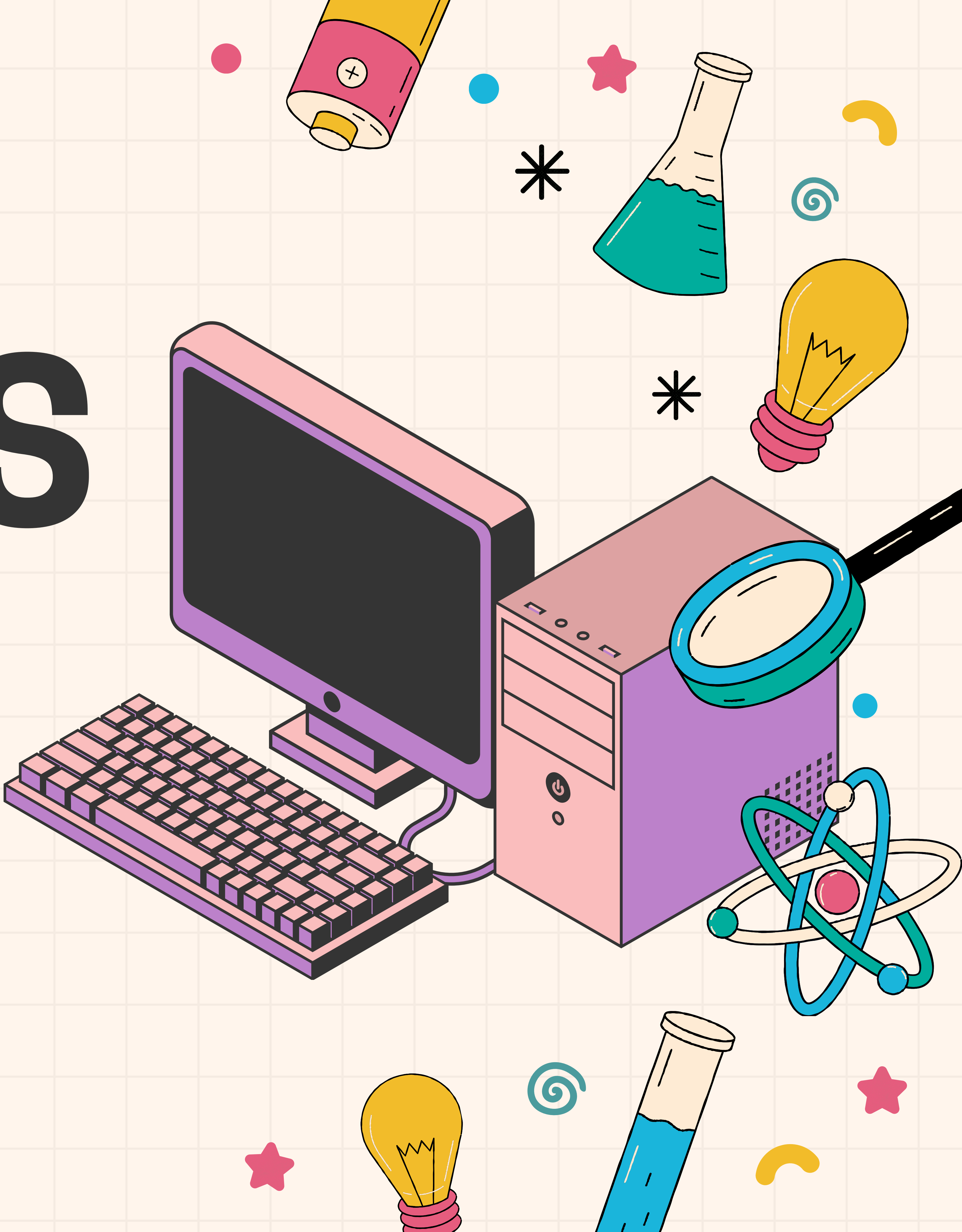
CS182

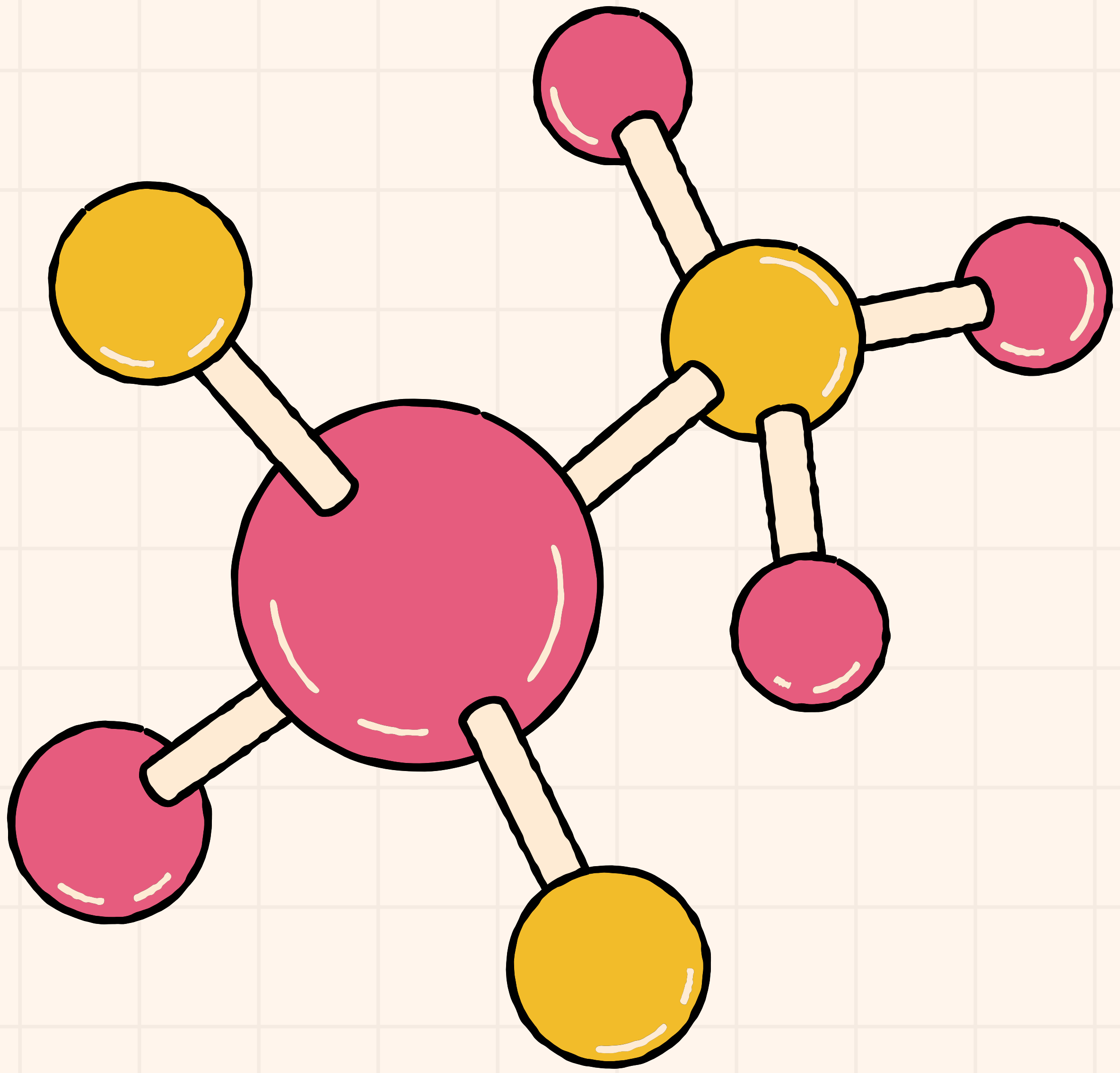
MOLECULES

PROPERTY CLASSIFICATION FEATURING **MULTI-MODALITY**



Team members: Zhangzhi Xiong Tianni Yang Yixuan Chen

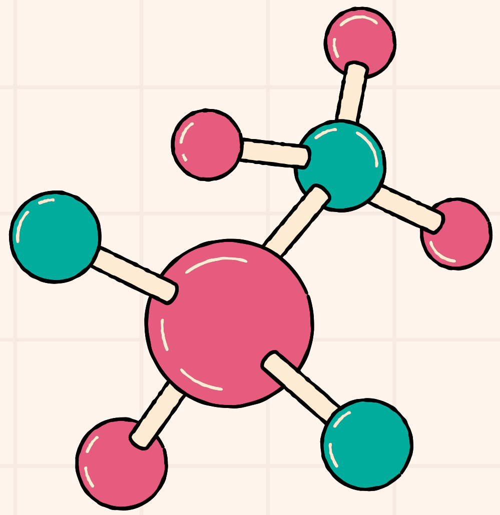
OUTLINE



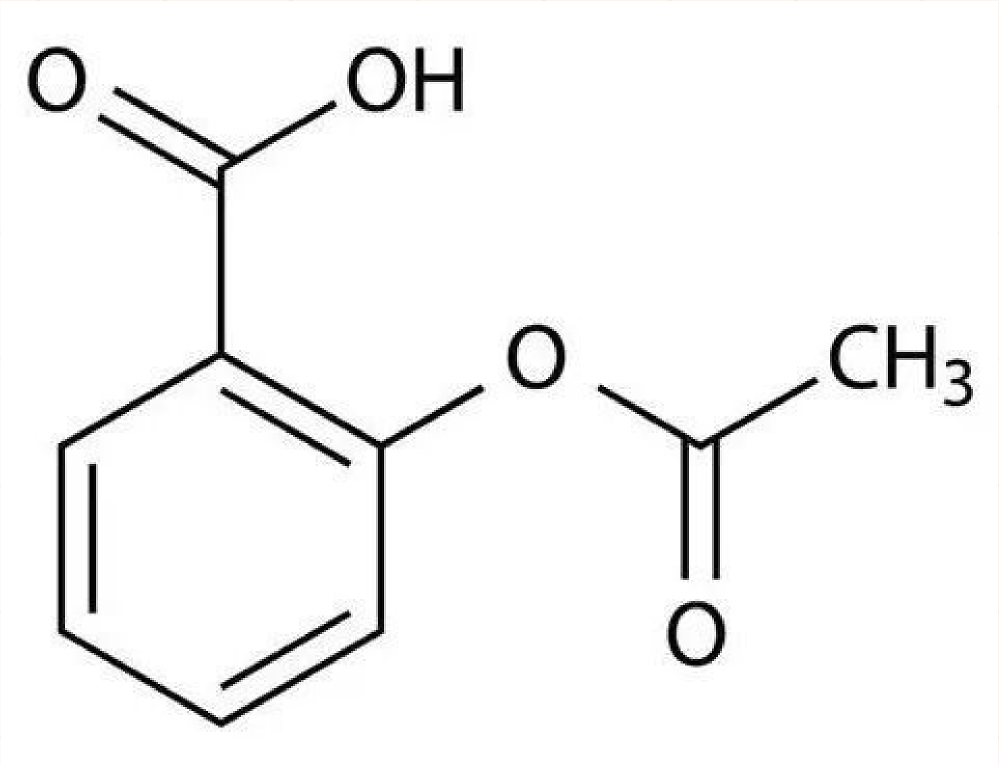
01. INTRODUCTION

02. METHODS

03. EXPERIMENTS



01. INTRODUCTION



Example: Aspirin

Output:

- Toxic?
- Soluble?
- ...

DATASET

Moleculenet

In our research, we choose a subset of datasets which are suitable for **classification**:

- **BACE:** β-secretase enzyme
- **HIV:** HIV activeness
- **Tox21:** multiple properties
- **BBBP:** Blood-Brain Barrier Penetration
- **Clintox:** drugs failed clinical trials for toxicity
- **SIDER:** multiple side effects

Data > MoleculeNet > hiv.csv

	A	B
1	smiles	HIV_active
2	CCC1=[O+][Cu-3]2([O+]=C(CC)C1)[O+]=C(CC...	0
3	C(=Cc1ccccc1)C1=[O+][Cu-3]2([O+]=C(C=Cc3...	0
4	CC(=O)N1c2ccccc2Sc2c1ccc1ccccc21	0
5	Nc1ccc(C=Cc2ccc(N)cc2S(=O)(=O)O)c(S(=O)(...	0
6	O=S(=O)(O)CCS(=O)(=O)O	0
7	CCOP(=O)(Nc1cccc(Cl)c1)OCC	0
8	O=C(O)c1ccccc1O	0
9	CC1=C2C(=COC(C)C2C)C(O)=C(C(=O)O)C1...	0
10	O=[N+]([O-])c1ccc(SSc2ccc([N+](=O)[O-])cc2[...	0
11	O=[N+]([O-])c1ccccc1SSc1ccccc1[N+](=O)[O-]	0
12	CC(C)(CCC(=O)O)CCC(=O)O	0
13	O=C(O)Cc1ccc(SSc2ccc(CC(=O)O)cc2)cc1	1
14	O=C(O)c1ccccc1SSc1ccccc1C(=O)O	0
15	CCCCCCCCCCCC(=O)Nc1ccc(SSc2ccc(NC(...	0

Data > MoleculeNet > clintox.csv

```
smiles,FDA_APPROVED,CT_TOX
*C(=O)[C@H](CCCCNC(=O)OCCOC)NC(=O)OCCOC,1,0
[C@@H]1([C@@H]([C@@H]([C@H]([C@@H]([C@@H]1Cl)Cl)Cl)Cl)Cl)Cl,1,0
[C@H]([C@@H]([C@@H](C(=O)[O-])O)O)([C@H](C(=O)[O-])O)O,1,0
[H]/[NH+]=C(/C1=CC(=O)/C(=C\C=c2ccc(=C([NH3+])N)cc2)/C=C1)\N,1,0
[H]/[NH+]=C(\N)/c1ccc(cc1)OCCCCCOc2ccc(cc2)/C(=[NH+]/[H])/N,1,0
[N+](=O)([O-])[O-],1,0
[N]=O,1,0
[O-][99Tc](=O)(=O)=O,1,0
[O-]P(=O)([O-])F,1,0
[O-]S(=O)(=O)[O-],1,0
[O-]S(=O)(=S)[O-],1,0
[Se],0,1
B([C@H](CC(C)C)NC(=O)[C@H](Cc1ccccc1)NC(=O)c2cnccn2)(O)O,1,0
B([C@H](CC(C)C)NC(=O)[C@H](CC1=CC=CC=C1)NC(=O)C2=NC=CN=C2)(O)O,0,1
B([C@H](CC(C)C)NC(=O)CNC(=O)C1=C(C=CC(=C1)Cl)Cl)(O)O,0,1
C#CC1(CCCCC1)OC(=O)N,1,0
C#CC[NH2+][C@@H]1CCc2c1cccc2,1,0
C#CCC(Cc1cnc2c(n1)c(nc(n2)N)N)c3ccc(cc3)C(=O)N[C@@H](CCC(=O)[O-])C(=O)[O-],1,0
C#N,1,0
```

RELATED WORK

Feature Engineering Model (Leveraging Different Fingerprint Feature):
Random Forest; XGBoost; AttentiveFP...

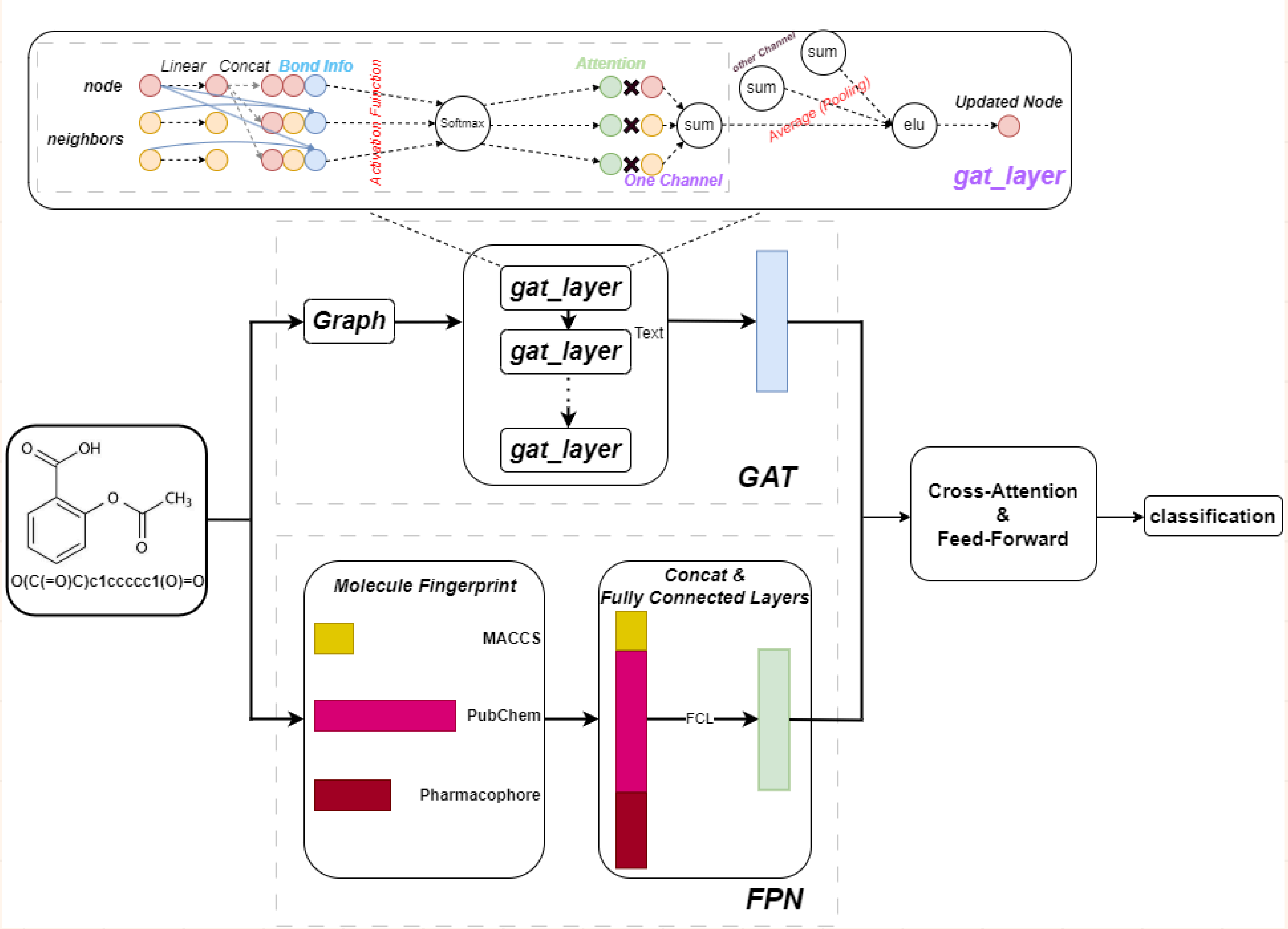
Graph Based Model:
Graph Convolution Model; Weave Model; Direct Acyclic Graph Model;...

Few Feature Mixture Model: FP-GNN

Multi-Modality
Engineering!

02. METHODS

PIPELINE



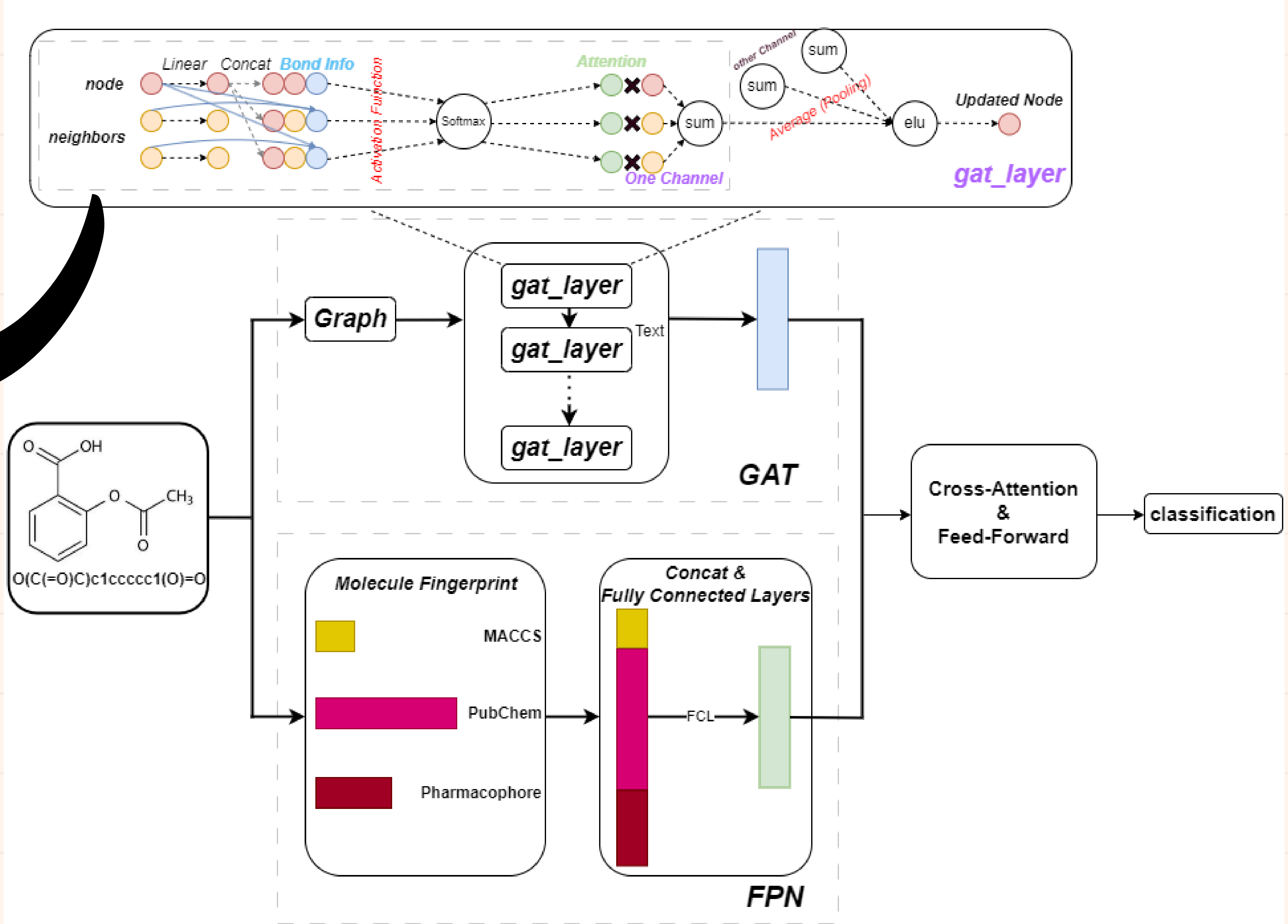
02. METHODS

PIPELINE

Graphical Information

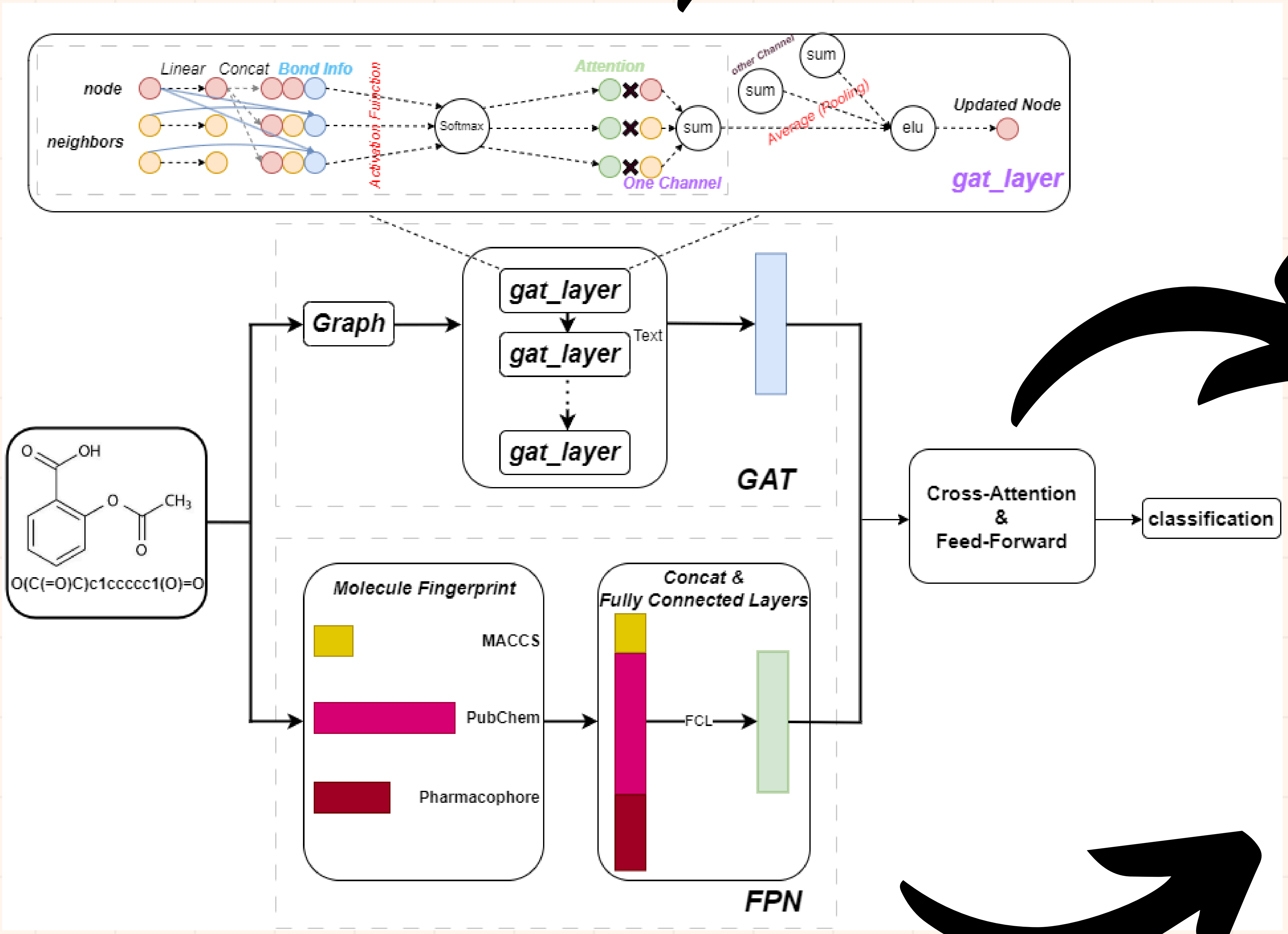
Graph features:

- Atom Features:
  - atom_symbol
  - degree
  - ...
- Adjacency Matrix
- Bond Features:
  - Single
  - Double
  - Triple
  - Aromatic
  - None



PIPELINE

Tried **GraphSAGE to replace graph attention,** but poor in performance



Cross-Attention

To Connect Features:

Previous methods:

- Naive concatenation

  Out method:

- Cross-Attention

Self-Attention

Tried **Self-Attention** on the fingerprints in FPN. Did not work.

03. EXPERIMENTS

Table 1: Model Comparison Results

Dataset Name	Split type	GC	WEAVE	AttentiveFP	XGBoost	FP-GNN	Ours
BACE	Random			0.876	0.889	0.881	**0.903**±0.007
	Scaffold	0.783	0.806			**0.860**	0.828±0.009
HIV	Random			0.824	0.816	0.825	**0.835**±0.016
	Scaffold	0.763	0.703			0.824	**0.825**±0.008
Tox21	Random	0.829	0.820	**0.852**	0.836	0.815	0.823±0.005
BBBP	Random			0.887	0.926	**0.935**	0.907±0.008
	Scaffold	0.690	0.671			0.916	**0.931**±0.012
Clintox	Random	0.807	0.832	0.904	**0.911**	0.840	0.870±0.011
SIDER	Random	0.638	0.581	0.623	0.642	0.661	**0.674**±0.005

Table 2: Ablation Study Results

Dataset Name	Split type	FP-GNN	Ours	Bond^{-}	CrossAttn^{-}	SelfAttn	GraphSAGE
BACE	Random	0.881	0.903±0.007	0.889±0.031	0.886±0.027	0.640±0.033	0.871±0.017
BBBP	Scaffold	0.916	0.931±0.012	0.901±0.023	0.911±0.025	0.816±0.085	0.871±0.011
SIDER	Random	0.661	0.674±0.005	0.672±0.013	0.679±0.008	0.584±0.014	0.627±0.021

THANK YOU