I. INTRODUCTION

The field of phylogenetics studies evolutionary relationships between biological entities such as species or individual organisms. A phylogenetic tree can be reconstructed based on the sequence alignment of the species (Felsenstein, 2004). A sequence alignment is writing two or more sequences in rows, where the columns of the alignment, called the sites, describe corresponding positions in the sequences (Chowdhury & Garai, 2017).

The phylogenetic tree contains nodes that are connected by branches (Yang & Rannala, 2012). The leaves of the tree depict the taxa or sequences while the internal nodes depict the ancestors of the taxa. The length of the branches indicates the evolutionary distances between sequences (Felsenstein 2004; Scott & Baum, 2016). The distance between sequences is given by the average number of mutations per alignment site (Yang, 2014).

If two sequences are closely related, the distance between taxa is possible to be estimated, as the number of expected nucleotide substitutions per site is in such cases similar as the observed substitutions (Lemey et al., 2009). However, some sites may have more than one substitution which causes the evolutionary change to be hidden. Hence, to correctly estimate the number of substitutions, it is necessary to have a probabilistic model to describe changes between nucleotides over evolutionary time. The substitution models are used to answer this problem (Yang, 2014).

Substitution models for pairwise sequence distance estimation assume how evolution takes place (Yang & Rannala, 2012). Some common examples include the Jukes-Cantor69 (JC69), Felsenstein81 (F81), and the General Time Reversible (GTR) model. Depending on how complicated these classical methods are, the distance estimation can be increasingly more tedious (Yang, 2014).

In the last years, a lot of neural network research have been applied to phylogenetics (Burgstaller-Muehlbacher et al., 2023; Leuchtenberger et al., 2020; Nesterenko et al., 2022; Suvorov et al., 2020). Since the classical distance estimation methods are tedious and complicated, applying neural networks to estimate evolutionary distance could solve this issue.

In this regard, the current study aims to investigate whether neural networks could successfully estimate the evolutionary distance of taxa within a tree. It is important to note that the aim is not to replace the phylogenetic tree reconstruction method but to have a better distance estimation that may be used later for distance-based methods such as the neighbor-joining algorithm. To achieve this aim, the evolutionary dataset for training, validation, and testing the network can be generated by choosing a phylogenetic tree and then simulating the evolutionary process according to a mathematical model of evolution. After that, the network can learn from the patterns of the simulated alignment data, and its performance is evaluated.