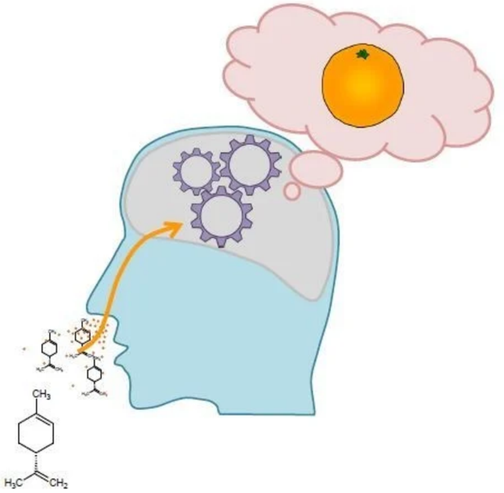
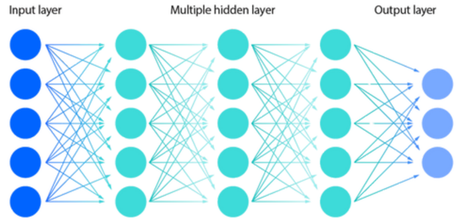
POLYTECH NANTES | Pôle Sciences et technologie Nantes Université



Molecular Odor Prediction Using GNN



Presentation by: Suman Basava
Supervisors: Prof. Fabric Guillet, Prof. Angélique Villière

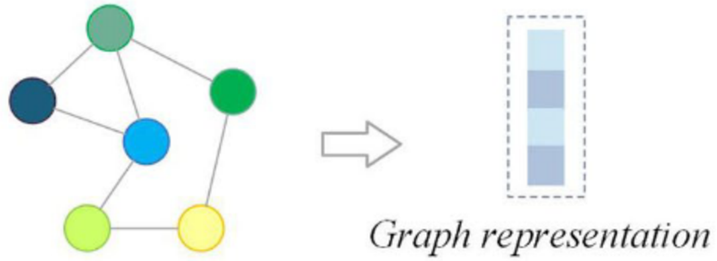
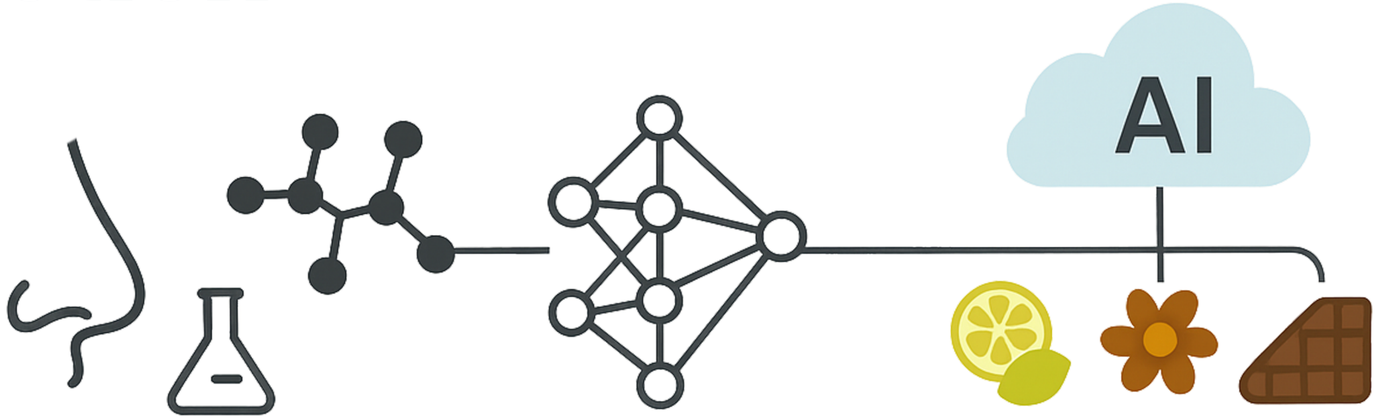


TABLE OF CONTENT

- Intorduction
- Problem Statement
- Objectives
- Related Works
- Thier Results
- Proposed Work
- Methodology
- Data Preprocessing
- Featurization
- Model Architecture
- Results
- Discussion
- Conclusion

Introduction



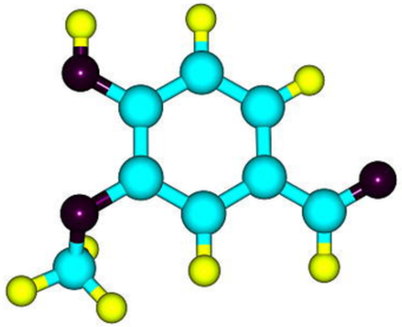
Why this matter?
Smell is complex, molecules trigger thousands of subtle odor perceptions.

What we did?
Used Graph Neural Networks (GNNs) to link molecule structure → Predict Odor.

Impact
Advances QSOR modeling through GNNs, enriched hierarchical datasets, and feature engineering, supporting applications in fragrance design.

*QSOR: Quantitative Structure Odor Relationship

"Smells Creamy, with hint of vanilla, slightly floral, some notes of woody and Chocolate."



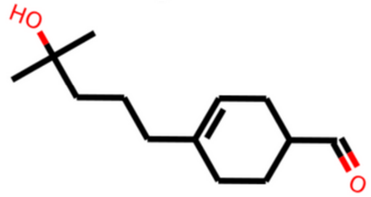
Predict →

creamy
fruite
chocolate
spicy
floral
citrus
baked
vanilla
woody

Odor Descriptors

Why it is hard to predict odor from molecule?

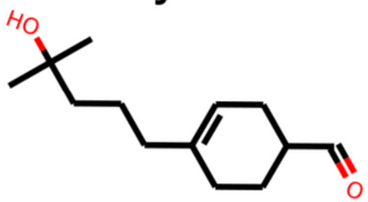
A Lyral



muguet, fresh,
floral, sweet

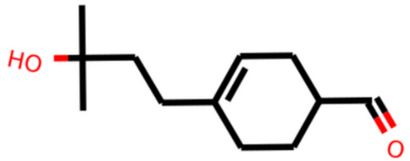
Why it is hard to predict odor from molecule?

A Lyral

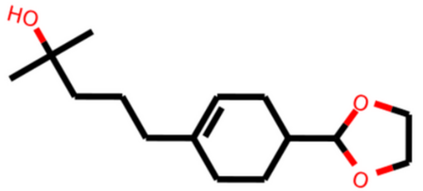


muguet, fresh, floral, sweet

B



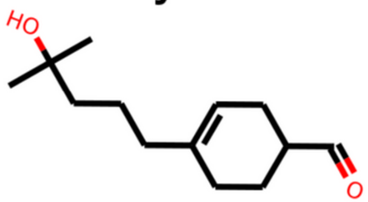
muguet floral



muguet, fresh, floral, sweet

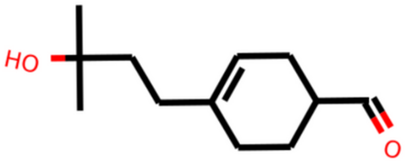
Why it is hard to predict odor from molecule?

A Lyral

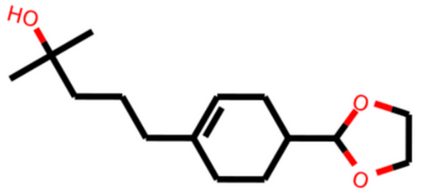


muguet, fresh, floral, sweet

B

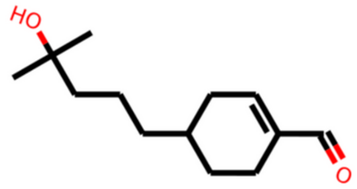


muguet floral



muguet, fresh, floral, sweet

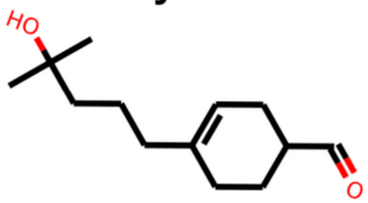
C



odorless

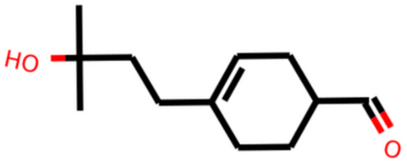
Why it is hard to predict odor from molecule?

A Lyral

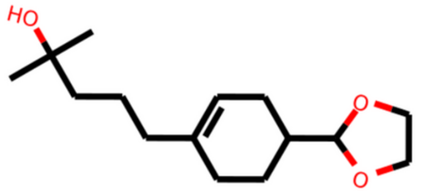


muguet, fresh, floral, sweet

B

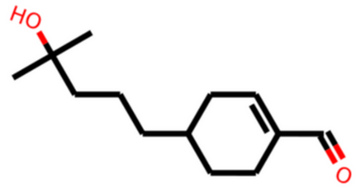


muguet floral



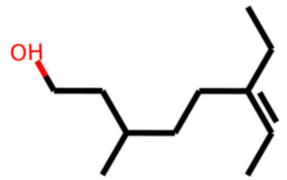
muguet, fresh, floral, sweet

C



odorless

D



muguet, fresh, floral, rosy

Objectives

1. Review state-of-the-art GNN approaches for odorant molecule mining
2. Enhance feature engineering with chemical functions (e.g., esters, ketones, aldehydes, lactones, etc.
3. Enrich dataset with odor hierarchies (e.g., fruity > apple)
4. Implement and evaluate the best-performing method
5. Publish findings

Literature Review

State of the Art

Machine Learning for Scent [1]	Principle Odor Map [2]	Multi-Feature Graph Attention Networks [3]	Mol-PECO [4]
Uses MPNN, GCN → better than fingerprint-based models. **Limitation:** Lacked chemical domain knowledge (e.g., functional groups, odor hierarchies).	MPNN trained to capture odor similarities & hierarchies. Achieved strong generalization beyond traditional methods. **Limitation:** Lacked chemical domain knowledge in feature engineering.	Uses Hierarchical Attention GCN (HAGCN). Integrated multiple features: atoms, bonds, functional groups, fingerprints. Adaptive focal loss handled label imbalance.	Combined Coulomb Matrix + spectral positional encoding. Improved representation of electrostatics & structure. **Limitation:** Outperformed traditional ML but not stronger than GCN/MPNN/GAT.

[1] Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules
[2] A Principal Odor Map Unifies Diverse Tasks in Human Olfactory Perception
[3] Molecular Odor Prediction Based on Multi-Feature Graph Attention Networks
[4] A deep position-encoding model for predicting olfactory perception from molecular structures and electrostatics

Model and Dataset used by previous works

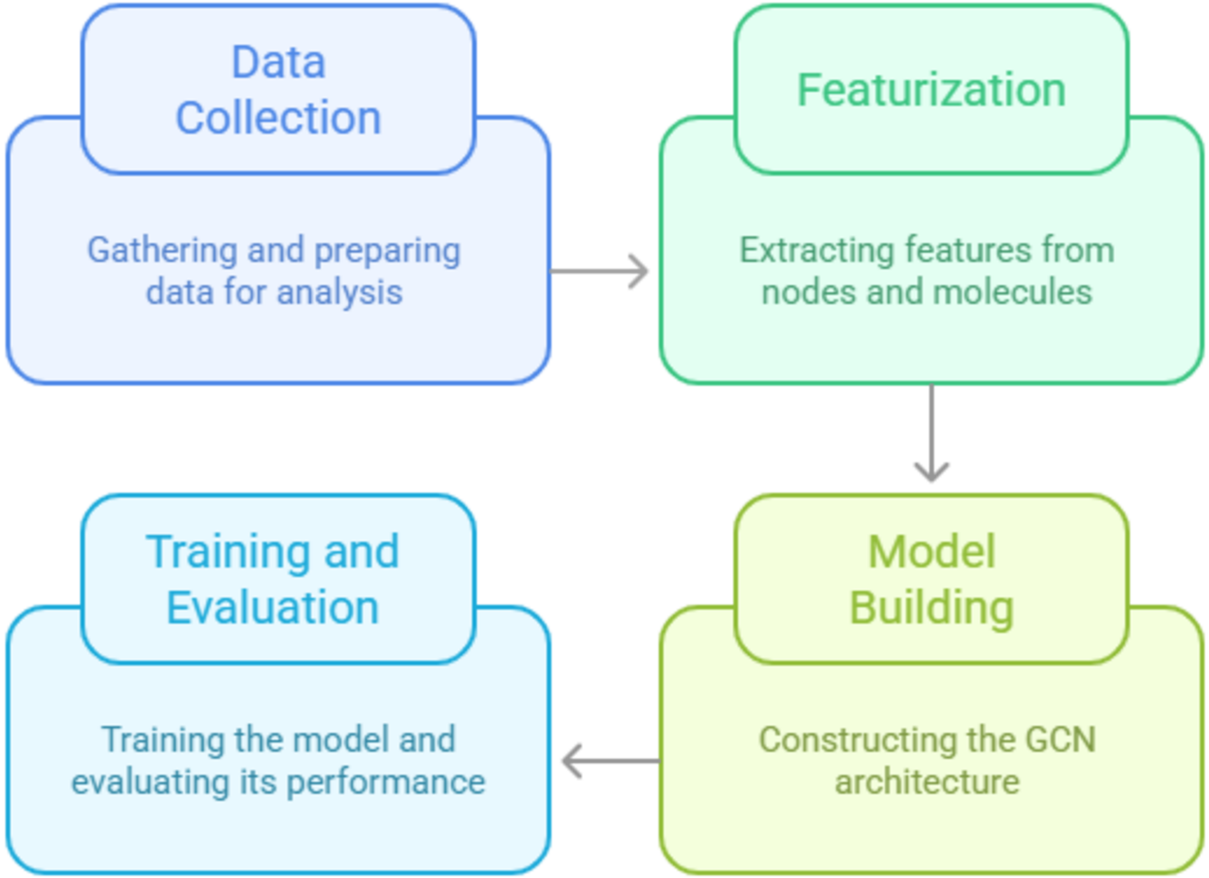
Paper	Model	Dataset Source	#Moleules	Odor Descriptors	Data Type
Machine Learning for Scent	GCN, MPNN	GS, LWF	5030	138	Flat
Principal Odor Map (POM)	MPNN	GS, LWF	~5000	138	Flat
Multi-Feature GAT	GAT	GS, LWF	5,788	154	Flat
MoI-PECO	GCN	Pyrfume	8503	118	Flat

GS: The good scents company - flavor, fragrance, food and cosmetics ingredients information.
LWF: John C Leffingwell. Leffingwell & associates
Pyrfume: Pyrfume Contributors. Pyrfume Data Repository

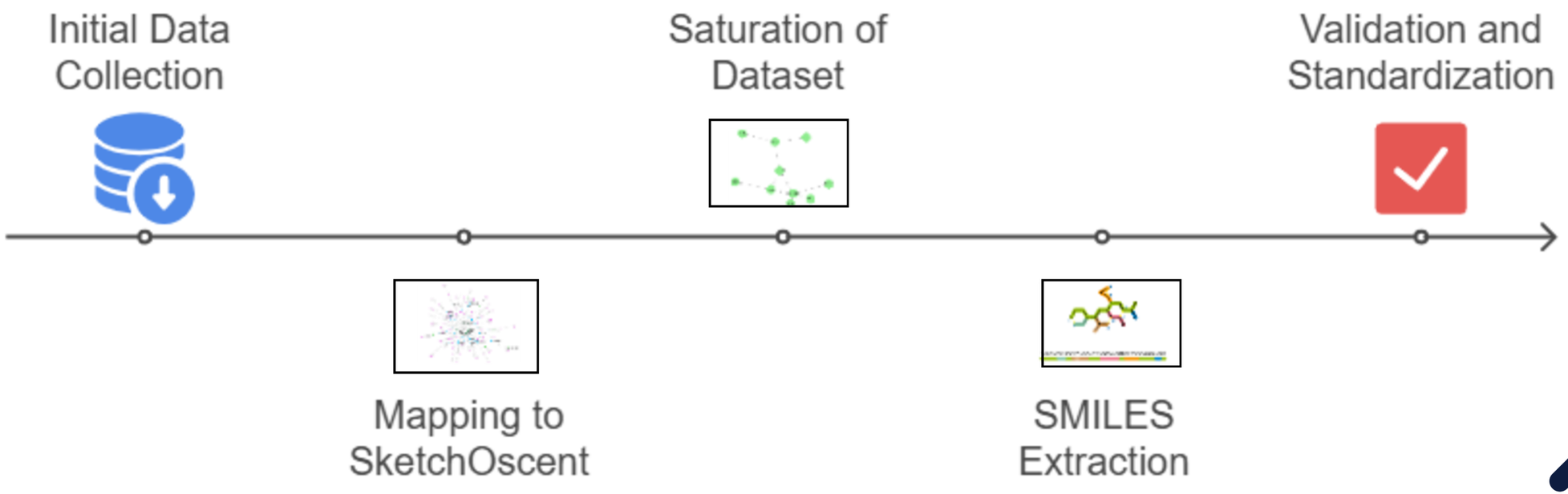
State of Art Results

Paper	AUROC	Precision	Recall	F1
Machine Learning for Scent	0.894	0.379	0.379	0.379
POM	0.854	0.578	0.557	0.555
Multi-Feature GAT	0.9294	-	-	0.4632
Mol-PECO	0.813	0.104	0.819	0.185

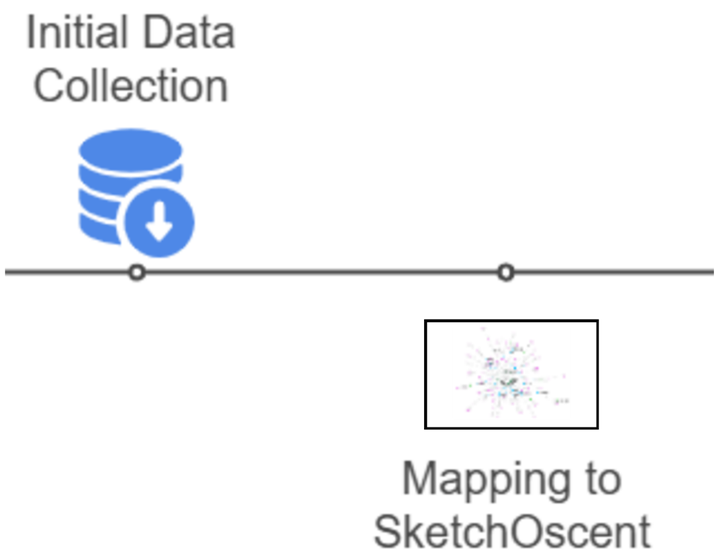
Proposed Work



Data Preprocessing and Analysis



Data Collection



- Initial Dataset was collected from GoodScent Company [ref]
- To enhance semantic consistency and reduce ambiguity, these terms were mapped into a standardized olfactory taxonomy known as the SketchOscent odor space.

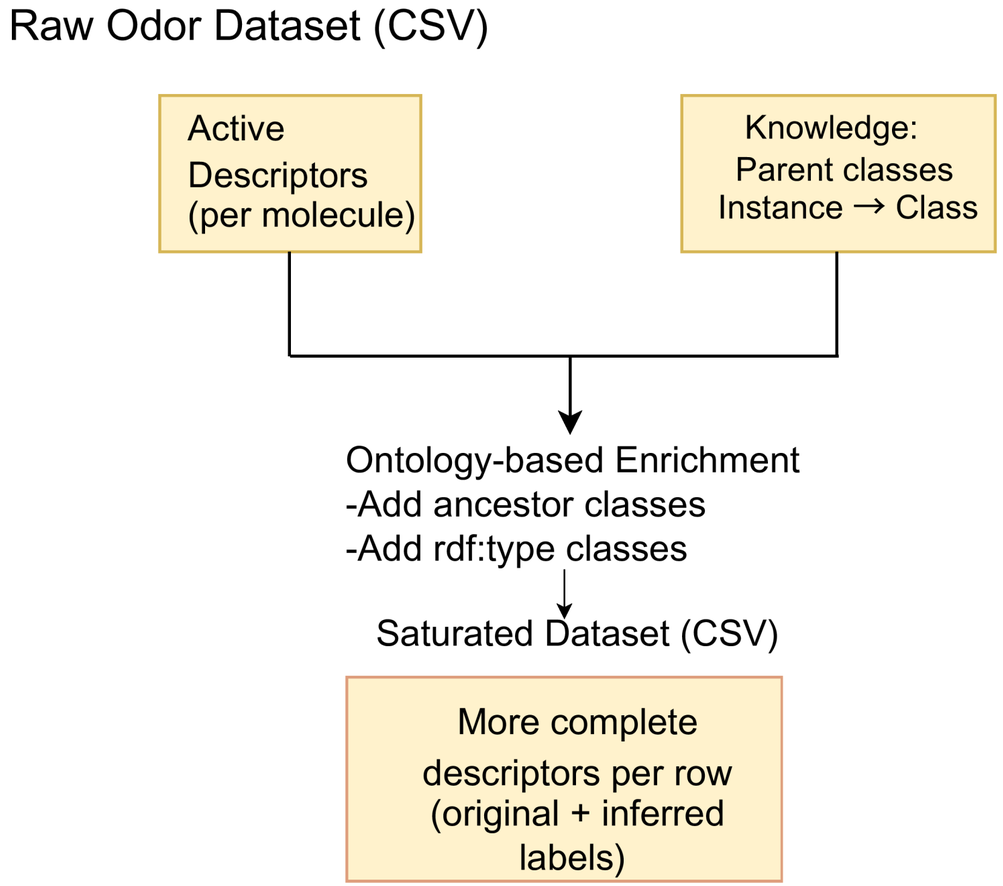
Raw Dataset Overview

Raw Dataset consited of 3,841 molecules and 394 odor descriptors.

	Acetique	Agrume	Aldehyde	Alliace	Boise	Baie	..... 388 other
50-70-4	0	0	0	0	0	1	..
51-67-2	0	0	0	0	1	0	..
65-85-0	0	0	0	1	0	0	..
56-40-6	0	0	0	0	0	0	..
..3837 other	..	..	..	..	..	..	..

*Structure of Scents (SoS) Dataset - https://oniris-polytech.univ-nantes.io/sketchoscent

Data Saturation



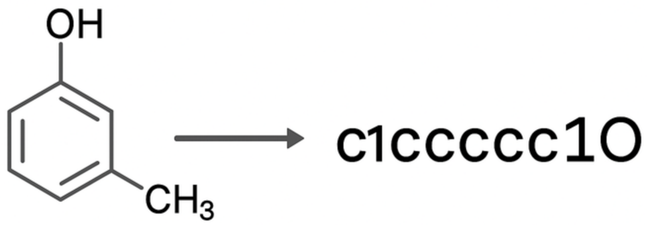
Saturation of Odor Labels

- Expanded and enriched odor labels using the SketchOscent ontology.
- Ontology provides structured relationships like “is-a” and “part-of”.
- Example: citrus → fruity
- Enables hierarchical, comprehensive odor recognition for each molecule.

SMILES Extraction

SMILES

Simplified Molecular Input Line Entry System



Benzene Ring

SoS Dataset (3,841 molecules, 394 descriptors)

↓

Step 1: CACTUS* (CAS → SMILES) [1]
- Parallel requests (ThreadPoolExecutor)
- > 2,000 SMILES retrieved

↓

Step 2: PubChemPy [2]
- get_cids(namespace='CAS') → get_properties('SMILES')
- Rate limiting (sleep)

↓

Filter & Merge
- Map SMILES back to rows
- Drop unresolved (141)

*NCI/CADD Chemical Identifier Resolver https://cactus.nci.nih.gov/chemical/structure

*NIH - National Library of Medicine https://pubchem.ncbi.nlm.nih.gov

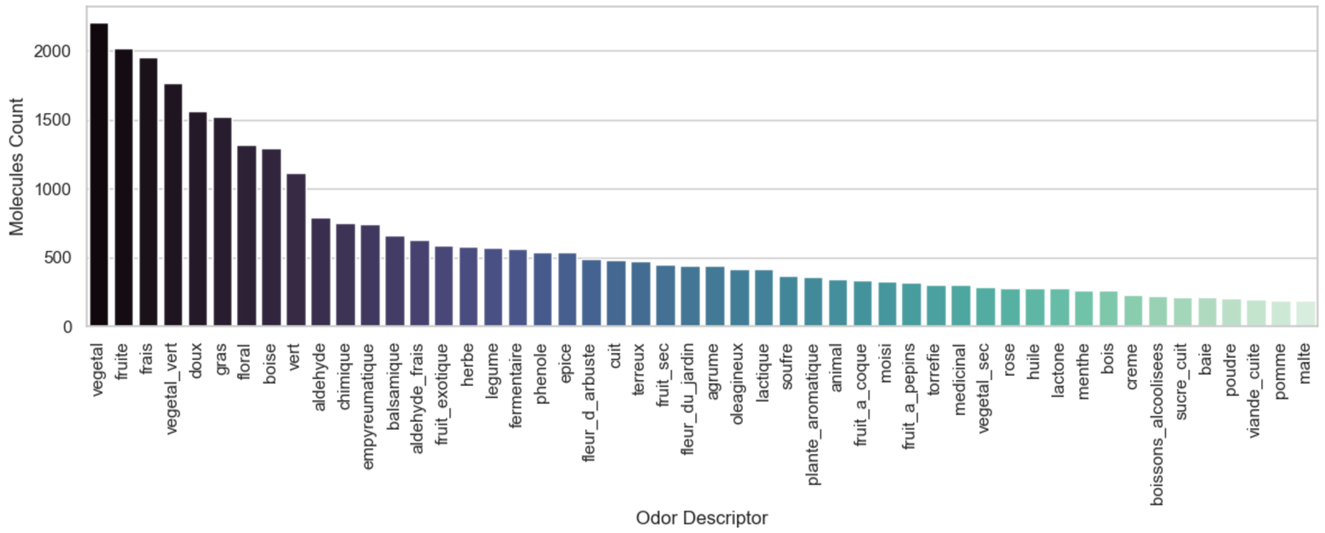
SoS Data

Dataset	#Molecules	#Odor Descriptors	Std. Dev	Min Count	Max Count	Mean
Raw SoS Data	3,841	394	129.94	1	1262	51.46
Data with SMILES	3,700	394	126.15	1	1220	49.99
Saturated Data	3,687	436	261.82	1	2210	97.97
Top138 Saturated Data	3,687	138	407.42	43	2208	284.05
OpenPOM Data	4,983	138	264.02	31	1902	176.17

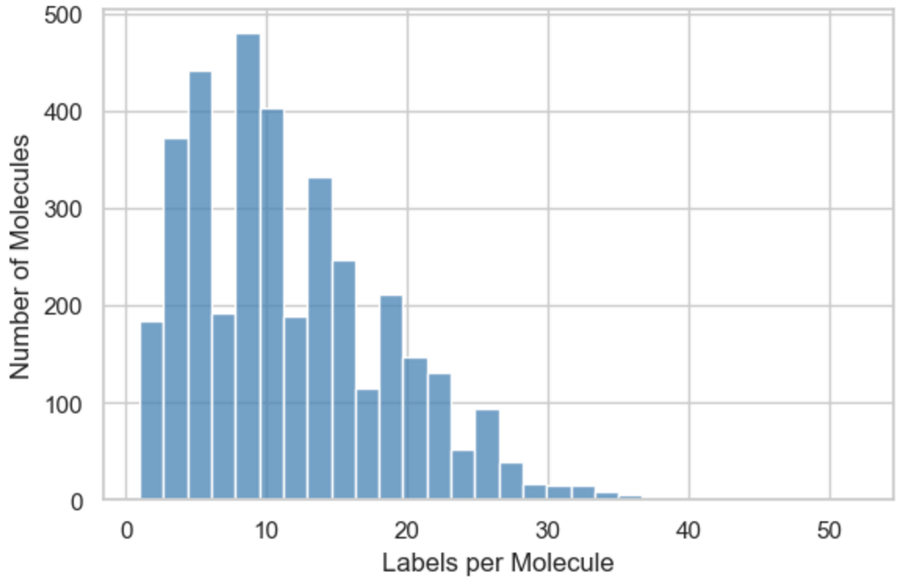
- 3,700 - Removed 141 molecules with no SMILES
- 3,687 - Removed 13 molecules with mol_wt > 600 g/mole

Dataset Details

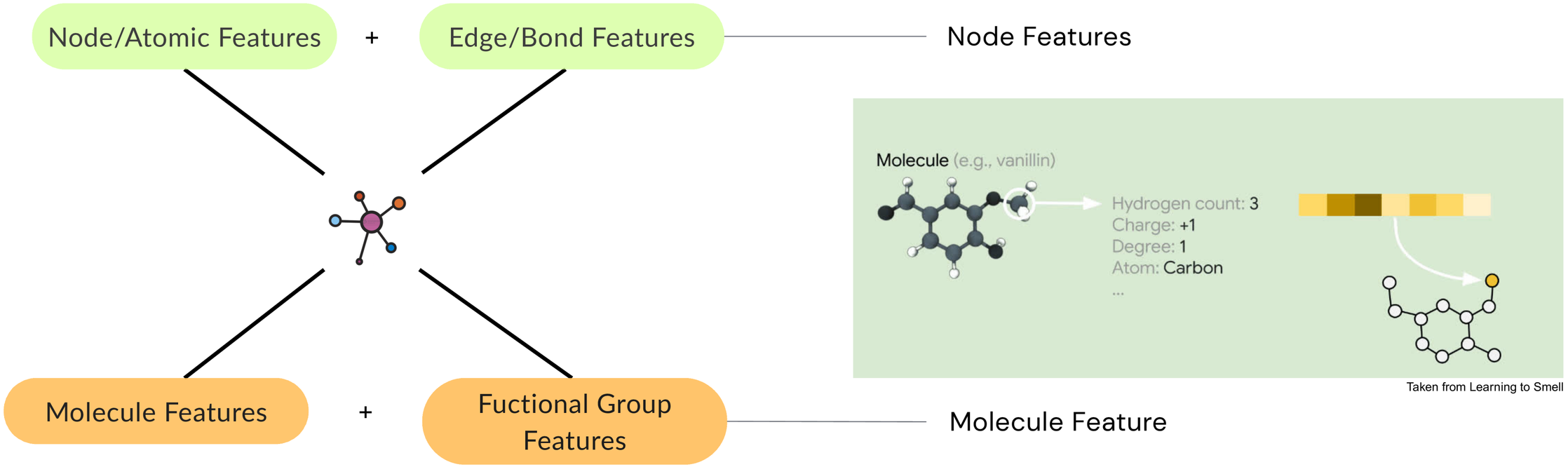
Distribution of odor descriptor frequencies



Distribution of label density.



Featurization



Taken from Learning to Smell

Featurization Strategy

A molecule can be represented as:

G = (V, E) $V = \{v_1, v_2, \ldots, v_n\}, \quad E = \{e_1, e_2, \ldots, e_m\}$

Total Features: 66

- **Node (Atomic) Features – 15**

$$F(A_i) = [Z_i, D_i, Q_i, H_i, R_i, V_i, A_i, Rg_i, S_i, C_i, Hy_i, B_i]$$

Z = Atomic Number	V = Valence
D = Degree	A = Aromatic
Q = Formal Charge	Rg = Is_in_Ring
H = No. of H	S = Smallest_Ring_Size
R = No. of radical e^-	C = Chirality
Hy = Hybridization	$B_i = [b_{single}, b_{double}, b_{triple}, b_{aromatic}] \in \{0,1\}^4$

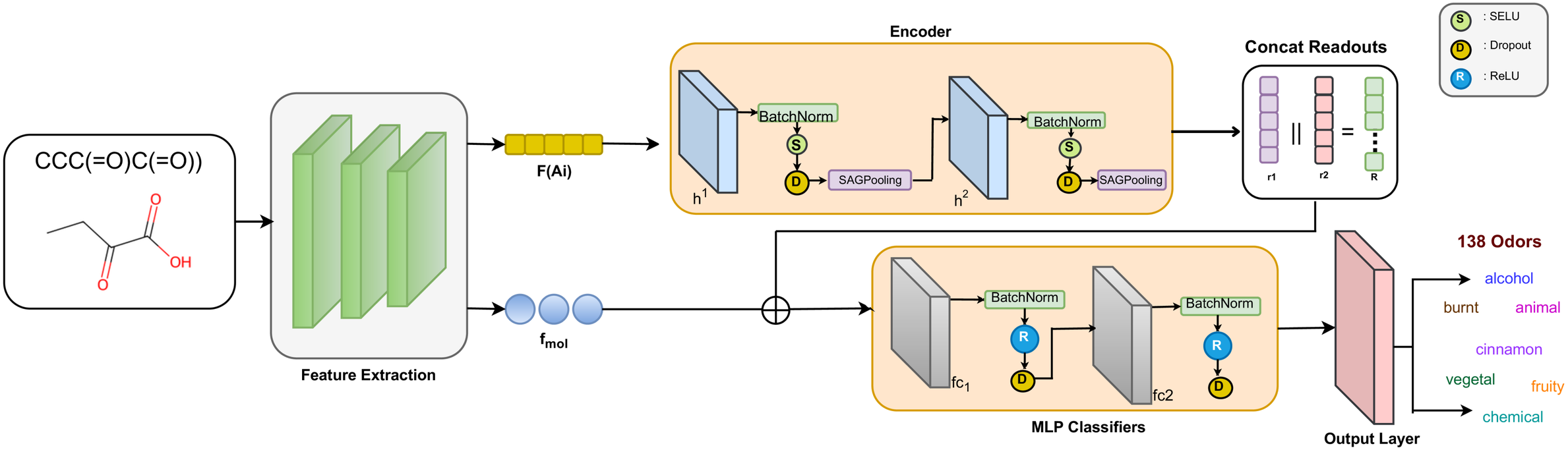
- **Molecular Features – 51**

$$f_{mol} = [MW, logP, TPSA, N_{rings}, N_{rot}, N_{donors}, N_{acceptors}, N_{heavy}, Q_{total}, C, LCC, f_{func}]$$

$$f_i = \text{count of occurrences of functional group } G_i \text{ in } M,$$

$$\mathbf{f}_{\text{func}}(M) = [f_1, f_2, \ldots, f_n] \in \mathbb{N}^n,$$

Model Architecture



Graph Isomorphism Network (GIN) [1]

The update rule for a node v in GIN is:

$$h_v^{(k)} = \mathrm{MLP}^{(k)}\left(\left(1+\epsilon^{(k)}\right)\cdot h_v^{(k-1)} + \sum\nolimits_{u\in\mathcal{N}(v)} h_u^{(k-1)}\right)$$

SAGPooling:

$\mathbf{y} = \mathrm{GNN}(X, A), \quad \mathbf{y} \in \mathbb{R}^N$ ——— Each node gets a scalar importance score via a GNN

$X' = (X \odot \sigma(\mathbf{y}))_{\mathbf{i}}, \quad A' = A_{\mathbf{i},\mathbf{i}}$ ——— Pooling (selection + rescaling + subgraph)

X → original node feature matrix

y → scalar importance scores (one per node) from the GNN.

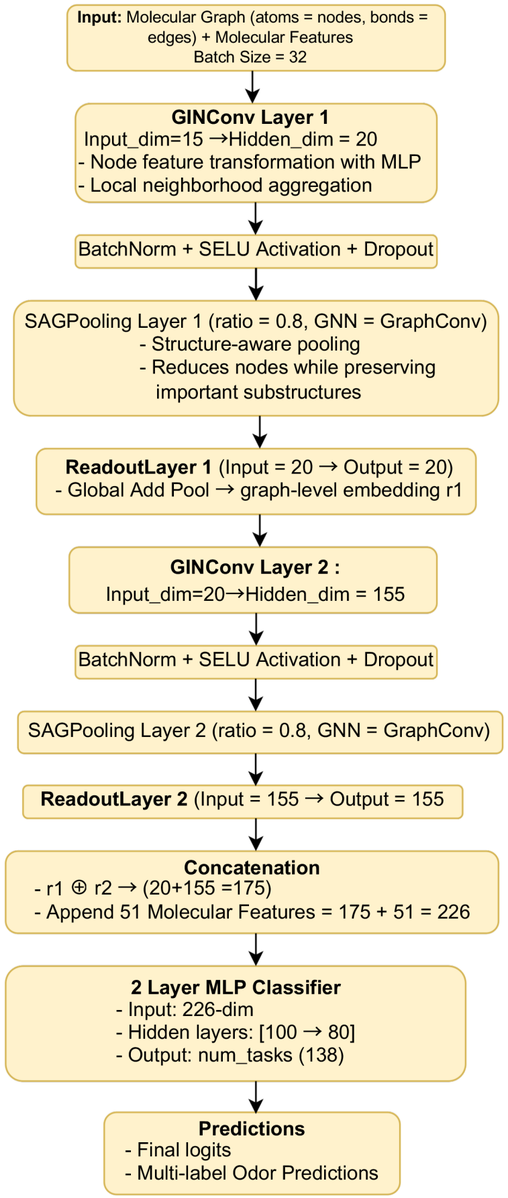
σ → nonlinearity that normalizes scores.

i → indices of selected nodes (top-k)

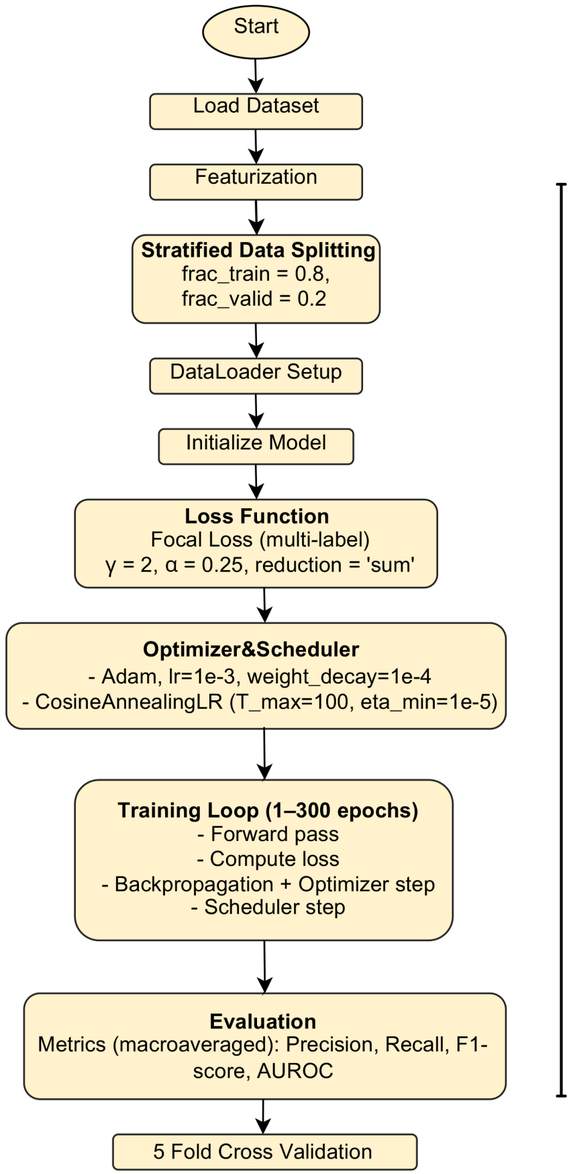
A → adjacency matrix of the graph.

[1] How Powerful are Graph Neural Networks?

Model Architecture Flowchart



Flow Diagram: Training and Evaluate



Major change made

Focal Loss

$$\mathcal{F}L(p) = -\Big(y(1-p)^{\gamma}\log p \;+\; (1-y)p^{\gamma}\log(1-p)\Big)$$

- Equivalent to cross-entropy loss when γ = 0
- y∈{0,1}: ground-truth label
- p = predicted probability (after sigmoid)
- γ : focusing parameter (controls down-weighting of easy examples)

Result

Paper	AUROC	Precision	Recall	F1
GIN Model(Our Work)	**0.879**	**0.376**	**0.512**	**0.384**
Machine Learning for Scent [1]	0.894	0.379	0.379	0.379
POM [2]	0.854	0.578	0.557	0.555
Multi-Feature GAT [3]	0.9294	-	-	0.4632
Mol-PECO [9]	0.813	0.104	0.819	0.185

Ablation: What Drives Performance?

#	Model Variant	AUROC [CI]	Precision [Min, Max]	Recall [Min, Max]	F1 Score [Min, Max]
1	Baseline GCN (BP) [1]	0.894 [0.888, 0.902]	0.379 [0.351–0.398]	0.390 [0.365–0.412]	0.360 [0.337–0.372]
2	OpenPOM_Git [17]	0.854	0.578	0.557	0.555
3	**GIN_Model**	**0.879 [0.869, 0.889]**	**0.376 [0.370, 0.381]**	**0.512 [0.494, 0.531]**	**0.384 [0.386, 0.394]**
4	Without Focal Loss	0.864 [0.863, 0.865]	0.648 [0.634, 0.662]	0.253 [0.243, 0.262]	0.277 [0.269, 0.283]
5	Without Normalizers	0.840 [0.836, 0.845]	0.496 [0.484, 0.508]	0.364 [0.358, 0.371]	0.271 [0.264, 0.278]
6	Without Regularizer	0.859 [0.857, 0.862]	0.372 [0.359, 0.385]	0.459 [0.453, 0.466]	0.332 [0.325, 0.338]
7	With Uniform Sampling	0.849 [0.844, 0.853]	0.383 [0.373, 0.432]	0.431 [0.421, 0.432]	0.311 [0.308, 0.314]
8	Without Global Features	0.844 [0.836, 0.853]	0.508 [0.454, 0.564]	0.274 [0.253, 0.295]	0.208 [0.190, 0.225]
9	Without Softmax at Readout	0.870 [0.869, 0.872]	0.352 [0.352, 0.356]	0.506 [0.499, 0.513]	0.370 [0.368, 0.373]
10	With Raw Dataset	0.842 [0.709, 0.868]	0.478 [0.401, 0.511]	0.291 [0.274, 0.375]	0.224 [0.219, 0.290]
11	Without Graphpooling	0.869 [0.868, 0.871]	0.343 [0.335, 0.353]	0.526 [0.523, 0.530]	0.376 [0.375, 0.376]
12	With GCN Model	0.863 [0.859, 0.866]	0.371 [0.353, 0.388]	0.483 [0.483, 0.483]	0.353 [0.348, 0.359]
13	With 4 Layers	0.860 [0.857, 0.863]	0.371 [0.357, 0.384]	0.463 [0.462, 0.464]	0.345 [0.342, 0.349]

Conclusion

GIN-based Framework for Multi-Label Odor Prediction

- Chemically-Informed Features → Adding functional groups + global descriptors improved performance by enriching molecular graphs.
- Targeted Training → Focal loss tackled label imbalance, boosting rare odor detection (F1 = 0.384).
- Hierarchical Semantics → Using Structure of Scent (SoS) captured relationships between odors, beyond flat classification.

Takeaway:

- Integrating graph learning + chemical knowledge + semantic hierarchy creates a robust, interpretable, and scalable QSOR model.

THANK YOU

For your attention