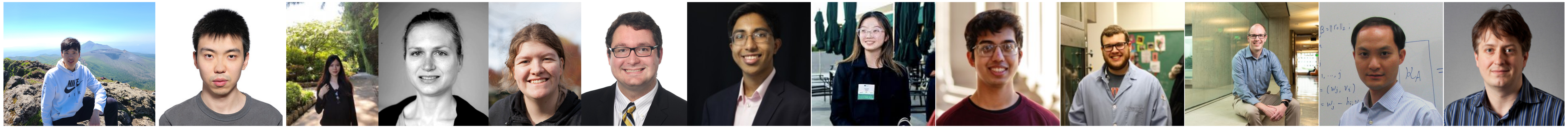
# Representing Molecules as Random Walks Over Interpretable Grammars

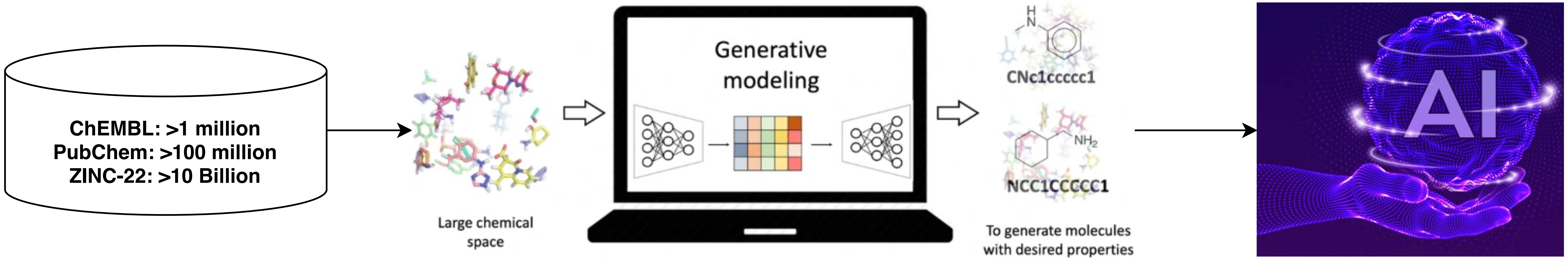
Michael Sun[1], Minghao Guo[1], Weize Yuan[3], Veronika Thost[4], Crystal Owens[1], Aristotle Grosz[2], Sharvaa Selvan[1], Katelyn Zhou[4], Hassan Mohiuddin[1], Benjamin Pedretti[2], Zachary Smith[2], Jie Chen[5], Wojciech Matusik[1]

[1]MIT CSAIL, [2]MIT Chemical Engineering, [3]MIT Chemistry, [4]Wellesley College, [5]MIT-IBM Watson AI Lab



# Background & Motivation

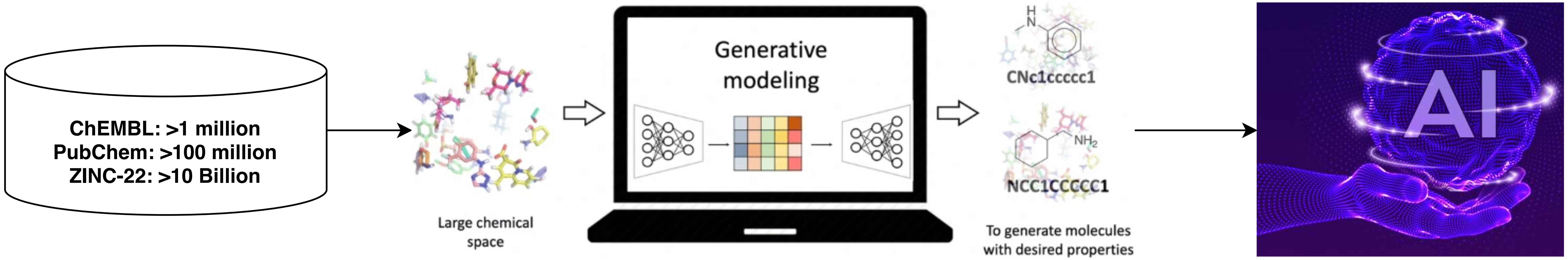
## Small Molecules



ChEMBL: >1 million
PubChem: >100 million
ZINC-22: >10 Billion

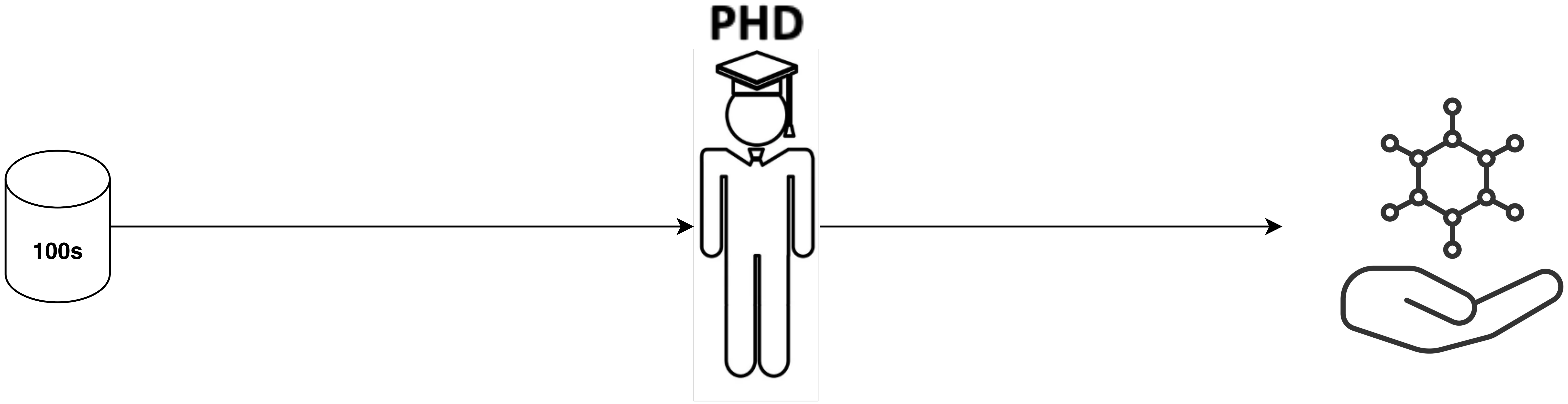
# Background & Motivation

## Small Molecules



ChEMBL: >1 million
PubChem: >100 million
ZINC-22: >10 Billion

## Big Molecules

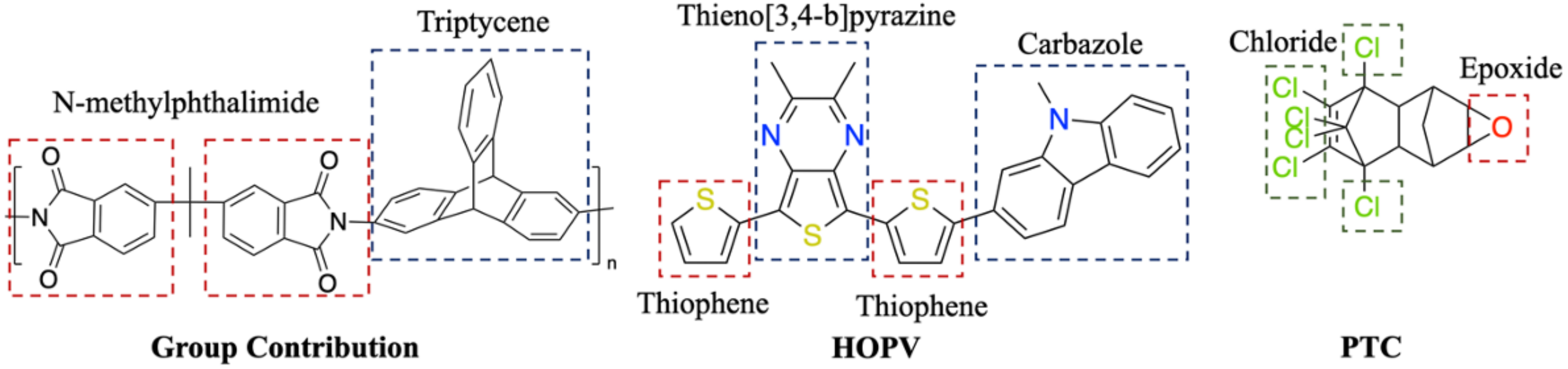


100s

# Problem Setting

## Traditional Approach

- Hand-designed by experts
- Uses known set of functional groups
- Only tens/hundreds of examples
- **Question 1**: How do we obtain these motifs?
- **Question 2:** Which motifs can attach to each other?

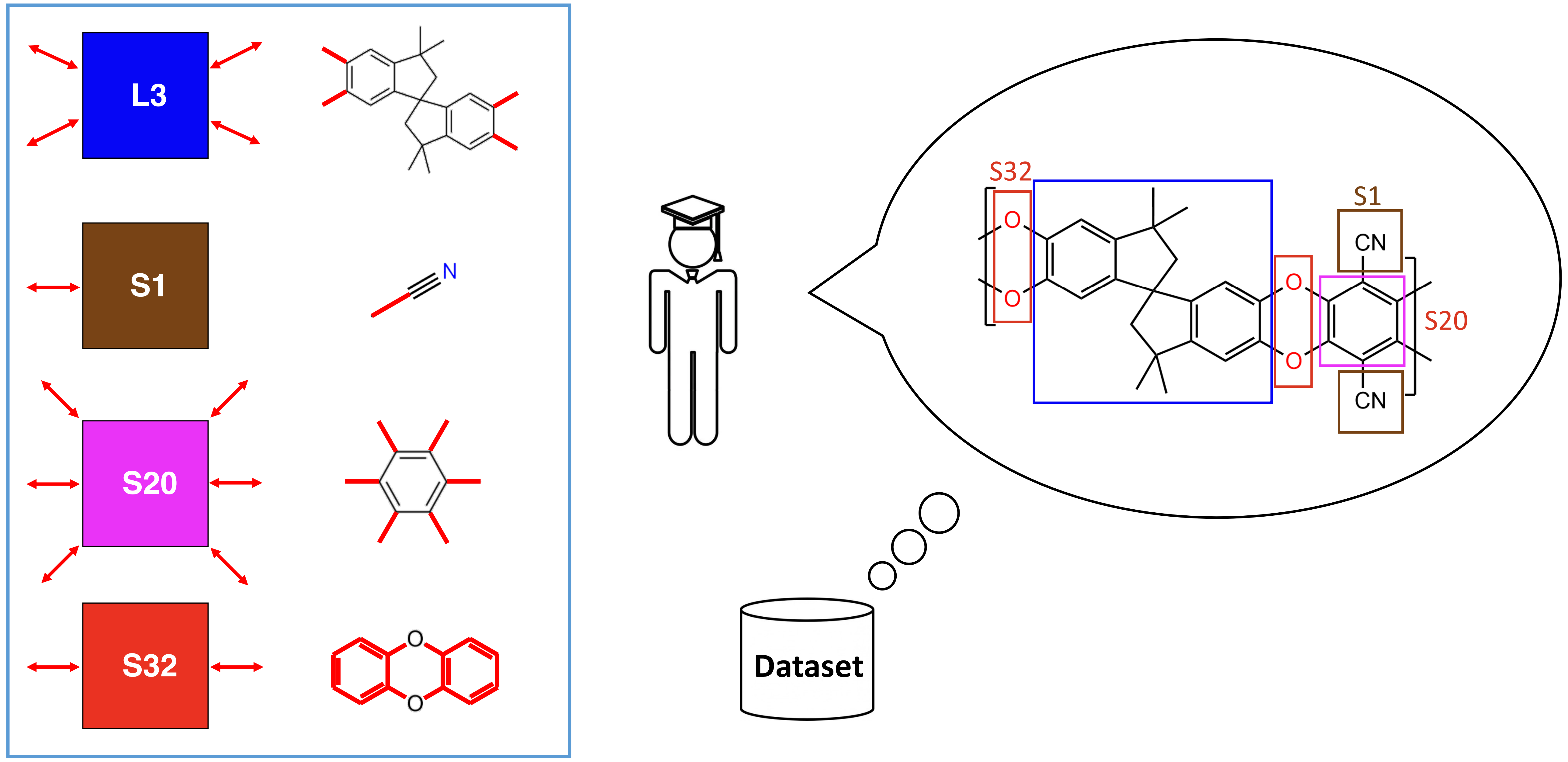


**Group Contribution** **HOPV** **PTC**

# Our Workflow

## Expert Approach (manual)

- Compile known motifs specific to domain
- Ask experts to annotate attachment *contexts* (red)
- Annotate occurrences of motifs in existing molecules



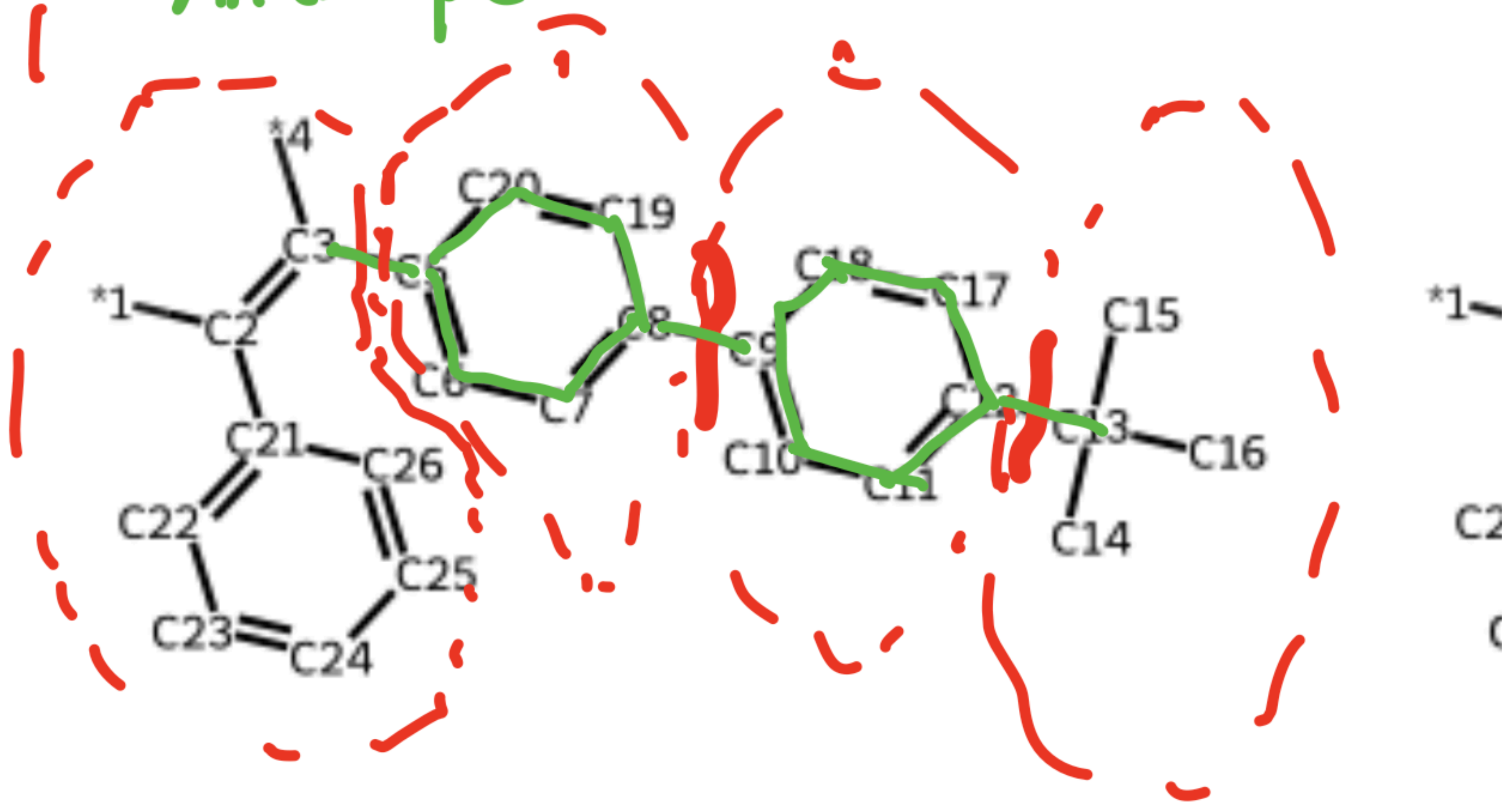
Context specifies what is required of a neighbor group.

# Our Workflow

## Expert Approach (semi-automated)

- Ask experts to fragment existing molecules (via breaking bonds)
- Extract the motifs programmatically
- Infer the contexts using dataset-specific rules

An example:



we have 4 black groups: #1: red group is 5,6,7,8,19,20

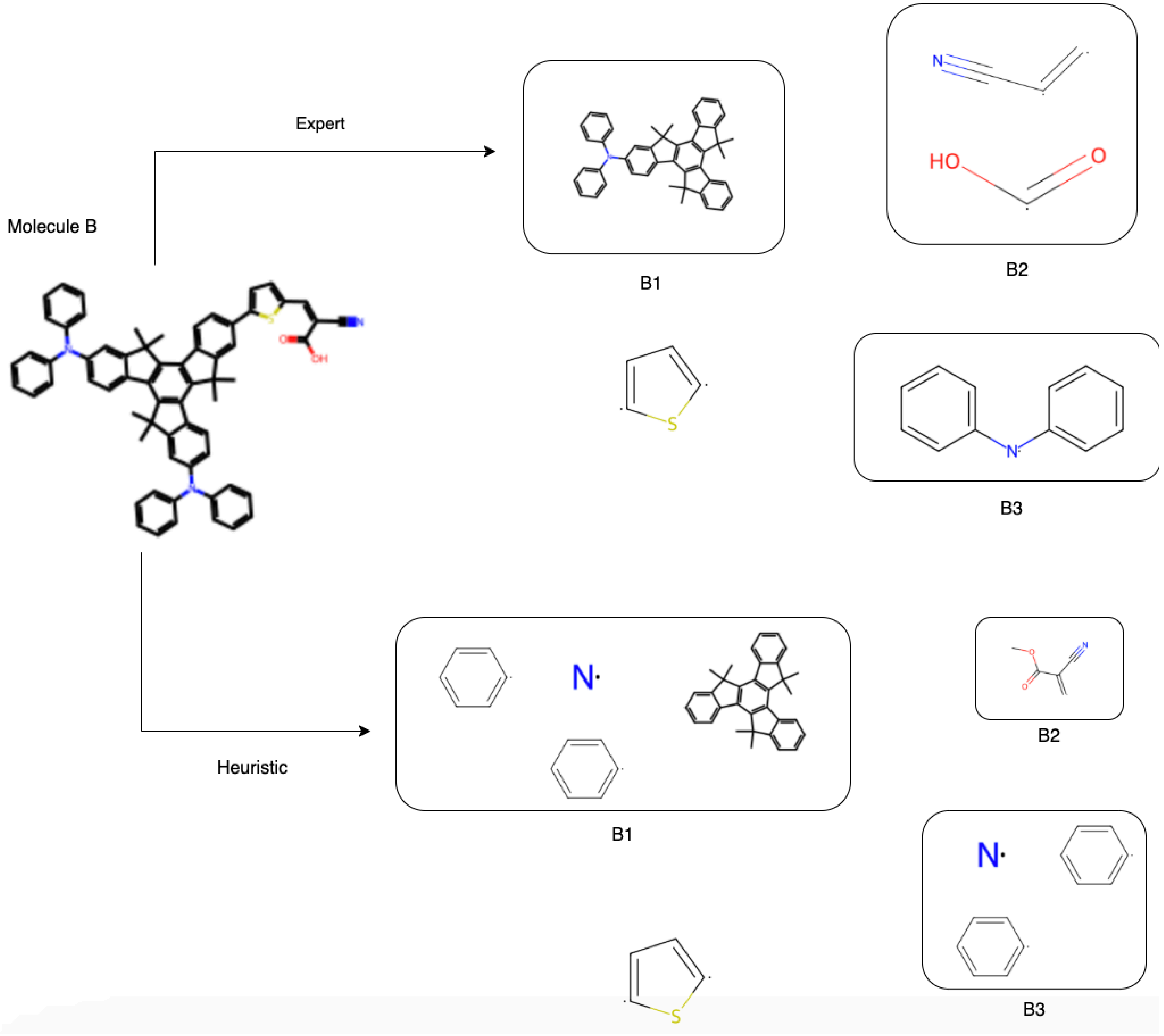
#2: red group is 3;9 #3: red group is 8;13

#4: red group is 9,10,11,12,17,18

# Our Workflow

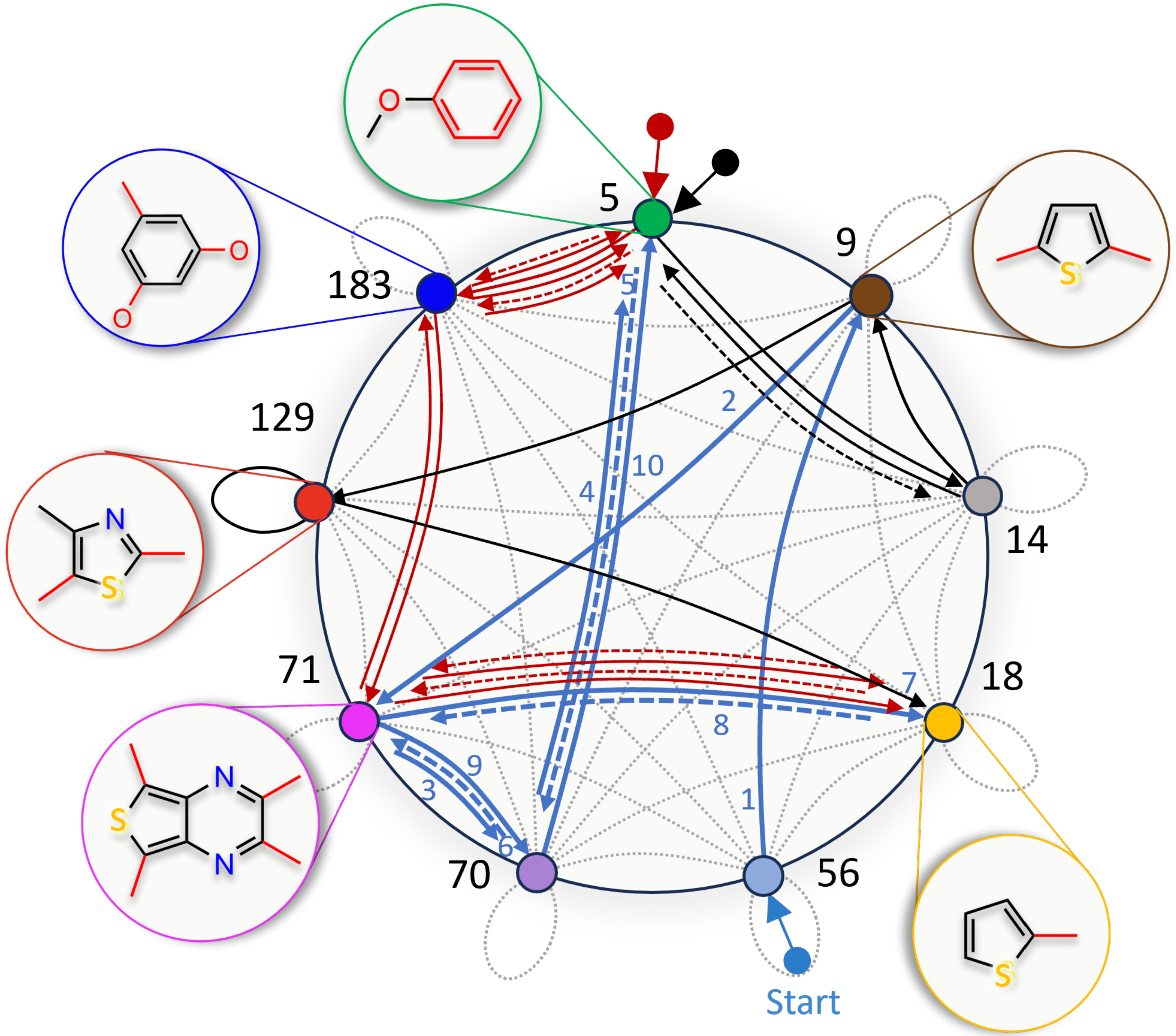
## Automated approach

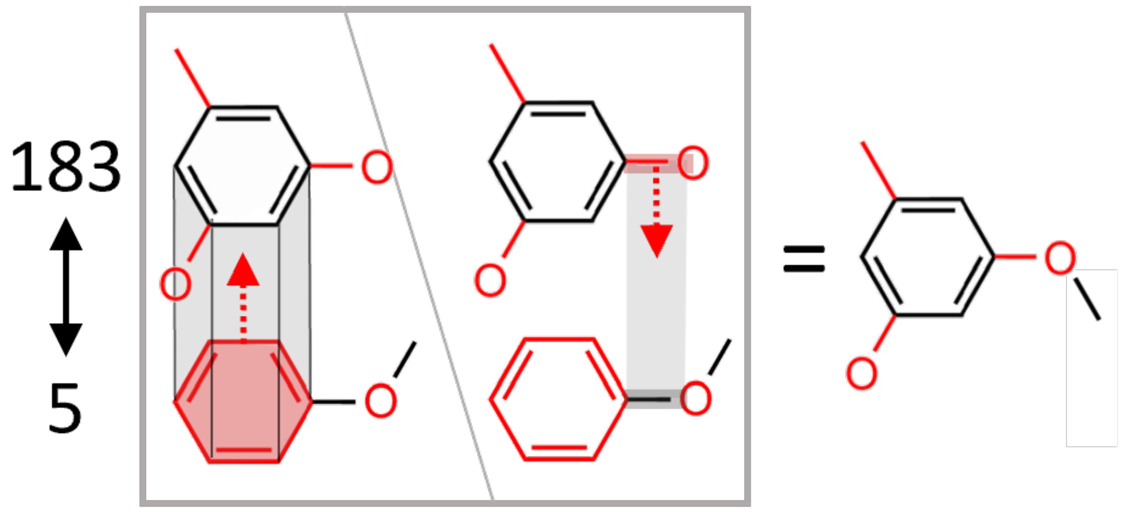
- Requires no expert input
- Heuristic fragmentation algorithm + pick simplest context (e.g. single atom)
- Other algorithms (e.g. BRICS)



# Method

## Transition Grammar Over Motif Graph





Each motif has black (base) and red (context).

We match motifs A and B iff:

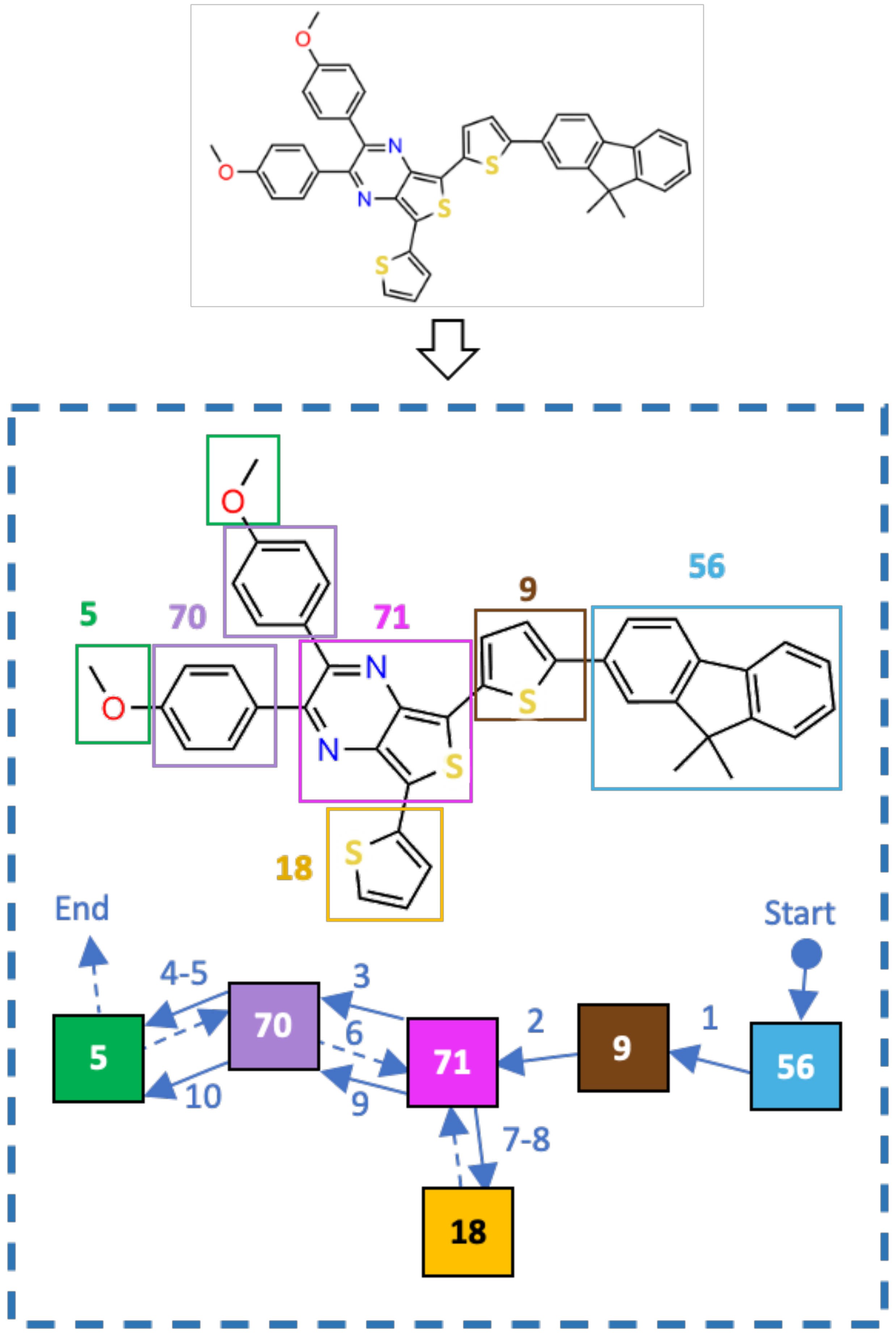
- B's red ~ subgraph of A's black (b1)
- A's red ~ subgraph of B's black (b2)
- A's red U b1 ~ B's red U b2
- A's red U b1 is connected

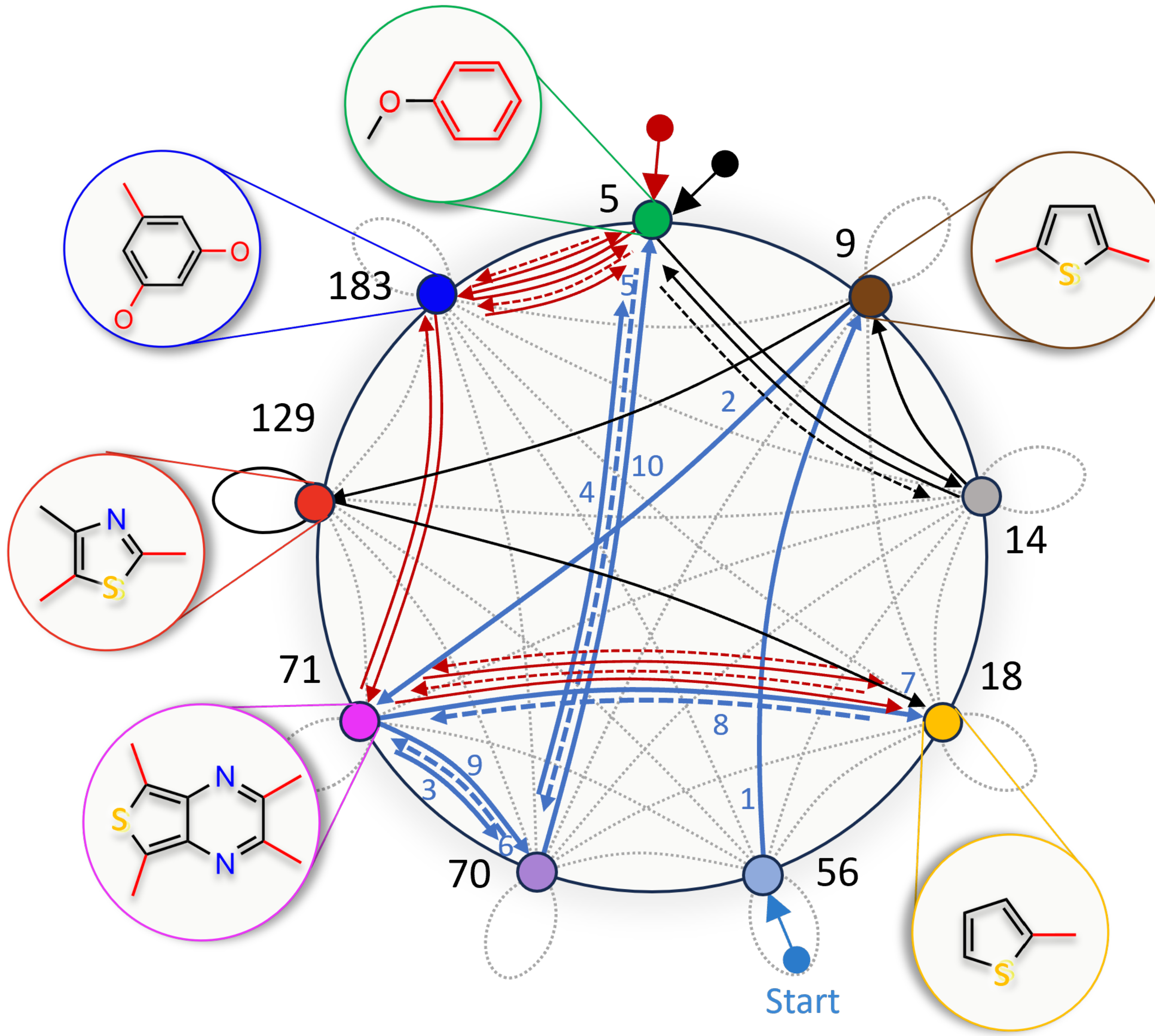
Obtain transition graph grammar rules:

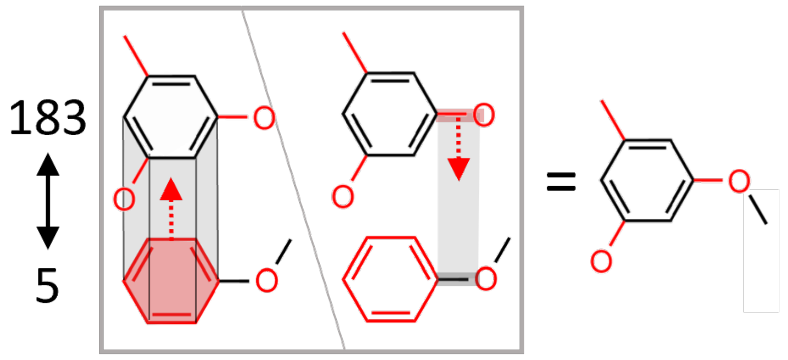
- A -> join(A, B)
- B -> join(B, A)

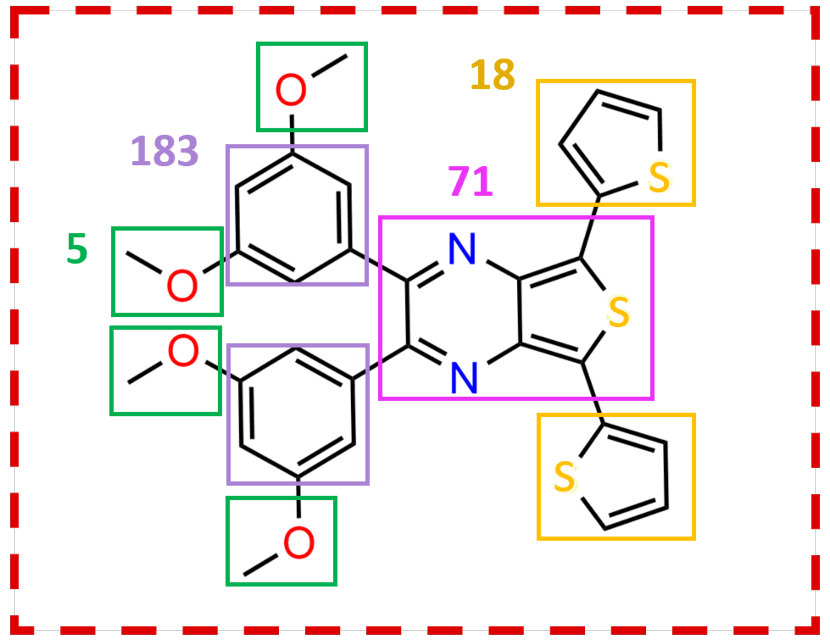
# Method (cont.)

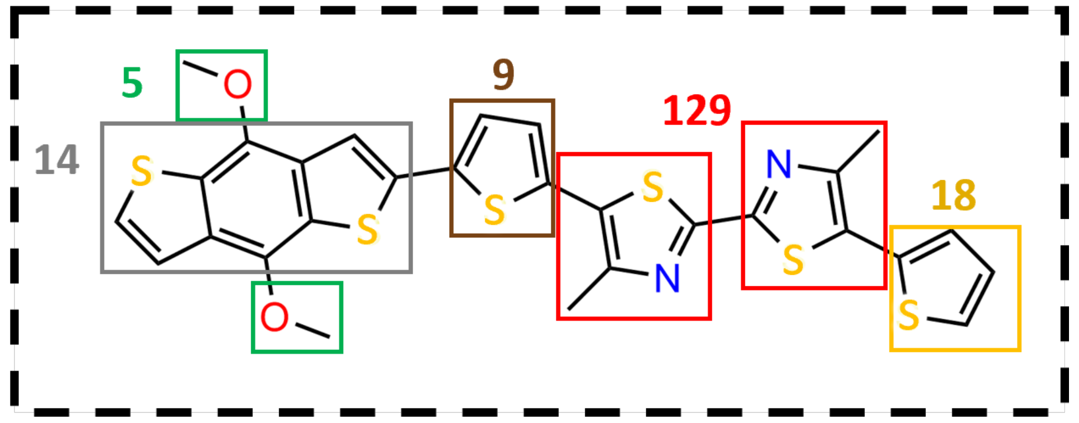
## Novel Representation Using Derivation of Grammar









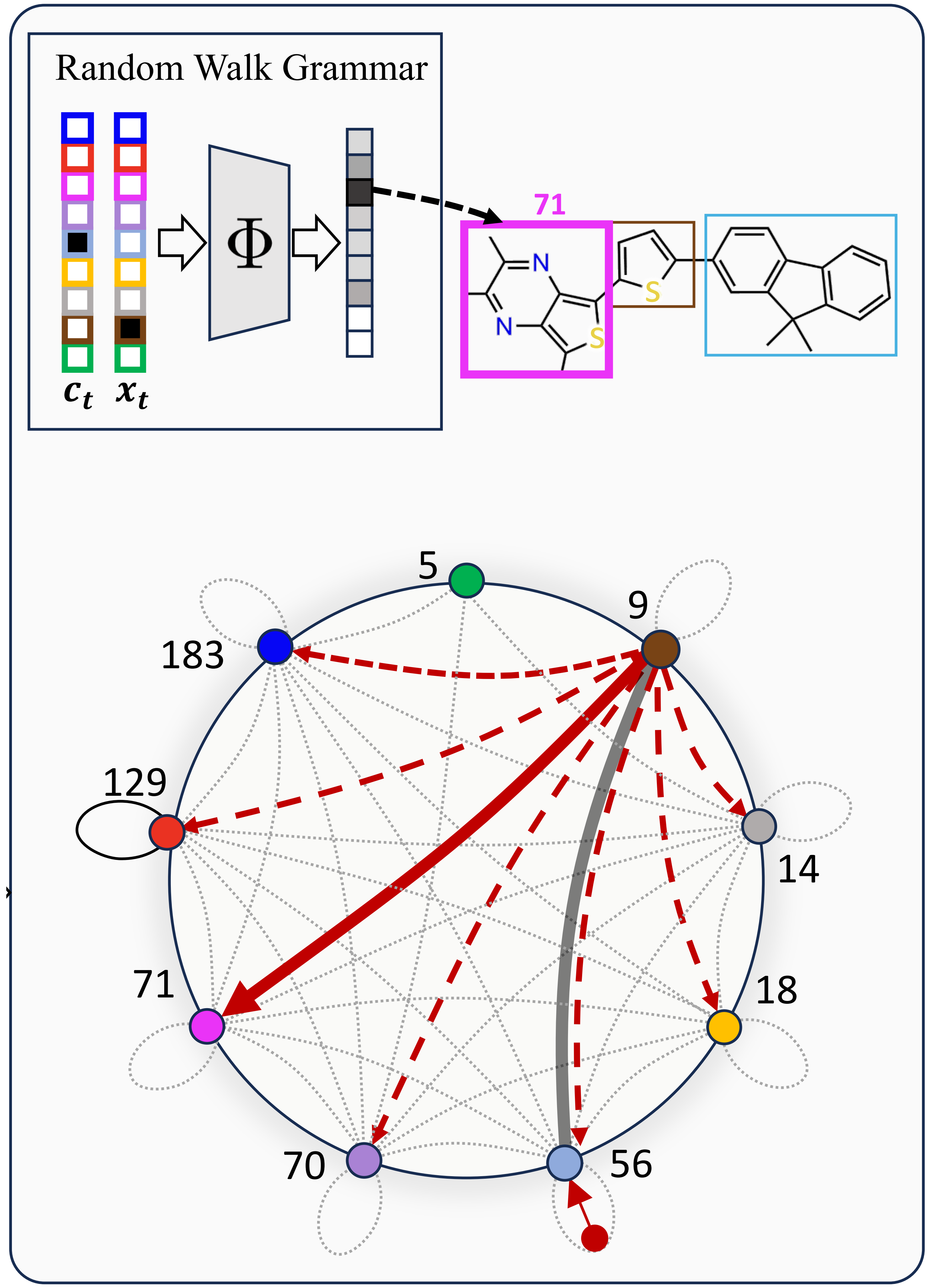


**Random Walk**: 56 → 9 → 71 → 70 → 5 → 70 → 71 → 18 → 71 → 70:1 → 5:1 (:1 means duplicate, not return)
**String Notation**: 56 → 9 → 71[→ 70 → 5, → 18] → 70:1 → 5:1

**Graph Theory Interpretation**: Euler path of an edge-induced subgraph of the Motif Graph

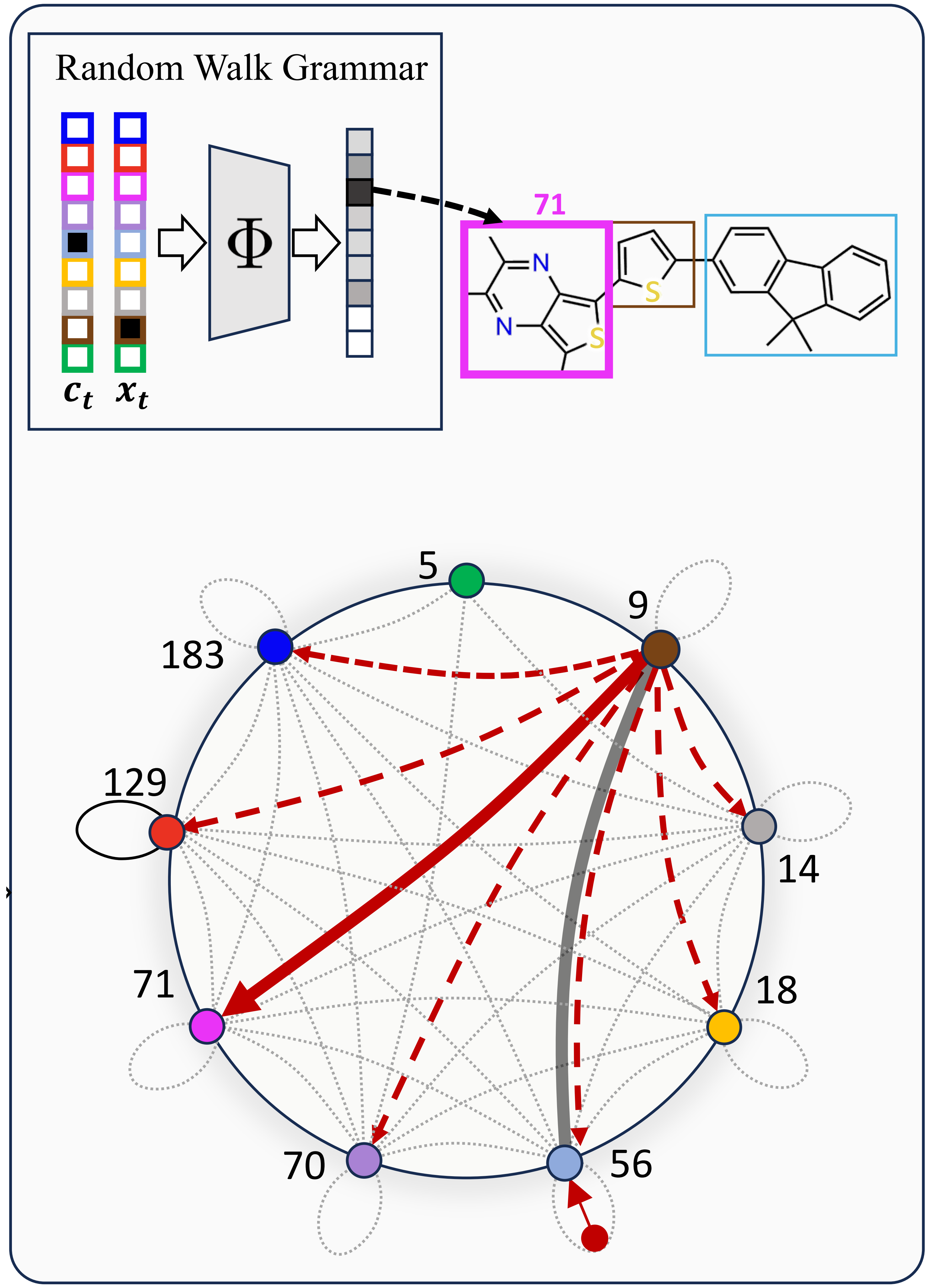
# Method (cont.)

## Learning Grammar by Taking Random Walks



# Method (cont.)

## Learning Grammar by Taking Random Walks

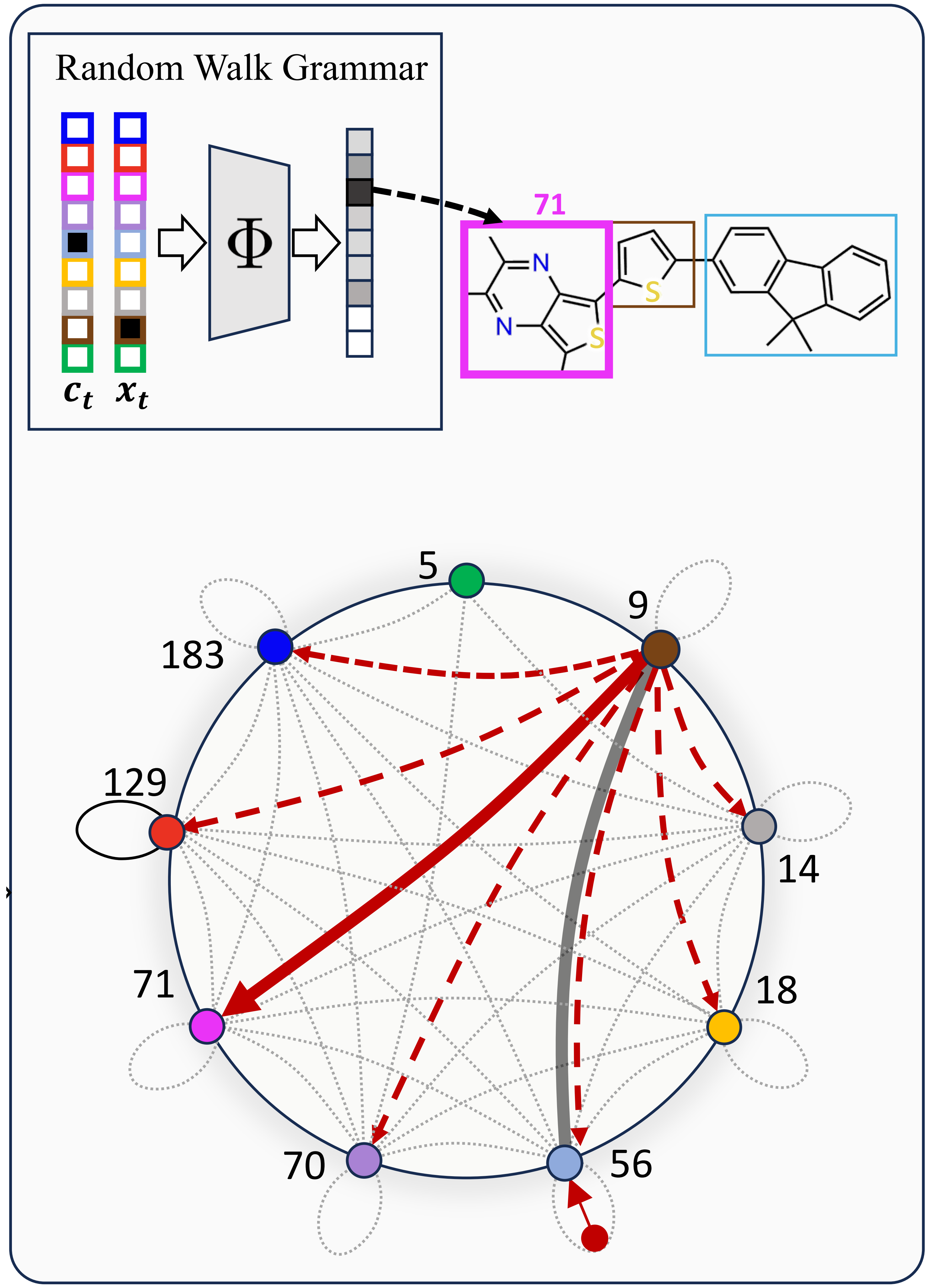


- **Idea 1:** Model random walks as stochastic discrete process with the Graph Heat Diffusion equation, where L is the Laplacian

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = L(\Phi, t)x_t$$

# Method (cont.)

## Learning Grammar by Taking Random Walks



- **Idea 1:** Model random walks as stochastic discrete process with the Graph Heat Diffusion equation, where L is the Laplacian

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = L(\Phi, t)x_t$$

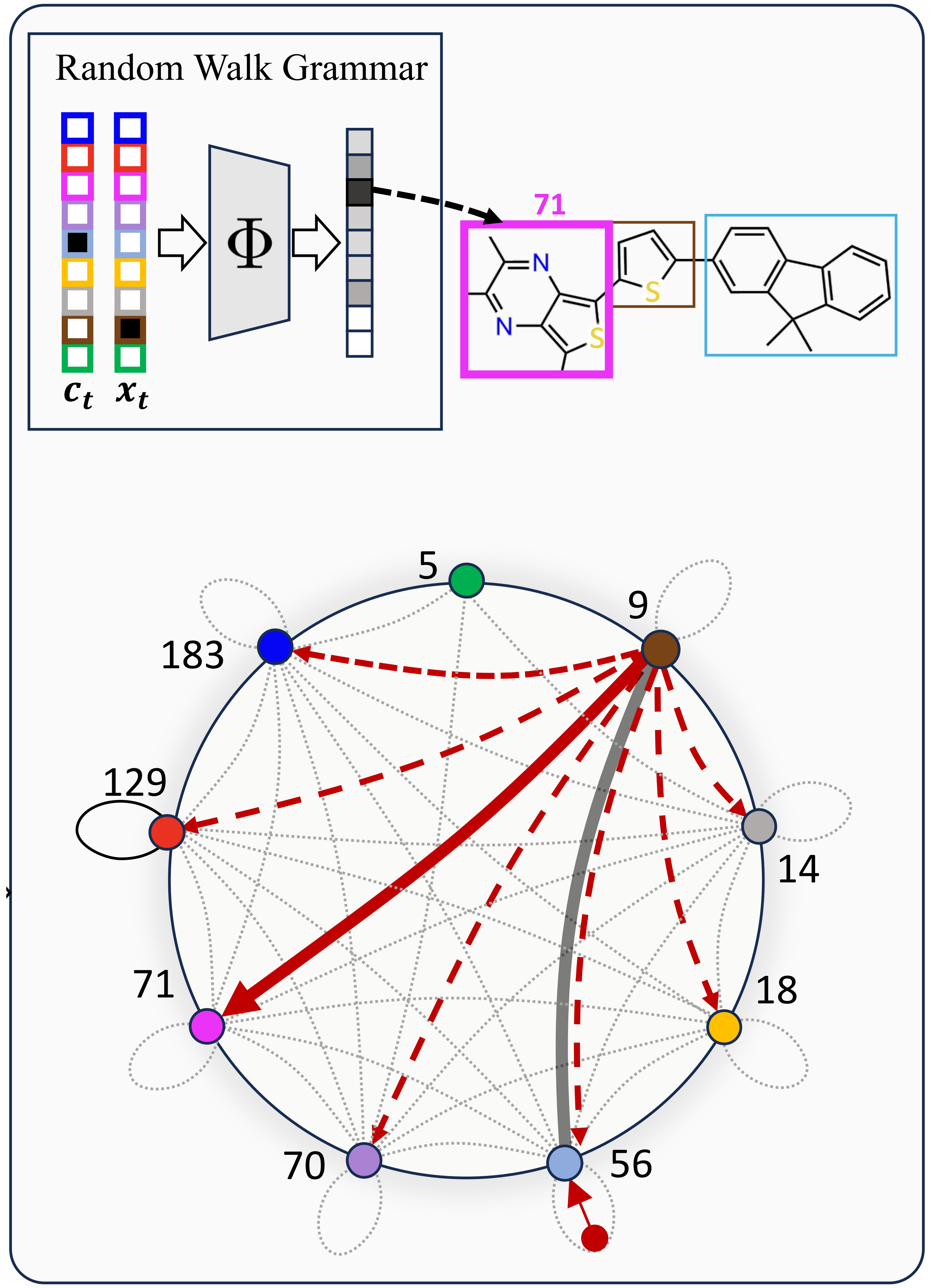
- **Idea 2:** Make the Laplacian learnable, conditioned on a set-based memory c.

$$L(\Phi, t) = D - \hat{W}(t), \hat{W}(t) = W + h(c_t; \phi)$$

$$c^{(t+1)} \leftarrow \frac{t}{t+1} \cdot c^{(t)} + \frac{1}{t+1} \cdot x^{(t)}$$

# Method (cont.)

## Learning Grammar by Taking Random Walks



- **Idea 1:** Model random walks as stochastic discrete process with the Graph Heat Diffusion equation, where L is the Laplacian

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = L(\Phi, t)x_t$$

- **Idea 2:** Make the Laplacian learnable, conditioned on a set-based memory c.

$$L(\Phi, t) = D - \hat{W}(t), \hat{W}(t) = W + h(c_t; \phi)$$

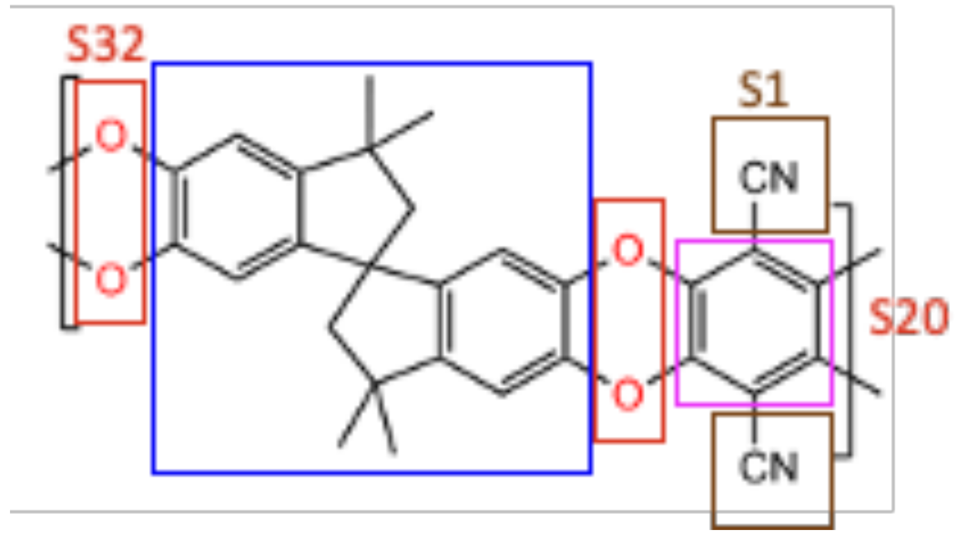
$$c^{(t+1)} \leftarrow \frac{t}{t+1} \cdot c^{(t)} + \frac{1}{t+1} \cdot x^{(t)}$$

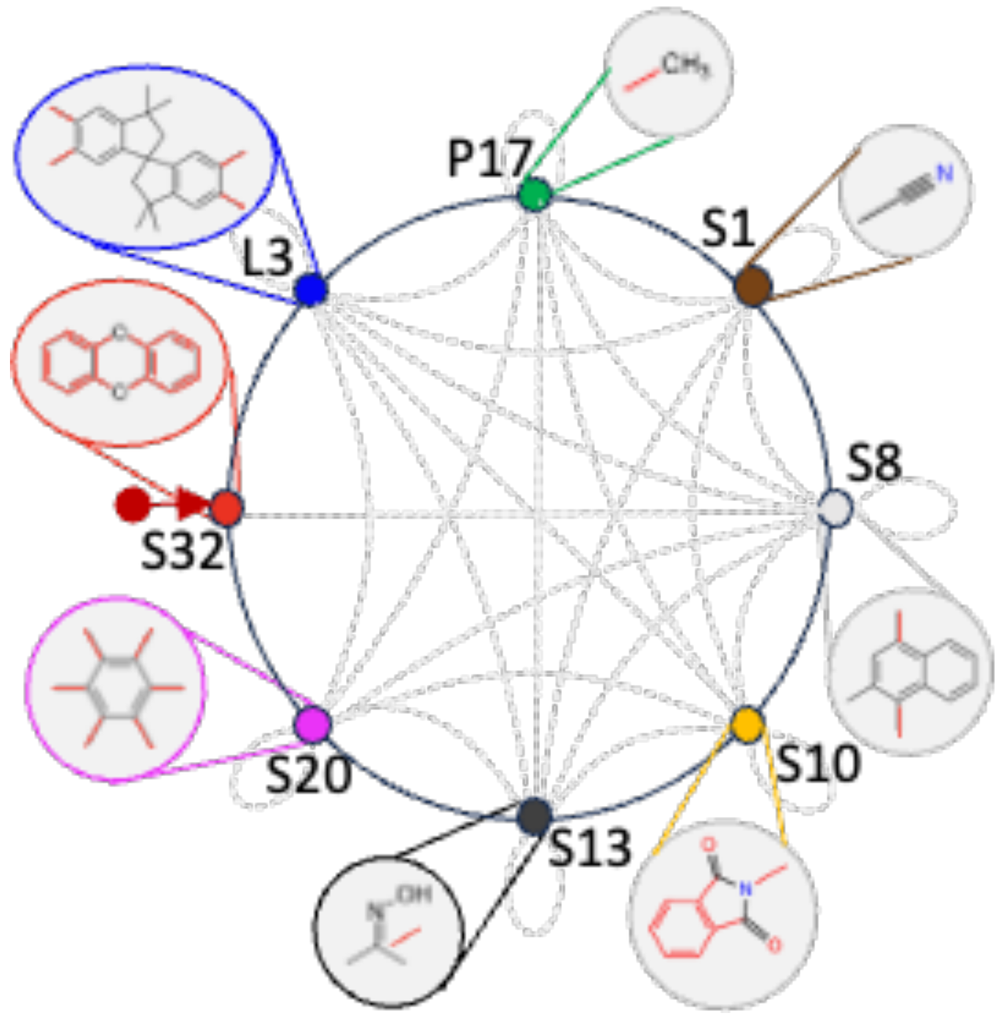
The learnable parameters are $\Phi = (W, \phi).$

Train parameters to maximize expectation of seeing the data.

# Method (cont.)

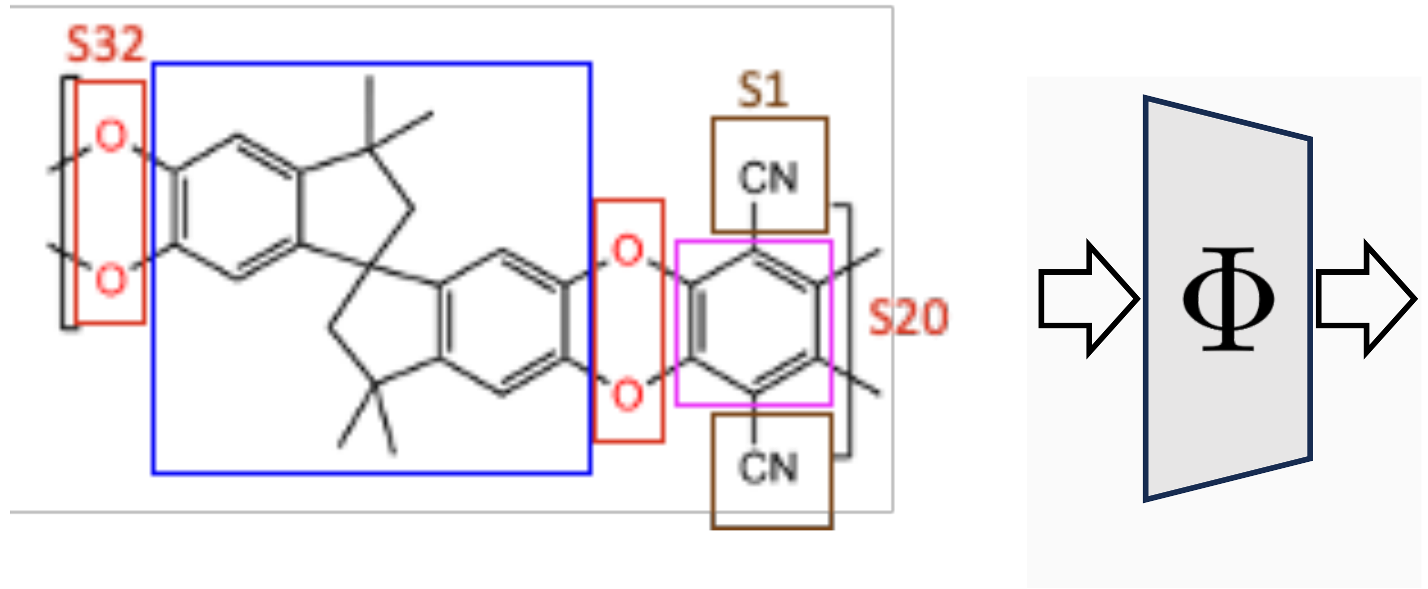
## Grammar-induced Representation for Property Prediction

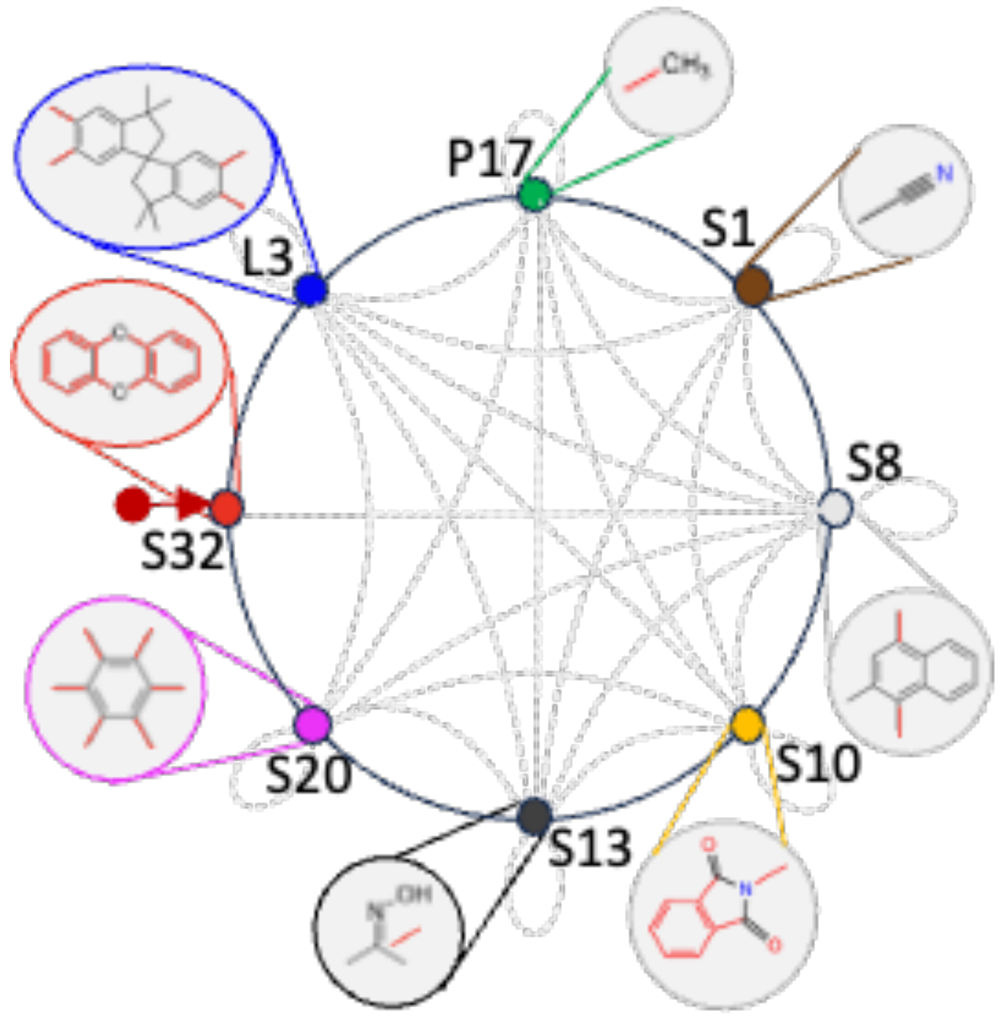




# Method (cont.)

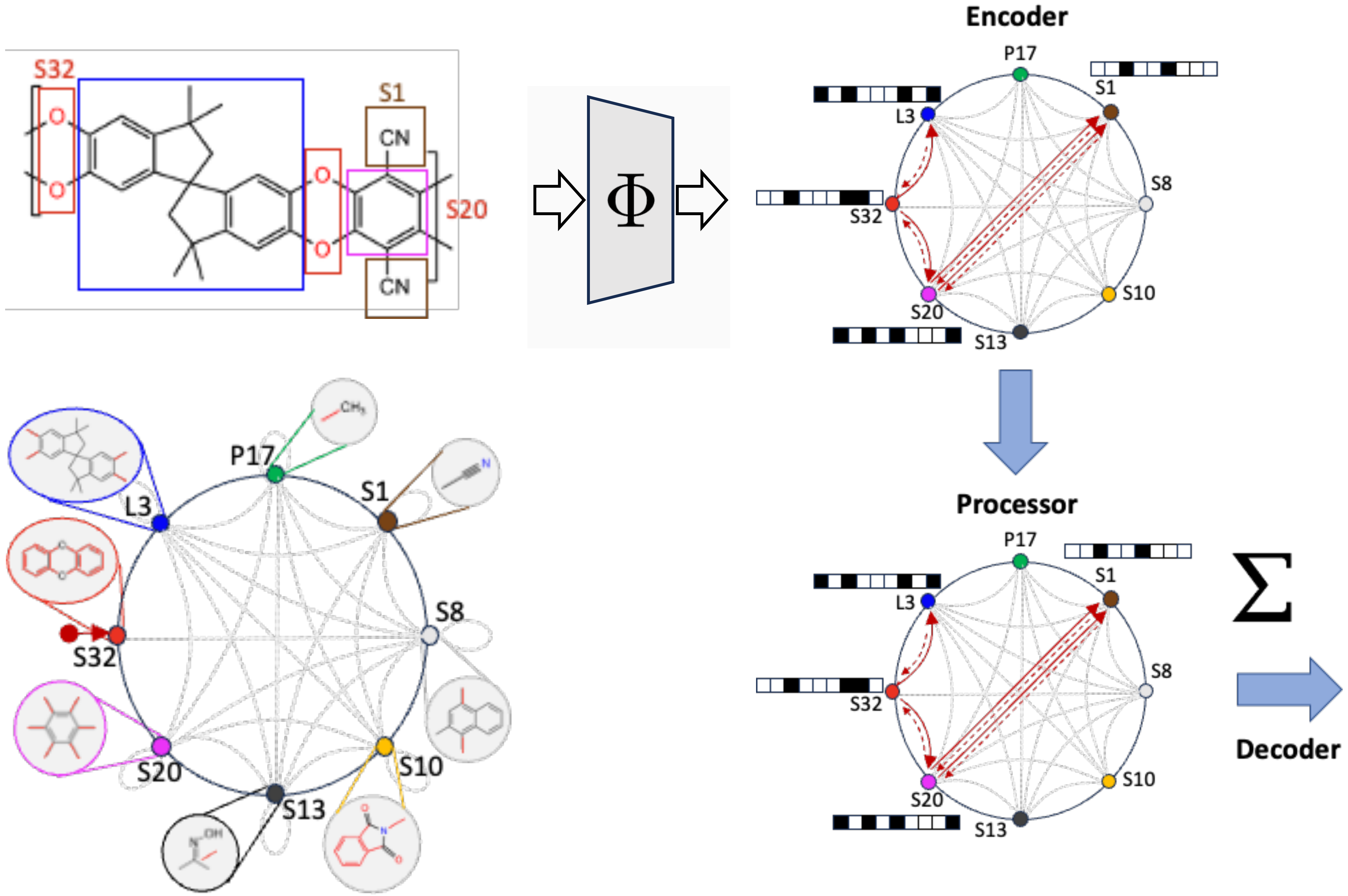
## Grammar-induced Representation for Property Prediction





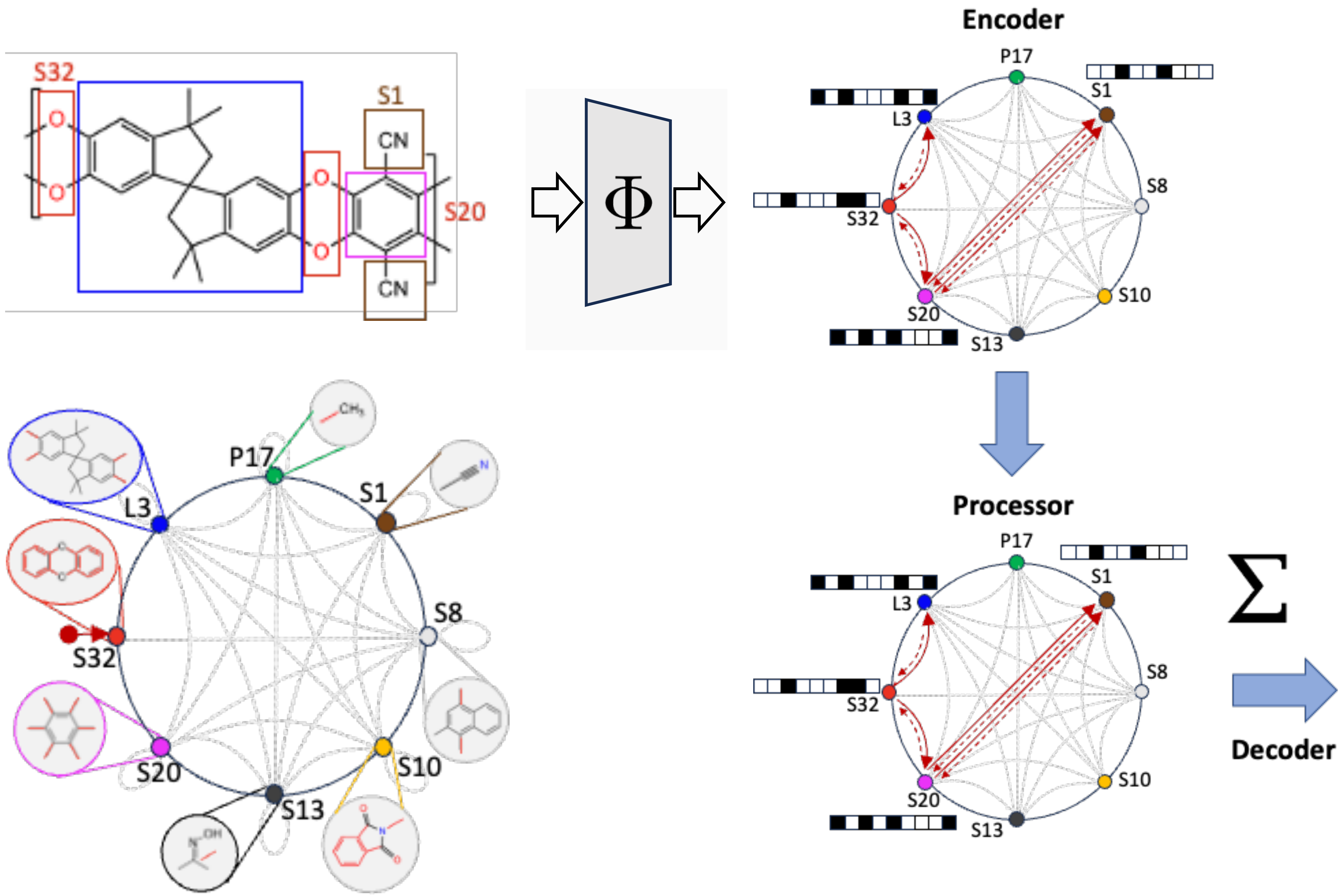
# Method (cont.)

## Grammar-induced Representation for Property Prediction



# Method (cont.)

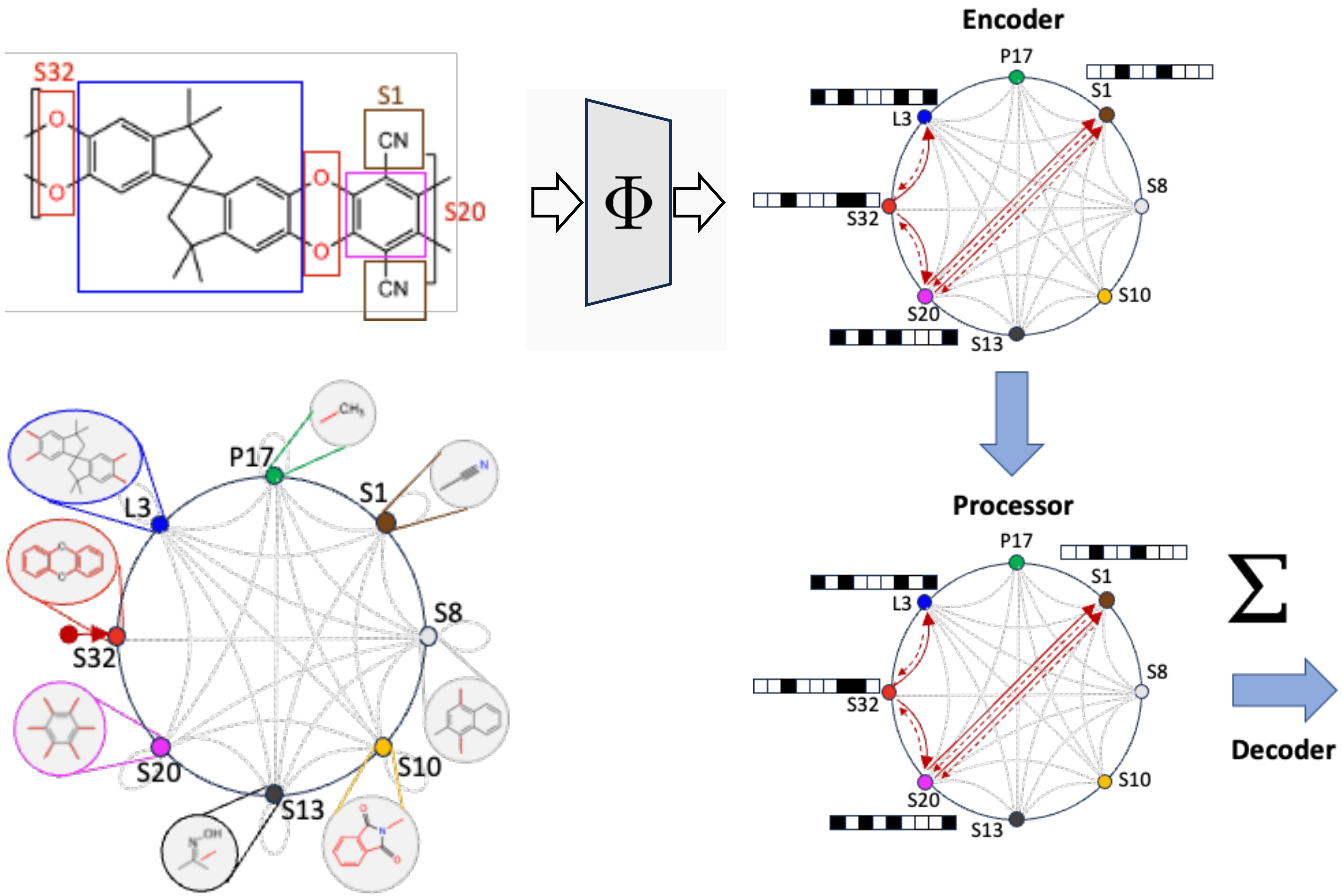
## Grammar-induced Representation for Property Prediction



- **Idea 1:** $\Phi$ induces a representation which can be fed into a downstream graph neural network(; $\Theta$, $\theta$) to predict properties.

# Method (cont.)

## Grammar-induced Representation for Property Prediction



- **Idea 1:** $\Phi$ induces a representation which can be fed into a downstream graph neural network(; $\Theta$, $\theta$) to predict properties.

- **Idea 2:** We can train ($\Phi$ ,$\Theta$, $\theta$ ) end-to-end!

representation

$$\tilde{\mathcal{L}}(D;\Theta,\theta,\Phi) = \mathbb{E}_{\hat{H}_M(\cdot;\Phi)}[\mathcal{L}(f_\theta(\mathcal{F}_\Theta(\hat{H}_M, y)]$$

$$= \frac{1}{|D|}\sum_{i=1}^{|D|}\mathcal{L}(f_\theta(\mathcal{F}_\Theta(\hat{H}_M^{(i)})), y^{(i)})$$

Dataset

properties

# Results

## Data-Efficient Molecule Generation

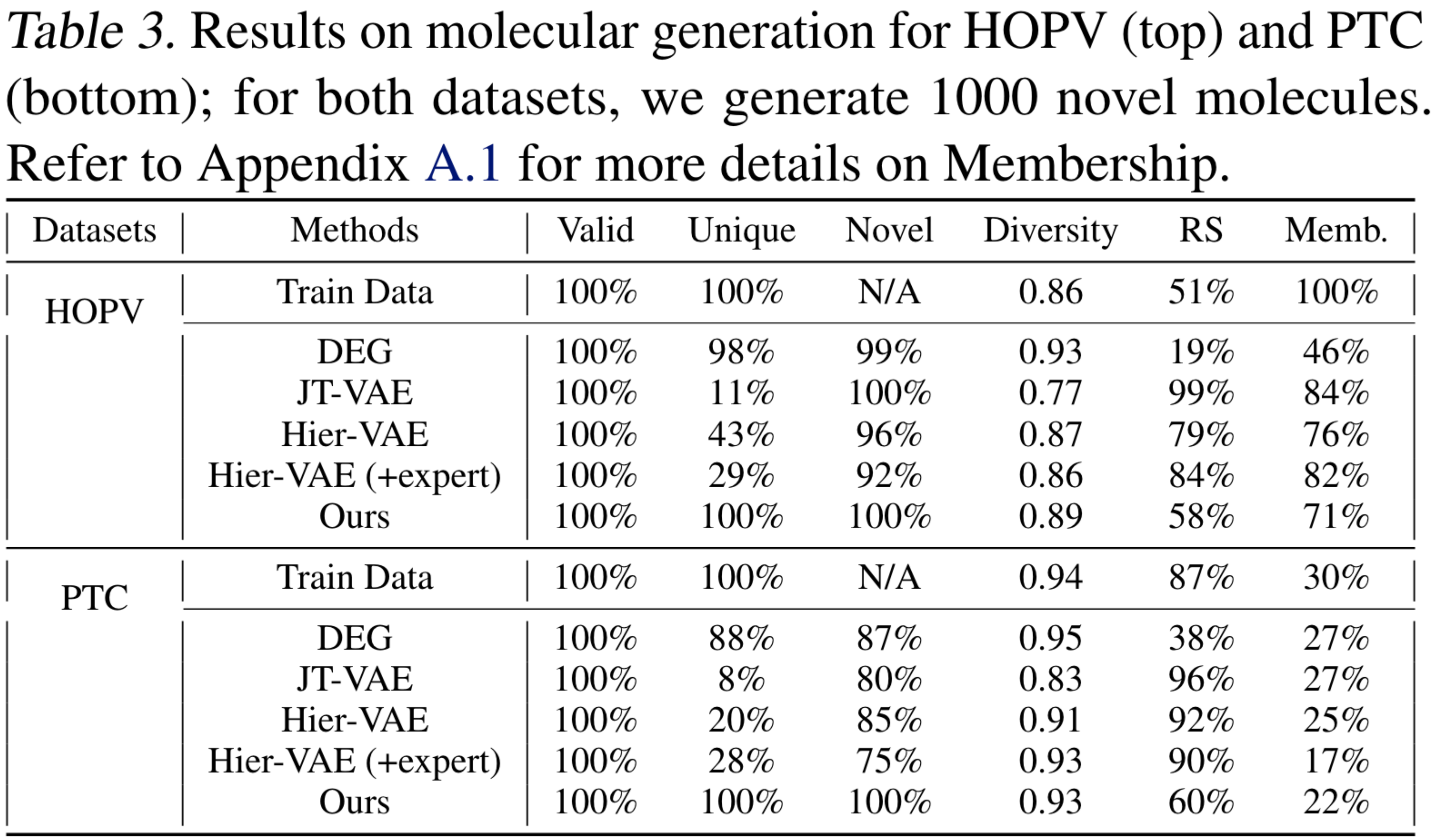


*Table 3.* Results on molecular generation for HOPV (top) and PTC (bottom); for both datasets, we generate 1000 novel molecules. Refer to Appendix A.1 for more details on Membership.

| Datasets | Methods | Valid | Unique | Novel | Diversity | RS | Memb. |
|---|---|---|---|---|---|---|---|
| HOPV | Train Data | 100% | 100% | N/A | 0.86 | 51% | 100% |
| | DEG | 100% | 98% | 99% | 0.93 | 19% | 46% |
| | JT-VAE | 100% | 11% | 100% | 0.77 | 99% | 84% |
| | Hier-VAE | 100% | 43% | 96% | 0.87 | 79% | 76% |
| | Hier-VAE (+expert) | 100% | 29% | 92% | 0.86 | 84% | 82% |
| | Ours | 100% | 100% | 100% | 0.89 | 58% | 71% |
| PTC | Train Data | 100% | 100% | N/A | 0.94 | 87% | 30% |
| | DEG | 100% | 88% | 87% | 0.95 | 38% | 27% |
| | JT-VAE | 100% | 8% | 80% | 0.83 | 96% | 27% |
| | Hier-VAE | 100% | 20% | 85% | 0.91 | 92% | 25% |
| | Hier-VAE (+expert) | 100% | 28% | 75% | 0.93 | 90% | 17% |
| | Ours | 100% | 100% | 100% | 0.93 | 60% | 22% |

- Ours generates *more diverse* molecules than training set

# Results

## Data-Efficient Molecule Generation

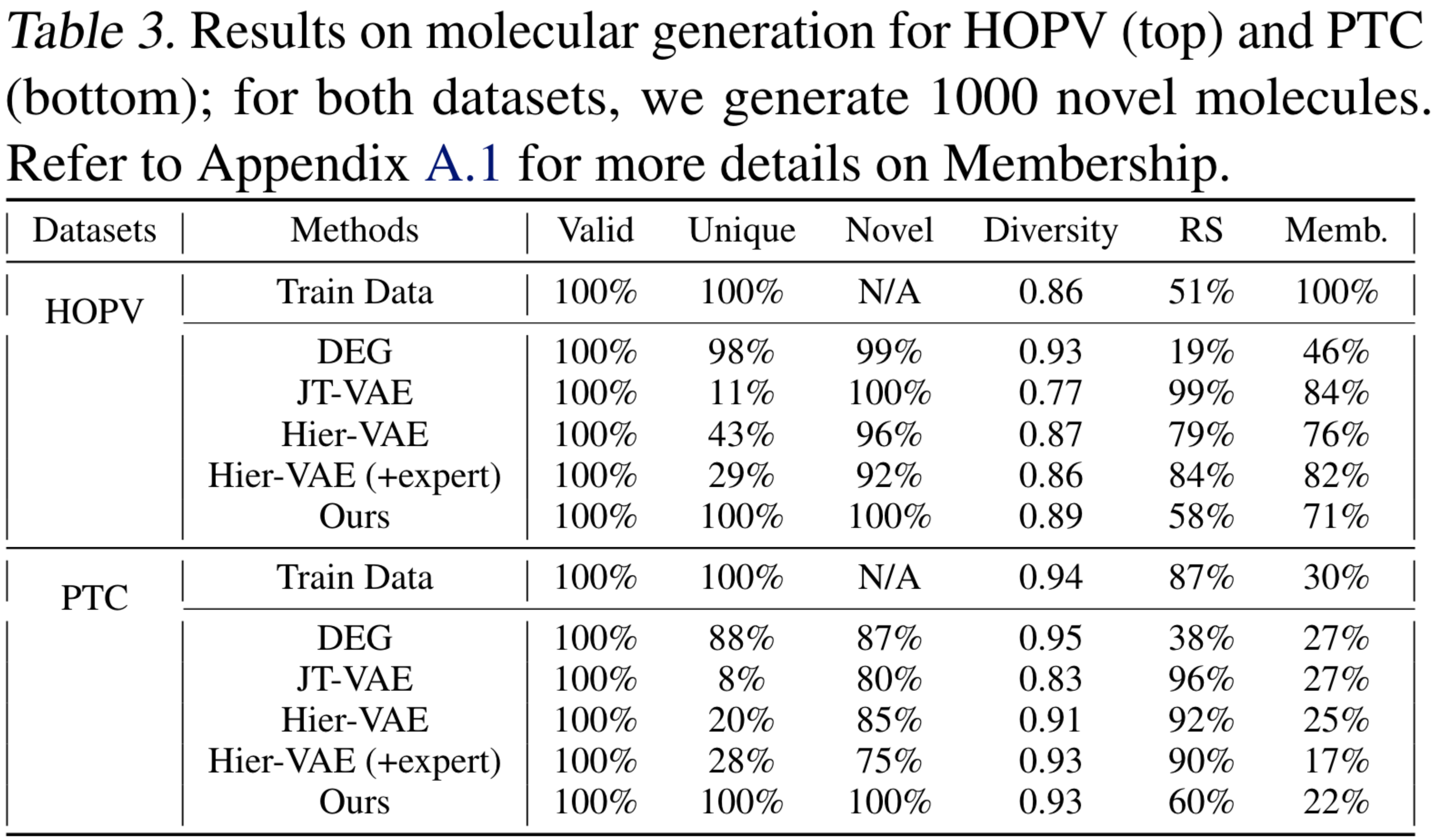


*Table 3.* Results on molecular generation for HOPV (top) and PTC (bottom); for both datasets, we generate 1000 novel molecules. Refer to Appendix A.1 for more details on Membership.

| Datasets | Methods | Valid | Unique | Novel | Diversity | RS | Memb. |
|---|---|---|---|---|---|---|---|
| HOPV | Train Data | 100% | 100% | N/A | 0.86 | 51% | 100% |
| | DEG | 100% | 98% | 99% | 0.93 | 19% | 46% |
| | JT-VAE | 100% | 11% | 100% | 0.77 | 99% | 84% |
| | Hier-VAE | 100% | 43% | 96% | 0.87 | 79% | 76% |
| | Hier-VAE (+expert) | 100% | 29% | 92% | 0.86 | 84% | 82% |
| | Ours | 100% | 100% | 100% | 0.89 | 58% | 71% |
| PTC | Train Data | 100% | 100% | N/A | 0.94 | 87% | 30% |
| | DEG | 100% | 88% | 87% | 0.95 | 38% | 27% |
| | JT-VAE | 100% | 8% | 80% | 0.83 | 96% | 27% |
| | Hier-VAE | 100% | 20% | 85% | 0.91 | 92% | 25% |
| | Hier-VAE (+expert) | 100% | 28% | 75% | 0.93 | 90% | 17% |
| | Ours | 100% | 100% | 100% | 0.93 | 60% | 22% |

- Ours generates *more diverse* molecules than training set
- Ours generates significantly *more synthesizable* molecules than previous grammar-based SOTA (DEG)

# Results

## Data-Efficient Molecule Generation

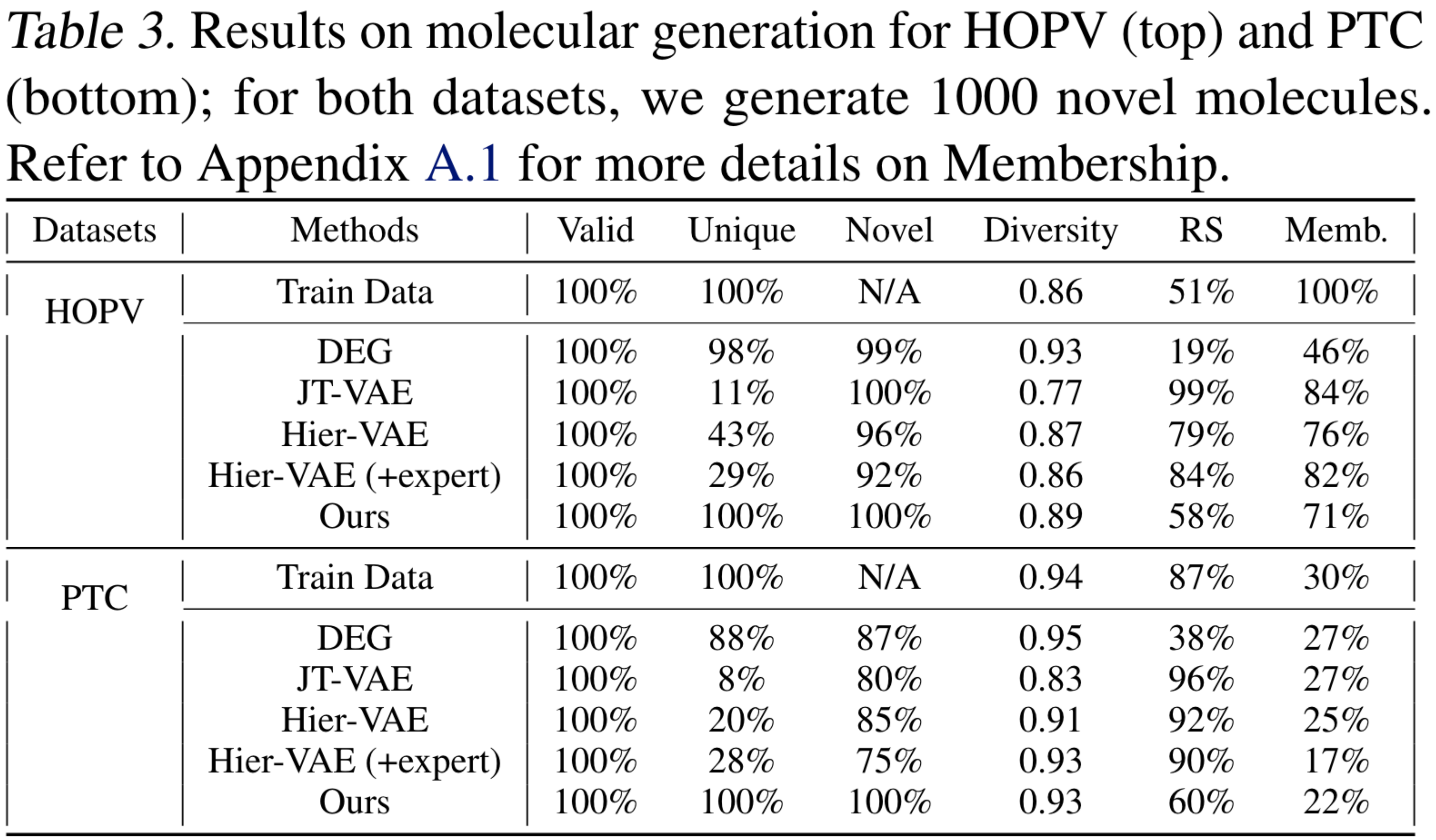


*Table 3.* Results on molecular generation for HOPV (top) and PTC (bottom); for both datasets, we generate 1000 novel molecules. Refer to Appendix A.1 for more details on Membership.

| Datasets | Methods | Valid | Unique | Novel | Diversity | RS | Memb. |
|---|---|---|---|---|---|---|---|
| HOPV | Train Data | 100% | 100% | N/A | 0.86 | 51% | 100% |
| | DEG | 100% | 98% | 99% | 0.93 | 19% | 46% |
| | JT-VAE | 100% | 11% | 100% | 0.77 | 99% | 84% |
| | Hier-VAE | 100% | 43% | 96% | 0.87 | 79% | 76% |
| | Hier-VAE (+expert) | 100% | 29% | 92% | 0.86 | 84% | 82% |
| | Ours | 100% | 100% | 100% | 0.89 | 58% | 71% |
| PTC | Train Data | 100% | 100% | N/A | 0.94 | 87% | 30% |
| | DEG | 100% | 88% | 87% | 0.95 | 38% | 27% |
| | JT-VAE | 100% | 8% | 80% | 0.83 | 96% | 27% |
| | Hier-VAE | 100% | 20% | 85% | 0.91 | 92% | 25% |
| | Hier-VAE (+expert) | 100% | 28% | 75% | 0.93 | 90% | 17% |
| | Ours | 100% | 100% | 100% | 0.93 | 60% | 22% |

- Ours generates *more diverse* molecules than training set

- Ours generates significantly *more synthesizable* molecules than previous grammar-based SOTA (DEG)

- Ours generates more unique, novel and diverse molecules compared to VAE-based methods

# Results

## Data-Efficient Molecule Generation

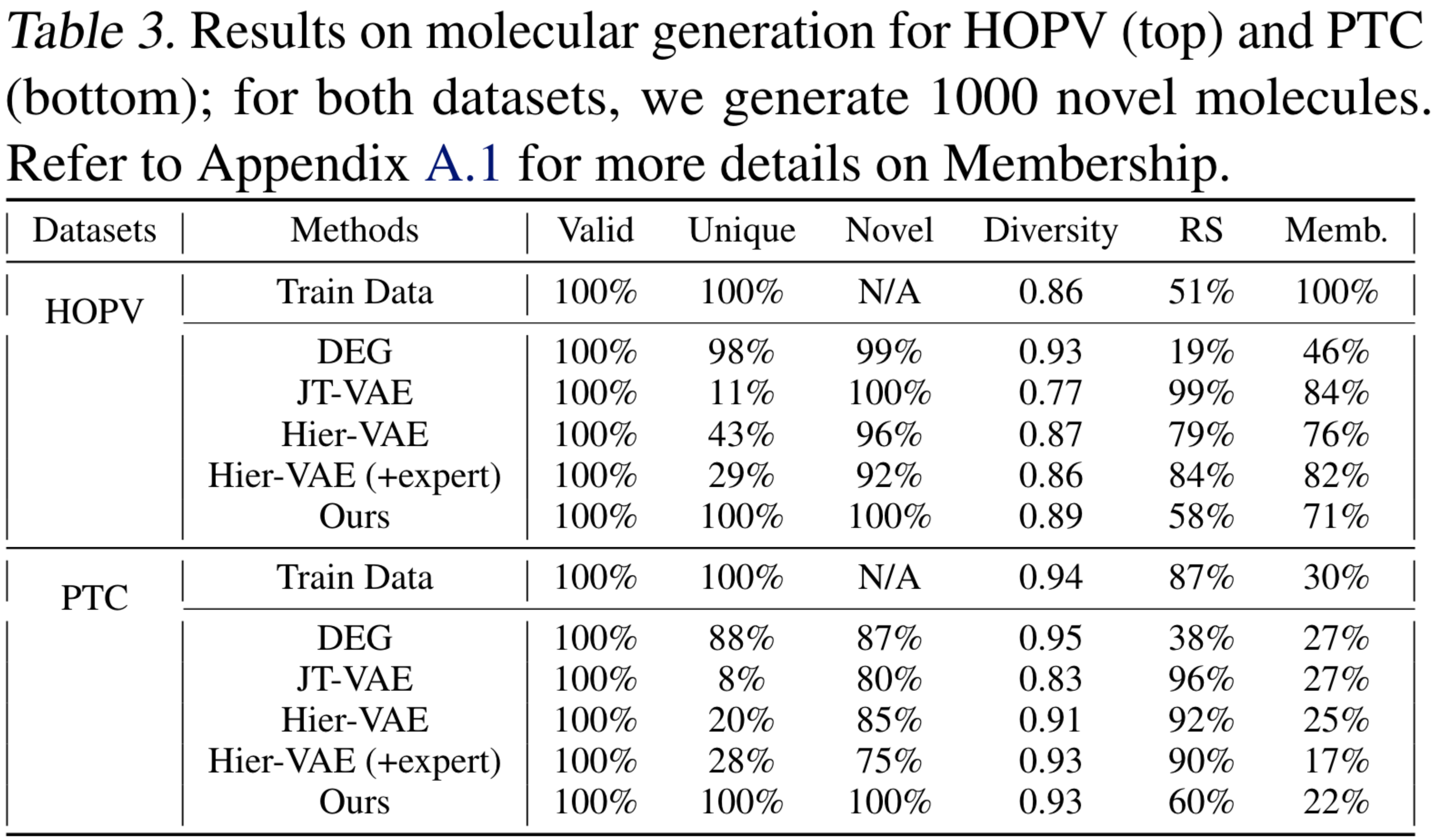


*Table 3.* Results on molecular generation for HOPV (top) and PTC (bottom); for both datasets, we generate 1000 novel molecules. Refer to Appendix A.1 for more details on Membership.

| Datasets | Methods | Valid | Unique | Novel | Diversity | RS | Memb. |
|---|---|---|---|---|---|---|---|
| HOPV | Train Data | 100% | 100% | N/A | 0.86 | 51% | 100% |
| | DEG | 100% | 98% | 99% | 0.93 | 19% | 46% |
| | JT-VAE | 100% | 11% | 100% | 0.77 | 99% | 84% |
| | Hier-VAE | 100% | 43% | 96% | 0.87 | 79% | 76% |
| | Hier-VAE (+expert) | 100% | 29% | 92% | 0.86 | 84% | 82% |
| | Ours | 100% | 100% | 100% | 0.89 | 58% | 71% |
| PTC | Train Data | 100% | 100% | N/A | 0.94 | 87% | 30% |
| | DEG | 100% | 88% | 87% | 0.95 | 38% | 27% |
| | JT-VAE | 100% | 8% | 80% | 0.83 | 96% | 27% |
| | Hier-VAE | 100% | 20% | 85% | 0.91 | 92% | 25% |
| | Hier-VAE (+expert) | 100% | 28% | 75% | 0.93 | 90% | 17% |
| | Ours | 100% | 100% | 100% | 0.93 | 60% | 22% |

- Ours generates *more diverse* molecules than training set
- Ours generates significantly *more synthesizable* molecules than previous grammar-based SOTA (DEG)
- Ours generates more unique, novel and diverse molecules compared to VAE-based methods
- VAE-based methods cannot utilize expert motifs as well

# Results (cont.)

## Data-Efficient Property Prediction

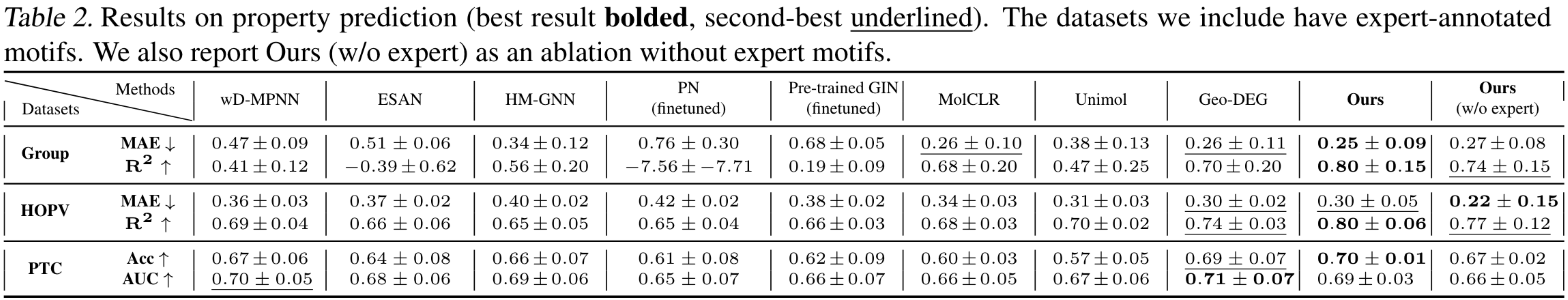


*Table 2.* Results on property prediction (best result **bolded**, second-best underlined). The datasets we include have expert-annotated motifs. We also report Ours (w/o expert) as an ablation without expert motifs.

| Datasets \ Methods | | wD-MPNN | ESAN | HM-GNN | PN (finetuned) | Pre-trained GIN (finetuned) | MolCLR | Unimol | Geo-DEG | **Ours** | **Ours** (w/o expert) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Group** | **MAE** ↓ | $0.47 \pm 0.09$ | $0.51 \pm 0.06$ | $0.34 \pm 0.12$ | $0.76 \pm 0.30$ | $0.68 \pm 0.05$ | $0.26 \pm 0.10$ | $0.38 \pm 0.13$ | $0.26 \pm 0.11$ | $\mathbf{0.25 \pm 0.09}$ | $0.27 \pm 0.08$ |
| | $\mathbf{R^2}$ ↑ | $0.41 \pm 0.12$ | $-0.39 \pm 0.62$ | $0.56 \pm 0.20$ | $-7.56 \pm -7.71$ | $0.19 \pm 0.09$ | $0.68 \pm 0.20$ | $0.47 \pm 0.25$ | $0.70 \pm 0.20$ | $\mathbf{0.80 \pm 0.15}$ | $0.74 \pm 0.15$ |
| **HOPV** | **MAE** ↓ | $0.36 \pm 0.03$ | $0.37 \pm 0.02$ | $0.40 \pm 0.02$ | $0.42 \pm 0.02$ | $0.38 \pm 0.02$ | $0.34 \pm 0.03$ | $0.31 \pm 0.03$ | $0.30 \pm 0.02$ | $0.30 \pm 0.05$ | $\mathbf{0.22 \pm 0.15}$ |
| | $\mathbf{R^2}$ ↑ | $0.69 \pm 0.04$ | $0.66 \pm 0.06$ | $0.65 \pm 0.05$ | $0.65 \pm 0.04$ | $0.66 \pm 0.03$ | $0.68 \pm 0.03$ | $0.70 \pm 0.02$ | $0.74 \pm 0.03$ | $\mathbf{0.80 \pm 0.06}$ | $0.77 \pm 0.12$ |
| **PTC** | **Acc** ↑ | $0.67 \pm 0.06$ | $0.64 \pm 0.08$ | $0.66 \pm 0.07$ | $0.61 \pm 0.08$ | $0.62 \pm 0.09$ | $0.60 \pm 0.03$ | $0.57 \pm 0.05$ | $0.69 \pm 0.07$ | $\mathbf{0.70 \pm 0.01}$ | $0.67 \pm 0.02$ |
| | **AUC** ↑ | $0.70 \pm 0.05$ | $0.68 \pm 0.06$ | $0.69 \pm 0.06$ | $0.65 \pm 0.07$ | $0.66 \pm 0.07$ | $0.66 \pm 0.05$ | $0.67 \pm 0.06$ | $\mathbf{0.71 \pm 0.07}$ | $0.69 \pm 0.03$ | $0.66 \pm 0.05$ |

- Our method *dominates* GNN baselines (including motif-based ones)

# Results (cont.)

## Data-Efficient Property Prediction

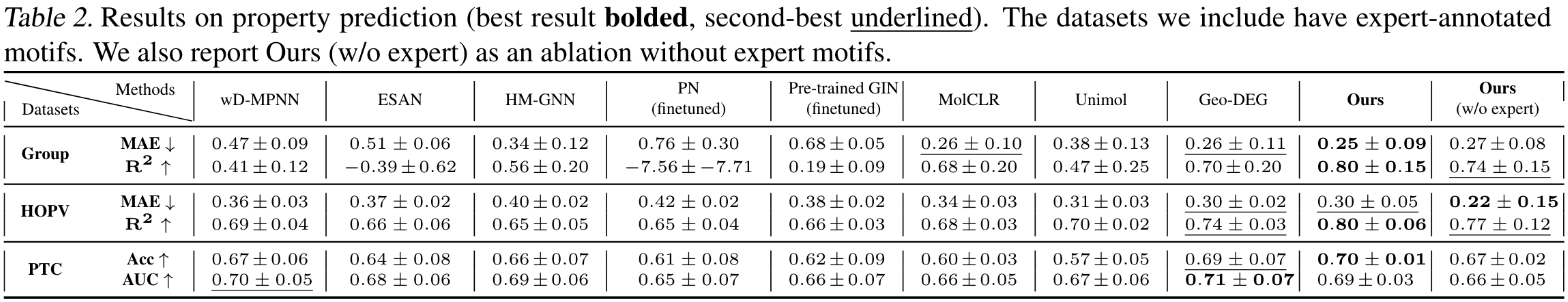


*Table 2.* Results on property prediction (best result **bolded**, second-best underlined). The datasets we include have expert-annotated motifs. We also report Ours (w/o expert) as an ablation without expert motifs.

| Methods<br>Datasets | | wD-MPNN | ESAN | HM-GNN | PN<br>(finetuned) | Pre-trained GIN<br>(finetuned) | MolCLR | Unimol | Geo-DEG | **Ours** | **Ours**<br>(w/o expert) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Group** | **MAE** ↓ | $0.47 \pm 0.09$ | $0.51 \pm 0.06$ | $0.34 \pm 0.12$ | $0.76 \pm 0.30$ | $0.68 \pm 0.05$ | $0.26 \pm 0.10$ | $0.38 \pm 0.13$ | $0.26 \pm 0.11$ | $\mathbf{0.25 \pm 0.09}$ | $0.27 \pm 0.08$ |
| | $\mathbf{R^2}$ ↑ | $0.41 \pm 0.12$ | $-0.39 \pm 0.62$ | $0.56 \pm 0.20$ | $-7.56 \pm -7.71$ | $0.19 \pm 0.09$ | $0.68 \pm 0.20$ | $0.47 \pm 0.25$ | $0.70 \pm 0.20$ | $\mathbf{0.80 \pm 0.15}$ | $0.74 \pm 0.15$ |
| **HOPV** | **MAE** ↓ | $0.36 \pm 0.03$ | $0.37 \pm 0.02$ | $0.40 \pm 0.02$ | $0.42 \pm 0.02$ | $0.38 \pm 0.02$ | $0.34 \pm 0.03$ | $0.31 \pm 0.03$ | $0.30 \pm 0.02$ | $0.30 \pm 0.05$ | $\mathbf{0.22 \pm 0.15}$ |
| | $\mathbf{R^2}$ ↑ | $0.69 \pm 0.04$ | $0.66 \pm 0.06$ | $0.65 \pm 0.05$ | $0.65 \pm 0.04$ | $0.66 \pm 0.03$ | $0.68 \pm 0.03$ | $0.70 \pm 0.02$ | $0.74 \pm 0.03$ | $\mathbf{0.80 \pm 0.06}$ | $0.77 \pm 0.12$ |
| **PTC** | **Acc** ↑ | $0.67 \pm 0.06$ | $0.64 \pm 0.08$ | $0.66 \pm 0.07$ | $0.61 \pm 0.08$ | $0.62 \pm 0.09$ | $0.60 \pm 0.03$ | $0.57 \pm 0.05$ | $0.69 \pm 0.07$ | $\mathbf{0.70 \pm 0.01}$ | $0.67 \pm 0.02$ |
| | **AUC** ↑ | $0.70 \pm 0.05$ | $0.68 \pm 0.06$ | $0.69 \pm 0.06$ | $0.65 \pm 0.07$ | $0.66 \pm 0.07$ | $0.66 \pm 0.05$ | $0.67 \pm 0.06$ | $\mathbf{0.71 \pm 0.07}$ | $0.69 \pm 0.03$ | $0.66 \pm 0.05$ |

- Our method *dominates* GNN baselines (including motif-based ones)
- Our method *outperforms* fine-tuning SOTA pretrained methods

# Results (cont.)

## Data-Efficient Property Prediction

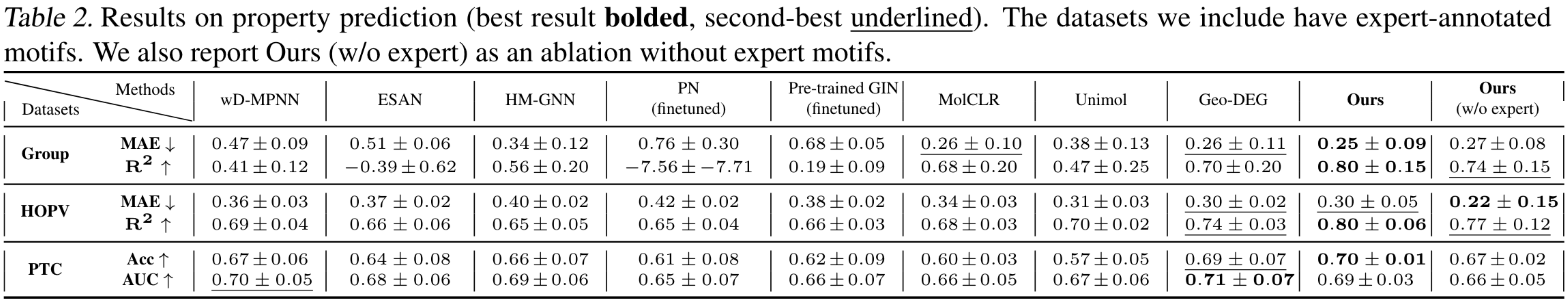


*Table 2.* Results on property prediction (best result **bolded**, second-best underlined). The datasets we include have expert-annotated motifs. We also report Ours (w/o expert) as an ablation without expert motifs.

| Methods<br>Datasets | | wD-MPNN | ESAN | HM-GNN | PN<br>(finetuned) | Pre-trained GIN<br>(finetuned) | MolCLR | Unimol | Geo-DEG | **Ours** | **Ours**<br>(w/o expert) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Group** | **MAE** ↓ | $0.47 \pm 0.09$ | $0.51 \pm 0.06$ | $0.34 \pm 0.12$ | $0.76 \pm 0.30$ | $0.68 \pm 0.05$ | $\underline{0.26 \pm 0.10}$ | $0.38 \pm 0.13$ | $\underline{0.26 \pm 0.11}$ | $\mathbf{0.25 \pm 0.09}$ | $0.27 \pm 0.08$ |
| | $\mathbf{R^2}$ ↑ | $0.41 \pm 0.12$ | $-0.39 \pm 0.62$ | $0.56 \pm 0.20$ | $-7.56 \pm -7.71$ | $0.19 \pm 0.09$ | $0.68 \pm 0.20$ | $0.47 \pm 0.25$ | $0.70 \pm 0.20$ | $\mathbf{0.80 \pm 0.15}$ | $\underline{0.74 \pm 0.15}$ |
| **HOPV** | **MAE** ↓ | $0.36 \pm 0.03$ | $0.37 \pm 0.02$ | $0.40 \pm 0.02$ | $0.42 \pm 0.02$ | $0.38 \pm 0.02$ | $0.34 \pm 0.03$ | $0.31 \pm 0.03$ | $\underline{0.30 \pm 0.02}$ | $\underline{0.30 \pm 0.05}$ | $\mathbf{0.22 \pm 0.15}$ |
| | $\mathbf{R^2}$ ↑ | $0.69 \pm 0.04$ | $0.66 \pm 0.06$ | $0.65 \pm 0.05$ | $0.65 \pm 0.04$ | $0.66 \pm 0.03$ | $0.68 \pm 0.03$ | $0.70 \pm 0.02$ | $\underline{0.74 \pm 0.03}$ | $\mathbf{0.80 \pm 0.06}$ | $\underline{0.77 \pm 0.12}$ |
| **PTC** | **Acc** ↑ | $0.67 \pm 0.06$ | $0.64 \pm 0.08$ | $0.66 \pm 0.07$ | $0.61 \pm 0.08$ | $0.62 \pm 0.09$ | $0.60 \pm 0.03$ | $0.57 \pm 0.05$ | $\underline{0.69 \pm 0.07}$ | $\mathbf{0.70 \pm 0.01}$ | $0.67 \pm 0.02$ |
| | **AUC** ↑ | $\underline{0.70 \pm 0.05}$ | $0.68 \pm 0.06$ | $0.69 \pm 0.06$ | $0.65 \pm 0.07$ | $0.66 \pm 0.07$ | $0.66 \pm 0.05$ | $0.67 \pm 0.06$ | $\mathbf{0.71 \pm 0.07}$ | $0.69 \pm 0.03$ | $0.66 \pm 0.05$ |

- Our method *dominates* GNN baselines (including motif-based ones)
- Our method *outperforms* fine-tuning SOTA pretrained methods
- Our method *is competitive with* SOTA data-efficient property predictor

# Results (cont.)

## Data-Efficient Property Prediction

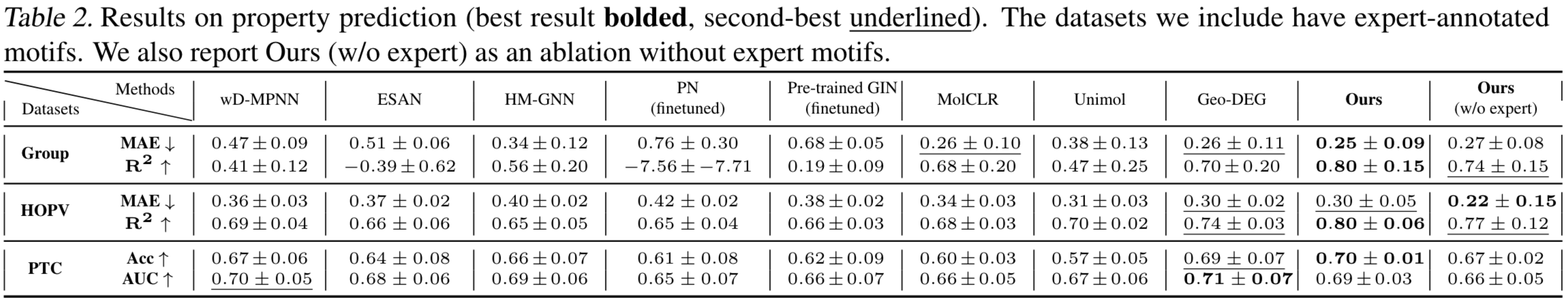


*Table 2.* Results on property prediction (best result **bolded**, second-best $\underline{\text{underlined}}$). The datasets we include have expert-annotated motifs. We also report Ours (w/o expert) as an ablation without expert motifs.

| Methods<br>Datasets | | wD-MPNN | ESAN | HM-GNN | PN (finetuned) | Pre-trained GIN (finetuned) | MolCLR | Unimol | Geo-DEG | **Ours** | **Ours** (w/o expert) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Group** | **MAE** ↓ | $0.47 \pm 0.09$ | $0.51 \pm 0.06$ | $0.34 \pm 0.12$ | $0.76 \pm 0.30$ | $0.68 \pm 0.05$ | $\underline{0.26 \pm 0.10}$ | $0.38 \pm 0.13$ | $\underline{0.26 \pm 0.11}$ | $\mathbf{0.25 \pm 0.09}$ | $0.27 \pm 0.08$ |
| | $\mathbf{R^2}$ ↑ | $0.41 \pm 0.12$ | $-0.39 \pm 0.62$ | $0.56 \pm 0.20$ | $-7.56 \pm -7.71$ | $0.19 \pm 0.09$ | $0.68 \pm 0.20$ | $0.47 \pm 0.25$ | $0.70 \pm 0.20$ | $\mathbf{0.80 \pm 0.15}$ | $\underline{0.74 \pm 0.15}$ |
| **HOPV** | **MAE** ↓ | $0.36 \pm 0.03$ | $0.37 \pm 0.02$ | $0.40 \pm 0.02$ | $0.42 \pm 0.02$ | $0.38 \pm 0.02$ | $0.34 \pm 0.03$ | $0.31 \pm 0.03$ | $\underline{0.30 \pm 0.02}$ | $\underline{0.30 \pm 0.05}$ | $\mathbf{0.22 \pm 0.15}$ |
| | $\mathbf{R^2}$ ↑ | $0.69 \pm 0.04$ | $0.66 \pm 0.06$ | $0.65 \pm 0.05$ | $0.65 \pm 0.04$ | $0.66 \pm 0.03$ | $0.68 \pm 0.03$ | $0.70 \pm 0.02$ | $\underline{0.74 \pm 0.03}$ | $\mathbf{0.80 \pm 0.06}$ | $\underline{0.77 \pm 0.12}$ |
| **PTC** | **Acc** ↑ | $0.67 \pm 0.06$ | $0.64 \pm 0.08$ | $0.66 \pm 0.07$ | $0.61 \pm 0.08$ | $0.62 \pm 0.09$ | $0.60 \pm 0.03$ | $0.57 \pm 0.05$ | $\underline{0.69 \pm 0.07}$ | $\mathbf{0.70 \pm 0.01}$ | $0.67 \pm 0.02$ |
| | **AUC** ↑ | $\underline{0.70 \pm 0.05}$ | $0.68 \pm 0.06$ | $0.69 \pm 0.06$ | $0.65 \pm 0.07$ | $0.66 \pm 0.07$ | $0.66 \pm 0.05$ | $0.67 \pm 0.06$ | $\mathbf{0.71 \pm 0.07}$ | $0.69 \pm 0.03$ | $0.66 \pm 0.05$ |

- Our method *dominates* GNN baselines (including motif-based ones)
- Our method *outperforms* fine-tuning SOTA pretrained methods
- Our method *is competitive with* SOTA data-efficient property predictor
- Expert motifs enhance performance, but heuristic-based motifs *remain competitive* with other methods

# Results (cont.)

## Data-Efficient Property Prediction

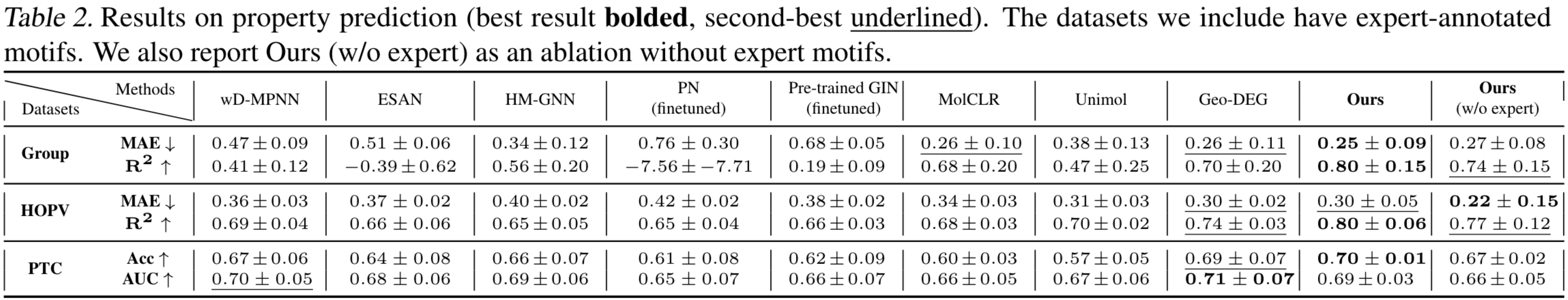


*Table 2.* Results on property prediction (best result **bolded**, second-best underlined). The datasets we include have expert-annotated motifs. We also report Ours (w/o expert) as an ablation without expert motifs.

| Datasets | Methods | wD-MPNN | ESAN | HM-GNN | PN (finetuned) | Pre-trained GIN (finetuned) | MolCLR | Unimol | Geo-DEG | **Ours** | **Ours** (w/o expert) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Group** | **MAE** ↓ | 0.47 ± 0.09 | 0.51 ± 0.06 | 0.34 ± 0.12 | 0.76 ± 0.30 | 0.68 ± 0.05 | 0.26 ± 0.10 | 0.38 ± 0.13 | 0.26 ± 0.11 | **0.25 ± 0.09** | 0.27 ± 0.08 |
|  | $\mathbf{R^2}$ ↑ | 0.41 ± 0.12 | −0.39 ± 0.62 | 0.56 ± 0.20 | −7.56 ± −7.71 | 0.19 ± 0.09 | 0.68 ± 0.20 | 0.47 ± 0.25 | 0.70 ± 0.20 | **0.80 ± 0.15** | 0.74 ± 0.15 |
| **HOPV** | **MAE** ↓ | 0.36 ± 0.03 | 0.37 ± 0.02 | 0.40 ± 0.02 | 0.42 ± 0.02 | 0.38 ± 0.02 | 0.34 ± 0.03 | 0.31 ± 0.03 | 0.30 ± 0.02 | 0.30 ± 0.05 | **0.22 ± 0.15** |
|  | $\mathbf{R^2}$ ↑ | 0.69 ± 0.04 | 0.66 ± 0.06 | 0.65 ± 0.05 | 0.65 ± 0.04 | 0.66 ± 0.03 | 0.68 ± 0.03 | 0.70 ± 0.02 | 0.74 ± 0.03 | **0.80 ± 0.06** | 0.77 ± 0.12 |
| **PTC** | **Acc** ↑ | 0.67 ± 0.06 | 0.64 ± 0.08 | 0.66 ± 0.07 | 0.61 ± 0.08 | 0.62 ± 0.09 | 0.60 ± 0.03 | 0.57 ± 0.05 | 0.69 ± 0.07 | **0.70 ± 0.01** | 0.67 ± 0.02 |
|  | **AUC** ↑ | 0.70 ± 0.05 | 0.68 ± 0.06 | 0.69 ± 0.06 | 0.65 ± 0.07 | 0.66 ± 0.07 | 0.66 ± 0.05 | 0.67 ± 0.06 | **0.71 ± 0.07** | 0.69 ± 0.03 | 0.66 ± 0.05 |

- Our method *dominates* GNN baselines (including motif-based ones)
- Our method *outperforms* fine-tuning SOTA pretrained methods
- Our method *is competitive with* SOTA data-efficient property predictor
- Expert motifs enhance performance, but heuristic-based motifs *remain competitive* with other methods
- Additional ablations showing *better runtime* and *data-efficiency* than Geo-DEG in paper

# Results (cont.)

## Comparison with Motif-based Property Predictors

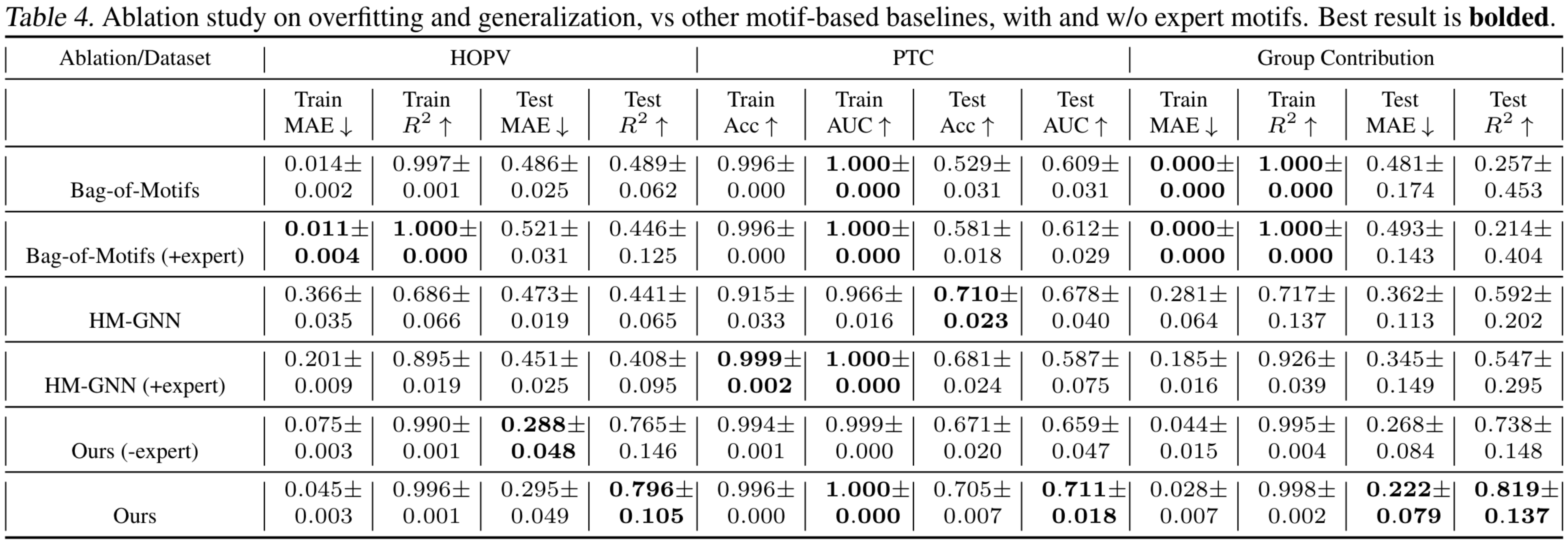


*Table 4.* Ablation study on overfitting and generalization, vs other motif-based baselines, with and w/o expert motifs. Best result is **bolded**.

| Ablation/Dataset | HOPV | | | | PTC | | | | Group Contribution | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train MAE ↓ | Train $R^2$ ↑ | Test MAE ↓ | Test $R^2$ ↑ | Train Acc ↑ | Train AUC ↑ | Test Acc ↑ | Test AUC ↑ | Train MAE ↓ | Train $R^2$ ↑ | Test MAE ↓ | Test $R^2$ ↑ |
| Bag-of-Motifs | 0.014± 0.002 | 0.997± 0.001 | 0.486± 0.025 | 0.489± 0.062 | 0.996± 0.000 | **1.000± 0.000** | 0.529± 0.031 | 0.609± 0.031 | **0.000± 0.000** | **1.000± 0.000** | 0.481± 0.174 | 0.257± 0.453 |
| Bag-of-Motifs (+expert) | **0.011± 0.004** | **1.000± 0.000** | 0.521± 0.031 | 0.446± 0.125 | 0.996± 0.000 | **1.000± 0.000** | 0.581± 0.018 | 0.612± 0.029 | **0.000± 0.000** | **1.000± 0.000** | 0.493± 0.143 | 0.214± 0.404 |
| HM-GNN | 0.366± 0.035 | 0.686± 0.066 | 0.473± 0.019 | 0.441± 0.065 | 0.915± 0.033 | 0.966± 0.016 | **0.710± 0.023** | 0.678± 0.040 | 0.281± 0.064 | 0.717± 0.137 | 0.362± 0.113 | 0.592± 0.202 |
| HM-GNN (+expert) | 0.201± 0.009 | 0.895± 0.019 | 0.451± 0.025 | 0.408± 0.095 | **0.999± 0.002** | **1.000± 0.000** | 0.681± 0.024 | 0.587± 0.075 | 0.185± 0.016 | 0.926± 0.039 | 0.345± 0.149 | 0.547± 0.295 |
| Ours (-expert) | 0.075± 0.003 | 0.990± 0.001 | **0.288± 0.048** | 0.765± 0.146 | 0.994± 0.001 | 0.999± 0.000 | 0.671± 0.020 | 0.659± 0.047 | 0.044± 0.015 | 0.995± 0.004 | 0.268± 0.084 | 0.738± 0.148 |
| Ours | 0.045± 0.003 | 0.996± 0.001 | 0.295± 0.049 | **0.796± 0.105** | 0.996± 0.000 | **1.000± 0.000** | 0.705± 0.007 | **0.711± 0.018** | 0.028± 0.007 | 0.998± 0.002 | **0.222± 0.079** | **0.819± 0.137** |

- Motif occurrence features (Bag-of-Motifs) overfits but does not generalize

# Results (cont.)

## Comparison with Motif-based Property Predictors

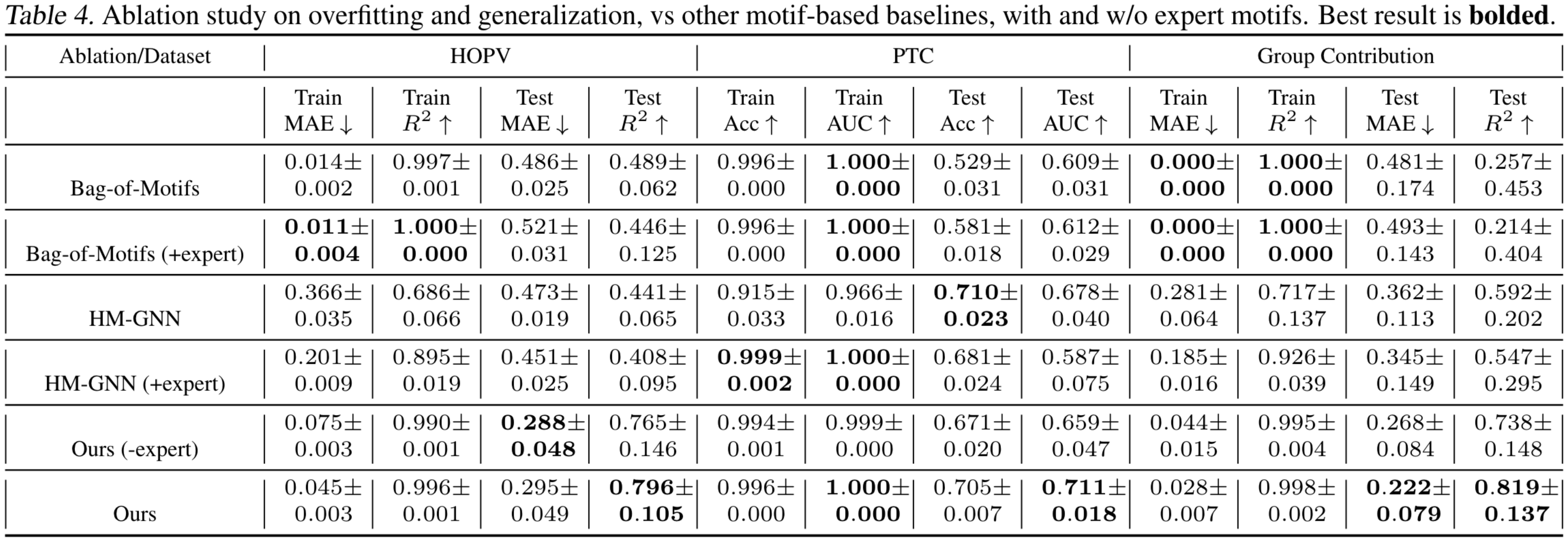


*Table 4.* Ablation study on overfitting and generalization, vs other motif-based baselines, with and w/o expert motifs. Best result is **bolded**.

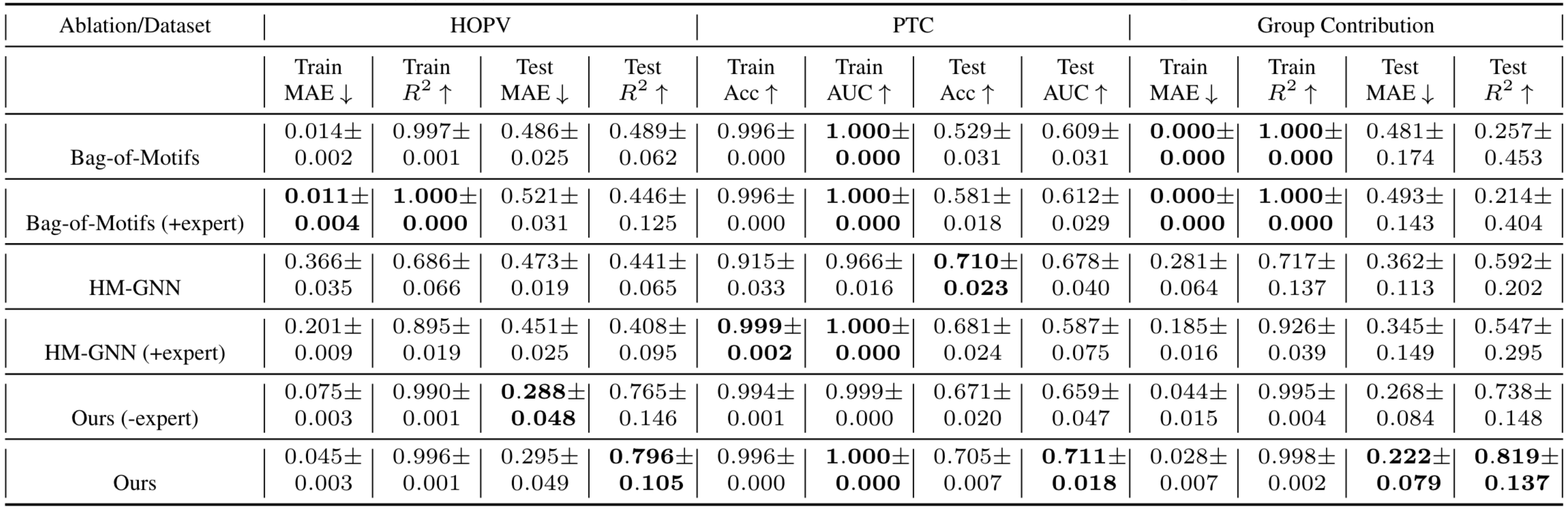
| Ablation/Dataset | HOPV | | | | PTC | | | | Group Contribution | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train MAE ↓ | Train $R^2$ ↑ | Test MAE ↓ | Test $R^2$ ↑ | Train Acc ↑ | Train AUC ↑ | Test Acc ↑ | Test AUC ↑ | Train MAE ↓ | Train $R^2$ ↑ | Test MAE ↓ | Test $R^2$ ↑ |
| Bag-of-Motifs | 0.014±<br>0.002 | 0.997±<br>0.001 | 0.486±<br>0.025 | 0.489±<br>0.062 | 0.996±<br>0.000 | **1.000**±<br>**0.000** | 0.529±<br>0.031 | 0.609±<br>0.031 | **0.000**±<br>**0.000** | **1.000**±<br>**0.000** | 0.481±<br>0.174 | 0.257±<br>0.453 |
| Bag-of-Motifs (+expert) | **0.011**±<br>**0.004** | **1.000**±<br>**0.000** | 0.521±<br>0.031 | 0.446±<br>0.125 | 0.996±<br>0.000 | **1.000**±<br>**0.000** | 0.581±<br>0.018 | 0.612±<br>0.029 | **0.000**±<br>**0.000** | **1.000**±<br>**0.000** | 0.493±<br>0.143 | 0.214±<br>0.404 |
| HM-GNN | 0.366±<br>0.035 | 0.686±<br>0.066 | 0.473±<br>0.019 | 0.441±<br>0.065 | 0.915±<br>0.033 | 0.966±<br>0.016 | **0.710**±<br>**0.023** | 0.678±<br>0.040 | 0.281±<br>0.064 | 0.717±<br>0.137 | 0.362±<br>0.113 | 0.592±<br>0.202 |
| HM-GNN (+expert) | 0.201±<br>0.009 | 0.895±<br>0.019 | 0.451±<br>0.025 | 0.408±<br>0.095 | **0.999**±<br>**0.002** | **1.000**±<br>**0.000** | 0.681±<br>0.024 | 0.587±<br>0.075 | 0.185±<br>0.016 | 0.926±<br>0.039 | 0.345±<br>0.149 | 0.547±<br>0.295 |
| Ours (-expert) | 0.075±<br>0.003 | 0.990±<br>0.001 | **0.288**±<br>**0.048** | 0.765±<br>0.146 | 0.994±<br>0.001 | 0.999±<br>0.000 | 0.671±<br>0.020 | 0.659±<br>0.047 | 0.044±<br>0.015 | 0.995±<br>0.004 | 0.268±<br>0.084 | 0.738±<br>0.148 |
| Ours | 0.045±<br>0.003 | 0.996±<br>0.001 | 0.295±<br>0.049 | **0.796**±<br>**0.105** | 0.996±<br>0.000 | **1.000**±<br>**0.000** | 0.705±<br>0.007 | **0.711**±<br>**0.018** | 0.028±<br>0.007 | 0.998±<br>0.002 | **0.222**±<br>**0.079** | **0.819**±<br>**0.137** |

- Motif occurrence features (Bag-of-Motifs) overfits but does not generalize
- SOTA motif-based property predictor (HM-GNN) avoids overfitting but does not generalize well

# Results (cont.)

## Comparison with Motif-based Property Predictors

*Table 4.* Ablation study on overfitting and generalization, vs other motif-based baselines, with and w/o expert motifs. Best result is **bolded**.

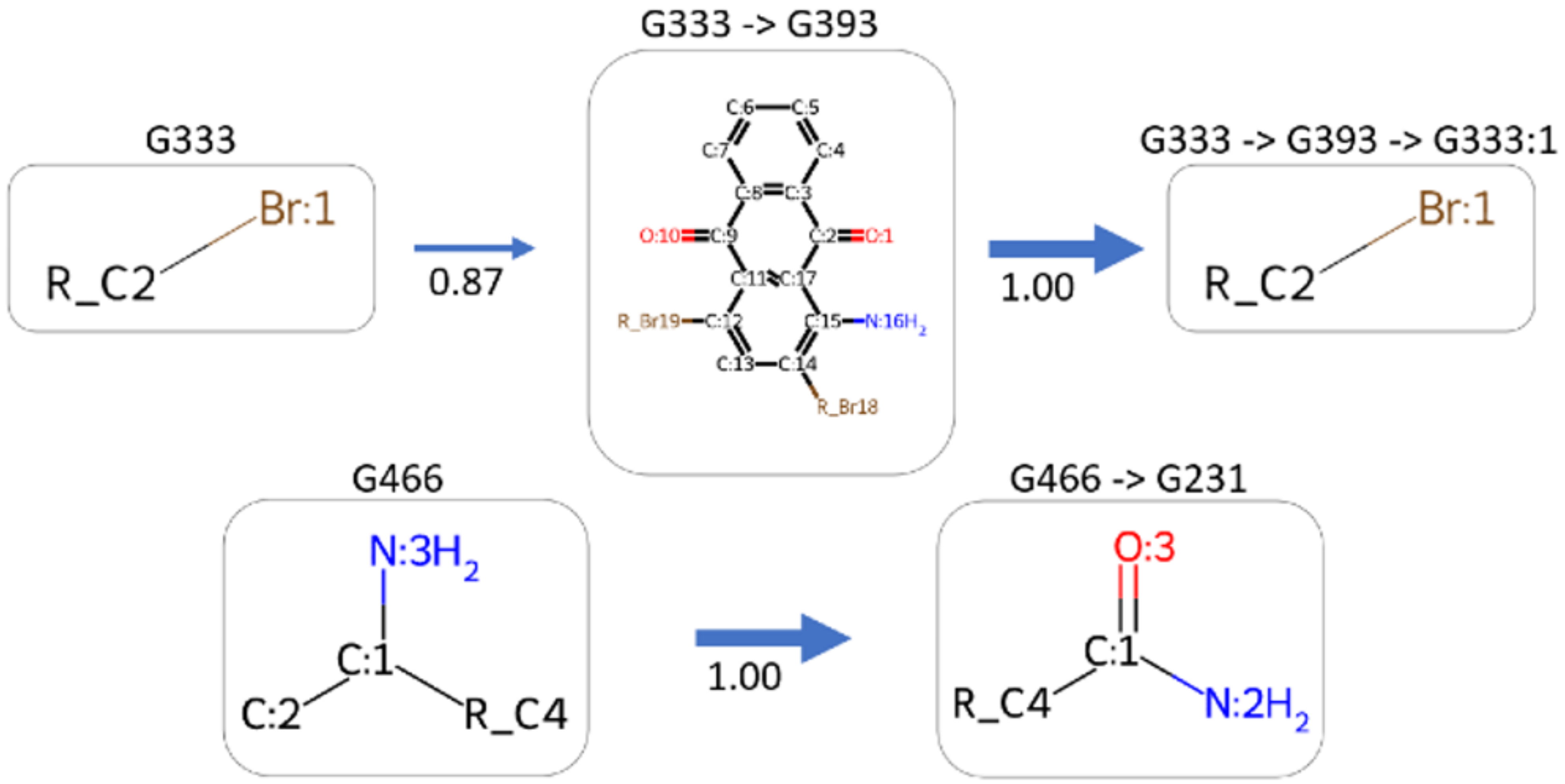


| Ablation/Dataset | HOPV | | | | PTC | | | | Group Contribution | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train MAE ↓ | Train $R^2$ ↑ | Test MAE ↓ | Test $R^2$ ↑ | Train Acc ↑ | Train AUC ↑ | Test Acc ↑ | Test AUC ↑ | Train MAE ↓ | Train $R^2$ ↑ | Test MAE ↓ | Test $R^2$ ↑ |
| Bag-of-Motifs | 0.014±0.002 | 0.997±0.001 | 0.486±0.025 | 0.489±0.062 | 0.996±0.000 | **1.000±0.000** | 0.529±0.031 | 0.609±0.031 | **0.000±0.000** | **1.000±0.000** | 0.481±0.174 | 0.257±0.453 |
| Bag-of-Motifs (+expert) | **0.011±0.004** | **1.000±0.000** | 0.521±0.031 | 0.446±0.125 | 0.996±0.000 | **1.000±0.000** | 0.581±0.018 | 0.612±0.029 | **0.000±0.000** | **1.000±0.000** | 0.493±0.143 | 0.214±0.404 |
| HM-GNN | 0.366±0.035 | 0.686±0.066 | 0.473±0.019 | 0.441±0.065 | 0.915±0.033 | 0.966±0.016 | **0.710±0.023** | 0.678±0.040 | 0.281±0.064 | 0.717±0.137 | 0.362±0.113 | 0.592±0.202 |
| HM-GNN (+expert) | 0.201±0.009 | 0.895±0.019 | 0.451±0.025 | 0.408±0.095 | **0.999±0.002** | **1.000±0.000** | 0.681±0.024 | 0.587±0.075 | 0.185±0.016 | 0.926±0.039 | 0.345±0.149 | 0.547±0.295 |
| Ours (-expert) | 0.075±0.003 | 0.990±0.001 | **0.288±0.048** | 0.765±0.146 | 0.994±0.001 | 0.999±0.000 | 0.671±0.020 | 0.659±0.047 | 0.044±0.015 | 0.995±0.004 | 0.268±0.084 | 0.738±0.148 |
| Ours | 0.045±0.003 | 0.996±0.001 | 0.295±0.049 | **0.796±0.105** | 0.996±0.000 | **1.000±0.000** | 0.705±0.007 | **0.711±0.018** | 0.028±0.007 | 0.998±0.002 | **0.222±0.079** | **0.819±0.137** |

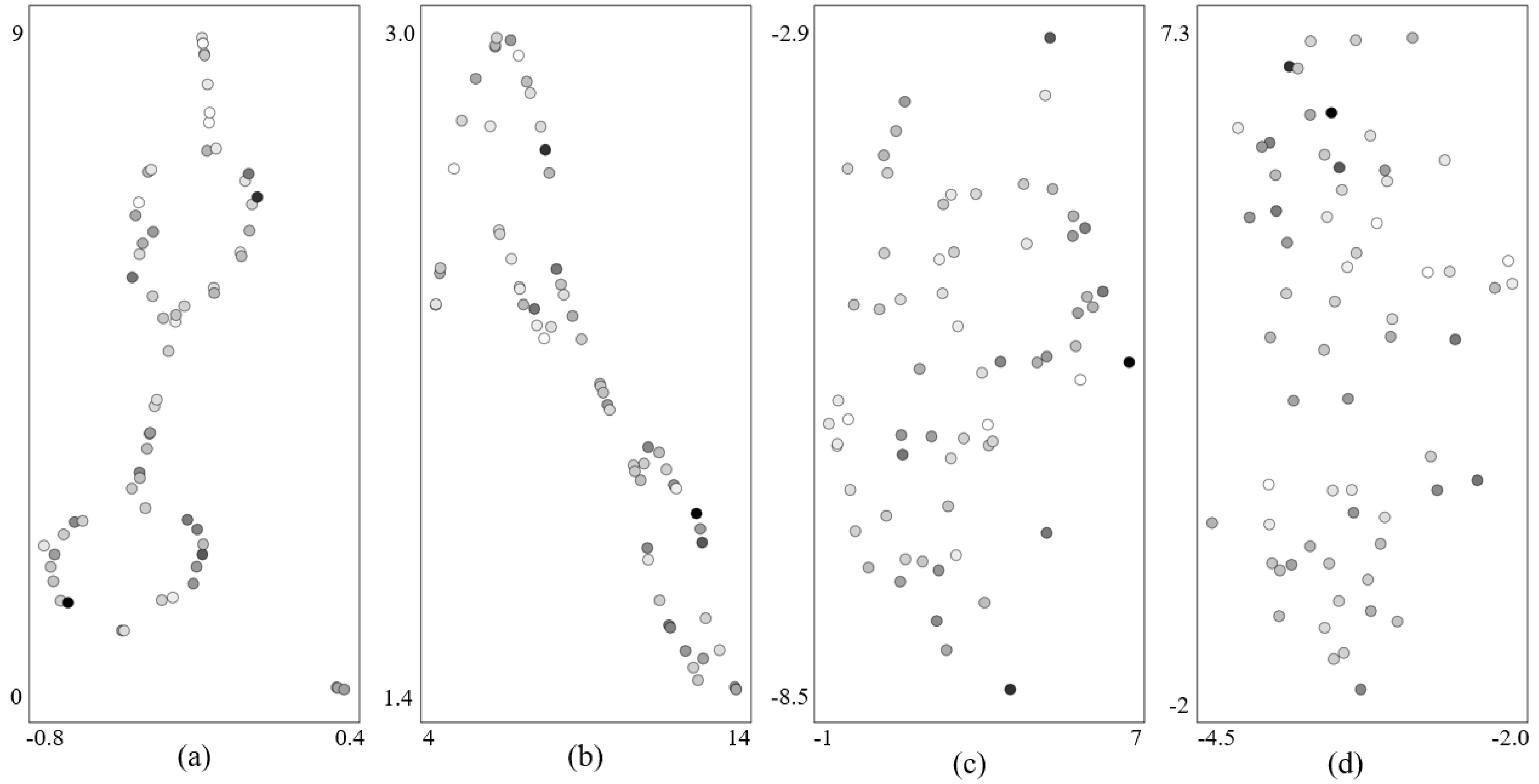
- Motif occurrence features (Bag-of-Motifs) overfits but does not generalize
- SOTA motif-based property predictor (HM-GNN) avoids overfitting but does not generalize well
- Both baselines cannot utilize expert motifs as well as Ours

# Discussion & Analysis

## Advantages in interpretability



We visualize two hard context-sensitive rules on PTC that correspond to design principles of the addition of halogen groups to further improve molecular toxicity.



Final layer representations from: a) Our method b) Our method (-expert) c) Pre-trained GIN d) HM-GNN. We apply a grayscale coloring map using the normalized value of the desired property (the darker the dot, the higher the HOMO).

Our learnt grammar enables the mining of design rules: transitions of probability 1.

Our grammar finds that this triple benzene derivative should have two symmetrical bromide moieties on the same aromatic ring to enhance its toxicity.

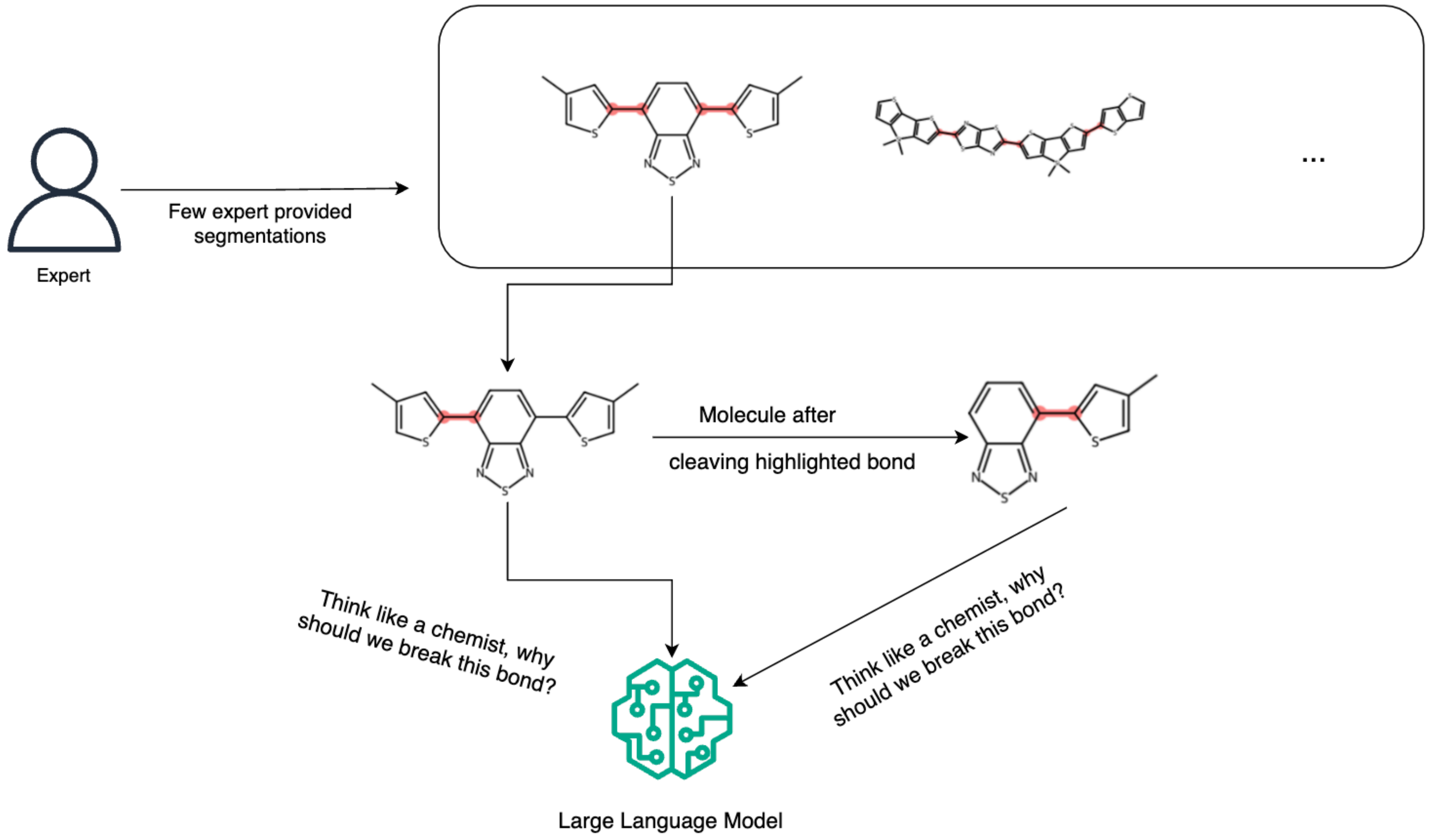
We enumerate a list of such design rules in our paper.

**Unexpected finding!** We find final layer representations of our GNN form two visually apparent clusters that correspond to the two primary ways to design molecules with high HOMO values.

# Future Work

## Integration of Large Language Models

- Teach Large Language Model to reason about expert annotations
- Teach Large Language Model to do motif extraction
- Induce graph grammars with Large Language Models
- Generalize to our method other domains beyond molecules


Expert
Few expert provided segmentations
...
Molecule after
cleaving highlighted bond
Think like a chemist, why should we break this bond?
Think like a chemist, why should we break this bond?
Large Language Model