Symbolic regression for the derivation of mathematical models directly from the data
by A. Murari¹, E.Peluso², M.Gelfusa², M.Lungaroni², P.Gaudio² and JET Contributors
EUROfusion Consortium, JET, Culham Science Centre, Abingdon, OX14 3DB, UK
¹) Consorzio RFX (CNR, ENEA, INFN, Università’ di Padova, Acciaierie Venete SpA), Corso Stati Uniti 4, 35127 Padova, Italy.
²) University of Rome “Tor Vergata”, Via del Politecnico 1,00133 Roma, Italy
§See the Appendix of F. Romanelli et al., Proceedings of the 25th IAEA Fusion Energy Conference 2014, Saint Petersburg, Russia

Nowadays, machine-learning tools are routinely used to extract knowledge from enormous amount of data; in fact they are typically deployed to handle theory-less applications, from the identification of trends in customer behaviour to voice recognition, fields where formulation of models in consolidated mathematical terms is not a priority. The penetration of these more traditional machine learning tools in many scientific fields, such as physics or engineering, has been more limited, given their limitations. Indeed, their results are often difficult to interpret and/or to express in terms of easily manageable mathematical forms, so their integration with theoretical models, based on first principles, proves problematic if not impossible. Moreover, they are not always capable of properly handling error bars in the measurements; the confidence in their results and their extrapolability can therefore be questioned. In this contribution, Symbolic Regression (SR) via Genetic Programming is introduced to prove its capability of deriving models, directly from the data, to be compared with theories. The power of the method is investigated with a series of systematic numerical tests and classification problems. To exemplify the potential of the approach in Nuclear Fusion, it is applied to the problem of deriving empirical scaling expressions directly from multimachine data bases, without the common “a priori” assumption that these have to be power laws. Indeed the results for the case of the energy confinement time and the power to access the H modes show how power laws are not necessarily the best expressions. More advanced applications of SR are also presented, in particular a method to determine the most adequate dimensionless quantity for the problem at hand. Information geometry concepts, such as Geodesic Distance on Gaussian Manifolds, are also applied to the problem of handling probability distributions and to increase the robustness of the results to noise and outliers. Particular attention is also devoted to the statistical properties of the presented new data analysis tools, to quantify their advantages and limitations.