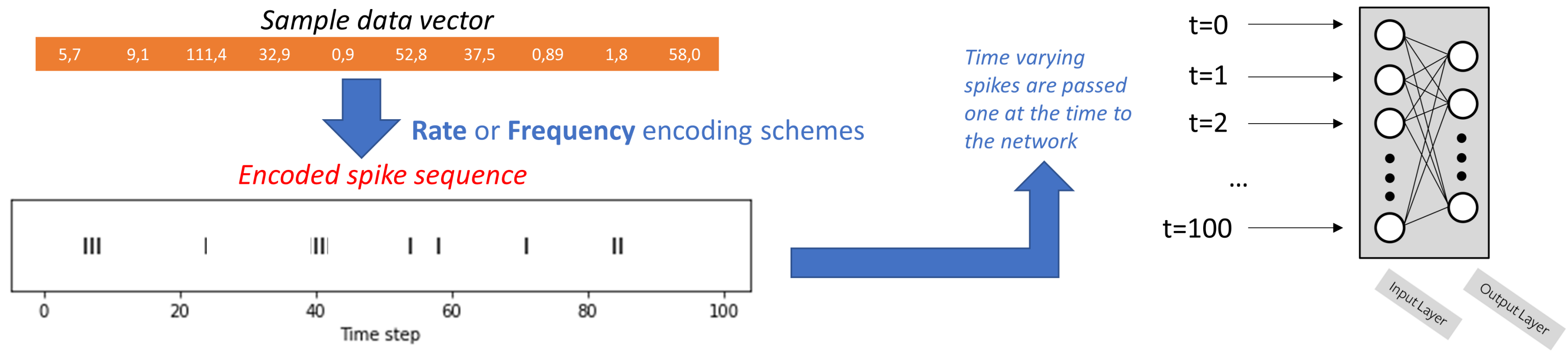
Molecular fingerprint-based spiking neural network QSAR for bioconcentration prediction

Mauro NASCIMBEN [1,3] Silvia SPRIANO [2] Manolo VENTURIN [1] Lia RIMONDINI [3]

[1]Enginsoft SpA, Padua, Italy
[2]Department of Applied Science and Technology, Polytechnic University of Turin, Turin, Italy
[3]Department of Health Sciences, University of Eastern Piedmont, Novara, Italy

Spiking Neural Networks

As happens in human brain neurons with action potentials, a neuron in SNN will fire when the accumulated stimuli reach a threshold value (e.g., membrane threshold). While traditional NN inputs are static analog values, SNNs operate based on dynamic binary spiking inputs as a function of time. The logic behind this is energy efficient because input data is suppressed in the absence of changes (zeros in the binary sequence). SNNs have become popular as a possible pathway to enable low-power event-driven neuromorphic hardware. SNNs are summarized by the “three S” characteristic: Spikes, Sparsity and Static suppression (a.k.a, event-driven processing).



Bio–concentration dataset

Bioconcentration is a natural biological phenomenon when organisms use proteins to uptake and sequester molecules in the intracellular space to utilize during cellular processes: for example, metal ions are involved in enzyme catalysis, signaling, stabilizing charges on biomolecules. In normal conditions, bioconcentration is reduced by kidneys performing osmotic work to concentrate the unwanted stuff for expulsion. However, the magnification of chemicals and nanomaterials in living organisms could have toxic effects on their tissues. In vivo and in vitro tests highlighted how bioaccumulation of metal ions affects cell viability and induces oxidative stress and DNA damage [3]. Inside human body, gut bacteria's accumulation of therapeutic drugs prevents their chemical transformation and metabolization [2]. This reduces drug availability and changes metabolite secretion leading to side effects or even affecting the mode of action of some drugs.

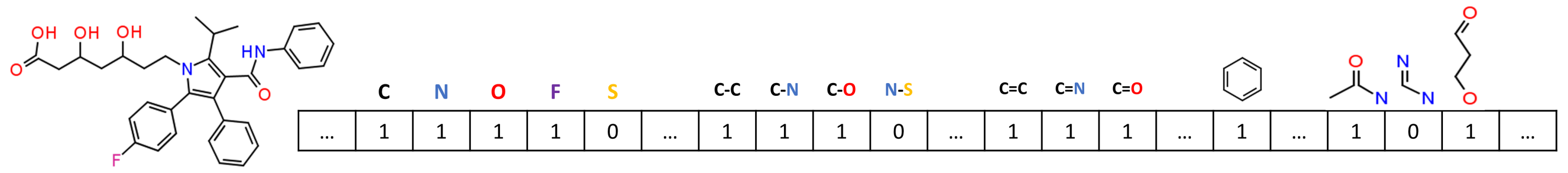
The dataset under investigation [1] was donated to the UCI repository in 2019, and collects SMILES and nine functional descriptors from 779 molecules to predict whether a chemical:

- is mainly stored within lipid tissues (lipophilicity)
- has additional storage sites (e.g., proteins) that increase the bioconcentration
- is metabolized/eliminated (thus reducing the bioconcentration)

Data comes already split randomly into a training set of 584 compounds (75%) and a test set of 195 compounds (25%).

Molecular fingerprints

In our analysis, SMILES ASCII strings were converted in 167 bit-vectors MACCS fingerprints using RDKit cheminformatics Python library. Binary sequences as fingerprints could be directly employed as time-varying input data for SNNs.



Feature engineering on molecular data

The initial nine functional descriptors of the original dataset were replaced by a set of 1613 2-D molecular attributes calculated with the Mordred Python library.

1. **Functional attributes reduction**:
   - removing incomplete attributes generated by the Mordred library (presence of NaN or all zero columns)
   - using the variance threshold approach (unsupervised attributes elimination)
   - finally, non-informative features were removed applying univariate feature selection (F-test for feature scoring)
2. **Training set class equalization**: the dataset comes with unbalanced examples from each class. The number of training samples for each class was equalized by oversampling until reaching a multiple of the batch size (384 examples for each of the three classes).
3. **Conversion from vectors to 2D matrices**: each example of structural (fingerprints) and functional (molecular attributes) data was converted into squared normalized grayscale images. In this way, the SNN was fed with a tensor of stacked 2D matrices equal to batch size.

Molecular attributes were encoded into Poisson spike trains and taking the neuron with the highest firing rate (or spike count) as the predicted class (rate coding).

Present SNN Network configurations

Structural and functional attributes were analyzed by two separate shallow SNN with fully connected neurons and one hidden layer of 3000 and 4000 neurons, respectively.

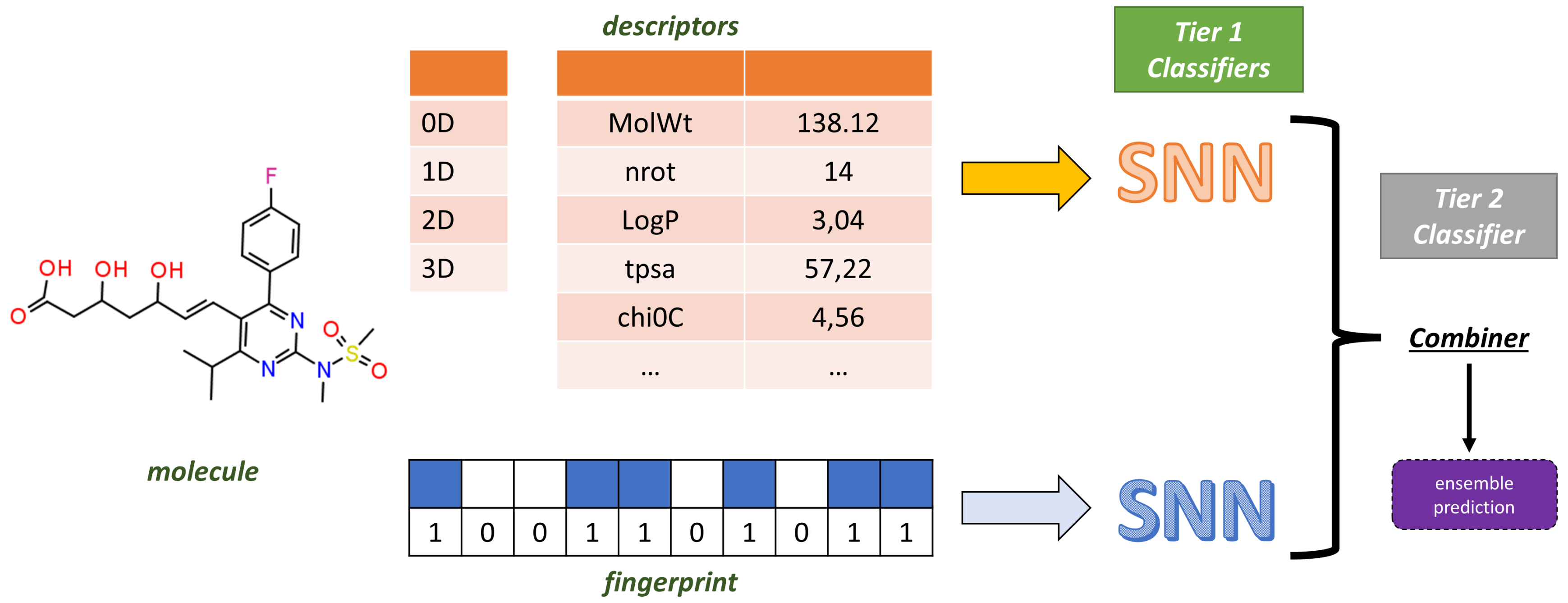
- The discrete nature of spikes makes it difficult for auto differentiation tools to calculate a gradient. To address the dead neuron problem are available for SNNs in the form of surrogate gradients.
  1. a **forward pass** with shifted Heaviside step function
  2. a **backward pass** with the gradient of fast sigmoid function
- Current configuration applied RMSprop in both SNN networks as weight updater at each backpropagation step.

Preliminary classification outcomes

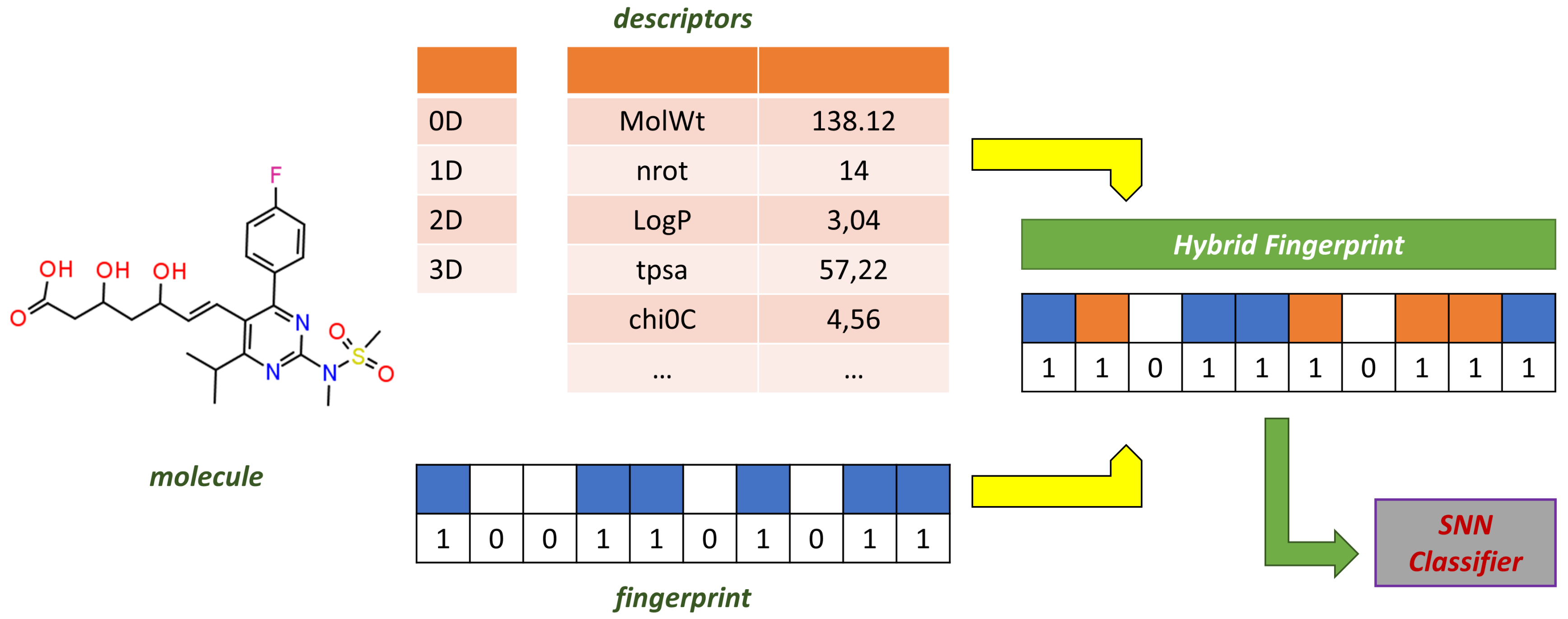
Data type	Neuron model	α	β	Bal. Acc.
Structural fingerprint	2nd-order IF	0.9	0.8	75.902%
Molecular attributes	2nd-order IF	0.8	0.5	62.103%

Chance level for this 3-class problem was estimated using the Scikit-learn Dummy classifier on functional attributes (32.99%).

Future developments



SNN ensemble to predict information coming from both structural and functional data



Creation of an hybrid fingerprint that merges structural and functional data as input sequence for SNN

[1] Francesca Grisoni, Viviana Consonni, Marco Vighi, Sara Villa, and Roberto Todeschini. Investigating the mechanisms of bioconcentration through qsar classification trees. *Environment international*, 88:198–205, 2016.

[2] Martina Klünemann, Sergej Andrejev, Sonja Blasche, Andre Mateus, Prasad Phapale, Saravanan Devendran, Johanna Vappiani, Bernd Simon, Timothy A Scott, Eleni Kafkia, et al. Bioaccumulation of therapeutic drugs by human gut bacteria. *Nature*, pages 1–6, 2021.

[3] Carlos Lopez-Chaves, Juan Soto-Alvaredo, Maria Montes-Bayon, Jörg Bettmer, Juan Llopis, and Cristina Sanchez-Gonzalez. Gold nanoparticles: distribution, bioaccumulation and toxicity. in vitro and in vivo studies. *Nanomedicine: Nanotechnology, Biology and Medicine*, 14(1):1–12, 2018.

European Union's Horizon 2020 research and innovation programme – MSC grant agreement No 860462

Mathematical Modelling and Control for Healthcare and Biomedical Systems Virtual Workshop — Italian National Research Council, September 28-30, 2021