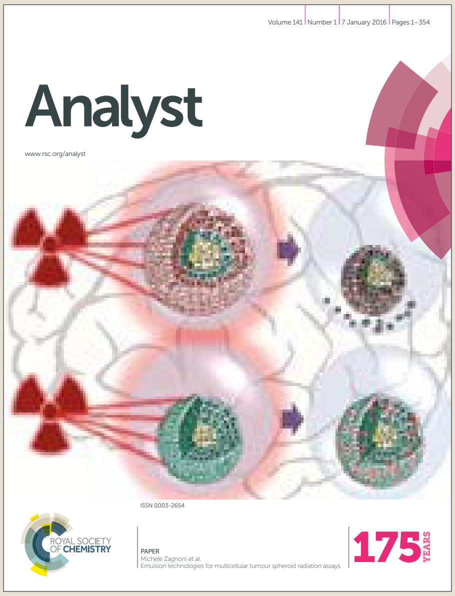
Volume 141 | Number 1 | 7 January 2016 | Pages 1–354
Analyst
www.rsc.org/analyst
ISSN 0003-2654
ROYAL SOCIETY OF CHEMISTRY
PAPER
Michele Zagnoni et al.
Emulsion technologies for multicellular tumour spheroid radiation assays
175 YEARS

View Article Online
DOI: 10.1039/C8AN00065D

P*ost hoc* support vector machine learning for impedimetric biosensors based on weak protein-ligand interactions

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Y. Rong[a], A.V. Padron[a], K. J. Hagerty[a], N. Nelson[b], S. Chi[c,d], N. O. Keyhani[d], J. Katz[e], S.P.A. Datta[f,g], C. Gomes[h], and E.S. McLamore[*a]

Impedimetric biosensors for measuring small molecules based on weak/transient interactions between bioreceptor and target analyte are a challenge for detection electronics, particularly in field studies or in analysis of complex matrices. Protein-ligand binding sensors have enormous potential for biosensing, but accuracy in complex solutions is a major challenge. There is a need for simple *post hoc* analytical tools that are not computationally expensive, yet provide near real time feedback on data derived from impedance spectra. Here, we show use of a simple, open source support vector machine learning algorithm for analyzing impedimetric data in lieu of using equivalent circuit analysis. We demonstrte two different protein-based biosensors to show that the tool can be used for various applications. We conclude with a mobile phone-based demonstration focused on measurement of acetone, an important biomarker related to onset of diabetic ketoacidosis. In all conditions tested, the open source classifier was capable of performing as well, or better, than equivalent circuit analysis for characterizing weak/transient interactions between a model ligand (acetone) and a small chemosensory protein derived from tsetse fly. In addition, the tool has a low computational requirement, facilitating use for mobile acquisition systems such as mobile phone. The protocol is deployed through Jupyter notebook (an open source computing environment available for mobile phone, tablet, or computer use) and the code was written in Python. For each of the applications we provide step-by-step instructions in English, Spanish, Mandarin, and Portuguese to facilitate widespread use. All codes were based on *scikit-learn*, an open source software machine learning library in the Python language, and were processed in Jupyter notebook, an open-source web application for Python. The tool can easily be integrated with mobile biosensor equipment for rapid detection, facilitating use by a broad range of impedimetric biosensor users. This post hoc analysis tool can serve as a launchpad for convergence of nanobiosensors in planetary health monitoring applications based on mobile phone hardware.

Introduction

Biosensors offer rapid analysis of targets ranging from small molecules, to biomolecules or cells, and can be applied across a wide variety of planetary health applications in medical, agricultural, and environmental analysis[1, 2]. With the advent of mobile phone electrochemical and plasmonic acquisition systems [3-5], the portfolio of biosensors used in applied field studies is rapidly expanding. Biosensor accuracy, speed, range, and limit of detection are a function of the nature of molecular interactions between target analyte and bioreceptor structure, the transduction mechanism, inclusion of nanomaterials which enhance transduction, type of detection hardware, and acquisition approach (including *post hoc* analysis).

Among the various transduction approaches, electrochemical biosensors are one of the most popular device types, and most current devices combine electroactive nanomaterials (e.g., graphene, nanometal, electropolymers) with biorecognition structures such as enzymes, antibodies, or aptamers, among others [6-10]. Use of transducer nanomaterials enhances signal acquisition, while the biological material is used to impart selective targeting and in some cases, catalyze a reaction[11-13]. Impedimetric biosensors are most commonly developed based on Faradaic impedance (with redox couple in solution), but label-free biosensors using non-Faradaic impedance (absence of redox couple) are increasing in popularity[14, 15]. In either case, the output impedance depends on changes in the interfacial electron transfer resistance and/or electrostatic repulsion that result from steric hindrance caused by interactions of the target and bioreceptor[16-18].

Interpretation of impedimetric biosensor data is often not trivial, particular for fast electron transfer processes in nanomaterial-modified electrodes, non-Faradaic impedance, or weak/transient interactions between bioreceptor and target. Impedance data are usually fit to an equivalent circuit model using Chi^2 testing, and parameters derived from the model to describe the underlying electrochemistry. Changes in equivalent circuit parameters are commonly reported as sensor output, although impedance at a single frequency is occasionally used as sensor output[14]. Equivalent circuit analysis is based on combinations of the Principle of Superposition, Ohm's Law, and Kirchoff's Laws, and is very accurate for simple electrode geometries with homogenous surfaces. However, as circuit models are assumed *a priori,* there is not necessarily a correspondence between circuit elements and underlying physico-chemical processes[19]. Furthermore, inclusion of transducer nanomaterials on the sensor surface complicates the equivalent circuit model, requiring additional “fitting”

[a] *Agricultural & Biological Engineering, Institute of Food and Agricultural Sciences, University of Florida.*
[b] *Biological & Agricultural Engineering, North Carolina State University.*
[c] *Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences; Key Laboratory of Microbial Resources, Ministry of Agriculture, Beijing, China.*
[d] *Department of Microbiology and Cell Sciences, Institute of Food and Agricultural Sciences, University of Florida.*
[e] *Department of Oral and Maxillofacial Diagnostic Sciences, University of Florida*
[f] *MIT Auto-ID Labs, Department of Mechanical Engineering, Massachusetts Institute of Technology,*
[g] *Biomedical Engineering Program, Department of Anaesthesiology, Massachusetts General Hospital, Harvard Medical School.*
[h] *Department of Mechanical Engineering, Iowa State University*

† Footnotes relating to the title and/or authors should appear here.

Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/x0xx00000x

View Article Online
DOI: 10.1039/C8AN00065D

elements. Thus, interpreting impedance data in complex solutions or with complex electrode geometries is challenging, and is sometimes more art than science[20]. The main challenge for planetary health biosensors is to balance enhancing conductivity with transducer nanomaterials (improving limit of detection) while limiting computational cost (maintaining speed), and at the same time developing simple label-free devices that can be used in diverse applications (ensuring robustness).

To improve limit of detection, many labs coat electrodes with nanomaterials such as graphene and/or nanometal, which is known to significantly enhance conductivity and electroactive surface area[21-23] while significantly decreasing charge transfer resistance (R_{ct}). This results in fast electron transfer processes, where Faradaic current is represented by a near-linear Nyquist plot. In a classic Randles-Ershler equivalent circuit, *post hoc* sensor analysis is usually constrained to R_{ct}, as other circuit parameters are a function of the solution resistance or inductance, which are not strong indicators of molecular interactions between bioreceptor and target analyte. This situation is particularly challenging for weak/transient interactions, where more complex circuit models with fitting parameters are required, increasing the computational cost while producing output parameters that have no physico-chemical meaning in the electrochemical circuit. There is a need for simple *post hoc* analytical techniques that can be used for point of need biosensors, particularly for field applications.

Machine learning has emerged as a powerful *post hoc* analytical tool for a wide range of sensor applications, including: flow cytometry[24], electronic tongue/nose[25-27], wearable sensors[28, 29], whole organism biosensing[30, 31], protein detection[32], sensor material optimization[33], food safety risk analysis[34], environmental pollutant monitoring[35] and multiplexing sensors arrays [36-38].

Here, we present an open source machine learning algorithm applied for label-free biosensors based on weak/reversible interactions that can be used with common mobile hardware such as a mobile phone or tablet (**Fig 1**). We first test the classifier for well-known binding interactions between proteins and DNA as a proof of concept. Next, we challenge the algorithm for classifying impedimetric data from a biosensor based on reversible interactions between a small molecule (acetone) and an insect-derived chemosensory protein. Acetone is an important biomarker in salivary diagnostics of diabetic ketoacidosis (DKA), which is a potentially fatal outcome from complications associated with diabetes. Rapid diagnostic tools are vital, as the overall mortality rate for DKA ranges from 1 to 10% of all patient admissions, and an even higher mortality rate is found among non-hospitalized patients and children under the age of 10 [59].

The machine learning tool is based on Jupyter notebook (open source computing environment available for mobile phone, tablet, or computer use) and the code was written in Python. For each of the applications we provide step-by-step instructions in English, Spanish, Mandarin, and Portuguese to facilitate widespread use for a variety of applications. The open source tool can easily be integrated with mobile biosensor equipment for rapid detection, facilitating use by a broad range of biosensor users.

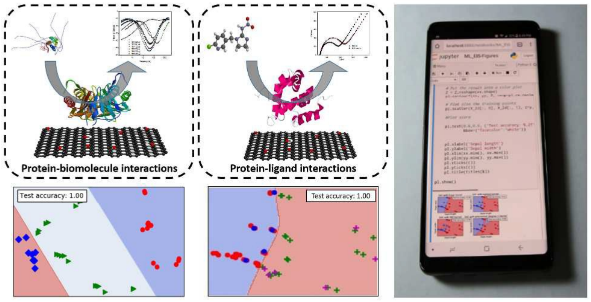

Figure 1. An open source support vector machine learning algorithm was developed for analyzing impedimetric biosensor data. Interactions. We tested the tool for analyzing weak/transient interactions including protein-DNA, protein-protein, and protein-small molecule. The cloud-based tool can be used for point of need applications with a mobile phone or tablet.

Methodology

Strains and reagents

Escherichia coli strain Rosetta DE3 (Promega, Madison WI, USA) was routinely grown in Luria-Bertani broth (LB) and/or on LB-agar (1.5%) plates containing 50 µg/mL kanamycin. All reagents and chemicals were purchased from Sigma-Aldrich (St. Louis, MO, USA) or Thermo Fischer Scientific (Waltham, MA, USA) except as noted. Potassium ferrocyanide (K_4FeCN_6), potassium ferricyanide ($K_3[Fe(CN)_6]$), and potassium chloride (KCl) were purchased from EMD chemicals (Billerica, MA, USA). Ni- and Co-NTA agarose was purchased from Gold Biotech (St. Louis, MO, USA). Thrombin was purchased from Amersham-Pharmacia Biotech (Little Chalford, UK). Polycrystalline diamond suspensions (3 and 1 mm) alumina slurry (0.05mm) were purchased from Buehler (Lake Bluff, IL, USA).

Electrochemical analysis

For all electrochemical analysis, a three-electrode system was used together with an electrochemical impedance analyzer (ERZ100, eDAQ, Colorado, USA). All electrochemical impedance spectroscopy (EIS) studies used Pt/Ir working electrodes (MF-2013, 1.6 mm diameter, BASi, West Lafayette, USA), Ag/AgCl reference electrodes (BASi, West Lafayette, USA) and platinum auxiliary electrodes (BASi, West Lafayette, USA) with nanoplatinum deposited as previously described[41, 42]. Before all experiments, the Pt/Ir working electrodes were polished with two sizes of polycrystalline diamond suspensions (3 and 1 μm), rinsed with methanol, polished with alumina slurry (0.05μm) and then rinsed with deionized water. Probes were cleaned in a sonication bath in DI water for 15 min, then with 0.1 M H_2SO_4 using cyclic voltammetry (CV) at a potential range of -1.0V to +1.0V until the peak current changed by less than 1%, and then finally cleaned in a sonication bath in DI water for 15 min. To ensure consistency during adsorption studies, electrodes were fitted with a plastic cap that was 3D printed on a Makerbot Replicator 2 Desktop 3D printer (see supplemental Figure S1 for specifications).

View Article Online
DOI: 10.1039/C8AN00065D

EIS analyses were conducted at 0.25V (DC), with a 100mV (AC) amplitude in the range of 100 kHz to 1 Hz in a solution with 2.5 mM potassium ferrocyanide ($K_4[Fe(CN)_6]$), 2.5 mM potassium ferricyanide ($K_3[Fe(CN)_6]$, and 100 mM potassium chloride (KCl). For equivalent circuit analysis, all EIS data was transformed to Nyquist Plots and analyzed using ZMAN (WonATech, South Korea) 2.2 software or support vector classification analysis as noted.

Protein expression and purification

Recombinant insect chemosensory proteins (CSP) derived from *Glossina morsitans* (Gmm, tsetse fly) were heterologously expressed and purified from *E. coli* hosts using the methods described in detail by Song et al[43]. Briefly, GmmCSP3 sequences were identified from genomic databases, codon optimized for *E. coli* expression, and synthesized/cloned into a pUC vector (Genewiz, Planefield, NJ). Expression constructs were synthesized with a 10X-histidine (10X C terminal His) tag and transformed into *E. coli* host cells. Single colony isolated via selection for ampicillin resistance on LB/ampicillin plates (100 µg/ml). Cells were harvested by centrifugation, washed with Co^{2+} equilibrium buffer, and suspended in the buffer. Protein purification was achieved using Co^{2+} affinity chromatography and elution of the bound protein with increasing concentrations of imidazole as described in detail by Song et al[43]. Purity of the protein was assessed by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) and pure fractions were pooled and dialyzed (3000kDa) against buffer (20 mM HEPES, pH 7.5). Protein concentration was quantified with SDS-PAGE, and samples were frozen at -80 °C until used.

Expression and purification of TATA binding protein (TBP) and multiprotein bridging factor 1 (MBF1) from *Beauveria bassiana* have been described in detail previously [43]. Briefly, the coding sequences for both genes were codon optimized for expression in *E. coli* and synthesized (Genewiz) as above for the CSP proteins. Expression plasmids were transformed into competent *E. coli* Rosetta DE3 cells for expression and purification as above. Purified proteins were dialyzed, aliquoted and stored at -80°C until used. After protein elution using, purity was confirmed by SDS-PAGE. Protein concentration was determined using Pierce™ BCA Protein Assay Kit (Thermo Scientific).

PCR products were digested, and then cloned into respective sites of an expression vector to produce plasmids. Expression plasmids were transformed into competent *E. coli* Rosetta DE3 cells and transformed *E. coli* were cultured in LB broth, harvested, lysed, and then purified using Ni- or Co-NTA agarose columns. Purified 10XHis-tagged proteins were aliquoted and stored at -80°C until used. After protein elution using imidazole buffers, purity was confirmed by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) and protein concentration was determined using Pierce™ BCA Protein Assay Kit (Thermo Scientific).

Protein adsorption onto sensor surface and sensor characterization

For characterizing acetone-CSP interactions, the optimal CSP concentration from previous studies[44] was used or all testing. Briefly, 2 µL of His-tagged GmmCSP3 (9.23 mg/mL) was drop cast on the surface of an electrodes, dried at 20°C for 5 minutes, and rinsed three times with DI water. A 5mM acetone stock solution was prepared in DI water, which is representative of salivary acetone levels for patients with DKA [45]. Where noted, aliquots (2 µL) of acetone stock solution were drop cast on the surface of the biosensor, stored at 20°C for 2 minutes, and rinsed with DI water three times prior to testing.

For biosensors based on protein-biomolecule interactions, the concentration in each experiment was based on Song et al [43]. A 2 µL aliquot of His-tagged TBP was first drop cast on the surface of the electrode, agitated gently, allowed to dry at 20°C for 5 minutes, and rinsed with DI water prior to impedance analysis. Next, 2 µL aliquots of MBF1 (no His-tag) or $TATA^1$ (a 40 bp DNA sequence containing two potential TATA motifs) was drop cast onto the TBP-functionalized electrode, dried at 20°C for 5 minutes, and then rinsed with deionized water three times prior to impedance analysis. Control experiments included using uncoupled (no TBP) surfaces as well as TBP-coupled + bovine serum albumin (BSA) solutions and TBP-coupled + $TATA^0$ (a 35 bp DNA sequence lacking the TATA sequence in $TATA^1$).

Data analysis and statistics

EIS plots were analyzed with ZMAN 2.2 using an equivalent circuit model based on Chi^2 analysis. Equivalent circuit parameters, namely solution resistance (R_s), charge transfer resistance (R_{ct}), Warburg impedance (Z_w), double layer capacitance (C_{dl}), and constant phase element (Q) were estimated using Chi^2 fitting in the ZMAN software.

Nyquist and Bode plots were generated with ZMAN 2.2 and several key values were extracted within the software from equivalent circuit analysis. Namely, the Nyquist with equivalent circuit analysis were used to extract R_s, R_{ct}, Z_w, and C_{dl} from a Randels-Ershler equivalent circuit. Bode plots were used to extract the impedance at a given cutoff frequency and associated phase angle. In addition to the Randels-Ershler circuit, various equivalent circuit models (shown in supplemental S5) were tested with the model search function in ZMAN software where noted.

Support vector machine (SVM) classification

For protein-ligand interactions, EIS data were exported and transformed into samples with 152 features that represent both real and imaginary impedance at frequencies from 100kHz to 1Hz. The number of features was selected to satisfy expected confidence levels for principle components analysis. A total of 54 EIS scans were randomly split into two groups, with 80% of the data used as the training set and 20% used as the testing set. Each of the 54 data sets were binary labeled, with baseline impedance data in the absence of acetone labeled as "0", and labeled 1 in the presence of 5mM acetone. EIS data for both baseline (no acetone) and positive (5mM acetone) experiments were standardized and transformed into a two-dimensional dataset, and then mapped in a new data space. To initially screen the data, the four most common types of SVM kernels [46] were used to screen the data. A shuffled K-fold cross validation was used for all applications of SVM in this study[47]; the training dataset was divided into ten folds and shuffled,, with 20% of the total data used for testing.

View Article Online
DOI: 10.1039/C8AN00065D

The test accuracy shown for each kernel is the percentage of the prediction accuracy based on the decision boundaries.

Prior to running the SVM algorithm, principal component analysis (PCA) was applied through singular value decomposition (SVD) to reduce the 152 features to two principal components. PCA was used to reduce the dimension of 152 features in the raw EIS data to a two-dimensional principal components matrix. Depending on the number of components to extract, full or randomized truncated SVD was used; this procedure was performed in LAPACK[48]. To ensure generalizability across other varied application-specific biosensors, code screens were prepared for four types of SVM kernels (linear, sigmoidal, radial basis function, and polynomial) to identify which approach best segregates the training data. This feature of the open source algorithm allows the user to select the most appropriate kernel for a given analysis by comparing the cross-validation results across kernel types. SVM hyperparameters (C and gamma) were optimized using grid search and random search methods[46, 49]. C is a tradeoff between misclassification and simplicity of the decision surface. Gamma is proportional to the radius of influence for selected support vectors [50]. All SVM codes were produced with “scikit-learn”, an open source machine learning library in Python [50], and were processed with Jupyter notebook, an open-source web application for Python (see supplemental section for step-by-step instructions in English, Spanish, Mandarin, and Portuguese and Python code). Heatmaps were generated using the built in visualization feature in scikit-learn (see user’s manual for details).

To validate the functionality of the SVM classifier for a well-known detection system, a TBP-protein and TBP-DNA biosensor were fabricated based on published methods[43]. The biosensor is based on interactions between TBP and either multiprotein bridging factor 1 (MBF1) (protein-protein binding) or TBP and TATA (protein-DNA). His-tagged TBP was first adsorbed to the electrode surface, and then EIS was used to study the interactions of TBP with either MBF (a 17 kDa protein) or TATA^{1} (a TBP-binding 40mer DNA sequence). As a control, EIS data were also recorded after addition of buffer, a non-binding protein (BSA), and a non-binding 35mer sequence (TATA^{0}). To challenge the approach for detection of small molecules, **a** CSP biosensor for detecting acetone was **also** developed. The experimental conditions were based on levels relevant to diagnosis of DKA.

All experiments were repeated in triplicate, resulting in a total of 54 data sets. Analysis of variance (one-way ANOVA with Games-Howell method and 99% confidence) and student’s t-test (two-sample t-test with 99.9% confidence) were performed for analyzing EIS data derived from equivalent circuit modeling as noted. All error bars represent the standard deviation of the arithmetic mean.

Results & Discussion

First, the functionality of the SVM classifier was validated for a well-known detection system using TBP-protein and TBP-DNA based on Song et al[43]. This well-documented biosensor produces large changes in impedance after target binding, and serves as a simple case study for the machine learning tool. The biosensor is based on interactions between TBP and either MBF (protein-protein binding) or TBP and TATA^{1}, a 40 mer nucleotide sequence containing the TATA motif that is the recognition sequence bound by TBP, (protein-DNA). His-tagged TBP was first adsorbed to the electrode surface, and then EIS was used to study the interactions of TBP with either MBF (a 17 kDa protein) or TATA^{1}. As a control, EIS data were also recorded after addition of buffer, a non-binding protein (BSA), and a 35-mer nucleotide sequence lacking the TATA motif (TATA^{0}).

Representative Nyquist plots show that adsorption of His-tagged TBP on the sensor surface caused a significant increase in R_{ct}, as expected (**Fig 2a**). Binding between TBP and MBF also resulted in a significant change in charge transfer resistance, as did binding between TBP and TATA^{1}. A Randles-Ershler equivalent circuit (Chi^{2}=1087 ± 212) was used to extract R_s, R_{ct}, Z_w, C_{dl} for each experiment (see supplemental Table S1 for details). Similar to other manuscripts in the literature [51] , R_{ct} was used as the most accurate parameter for characterizing protein-biomolecule interactions (**Fig 2b**). For comparison, addition of BSA or buffer did not result in any significant change in impedance due to non-specific binding (see supplemental Figure S2). EIS data was further analyzed by SVM classification by dividing the dataset into groups of TBP, TBP+MBF, and TBP+TATA^{1}. Of the screened kernels, the linear type successfully classified each of the interactions (test accuracy =100%) and was the simplest of the considered kernel types. Such accuracy was not surprising given the large change in R_{ct} for a protein-biomolecule interaction of this type. Results associated with the other common, but more complex, kernels and their optimization parameters are shown in supplemental Figure S3-S4.

The molecular interactions between TBP and MBF/TATA^{1} shown in **Fig 2** can be viewed as between transient and permanent interactions[52]. For these moderate to tight biomolecule interactions, a basic Randles-Ershler equivalent circuit or SVM (linear kernel) analysis can be used to analyze the data. In the following section, we show that for analysis of much weaker reversible protein-ligand interactions, equivalent circuit analysis is not sufficient and more complicated SVM classification must be used for accurate analysis.

View Article Online
DOI: 10.1039/C8AN00065D

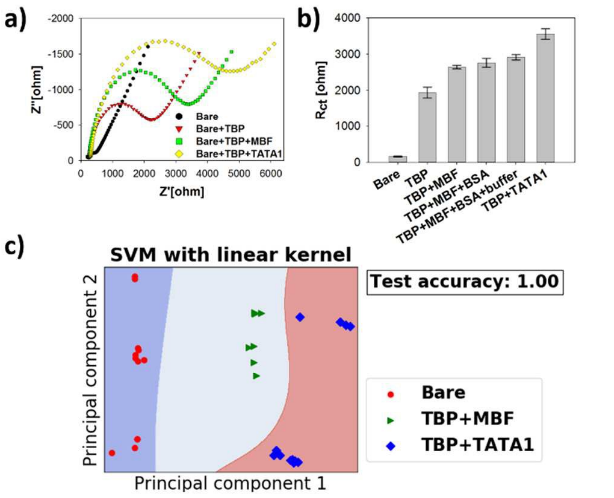

Figure 2. Impedimetric biosensor for the detection of protein-protein or protein-DNA interactions. **a)** Representative Nyquist plots for TBP-MBF interactions or TBP-TATA interactions clearly show an increase in charge transfer resistance after addition of target. **b)** Average R_{ct} derived from Randles-Ershler equivalent circuit model show significant results for TBP-biomolecule interactions but no significant change after addition of buffer or BSA (p values shown for each test group). **c)** Support vector machine (SVM) classification results with linear kernel. Additional Nyquist plots, average charge transfer resistance, and other SVM kernels are shown in the supplemental section.

Reversible protein-ligand interactions

To further challenge the machine learning tool, a CSP biosensor for detecting acetone at levels relevant for DKA triage diagnosis was developed and tested. DKA is a potentially fatal outcome from complications associated with diabetes, and accurate measurement of acetone is challenging. CSP are an excellent candidate for binding volatiles such as acetone, but to date the technology has not been proven. *In vivo*, CSP solubilize volatile odorants and facilitate transport to downstream odorant receptors (ORs) through reversible association/disassociation [39, 40]. This represents a model impedimetric biosensor based on interactions between low molecular weight binding proteins and small molecules. Biosensors based on CSP are becoming popular, but the transient ligand interactions and relationship to underlying electrochemistry are not well documented. Protein size, surface charge, and the nature of any conformation changes/ligand displacement upon binding have not yet been described in detail, although CSPs in general are smaller than 15kDa and the acetone levels critical to DKA (> 5mM) are below protein denaturing levels.

Representative Nyquist plots (**Fig 3a**) and Bode plots (**Fig 3b**) show that the adsorption of CSP onto the electrode caused a significant change in EIS spectra, but the change after addition of clinically relevant acetone (5 mM) was less pronounced. Phase plots (see supplemental Figure S5) had similar behavior, with the most pronounced change in phase angle at a frequency of approximately 1kHz. A Randles-Ershler equivalent circuit was used to derive R_s, R_{ct}, Z_w, and C_{dl} as previously described (**Fig 3c**). In addition, net impedance at various cut-off frequencies was extracted from Bode plots (**Fig 3d**). Using a 99.9% confidence level, there was no significant difference between baseline measurements and average R_s (p=0.015), R_{ct} (p=0.002) Z_w (p=0.016), or impedance at any cut-off frequency (p<0.002) after addition of 5mM acetone.

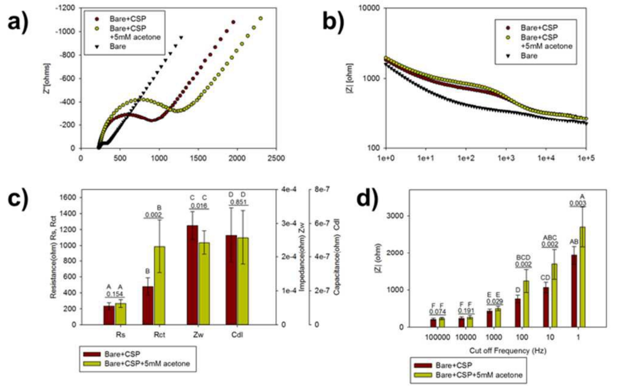

Figure 3. EIS analysis of CSP-acetone interactions in the presence and absence of 5mM acetone. Representative a) Nyquist plots and b) Bode plots. c) Average parameters from Randles-Ershler equivalent circuit analysis (R_s, R_{ct}, Z_w and C_{dl}) from Randel's equivalent circuit, baseline data and EIS in the presence of 5mM acetone. d) Net impedance at representative cut off frequencies. In panels ac and d, numbers denote the p-value and uppercase letters denote statistically significant groups.

To further analyze the spectra, more complex equivalent circuit models were analyzed using ZMAN software with Chi^2 fitting. All equivalent circuits with improved Chi^2 fit (relative to Randles-Ershler) had more than four elements in various parallel/series connections, including at least one resistive element(R), capacitive element(C), constant phase element (Q) and inductive element (I) (see supplemental Fig S6). However, statistical analysis of the output parameters for these circuits also showed no significant difference in baseline and in the presence of acetone for replicate biosensors. Furthermore, there is no direct physical analogous biological structure to the constant phase elements (Q) produced by the model, further complicating the interpretation of the results and inducing bias on the interpretation. Two important factors that could lead to the lack of statistical significance in equivalent circuit parameters are the possibility of CSP conformation changes upon binding, or dislocation of the ligand. For this study, we assume that ligand dislocation is insignificant due to the relatively high concentration of acetone (5mM), but conformation change cannot be ruled out. While these acetone concentrations are significantly lower than denaturing conditions, the levels are high enough to possibly induce CSP conformation changes. Khabiri stet al (2013) have shown in other protein-ligand systems that repulsion of water molecules from the first solvation shell of the protein causes polar amino acid side chains to be more rigid and less likely to interact with water (Khabiri et al., 2013, J. Mol. Model, 19: 4701-4711). More detailed studies are needed to understand the detailed interactions between acetone and CSPs, although in the next section we show that CSPs are a useful biorecognition structure for acetone detection when applying the machine learning tool developed here.

The case study in **Fig 3** represents a common issue in non-Faradaic impedimetric biosensing where the device is based on interaction of proteins and small molecules. In such a case, the individual biosensor responds to target analyte, but variability of replicate sensors is high and interpretation of results at relevant levels is challenging. This is particularly true for weak/reversible interactions between small molecules and proteins where there is not an inherent reaction (as is the case for CSP-ligand binding). The CSP

View Article Online
DOI: 10.1039/C8AN00065D

biosensor system is a promising biomimetic sensor system, but more accurate *post hoc* tools are needed for accurate detection of target biomarkers. As described by Liu et al [53] , the underlying cause for this challenge is likely a result of the nature of CSP-ligand binding in sensors. Liu et al showed that protein conformation change (backbone displacement) plays a major role in the electrical (Faradaic) properties of the sensor; this work was based on the honeybee protein Ac-ASP3. Since conformation changes can occur with non-specific interactions such as hydrogen bonding, CSP biosensors are subject to erroneous outputs due to non-specific interactions. To alleviate the false negative issue shown in **Fig 3**, EIS data was further analyzed by SVM classification.

SVM classification for acetone-CSP interactions

The decision boundaries for each kernel are shown in **Fig 4**, where testing data that fall into blue areas is predicted as negative (no acetone) and those that fall in red areas as positive (≥ 5mM acetone). As discussed by Liu et al [54], other *post hoc* algorithms not analyzed here, such as random forest, may be more accurate in some cases. However, these approaches often increase accuracy by overfitting the data[55], which ultimately decreases the robustness of the classifier. Moreover, many of these are computationally expensive and cannot be analyzed using mobile hardware such as a mobile phone or tablet. The Gaussian radial base function (RBF) kernel (accuracy = 98%) had the highest test accuracy for classifying the training dataset. However, using the default kernel settings the dataset was not linearly separable and the RBF kernel had an overfitting issue, requiring further analysis and tuning of the parameters.

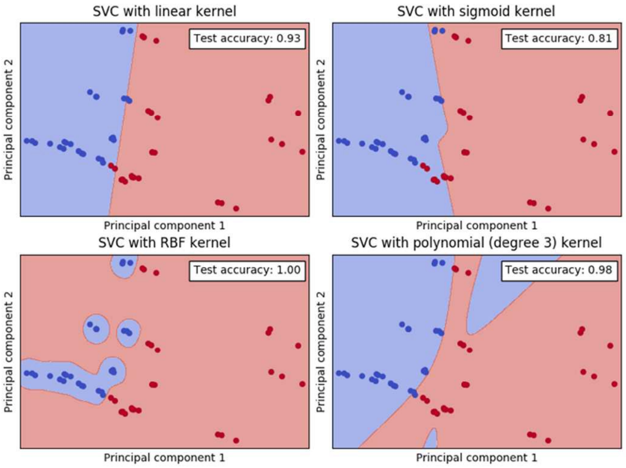

Figure 4. SVM classification for CSP-acetone biosensors using four common kernels. **a)** linear kernel (test accuracy =96%), **b)** sigmoidal kernel (test accuracy =83%), **c)** radial base function kernel (test accuracy =98%), and **d)** polynomial kernel (test accuracy =96%). Blue dots represented baseline EIS signals (no acetone in samples) and red dots represented positive EIS signals (5mM acetone in samples). The decision surface of these four SVM classifiers are plotted by red and blue regions.

To tune the RBF kernel parameters, a grid search and cross validation were performed. In cross validation, the original dataset was shuffled and divided into ten different training and testing sets, with 20% of the total data used for testing. Next, each training set was used to fit the SVM classifier and average test accuracy calculated for each split training set. In the RBF kernel, the two governing hyper-parameters are the penalty parameter (C) and non-linear kernel coefficient (γ). The penalty hyper-parameter trades off misclassification against simplicity of decision surface, where lower C values tolerate more mistakes. The non-linear parameter defines the influence of a single training example on the output, and can be seen as the inverse of the radius of the influence of support vectors[50]. Each of these parameters were optimized using a grid parameter search function using the RBF kernel (**Fig 5**). In the top left panel of **Fig 5**, where C is low, the penalty for misclassification is small and the decision surface is simple relative to values in the first column with higher C values. As the nonlinear hyper-parameter increases (from left to right in **Fig 5**), the influence radius decreases, causing over-fitting. The protocol described herein resolves this issue by creating a visualization tool to select the optimum hyper-parameters.

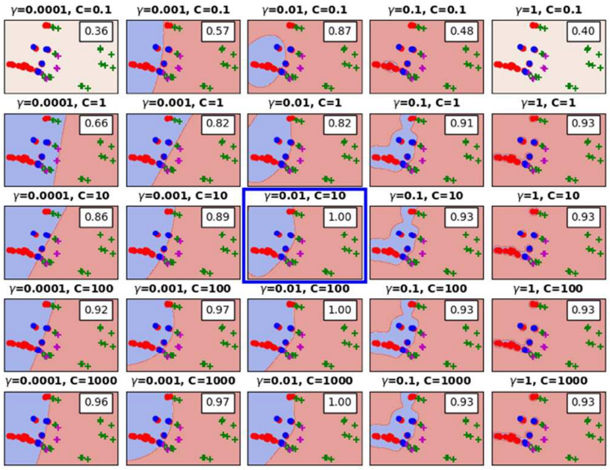

Figure 5. Tuning of RBF hyper-parameters (C and gamma) for CSP acetone interactions. Representative SVM classification results for one training and testing set show the effects of parameters C and g in the output of the RBF kernels. Red and blue circles represent the baseline samples in training and testing sets; green and purple plus symbols represent the positive signals in training and testing sets. The background blue and red region indicated the classifier decision surface, where all data fall into the red region are predicted as positive. Cross-validation scores are shown in the top right corner of each subplot. The optimal classifier zone is highlighted with a blue rectangle in the center of the image.

The Python code has a built-in function to optimize the hyper-parameters from data such as that shown in **Fig 5**. Based on this heat map (**Fig 6**), the optimum value of γ was 0.01, and the optimum value of C was 10. Using these parameters, the Python code is then modified (see details in step-by-step user guide) and the data is analyzed. Using the optimized kernel selection and hyper-parameters, the SVM demonstrated an accuracy of 95 ± 4% in cross validation and prediction of test samples.

View Article Online
DOI: 10.1039/C8AN00065D

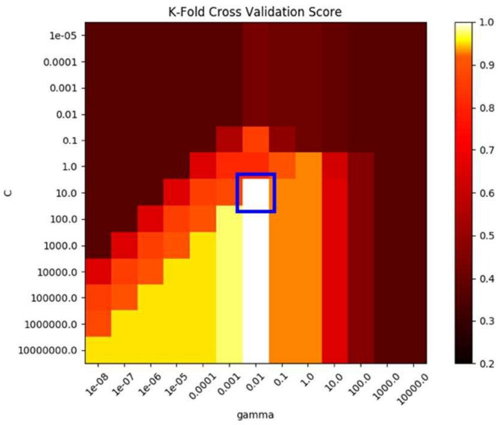

Figure 6. Heatmap of validation accuracy as a function of RBF parameters C and γ. The color indicates the cross-validation accuracy, where lighter colors represent a higher cross-validation score. Optimal parameters are highlight with a blue rectangle in the center

Decomposition of high dimensional EIS data to two-dimensional data with PCA is known to improve sensor/detector accuracy due to identification of uncorrelated variables from a large set of data[56]. PCA explains the maximum amount of data variance with the fewest number of principal components. For a semi-quantitative biosensor application such as the data in **Fig 1-6**, the use of only two principal components can lead to loss of useful information during data decomposition. However, for the RFP kernel with optimal tuning parameters the results were statistically significant at the 95% confidence interval. To further analyze the dataset, classifiers with 3 and 10 principal components were built and the cross-validation accuracy was improved (97 ± 3%), which is expected as less information was lost during decomposition (a 3D data representation for data analyzed with three principal components is shown in supplemental Fig S7). This result was expected, as use of reductionist clustering (i.e., using two-dimensional PCA) increases the risk of eliminating important outliers within the data. For example, over-clustering could result in important deviations from the "normal", for example in the case of silent ischemia[57]. In this case the data curation can be improved by analyzing polar coordinates in lieu of, or in addition to, Cartesian coordinates from impedimetric sensor data. However, analysis of classifiers with a dimension larger than two is computationally expensive, and can make use of mobile phone based analytical systems challenging. Care should be taken to discern as to whether the computational need outweighs the ability to analyze data on site using mobile equipment such as a tablet or mobile phone. To maintain focus on mobile-enabled diagnostic systems in this study, we used a two-dimensional PCA analysis, which is valid for semi-quantitative biosensor data where a regulatory or diagnostic metric is known (such as the case of DKA salivary biomarkers shown here).

Although not used here, computational speed and memory requirement can be improved by using more advanced computational tools such as the tensor compiler by Kjolstad et al[58]. This approach is particularly useful for multidimensional data analysis, and provides a generic mechanism that can generate code for compound tensor operations with sparse tensors, eliminating the need for writing optimized code for a specific problem. This tensor algebra compiler library represents an excellent next step forward to improve the work herein.

The SVM tool shown here is highly useful for point of need small molecule analysis using mobile detection and analysis systems (see supplemental Figure S8). Rapid triage analysis of breath disease state biomarkers is vital for triage analysis, and mobile phone solutions can bring this diagnosis to rural areas where health care is limited. Convergent technologies for triage diagnostics require systems-level solutions that are based on readily accessible hardware such as mobile phones or tablets [60].

Conclusions

Biosensors based on weak/transient interactions between small molecules and bioreceptors are a challenge for detection electronics, particularly in field studies or in analysis of complex matrices (e.g., body fluids, food, river water, etc.) using non-Faradaic impedimetric sensors. Support vector machine learning tools are facile *post hoc* analysis tools that do not require significant computational power and can be used for *in situ* analysis with mobile hardware such as a mobile phone or tablet. Here, we show use of a simple, open source machine learning algorithm for analysing such impedimetric data, and we show that the tool can be used for point of need applications.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors would like to thank the Agricultural & Biological Engineering PhD Fellowship (YR), the UF Opportunity Fund grant (ESM, NOK), the National Science Foundation Nanobiosensors program (ESM; Grant No. CBET-1511953, Nanobiosensing) and China Scholarship Council Fellowship (CS; No. 201303250074). The authors thank Dr. Diana Vanegas for assistance with the translation of the user's manual.

References

1. A. P. F. Turner, *Chemical Society reviews*, 2013, **42**, 3184-3196.
2. A. R. Demaio and J. Rockstroem, *Lancet*, 2015, **386**, E36-E37.
3. D. Z. Ji, L. Liu, S. Li, C. Chen, Y. L. Lu, J. J. Wu and Q. J. Liu, *Biosens Bioelectron*, 2017, **98**, 449-456.
4. L. Liu, D. M. Zhang, Q. Zhang, X. Chen, G. Xu, Y. L. Lu and Q. J. Liu, *Biosens Bioelectron*, 2017, **93**, 94-101.
5. D. M. Zhang, J. Jiang, J. Y. Chen, Q. Zhang, Y. L. Lu, Y. Yao, S. Li, G. L. Liu and Q. J. Liu, *Biosens Bioelectron*, 2015, **70**, 81-88.
6. A. Hayat and J. L. Marty, *Front Chem*, 2014, **2**.

View Article Online
DOI: 10.1039/C8AN00065D

7. D. Vanegas, C. Gomes and E. McLamore, *Biosens J*, 2016, **5**, 2.

8. A. Walcarius, S. D. Minteer, J. Wang, Y. Lin and A. Merkoçi, *J Mater Chem B*, 2013, **1**, 4878-4908.

9. A. Bonanni, A. H. Loo and M. Pumera, *TrAC Trends in Analytical Chemistry*, 2012, **37**, 12-21.

10. T. Yin and W. Qin, *TrAC Trends in Analytical Chemistry*, 2013, **51**, 79-86.

11. M. A. Daniele, M. Pedrero, S. Burrs, P. Chaturvedi, W. W. A. Wan Salim, F. Kuralay, S. Campuzano, E. McLamore, A. A. Cargill, S. Ding and J. C. Claussen, in *Nanobiosensors and Nanobioanalyses*, eds. M. d. C. Vestergaard, K. Kerman, I. M. Hsing and E. Tamiya, Springer Japan, Tokyo, 2015, DOI: 10.1007/978-4-431-55190-4_8, pp. 137-166.

12. C. Z. Zhu, G. H. Yang, H. Li, D. Du and Y. H. Lin, *Anal Chem*, 2015, **87**, 230-249.

13. E. S. McLamore, M. Convertino, I. Ocsoy, D. C. Vanegas, M. Taguchi, Y. Rong, C. Gomes, P. Chaturvedi and J. C. Claussen, in *Semiconductor-Based Sensors*, WORLD SCIENTIFIC, 2016, DOI: 10.1142/9789813146730_0002, pp. 35-67.

14. J. S. Daniels and N. Pourmand, *Electroanal*, 2007, **19**, 1239-1257.

15. E. B. Bahadir and M. K. Sezginturk, *Artif Cells Nanomed Biotechnol*, 2016, **44**, 248-262.

16. M. I. Prodromidis, *Electrochim Acta*, 2010, **55**, 4227-4233.

17. J.-G. Guan, Y.-Q. Miao and J.-R. Chen, *Biosensors and Bioelectronics*, 2004, **19**, 789-794.

18. R. Elshafey, A. C. Tavares, M. Siaj and M. Zourob, *Biosens Bioelectron*, 2013, **50**, 143-149.

19. D. D. Macdonald, *Electrochim Acta*, 2006, **51**, 1376-1388.

20. M. E. Orazem, P. Agarwal and L. H. Garciarubio, *J Electroanal Chem*, 1994, **378**, 51-62.

21. S. R. Das, Q. Nian, A. A. Cargill, J. A. Hondred, S. W. Ding, M. Saei, G. J. Cheng and J. C. Claussen, *Nanoscale*, 2016, **8**, 15870-15879.

22. W. J. Yuan, Y. Zhou, Y. R. Li, C. Li, H. L. Peng, J. Zhang, Z. F. Liu, L. M. Dai and G. Q. Shi, *Sci Rep-Uk*, 2013, **3**.

23. I. I. Suni, *TrAC Trends in Analytical Chemistry*, 2008, **27**, 604-611.

24. H. Song, Y. Wang, J. M. Rosano, B. Prabhakarpandian, C. Garson, K. Pant and E. Lai, *Lab Chip*, 2013, **13**, 2300-2310.

25. R. Kumar, A. P. Bhondekar, R. Kaur, S. Vig, A. Sharma and P. Kapur, *Sensor Actuat B-Chem*, 2012, **171**, 1046-1053.

26. Q. Dong, L. Du, L. Zhuang, R. Li, Q. Liu and P. Wang, *Biosens Bioelectron*, 2013, **49**, 263-269.

27. L. Lu, S. P. Deng, Z. W. Zhu and S. Y. Tian, *Food Analytical Methods*, 2015, **8**, 1893-1902.

28. Y. X. Dai, X. Wang, P. B. Zhang and W. H. Zhang, *Measurement*, 2017, **109**, 408-424.

29. X. Ding, Z. Lv, C. Zhang, X. Gao and B. Zhou, *IEEE Access*, 2017.

30. Z. Qin, B. Zhang, K. Gao, L. Zhuang, N. Hu and P. Wang, *Sensors and Actuators B: Chemical*, 2017, **239**, 746-753.

31. O.-P. Smolander, A. S. Ribeiro, O. Yli-Harja and M. Karp, *Sensors and Actuators B: Chemical*, 2009, **141**, 604-609.

32. E. Akbari, Z. Buntat, E. Shahraki, R. Parvaz and M. J. Kiani, *Journal of biomaterials applications*, 2016, **30**, 677-685.

33. F. F. Gonzalez-Navarro, M. Stilianova-Stoytcheva, L. Renteria-Gutierrez, L. A. Belanche-Muñoz, B. L. Flores-Rios and J. E. Ibarra-Esquer, *Sensors-Basel*, 2016, **16**, 1483.

34. Z. Q. Geng, S. S. Zhao, G. C. Tao and Y. M. Han, *Food Control*, 2017, **78**, 33-42.

35. S. De Vito, E. Esposito, M. Salvato, O. Popoola, F. Formisano, R. Jones and G. Di Francia, *Sensors and Actuators B: Chemical*, 2017, DOI: https://doi.org/10.1016/j.snb.2017.07.155.

36. O. Sadik, W. H. Land, A. K. Wanekaya, M. Uematsu, M. J. Embrechts, L. Wong, D. Leibensperger and A. Volykin, *Journal of chemical information and computer sciences*, 2004, **44**, 499-507.

37. T. Alizadeh and S. Zeynali, *Sensors and Actuators B: Chemical*, 2008, **129**, 412-423.

38. Y. Zuo, S. Chakrabartty, Z. Muhammad-Tahir, S. Pal and E. C. Alocilja, *Ieee Sens J*, 2006, **6**, 1644-1651.

39. R. G. Vogt, in *Comprehensive Molecular Insect Science*, Elsevier, Amsterdam, 2005, DOI: https://doi.org/10.1016/B0-44-451924-6/00047-8, pp. 753-803.

40. A. Sanchez-Gracia, F. G. Vieira and J. Rozas, *Heredity*, 2009, **103**, 208-216.

41. D. C. Vanegas, M. Taguchi, P. Chaturvedi, S. Burrs, M. Tan, H. Yamaguchi and E. S. McLamore, *Analyst*, 2014, **139**, 660-667.

42. S. L. Burrs, D. C. Vanegas, Y. Rong, M. Bhargava, N. Mechulan, P. Hendershot, H. Yamaguchi, C. Gomes and E. S. McLamore, *Analyst*, 2015, **140**, 1466-1476.

43. C. Song, A. Ortiz-Urquiza, S. H. Ying, J. X. Zhang and N. O. Keyhani, *Plos One*, 2015, **10**.

44. Y. Rong, J. Kieran-Lewis, N. O. Keyhani and E. S. McLamore, 2016.

45. S. Fujii, T. Maeda, I. Noge, Y. Kitagawa, K. Todoroki, K. Inoue, J. Z. Min and T. Toyo'oka, *Clin Chim Acta*, 2014, **430**, 140-144.

46. C.-W. Hsu, C.-C. Chang and C.-J. Lin, 2003.

47. T. Hastie, R. Tibshirani and J. Friedman, in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer New York, New York, NY, 2009, DOI: 10.1007/978-0-387-84858-7_7, pp. 219-259.

48. N. Halko, P. G. Martinsson and J. A. Tropp, *Siam Rev*, 2011, **53**, 217-288.

49. J. Bergstra and Y. Bengio, *J Mach Learn Res*, 2012, **13**, 281-305.

50. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J Mach Learn Res*, 2011, **12**, 2825-2830.

51. S. W. Ding, C. Mosher, X. Y. Lee, S. R. Das, A. A. Cargill, X. H. Tang, B. L. Chen, E. S. McLamore, C. Gomes, J. M. Hostetter and J. C. Claussen, *Acs Sensors*, 2017, **2**, 210-217.

52. I. M. Nooren and J. M. Thornton, *Journal of molecular biology*, 2003, **325**, 991-1018.

53. Q. J. Liu, H. Wang, H. L. Li, J. Zhang, S. L. Zhuang, F. N. Zhang, K. J. Hsia and P. Wang, *Biosens Bioelectron*, 2013, **40**, 174-179.

54. M. Liu, M. Wang, J. Wang and D. Li, *Sensors and Actuators B: Chemical*, 2013, **177**, 970-980.

55. T. Hastie, R. Tibshirani and J. Friedman, in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer New York, New York, NY, 2009, DOI: 10.1007/978-0-387-84858-7_2, pp. 9-41.

56. H. Abdi and L. J. Williams, *Wiley interdisciplinary reviews: computational statistics*, 2010, **2**, 433-459.

57. P. F. Cohn, K. M. Fox and C. Daly, *Circulation*, 2003, **108**, 1263-1277.

View Article Online
DOI: 10.1039/C8AN00065D

58. F. Kjolstad, S. Kamil, S. Chou, D. Lugato and S. Amarasinghe, 2017.

59. R. Ganesh, N. Suresh and J. Ramesh, *The National medical journal of India*, 2006, **19**, 155-158.

60. R. Chiu, C. Ho, S. Tong, K. Ng and C. Lam, *Hong Kong medical journal= Xianggang yi xue za zhi/Hong Kong Academy of Medicine*, 2002, **8**, 172-176.