b:e]$
  procedure BUILD($S, b, e$)
    $R \leftarrow nil$
    if $e = b$ then
      $R \leftarrow$ new Node
      $R.index \leftarrow b$
      $R.left, R.right \leftarrow nil, nil$
    else if $e > b$ then
      $R \leftarrow$ new Node
      $R.index \leftarrow argmax_{i=b}^{e} Score(\mathbf{h_i})$
      $R.left \leftarrow$ BUILD($S, b, R.index - 1$)
      $R.right \leftarrow$ BUILD($S, R.index + 1, e$)
    end if
    return $R$
  end procedure
```

Fig. B10 Pseudo-code of the model architecture.

Appendix C Interpretation Details

C.1 BACE Classification Analysis (cont'd)

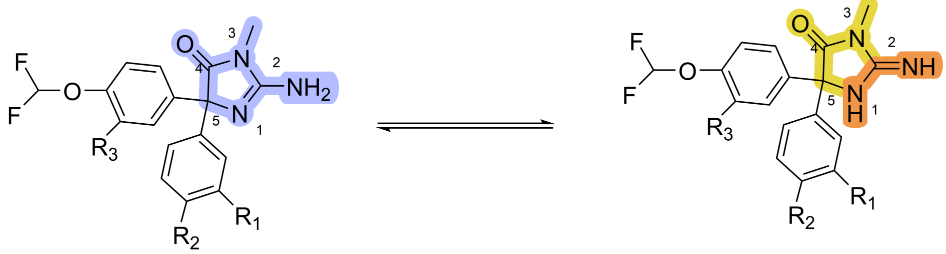


Fig. C11 The scheme of this imine-amine tautomerism. The aminohydantoin core is shown in purple and its iminohydantoin tautomeric form shown in mustard yellow and orange.

C.2 BBBP Analysis

The fragments extracted by the workflow from this task **Figure A2** tend to be smaller chemical groups, which necessarily may not be in direct relation with the BBB penetrability of the larger parent molecule, from which the fragment was extracted. In other words, these specific smaller fragments may not directly improve or worsen the molecule's ability to pass the BBB, as in some cases that specific fragment may decrease lipophilicity and in turn decrease BBBP, while that same fragment found on a different molecular skeleton may increase lipophilicity and BBBP. For this reason, and seeing as the given fragments are not large enough chemical moieties to appoint accurate qualities to a specific fragment in terms of being able to change the main molecule's BBBP, it would be rational to investigate the lipophilicity of each molecule containing a specific fragment according to that molecule only and refrain from making broader assessments based on the fragments.

Fragments 1 and 12 (benzene and toluene moieties) are known to increase the lipophilicity of molecules based on fundamental chemical knowledge. The same is true for tertiary amines (fragments 13, 14, and 19), as a decrease in the number of hydrogen atoms bound to an amine's nitrogen atom leads to a decrease in that amine's ability to form H-bonds, which results in a decrease in lipophilicity, sequentially putting tertiary amines above secondary amines and above primary amines in the lipophilicity hierarchy. However, this order is not followed here as 15 (a secondary amine) comes after 14 (a tertiary amine). This is due to the fact that the workflow used here is solely based on methodological approaches aimed at making the workflow extract the most repeated meaningful words, and is not based on chemical rules and features of each dataset. Moreover, fragments 17, 18, and 20 are undesirable fragments when designing molecules that cross the BBB, because, they typically decrease rather than increase a molecule's lipophilicity due to their H-bond forming abilities [55].

C.3 ClinTox Analysis

Drugs and drug candidates may fail to pass clinical trials due to toxicity, which can be caused by the formation of toxic metabolites after biotransformation. Certain functional groups, such as nitro groups, aromatic amines, and polyhalogenated groups, may contribute to the production of these toxic metabolites. Some fragments found by the workflow are marked on the drug molecules in **Figure C12**. When bound to a molecule, the fragment 20 (`O=[N+][O-]`) in **Figure A3** represents the nitro group. Although the nitro group is found in many active molecules, it is toxic and is frequently classified as a structural pharmacophore and/or toxicophore group. Many studies on molecules containing nitro groups in the literature show toxicity issues such as carcinogenicity, hepatotoxicity, mutagenicity, and bone marrow suppression [56]. Nitro radical anion, nitrozo derivative, nitroxyl radical, hydroxylamine and primary amine derivatives are formed in the biotransformation process by reduction of the aromatic nitro group [57]. Studies have revealed that intermediate products are responsible for toxicity. Especially hydroxylamine derivatives are responsible for methemoglobinemia, while other intermediates have shown mutagenicity and carcinogenicity [58]. **Figure C13** depicts the mechanism of aromatic nitro molecule reductive biotransformation.

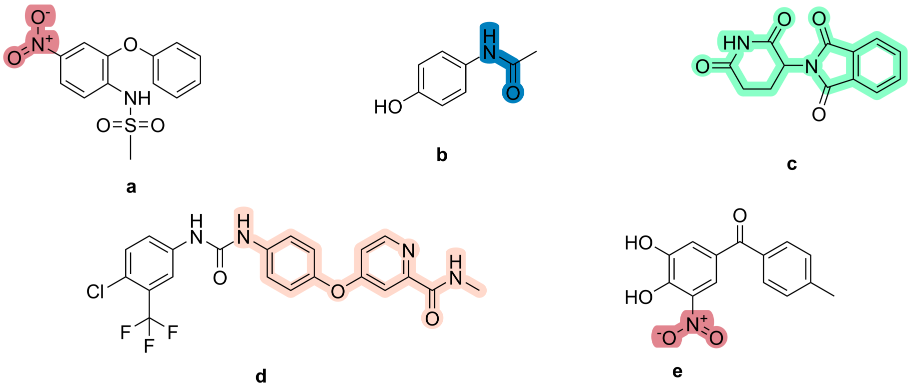


Fig. C12 2D representation of nimesulide **(a)**, acetaminophen **(b)**, thalidomide **(c)**, sorafenib **(d)**, and tolcapone **(e)**. The fragments found by the workflow are labelled on the drug molecules.

Nimesulide (4-nitro-2-phenoxymethane-sulfo-anilide), shown in **Figure C12a**, is selective cyclooxygenase $(COX) - 2$ inhibitor nonsteroidal anti-inflammatory drug (NSAID) used in the treatment of various inflammatory and pain conditions. Nimesulide's aromatic nitro group undergoes reductive metabolism and shows hepatotoxic effect. As shown in **Figure C13**, nimesulide bioactivation in the liver generates reactive intermediates that bind to macromolecules in the body and causes oxidoreductive stress [59].

The formamide structure is represented by fragment 3 (`NC=O`) discovered by the workflow. Intermediates formed as a result of N-oxidation of aromatic amides show carcinogenic and cytotoxic effects by covalent bonding with biological macromolecules as in aromatic amines. Acetaminophen **Figure C12b** is an analgesic. The hepatoxic effect of the compound is due to N-acetyl-p-benzoquinonimine (NAPQI), a toxic intermediate formed during its metabolism. Under normal conditions, NAPQI is rapidly neutralised by an antioxidant molecule called glutathione. However, when high doses are taken, glutathione stores are depleted and the accumulation of NAPQI can cause damage to liver cells [60]. **Figure C14** shows the reaction mechanism.

Aldehydes, which constitute a large class of electrophilic carbonyl compounds, are toxic structures for the body, although they are important functional groups for drug-active molecules. They react with biologic macromolecules (e.g., protein, nucleic acid) in the body and form adducts. Thus, they cause cytotoxicity, mutagenicity, carcinogenicity, and oxidative stress [61].

Fragment 1 representing the acetaldehyde structure found by the workflow is a genotoxic substance that causes nasopharyngeal cancer upon long-term exposure. This toxicity can be dehydrated first to aminol and then to imines by 1,2-addition reaction with amines. Thus, it has been shown that deoxyguanosine reacts with nucleophiles such as N_2 nitrogen [62]. Moreover, fragment 15, `CCO` represents ethanol structure. The metabolism of ethanol occurs in the liver. Toxic intermediates such as acetaldehyde and reactive oxygen species (ROS) are produced. Acetaldehyde causes damage to liver cells by the mechanism described above. ROS may cause cell death and damage by binding to macromolecules in the body [63].

Even though fragments 2, 6, 8, and 16 could not provide sufficient information about toxicity, it can be said that they have an electrophilic structure due to the double bond they carry. It can be interpreted that these electrophilic structures can bind to nucleophilic structures in the body and develop toxic effects. Finally, fragments 4 and 5 found by the workflow show relatively large structures. The compound thalidomide **(c)** and sorafenib **(d)** shown in **Figure C12** carry these ring structures to a large extent. Studies on various toxic metabolites of these drug molecules are available in the literature [64, 65]. Even if no interpretation can be made in terms of metabolism reaction due to the size of the fragments, it can be said that they represent a substructure of drug molecules with toxicity.

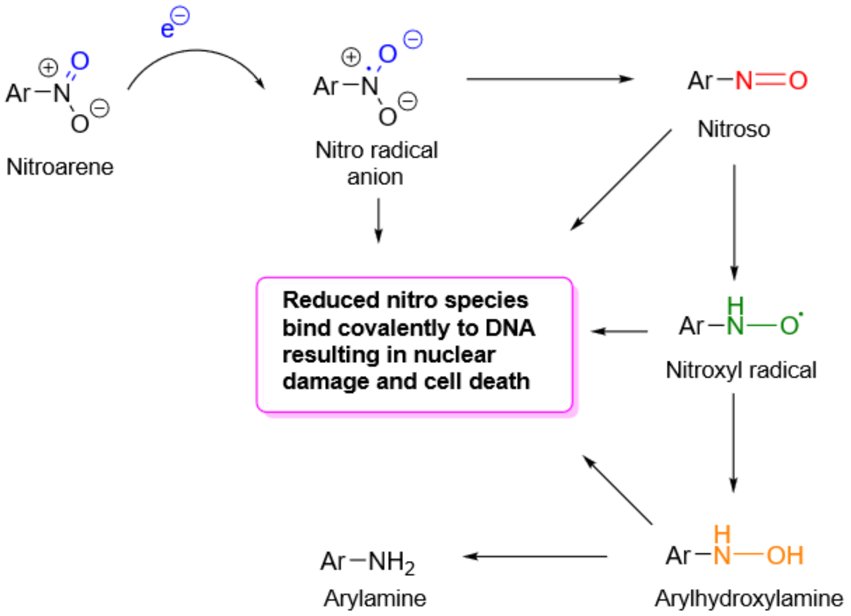

Nitroarene
Nitro radical anion
Nitroso
Reduced nitro species bind covalently to DNA resulting in nuclear damage and cell death
Nitroxyl radical
Arylamine
Arylhydroxylamine

Fig. C13 Reductive biotransformation mechanism of aromatic nitro molecules.

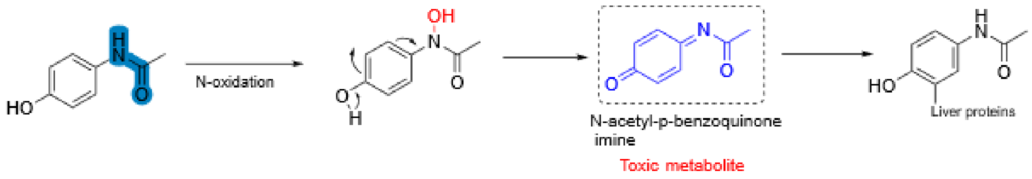


Fig. C14 Reductive biotransformation mechanism of acetaminophen.

C.4 Tox21 Analysis

In **Subsection C.3**, it is mentioned the contribution of polyhalogen groups to toxicity. Organic halogenated molecules should be used with extreme caution. They have a high potential for toxicity when they accumulate in adipose tissue and cause cancer. Because they increase lipophilicity, halogens facilitate passage through the blood-brain barrier [66]. Organic halogenated molecules with high toxicity include thyroxine (**Figure C15a**), teflon (**Figure C15b**), and halothane (**Figure C15c**). The workflow discovered eight halogenated molecules among the top 20 fragments. The high amount of fluorine halogen found in fragments 4, 6, 9, and 15 is said to have toxicity similar to the teflon molecule. Furthermore, fragments 11 and 19 with aromatic amine and amide structures may be toxic via the mechanism depicted in **Figure C14**. Studies have shown that NSAIDs with acidic structure can covalently bind to hepatic proteins.

NSAIDs containing carboxylic acid structure form electrophilic intermediates which are acyl glucuronides and bind to nucleophilic amino acids. Although not proven in vivo, this covalent binding mechanism to these macromolecules has been suggested to play a role in the hepatic toxicity of many NSAIDs [67]. The workflow successfully found the carboxylic acid structure in fragments 2, 3, and 13.

In one study, the halogen substituents of the antifungal drug UK 47265 were modified to find a compound that is less toxic to the liver. In this case, the chlorobenzene fragment in fragment 5 was removed and replaced by the 1,3-dichlorobenzene structure, yielding the less toxic fluconazole molecule [68, 69]. **Figure C16** depicts 2D representations of UK47265 and fluconazole.

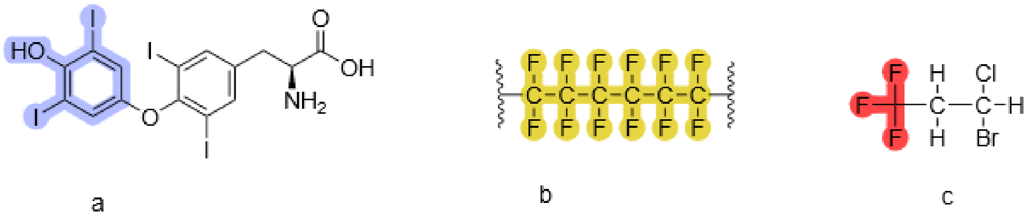


Fig. C15 2D representation of thyroxine **(a)**, teflon **(b)**, and halothane **(c)**. The fragments found by the workflow are marked on the compounds in the figure.

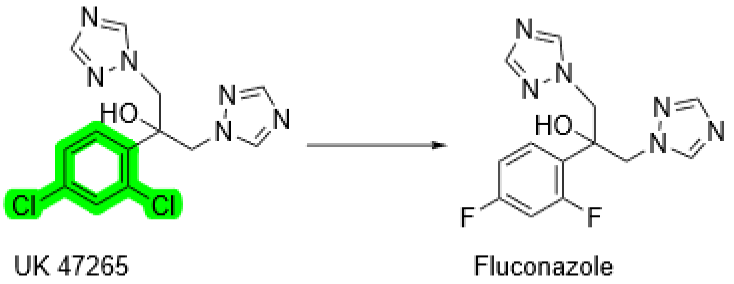


Fig. C16 2D representation of UK47265 and fluconazole. Toxicity was reduced by the diversification of aromatic substituents. The `Clc1ccccc1` fragment found by the workflow is the subfragment of the 1,3-dichlorobenzene structure marked in green on the molecule.

C.5 BACE Regression Analysis

The fragment 14 (`CN(C)C(=N)N`) represents 1,1-dimethylguanidine structure. Guanidine fragments can be found among acyclic guanidine-containing BACE1 inhibitors in the studies reported by Boy et al. [70] and Gerritz et al. [71]. The guanidine fragment extracted by the workflow is slightly different from the guanidine fragments found in these two studies. The guanidine groups in these studies contain a guanidine base that has one alkyl side chain, while the extracted fragment carries two alkyl side chains (dimethyl). Although the extracted fragment is not fully identical with the guanidine bases in these studies, this fragment still denotes an important moiety since acyclic guanidine-containing BACE1 inhibitors comprise a large corpus of the studies focused on discovering BACE1 targeting agents. Moreover, as mentioned in **Subsection 2.2.1**, cyclic guanidines (aminohydantoins, iminohydantoins) are also common structural cores among BACE1 inhibiting compounds [35–37], and fragment 4 may also be pointing to the cyclic guanidine fragment present in iminohydantoins. Altogether, the ability of the workflow to find and highlight the guanidine moiety as an important fragment, which is used widely as a common motif across BACE1 inhibitors, once more marks the workflow's success in finding meaningful chemical fragments. As it can be seen in **Figure A6** (fragments 16, 18, 19, and 20), discussed in **Subsection 2.2.1**, the workflow was also able to extract fragments that carry an aminohydantoin moiety (the aminohydantoin core is shown in purple in **Figure C11**) for the BACE Regression task as well.

C.6 ESOL Analysis

In general, the factors affecting the solubility of a compound in water depend on the polarity of the molecule and the presence of functional groups that can form hydrogen bonds with water molecules [72]. One of the most widely used methods in drug design and development studies is increasing the polarity of a molecule by adding hydrophilic groups (e.g., hydroxyl, amino, carboxylic acid). For example, polar hydroxyl group and polar heterocyclic rings were added to thioconazole compound which is used only in skin effects due to its high lipophilicity. In this way, fluconazole compound whose solubility was improved and which can be used against systemic infections was synthesized [69]. **Figure C17** shows the groups responsible for high polarity in an antibacterial compound. Seven of the top 20 molecules discovered by the workflow contain hydroxyl (OH) groups, while one contains phenol groups (shown in pink and yellow in **Figure C17**). Ether (ROR), amino (NH_2), and carbonyl ($C = O$) structures are found in fragments 6, 7, 9, and 18. The pyrazine ring is represented by fragment 20 discovered by the workflow. Pyrazine ring freely soluble in water.

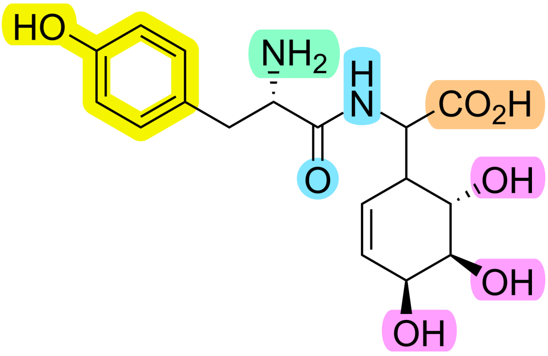


Fig. C17 Functional groups that contribute to polarity are colored on the compound.

C.7 Lipophilicity Analysis

Similar to the BBBP task, in the lipophilicity task also, the extracted fragments in **Figure A9** cannot be said to be in direct relation with increasing the overall lipophilicity of the compound they were extracted from. However, fragments 1, 4, 5, and 9, contain phenyl moieties, which increase the lipophilicity of chemical compounds in general. The trifluoromethyl, fragment 13, is also well known for increasing the lipophilic characteristics of molecules [55]. The presence of the bulky adamantly group in fragment 15 is expected to increase lipophilicity. Fragments 17, 18, and 19 also contain bulky hydrophobic structures, which amplify the molecule's lipophilic properties. Similar to the case in the BBBP task, the workflow is not aimed at rating the fragments in terms of their physicochemical differences, but rather to assist the method in finding the largest meaningful fragments.

Appendix D Dataset Details

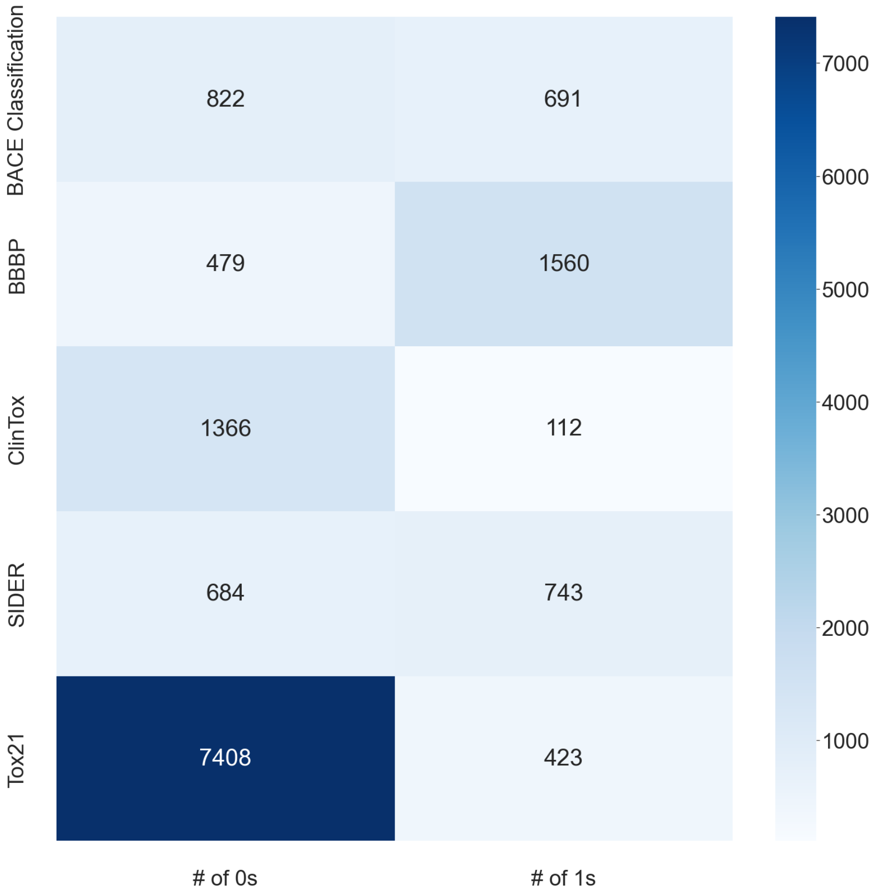


Fig. D18 Balance of classification tasks on heatmap.

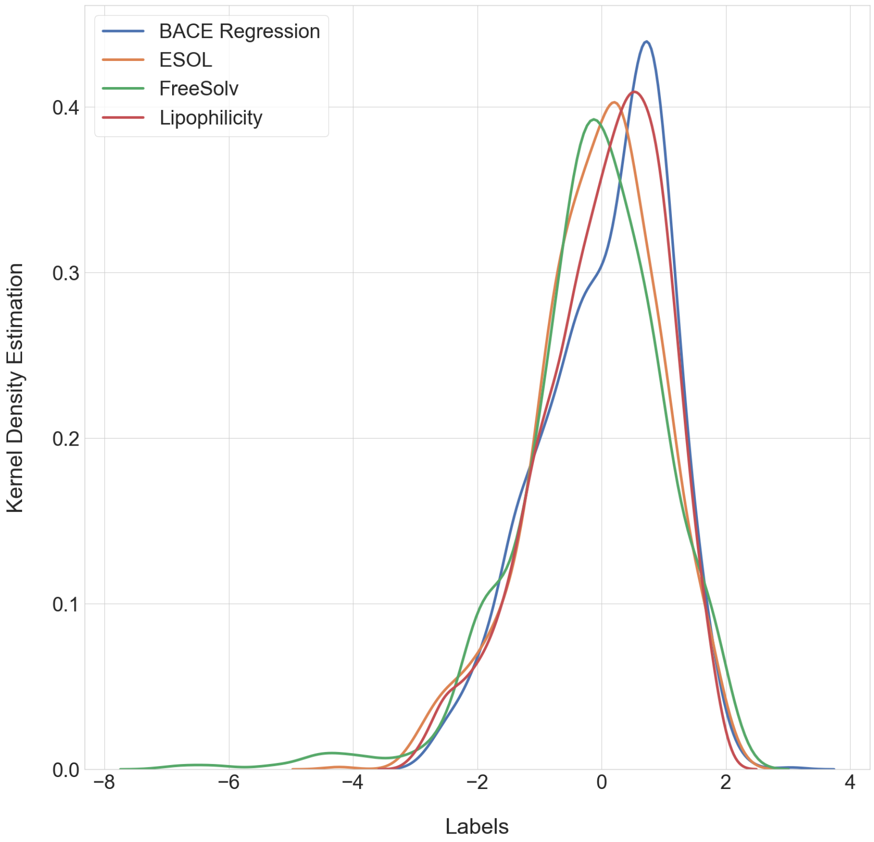


Fig. D19 Balance of regression tasks on KDE plot.

Table D1 Informations about datasets.

Dataset Name	# of Molecules	# of Tasks	Task Type
BACE	1513	1	clf, reg
BBBP	2039	1	clf
ClinTox	1478	2	clf
SIDER	1427	27	clf
Tox21	7831	12	clf
ESOL	1128	1	reg
FreeSolv	642	1	reg
Lipophilicity	4200	1	reg

clf and reg indicate classification and regression, respectively.

Appendix E Best Hyperparameters

Table E2 Best hyperparameters for classification tasks.

Hyperparameter	BACE Clf	BBBP	ClinTox	SIDER	Tox21
optimizer	adadelta	adadelta	adadelta	adagrad	adadelta
dropout ratio	0.3	0.3	0.5	0.9	0.5
# of DNN layers	1	5	5	1	3
# of nrns in DNN	6144	6144	2048	2048	2048
learning rate	1e-3	1e-3	1e-3	1e-3	1e-3
# of nrns in TBL	300	300	1500	500	100
use of batchnorm	1	1	0	0	1
rank of input	h	h	w	w	h
# of epochs	200	200	400	150	100

nrns, DNN and TBL indicate neurons, Dense Neural Network and Tree-Bidirectional-LSTM, respectively.

Table E3 Best hyperparameters for regression tasks.

Hyperparameter	BACE Reg	ESOL	FreeSolv	Lipophilicity
optimizer	adagrad	adagrad	adagrad	adam
dropout ratio	0.8	0.3	0.3	0.3
# of DNN layers	1	3	1	1
# of nrns in DNN	3072	128	1024	512
learning rate	1e-3	1e-3	1e-3	1e-3
# of nrns in TBL	1500	50	1500	500
use of batchnorm	1	0	1	1
rank of input	w	h	h	h
# of epochs	100	200	100	150

nrns, DNN and TBL indicate neurons, Dense Neural Network and Tree-Bidirectional-LSTM, respectively.

Table E4 Tried hyperparameters and their values.

Hyperparameter	Values
optimizer	adadelta, adam, adagrad
dropout ratio	0.1, 0.3, 0.5, 0.7, 0.8, 0.9
# of DNN layers	1, 2, 3, 5, 7, 10
# of nrns in DNN	8, 16, 32, 64, 128, 256, 512, 1024, 2048, 3072, 4096
learning rate	1e-3, 1e-4, 1e-5
# of nrns in TBL	50, 100, 200, 300, 500, 1000, 1500
use of batchnorm	1, 0
rank of input	w, h
# of epochs	500

nrns, DNN and TBL indicate neurons, Dense Neural Network and Tree-Bidirectional-LSTM, respectively.

Appendix F Representation Details

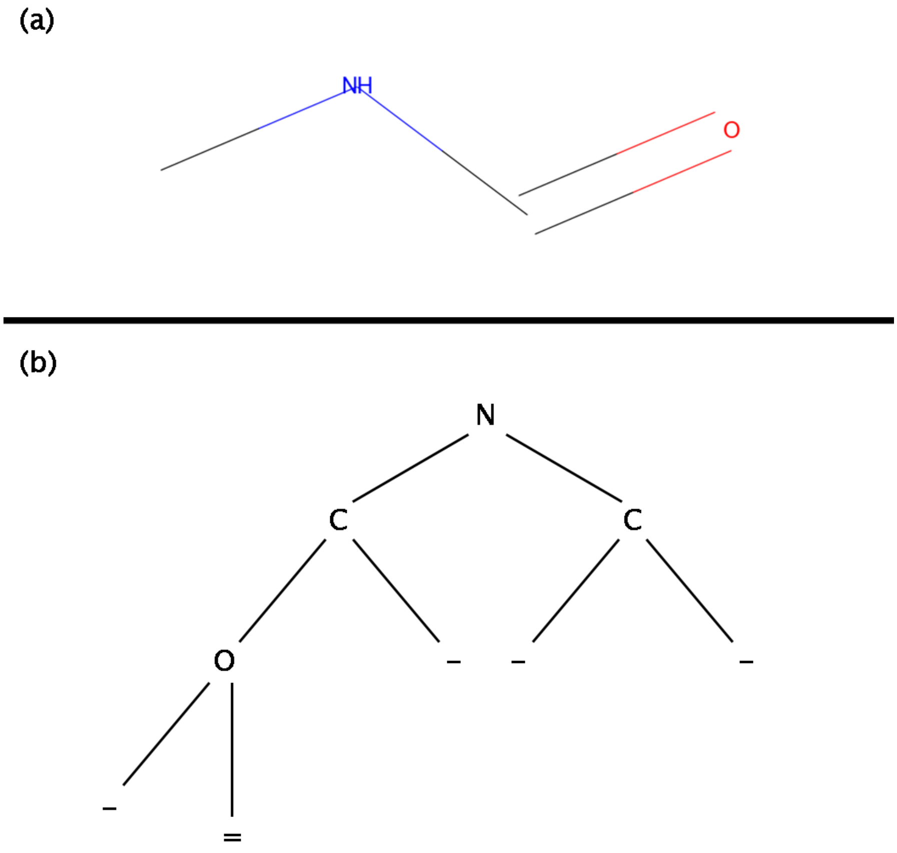


Fig. F20 Tree representation of a molecule.

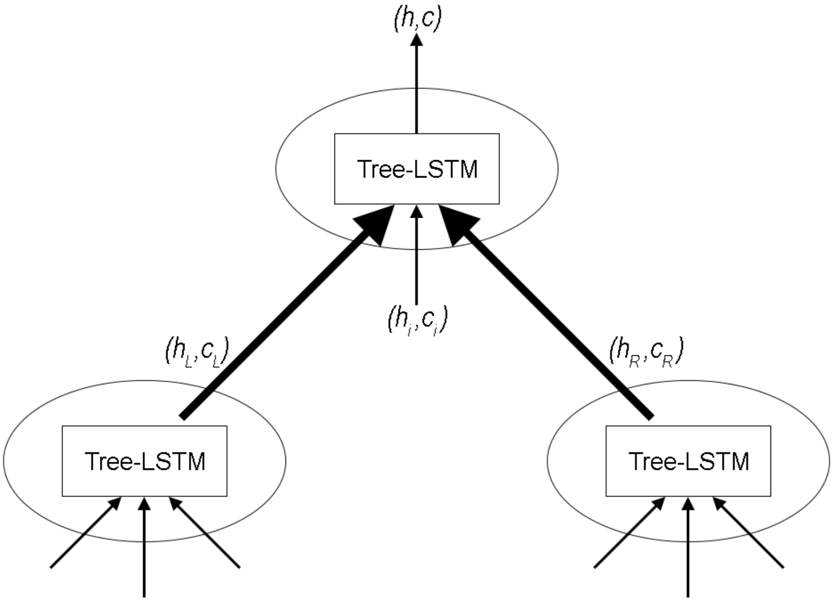


Fig. F21 Bottom-up embedding.

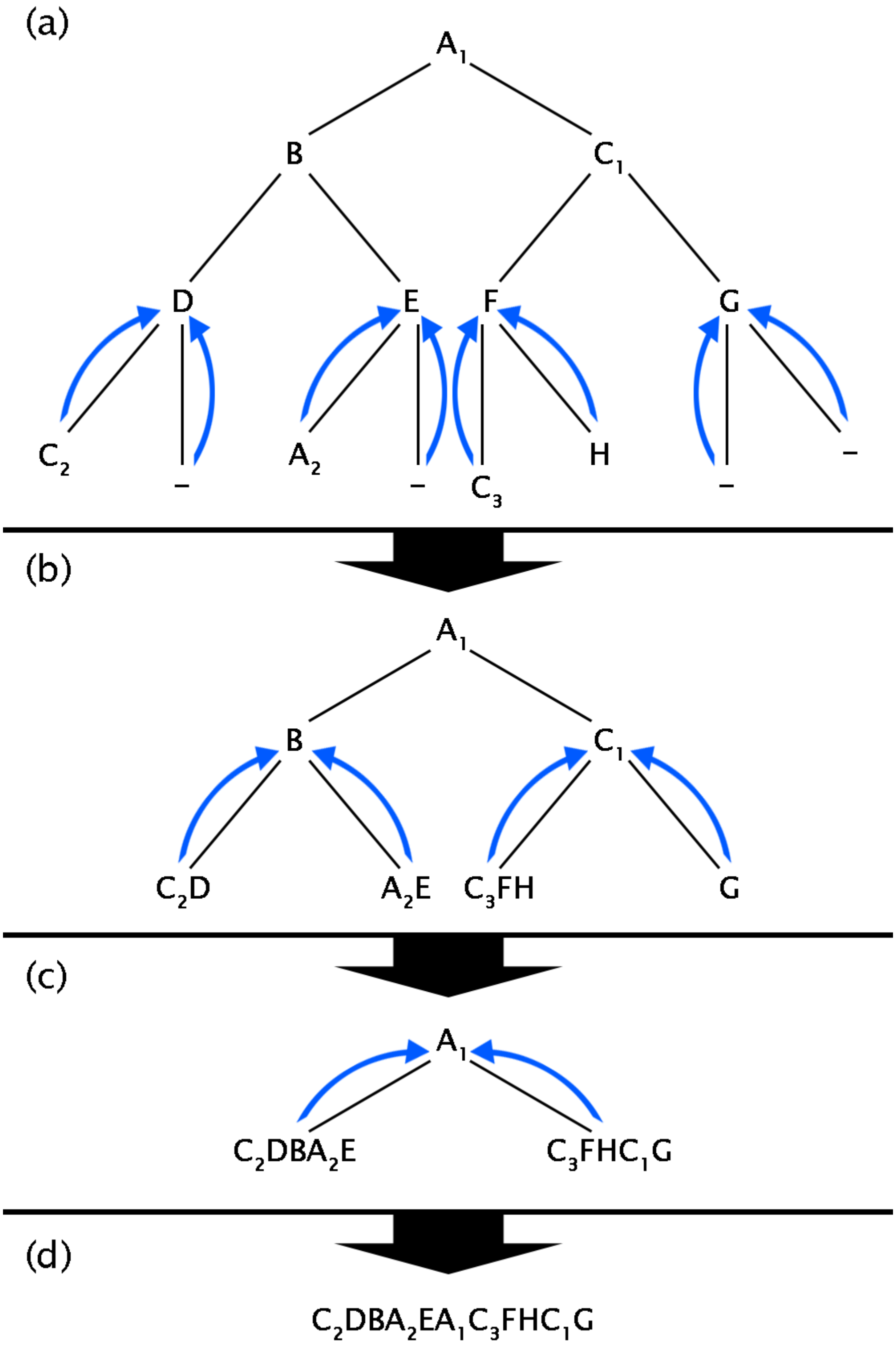


Fig. F22 Fragment formation procedure.

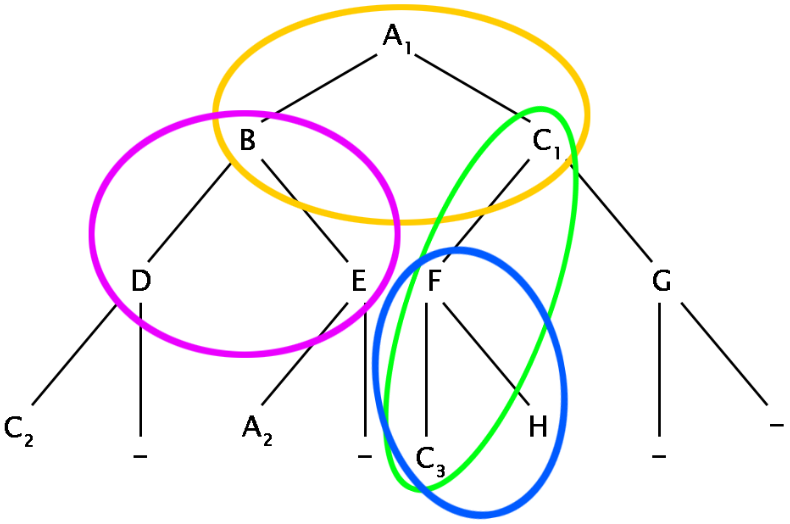


Fig. F23 Different subtree locations on a tree.

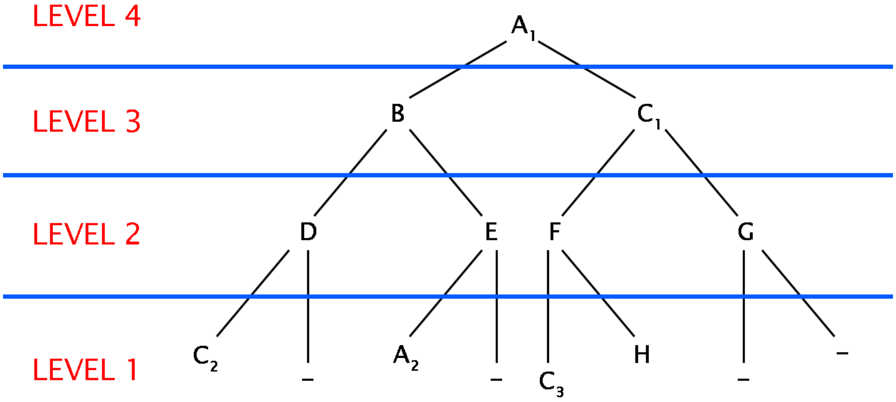


Fig. F24 Point levels of a tree for scoring.

Appendix G Fragment Elimination Details

1 - The fragment must pass the validation check created with the RDKit library. It is essentially a sanitizability test that determines whether the fragment contains open rings, branches, illegal atom types, and so on. As a result, the fragment can be assumed to be both syntactically and chemically plausible.

2 - Taking hydrogen atoms out of the equation and treating salt structures as one atom, the total number of atoms in the fragment must be greater than or equal to three. As a result, more interpretable and chemically significant fragments can be targeted for investigation.

3 -

For classification tasks,

The task label of the chemical containing the fragment must be 1. As a result, only chemical fragments that have a positive relationship with their corresponding task can be investigated.

For regression tasks,

The task label of the chemical containing the fragment must be greater than or equal to the label set's average value. As a result, only chemical fragments that have a positive relationship with their corresponding task can be investigated.

Fig. G25 Fragment elimination criteria.