

# Data Integration and Knowledge transfer: Application to the tissue:air partition coefficients

Cédric Gaudin<sup>1, 2</sup>, Gilles Marcou<sup>1</sup>, Philippe Vayer<sup>2</sup>, Igor Tetko<sup>3</sup>, Igor Baskin<sup>4</sup>, Alexandre Varnek<sup>1</sup> \*

<sup>1</sup>Laboratoire d'Infochimie (UMR 7177 CNRS), Strasbourg, France

<sup>2</sup>Technologie Servier, Orléans, France

<sup>3</sup>GSF-Institute for Bioinformatics, Neuherberg, Germany

<sup>4</sup>Department of Chemistry, Moscow State University, Moscow, Russia

Conventional QSAR/QSPR models are built only for one target property without exploiting any *a priori* knowledge stored in datasets of related properties. Here, individual models are not viewed as separate entities but as nodes in the network of interrelated models. Such interrelated models can be built *in parallel* by means of *multitask learning (MTL)*, or *sequentially* using *feature nets (FN)*. MTL and FN are kinds of data integration, as opposed to traditional *single-task learning (STL)*, in which all models are built separately. We apply this strategy to model Human blood:air, human and rat tissue:air partition coefficients of organic compounds using diverse and relatively small datasets