# Introduction to Molecular Fingerprints (by Famke and Kirsten)

Molecular fingerprints are essentially a series of bits, often in binary form, that encode the presence or absence of specific chemical substructures or molecular properties in a molecule (so a molecule could look like 0 1 1 1 0 0 0 1 0 1 0 etc.) [1]. Molecular descriptors (so the features we are currently using) are the experimental or theoretical physicochemical properties of a compound. Molecular fingerprints encode molecules as bit or count vectors. The information that is encoded depends on the type of fingerprint [2]. Some general categories of molecular fingerprints include:

* Structural pattern fingerprints (e.g. MACCS) - sets of functional group patterns (e.g. alkyl chains, aromatic rings, ketones, ..) one bit or count per functional group in the fingerprint.
* Daylight-style linear fingerprints - take the atom, and it's linear neighbors
* Circular-style fingerprints (e.g. ECFP) - generating a radius or diameter around the neighbors of an atom. Like the linear fingerprints, these can essentially generate an infinite number of patterns, so it's hashed to a fixed bit vector
* 3D property fingerprints - interatomic distances, surfaces, electrostatic fields, etc. [3]

## ECFP (the fingerprint we propose)

### Reasoning/background info

Daylight fingerprints and extended connectivity fingerprints (ECFP) extract chemical patterns of up to a specified length or diameter from a chemical graph [4]. According to the forum [3], the Morgan circular / ECFP fingerprints are really good, especially with a fairly large bit vector (e.g. 2048 bits per molecule) to hash.

Among the different types of fingerprints, substructure fingerprints perform best for small molecules such as drugs, while atom-pair fingerprints are preferable for large molecules such as peptides [5].

They (ECFPS) were originally designed for substructure search in databases, but later gained popularity for similarity searching and molecule clustering. Extended-connectivity fingerprints (ECFPs) are a type of molecular fingerprint specifically designed for predicting and analyzing molecular activity and properties [6].

I think we might prefer ECFPs over Morgan Algorithm, it seems like an improved version [7].

### ECFP Process

The ECFP generation process has three sequential stages:

|  |  |
| --- | --- |
| 1. | An *initial assignment stage* in which each atom has an integer identifier assigned to it. |
| 2. | An *iterative updating stage* in which each atom identifier is updated to reflect the identifiers of each atom’s neighbors, including identification of whether it is a structural duplicate of other features. |
| 3. | A *duplicate identifier removal stage* in which multiple occurrences of the same feature are reduced to a single representative in the final feature list. (The occurrence count may be retained if one requires a set of counts rather than a standard binary fingerprint.) |

The above process is further described as follows. First, atoms are assigned integer identifiers (for example, atoms might use their atomic number). These initial atom identifiers are collected into an initial fingerprint set. Next, each atom collects its own identifier and the identifiers of its immediately neighboring atoms, into an array (the neighbors are ordered using their identifiers, and the order of the attaching bonds, to avoid order-dependence). A hash function is applied to reduce this array back into a new, single-integer identifier. Once all atoms have generated their new identifiers, they replace their old identifiers with their new identifiers. The new atom identifiers are added into the fingerprint set. This iteration is repeated a prespecified number of times [8].

## Some decisions we made regarding fingerprints

* The type: ECFP. Reasoning as mentioned before.
* Binary or integers: Binary, because that was recommended by chatgpt 😊
* Amount of bits: 1024 instead of 2048, because 1024 is recommended and people say that you will get enough info out of 1024. Famke will make both datasets however, just in case.
* Radius: 3. 3 is recommended by rdkit themselves for machine learning.

## Some notes about our plan of action (in general):

Extensive experimental results reveal that (1) Descriptor-based ML models generally slightly outperform fingerprint-based ML models in terms of predictive performance. RF as an ensemble learning approach displays the overall best predictive performance. (2) Single-task graph-based DL models are generally inferior to conventional descriptor- and fingerprint-based ML models, however, the corresponding multi-task models generally improves the average accuracy of kinase profile prediction.

Combining descriptors and fingerprints could improve the performance of models, especially for the fingerprint-based models [9].