Molecular Representation: **Initial Representation $\mathcal{M}$ (e.g. SMILES String, ...)**
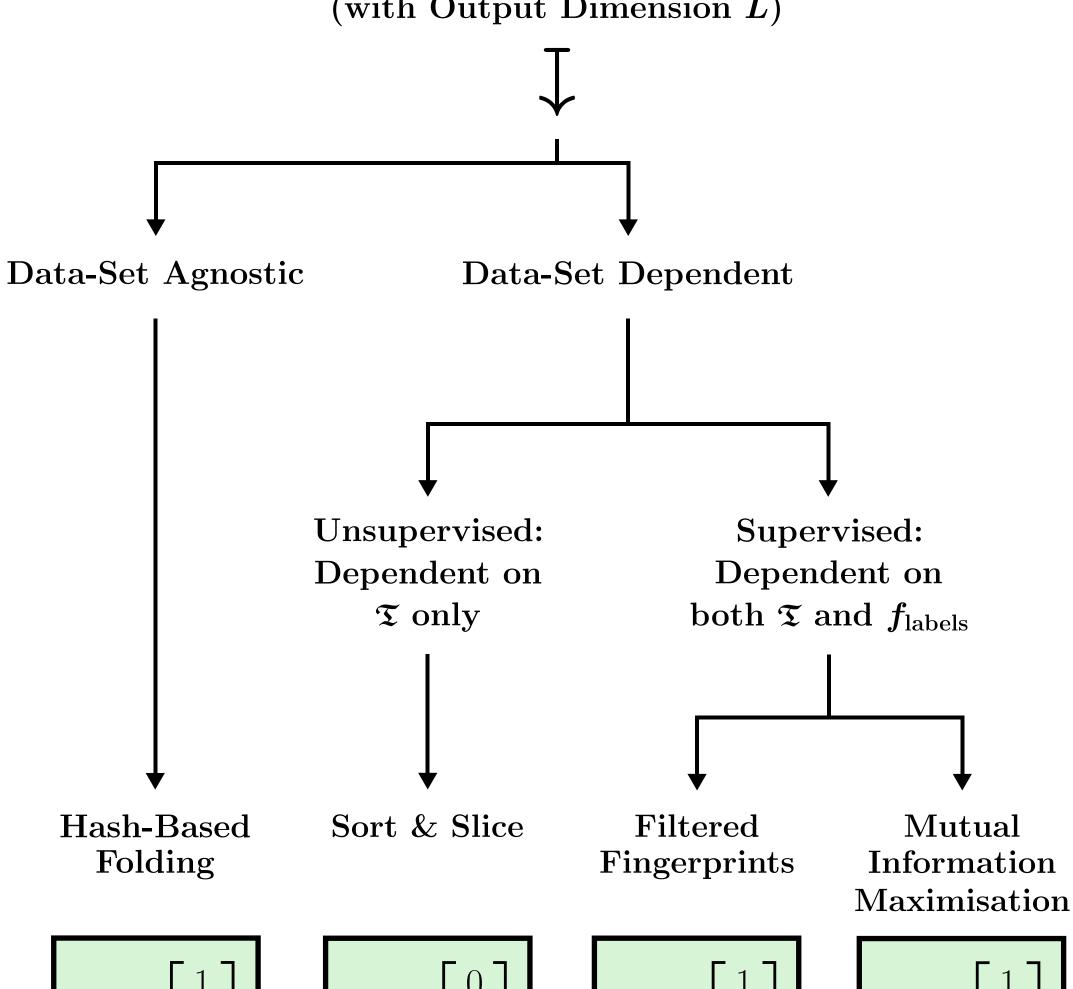
$\mathcal{M} = $ N=C(N)c1ccc2[nH]c(-c3cccc(-c4ccccc4)c3)cc2c1

ECFP-Substructure Enumeration
(with Maximum Diameter D and Atomic Invariants A)

Molecular Representation: **Set of Integer Substructure Identifiers $\mathcal{I}$**

$\mathcal{I} = \Big\{$ NH 734603939 , NH 1559650422 , NH -1100000244 , NH 131006455 , $\ldots\Big\}$

Substructure Pooling
(with Output Dimension L)

Data-Set Agnostic

Data-Set Dependent

Unsupervised: Dependent on $\mathfrak{T}$ only

Supervised: Dependent on both $\mathfrak{T}$ and f_{labels}

Four Investigated Substructure-Pooling Methods to Vectorise $\mathcal{I}$

Hash-Based Folding

Sort & Slice

Filtered Fingerprints

Mutual Information Maximisation

Molecular Representation: **Binary Vector $\mathcal{F}$**

$\mathcal{F} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 1 \\ 1 \end{bmatrix}$

$\mathcal{F} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \\ 0 \end{bmatrix}$

$\mathcal{F} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$

$\mathcal{F} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$ $\begin{matrix} i=1 \\ \vdots \\ \vdots \\ \vdots \\ i=L \end{matrix}$