## Natural compounds database collection

A database of 437,859 natural products was compiled to perform virtual screening. The data to construct this database was extracted from different online sources containing information about natural products and nutraceuticals: 406,747 molecules from COCONUT (Sorokina et al., 2021), 107 nutraceuticals from Drugbank(Wishart et al., 2018), 20 nutraceuticals from Drugs.com(*Drugs.Com | List of Nutraceutical Products*, n.d.), 59 nutraceuticals from FooDB(The Metabolomics Innovation Centre, n.d.)and 30,926 natural products from NPASS(Zeng et al., 2018). A curating process was performed on the initial database to remove duplicated molecules, salts, inorganic compounds, and mixtures, obtaining a resulting database of 429,460 compounds.

Exploring the complete dataset of natural compounds would be time consuming. Thus, to reduce the time necessary to perform the virtual screening, the natural compounds database was reduced to 10% of its original size by clustering to select the most varied compounds. First, to identify each compound, MACCS fingerprints were calculated using RDkit(Landrum, 2010). Later, the clustering of the molecules was performed using the density-based OPTICS (Ordering Points To Identify the Clustering Structure) method(Ankerst et al., 1999) with the Jaccard index as the metric to evaluate molecular similarity, implemented in Scikit-learn(Pedregosa et al., 2011). This process resulted in 43,338 structurally varied molecules to perform the virtual screening, using the models for predicting the inhibitory capacity of small molecules implemented in the server.

Refs:

Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). OPTICS. *ACM SIGMOD Record*, *28*(2), 49–60. https://doi.org/10.1145/304181.304187

*Drugs.com | List of nutraceutical products*. (n.d.). Retrieved December 15, 2021, from https://www.drugs.com/drug-class/nutraceutical-products.html

Landrum, G. (2010). *RDKit: Open-source cheminformatics.* https://www.rdkit.org/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning. *Journal of Machine Learning Research*, 2825–2830. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.OPTICS.html

Sorokina, M., Merseburger, P., Rajan, K., Yirik, M. A., & Steinbeck, C. (2021). COCONUT online: Collection of Open Natural Products database. *Journal of Cheminformatics*, *13*(1), 1–13. https://doi.org/10.1186/S13321-020-00478-9/FIGURES/4

The Metabolomics Innovation Centre. (n.d.). *FooDB*. Retrieved December 10, 2021, from https://foodb.ca/

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., MacIejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, Di., … Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, *46*(D1), D1074–D1082. https://doi.org/10.1093/NAR/GKX1037

Zeng, X., Zhang, P., He, W., Qin, C., Chen, S., Tao, L., Wang, Y., Tan, Y., Gao, D., Wang, B., Chen, Z., Chen, W., Jiang, Y. Y., & Chen, Y. Z. (2018). NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Research*, *46*(D1), D1217–D1222. https://doi.org/10.1093/NAR/GKX1026