```
C:\Users\Minh\anaconda3\envs\my-rdkit-env\python.exe D:\Minh\Books\
Research\Accelerated-Paper\Github\scripts\prepare_dataset.py
Python: 63230 molecules.
Excel: 63230 molecules.
Python: 63230 molecules.
Excel: 62861 molecules.
Identical: 62861 molecules.
Python only: 369 molecules.
Python-only molecules: {'OCCC1NCNCN1', 'O=C[C@H](O)c1ccccc1', 'CSc1ccc(O)
cc1', 'COC1CCC(C)CC1', 'NNCC1CCCCC1', 'CC(=O)C1CCC(C)CC1', 'NN1CCOC1=O', '
C#CC#Cc1ccccc1', 'NC(=O)Cn1ccsc1=O', 'NC(=S)NC1CCCCC1', 'N#CNC1CCCCC1', '
CC(=O)CC1CCCCC1', 'C=C[C@H](O)C1CCCCC1', 'CC1CCC(C=O)CC1', 'C=C=CC1CCCCC1
', 'OCC1CCCCC1', 'CN1CCN(C(N)=O)C1=O', 'COC(=O)C1CCCCC1', 'C=CCn1cccc1', '
C#CC1CCNCC1', 'CC(=O)C1CCNCC1', 'C#C/C=C/n1cccc1', 'CC(C)C1CCC(N)CC1', '
CCCCC1CCCCC1', 'C=CC(=C)c1ccccc1', 'O=CCC1CCCCC1', 'CCn1cccnc1=O', 'C#
CCOC1CCCCC1', 'C#C/C=C/C1CCCCC1', 'CCC(=O)C1CCCCC1', 'O=C(O)Nc1ccccc1', '
FC(F)Oc1ccncc1', 'CN(C)Cc1ccncc1', 'CN1Cc2cccnc2C1', 'C=C1CCC(=C)CC1', 'O=
NOC1CCCCC1', 'C=C(C)CC1CCCCC1', 'CNc1ccc(N)cc1', 'CCOC1CCC(O)CC1', 'C=C[C@
H](C)N1CCCC1', 'CNCCn1cccc1', 'NOC1CCCCC1', 'CC(C)CC1CCCCC1', '
CSCSc1ccccc1', 'C[C@H](CO)C1CCCCC1', 'NCCSc1ccccc1', 'c1cocn1', 'C#C[C@H](
O)C1CCCCC1', 'CS(=O)(=O)C1CCNCC1', 'C1CCNCC1', 'C[S@](=O)c1ccc(N)cc1', '
NCCC1CCCCC1', 'COCc1ccncc1', 'C#CCC1CCCCC1', 'CC[C@H](N)c1ccccc1', '
COCCc1ccccc1', 'NC(=O)OC1CCCCC1', 'CCCn1ccoc1=O', 'CCCCn1cccc1', 'C[C@@H](
S)C1CCCCC1', 'CSCC1CCC(C)CC1', 'O=c1ccocc1', 'O=CC#Cc1ccccc1', 'CCCn1cccc1
', 'OCCSC1CCCCC1', 'C=C(C)Cn1cccc1', 'COCCn1cccc1', 'C[C@H](CN)N1CCCC1', '
C=CCOc1ccncc1', 'CCOCc1ccncc1', 'C[C@H](N)CC1CCCCC1', 'CC1CCC([C@H](C)O)
CC1', 'N#Cc1ccc(C=O)cc1', 'C[S@@](=O)c1ccc(N)cc1', 'O=C(O)C1CCC(F)CC1', '
COc1ccc(CN)cc1', 'CNC(=O)C1CCCCC1', 'C/C=N/C1CCCCC1', 'C=CCC1CCC(C)CC1', '
NCC1CCC(O)CC1', 'NNC(=O)C1CCCCC1', 'O=C(O)N1CCCC1', 'CC(C)SC1CCCCC1', 'NC[
C@H](O)c1ccccc1', 'O=CCOC1CCCCC1', 'CCC(=O)n1cccc1', 'OC(O)C1CCCCC1', '
c1ccc2ccccc2c1', 'CCCNc1ccccc1', 'c1ccccoccc1', 'ONCc1ccccc1', 'ONC1CCCCC1
', 'NN1CCc2ccccc21', 'CC1CCC(C(N)=O)CC1', 'O=CNC1CCCCC1', 'NC(=O)c1ccc(O)
cc1', 'CCCSc1ccccc1', 'CCCN1CCC(=O)CC1', 'O=C1OCCN1CCO', 'CC(=O)N1CCC(=O)
CC1', 'CNCCC1CCNCC1', 'CCCOC1CCNCC1', 'OCCNc1ccncc1', 'CN1CCCCC1=N', '
CC1CCC(OO)CC1', 'CCSC1CCCCC1', 'CCC#CC1CCCCC1', 'C=C1C[C@@H]1C1CCCCC1', 'N
#CC1CCC(CN)CC1', 'CCCC(=O)n1cccc1', 'CSCCC1CCCCC1', 'CS(=O)(=O)C1CCCCC1
', 'NC(N)=NC1CCCCC1', 'CC1CCC(CN)CC1', 'C[C@H](O)Cc1ccccc1', 'c1cncnc1', '
C[C@H](C#N)n1cccc1', 'CCC1CCC(C=O)CC1', 'CN(C)CCn1cccc1', 'N=C1CC1', 'C[C@
H](CO)N1CCCC1', 'O=C1CCCCCC1', 'CCOC(=O)n1cccc1', 'O=CN1CCCCC1=O', 'CC(C)
C1CCCCC1', 'C[C@H](S)c1ccccc1', 'C=CC1CCC(O)CC1', 'Cc1ccncc1', 'CC(C)n1ccc
(=O)cc1', 'FC(Cl)(Cl)c1ccccc1', 'NCCC1CCNCC1', 'OOC1CCCCC1', 'O=COC1CCCCC1
', 'N=C(N)SC1CCCCC1', 'CCCC1CCNCC1', 'Cn1ccc(=NN)cc1', 'C1CNC2NCCCC2C1', '
OCCn1ccnc1', 'OCc1ccncc1', 'CCn1ccc(=O)cc1', 'CC(C)OC1CCNCC1', 'CS(=O)(=O)
n1cccc1', 'N[C@@H](CF)c1ccccc1', 'NS(=O)(=O)c1ccncc1', 'CCN(C)c1ccccc1', '
CCCN1CCCCC1=O', 'CCc1ccc(N)cc1', 'Cn1ccc(=O)o1', 'C1CCCCC1', 'Cn1ccn(CO)c1
=O', 'COC1CCCCC1', 'CNCc1ccc(C)cc1', 'C[C@]1(C2CCCCC2)CO1', 'C1CCC(C2CNC2)
CC1', 'CCNCCn1cccc1', 'O=C(O)c1ccccc1', 'CCNCc1ccncc1', 'C=C(C)c1ccccc1
', 'CC(C)Cn1cccc1', 'Cn1cnc(=O)n(C)c1=O', 'C/N=c1\\n(C)ccn1N', 'CNC1CCC(C)
CC1', 'COC(=O)n1ccoc1=O', 'C#CCOc1ccncc1', 'CCN1CCCC1', 'CCCCN1CCNC1', 'CC
(C)SC1CCNCC1', 'CCN(C)c1ccncc1', 'Cc1ccc(NN)cc1', 'C[C@H](O)Cc1ccncc1', '
CSCOC1CCCCC1', 'C[C@H](O)C1CCNCC1', 'O=[N+]([O-])C1CCCCC1', 'C=CCSc1ccccc1
```

```
 8 ', 'CCCc1ccc(C)cc1', 'COCc1ccc(N)cc1', 'C=CCCC1CCCCC1', 'CCCC1CCCCC1', '
   C1CCn2nnnc2C1', 'O=[N+]([O-])CC1CCCCC1', 'NCCNC1CCCCC1', 'CCCNc1ccncc1', '
   N#Cc1ccc(C#N)cc1', 'OCCCCn1cccc1', 'CC[C@H](O)C1CCNCC1', 'CN(C)CC1CCCCC1
   ', 'CCS(=O)(=O)N1CCCC1', 'CCOC1CCCCC1', 'O=C(O)Cn1ccsc1=O', 'NCCc1ccc(N)
   cc1', 'CNC1CCC(CO)CC1', 'C[C@H](C=O)N1CCCC1', 'OCCc1ccc(O)cc1', 'CNC(=O)
   n1cccn1', 'COC(=N)c1ccccc1', 'O=C1CCC1=O', 'COCC1CCCCC1', 'NC(=S)Cc1ccccc1
   ', 'CC(C)(O)C1CCCCC1', 'c1cnccn1', 'CCCCc1ccncc1', 'CCc1ccc(CC)cc1', 'C=
   CCN1CCCCC1=O', 'CCSCC1CCCCC1', 'CCCOC1CCCCC1', 'CC(=O)c1ccc(O)cc1', 'CC[C@
   H](O)c1ccccc1', 'C/C=C/Cc1ccccc1', 'C=CCC1CCCCC1', 'CO[C@H](C)C1CCCCC1', '
   O=C1CCCCN1CO', 'NCCCn1cccc1', 'c1ccoc1', 'O=C(OO)c1ccccc1', 'O=S(=O)(O)
   c1ccccc1', 'N#Cc1ccc(O)cc1', 'CCc1ncccn1', 'FC(F)CC1CCNCC1', 'C1CCC(NC2CC2
   )CC1', 'COC(=O)c1ccncc1', 'COC(=O)n1cccc1', 'CCc1ccncc1', 'CC(=O)C1CCCCC1
   ', 'CN(C)N1CCCC1', 'CCNCc1ccccc1', 'C1CCC(N2CCC2)CC1', 'C=COC1CCCCC1', '
   OCCc1ccccc1', 'C/C=C/CN1CCCC1', 'CC(C)CCn1cccc1', 'CN1CCc2cnncc21', 'C#
   CCn1cccc1', 'NCc1ccccc1', 'NCc1ccncc1', 'NCCN1CCCCC1=O', 'NN1CCCCC1=O', '
   NC(=S)c1ccncc1', 'C=C[C@H](C)C1CCCCC1', 'CCOc1ccncc1', 'CSCC1CCCCC1', '
   NNCCc1ccccc1', 'OOCC1CCCCC1', 'COCOC1CCCCC1', 'O=C(NO)C1CCCCC1', 'CC(F)(F)
   c1ccncc1', 'NC1CCCCC1', 'C[C@H](N)c1ccccc1', 'CCC1CCCCC1', 'C[S@@](=O)
   CC1CCCCC1', 'N#C/C=C/C1CCCCC1', 'C[C@H](C=O)C1CCCCC1', 'CNCC1CCC(N)CC1', '
   CC(=O)Oc1ccncc1', 'C1CNOC1', 'FCCSC1CCCCC1', 'N[C@H](CF)c1ccccc1', 'O=C(O)
   Cn1cccc1', 'c1ccc([C@@H]2CO2)cc1', 'O=CC1CCC(C=O)CC1', 'OCCOc1ccncc1', '
   C1Cn2nnnc2CN1', 'C#CC1CCCCC1', 'CO[C@H](C)C1CCNCC1', 'COc1ccc(C#N)cc1', '
   CSCCc1ccncc1', 'OC[C@H](O)c1ccncc1', 'CN1CCN(C)C1=N', 'C#CC(=C)C1CCCCC1
   ', 'CCCc1ncccn1', 'CC[S@](=O)C1CCCCC1', 'OCC1CCC(O)CC1', 'C[C@H](O)c1ccc(O
   )cc1', 'NCCCc1ccccc1', 'C1CCN(C[C@@H]2CO2)C1', 'OCN1CCCC1', 'C[S@](=O)
   CC1CCCCC1', 'CN1CCc2ccccc21', 'O=C=NCC1CCCCC1', 'C1CCNNC1', 'CC(C)N1CCCC1
   ', 'O=C1CCCCN1O', 'OC[C@H](O)c1ccccc1', 'CC1CCC(ON)CC1', 'COn1ccccc1=O', '
   CC[C@H](C)C1CCCCC1', 'N[C@@H]1C[C@@H]1c1ccccc1', 'Cn1cc2ccccc2c1', 'CC(=O)
   n1cccc1', 'CN1CCc2cnccc21', 'CCOC1CCC(N)CC1', 'Cn1cnc2ccccc21', 'NS(=O)(=O
   )c1ccccc1', 'COc1ncc(C=O)cn1', 'O=Nc1ccncc1', 'COCn1cccc1', 'C=CC(=O)
   c1ccccc1', 'O=C(CF)c1ccccc1', 'C[C@H](C#N)C1CCCCC1', 'CNCCC1CCCCC1', 'N#
   CCCN1CCCC1', 'CO/C=C/C1CCCCC1', 'CNN1CCCC1', 'C=CCOC1CCCCC1', 'C[C@H](O)
   c1ccccc1', 'CN(C)C(=O)N1CCCC1', 'O=CN1CCCC1', 'CCN(O)c1ccccc1', 'C#CC(=O)
   C1CCCCC1', 'CN1CCC(=N)CC1', 'CC(C)(N)C1CCCCC1', 'CONCC1CCCCC1', 'C/C=C/
   c1ccccc1', 'NN1CCCC1', 'CN(O)c1ccccc1', 'C=CC1CCCCC1', 'COn1ccc(=O)cc1', '
   N=C1SCCN1CC(=O)O', 'C[C@H](N)c1ccc(F)cc1', 'NCC1CCC(N)CC1', 'O=S(=O)(O)
   c1ccncc1', 'C[C@H](O)Cn1cccc1', 'C=CCn1ccc(=O)cc1', 'C1CCC(C2CC2)CC1', 'N#
   CC#CC1CCCCC1', 'Oc1cncnc1', 'O=C/C=C/C1CCCCC1', 'N=C=NC1CCCCC1', 'C=
   CCNC1CCCCC1', 'CCC1CCC(C)CC1', 'COc1ccncc1', 'Cn1ccn(C)c1=NN', '
   CN1Cc2ccncc2C1', 'OCCCn1cccc1', 'CCNC1CCC(N)CC1', 'CC(=O)OC1CCCCC1', 'C=C(
   C)C1CCC(C)CC1', 'O=C1NCCCN1O', 'C=CCCn1cccc1', 'C#CCCCn1cccc1', 'N#
   CCOC1CCCCC1', 'CS(=O)(=O)c1ncccn1', 'CN(C)c1ccc(N)cc1', 'C#COC1CCCCC1', 'C
   [C@H](NN)c1ccccc1', 'ON=C1CCCCCC1', 'CN1CCC(=O)N1C', 'COc1ccc(O)cc1', 'CC[
   S@@](=O)c1ccccc1', 'N#C[C@H](N)C1CCCCC1', 'CC(C)c1ncccn1', 'S=C(S)
   NC1CCCCC1', 'CNCC1CCCCC1', 'CC(=O)n1ccoc1=O', 'NCCCC1CCNCC1', 'O=CC1CCCCC1
   ', 'O=C1CCC(=O)C1=O', 'NC(=O)Cc1ccccc1', 'CC(C)(O)c1ccncc1', 'CC(C)=
   Cc1ccncc1', 'CC(C)C(=O)n1cccc1', 'CN[C@H](C)c1ccccc1', 'C1=C(n2cccc2)CCC1
   ', 'N#CCCCn1cccc1', 'C[C@@H](N)c1ccc(F)cc1', 'CC(=O)N1CCNC1', 'CNCc1ccncc1
   ', 'c1cnn2nnnc2c1', 'CN[C@H](C)C1CCNCC1', 'CCNC(=O)n1cccn1'}
 9 Excel only: 0 molecules.
10 Excel-only molecules: set()
```

```
11
12 Process finished with exit code 0
13
```