Department of Applied Math and Physics, Discrete Math Research Group

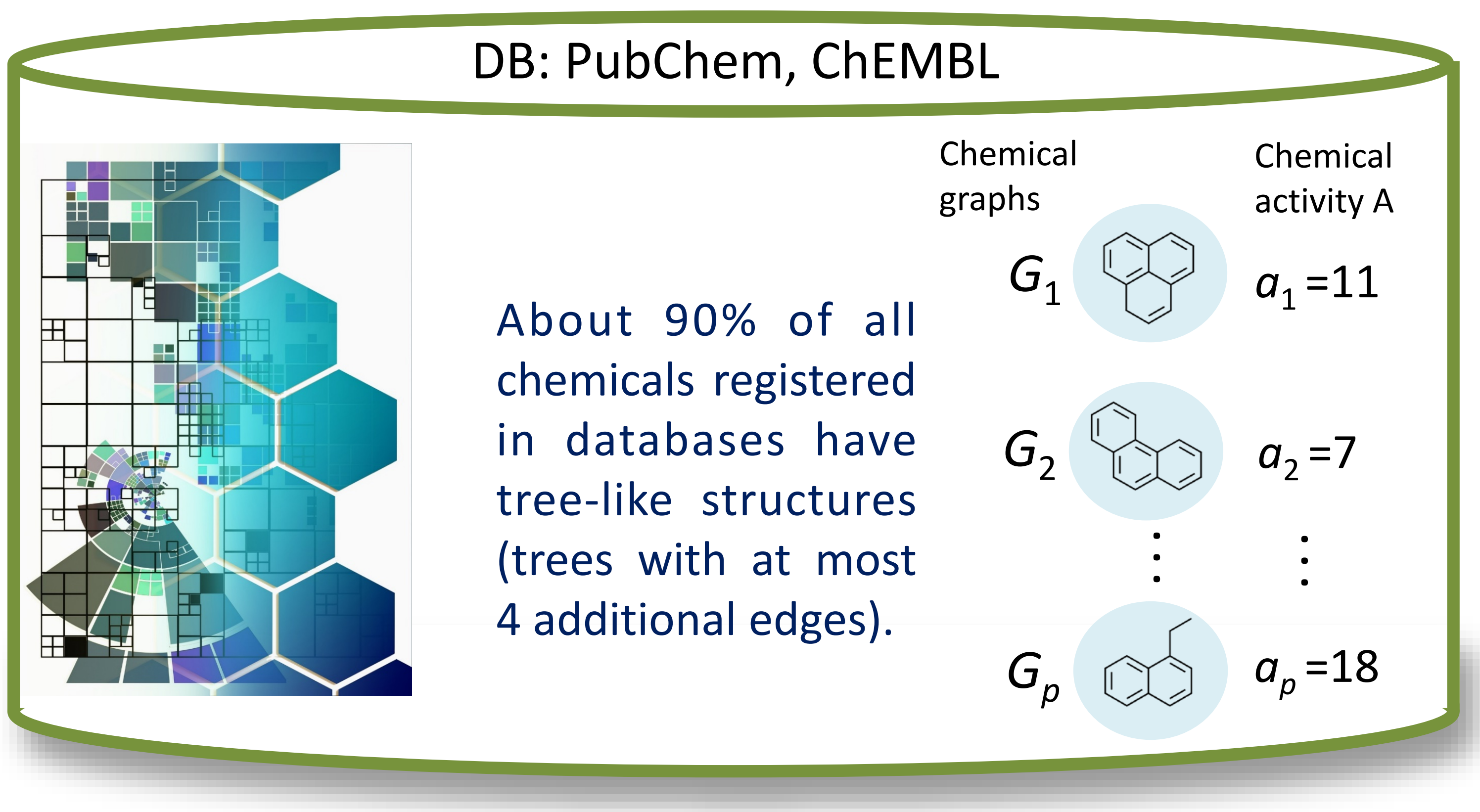
Machine Learning and Discrete Optimization for Designing Novel Chemical Compounds

Joint project with Prof Akutsu’s Lab, Kyoto University Institute for Chemical Research

We develop computational methods in cheminformatics for discovering novel substances with useful chemical activities.

1

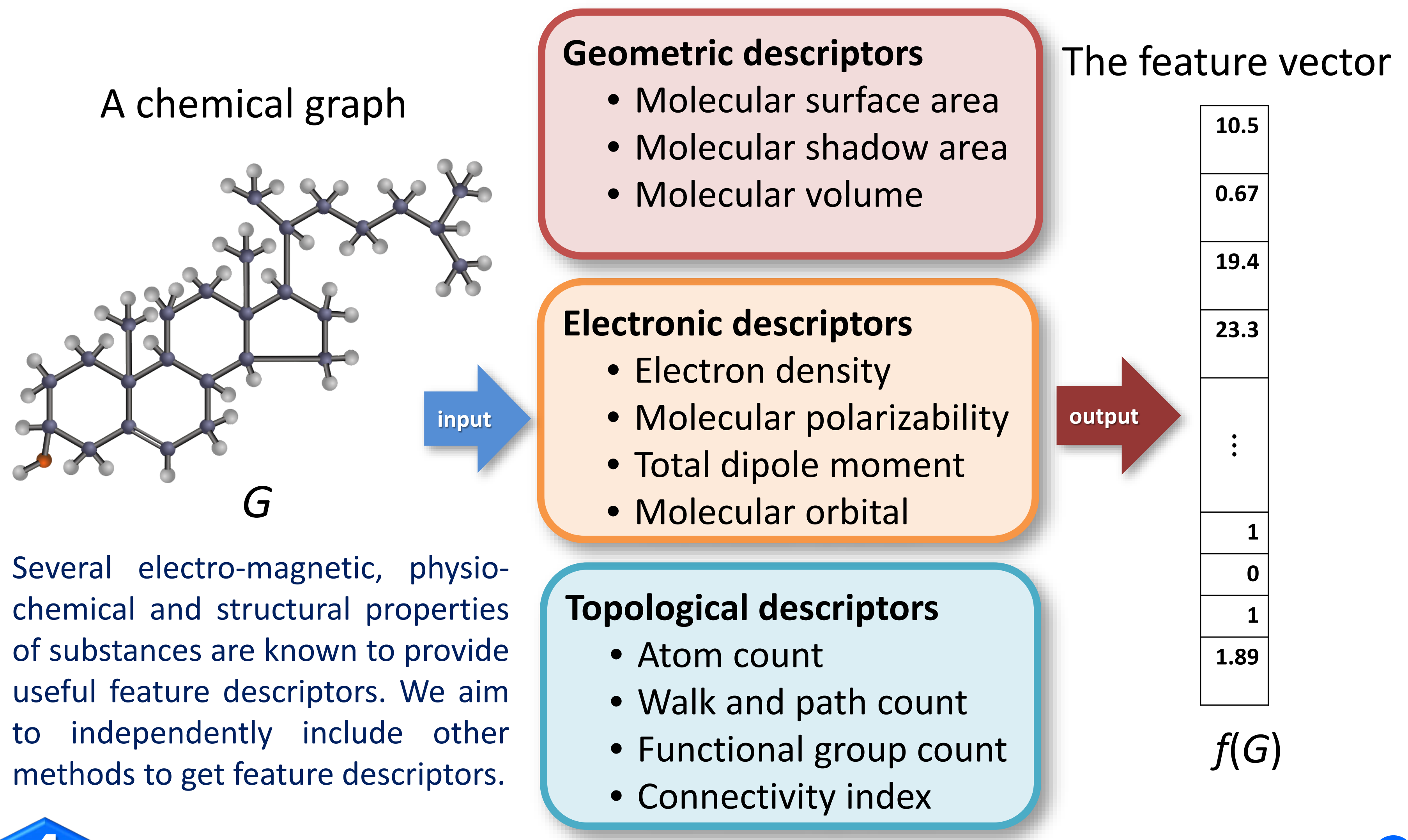
First, choose a chemical activity A, such as corrosiveness, solubility, or medicinal effects, and collect from a chemical DB compounds G_1, G_2, ..., G_p whose values for activity A are a_1, a_2, ..., a_p, respectively.



About 90% of all chemicals registered in databases have tree-like structures (trees with at most 4 additional edges).

2

Next, from the structure of each graph G_i compute several feature descriptors likely to be related to activity A and prepare a feature vector $f(G_i)$.



Geometric descriptors

- Molecular surface area
- Molecular shadow area
- Molecular volume

Electronic descriptors

- Electron density
- Molecular polarizability
- Total dipole moment
- Molecular orbital

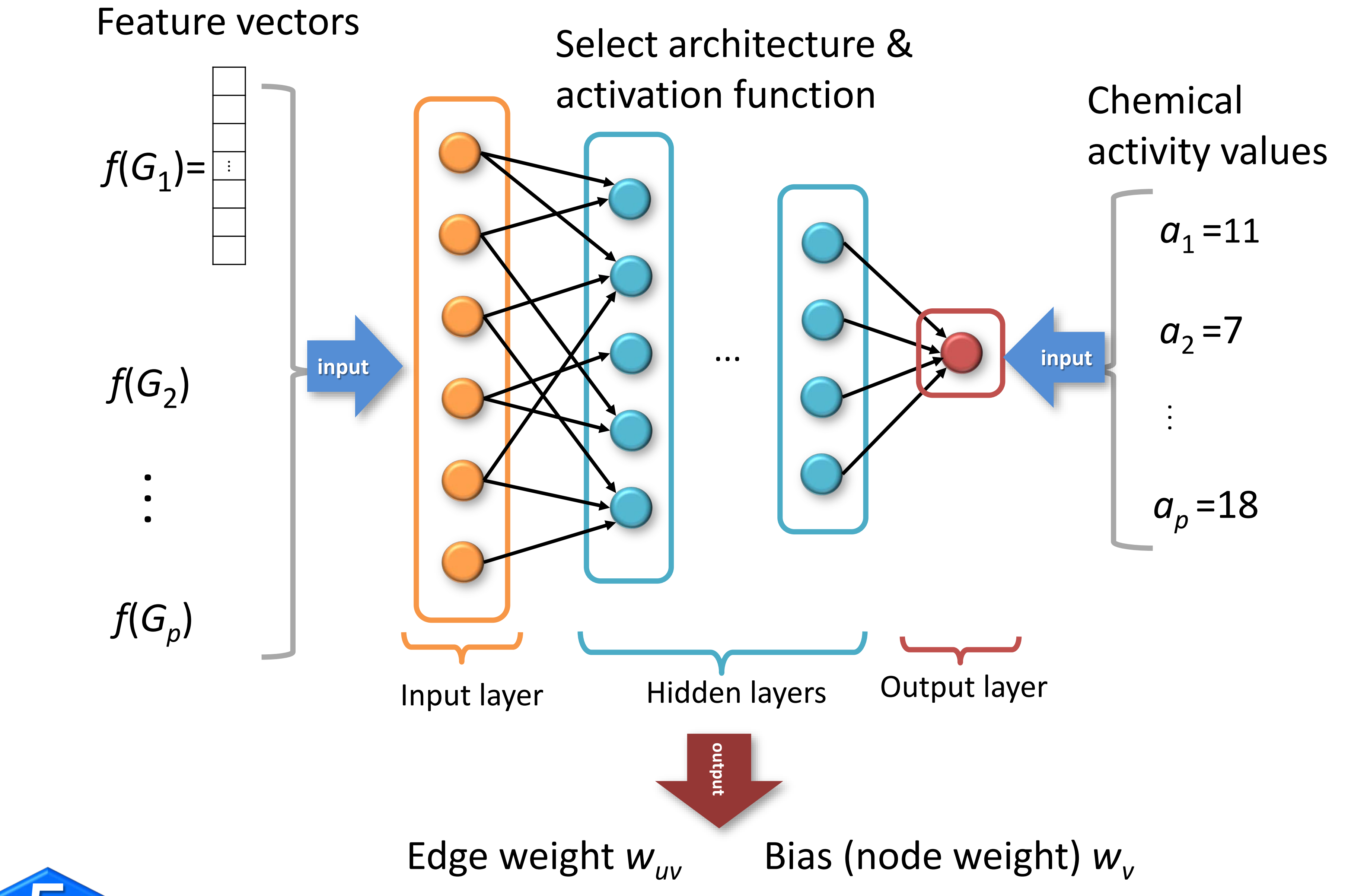
Topological descriptors

- Atom count
- Walk and path count
- Functional group count
- Connectivity index

Several electro-magnetic, physio-chemical and structural properties of substances are known to provide useful feature descriptors. We aim to independently include other methods to get feature descriptors.

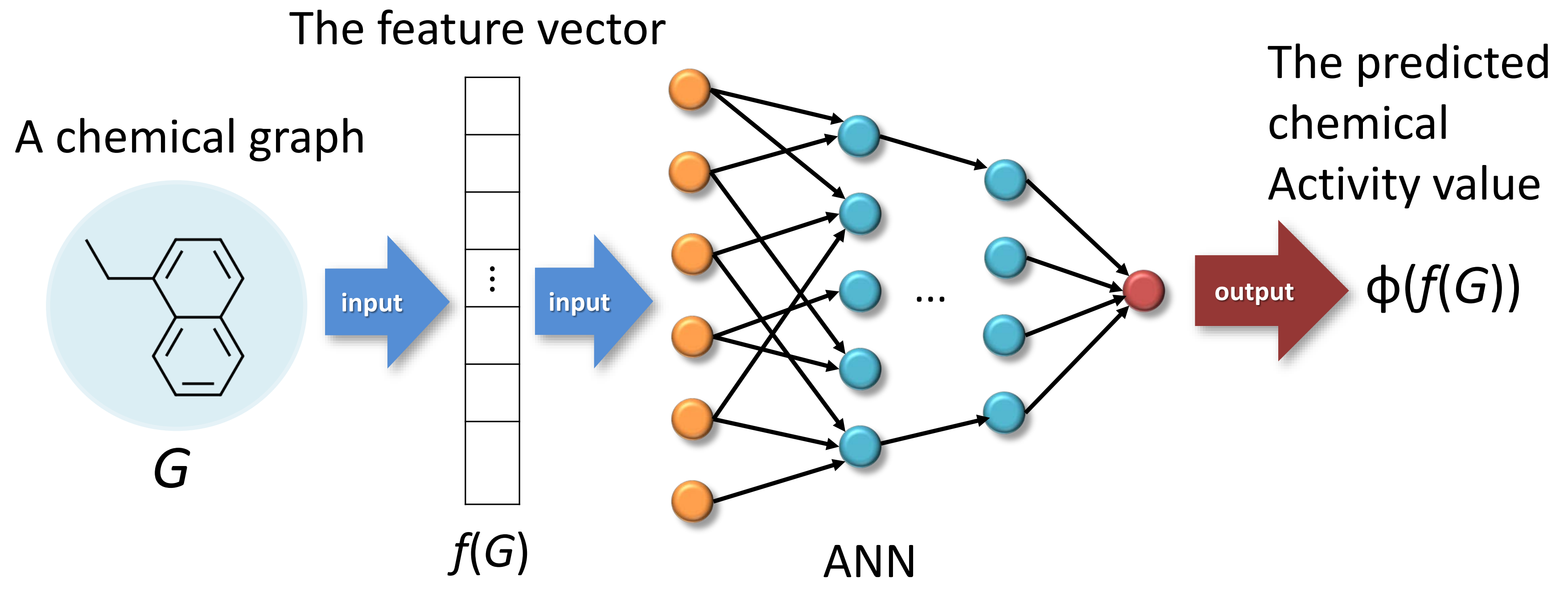
3

Using the pairs of a feature vector and activity value ($f(G_1)$, a_1), ($f(G_2)$, a_2), ..., ($f(G_p)$, a_p) as training data for an Artificial Neural Network (ANN), calculate edge weights and node biases for a given network architecture that best capture the relationship between the feature vectors and activity values.



4

Using the trained ANN we can get a prediction ϕ($f(G)$) for the value of activity A of an unknown chemical compound *G*.

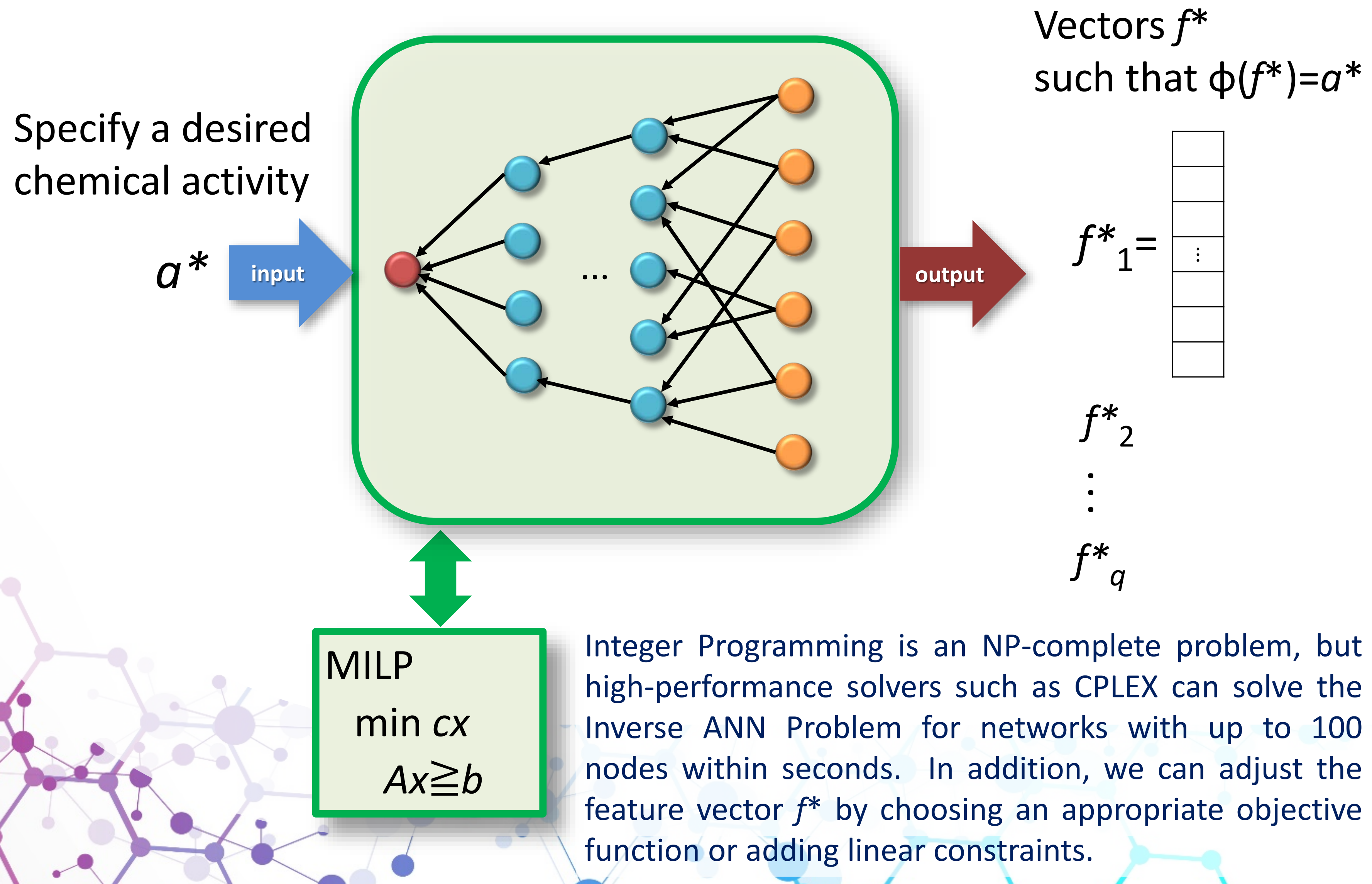


Conventionally, machine learning techniques have been used until this step. However, the aim of this project is to devise a method that given a desired value ϕ as an image of an unknown chemical compound, calculates a pre-image chemical compound that has the desired activity value.

5

In order to do this, we must solve the “inverse ANN problem,” that is, given a target value a^*, find a feature vector f^* such that $\phi(f^*) = a^*$. We have recently proposed a way of solving the inverse ANN problem as a Mixed Integer Programming Problem.

T. Akutsu and H. Nagamochi, A Mixed Integer Linear Programming Formulation to Artificial Neural Networks, Technical Report 2019-001. http://www.amp.i.kyoto-u.ac.jp/tecrep/index.html


Specify a desired chemical activity
a^*
input
…
output
Vectors f^* such that $\phi(f^*)=a^*$
f^*_1=
⋮
f^*_2
⋮
f^*_q
MILP
min cx
$Ax \geqq b$

Integer Programming is an NP-complete problem, but high-performance solvers such as CPLEX can solve the Inverse ANN Problem for networks with up to 100 nodes within seconds. In addition, we can adjust the feature vector f^* by choosing an appropriate objective function or adding linear constraints.

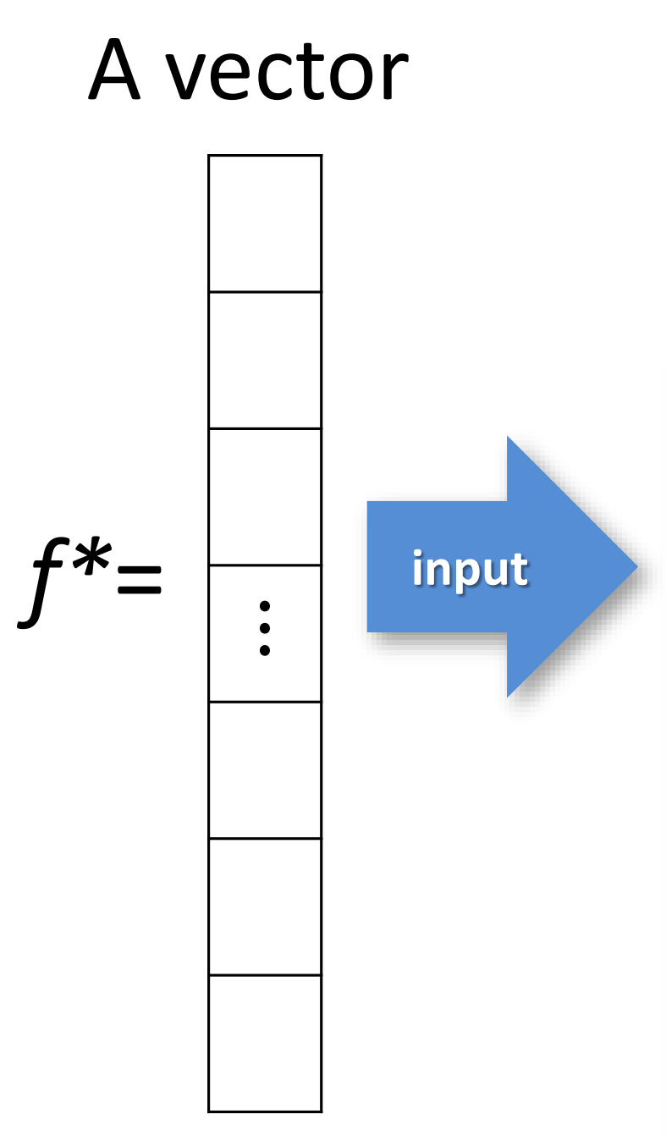
6

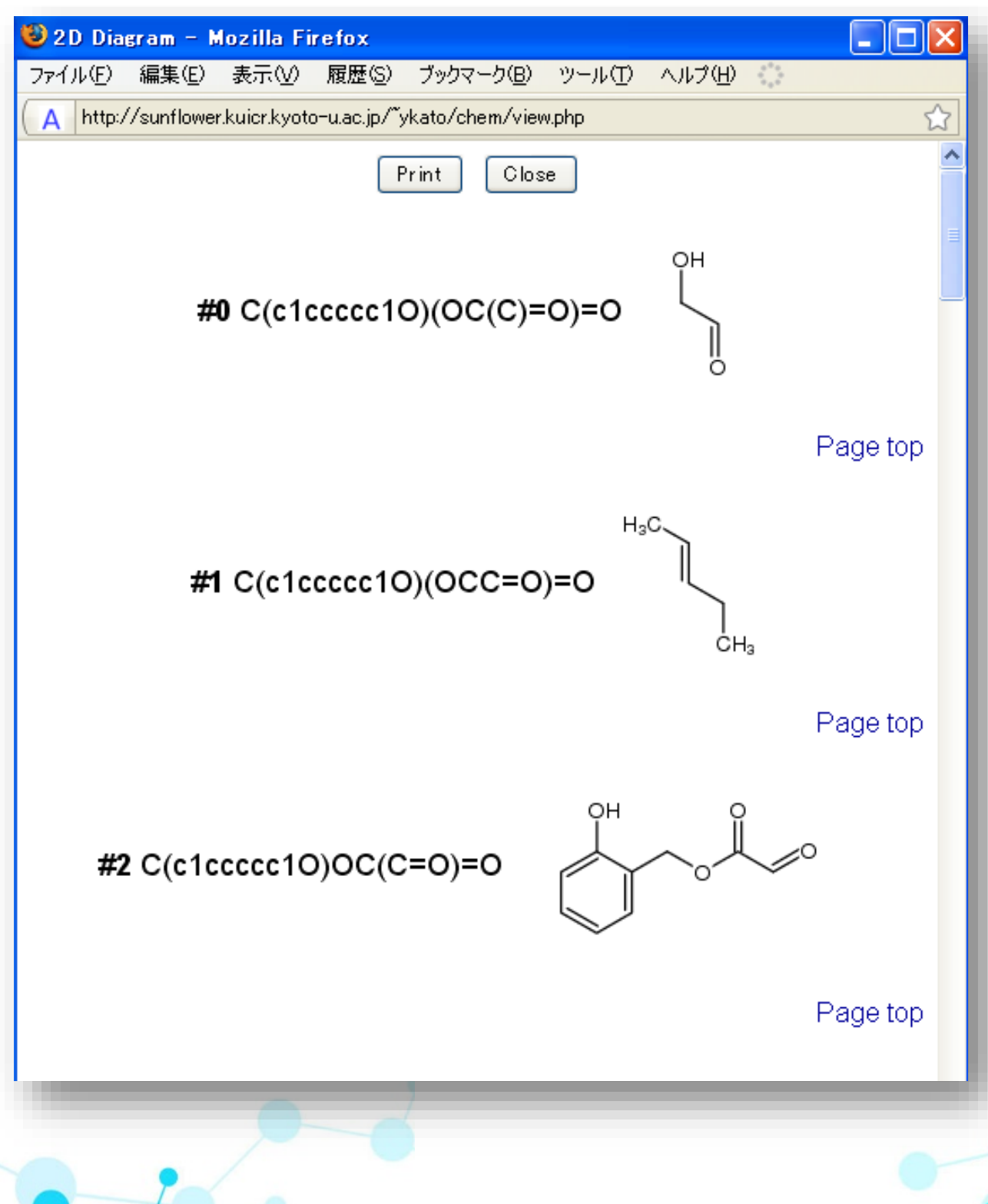
Finally, once we have computed a feature vector f^*, we enumerate all possible chemical graphs G^* such that $f(G^*) = f^*$.
For this purpose we design algorithms based on the branch-and-bound and the dynamic programming paradigms.

A Graph Enumeration Algorithm

We have made our algorithm that generates tree structures publically available as the EnuMol solver.

EnuMol: Enumeration of tree-like chemical graphs
http: //sunflower.kuicr.kyoto-u.ac.jp/tools/enumol2

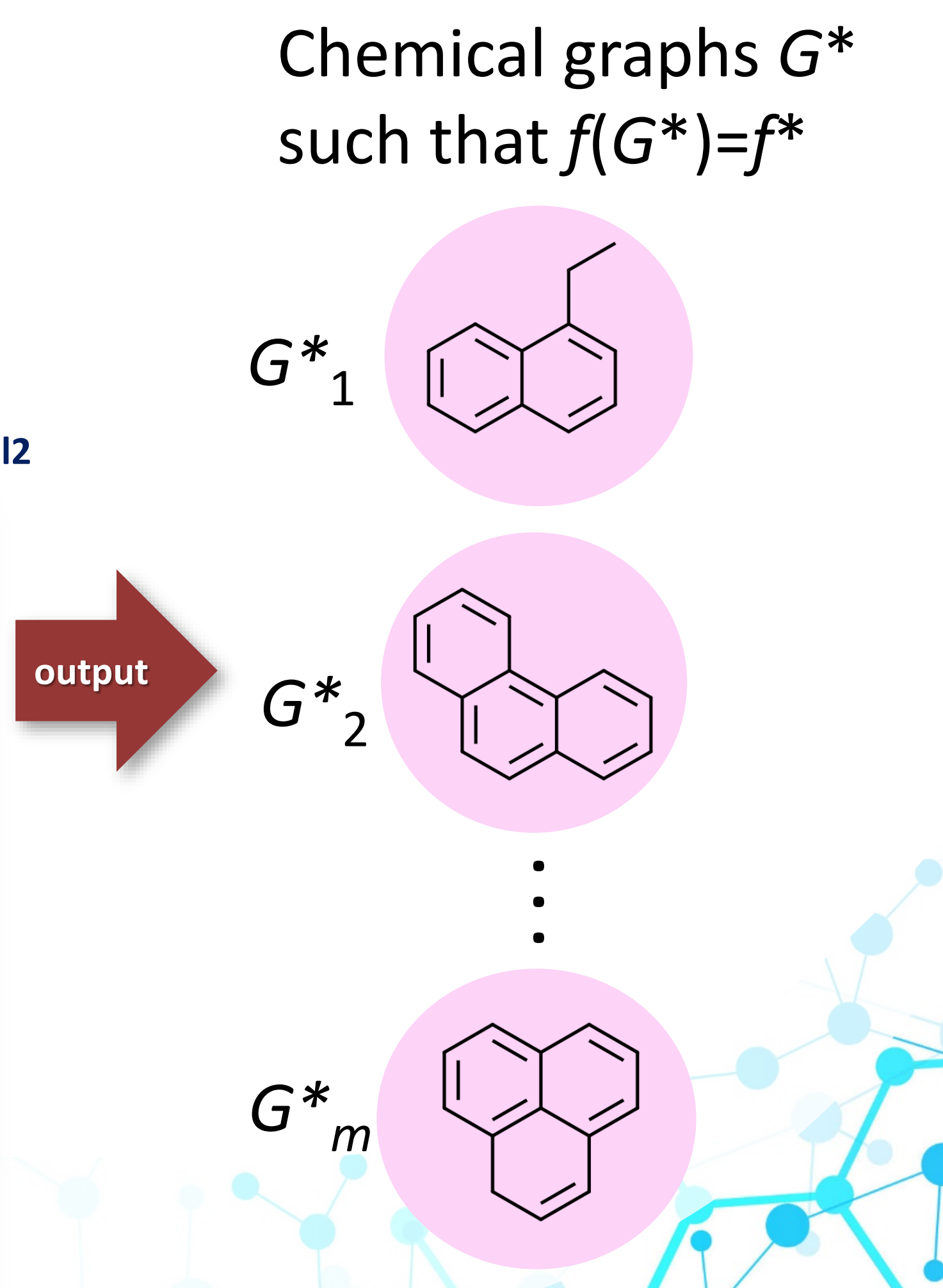




#0 C(c1ccccc1O)(OC(C)=O)=O

#1 C(c1ccccc1O)(OCC=O)=O

#2 C(c1ccccc1O)OC(C=O)=O


output
Chemical graphs G^* such that $f(G^*)=f^*$
G^*_1
G^*_2
⋮
G^*_m