Unsupervised Drift Compensation Based on Information Theory for Single-Molecule Sensors

Mohamed Ouqamra [1,2] Delphine Bouilly [1,3,4]

(1) Institute for Research in Immunology and Cancer (IRIC), Université de Montréal
(2) Institute for Biomedical Engineering, Department of Pharmacology and Physiology, Faculty of Medicine, Université de Montréal
(3) Department of Physics, Faculty of Arts and Sciences, Université de Montréal
(4) RQMP Regroupement Québécois sur les Matériaux de Pointe

Laboratory of Nanobioelectronics

MLSP2019 IEEE International Workshop on MACHINE LEARNING FOR SIGNAL PROCESSING

Multiscale Signal Compression to Multicomponent Signal Decomposition

- **Single-molecule sensors based on carbon nanotubes** transducer, enable to probe **stochastic molecular dynamics** thanks to long acquisition periods and high-throughput measurements. With such sampling conditions, the sensor baseline may drift significantly and induce fake states & transitions in the recorded signal, leading to wrong kinetic models.
- We present **MDL-AdaCHIP:** a **multiscale signal compression** technique based on the **Minimum Description Length (MDL)** principle, combined with an **adaptive piecewise cubic Hermite interpolation (AdaCHIP)**, both implemented into a **blind source separation framework** to compensate the parasitic **baseline drift** in **single-molecule biosensors**.

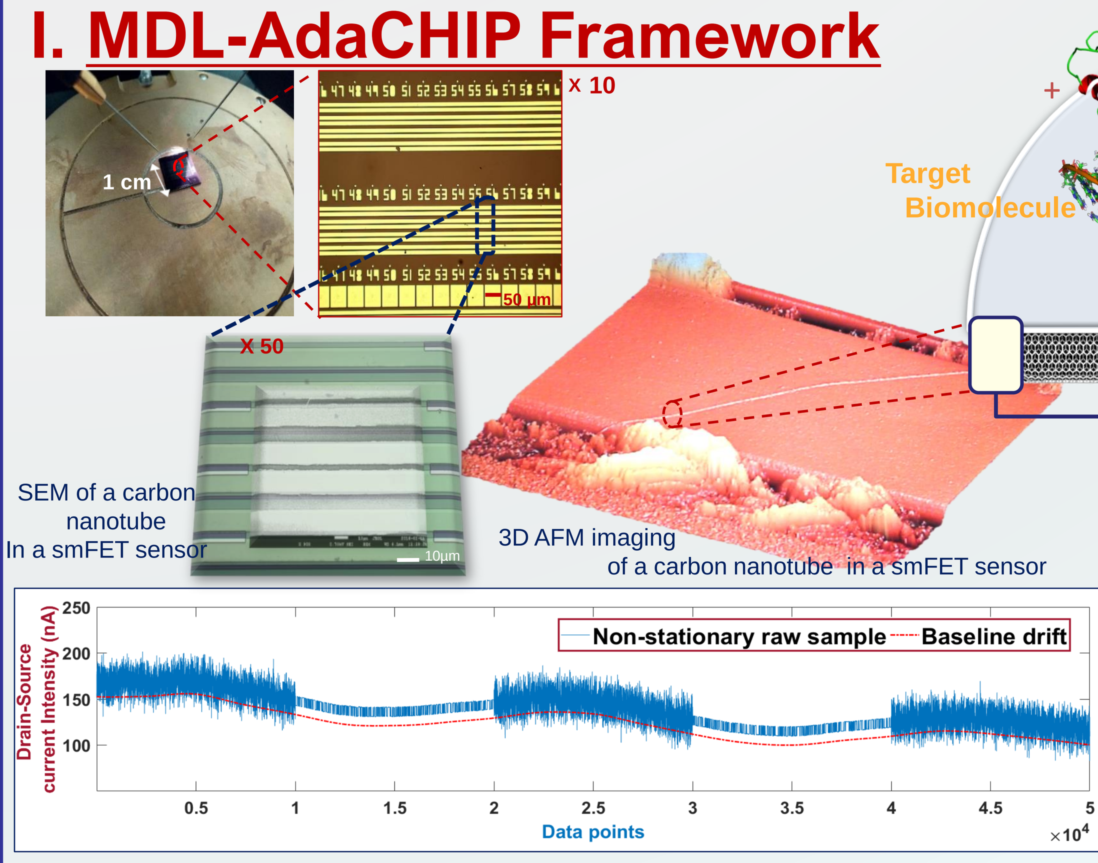
Signal Constraints & Specificities

1. **Non-stationarity**
2. **Multi-source signals**
3. **$(SNR < 1dB)$ + mixed colored noises**
4. **Data size (sampling rate 25kHz during ~1h)**

Contributions

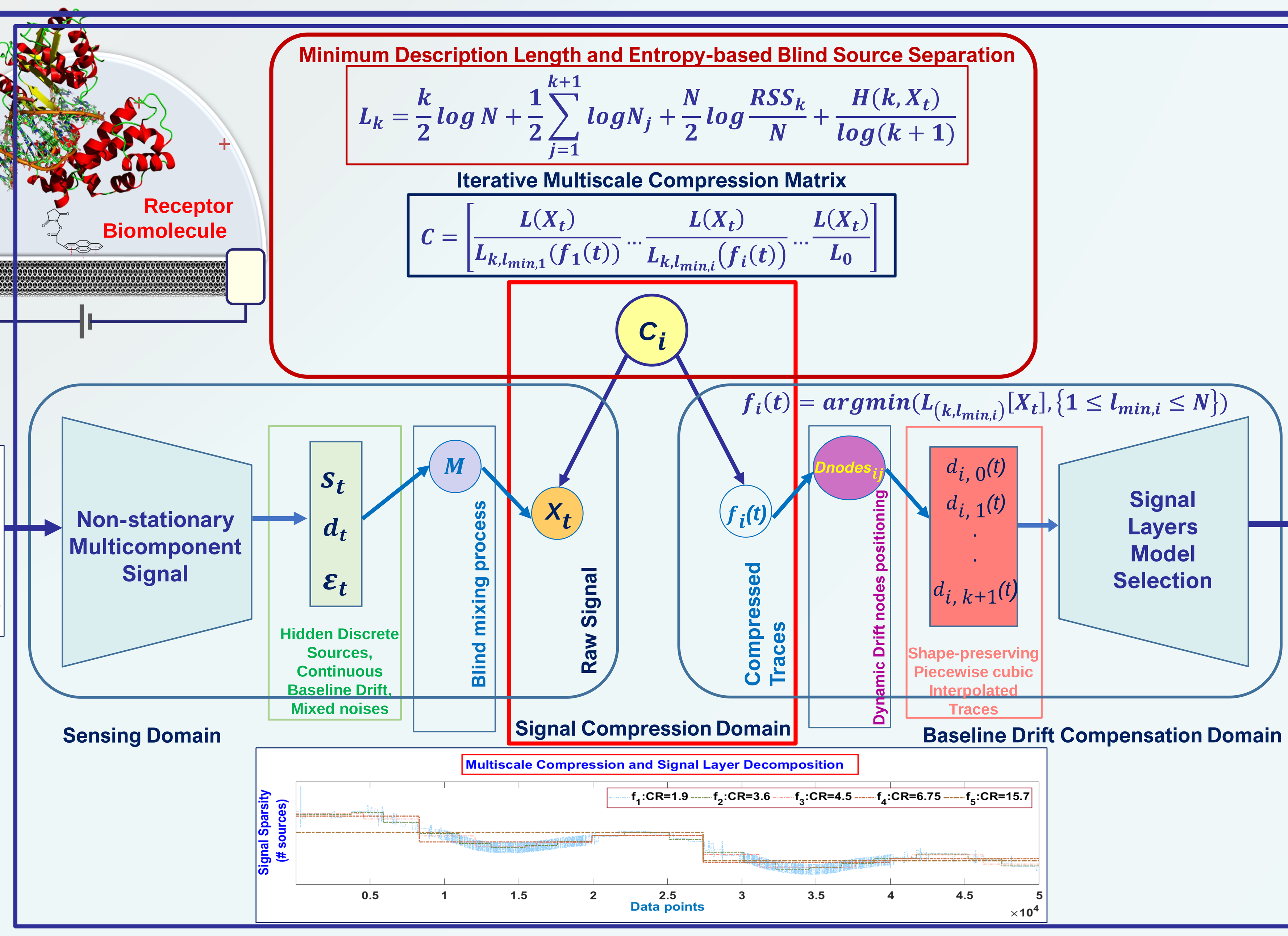
1. **New Baseline Drift Compensation tool** tailored for a wide range of **non-stationary** biosignals (FRET, smFET, ECG, EEG, PCG, EMG)
2. **Automatic baseline wander correction** without signal prefiltering , data preparation nor post-processing.
3. **Model-free unsupervised** baseline drift compensation method, where baseline parameters are learned from the raw signal without any prior knowledge on the sensor characteristics nor on the underlying kinetics of the probed phenomenon.
4. **Robustness to high noise level $(SNR < 1dB)$ and mixed colored noises**
5. **Fast computational time $\mathcal{O}(nlogn)$, user-friendly implementation**

I. MDL-AdaCHIP Framework

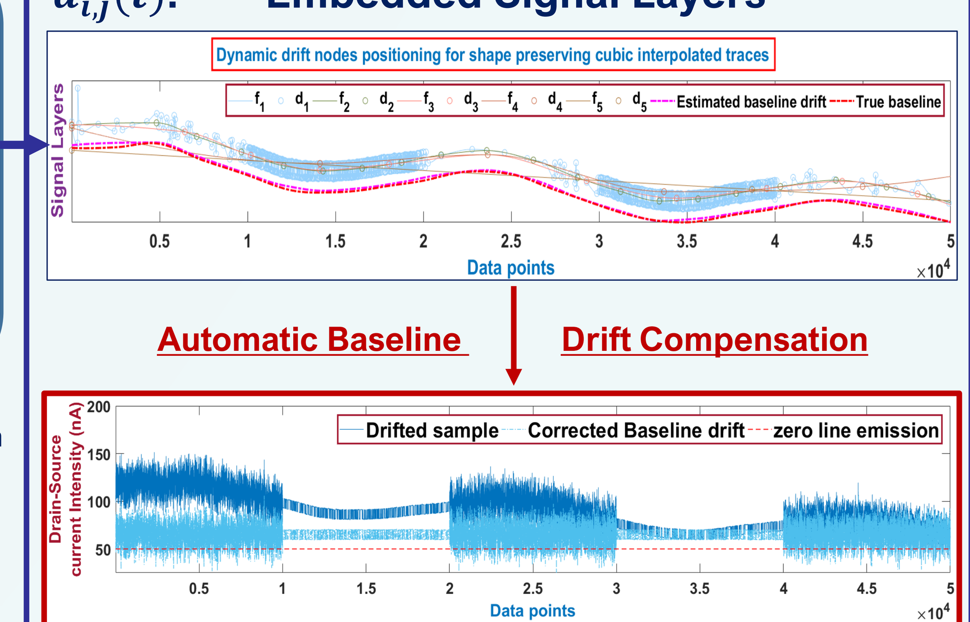


A 3-steps **Signal Decomposition** implementation into a **Blind Source Separation** framework to compensate **the baseline drift** of **single-molecule sensors** under **non-stationary environments:**

① Iterative Multiscale Signal Compression
② Unsupervised Dynamic Drift Nodes Positioning
③ Adaptive Piecewise Cubic Hermite Interpolation



s_t: Hidden sources
d_t: Sensor Baseline Drift
ε_t: White Gaussian + Flicker noises
M: Blind Mixing Matrix
X_t: Raw Signal
L_k: Codelength of the Compressed Trace
C: Multiscale Compression Matrix
C_i: Compression Ratio
$f_i(t)$: Compressed Trace
$Dnodes_{ij}$: Drift nodes
$d_{i,j}(t)$: Embedded Signal Layers



II. Information Theory to Sensor Baseline Drift Modeling

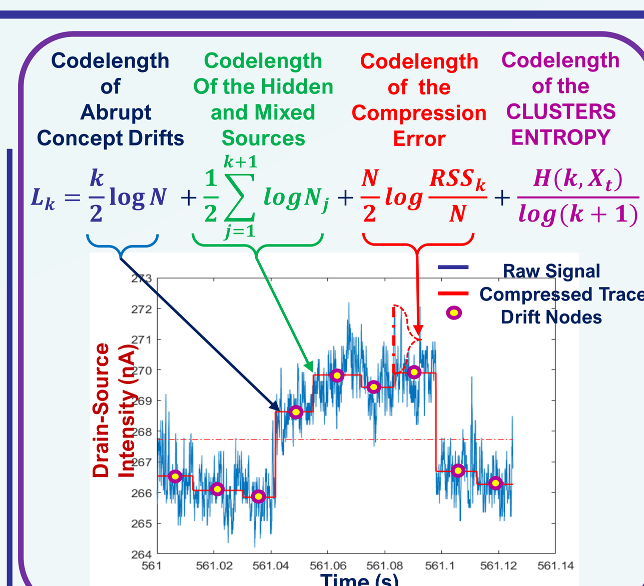
2.1 Multiscale Compression to Signal Decomposition:

An entropy-based signal compression cost function L_k acts as a **soft clustering-based source separation method**, by assigning to each data point a source membership coefficient, enabling to handle mixtures of sources emitting both simultaneously or sequentially:

$$L_k = \frac{k}{2}\log N + \frac{1}{2}\sum_{j=1}^{k+1} logN_j + \frac{N}{2}log\frac{RSS_k}{N} + \frac{H(k,X_t)}{log(k+1)} \quad (1)$$

For a compression ratio C_i, the compressed trace $f_i(t)$ is obtained in (2) by minimizing the cost function L_k described in (1). To do so, we scan the signal X_t for k concept drifts given a source propensity threshold of $l_{min,i}$

$$f_i(t) = argmin(L_{(k,l_{min,i})}[X_t], \{1 \le l_{min,i} \le N\}) \quad (2)$$



Source having the longest propensity serve as the minimum compression support $l_{min,i+1}$ in the next iteration

$$l_{min,i+1} = \max(steplength[f_i(t)]), l_{min,i+1} \in [1,N] \quad (3)$$

The compression ratio thus increases at each iteration, while the compression bounds are:

$$L(X_t): L_{k,l_{min,1}}(f_1(t)) \le C_i \le L(X_t): L_0 \quad (4)$$

$L(X_t)$ being the average number of bits of information per symbol in the raw signal X_t, calculated using a Huffman coding scheme

$$H(X_t) \le L(X_t) = \sum_{a=1}^{M} CW(a)\, p_a \le H(X_t) + 1$$
$$H(X_t) = \sum_{a=1}^{M} p_a log_2[\frac{1}{p_a}] \quad (5)$$

2.2 Dynamic Drift Nodes Positioning

The middle of the steplength of each separated source serves to position nodes for a cubic Hermite interpolating polynomial function $d_i(t)$ in $(k+2)$ intervals:

$$d_{i,j}(t) = a_j + b_j(t-x_j) + c_j(t-x_j)^2 + d_j(t-x_j)^2(t-x_{j+1}) \quad (6)$$

Each $d_{i,j}(t)$ is the local cubic polynomial interpolating function of $f_i(t)$ in the j^{th} of $(k+2)$ intervals, while a_j, b_j, c_j, d_j are the polynomial coefficients to estimate.

2.3 Adaptive Model Selection of Signal Layers

A **shape-preserving piecewise cubic Hermite interpolation** on each compressed trace $f_i(t)$ is applied with one interpolation node $Dnodes_{i,j}$ per separated source $s_{i,j}$ in each compressed trace, using the "pchip" Matlab® function on the set of nodes:

$$d_i(t) = pchip(Dnodes_{i,j}, f_i(t), X_t) \quad (7)$$

$$Dnodes_{i,j} = 0,5 \times steplength(s_{i,j}) \quad (8)$$

The baseline drift $d(t)$ is thus inferred by a **model selection of signal layers** having frequencies of concept drifts between sources comparable to baseline oscillations rate:

$$d(t) = \frac{1}{m-2}\sum_{i=2}^{m-1} d_i(t) \quad (9)$$

III. Evaluations & Validation

✓ We tested a large variety of **baseline drift scenarios**, using a multi-parametric simulated baseline drift function $d_{sim}(t)$:

$$d_{sim}(t) = \sum_{i=1}^{n} A_i \sin(\omega_i t + \varphi_i) - C_t + D \quad (10)$$

(t: the time, C: the slope of the linear component, ω_i and φ_i: the frequency and phase of the n^{th} oscillating component, A_i and D: coefficients allowing to modulate baseline function amplitude)

✓ **The Robustness of MDL-AdaCHIP** on a wide range of **real biosignals** (ECG, EMG, foetal PCG, FRET) confirms its **reliability** to compensate the baseline drift under challenging situations.

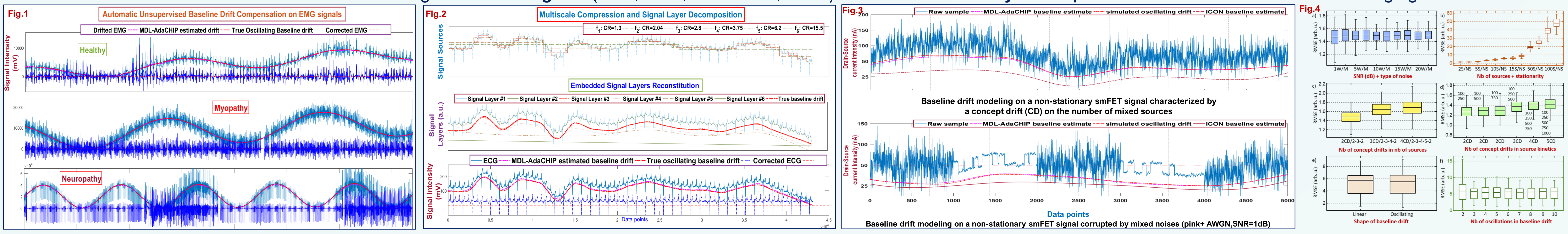


Fig1: Baseline drift compensation on EMG traces from a healthy (top), myopathy (middle) and neuropathy (top) patients recordings (EMG-DB PhysioBank) , corrupted by a baseline wander with changing direction and amplitude over time **Fig2**: Baseline drift removal on ECG recording (ECG-IDDB PhysioBank) The MDL-AdaCHIP is performed to highlight the embedded signal components through the iterative multiscale compression ratios (CR) (top). The baseline drift is then inferred thanks to a model selection of the signal layers, excluding the first and last compression levels containing the sensed molecular dynamics and the signal average intensity respectively. (middle and bottom panels) **Fig3**: Comparison between the proposed MDL-AdaCHIP and a Bayesian non-parametric approach for baseline modeling of non-stationary signals with CD in the number of sources (top) and the type of noise (bottom) **Fig4**: Performance of the baseline drift modeling given various sensor and signal parameters. Each boxplot represents the distribution of RMSE between estimated and true baseline drift for 200 traces of 5000 points.

IV. Conclusion & Future Work

- The originality of **MDL-AdaCHIP** relies on an **alternative blind source separation method to the decomposition of non-stationary multicomponent signals** based on the **minimum description length principle.**
- Thanks to an **iterative multiscale signal compression**, we manage to **segregate the signal layer containing the observations corresponding to the sensed phenomenon, from the signal layer corresponding to the sensor baseline, without any signal prefiltering nor supervision.**
- Finally we intend to implement the **on-line version of MDL-AdaCHIP** through a **compressive sensing approach.**

Reference:

[1] R. Gnanasambandam, M. S. Nielsen, C. Nicolai, F. Sachs, J. P.Hofgaard, and J.K.Dreyer, "Unsupervised Idealization of Ion Channel Recordings by Minimum Description Length: Application to Human PIEZO1-Channels," *Frontiers in Neuroinformatics*, vol. 11, pp.1–16, 2017.
[2] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, Sept. 1978.
[3] Luo Yang and Zhu Huiyan, "Shape preserving piecewise cubic interpolation," *Applied Mathematics*, vol.11, no. 4, pp. 419–424, Dec. 1996.

Acknowledgements:

- Colleagues in the Bouilly's Nanoelectronic Laboratory
- IRIC doctoral fellowship
- Bioinformatics Laboratory-IRIC

INSTITUT DE RECHERCHE EN IMMUNOLOGIE ET EN CANCÉROLOGIE
Université de Montréal
CANADA 150 CHAIRES DE RECHERCHE RESEARCH CHAIRS
INNOVATION.CA CANADA FOUNDATION FOR INNOVATION FONDATION CANADIENNE POUR L'INNOVATION