Title
Computational Sensing in Biophotonics and Diagnostics

Author
Ballard, Zachary

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Computational Sensing in Biophotonics and Diagnostics

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Electrical and Computer Engineering

by

Zachary Scott Ballard

2019

ABSTRACT OF THE DISSERTATION

Computational Sensing in Biophotonics and Diagnostics

by

Zachary Scott Ballard

Doctor of Philosophy in Electrical and Computer Engineering

University of California, Los Angeles, 2019

Professor Aydogan Ozcan, Co-Chair

Professor Asad Mohamed Madni, Co-Chair

Over the past decades the dramatic increase in computational resources coupled with the advent of machine learning and artificial intelligence has profoundly impacted sensor technology. Together, these trends present a new opportunity for data-driven computational sensor design, where acquisition hardware is fundamentally changed to "lock-in" to the optimal sensing data with respect to a user defined cost function. This dissertation addresses this emerging opportunity, referred to herein as computational sensing, with a focus on applications in biophotonics and diagnostics where there is currently high-demand for cost-effective and
compact sensing systems for democratizing global health. Therefore, several examples of computational sensing systems will be presented, all of which leverage machine learning and data-

driven design principles to enable low-cost, compact, and sparse implementations well-suited for the point-of-care among other settings.

Firstly, I will discuss a multiplexed paper-based vertical flow assay and reader for biomarker quantification at the point-of-care. I will show how deep learning and statistical analysis of the multiplexed sensor response can be leveraged to improve accuracy, dynamic range, and limit-of-detection, greatly enhancing the sensing capabilities beyond that of the traditionally employed lateral flow assay. Additionally, I will demonstrate feature selection techniques which can serve as a toolbox for iterative assay development, sensor design, and cost-performance optimization. Two applications in point-of-care diagnostics will be presented with this platform, one being a clinical study for early-stage Lyme Disease diagnoses, and the other, a clinical study for cardio-vascular risk assessment by the high-sensitivity C-Reactive Protein test. Secondly, Localized plasmon resonance (LSPR) sensors and their corresponding readers will be discussed as computational sensing systems for label-free biomolecular sensing. Through machine learning of spectral characteristics, I will present how the corresponding LSPR sensor reader can be jointly designed with the LSPR sensor to optimize read-out accuracy while minimizing cost by computationally selecting from cost effective illumination optics. Lastly, a fundamentally different hardware design for spectroscopy, enabled by deep learning, will be demonstrated and discussed for application-specific performance advantages in terms of acquisition rate, form factor, and cost.

The dissertation of Zachary Scott Ballard is approved.

Dino Di Carlo

Sam Emaminejad

Oscar M. Stafsudd

Asad Mohamed Madni, Committee Co-Chair

Aydogan Ozcan, Committee Co-Chair

University of California, Los Angeles

2019

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would first like to thank my advisor, Professor Aydogan Ozcan. Professor Ozcan granted me the opportunity to pursue research at the highest-caliber, to chase curiosities while maintaining an overall focus on technology development that can have real-world impact. My PhD studies have been an incredible privilege, and I am deeply grateful for Professor Ozcan's patience, passion, guidance, care, and commitment to scholarship. I would also like to thank Professor Dino DiCarlo and Professor Omai Garner for their teaching, thoughtfulness, and the incredible opportunity to collaborate across disciplines, which cultivated much of the work forming my dissertation. I also want to give special thanks to my co-advisor, Professor Asad Madni, for his encouragement, guidance, and perspective during the last two years of my PhD along with the other members of my PhD committee: Professor Oscar Staffsud and Professor Sam Emaminejad.

I have also had the immense joy of working with a rich community of brilliant scientists from all parts of the world. I have learned immensely from their intellect, acumen, and willingness to work together to solve problems. I owe a special thanks to Dr. Hyou-Arm Joung, who worked diligently and with stellar optimism to bring his vision of a new point-of-care technology into reality. The development of this sensing platform led to a fruitful collaboration, forming two chapters of this dissertation. Of my many supportive colleagues, I would also like to thank Professor Gerard Cote, Professor Kristen Maitlan, Dr. Zoltan Gorocs, Dr. Yair Rivenson, Dr. Jing Wu, Dr. Rong Zhen, Dr. Hatice Koydemir, Derek Tseng, Calvin Brown, Artem Goncharov, Deniz Mengu, Dr. Mo Sun, Chuanzhen Zhao, Dr. Yibo Zhang, Dr. Aniruddha Ray, Dr. Vittorio Bianco, Dr. Mustafa Daloglu, Dr. Daniel Shir, Jesse Liang, Dr. Yichen Wu, Muhammed Veli, Kevin DeHann, Hongda Wang, Professor Euan McLeod, Dr. Lucas Shaw, Eric Welch, Dr. Kahlen Ouyang, Paul Gordon, Ming Lei, Mason Fordham, Ashley Clemens, Farhana Haque, Sarah

Bazargan, Diane Jung, Aashish Bhardwaj, Katie Barajas, Karina Nugroho, Hailie Teshome, Mackenzie Lucero, and Shyama Sathianathan.

I must also express my deepest gratitude and love to my family for their unwavering support: to my wife to be, Natali Senocak for her daily dose of understanding, kindness, and encouragement, and to my parents and brother to which I owe everything.

I would like the thank American Chemical Society, for allowing the permission to include a previous publication into a chapter. Lastly, I would like to acknowledge the National Science Foundation and Society for Photo-optical Engineers for their support as a Graduate Research Fellow, as well as the PATHSUP research center.

Vita

Zachary S. Ballard received his B.S. degree in Engineering Physics from Brown University in 2013. After graduating he was a visiting researcher at the Max Planck Institute for the Science of Light in Erlangen, Germany where he worked with free-space coupled micro-resonators for bio-sensing applications. He is currently a NSF fellow and PhD candidate in the Electrical and Computer Engineering Department at University of California, Los Angeles. He is also a Fellow in the NSF Emerging Research Center for Precise Advanced Technologies and Health Systems for Underserved Populations (PAHTS-UP). His work aims to develop computational sensing platforms for biophotonics and point-of-care diagnostic applications.

During his education at UCLA, Zach received the 2015 Dr. Ursula Mandel Scholarship for graduate students in scientific fields allied to medicine, the 2017 Podium Presentation award at the UC System-wide Bioengineering Symposium, 3rd place for Oral Presentation at the 2018 Emerging Researchers National Conference, the 2018 SPIE Optics and Photonics Education Scholarship, and the Audience Choice Award at the 2018 UCLA Grad Slam. Zach has also co-authored 15 publications in peer-reviewed journals and international conferences as well as a one book chapter.

Selected Publications

1. Z. S. Ballard and A. Ozcan, "Wearable Optical Sensors," in *Mobile Health*, Springer, Cham, 2017, pp. 313–342.
2. Z. S. Ballard, D. Shir, A. Bhardwaj, S. Bazargan, S. Sathianathan, and A. Ozcan, "Computational Sensing Using Low-Cost and Mobile Plasmonic Readers Designed by Machine Learning," *ACS Nano*, vol. 11, no. 2, pp. 2266–2274, Feb. 2017.
3. Z. S. Ballard *et al.*, "Deep Learning-Enabled Point-of-Care Sensing Using Multiplexed Paper-Based Sensors," *bioRxiv*, p. 667436, Jun. 2019

Chapter 1. Introduction to computational sensing and machine learning

Computational sensing, as an analogue to computational imaging, is defined by the joint engineering of sensing system hardware and software at all levels of operation, from acquisition to the reporting of results. In this dissertation, I report several computational sensing approaches for applications in bio-photonics and point-of-care diagnostics, with each chapter being a seperate contained study.

After a brief discussion on computational sensing as a whole in Chapter 1, Chapter 2 begins by proposing an emerging point-of-care sensing technology, multiplexed paper-based immunoassays as a computational sensing platform, detailing a proof-of-concept clinical study performed for a common cardiac health assessment test. Chapter 3 further expands upon the potential of this paper-based computational point-of-care sensor, by introducing a multi-antigen test and deep-learning diagnostic algorithm for early-stage Lyme Disease. Both chapters showcase the benefits to cost and performance resulting from the computational sensing approach, and discuss how machine learning, specifically neural networks, can be best harnessed to combat real-world sources of noise inherent in biosensing.

Chapter 4 explores computational design of the read-out instrumentation for emerging plasmonic sensing applications, where the final hardware contains computationally selected illumination bands. Chapter 4 shows how data-driven analysis of low-cost plasmonic sensors fabricated across multiple batches, can produce more accurate read-out instrumentation that can learn statistical variation inherent in the fabrication tolerances of the corresponding sensors.

Lastly, Chapter 5 proposes a fundamentally different spectrometer design enabled by deep learning. Utilizing the low-cost fabrication of plasmonic filters discussed in chapter 4, here I

demonstrate an encoder-based approach to optical spectroscopy, where deep learning is used for spectral reconstruction from the encoded features.

1.1 Background

Computation has profoundly shaped the way we approach sensing. In the realm of bio-sensing, for example, signals are acquired, often at high cost, with various sources of noise including the stochastic behavior of molecular interactions, imperfections in fabrication, chemical and/or optical signal transduction mechanisms, and human variation in terms of sample handling as well as physiological differences and natural variations inherent in large populations under test. With such noisy and sparse sensing landscapes, computational methods have evolved helping us garner meaningful information from raw sensor data. Naturally, as Moore's law has progressed, and computation has become more powerful, cheaper, and widely accessible, it has in many ways handled an even larger share of the noise burden when compared to the sensing hardware itself. For example, Support Vector Machine (SVM) based algorithms have more and more been employed in numerous sensing fields such as remote hyperspectral imaging for geological and environmental mapping and classification, cross-reactive sensor arrays for e.g. 'electronic/optoelectronic nose' sensors for identification of trace amounts of explosives and toxins among other analytes, as well as for diagnostics and genomics applications, including pattern recognition of biological pathways for cancer and disease prediction [1]–[3]. More recently, deep learning and the implementation of Neural Networks (NNs) have shown immense promise in computer vision and image processing, beyond the capabilities of traditional machine learning approaches, inferring complex nonlinear patterns in high-dimensional data [2], [4], [5]. Furthermore, the feed-forward nature of NNs provide a major advantage in computational speed

compared other traditional signal recovery approaches based on e.g., compressive sampling, and can be readily integrated into common processors on mobile phones and tablet PCs, paving the way for cost-effective, mobile and yet powerful deep-learning enabled sensing and diagnostic systems [6], [7].

Despite all this progress in the field of sensing at large, there is an important opportunity that has not been extensively explored yet: *computation and machine learning methods can fundamentally change the hardware designs of traditional sensors and can be used to holistically design intelligent sensor systems*. In such designs, computation and statistical learning tools need to be utilized at the design phase of the sensing instrument, carefully taking into account various sources of noise inherent in a given application of interest in order to design a smart read-out instrument, also impacting the signal generation and decoding schemes that are employed. In many of the examples listed earlier, this is not the case: the sensing framework has been designed based on a "sequential" merger of the hardware output and computational analysis applied on this output, following its acquisition with a traditional sensor hardware. This is a natural result of the analog to digital transition, where conventional sensor designs were, later in the sensing scheme, empowered by computers and computational analysis. Though this is an exciting and extensive field of sensing it falls outside the scope of this Dissertation. *Here I specifically focus on an emerging opportunity, namely, "**computational sensing systems**" that merge computation and machine learning based statistical analysis of signals as a fundamental part of their hardware design to optimize sensing performance based on the full integration of computational processing, feature optimization and statistical learning at the hardware level.* These systems that are of interest for my Dissertation try to answer the following question: if today's powerful computers, computational resources and algorithms had existed before a well-known traditional sensor system

or instrument was designed, how would it have fundamentally changed and improved the sensing system?

One can ask the same question for e.g., imaging systems and microscopes/nanoscopes, the answers of which fall under the category of computational imaging field. While computational imaging field will be left outside the scope of this Dissertation in general, one exemplary answer to the former question that I posed is given by the design of a single pixel camera/imager [8]. This single pixel imaging system that is based on a compressive sampling framework, has entirely changed the design of a traditional imaging system based on the merger of computation at the hardware level, drastically simplifying the imaging hardware, in addition to efficiently releasing the requirement for a focal-plane array in an imager, which is especially important for designing cameras that operate at longer wavelengths. There are other examples from computational imaging field highlighting this emerging opportunity to use computational techniques and statistical learning for designing intelligent imaging systems; however in the in the following chapters, I will specifically focus on how the same underlying principles can manifest a transformation in the design and operation principles of computational sensing systems.

1.2 Discussion on computational sensing systems

Computational sensing, in some ways is analogous to 'inverse' design methodology. With inverse design, a desired performance 'target' is first defined, then computation is utilized to converge to a given set or subset of parameters, which most accurately yield this target. Such a computational design methodology can drastically improve the overall performance of a given system, through the implementation of locally optimal, but perhaps *non-intuitive* design choices. Additionally, inverse design can be used to rapidly provide a blueprint for a dedicated device that

can out-perform, sometimes at a lower cost, a more general purpose device design that has perhaps been modified to execute a given sensing task. Examples of inverse design can be found in mechanics, antenna design, and nanophotonics, among other engineering disciplines. For example, wavelength demultiplexers for telecommunications applications and recently metasurface gratings have been shown to benefit from topological optimization methods in-line with the inverse design methodology [9], [10]. This approach has important implications for designing sensing hardware and sub-components, specifically for nanophotonic sensors, where the electric field patterns are complex and delicate functions of 3D device geometry, fabrication artifacts and tolerances, interacting wave properties and material dispersion, among others. All these variables need to be taken account, along with their statistical variation, in the design of intelligent sensor hardware and signal decoding strategies.

One powerful mathematical framework that can be used for this purpose is compressive sensing (also known as compressive sampling). The general goal of compressive sensing is to 'encode' a given signal, x, with a sparse sampling operator, θ. By solving an under-determined matrix equation representing this sparse sampling/sensing operation, i.e., $y = \theta x$, the original signal (x) can be reconstructed from a fewer number of measurements (compared to what is dictated by the Nyquist sampling theorem) assuming that the signal can be represented as a sparse vector in some mathematical domain, such as the wavelet domain [11]. The field of spectroscopy exemplifies how this compressive signal recovery framework can transform traditional grating and line-scan CCD based spectrum analyzer designs into much more compact computational spectroscopy tools, which use e.g., Liquid Crystal Displays (LCDs), and even disordered diffractive nano-structures to encode a given spectral signal and then computationally reconstruct

it with a minimum number of measurements [12], [13]. This approach saves measurement time while bringing in compactness advantages to the final instrument design.

These previously mentioned frameworks showcase powerful examples of how new sensing hardware, with clear engineering advantages, can be enabled by computation; however they are also limited by their inability to *learn and properly take into account* characteristic noise inherent in real-world sensing systems. Therefore, other computational sensing approaches based on statistical learning (and data-centric training) must be evoked to tackle this noise problem [14]. Such learning frameworks rely on large, statistically significant sets of sensing data, often acquired by 'gold-standard' sensing systems used in the training phase of the system design. This acquired training data set is then mapped to a verified target, and a mathematical cost function, $J(\boldsymbol{y}, \boldsymbol{y}')$, is defined based on the difference between the verified target (gold standard, $\boldsymbol{y}$) and the algorithmic output of the sensor that is being explored and designed, i.e. $\boldsymbol{y}'$. This cost function can be iteratively minimized while the design of the sensor evolves based on the target criteria, budget and performance expectations. Statistical learning and modern machine learning approaches, including deep learning techniques, provide powerful tools for this purpose, helping design intelligent sensor systems that learn how to digitally separate out noise features from the target sensing signal by using large numbers of well-characterized training data.

While quite powerful in general, one has to be careful when using such a data-driven instrument or sensor design approach: sometimes, the high dimensional space of training data might drown-out the meaningful correlations to the target sensing information. This phenomenon, which is also referred to as 'the curse of dimensionality', presents design engineers with an ultimatum to either acquire exponentially larger training data sets, or to reduce the dimensionality of their sensing systems. Machine learning inspired sensor design, therefore, attempts to

systematically prioritize or select potential measurement features, in terms of their statistical values/contributions for accurately predicting the desired sensor output. This process, known as 'feature selection' can be thought of as an '*elite democracy*' of measurement features, which computational sensing systems can use as optimal building blocks in subsequent iterations of the sensing hardware, fully realizing the merger of computation, statistical learning and sensor hardware and read-out design. Optimization and engineering of this feature selection process (determining the voting power of individuals in an elite democracy, following the same analogy) can benefit sensing systems in a myriad of ways: by reducing the complexity, cost, bulkiness and weight of the sensing instrument, while also mitigating various noise sources as well as the data acquisition burden, which is increasingly becoming an issue with the proliferation of high-throughput sensor systems driven by the internet of things (IoT) and the related 'big data' paradigm [15].

Some of these emerging ideas have already started to be used in the field of spectroscopy, by incorporating feature selection as a computational sensing tool. Several learning algorithms such as Genetic Algorithms (GAs), SVMs, as well as L-1 norm based regularization methods, have been investigated as means with which to select an optimal subset of spectral bands from a larger hyperspectral data set for remote sensing applications, empowering classification algorithms as well as subsequent designs for targeted applications [2], [16], [17]. Stated differently, data-driven designs can in general outperform intuitive designs, especially if well-characterized training data are available to engineer and select features that can statistically separate out various noise terms from the target signal of interest for the sensor. Such frameworks can naturally be extended to other spectroscopy applications including fluorescence-based sensing systems, comprised of e.g., a complex mixture of spectrally overlapping, multiplexed exogenous fluorophores, as well as

endogenous auto-fluorescent or absorbing media for sub-dermal, noninvasive and wearable bio-sensing applications, among others [18].

It is important to note that feature selection is not solely a computational approach used for improving sensor component performance. In fact, utilizing a subset of the physically measurable features can sometimes, especially in low-dimensional feature spaces, lead to poorer sensor performance due to the information discarded during the exclusive selection process. However, performance trade-offs such as this are inherent in most engineering applications, and should be considered on a case-by-case basis, ultimately converging to design choices that embody the most appropriate sensing technology given a set of performance and budget constraints of the target application. Furthermore, the above described data-driven design methods can quantify such trade-offs in performance and help optimize the engineering of the sensor system. For example, some sensing systems, especially those used for environmental monitoring, are much more powerful and useful, provided they can operate in a widely distributed format. Therefore, the added benefit of the expanded spatio-temporal data collection capability in such a distributed sensing network might practically outweigh the decrease in performance for an individual sensing node. Cost-sensitive feature selection implementations can therefore exist as a systematic way of making such design choices, given an appropriately defined threshold of needed resolution, sampling rate, power requirement, etc. Furthermore, constraining systems to be constructed from existing commercial electronics (such as e.g., mobile phones, low cost LEDs) can benefit from the economies of scale and greatly reduce costs, improve accessibility and throughput of these sensing systems, broadly benefiting the overall sensing goal [19]–[21].

Also of strong interest are computational sensing systems which can re-configure on demand e.g., their sensing architecture and/or path to best suite a specific application. Such dynamic and

intelligent sensor designs can be applied to environmental monitoring with distributed sensing networks that can autonomously decide where to sample and how to sense (e.g., which modality to use), whether they are doing so under a supervised learning framework with a concrete sensing goal (e.g., a well-defined cost function to map hydrocarbons and other pollutants in the air or ocean after anthropogenic disturbances), or in an unsupervised capacity for general discovery or surveillance [25]. Similarly, wearable sensors can greatly benefit from reconfigurable computational sensing designs as a means to optimize signal acquisition for different body types, motion artifacts, and health states [26]. Specifically, optical or bio-impedance based sensor arrays for e.g., blood pressure monitoring could quickly be computationally reconfigured, by optimizing the relative weights of different signals within the array to converge to a reliable and accurate read-out for photoplethysmogram (PPG) or electro-cardio gram (ECG) signals. Additionally, systems for outpatient monitoring, for example, can incorporate dynamic configurability for extending the monitoring life-time by autonomously adjusting the sampling rate, or activating key sensor functions only in instances where the patient is deviating from a nominal or pre-defined health state. Furthermore, such computational sensing systems, if connected in a widely distributed and cost-effective manner, as part of an IoT network, will have the major advantage of collectively learning 'on-line' from evidence based sensing outcomes, thereby solving and converging to sensing solutions otherwise intractable with a single sensing unit [27], [28].

References

[1] K. P. Soh, E. Szczurek, T. Sakoparnig, and N. Beerenwinkel, "Predicting cancer type from tumour DNA signatures," *Genome Med.*, vol. 9, Nov. 2017.

[2] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep Learning-Based Classification of Hyperspectral Data," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[3] "An optoelectronic nose for identification of explosives - Chemical Science (RSC Publishing)." [Online]. Available: http://pubs.rsc.org/en/content/articlelanding/2016/sc/c5sc02632f#!divAbstract. [Accessed: 18-Dec-2017].

[4] Y. Rivenson, Z. Göröcs, H. Günaydin, Y. Zhang, H. Wang, and A. Ozcan, "Deep learning microscopy," *Optica*, vol. 4, no. 11, pp. 1437–1443, Nov. 2017.

[5] Y. Rivenson, Y. Zhang, H. Gunaydin, D. Teng, and A. Ozcan, "Phase recovery and holographic image reconstruction using deep learning in neural networks," *Light Sci. Appl.*, vol. Accepted Article Preview, May 2017.

[6] N. D. Lane and P. Georgiev, "Can Deep Learning Revolutionize Mobile Sensing?," in *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, New York, NY, USA, 2015, pp. 117–122.

[7] Y. Rivenson *et al.*, "Deep learning enhanced mobile-phone microscopy," *ArXiv171204139 Phys.*, Dec. 2017.

[8] "Single-pixel imaging via compressive sampling - IEEE Journals & Magazine." [Online]. Available: http://ieeexplore.ieee.org/abstract/document/4472247/. [Accessed: 17-Dec-2017].

[9] A. Y. Piggott, J. Lu, K. G. Lagoudakis, J. Petykiewicz, T. M. Babinec, and J. Vučković, "Inverse design and demonstration of a compact and broadband on-chip wavelength demultiplexer," *Nat. Photonics*, vol. 9, no. 6, p. 374, Jun. 2015.

[10] J. Yang and J. A. Fan, "Topology-optimized metasurfaces: impact of initial geometric layout," *Opt. Lett.*, vol. 42, no. 16, pp. 3161–3164, Aug. 2017.

[11] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.

[12] R. French, S. Gigan, and O. L. Muskens, "Speckle-based hyperspectral imaging combining multiple scattering and compressive sensing in nanowire mats," *Opt. Lett.*, vol. 42, no. 9, pp. 1820–1823, May 2017.

[13] Y. Oiknine, I. August, D. G. Blumberg, and A. Stern, "Compressive sensing resonator spectroscopy," *Opt. Lett.*, vol. 42, no. 1, pp. 25–28, Jan. 2017.

[14] A. Adler, M. Elad, and M. Zibulevsky, "Compressed Learning: A Deep Neural Network Approach," *ArXiv161009615 Cs*, Oct. 2016.

[15] A. Abdi and F. Fekri, "Optimal sensor selection in the presence of noise and interference," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 2378–2382.

[16] M. Pal, "Hybrid genetic algorithm for feature selection with hyperspectral data," *Remote Sens. Lett.*, vol. 4, no. 7, pp. 619–628, Jul. 2013.

[17] Z. S. Ballard, D. Shir, A. Bhardwaj, S. Bazargan, S. Sathianathan, and A. Ozcan, "Computational Sensing Using Low-Cost and Mobile Plasmonic Readers Designed by Machine Learning," *ACS Nano*, vol. 11, no. 2, pp. 2266–2274, Feb. 2017.

[18] Z. Göröcs *et al.*, "Quantitative Fluorescence Sensing Through Highly Autofluorescent, Scattering, and Absorbing Media Using Mobile Microscopy," *ACS Nano*, vol. 10, no. 9, pp. 8989–8999, Sep. 2016.

[19] U. Jensen, P. Kugler, M. Ring, and B. M. Eskofier, "Approaching the accuracy–cost conflict in embedded classification system design," *Pattern Anal. Appl.*, vol. 19, no. 3, pp. 839–855, Aug. 2016.

[20] J. Bian, X. Peng, Y. Wang, and H. Zhang, "An Efficient Cost-Sensitive Feature Selection Using Chaos Genetic Algorithm for Class Imbalance Problem," *Mathematical Problems in Engineering*, 2016. [Online]. Available:

https://www.hindawi.com/journals/mpe/2016/8752181/abs/. [Accessed: 14-Dec-2017].

[21] A. Ozcan, "Mobile phones democratize and cultivate next-generation imaging, diagnostics and measurement tools," *Lab. Chip*, vol. 14, no. 17, pp. 3187–3194, Sep. 2014.

[22] P.-C. Hsieh, H.-T. Lin, W.-Y. Chen, J. J. P. Tsai, and W.-P. Hu, "The Combination of Computational and Biosensing Technologies for Selecting Aptamer against Prostate Specific Antigen," *BioMed Research International*, 2017. [Online]. Available: https://www.hindawi.com/journals/bmri/2017/5041683/abs/. [Accessed: 15-Dec-2017].

[23] M. Netzer *et al.*, "Profiling the human response to physical exercise: a computational strategy for the identification and kinetic analysis of metabolic biomarkers," *J. Clin. Bioinforma.*, vol. 1, p. 34, Dec. 2011.

[24] F. F. Gonzalez-Navarro, M. Stilianova-Stoytcheva, L. Renteria-Gutierrez, L. A. Belanche-Muñoz, B. L. Flores-Rios, and J. E. Ibarra-Esquer, "Glucose Oxidase Biosensor Modeling and Predictors Optimization by Machine Learning Methods," *Sensors*, vol. 16, no. 11, p. 1483, Oct. 2016.

[25] B. Brunton, S. Brunton, J. Proctor, and J. Kutz, "Sparse Sensor Placement Optimization for Classification," *SIAM J. Appl. Math.*, vol. 76, no. 5, pp. 2099–2122, Jan. 2016.

[26] R. Saeedi, H. Ghasemzadeh, and A. H. Gebremedhin, "Transfer learning algorithms for autonomous reconfiguration of wearable systems," in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 563–569.

[27] Y. Zuo, S. Chakrabartty, Z. Muhammad-Tahir, S. Pal, and E. C. Alocilja, "Spatio-temporal processing for multichannel biosensors using support vector machines," *Ieee Sens. J.*, vol. 6, no. 6, pp. 1644–1651, Dec. 2006.

[28] M. P. McRae, G. Simmons, J. Wong, and J. T. McDevitt, "Programmable Bio-nanochip Platform: A Point-of-Care Biosensor System with the Capacity To Learn," *Acc. Chem. Res.*, vol. 49, no. 7, pp. 1359–1368, Jul. 2016.

Chapter 2. Deep learning-enabled point-of-care sensing using multiplexed paper-based sensors

This chapter introduces a new paper-based point-of-care sensor which utilizes traditional colorimetric based immunochemistry, but is uniquely well-equipped as a computational sensor due to its multiplexed, vertical flow format. This chapter shows how multiplexed sensors can be iteratively developed through data-driven computational design to improve performance, expand the dynamic range, and lower the cost per-test. As a use-case, we demonstrate a low-cost and rapid paper-based vertical flow assay (VFA) for high sensitivity C-Reactive Protein (hsCRP) testing, commonly used for assessing risk of cardio-vascular disease (CVD). A machine learning-based framework was developed to (1) determine an optimal configuration of immunoreaction spots and conditions, spatially-multiplexed on a sensing membrane, and (2) to accurately infer target analyte concentration. Using a custom-designed handheld VFA reader, a clinical study with 85 human samples showed a competitive coefficient of variation of 11.2% and linearity of $R^2 = 0.95$ among blindly-tested VFAs in the hsCRP range (i.e., 0-10 mg/L). We also demonstrate a mitigation of the hook-effect due to the multiplexed immunoreactions on the sensing membrane. This paper-based computational VFA could expand access to CVD testing, and the presented framework can be broadly used to design cost-effective and mobile sensors.

This chapter is part of a manuscript which is currently under review, and is currently on-line in pre-print form: Z. Ballard, H.A. Joung, A. Goncharov, J. Liang, K. Nugroho, D. Di Carlo, O. B. Garner, A. Ozcan, "Deep Learning-Enabled Point-of-Care Sensing Using Multiplexed Paper-Based Sensors," *bioRxiv*, p. 667436, Jun. 2019. (doi: http://dx.doi.org/10.1101/667436.)

2.1 Introduction

Computation has great potential for improving diagnostics. By identifying complex and nonlinear patterns from noisy inputs, computational tools present an opportunity for automated and robust inference of medical data. For example, several studies have shown deep learning as a method to automatically identify tumors from an image, potentially enabling diagnostics in low-resource settings that lack a trained diagnostician [1]–[3]. Additionally, computational solutions have been demonstrated earlier in the diagnostics pipeline to virtually stain pathology slides and enhance image resolution through the use of convolutional neural networks [4]–[6]. Though much of this recent success is within the field of imaging, diagnostics that rely on biosensing can similarly leverage computational tools to improve sensing results and design future systems.

Point-of-care (POC) testing can especially benefit from computational sensing approaches. Due to their low-cost materials, compact designs, and requirement for rapid and user-friendly operation, POC tests are often less accurate when compared to traditional laboratory tests and assays [7]–[12]. For example, paper-based immuno-assays such as rapid diagnostic tests (RDTs) offer an affordable and user-friendly class of POC tests which have been developed for malaria, HIV-1/2, and cancer screening, among other uses [13]–[17]. However, these RDTs lack the sensitivity and specificity needed for certain diagnostic applications largely due to issues of reagent stability, fabrication and operational variability, as well as matrix effects present in complex samples such as blood [15], [16], [18]. Additionally, a well-known competitive binding phenomenon called the hook-effect can lead to false reporting of results, specifically in instances where the sensing analyte can be present over a large dynamic range [19]–[24]. Therefore, computational tools alongside portable and cost-effective assay readers present a unique opportunity to compensate for some of these constraints [25]–[31]. By quantifying the signals

generated on paper-based substrates, machine learning algorithms have the potential to significantly improve the performance of POC sensors, without a significant hardware cost or increased complexity to the assay protocol.

As a unique demonstration of this emerging opportunity at the intersection of computational sensing and machine learning, here we report a computational paper-based vertical flow assay (VFA) for cost-effective high-sensitivity C-Reactive Protein (hsCRP) testing, also referred to as cardiac CRP testing (cCRP) [32]. This low-cost and rapid (< 12 min) VFA uses a multiplexed sensing membrane and diagnostic algorithm based on neural networks to accurately quantify CRP concentration in the high-sensitivity range (i.e. 0-10 mg/L), as well as to identify samples outside of this range despite the presence of the hook-effect.

CRP is a general biomarker of inflammation, however slightly elevated CRP levels in blood can be an indicator of atherosclerosis, and have been shown to be a predictor for heart attacks, stroke, and sudden cardiac death for patients with and without a history of CVD [33]–[36]. Therefore, the hsCRP test is a quantitative test commonly ordered by cardiologists to stratify certain patients into low, intermediate, and high risk groups for CVD based off of clinically defined cut-offs: below 1 mg/L is considered low risk, between 1 and 3 mg/L is intermediate risk, and above 3 mg/L is high-risk [37]. As a result, the hsCRP test requires a high degree of accuracy and precision, especially around the clinical cut offs, putting it out-of-reach of traditional paper-based systems [38]. Additionally, in the presence of infection, tissue injury, or other acute inflammatory events, CRP levels can rise nearly three orders of magnitude, making hsCRP testing with immuno- and nephelometric- assays vulnerable to the hook-effect [34], [36], [39]. As a result, samples with greatly elevated CRP levels can be falsely reported as within the hsCRP range (i.e., < 10 mg/mL), and therefore wrongly interpreted for CVD risk stratification.

To address these existing challenges of POC hsCRP testing, in this work we implemented a computational VFA-based sensing framework to jointly develop the CRP quantification algorithm and multiplexed sensing membrane configuration, computationally selecting the most robust subset of sensing channels with which we accurately infer the CRP concentration. We performed a clinical study with 85 patient serum samples and >250 VFA tests created over multiple fabrication batches, and compared the sensor performance to an FDA-approved assay and nephelometric reader (Dimension Vista System, Siemens). Our blind testing results yielded an average coefficient of variation (CV) of 11.2% and a coefficient of determination (R^2) of 0.95 over an analytical measurement range of 0 mg/L to 10 mg/L. It is important to note that although there is no FDA-approved POC hsCRP sensor, various systems have been demonstrated in the literature[40]–[44]. However, the tests which report accurate quantification in the high-sensitivity range employ fluorescent-based chemical assays and benchtop readers to overcome the performance limits of their traditional colorimetric counterparts. In contrast, this work uniquely demonstrates a new data-driven sensor design and read-out framework, powered by deep learning, for improving POC testing. We applied this machine learning-enabled sensing framework to a colorimetric paper-based multiplexed test for quantification of hsCRP as a use-case, and demonstrated its competitive quantitative performance using a mobile reader, without the need for more advanced and sensitive molecular assays and their corresponding benchtop read-out systems.

We believe that the presented POC hsCRP sensor platform could provide a rapid and cost-effective means to obtain valuable diagnostic and prognostic information for CVD, expanding access to actionable health information, especially for at-risk populations that often go underserved (34,35). Broadly, our results also highlight computational sensing as an emerging opportunity for iterative assay and sensor development. Given a training data set, machine learning-based feature

selection algorithms can be implemented to determine the most robust sensing channels for a given multiplexed system such as protein micro-array, well-plate assay, or multi-channel fluidic device, among others. This can therefore lead to optimized and cost-effective implementations of multiplexed bio-sensing systems for future POC diagnostic applications.

2.2 Materials and methods

Multiplexed VFA

Overview

The multiplexed VFA platform is comprised of functional paper layers stacked within a 3D-printed plastic cassette. These layers contain different paper materials and wax printed structures which have been optimized to support uniform vertical flow of serum across a two-dimensional nitrocellulose sensing membrane (Fig. 2.1a, Table S1). Similar to conventional paper-based immunoassays, the VFA works by immobilizing a target analyte onto a paper substrate through binding to a complimentary capture antigen or antibody previously adsorbed within the porous structure [45]. Gold nanoparticles conjugated with a secondary antibody are then introduced and bound to the immobilized analyte in a sandwich structure, resulting in a color signal on the sensing membrane. The operation of our VFA test involves three sequential injection steps: 1) the running buffer, 2) the sample serum and nanoparticle conjugate, and 3) the washing buffer (Fig. 2.1c). After a 10-minute wait-period, the assay is complete and the VFA cassette is opened by twisting apart the top and bottom case, revealing the multiplexed sensing membrane on the top layer of the bottom case (Fig. 2.1c). This bottom case is then inserted into a custom-designed mobile-phone reader. An image of the activated multiplexed sensing membrane is subsequently captured and

analyzed via a fully-automated image processing and deep learning-based CRP quantification algorithm (Fig. 2.1d).

Multiplexed sensing membrane fabrication and VFA assembly

The multiplexed sensing membrane contains up to 81 spatially isolated immunoreaction spots that are each defined by a 'spotting condition' which refers to the capture protein and the associated buffer dispensed onto the nitrocellulose sensing membrane prior to assembly and activation. Therefore, to design the multiplexed sensing membrane for computational analysis, a custom spot-assignment algorithm was developed to generate a 'spot map' within the active area of the sensor. Based on a given grid spacing and number of spotting conditions, the assignment algorithm distributes spotting conditions such that no single spotting condition is disproportionately positioned near the center or the edge of the sensing membrane. Because the vertical flow rate can vary radially across the sensing membrane, leading to variations of each immunoreaction across the sensor area, this step mitigates a potential bias on any given spotting condition. With seven spotting conditions (see Table 2.2.) in a 9x9 grid format (1.3 mm periodicity), the spot-assignment algorithm produced the map shown in Figure 2.1c, which was implemented as the initial design for this study.

An automated liquid dispenser (MANTIS, Formulatrix®) was used to deposit 0.1 μL of the different protein conditions directly onto a nitrocellulose (NC) membrane in the algorithmically determined pattern shown in Fig. 2.1c. During the spotting process, up to 24 NC sensing membranes were produced on a single connected sheet, constituting one fabrication batch, and up to three batches were produced on a given day. In order to evaluate batch-to-batch variations, we intentionally produced sensing membranes over multiple fabrication batches as well as with two

reagent batches (*i.e.* sets of reagents which had unique storage times and/or lot numbers). Each sensing membrane was therefore tagged with a corresponding fabrication batch ID (FID, *e.g.*, 1, 2 or 3,) and reagent batch ID (RID, *e.g.*, 1 or 2).

Following the automated spotting procedure, the NC sheets were incubated at room temperature for 4 hours after which they were submerged in 1% BSA blocking solution and allowed to incubate at room temperature for 30 min. The NC sheets were then dried in an oven at 37°C for 10 min, after which they were cut into individual sensing membranes (1.2 x 1.2 cm) using a razor. The remaining paper materials contained in the VFA were produced following the methods outlined in a previous publication [45]. All the paper materials, including the NC sensing membrane were then assembled within the top and bottom cases of a 3-D printed VFA cassette, with foam tape holding together the paper stack (see Table 2.1).

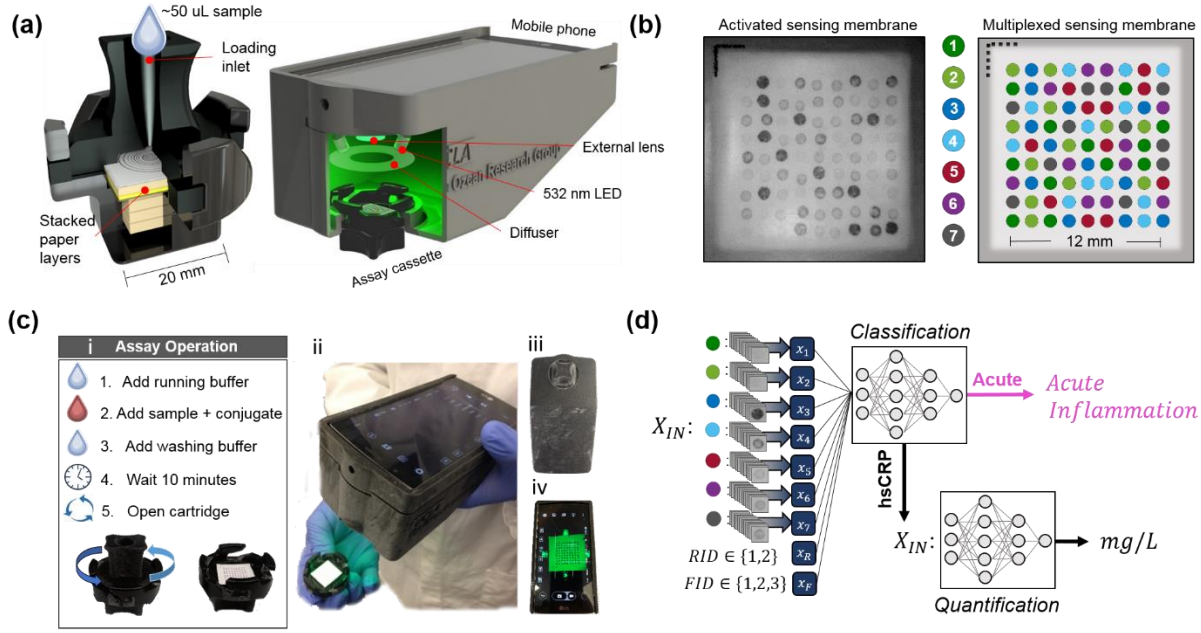


Figure 2.1. The multiplexed vertical flow assay (VFA) (a) The VFA cassette cross-section and mobile-phone reader with the inserted VFA cassette to be tested. (b) The multiplexed sensing membrane contained within the VFA cassette. The algorithmically determined immunoreaction spot layout (right) contains seven unique spotting conditions, each

of which uniquely reacts with the sensed analyte and the signal-forming Au NPs. A raw image of an activated sensing membrane taken with the mobile-phone reader is shown to the left. (c) (i) The VFA assay operation protocol (ii) The VFA cassette and mobile phone reader after the assay completion. The VFA cassette inserted into the mobile phone reader from the (iii) bottom and (iv) top view. (d) Block diagram of the computational analysis, showing the input features X_{IN} which contain, the average signals from like-spotting conditions along with the reagent batch ID (RID) and the fabrication batch ID (FID).

hsCRP assay procedures

Each hsCRP measurement with our VFA test is performed as follows: first 5 µL of serum sample is diluted 10 times in a running buffer (3% tween 20, 1.6% BSA in PBS) resulting in a 50 µL sample solution. Then 200 µL of running buffer is injected into the VFA inlet and allowed to absorb. After absorption into the VFA paper-stack (~ 30 sec), 50 µL of sample solution is mixed with 50 µL of the gold-nanoparticle (Au NP) conjugate solution (see Appendix for synthesis), and the mixture is pipetted into the inlet and allowed to absorb. Lastly, after absorption of the sample solution, 400 µL of the running buffer is added to wash away the nonspecifically bound proteins and Au NPs. After a 10-minute reaction time, the VFA cassette is then opened, and inserted into the bottom of the mobile-phone reader (Fig. 2.1a.). This mobile reader images the multiplexed sensing membrane using the standard Android camera app (ISO: 50, shutter at 1/125, autofocused), and saves a raw image of the VFA sensing membrane (.dng file) for subsequent processing and quantification of the CRP concentration.

Data processing

Custom image processing software was developed to automatically detect and segment the immunoreaction spots in each mobile-phone image of the activated VFA cassette (see Appendix, Fig. 2.5). After segmentation, the pixel average of each spot is calculated and subtracted by the pixel-average of a locally defined background containing BSA blocked NC membrane. Each background-subtracted spot signal is then normalized to the sum of all the spots on the sensing membrane. The final spot signal $s'_{m,p}$ is therefore described by,

$$s'_{m,p} = \frac{s_{m,p} - b_{m,p}}{\sum_p \sum_m (s_{m,p} - b_{m,p})} \quad \text{Eq. 2.1}$$

where m represents the spotting condition, and the p represents the p^{th} redundancy on the VFA per condition. $s_{m,p}$ is the pixel average of a given segmented spot, and $b_{m,p}$ is the local background signal. The final VFA signal per condition can then be calculated as:

$$x_m = \frac{1}{P_m} \sum_{p=1}^{P_m} s'_{m,p} \quad \text{Eq. 2.2}$$

where P_m is the number of redundancies for a given spotting condition. The normalization step in Eq. (1) helps us to account for sensor-to-sensor variations borne out of pipetting errors, fabrication tolerances, as well as operational variances.

Clinical testing

We procured remnant human serum samples (under UCLA IRB #19-000172) for hsCRP testing using our VFA platform. Each clinical sample was previously measured within the standard clinical workflow as part of the UCLA Health System using the CardioPhase *hs*CRP Flex® reagent cartridge (Cat. No. K7046, Siemens) and Dimension Vista System (Siemens). In total, we measured 85 clinical samples in triplicate with our VFA sensors. All but one sample was within

the standard hsCRP range of 0 to 10 mg/L, with the outlier having a concentration of 83.6 mg/L. In addition to testing these clinical samples, nine CRP-free serum samples (Fitzgerald Industries International, 90R-100) were measured as well as nine artificial samples created by spiking 200, 500, and 1000 mg/L CRP into CRP-free serum samples. These artificial samples were tested to simulate serum samples from patients undergoing acute inflammatory events. Though relatively rare in the context of hsCRP testing, such high concentration samples can be falsely reported as having a low CRP concentration due to the hook-effect. Therefore, these samples were included to test if our multiplexed computational VFA could avoid such false reporting. Among different batches of 273 fabricated VFA sensors, we removed one VFA test from the data-set due to a fabrication error (misalignment, see Appendix, Fig. 2.6a), and removed two triplicates due to abnormally high levels of non-specific binding, which was immediately obvious in the low signals on the sensing membrane and unusual pink color observed on the top case (Appendix, Fig. 2.6b).

Computational VFA sensor analysis

After the clinical study was completed the image data from the activated VFA tests were partitioned into a training set (N_{train} = 209) and testing set (N_{test} = 57). This data partition was structured to ensure that the testing samples would be distributed linearly over the hsCRP range, and that samples were pulled proportionally from the different fabrication batches within each cardiovascular risk stratification group. The raw background-subtracted pixel average values are shown in the Appendix (Fig. 2.7), where the marker color and shape indicate the fabrication batch ID and the reagent batch ID, respectively.

Model and cost function selection

The training set was analyzed via a k-fold cross-validation (k=5) to determine the optimal learning algorithm for quantification of CRP concentration from the inputs X_{IN}. We evaluated different fully connected networks through a random hyper-parameter search, where the number of nodes, layers, regularization, dropout, batch-size, and cost-function were each randomly selected from a user-constrained list. A tiered neural network architecture (see Appendix, Fig. 2.8) with a cost function of mean-squared logarithmic error (MSLE) yielded the best performance over the random iterations of the cross-validation. As an alternative, a single neural network with multiple hidden layers, in contrast to the tiered structure, could also be used in providing an accurate and generalizable model.

2.3 Results

Optimization of VFA spots and conditions using machine learning

Machine learning-based optimization and feature selection of our VFA platform was performed in two distinct steps: spatial spot selection and condition selection, illustrated in Figure 2.2a and Figure 2.2b, respectively. For the spot selection process, a cost function, $j_{m,p,}$ was defined per sensing spot to represent the normalized distance from the mean of like-spots (*i.e.* spots that share the same condition) averaged over the samples in the training set,

$$j_{m,p} = \sum_{n=1}^{N_{Train}} \frac{\left| s'_{m,n,p} - \overline{s'}_{m,n} \right|}{\overline{s'}_{m,n}} \qquad \text{Eq. 2.3}$$

where $s'_{m,n,p}$ is defined in Eq. 2.1 with the added index n indicating the n^{th} sample in the training set. $\bar{s}'_{m,n}$is the spot signal averaged over each condition within a single test, *i.e.* $\bar{s}'_{m,n} = \frac{1}{P_m}\sum_{p=1}^{P_m} s'_{m,n,p}$.

The heat map in Figure 2.2a, which is interpolated from a 9x9 matrix of the cost function defined at each spot of the VFA, visualizes the statistically robust active areas of the VFA sensing membrane. To select a subset of spots from the 9x9 grid configuration, we then performed a k-fold (k=5) cross-validation. The cross validation was performed over 75 iterations where the input to the neural network, X_{IN}, was defined by incrementally smaller subsets of the original 81 spots for each iteration. The spot with the maximum cost $j_{m,p}$ was eliminated at each iteration, resulting in the last iteration containing a subset of 7 spots, each corresponding to a different condition. The MSLE value from the cross validation was then plotted for every iteration to visualize the trade-off between the number of spots and the error of the network inference (Fig. 2.2a). Due to the random training process of the neural network, there is noise associated with this curve, however a clear performance benefit can be seen after the elimination of the first 30 to 40 spots corresponding to the highest $j_{m,p}$. It is also clear that further reducing the number of spots results in substantial increase in quantification error. Therefore, the approximate minimum of the MSLE curve was used to define a subset of 38 spots for subsequent analysis.

After this initial spot selection (Fig. 2.2a), this subset of 38 spots was further subject to a condition selection step to further optimize the performance of our computational VFA for hsCRP. This second phase of the feature selection aims to select the most robust sensing channels as defined by the unique chemistry attributed to the different spotting conditions. To this end, we performed a second iterative k-fold (k=5) cross-validation analysis, eliminating one spotting

condition each iteration and tracking the cross-validation error as a result of each elimination. This process was repeated for incrementally smaller subsets of conditions defined by the minimum MSLE result from the previous iteration. Resulting from this analysis, Fig. 2.2b reports the MSLE and coefficient of determination as function of the number of spotting conditions, suggesting that eliminating the Mix 1 and Ag-low condition can lead to slightly better or equivalent performance when compared to the inclusion of all the original spotting conditions.

Taken together, this machine learning-based optimization of the VFA leads to the statistical selection of the best combination of spots and conditions (Fig. 2.2c inset) that can computationally determine the analyte concentration. The cross-validation results, compared to the gold standard hsCRP measurements, are also reported in Figs. 2.2c and d. Here the inputs to the neural network, X_{IN}, are defined by the optimal spot configuration as determined by the spot and condition selection (see Fig. 2.2c inset), and also include two additional integer features which correspond to the reagent ID ($RID \in \{0,1\}$) and the fabrication batch ID, ($FID \in \{1,2,3\}$).

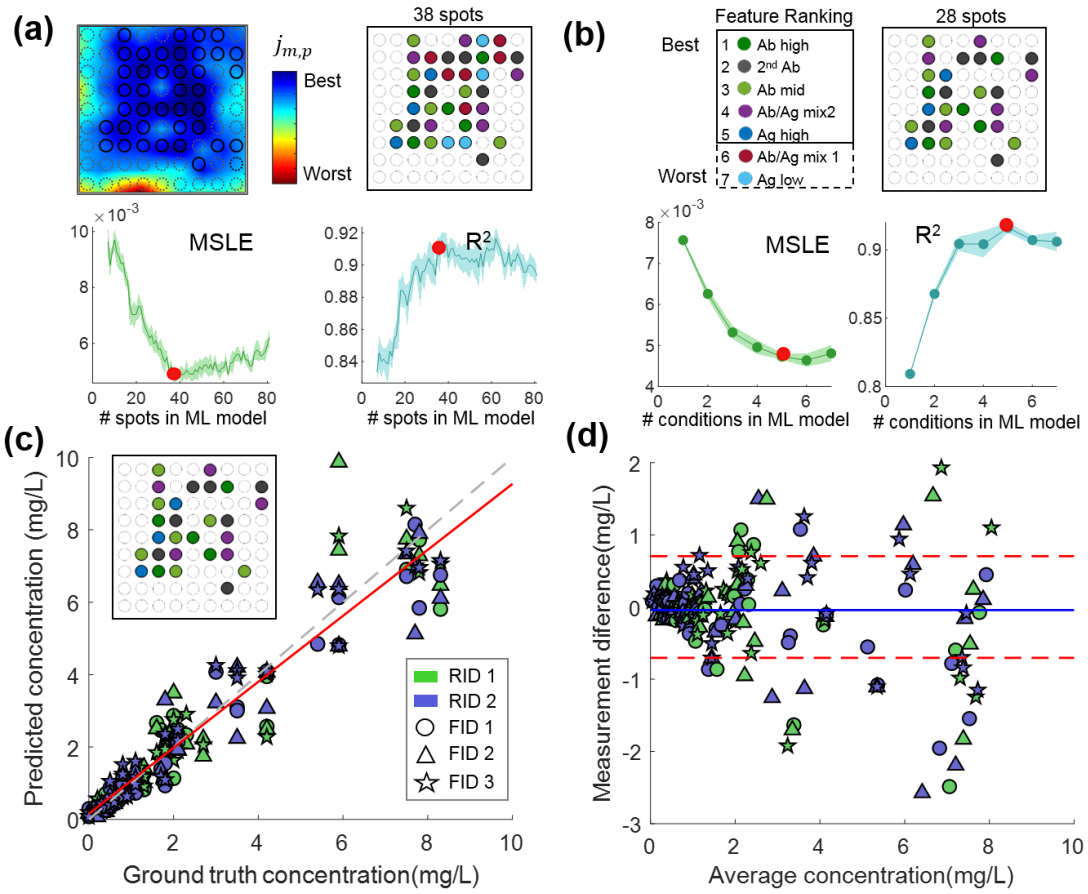


Figure 2.2. Cross-validation and feature selection analysis using the training data set of clinical samples (N_{Train} = 209) (a) The spot selection process. A heat-map (top left) is generated by plotting the cost function $j_{m,p}$ across the sensing membrane. The cross-validation performance, both MSLE and the coefficient of variation (R^2), is then plotted against the number of spots selected based off of $j_{m,p}$ (bottom). The optimal subset of spots (top right) is then selected based off the optimal quantification performance indicated by the solid red marker. (b) The condition selection process. Conditions are ranked based off of an iterative elimination method (top left), and the cross-validation performance is plotted against the number of conditions input into the quantification network. The optimal subset of conditions (top right) is then selected based off the optimal quantification performance indicated by the solid red marker. (c) The cross validation results using the selected features, where the ground truth CRP concentration is plotted against the predicted CRP concentration. The marker color and shape represent the different reagent batch ID (RID) and the fabrication batch ID (FID), respectively. (d) Bland-Altman plot of the same cross-validation results, where the dashed red lines represent the ± standard deviation of the measurement difference from the tested VFAs.

After this feature selection and cross-validation analysis reported in Figure 2.2, the final CRP quantification algorithm was trained using the entire training set (N_{train}=209) and the optimal spot configuration (Fig. 2.2c inset). In addition to the CRP quantification algorithm, a second classification algorithm was trained to identify the CRP samples representing an acute inflammation event, with a CRP concentration of > 10 mg/L ($\widehat{N}_{train}$= 6, $\widehat{N}_{test}$ = 6) (Fig. 2.1d). Next, we report our blind testing results using this optimized CRP VFA platform.

Validation of computational VFA performance for CRP measurements

Our computational VFA results from the blind testing set (N_{test}=57) correlated well to the quantification results of the gold-standard hsCRP Flex cartridge run on the Dimension Vista System (see Fig. 2.3). These samples were analyzed using only the pixel information contained within the computationally determined subset of 28 spots and 5 conditions (Fig. 2.3a). The x_m signals (Eq. 2) along with the FID and RID of each test sample were first classified by an initial neural network to determine if the test was in the hsCRP range (<10 mg/L) or the acute inflammation range (> 10 mg/L); we achieved 100% classification accuracy, and correctly classified 6 samples as acute and the rest (51 samples) as in the hsCRP range. The samples classified in the hsCRP range were then routed to a quantification neural network, whereas the acute samples were simply reported as acute along with a confidence score, as summarized in Fig. 2.3c.

The quantification accuracy of the hsCRP samples using our computational VFA was characterized by a direct comparison to the gold-standard values (Fig. 2.3b-c). With 51 tests quantified in the hsCRP range, the R^2 value was found to be 0.95, with a slope and intercept of the

linear best-fit line being 0.98 and 0.074 respectively. The overall average CV of the blind testing data was found to be 11.2% with the average CV for the low-risk, intermediate-risk, and high-risk stratified samples quantified as 11.5%, 10.1%, and 12.2 %, respectively. As a reference point, the FDA review criteria for hsCRP testing state an acceptance criterion of ≤ 20% overall CV, with a specific CV of ≤ 10% for samples in the low risk category (*i.e.* < 1 mg/L) [38].

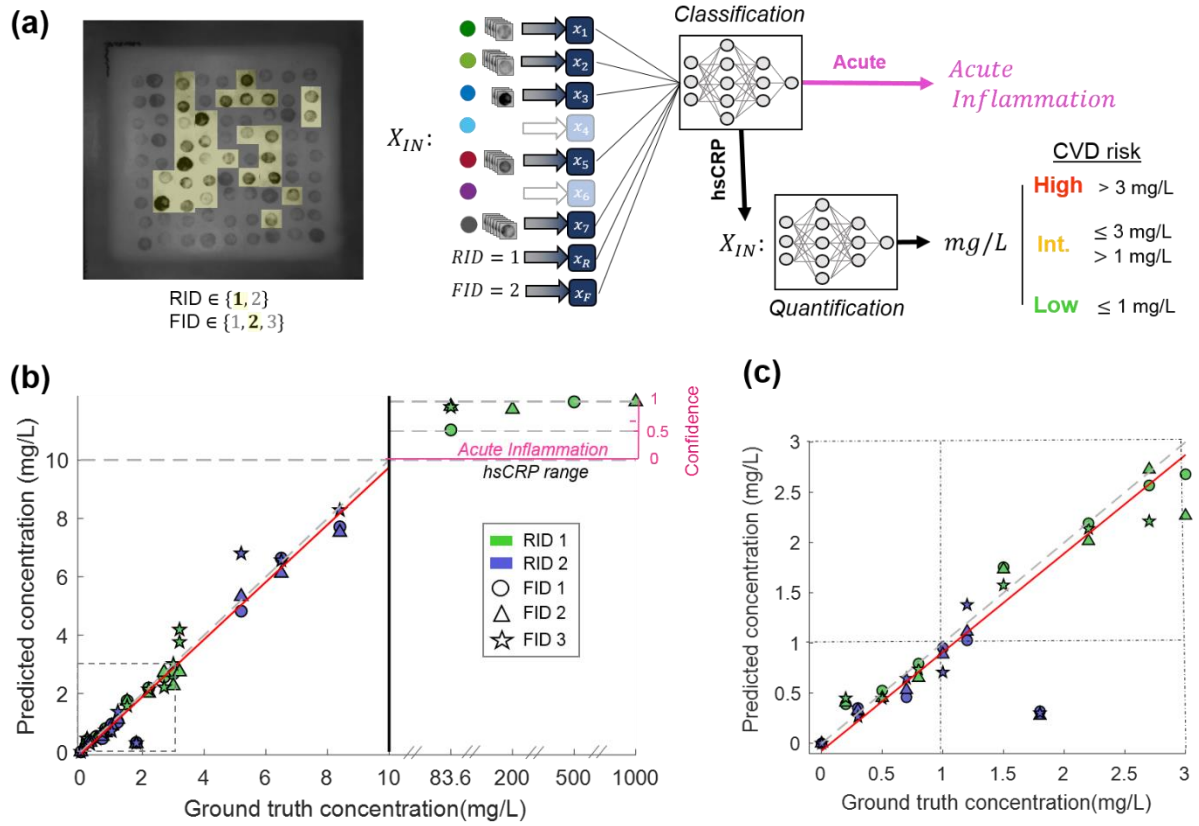


Figure 2.3. Blind testing results of clinical samples (Ntest = 57) (a) The features selected from the cross-validation analysis are extracted from a blind testing image and input into the neural network-based processing which infers the final CRP concentration. The clinical cutoffs for stratifying patients in terms of cardiovascular disease (CVD) risk are shown on the right. (b) The ground truth CRP concentration plotted against the VFA predicted CRP concentration (left y-axis) from blindly tested clinical samples. The dotted line represents a perfect match (y = x) and the red line represents the linear best fit. The confidence score is plotted (right y-axis) for the samples classified as acute. The marker color and shape represent the different reagent batch ID (RID) and fabrication batch ID (FID), respectively.

(c) The blind testing results for the low and intermediate CVD risk regimes, where the dotted lines represent the clinical cutoffs at 1 and 3 mg/L.

2.4 Discussion

Our VFA-based hsCRP test benefits from machine learning in several ways. Firstly, using neural networks to select optimal spots and infer analyte concentration from the highly multiplexed sensing channels greatly improves our quantification accuracy when compared to *e.g.* a standard multi-variable regression (see Appendix, Fig. 2.9). Deep learning algorithms such as the fully-connected network architecture used in this work, contain a much larger number of learned/trained coefficients along with multiple layers of linear operations and non-linear activation functions when compared to standard linear regression models. These added degrees of freedom enable neural networks to converge to robust models which can learn non-obvious patterns from a confounding set of variables, making them a powerful computational tool for assay interpretation and calibration. However, one concern with deep learning approaches is the possibility of overfitting to the given training set, especially in the instance of limited data. To mitigate this issue, we incorporated regularization terms in the hyper-parameter search (both L2 regularization and dropout), and found via cross-validation that the lowest error model employed the maximum degree of dropout regularization (*i.e.* 50%) [46], [47]. However, we observed better quantification results in the blindly tested samples when compared to the cross validation analysis, suggesting that our model appropriately generalized over the operational range of the hsCRP sensor.

Secondly, by incorporating fabrication information using RID and FID input features, the neural network was able to learn from batch-specific patterns and signals. This resulted in a 12.9% reduction in the blindly tested MSLE when compared to the performance of a network trained

without these fabrication batch input features. Similarly, incorporating the fabrication information reduced the overall CV from 16.64% to 11.2% and increased R^2 value from 0.92 to 0.95. It is important to note that these VFA tests (N=273) were fabricated without the use of industry-grade production equipment such as humidity and temperature controlled chambers, and in addition, several fabrication steps involved manual assembly. Taken together, these simple input features can benefit the performance and quality assurance of future computational POC tests following the methodology of this work. For example, the fabrication information could be included for each test in the form of a Quick Response (QR) code or could alternatively be logged into a GUI by the user before the measurement data are sent to the quantification network (running on a local or remote computer).

Another benefit of our computational VFA platform is the mitigation of false sensor response due to the hook effect. The VFA format importantly enables rapid computational analysis of highly multiplexed immunoreaction spots with minimal cross talk or interference among spots, which is inevitable for the case of standard lateral flow assays or RDTs. The multiplexed information reported by the different spotting conditions therefore allows for unique combinatorial signals to be generated over a large dynamic range (see Fig. 2.3b). The hook effect is clearly seen in our raw sensor data, exhibited by the capture antibody (Ab) condition (see Appendix Fig. 2.4, Fig. 2.7), illustrating how this condition alone can lead to false reporting of high analyte concentrations, *i.e.* in the case of acute inflammation. Therefore, without the incorporation of the monotonically responsive CRP antigen (Ag) spotting condition as one of the multiplexed channels in the VFA, high-concentration CRP samples can be falsely reported as low concentration due to the hook effect. This conclusion would still be true even if we trained another neural network that used a limited number of conditions as input; for example, by re-training the classification network using

only the Ab and Secondary Ab spotting conditions as inputs, we found that the 83.6, 200, and 1000 mg/L samples are falsely reported as having CRP concentrations of 7.81, 7.34, and 3.84 mg/L respectively. In the case of analyzing only the Ab channel, all of the high-concentration CRP samples would have been falsely reported as having concentrations below 10 mg/L. These results highlight the importance of multiplexed sensing in our computational VFA platform to mitigate the limitations induced by the hook effect in order to algorithmically enhance the dynamic range of our sensor.

A comparison of the computational xVFA to other commercially available hsCRP tests is shown in Table 2.3 in the Supporting Materials, highlighting comparable performance along with some major advantages of our platform such as its portability, low-cost, low sample-volume, and significantly extended dynamic range. It is important to note that though there are commercially available tests, no FDA approved POC test exists for CRP.

Computational sensing for assay development

Computational sensing broadly refers to the joint design and optimization of sensing hardware and software, and as implemented in this study, provides a framework for data-driven assay development where the diagnostic or quantification algorithm informs the multiplexed sensor design and vice versa. As detailed in the Methods section, the computational sensing approach begins with the selection of a neural network architecture and associated cost function. This first step is paramount to the computational sensor design, as it defines the model and error metric with which the subsequent feature selection is performed. The determination of the cost function therefore poses an interesting question for future computational sensors and diagnostic tests: because the selection of the cost function defines the training of a neural network, what are the

most clinically appropriate error functions with which one should design a computational sensing system? For example, in the case of cardiovascular risk stratification with the hsCRP test, an error of ± 0.1 mg/L is more problematic for samples that are in the range of the clinically defined cutoffs (*i.e.* 1 and 3 mg/L) when compared to samples with relatively higher CRP concentrations, such as 8 mg/L. Therefore, a traditional cost function for regression such as the mean-squared-error may not be as appropriate as the mean-squared-logarithmic-error or mean-absolute-percentage error, which take into account the relative ground-truth concentration for each error calculation. Therefore, special consideration must be given to the cost functions employed, and custom cost functions defined jointly by physicians/clinicians and engineers should be considered.

Feature selection and machine learning based optimization can similarly be used to inform the sensing membrane design. POC sensors can especially benefit from feature selection to circumvent noise borne out of their low-cost materials (such as paper used in our VFA) and operational variations. For example, the heat-map in Fig. 2.2a very well reveals how the immunoreaction spots closest to the edges of the sensing membrane contain the most variation in their normalized signals. This most likely results from the position-dependent vertical flow variations inherent in the inexpensive VFA format, which uses paper materials totaling <$0.2 per CRP test (Table. 2.1). These areas can therefore be avoided in future iterations of the sensor development, saving reagent costs and fabrication time, while also preserving robust sensing channels. Furthermore, identifying these areas of statistical variation can also inform the fabrication process. For example, Fig. 2.2a also shows that the top edge of the VFA sensing membrane as statistically more robust than the bottom and sides of the sensing membrane. Therefore, this spot selection analysis indicates a unidirectional fabrication bias in the lateral alignment of the sensing membrane within the VFA stack, which can be addressed in future iterations of the batch fabrication process.

Complementing the spot selection, the statistical condition selection process investigates the efficacy of the sensing channels and the unique immunoreactions defined by their spotting condition. Inherent complexities of the underlying chemistry such as the stochastic arrangement of the capture proteins within the porous NC membrane, as well as the effects of steric hindrance, pH, humidity, and temperature can obscure intuition behind the selection of spotting conditions for a given sensing application. Therefore, computational sensing systems can benefit from data-driven selection of sensing channels. For example, Fig. 2.2b shows that the quantification performance improves slightly upon the out-right elimination of the Mix 1 and Ag-low conditions. This suggests that their signal response is redundant or less stable when compared to the other conditions, and is confirmed by the poor repeatability of the Ag signal between the reagent and fabrication batches (see Appendix, Fig. 2.7). Such a feature selection procedure in a highly multiplexed format like the VFA could therefore be used to computationally screen spotting conditions from a large number of differing capture chemistries including, but not limited to, different structures of capture antibodies/antigens (i.e., polyclonal vs. monoclonal) as well as varying buffer conditions and reagent concentrations. Conditions which do not empirically benefit sensor performance can be replaced by new conditions in another iteration of the development phase, or be replaced by additional redundancies of effective conditions in order to benefit from signal averaging.

Additionally, this statistical feature selection process can inform cost-performance trade-offs to help design the most robust and cost-effective implementations of POC assays. For example, the reagent cost for the immunoreaction spots contained in the hsCRP VFA test is reduced by 62%, from $2.61 to $0.97 per test, by implementing only the computationally selected chemistries. Additionally, certain spotting conditions might have an optimal capture protein concentration due

to steric hindrance effects or higher degrees of nonspecific binding. Therefore, in a computational sensor, reagent costs can be significantly reduced without sacrificing assay performance by employing these statistically optimized capture-protein concentrations. One should also note here that these reagent costs per test would be significantly reduced under large scale manufacturing, benefiting from economies of scale, which is expected to bring the total cost per test (including all the materials and reagents) to <$0.5.

Taken together, we showed a data-driven sensor design and read-out framework, enabled by deep learning, for improving POC tests. As a use-case scenario, we demonstrated hsCRP testing with a colorimetric paper-based multiplexed VFA and clinical samples covering a large dynamic range. The multiplexed sensing membrane contained in the VFA was jointly developed with a quantification algorithm based on a fully-connected neural network architecture. First, a training data-set was formed by measuring human serum samples with the VFA. Then, through cross-validation of the training set, the most robust subset of sensing channels was selected from the multiplexed sensing membrane and used to train a CRP quantification network. The network was then blindly tested with additional clinical samples and compared to the gold standard CRP measurements, showing very good agreement in terms of quantification accuracy and precision. Additionally, the multiplexed channels and computational analysis helped us overcome limitations to the operational range of the CRP test borne out of the hook-effect. Our results demonstrate how a computational sensing framework and multiplexed sensor design can be used to engineer robust and cost-effective POC tests that have the potential to democratize diagnostics and expand access to care.

2.5 Appendix

The gold nanoparticle-C-Reactive Protein antibody (AuNP-antiCRP) conjugate is synthesized using the following protocol:

1. Mix 900 µl of 40 nm AuNP solution (Ted Pella Inc., 15707-1), 100 µl 0.1M Borate buffer (pH 8.5), and 5 µl anti-CRP mouse IgG antibody (Abcam, ab8278). Incubate the mixture at 25^{o}C for 1hr.
2. Following the 1-hour incubation, add 100 µl of 1% BSA in PBS solution and mix by vortexing. Then incubate the mixture at 25^{o}C for 30 minutes.
3. Transfer the mixture to the fridge and incubate at 4℃ for 2 hours.
4. Centrifuge the mixture in a tube at 8000 rpm at 4 ℃ for 15 minutes.
5. After centrifugation, open the tube and discard the supernatant.
6. Add 1 ml of 10 mM tris buffer (pH 7.4) to the microcentrifuge tube containing the AuNP-antCRP mixture and mix by vortexing.
7. Repeat the centrifugation and wash steps (steps 4,5,6) twice to enhance the purity of the mixture.
8. Add 100 µL of storage buffer (0.1 M borate buffers, pH 8.5 with 0.1% BSA and 1% sucrose) to the supernatant and mix via pipetting.

The final concentration of AuNP antibody conjugates (5 OD) was confirmed by optical density measurements at 525 nm.

Supplementary Figures

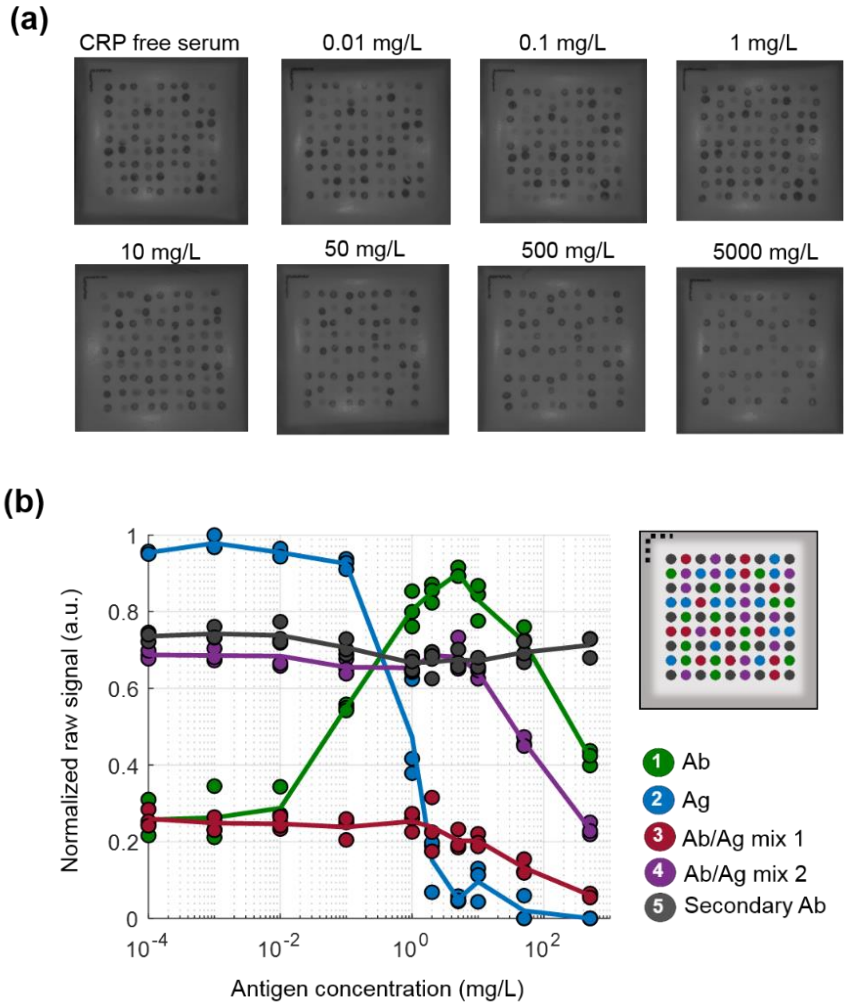


Figure 2.4. Response of the multiplexed sensing channels to different analyte concentrations. (a) Black-and-white images from the mobile-reader of the sensing membranes activated at different CRP concentrations. CRP was spiked into CRP-free serum and run without dilution. The spot-map corresponding to these sensing membranes is shown on the right with a color-coded legend below. (b) The normalized raw signals of five different spotting conditions implemented into the VFA as the sensors are activated with varying CRP concentrations, which were spiked into CRP-free serum. The following spotting conditions were used within PBS buffer: 1) Primary CRP antibody (Ab) at 1 mg/mL; 2) the CRP antigen itself (Ag) at 2.1 mg/mL; 3) A mixture of the CPR Ab and Ag at 0.8 and 0.08 mg/L, respectively; 4) A mixture of the CPR Ab and Ag at 0.8 and 0.24 mg/mL; and 5) the CRP secondary Ab at 0.2 mg/mL.

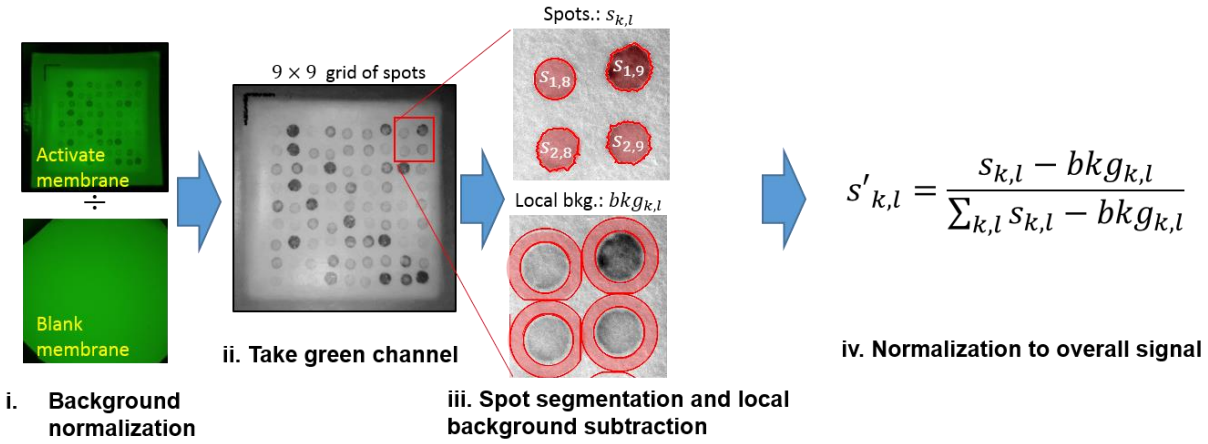


Figure 2.5. Overview of the image processing. i) The image of the sensing membrane is normalized to a universal blank background image (a blank NC membrane), and ii) the green channel is taken. iii) The spots are segmented through an automated algorithm and a local background is taken in a donut-shape outside of the segmented area. iv) The average pixel intensity of the local background $bkg_{k,l}$ is subtracted from the average pixel intensity of the segmented spot $s_{k,l}$ and normalized to the sum of all the background subtracted spot signals. Here, the indices k and l correspond to the row and column locations of the spots on the 9x9 grid, respectively.

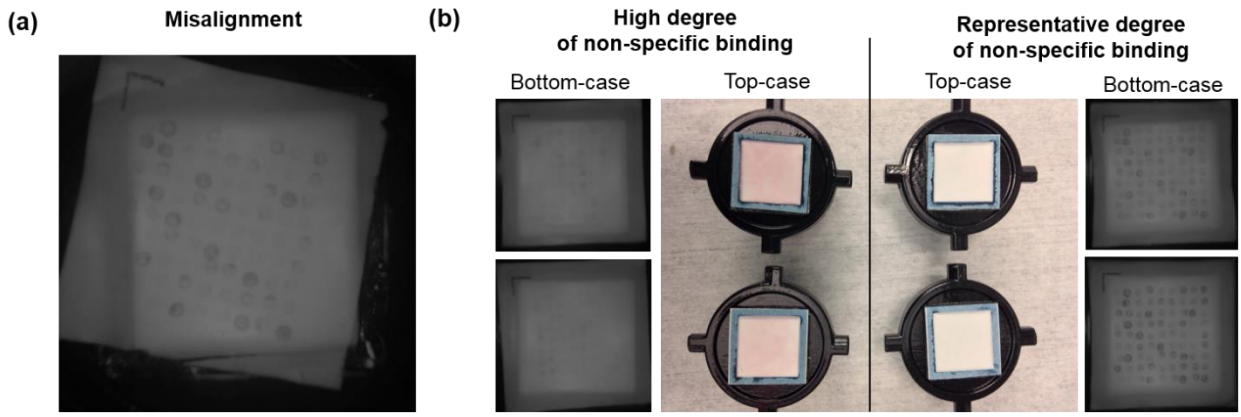


Figure 2.6. Rejected samples from the clinical testing. (a) A misaligned sensing membrane (fabrication error). (b) High non-specific binding in two clinical serum samples which resulted in a pink-color to the paper-layers in the top case and a low overall signal on the sensing membrane (bottom-case). A comparison to a representative top and bottom case for a normal test is shown to the right.

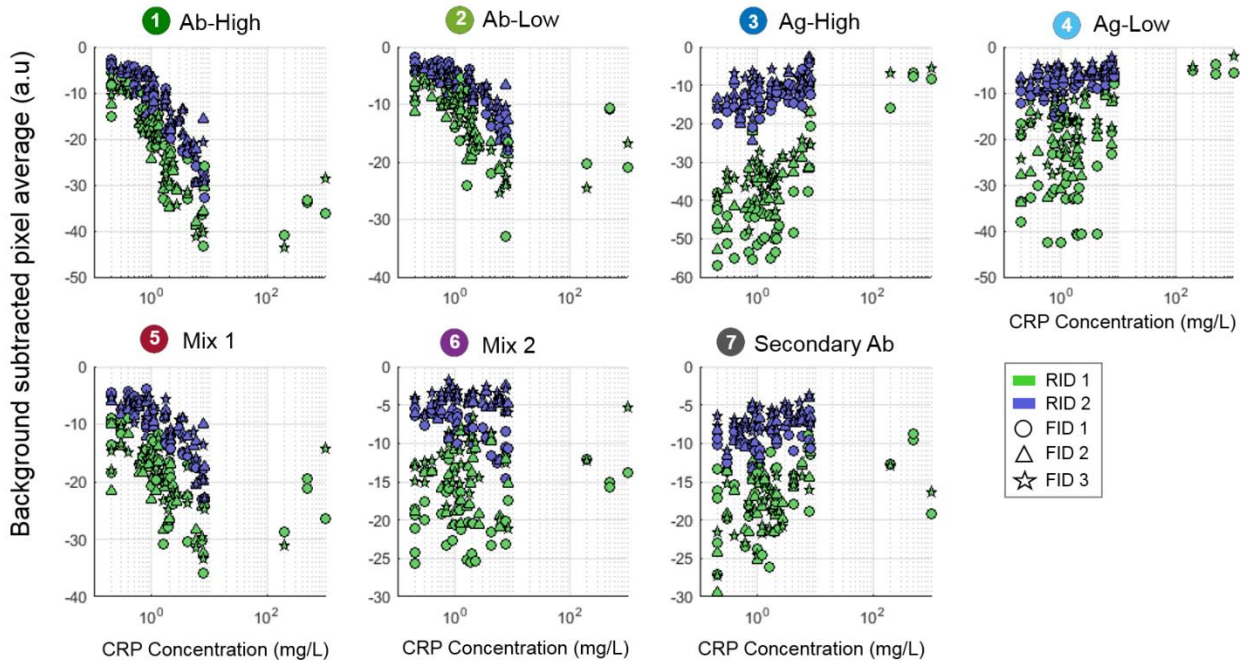


Figure 2.7. Raw data from the training data set of clinical samples. The background-subtracted pixel averages of the immunoreaction spots are plotted against the CRP concentration. Each data point represents the average of like-spots and plotted per spotting condition. The marker color and shape represent the different reagent batch ID (RID) and the fabrication batch ID (FID), respectively.

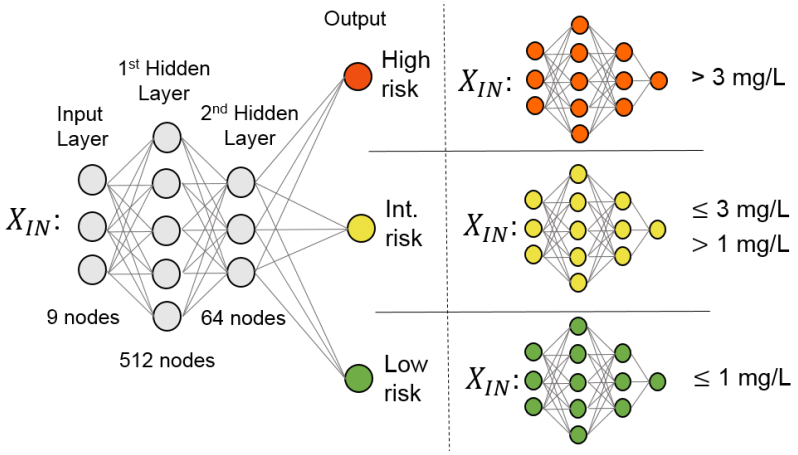
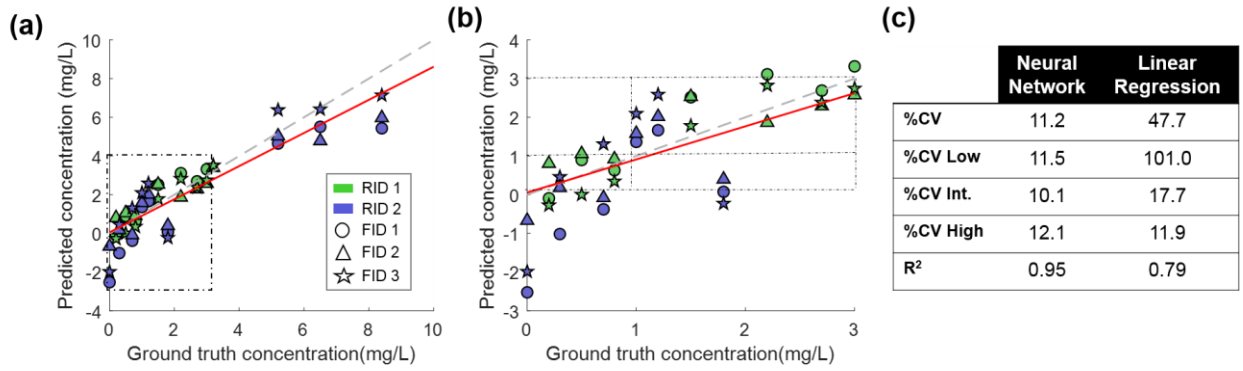


Figure 2.8. The quantification algorithm with a tiered network structure used for cross-validation. The first part of the algorithm (left, in grey) classifies a given sample, defined by the input X_{IN}, into the high, intermediate, or low risk hsCRP regime based off of the clinical cut-offs of 1 and 3 mg/L. The second part of the algorithm then uses separate networks trained with samples within each regime to quantify the CRP concentration of the sample. To avoid edge effects, each quantification network is trained with samples within their cut-off as well as with samples within

±50% of the corresponding cut-off value. Each layer was trained with 50% dropout, ReLu (Rectified Linear Unit) activation function, and a batch size of 22, as determined via a hyper-parameter search. For simplicity, every neural network used the same architecture and hyper-parameters, differing only in the output layer (*i.e.* classification or quantification) and the training data.

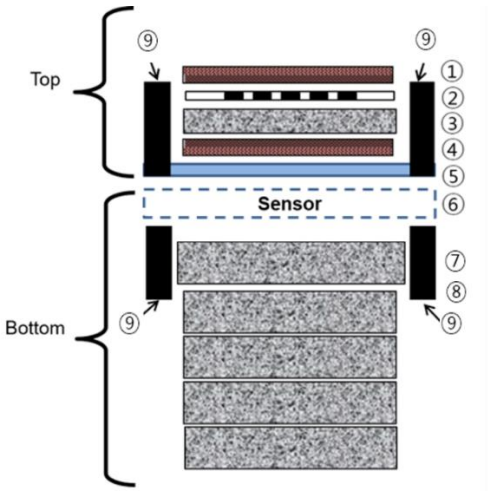


	Neural Network	Linear Regression
%CV	11.2	47.7
%CV Low	11.5	101.0
%CV Int.	10.1	17.7
%CV High	12.1	11.9
R^2	0.95	0.79

Figure 2.9. Blind testing results of the clinical samples using a multi-variable regression trained on the full x_m inputs (*i.e.* input data defined by all the 81 spots). (shown here for comparison to our approach detailed in the main text). (a) The ground truth CRP concentration plotted against the predicted CRP concentration from blindly tested clinical samples. The marker color and shape represent the different reagent batch ID (RID) and the fabrication batch ID (FID), respectively. (b) The blind testing results for the low and intermediate CVD risk regime, where the dotted lines represent the clinical cut-offs at 1 and 3 mg/L. (c) Table comparing the % coefficient of variation (%CV) and coefficient of determination (R^2) between the neural network based analysis (results show in Figure 2.3 of the main text) and the multivariable regression, which clearly demonstrate the major advantages of the deep learning based neural net infer.

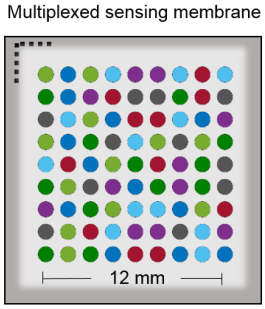
Supplementary Tables

Table 2.1. Material specification and cost breakdown of the dry VFA contents. A cross-sectional diagram of the paper layers contained in the VFA is shown on the left with numbers corresponding to the different materials in the table. The multiplexed sensing membrane of the VFA is denoted by the blue dotted outline, contained on the top layer of the bottom case.



	Layer	Material specification	Cost (¢)
①	Asymmetric membrane	Vivid GX, (1.2x1.2 cm)	0.5
②	Vertical flow diffuser	NC membrane, 0.45 µm pore size (1.2x1.2 cm)	2.9
③	Absorption pad	(1.2x1.2x0.1 cm)	1.4
④	Asymmetric membrane	Vivid GX (1.2x1.2 cm) reverse orientation	0.5
⑤	Supporting membrane	NC membrane, 0.22 µm pore size (1.7x1.7 cm)	2.9
⑥	Multiplexed sensing membrane	NC membrane, 0.22 µm pore size (1.7x1.7cm)	2.9
⑦	Absorption pad	(1.4x1.4x0.18 cm)	1.4
⑧	Absorption pad (stack)	(1.2x1.2x1.2x0.18 cm) x 4 pads	5.6
⑨	Foam tape		0.7
		TOTAL COST	**18.8¢**

Table 2.2. The seven spotting conditions implemented for the clinical testing with our computational VFA platform (right). The algorithmically determined spot map of the multiplexed sensing membrane (left). Specific colors encode the conditions.

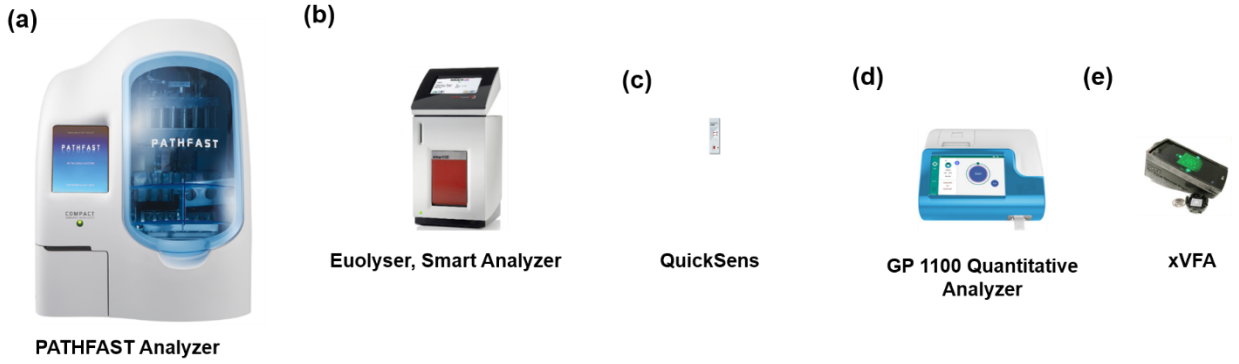


	Condition Abbreviation	Protein	Vendor, Product Info.	Concentration	Cost per spot (¢)	Spots per test
1	Ab-High	CRP Capture Antibody (Ab)	Abcam, ab136176	1 mg/mL	8.9	12
2	Ab-Low	CRP Capture Antibody (Ab)	Abcam, ab136176	0.4 mg/mL	3.6	12
3	Ag- High	CRP Antigen (Ag)	Fitzgerald Industries International, 30-AC05AF	2.1 mg/mL	1.5	12
4	Ag- Low	CRP Antigen (Ag)	Fitzgerald Industries International, 30-AC05AF	1.05 mg/mL	0.8	12
5	Mix 1	Capture Ab + CRP Ag	Combination (see above)	0.8 + 0.08 mg/mL	3.7	11
6	Mix 2	Capture Ab + CRP Ag	Combination (see above)	0.8 + 0.24 mg/mL	3.8	11
7	Secondary Ab	Secondary CRP Antibody	SouthernBiotech, anti-mouse IgG: 1036-01	0.2 mg/mL	0.1	11

Table 2.3. Comparison of point-of-care (POC) xVFA to other high-sensitivity CRP (hsCRP) testing systems. It is important to note that thought these tests (besides the xVFA) are commercially available, there is no FDA approved POC test for hsCRP. A to-scale comparison of the readers and tests is also shown (bottom) along with hyperlinks to the references (Ref) cited in the last column. Table Hyperlinks: [1] PATHFAST package insert, [2] PATHFAST retail

cost, [3] Eurolyser package insert, [4] Eurolyser retail cost, [5] QuickSens package insert, [6] GP 1100 Analyzer package insert, [7] GP 1100 Analyzer retail cost

	Test Type	Portable?	%CV	R^2	Dynamic Range	Susceptible to Hook-Effect?	Sample Volume	Assay time	Analyzer size	Analyzer weight	Cost (per-Reader, per-Test)	Power Requirement	Ref
PATHFAST Analyzer (PATHFAST)	Immunoturbidimetric	No	4.1% (blood)	0.991 (blood)	0.05 – 30 mg/L	Yes	100 µL (serum or blood)	< 17 minutes (serum or blood)	34.3 x 56.9 x 47.5 cm	28 kg	~$17,200, ~$9.75 (retail)	Wall power	[1][2]
Eurolyser CCA180 (Eurolyzer)	Immunoturbidimetric	No	5.2% (serum) 8.6% (blood)	0.9995 (serum) 0.9988 (blood)	0.10 - 320 mg/L	Yes	100 µL (serum or blood)	~3 min (serum) ~4 min (blood)	26 x 14.5 x 14 cm	3.4 kg	~$3,299, ~$5.00 (retail)	Wall power	[3] [4]
QuickSens, hsCRP (Biognostic)	Colorimetric	Yes	Not Quantitative	Not Quantitative	0.5 – 10 mg/L	Yes	120 µL (serum or blood)	~15 minutes	No Reader	No Reader	No Reader, Not found	None	[5]
GP 1100 Analyzer (GeTein)	Immunofluorescent	Somewhat	≤15% (serum)	0.941 (serum)	0.5–200.0 mg/L	Yes	10 µL (serum or blood)	<20 minutes	26.1× 24.1 × 11.5 cm	1.9 kg	$2,300 Not found (retail)	Wall power	[6] [7]
xVFA	Colorimetric	Yes	11.2% (serum)	0.950 (serum)	0.1 – 10 mg/L 85 – 1000 mg/L	No	5 µL (serum)	~12 min	15.5 x 8.0 x 6.2 cm	0.36 kg	~$200 ~$1.16 (material cost)	Battery powered	



References

[1] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017.

[2] B. E. Bejnordi *et al.*, "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, Dec. 2017.

[3] L. Hu *et al.*, "An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening," *J Natl Cancer Inst*.

[4] Y. Rivenson, Z. Göröcs, H. Günaydin, Y. Zhang, H. Wang, and A. Ozcan, "Deep learning microscopy," *Optica, OPTICA*, vol. 4, no. 11, pp. 1437–1443, Nov. 2017.

[5] Y. Rivenson, T. Liu, Z. Wei, Y. Zhang, K. de Haan, and A. Ozcan, "PhaseStain: the digital staining of label-free quantitative phase microscopy images using deep learning," *Light: Science & Applications*, vol. 8, no. 1, p. 23, Feb. 2019.

[6] Y. Rivenson *et al.*, "Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning," *Nature Biomedical Engineering*, p. 1, Mar. 2019.

[7] J. L. V. Shaw, "Practical challenges related to point of care testing," *Practical Laboratory Medicine*, vol. 4, pp. 22–29, Apr. 2016.

[8] M. Anastassova Dineva, L. Mahilum-Tapay, and H. Lee, "Sample preparation: a challenge in the development of point-of-care nucleic acid -based assays for resource-limited settings," *Analyst*, vol. 132, no. 12, pp. 1193–1199, 2007.

[9] S. Wang, M. A. Lifson, F. Inci, L.-G. Liang, Y.-F. Sheng, and U. Demirci, "Advances in addressing technical challenges of point-of-care diagnostics in resource-limited settings," *Expert Review of Molecular Diagnostics*, vol. 16, no. 4, pp. 449–459, Apr. 2016.

[10] M. Schito *et al.*, "Opportunities and Challenges for Cost-Efficient Implementation of New Point-of-Care Diagnostics for HIV and Tuberculosis," *J Infect Dis*, vol. 205, no. suppl_2, pp. S169–S180, May 2012.

[11] P. Yager, G. J. Domingo, and J. Gerdes, "Point-of-Care Diagnostics for Global Health," *Annual Review of Biomedical Engineering*, vol. 10, no. 1, pp. 107–144, 2008.

[12] T. R. Kozel and A. R. Burnham-Marusich, "Point-of-Care Testing for Infectious Diseases: Past, Present, and Future," *Journal of Clinical Microbiology*, vol. 55, no. 8, pp. 2313–2320, Aug. 2017.

[13] A. M. López-Marzo and A. Merkoçi, "Paper-based sensors and assays: a success of the engineering design and the convergence of knowledge areas," *Lab Chip*, vol. 16, no. 17, pp. 3150–3176, Aug. 2016.

[14] A. W. Martinez, S. T. Phillips, G. M. Whitesides, and E. Carrilho, "Diagnostics for the Developing World: Microfluidic Paper-Based Analytical Devices," *Anal. Chem.*, vol. 82, no. 1, pp. 3–10, Jan. 2010.

[15] K. Mahato, A. Srivastava, and P. Chandra, "Paper based diagnostics for personalized health care: Emerging technologies and commercial aspects," *Biosensors and Bioelectronics*, vol. 96, pp. 246–259, Oct. 2017.

[16] S. Smith, J. G. Korvink, D. Mager, and K. Land, "The potential of paper-based diagnostics to meet the ASSURED criteria," *RSC Adv.*, vol. 8, no. 59, pp. 34012–34034, Sep. 2018.

[17] "Paper Diagnostics Market Worth $10.50 Billion by 2025 | CAGR: 8.0%." [Online]. Available: https://www.grandviewresearch.com/press-release/global-paper-diagnostics-market. [Accessed: 10-Jan-2019].

[18] E. Primiceri *et al.*, "Key Enabling Technologies for Point-of-Care Diagnostics," *Sensors (Basel)*, vol. 18, no. 11, Oct. 2018.

[19] A. N. Hoofnagle and M. H. Wener, "The fundamental flaws of immunoassays and potential solutions using tandem mass spectrometry," *Journal of Immunological Methods*, vol. 347, no. 1, pp. 3–11, Aug. 2009.

[20] S. Amarasiri Fernando and G. S. Wilson, "Studies of the 'hook' effect in the one-step sandwich immunoassay," *Journal of Immunological Methods*, vol. 151, no. 1, pp. 47–66, Jul. 1992.

[21] "The hook effect: a need for constant vigilance," *Ann Clin Biochem*, vol. 43, no. 4, pp. 314–317, Jul. 2006.

[22] E. Rey, D. O'Dell, S. Mehta, and D. Erickson, "Mitigating the hook effect in lateral flow sandwich immunoassays using real-time reaction kinetics," *Anal Chem*, vol. 89, no. 9, pp. 5095–5100, May 2017.

[23] J. Oh, H.-A. Joung, H. S. Han, J. K. Kim, and M.-G. Kim, "A hook effect-free immunochromatographic assay (HEF-ICA) for measuring the C-reactive protein concentration in one drop of human serum," *Theranostics*, vol. 8, no. 12, pp. 3189–3197, May 2018.

[24] Y. Kyoung Oh, H.-A. Joung, H. Han, H.-J. Suk, and M.-G. Kim, "A three-line lateral flow assay strip for the measurement of C-reactive protein covering a broad physiological concentration range in human sera," *Biosensors and Bioelectronics*, vol. 61, pp. 285–289, Jan. 2014.

[25] B. Berg *et al.*, "Cellphone-Based Hand-Held Microplate Reader for Point-of-Care Testing of Enzyme-Linked Immunosorbent Assays," *ACS Nano*, vol. 9, no. 8, pp. 7857–7866, Aug. 2015.

[26] M. P. McRae, G. Simmons, J. Wong, and J. T. McDevitt, "Programmable Bio-nanochip Platform: A Point-of-Care Biosensor System with the Capacity To Learn," *Acc. Chem. Res.*, vol. 49, no. 7, pp. 1359–1368, Jul. 2016.

[27] X. Xu *et al.*, "Advances in Smartphone-Based Point-of-Care Diagnostics," *Proceedings of the IEEE*, vol. 103, no. 2, pp. 236–247, Feb. 2015.

[28] H. Zhu, S. O. Isikman, O. Mudanyali, A. Greenbaum, and A. Ozcan, "Optical Imaging Techniques for Point-of-care Diagnostics," *Lab Chip*, vol. 13, no. 1, pp. 51–67, Jan. 2013.

[29] Z. S. Ballard, D. Shir, A. Bhardwaj, S. Bazargan, S. Sathianathan, and A. Ozcan, "Computational Sensing Using Low-Cost and Mobile Plasmonic Readers Designed by Machine Learning," *ACS Nano*, vol. 11, no. 2, pp. 2266–2274, Feb. 2017.

[30] A. Ozcan, "Mobile phones democratize and cultivate next-generation imaging, diagnostics and measurement tools," *Lab Chip*, vol. 14, no. 17, pp. 3187–3194, Jul. 2014.

[31] O. Mudanyali, S. Dimitrov, U. Sikora, S. Padmanabhan, I. Navruz, and A. Ozcan, "Integrated rapid-diagnostic-test reader platform on a cellphone," *Lab Chip*, vol. 12, no. 15, pp. 2678–2686, Aug. 2012.

[32] P. M. Ridker, "A Test in Context: High-Sensitivity C-Reactive Protein," *J. Am. Coll. Cardiol.*, vol. 67, no. 6, pp. 712–723, Feb. 2016.

[33] D. M. Lloyd-Jones *et al.*, "Framingham risk score and prediction of lifetime risk for coronary heart disease," *Am. J. Cardiol.*, vol. 94, no. 1, pp. 20–24, Jul. 2004.

[34] D. Adukauskienė, A. Čiginskienė, A. Adukauskaitė, D. Pentiokinienė, R. Šlapikas, and I. Čeponienė, "Clinical relevance of high sensitivity C-reactive protein in cardiology," *Medicina*, vol. 52, no. 1, pp. 1–10, Jan. 2016.

[35] W. Koenig *et al.*, "C-Reactive protein, a sensitive marker of inflammation, predicts future risk of coronary heart disease in initially healthy middle-aged men: results from the MONICA (Monitoring Trends and Determinants in Cardiovascular Disease) Augsburg Cohort Study, 1984 to 1992," *Circulation*, vol. 99, no. 2, pp. 237–242, Jan. 1999.

[36] A. K. Shrivastava, H. V. Singh, A. Raizada, and S. K. Singh, "C-reactive protein, inflammation and coronary heart disease," *The Egyptian Heart Journal*, vol. 67, no. 2, pp. 89–97, Jun. 2015.

[37] "2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults | Circulation." [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/01.cir.0000437738.63853.7a. [Accessed: 17-Sep-2018].

[38] C. for D. and R. Health, "Guidance Documents (Medical Devices and Radiation-Emitting Products) - Review Criteria for Assessment of C Reactive Protein (CRP), High Sensitivity C-Reactive Protein (hsCRP) and Cardiac C-Reactive Protein (cCRP) Assays - Guidance for Industry and FDA Staff." [Online]. Available:

https://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm077167.htm. [Accessed: 17-Sep-2018].

[39] G. J. Blake and P. M. Ridker, "Inflammatory bio-markers and cardiovascular risk prediction," *J. Intern. Med.*, vol. 252, no. 4, pp. 283–294, Oct. 2002.

[40] M. Dong *et al.*, "Rapid and Low-Cost CRP Measurement by Integrating a Paper-Based Microfluidic Immunoassay with Smartphone (CRP-Chip)," *Sensors*, vol. 17, no. 4, p. 684, Apr. 2017.

[41] R. Wu *et al.*, "Quantitative and rapid detection of C-reactive protein using quantum dot-based lateral flow test strip," *Analytica Chimica Acta*, vol. 1008, pp. 1–7, May 2018.

[42] Y. Cai, K. Kang, Y. Liu, Y. Wang, and X. He, "Development of a lateral flow immunoassay of C-reactive protein detection based on red fluorescent nanoparticles," *Analytical Biochemistry*, vol. 556, pp. 129–135, Sep. 2018.

[43] S. W. Oh *et al.*, "Evaluation of fluorescence hs-CRP immunoassay for point-of-care testing," *Clin. Chim. Acta*, vol. 356, no. 1–2, pp. 172–177, Jun. 2005.

[44] H.-A. Joung, Y. K. Oh, and M.-G. Kim, "An automatic enzyme immunoassay based on a chemiluminescent lateral flow immunosensor," *Biosensors and Bioelectronics*, vol. 53, pp. 330–335, Mar. 2014.

[45] H.-A. Joung *et al.*, "Paper-based multiplexed vertical flow assay for point-of-care testing," *Lab Chip*, Jan. 2019.

[46] P. Baldi and P. J. Sadowski, "Understanding Dropout," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 2814–2822.

[47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

Chapter 3. Point-of-care serodiagnostic test for early-stage Lyme disease using a multiplexed paper-based immunoassay and machine learning

In this chapter, I describe another use-case for the VFA computational sensor platform. Similar to Chapter 2, I discuss the methodology behind computational assay development of another diagnostic tests followed by a validation study with clinical samples. However, in this chapter, the assay is for Lyme Disease, i.e. requiring a binary positive/negative result, and in contrast to CRP quantification relies on the detection of numerous different analytes for accurate patient assessment. Here I will discuss the results of a large fully blinded clinical study performed in collaboration with the Lyme Disease Biobank, the results of which show immense promise for computational sensing to enable new and in-demand point-of-care technologies. In addition, this chapter discusses how computational sensing systems can account for drift in the multiplexed inputs that may not be accounted for in the fabrication batch information and training data set, i.e. changes that might occur naturally due to instabilities in the stored reagents.

Caused by the tick-borne spirochete, *Borrelia burgdorferi,* Lyme disease (LD) is the most common vector-borne infectious disease in North America and Europe. Though timely diagnosis and treatment are effective in preventing disease progression, current tests are insensitive in early-stage LD, with a sensitivity <50%. Additionally, the serological testing currently recommended by the US Center for Disease Control has high costs (>$400/test) and extended sample-to-answer timelines (>24 hrs). To address these challenges, we created a cost-effective and rapid point-of-care (POC) test for early-stage LD that assays for antibodies specific to nine *Borrelia* antigens and a synthetic peptide in a paper-based multiplexed vertical flow assay (xVFA). We trained a deep learning-based diagnostic algorithm to select an optimal subset of antigen/peptide targets, and then

blindly-tested our xVFA using human samples ($N_{(+)}$ = 42, $N_{(-)}$= 54), achieving an area-under-the-curve (AUC), sensitivity, and specificity of 0.950, 90.5%, and 87.0% respectively, outperforming previous LD POC tests. With batch-specific standardization and threshold tuning, the specificity of our blind-testing performance improved to 96.3%, with an AUC and sensitivity of 0.963 and 85.7%, respectively.

Part of this chapter has been drafted as a manuscript and is currently in peer-review: Z. Ballard, H. A. Joung, J. Wu, D. Tseng, H. Teshome, L. Zhang, E. J. Horn, P. M. Arnaboldi, R. J. Dattwyler, O. B. Garner, D. Dicarlo, A. Ozcan, "*Point-of-care serodiagnostic test for early-stage Lyme disease using a multiplexed paper-based immunoassay and machine learnin*g," (In review, Nature Biomedical Engineering).

3.1 Introduction

Lyme disease (LD) is the most common vector-borne infectious disease in both North America and Europe, causing ~300,000 infections annually in the United States[1], [2]. It is caused by infection with the spirochete *Borrelia burgdorferi* (Bb) transmitted by black-legged ticks (*Ixodes genus)*. Early disease is associated with a characteristic skin lesion, erythema migrans (EM) along with other symptoms[3]–[5]. If not diagnosed and treated with appropriate antibiotics, the infection can disseminate to distal sites including the nervous system, heart, and joints causing an array of symptoms, including e.g., lymphocytic meningitis, cranial neuropathy, facial nerve palsy, radiculopathy, A/V node heart block [3], [6], [7], and arthritis[4], [8].

Although a presenting EM is diagnostic, the characteristic lesion is absent in 10-20% of infected persons and is frequently atypical thus escaping recognition. This makes laboratory testing critical to confirm the diagnosis and guide treatment[3], [5], [6], [9] Despite recent advances in direct detection of Bb through e.g. Nucleic Acid Amplification Testing (NAAT), these methods

remain inadequate due to the low concentration and transient presence of Bb in the blood. Culturing Bb is also not practical due the slow growth of the bacteria, as well as the need for specialized growth media [9]–[11]. Therefore, current testing methods work indirectly by detecting specific antibodies produced by the body's immune response to the infection.

The United States Center for Disease Control and Prevention (CDC) recommends a 'two-tier' testing method, where the first-tier consists of a sensitive Enzyme Immunoassay (EIA) or immunofluorescence assay (IFA). If the first-tier is positive or equivocal, a Western Blot (WB) is then recommended for confirming the presence of 2 of 3 Immunoglobulin M (IgM) antibodies and/or 5 of 10 Immunoglobulin G (IgG) antibodies targeting Bb associated antigens[12]. A number of reports have also showed the efficacy of a modified two-tier test (MTTT) format, where the WB is replaced by a second complimentary EIA, and as a result, the FDA has recently approved the use of some EIAs as viable tests for the second tier[13]–[17].

Despite being the standard for the laboratory diagnosis of LD, the two-tier serological testing method has multiple drawbacks. Although there is a high specificity (>98%) and sensitivity (70-100%) in *late* LD, the two-tier test has poor sensitivity in *early-stage* LD, seldom exceeding 50% at the time when most patients seek medical care[6], [9], [14], [18], [19]. This is also the time when treatment is the least costly and most effective at preventing disease sequela. The poor sensitivity can be attributed to the underdeveloped immune response within the first weeks of infection in which a limited IgM antibody response is followed by an IgG antibody response. However, it is also exacerbated by the limited number of antigen-targets in the first tier test that may miss detection of antibodies produced during the earliest stage of infection.[20] Specifically, the earliest responses are to Flagellin B (FlaB) and p66 with responses to a number of additional

antigens such as OspC (25kd), VlsE, BBK32, FlaA (37kd), BmpA (39kd) and DbpA proteins developing as *B. burgdorferi* disseminates[20]–[23] .

The two-tier testing method also suffers from slow turn-around time (>24 hrs) and high costs (>$400/test), with estimated expenses exceeding $492 million *annually* just in the United States[24]. Additionally, the standard testing used in the two-tier format must be performed in a centralized testing facility by trained technicians, requiring bulky and expensive clinical analyzers. These drawbacks therefore limit accessibility to accurate LD testing, especially impacting populations far from clinical laboratories as well as populations in rural and forested areas where tick bites are common. Therefore, accurate and affordable LD testing methods for use at the point-of-care (POC) are in high demand[13], [18].

Paper-based lateral flow assays (LFAs), also known as Rapid Diagnostic Tests (RDTs), are appealing for POC serological analysis due to their low-cost, ease-of-use, and rapid nature[25]–[27]. These tests use e.g., colorimetric or fluorescent conjugates embedded in one-time-use cassettes to rapidly and cost-effectively detect the presence of antibodies specific to disease. LFAs however, are not conducive to the detection of multiple analytes due to their in-line geometry, measuring only one or two antibodies in a single test[28]. This restrictive design inherently limits the potential sensitivity and specificity of traditional RDTs for LD. For example, tests that rely on single antibody measurements (like many EIAs in the first tier) can be less robust to false positive results due to antigens (p66 and FlaB in particular) which contain epitopes that are highly cross-reactive with epitopes found in multiple other bacteria[29], [30]. They can also have low sensitivity if the target antigen is mismatched with the underlying immunodominance. This paradigm is the reason that the performance of the two-tier algorithm can depend on which EIA is

used in the first tier, as well as what strain of Bb the EIA test was designed against, B31 being the most common[31].

Therefore, to overcome this limitation, large-scale screening efforts alongside new epitope mapping and peptide synthesis are focused on developing a universal multi-antigen detection panel, with e.g. 5 to 10 LD-specific antigen targets being suggested for improving diagnostic performance for early LD[17], [32], [33]. Consequently, we have a unique opportunity to leverage advances in computation and machine learning, to train novel serodiagnostic algorithms with these rich, multi-antigen measurements derived from well-characterized clinical samples, to ultimately create new decision algorithms that outperform the traditional two-tier test[32]–[34]. Deep learning, which refers to the use of artificial neural networks with multiple hidden-layers, can be especially effective in developing nonlinear yet robust inference models from noisy data-sets with complex and confounding variables[35]–[37].

Here, we introduce a paper-based multi-antigen POC test powered by deep learning for serological diagnosis of early-stage LD. Our test is a multiplexed vertical flow assay (xVFA), composed of a stack of functional paper layers, which in contrast to the more common LFA, allows for a multi-antigen detection panel for measuring an array of LD-specific antibodies on a single sensing membrane. Containing 13 spatially-separated immunoreaction spots, the sensing membrane is functionalized with Bb-specific antigens (OspC, BmpA, P41, DbpB, Crasp1, P35, Erpd/Arp37) as well as a peptide (Mod-C6) composed of a C6-like epitope linked to a specific p41 epitope. The xVFA can be operated in 15 min, after which the assay cassette is opened and the sensing membrane is imaged by a custom-designed mobile-phone based reader. Computational analysis then quantifies the colorimetric signals on the sensing membrane through automated image processing, and a neural network is used to automatically infer a diagnosis from the

multiplexed immunoreactions. The diagnostic algorithm in this work was trained with 50 human serum samples (25 early-stage LD and 25 endemic controls), obtained from the Lyme Disease Biobank (LDB), run in duplicate for both IgM and IgG antibodies, resulting in 200 individually-activated xVFAs composing the training data-set. This training data-set was also used to computationally select a subset of detection antigens from the larger panel using a feature selection technique, improving the diagnostic performance and reducing the per-test cost. The computational xVFA was evaluated through blind testing of an additional 50 human serum samples (25 early stage LD and 25 endemic controls), obtained fully-blinded by the LDB. Testing entirely early-stage LD samples, we achieved an AUC of 0.95, and by equally weighing the false positive and false negative results, we obtained a sensitivity and specificity of 90.5% and 87.0% with respect to the gold-standard two-tier serological testing. By adjusting the diagnostic cut-off value to favor high-specificity during the training phase and incorporating batch-specific standardization for the network inputs, our blind testing specificity improved to 96.3%, with a small drop in our sensitivity (85.7%) in relation to the gold-standard two-tier serological testing.

There is currently a *first tier* POC LD test, but *no* FDA-approved POC test for LD that can serve as a replacement to the two-tier method[28], [38]. However, development of a POC two-tier replacement test will allow for more rapid diagnosis and better treatment outcomes. This is especially important as LD is projected to increase over the next decades as the geographic areas of tick-populations continue to expand [39]–[41]. Although multi-target POC sensing approaches for LD have been explored in the literature, the methods proposed either exhibited poor performance or have not undergone validation with a clinical study[28], [42]. We believe our platform demonstrates a leapfrog improvement over existing POC LD testing approaches,

reporting a cost-effective and rapid (15 min) paper-based multiplexed assay powered by deep learning for serological diagnosis of early-stage LD at the POC.

3.2 Materials and Methods

The multiplexed vertical flow assay (xVFA)

Overview

Our multi-target vertical flow assay (xVFA) is composed of a stack of functional paper layers and a sensing membrane contained within a 3-D printed plastic cassette. The cassette is divided into a top and bottom-case which can be separated through a twisting mechanism, revealing the multiplexed sensing membrane on the top layer of the bottom-case. The sensing membrane contains 13 immunoreaction spots defined by a black wax-printed barrier, where each spot is pre-loaded with a different capture-antigen or antigen epitope-containing peptide as well as proteins serving as positive- and negative-controls to enable multiplexed sensing information within a single test (Fig. 3.1B). For each xVFA, there are two top-cases used during the operation (Fig. 3.1C). The first top-case facilitates the uniform flow of a serum sample from the loading inlet to the sensing membrane, where LD-specific antibodies are bound to the detection antigens immobilized on the nitrocellulose surface. The second top-case is then used for color signal generation, where a conjugate pad, upon wetting, releases embedded gold-nanoparticles (AuNPs) conjugated to anti-human IgM or IgG antibodies. The AuNPs then bind to the LD-specific IgM or IgG antibodies previously captured on the sensing membrane, resulting in a color signal in response to the captured amount. After completion of these sandwich immunoreactions, both IgM and IgG running in parallel, the sensing membrane is immediately imaged by a custom-designed mobile-phone reader (Fig. 3.1D-E), which captures the background image (taken before the assay

operation) and the signal image (taken after the assay operation) of the sensing membrane for subsequent analysis in a computer, where a neural network is used to ultimately determine the final result (seropositive or seronegative). The general concept of paper-based vertical flow design was reported in our previous work[43]; a detailed breakdown of the functional paper layers and assay reagents can be found in the Appendix (Fig. 3.6, Table 3.1, 3.2) along with further discussion on the wax-printed sensing membrane design and optimized operational protocol (Fig. 3.7).

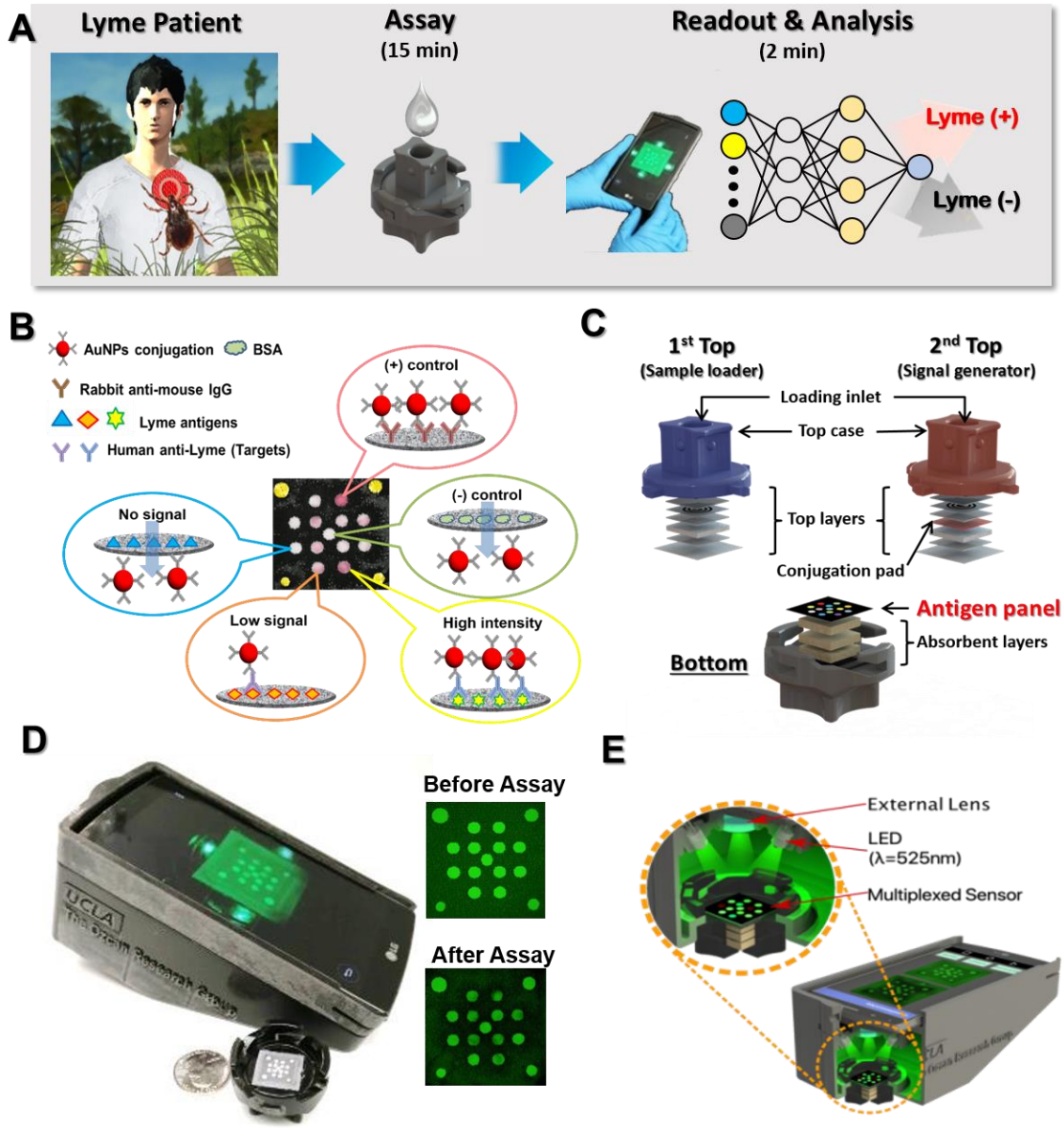


Figure 3.1. Overview of POC Lyme disease diagnostic testing using xVFA and machine learning, shown in **(A).** **(B)** Illustration of the multiplexed immunoreactions which occur on the sensing membrane during the xVFA operation. **(C)** Exploded diagram of the paper materials within the xVFA showing the sample loading top-case (in blue) and the signal generating top-case (in red). The bottom-case is shown below with the sensing membrane containing the multi-antigen panel. **(D)** Photograph of the mobile-phone reader with an opened xVFA cassette and example images of the sensing membrane (shown to the right). **(E)** Cross-section of the xVFA mobile-phone reader with an inset showing the illumination of the sensing membrane using an inexpensive opto-mechanical attachment to the smartphone.

The mobile phone-based assay reader (Fig. 3.1D-E) is constructed from a smartphone (LG G4H810) and 3-D printed (Dimension Elite, Stratasys) attachment containing four 525 nm wavelength light emitting diodes (LEDs) for even illumination of the sensing membrane. An external lens was also mounted in the 3-D printed attachment below the built-in phone camera lens system for enabling an in-focus field of view. All the images were obtained in raw dng format using the standard Android camera app of the smartphone.

Assay operation

First, a background image of the blank sensing membrane is taken with our mobile-phone reader. Then the first top-case is mated with the bottom-case, and 200 µL of running buffer is introduced to fully wet the paper layers in the xVFA. After the buffer is absorbed fully into the xVFA cassette (~20 seconds), 20 µL of serum sample is pipetted into the loading inlet and allowed to absorb. Then, a second addition of running buffer is introduced to the loading inlet, followed by a 6-minute wait period, during which the serum sample reacts with the sensing membrane and the unreacted sample is washed away to the lower absorbent pads. The first top-case is then exchanged with the second top-case, and 450 µL of running buffer is added to release the AuNP conjugates

responsible for color signal generation. After an 8-minute wait period, the xVFA cassette is opened and imaged by the mobile-phone reader to get the multiplexed signal.

Separate xVFAs are run in parallel for IgM and IgG antibody detection, where the only difference between the two xVFAs is in the conjugate pad, i.e. one conjugate pad contains AuNPs conjugated to anti-IgM antibodies and the other contains AuNPs conjugated to anti-IgG antibodies (Fig. 3.2A).

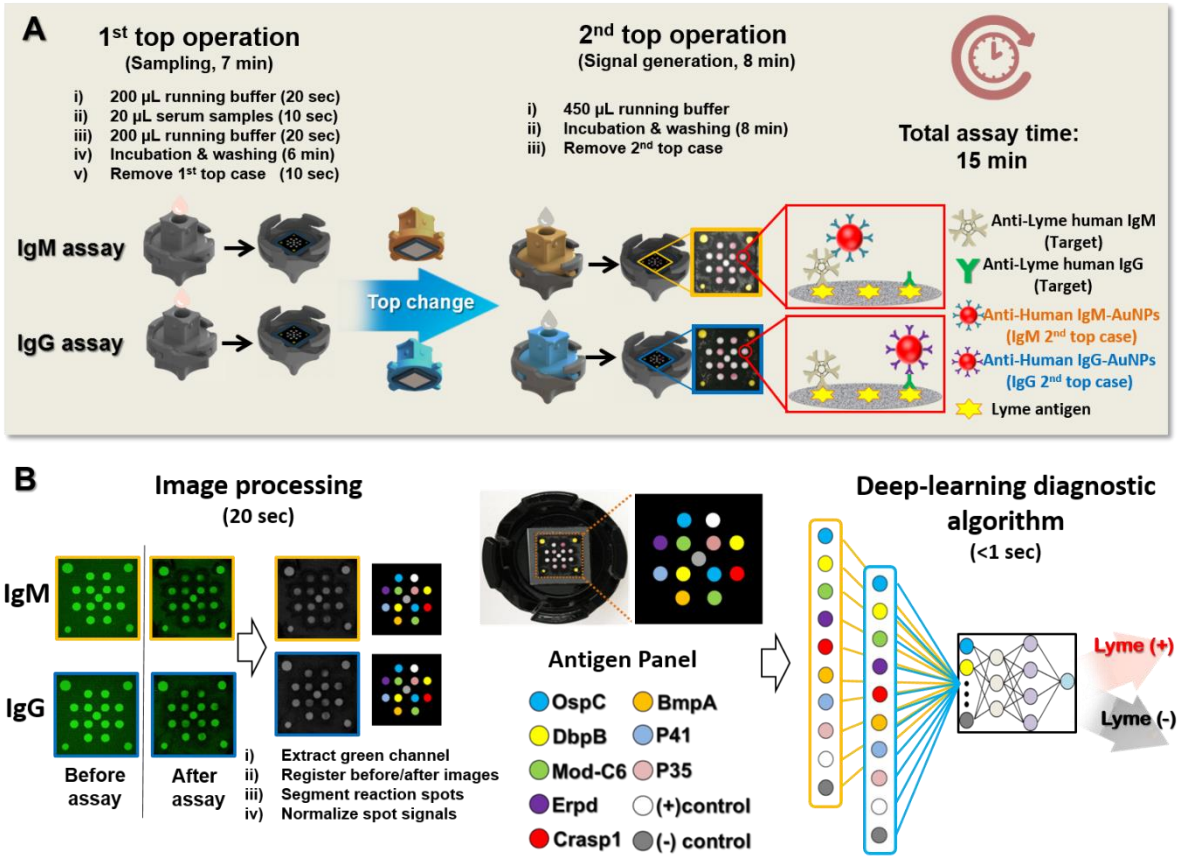


Figure 3.2. Overview of xVFA Lyme Disease assay operation (A) Detailed xVFA operation for the IgM and IgG assays, which are performed in parallel. A depiction of the IgM and IgG immunoassays is shown to the right**. (B)** Image processing (left), capture antigen panel (middle), and deep learning-based analysis (right) of the multiplexed sensing membrane.

Image processing and deep learning-based analysis

Raw dng images, captured by our mobile reader, of the sensing membrane taken before (background image) and after (signal image) the assay are converted to tiff format. The green pixels are then extracted, and the background and signal images are registered to each other via a rigid-transformation. The immunoreaction spots are then identified in the background image and a fixed-radius mask is defined per-spot at approximately 80% of the immunoreaction spot area. The pixel intensity within this immunoreaction spot mask is then calculated for the registered signal image and normalized by the pixel average of the corresponding immunoreaction spot in the registered background image,

$$R_m = \frac{\sum_{x,y\,\epsilon\,\Omega_m} I_{signal}\,(x,y)}{\sum_{x,y\,\epsilon\,\Omega_m} I_{background}\,(x,y)} \qquad \text{Eq. 3.1}$$

where R_m defines the normalized signal per immunoreaction spot (m) and Ω_m defines the x-y bounds of the fixed-radius mask per immunoreaction spot. This background normalization procedure helps account for non-uniformities in the illumination as well as local defects that might exist within the immunoreaction spots on each xVFA sensing membrane. Immunoreaction spots functionalized by the same capture antigen are averaged together, and each of the unique R_m signals, derived from both the IgM and IgG xVFAs, are then used for deep learning analysis (Fig. 3.2B).

Lastly, before being input into the diagnostic decision neural network, the R_m signals from both the IgM and IgG xVFAs are standardized to the mean, $\overline{R_m}$, and standard deviation, σ_m taken over the *training* set,

$$R'_m = \frac{R_m - \overline{R_m}}{\sigma_m}. \qquad \text{Eq. 3.2}$$

The decision neural network contains an input layer with M nodes (e.g., M = 20 with the full IgM/IgG antigen panel), three fully connected hidden layers with 128, 64, and 32 nodes, in the first, second, and third layers, respectively. Each layer contains batch normalization, a 50% dropout, and a Rectified Linear Unit (ReLU) activation function, defined by $f(x) = \max(0, x)$, with the exception of the final output layer, which uses a sigmoid activation function, yielding a network output as a numerical value between 0 and 1. A final binary diagnosis is then made by evaluating this numerical output with a blind cutoff value of 0.5. A binary cross-entropy cost function, defined as:

$$H(\boldsymbol{y}, \boldsymbol{y}') = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log y'_n + (1 - y_n)\log(1 - y'_n)\right] \qquad \text{Eq. 3.3}$$

was used during the training phase (learning rate = 0.001, batch-size = 32) with the two-tier based seropositive and seronegative diagnosis used as the gold-standard label, $y_n \in \{0,1\}$. Here, N represents the number of training samples, and $y'_n \in (0,1)$ represents the neural network prediction. The hyper-parameters mentioned above (except the diagnostic threshold) were determined by a random parameter optimization in which 2 and 3 hidden layers with varying number of nodes were tested via k-fold cross-validation (k =5) along with different dropout, learning rate, and batch size parameters randomly selected from a predefined list. All networks were trained for up to 1000 epochs with early stopping criterion defined by the stagnation of training accuracy not exceeding a change of 0.1% over 100 training epochs. The network weights at the last epoch (defined by the early stopping criterion) were then stored in the final model.

Clinical study

In total, 106 unique human serum samples were obtained from the LDB (collected under Advarra IRB Pro00012408). Out of these sample, 50 were used for training, 50 for blind testing, with the additional 6 used for precision evaluation. All the samples used were *early-stage* LD, having been obtained < 30 days since symptoms or the initial tick bite. All the cases and endemic controls were confirmed to be Lyme-positive or negative through standard two-tier testing methods, or in some cases quantitative Polymerase Chain Reaction (qPCR) or convalescent draws (seroconversion). For the first tier, a combination of Whole Cell Lysate Enzyme-linked Immunosorbent Assay (ELISA), C6 Peptide EIA, or VlsE/PepC10 ELISA testing was used. The second tier, performed regardless of the first-tier results, was comprised of the standard IgM and IgG WB. For this work, samples were considered seropositive if any of the three EIA tests in the first tier had a positive or equivocal (borderline) result and the second tier had a positive result for either the IgM or IgG WB as defined by the CDC recommendation (≥ 2 of 3 bands for IgM WB, and ≥ 5 of 10 bands for IgG)[9]. Samples were also considered seropositive by MTTT guidelines, where a seropositive diagnosis was called from two positive or equivocal EIA tests along with the presence of EM, without the need for a positive WB result[16], [31]. Additionally, all samples were confirmed negative for coinfections of Anaplasmosis and Babesia, both of which are infections also transmitted by the *Ixodes* tick and can produce similar constitutional symptoms to LD.

The clinical samples used for training and testing were obtained and tested in two separate sample-pulls. The first sample-pull contained 25 LD cases and 25 endemic controls from a collection site in East Hampton, New York between 2014 and 2016 (see Table 3.3). 24 of the 25 LD cases were seropositive, with the one exception confirmed Lyme-positive through *B.*

burgdorferi qPCR. All seropositive samples were early-stage LD, with 3 out of 24 being early disseminated, defined by the presence of multiple EMs. All samples in the first sample-pull were tested with the IgM and IgG xVFAs in duplicate and used as the *training* data-set ($N_{(+)} = 48$, $N_{(-)} = 52$). We should emphasize that since our *training ground truth* for this work is the two-tier test (which we aim to replace using our paper-based POC xVFA), one of the Lyme-positive samples here (tested in duplicate) has been added to the negative set ($N_{(-)} = 52$) since it was seronegative (although being qPCR positive).

The second sample-pull, used for blind testing of our platform, also contained 25 LD cases and 25 endemic controls, but was obtained fully blinded (i.e. without case and control labels), and tested with our xVFAs and the associated serodiagnostic algorithm three months after the first sample-pull (see Table 3.4). The second sample-pull also contained cases and endemic controls from collection sites in East Hampton (USA), New York (USA), but also included samples from several sites in Wisconsin (USA), which were *not* included in our training phase. In this blind testing set, 23 of the 25 LD cases were seropositive, with the two exceptions confirmed Lyme-positive through a convalescent blood-draw taken approximately 3 months after the first draw that revealed seroconversion via two-tier testing methods. All seropositive samples were early-stage LD, with 9 out of 23 being early disseminated. All the samples in the second sample-pull were tested with the IgM and IgG xVFAs in duplicate and used as a blinded test set ($N_{(+)} = 42$, $N_{(-)} = 54$), with four IgM tests removed due to fabrication error (see Appendix, Fig. 3.8). The seropositive and negative diagnostic predictions resulting from our xVFA testing and machine learning algorithm were blindly sent to the LDB prior to receiving the gold-standard testing labels (to determine our blind testing performance). Six sensing membranes were additionally processed (compared to other activated xVFAs) using an affine image transformation to correct

misalignments resulting from unforeseen expansion and/or contraction of the nitrocellulose pad, during the automated image registration step (see Appendix, Fig. 3.9).

Lastly six separate serum samples, obtained at the same time as the second sample pull were used for precision evaluation performed by a single assay operator as well as multiple, newly trained personnel. In summary, 106 unique human serum samples obtained from the LDB have been used for training, testing and further inspection of our xVFA LD detection platform.

3.3 Results

Training and cross-validation

The first sample-pull ($N_{(+)}$ = 48, $N_{(-)}$= 52) was used entirely for cross-validation and training of our xVFA platform. Figure 3.3 summarizes the signals from the multi-antigen detection panel across all the seropositive samples in the training set, with the raw $(1 - R)$ signals of all the tests shown in Figure 3.10 (Appendix). The statistical performance of each antigen in terms of its t-score (Appendix, Eq. 3.5) is also shown, with the Mod-C6 peptide and OspC antigen being the top-ranked immunoreaction spots for the IgG and IgM sensing membranes respectively. As will be detailed and discussed later on, the individual performance and therefore the value of different antigen-targets as a single disease discriminator is highly-limited in the context of LD diagnosis, and our multiplexed xVFA platform computationally selects a complementary set of antigen-targets that collectively diagnose early-stage LD, providing a cost-effective and rapid POC replacement for the two-tier test.

Before training the final serodiagnostic algorithm to be used for the detection of early-stage LD, the training set was used for selecting an optimal subset of antigens. We implemented a sequential forward feature selection (SFFS) method where the signals from each sensing channel

were added one at a time, into the input layer of the neural network and then trained via k-fold (k=5) cross-validation. After the addition of each input feature, the performance of the network was evaluated based off a mean square error cost function, i.e.,

$$J(\boldsymbol{y}, \boldsymbol{y}') = \frac{1}{N}\sum_{k=1}^{5}\sum_{i=1}^{\frac{N}{5}}\left(y_{i,k} - y'_{i,k}\right)^2 \qquad \text{Eq. 3.4}$$

where $y_{i,k} \epsilon \{0,1\}$ is the ground truth binary result (*i.e.* seropositive or negative) and $y'_{i,k} \epsilon (0,1)$ is the numerical output of the network for the i[th] test in the k[th] fold of the k-fold cross validation (k=5). *N* is the number of tests in the training data-set. The input feature which yielded the best network performance for that iteration was then kept as an input feature until all the 20 sensing channels were included as inputs (for ranking of their collective value for LD diagnosis); see e.g., Fig. 3.4C.

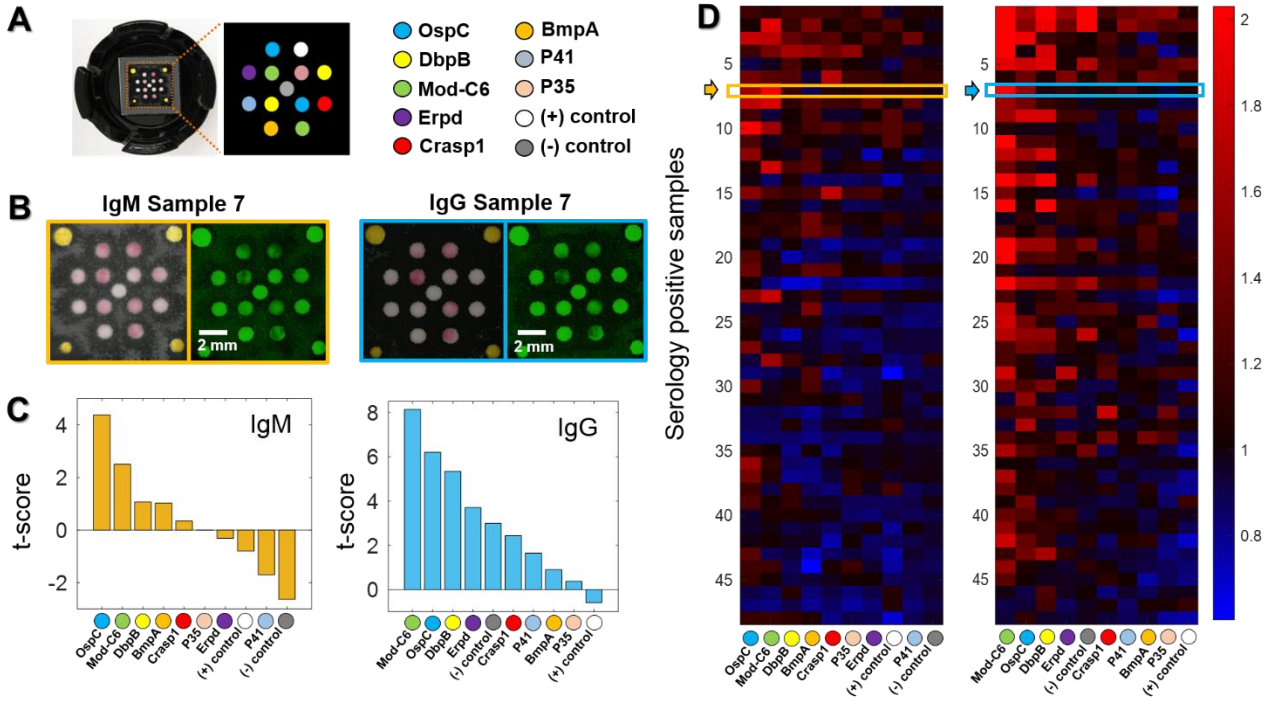


Figure 3.3. The xVFA training data set (A) The sensing membrane of our xVFA and the map of the multi-antigen panel. **(B)** Example images of IgM (left) and IgG (right) sensing membranes after activated by a human serum sample

(sample #7) under ambient lighting conditions and using the mobile-phone reader (under green illumination). **(C)** The t-score (Appendix, Eq. 3.5) of the multi-antigen channel calculated over the training data-set for the IgM (left) and IgG (right) sensing membranes ranked in descending order. **(D)** A heat map showing the signals from the multi-antigen panel of the IgM (left) and IgG (right) sensing membranes activated by the seropositive samples in the training set. The color bar represents the (1-R) signal of each sensing channel normalized to the mean signal of the same sensing channel across all the seronegative samples in the training data-set. The normalized IgM and IgG signals for sample #7 are outlined by the orange and blue boxes respectively.

Figure 3.4 shows the area under the receiver operator characteristic (ROC) curve (i.e., AUC) at each round of the SFFS antigen-selection algorithm, revealing a local maximum (AUC = 0.969) created by a panel of 6 antigens and the 3 control spots: OspC, P41, and the positive control spot (anti-mouse IgM) from the IgM sensing membrane, as well as the Mod-C6 peptide, Crasp1, BmpA, P41, positive control spot (anti-mouse IgG), and the negative control spot (Bovine Serum Albumin, BSA) from the IgG sensing membrane. Using this subset of sensing channels as the detection panel, a sensitivity of 91.7% and specificity of 96.2% was found via the cross-validation analysis (Fig. 3.4A, inset). The final LD diagnosis network was then trained using this computationally-selected subset as the antigen panel (M = 9), incorporating the tests (i.e., 100 activated xVFAs) from the first sample-pull (N_{Train} = 100, i.e., $N_{(+)}$ = 48 and $N_{(-)}$= 52), using the same network parameters outlined in the Methods.

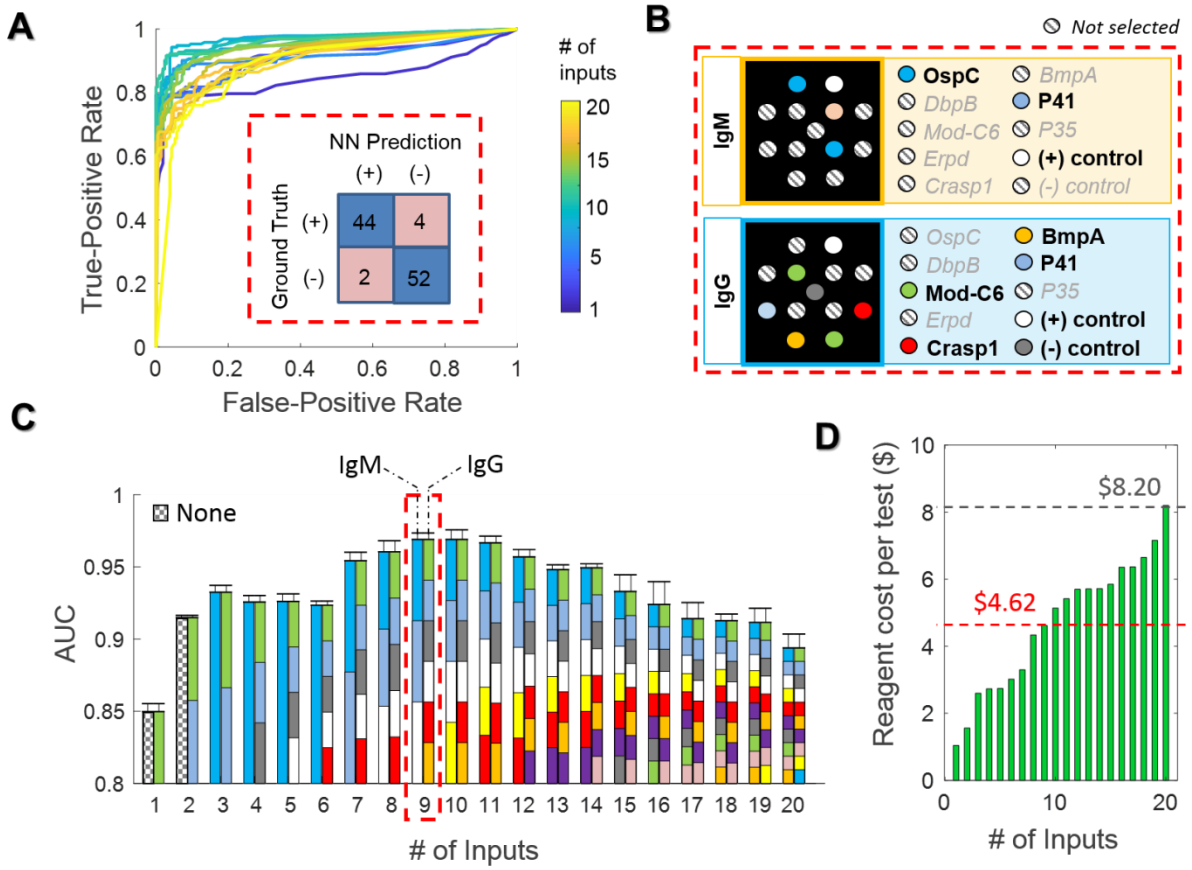
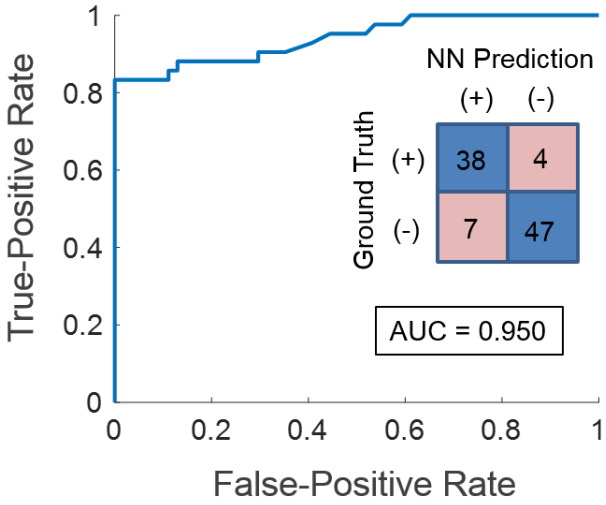


Figure 3.4. The feature selection process for the xVFA IgM and IgG antigen panel. (A) Receiver operator characteristic (ROC) curves resulting from the neural network inference during cross validation. The color bar represents the output from networks trained with different number of input features. The inset defined by the red dotted line shows the confusion matrix for all the samples in the training set predicted during cross-validation using the 9 antigen inputs shown in **(B)**, which is selected based on the optimization reported in **(C)**. **(C)** The area-under-the-curve (AUC) plotted for various networks trained with a different number of inputs of the multi-antigen panel (also see **(A)**). Each bar plot corresponds to the ROC curves shown in A, and is color-coded to represent which members of the multi-antigen panel are included as the inputs with the left and right side of each bar showing the members from the IgM and IgG sensing membrane, respectively. The red-dotted line shows the local optimum AUC (0.969) and the resulting 9 selected features. The error bars show the standard deviation between 4 different training instances of the same network. **(D)** The reagent cost of the sensing membrane versus the number of selected members in the multi-antigen panel as they are included in the order shown in **(C)**. Under large-volume manufacturing the reagent cost per

test is expected to drop by more than an order of magnitude. The red-dotted line represents the reagent cost for the 9 selected antigens, and the grey-dotted line shows the reagent cost of the whole antigen panel before feature selection.

Clinical blind testing

The second sample-pull ($N_{(+)}$ = 42, $N_{(-)}$= 54) was then entirely used for blind testing the performance of our xVFA diagnostic platform, yielding an AUC of 0.950, sensitivity of 90.5%, and specificity of 87.0% with respect to the seropositive and seronegative results, as summarized in Figure 3.5. This reported diagnostic performance of our xVFA platform achieved during the blind testing phase, to the best of our knowledge, exceeds previous POC tests for early-stage LD[31], [42]. As a point of reference, Table 3.5 shows a comparison with the recently FDA-approved POC test from Quidel, which can be used as a first-tier test, detecting IgM and IgG antibodies using the C6 antigen. Though FDA-approved, the Quidel test is *not* recommended as a replacement of the two-tier testing. The performance of our xVFA also outperforms a number of previous clinical studies investigating diagnostic performance of standard two-tier as well as modified two-tier testing with respect to the ultimate clinical diagnosis[9], [17], [38].



LD Stage	Sensitivity*	Positive Samples, $N_{(+)}$
Early	84.0%	25
Early Disseminated	100.0%	17
Overall	90.5%	42
	Specificity*	**Negative Samples, $N_{(-)}$**
Overall	87.0%	54

Figure 3.5. ROC curve for the blind testing data (N_{Test} = 96) as output from the neural network trained with the 9 selected antigens (see Fig. 3.4) from the training set (N_{Train} = 100). The inset shows the confusion matrix and area under the ROC curve (AUC). The table to the right summarizes the performance over the blindly-tested LD human serum samples with respect to the two-tier testing method.

Next, to achieve higher specificity, we tuned the decision threshold of the diagnostic network during the *training* phase such that the cross-validation specificity reached > 98%, resulting in a decision threshold of 0.66. Implementing this threshold in the blind testing phase (with 192 activated xVFAs), along with batch-specific standardization (also see the Discussion section), we achieved an AUC, sensitivity and specificity of 0.963, 85.7% and 96.3% with respect to the two-tier serology results (see Table 3.6). Such optimization of the decision biases can be utilized to achieve a desired false-positive and false-negative trade-off, depending on the clinical setting where the LD test is administered.

With this fine-tuned decision threshold, we achieved two false positive and six false negative results out of the 96 individual tests run through our xVFA platform, reaching an overall accuracy of 91.7%. Interestingly, some intuitive reasoning can be attributed to these instances of misdiagnosed tests. For example, four out of the six the instances of false negative tests were from

two patients that self-reported the *shortest* duration of LD symptoms (≤ 1 day) indicating that these samples may have the least developed immune response in our blind testing set. Additionally, one of these two false-negative duplicate pairs (LD188, see Table 3.4) was clinically positive in the first-tier due only to an equivocal result in the VlsE/PepC10 EIA. Because VlsE was eliminated from our multi-antigen detection panel during the pre-screening process (see Appendix, Fig. 3.11), our xVFA is incapable of detecting antibodies to VlsE. This can be addressed in the future by further optimizing the binding properties of VlsE antigen in the nitrocellulose substrate as well as expanding the number of diverse sera in the training set prior to the computational antigen selection.

It is also important to note that two seropositive samples in our testing set, i.e., LD185 and LD158 (see Table 3.4), were considered negative through standard two-tier testing (STTT), i.e. where the second-tier is IgG/IgM WB. Nevertheless, our POC test correctly called these samples positive with respect to the MTTT gold-standard label, which has demonstrated greater efficacy compared to the STTT in recent reports and has in fact lead the CDC to amend their recommendation for LD testing[16,17,38].

Lastly, half of the misclassified samples (4 out of 8) are from single discordant tests among the duplicate pairs. Therefore, to shed more light on this result, precision testing was performed with an additional six samples ($N_{(+)} = 3$, $N_{(-)} = 3$) obtained from the LDB, where each sample was measured with six repeated xVFA tests following the same operational protocol as in the training and testing phases (see Appendix, Fig. 3.12, 3.13, Table 3.7). To more realistically assess how the xVFA performs as a POC test, we also examined the precision between three different users/operators, two of which were new to xVFA operation and given only 5 minutes of training. With the batch-specific standardization and threshold tuning, the overall accuracy during this

precision testing, 91.7%, exactly matched the accuracy obtained during the blind testing. Additionally, no difference in accuracy was observed between the precision testing from a single operator and the testing completed by the multiple newly trained operators (Table 3.7). The precision of the individual immunoreactions was also investigated, yielding an average coefficient of variation of 8.9% for the same operator and 9.3% for the various newly trained operators (Appendix, Fig. 3.12, 3.13). While this shows good overall repeatability in the underlying immunochemistry, some antigen spots exhibited a coefficient of variation exceeding 20% which can ultimately lead to poor precision in the output of the decision algorithm. Therefore, despite the demonstration of interoperability, these testing results suggest that precision might be a limiting factor of our performance. Further improvements can be made by automating the fabrication processes and implementing humidity and temperature controlled environments, in order to improve the coefficient of variation of the individual immunoreactions as part of our xVFA. Additionally, ensuring that a large number of samples are used during the batch-specific standardization could potentially further improve our testing precision.

3.4 Discussion

Optimization of antigen selection in early-stage LD xVFA platform

Computationally selecting the detection panel from a larger set of antigen/peptide targets improves the performance of the serodiagnostic algorithm. The capture antigens functionalized to the sensing membrane produce varying degrees of statistical variance in their optical signals, especially over different batches of fabricated sensors. This can stem from fabrication tolerances borne out of the low-cost materials, or from operational variance. Some capture antigens can also exhibit varying degrees of cross-reactivity with other antibodies native to human sera. Our feature

selection procedure helps eliminate the least reliable discriminators while conserving an ensemble of reactions that can most reliably detect the immune response. A very good example of this phenomenon can be seen in this comparison: when forgoing the feature selection process and implementing the full antigen selection panel (M = 20) as inputs to the network, the cross-validation testing reveals that AUC, sensitivity and specificity are 0.894, 72.9% and 92.3% respectively, compared to 0.969, 91.7% and 96.2% respectively, from the network trained on the computationally-selected subset of 9 features (see Table 3.8 and Fig. 3.4); this clearly emphasizes that the use of more antigen-targets does not offer a better solution for LD diagnosis. Interestingly, using the single antigen with the highest t-score (the Mod-C6 like in the IgG membrane) alone as an input to the network, results in an AUC of 0.850 and a sensitivity and specificity of 77.1% and 98.1%, respectively. Such performance for early-stage LD samples is characteristic of EIAs that detect only a single antibody, as the antigen-targets employed, such as the Mod-C6, can be synthesized to limit the presence of cross-reactive epitopes while failing to detect less prevalent antibodies, such as OspC or BmpA that may be produced at the beginning of the infection, albeit at lower concentratrions[20], [44]. The significant benefit of the computationally-determined multi-antigen panel is also clearly reflected in the blind testing set, showing the highest AUC when compared to networks trained with the full antigen panel and the Mod-C6 alone (see Table 3.8).

It is also important to note that the SFFS method implemented in this work, which is referred to as a wrapper feature selection technique, does not simply select the top individual discriminators (i.e. the sensing channels with the highest t-score). Instead, it iteratively adds input features and assesses their performance as an ensemble of inputs to a neural network. We should emphasize that filtering the input features based on the top 9 individual t-scores results in *poorer* cross-validation performance (AUC = 0.900, Sensitivity = 79.2%, Specificity = 92.3%) when compared

to a network with 9 features selected by our SFFS method. This can be partly attributed to the relative unimportance of redundant information. In other words, antigen-targets associated with the same patient population or stage of infection as an already implemented antigen-target are of less value to the diagnostic performance, despite having good performance as a single discriminator; in fact, this is a *very important* conclusion for the design of multiplexed sensors in general, which certainly applies to our early-stage LD xVFA platform.

Additionally, the positive and negative control spots, while not intended for unilaterally discriminating seropositive and negative samples can contain pertinent relative information for a computational POC sensor. For example, the relative concentration of AuNPs in the conjugation pad as well as instances of strong non-specific binding from matrix effects inherent in human sera can be represented by the positive and negative control immunoreactions and thus factored into the logistical classifier.

Alternative to the wrapper SFFS technique used here, a global search could potentially be implemented in which the network is trained by *every combination* of possible sensing channels. However, this requires training a network for every possible subset of M sensing channels, multiplied by the number of folds in the k-fold cross validation, i.e. $k * \sum_{r=1}^{M} \frac{M!}{r!(M-r)!}$. With our multiplexed xVFA platform (M = 20) and 5-fold validation, this results in over 5.2 million training instances, which was prohibitive due to the computation time required (> 10,000 hours at 7.2 seconds per training instance on an NVIDIA GeForce FTX1080 Ti GPU).

In addition to the performance advantages discussed above, our feature selection process can also be used to reduce the cost per-test. For example, the reagent cost for the full antigen panel (M=20) would be reduced by 44% by implementing the subset of 9 immunoreaction spots selected during the training phase (Fig. 3.4D). Although not considered here, future methodologies could

even incorporate reagent cost into the feature selection cost function, $J(\boldsymbol{y}, \boldsymbol{y}')$ to better illustrate cost-performance trade-offs. Additionally, immunoreaction spots with antigen-targets that were not computationally selected could be replaced in future sensing membrane designs by redundancies of positive and negative control spots as well as already selected capture antigens, or even alternative antigen-targets not yet tested. This type of data-driven iterative assay development is universally applicable, and could be a powerful framework for improving multiplexed sensors, especially for complex diagnoses like LD as well as for POC tests that must balance cost and performance.

Optimization of the false-positive and false-negative rate

Another important aspect of the training and feature selection phase is the degree of trade-off between sensitivity and specificity. The diagnostic algorithm can be influenced during the training phase by penalizing instances of false negatives more heavily than instances of false positives, e.g. through tunable weights on the two terms in the binary cross-entropy cost function used to train the network (Eq. 3.3), or by adjusting the threshold which discriminates between positive and negative samples. In practice, it may be more beneficial for a POC assay that is intended to be used at the first line of patient assessment, to have a greater portion of false positives over false negatives, especially in the case of contagious diseases. However, another important consideration is the pre-test likelihood, which can be low for samples submitted for serological LD testing (<20%)[24], [45]. Therefore, it may also be important to ensure a lower portion of false positives in order to reduce the overall number of misdiagnoses. Ultimately however, due to the small number of misclassified samples observed during the training phase of this work (only 6), we did not implement any of these possible adjustments before the initial blind testing. With larger data-

sets, where the empirical effect of tuning the bias can be better modeled, these approaches should be considered and jointly investigated with experts in diagnostic testing.

Batch-specific standardization

Although our blind testing sensitivity is comparable to the cross-validation sensitivity during the training phase, indicating that our deep learning-based diagnostic algorithm did not over-fit to the training data-set, the network did exhibit a drift in its numerical output from the training phase to the testing set (see Appendix Fig. 3.14). The mean output for negative samples shifted from 0.060 ± 0.156 to 0.278 ± 0.246 (mean ± SD) between the training and testing samples (also shown in Fig. 3.14). This effect is likely due to the instability of the antigen-targets, which produced underlying statistical differences of the R_m signals between the training and testing phases of the clinical study, which were conducted 3 months apart due to sample availability from the LDB. Additionally, differences in temperature and humidity at the time of manual fabrication among other variables can impact repeatability across fabrication batches. Such issues are also pervasive in commercial assays, demanding production-grade equipment to establish environmental controls during automated fabrication as well as rigorous quality assurance to verify performance standards.

To mitigate such issues, future batches of sensors can be standardized to sample means ($\overline{R_m}$) and standard deviations (σ_m) as a characteristic signature of their fabrication batch (see Eq. 3.2). For example, during the initial blind testing phase reported in this work, the mean and standard deviation of the *training* data were used for the input standardization. This resulted in a drift of the $R_m{}'$ inputs away from zero, which can in turn manifest as a drift in the network output. Therefore, by standardizing the blind testing inputs to the sample mean and standard deviation of the *testing* batch, the drift in negative sample outputs is reduced by over 80% (see Appendix Fig. 3.14). As a

result, three of the false positive samples are subsequently classified as true negatives (with a decision threshold of 0.5), increasing the specificity from 87.0% to 92.6% and the AUC from 0.950 to 0.963 with respect to the two-tier serology, while only incurring one extra false negative sample. Practically, batch-specific sample means and standard deviations could be calculated as a running average of the activated sensors in a batch, and could even be exclusively determined using endemic control sera, which is more readily available. An alternative approach is to include fabrication batch information as inputs to the network in the form of external labels[46]. In this case, the network could learn inherent batch-to-batch variations and compensate for these differences through its internal tunable parameters.

Outlook

In conclusion, we demonstrated a paper-based computational multi-antigen xVFA platform capable of diagnosing early-stage LD at the POC. Our xVFA has a material cost of $0.42 per-test and can be performed in 15 min by an individual with minimal training. A low-cost and handheld optical reader enables automated analysis to quantify colorimetric signals generated on a nitrocellulose membrane, followed by analysis with a neural network for inferring a diagnosis from the multi-antigen sensing information. By computationally selecting a panel of detection antigens for IgM and IgG antibodies specific to LD and performing a fully-blinded clinical study with early-stage LD samples, we report an AUC, sensitivity, and specificity of 0.950, 90.5%, and 87.0%, respectively, with respect to the two-tier serological testing. Using batch-specific standardization and threshold tuning, we improved the specificity of our blind-testing performance to 96.3%, with an AUC and sensitivity of 0.963 and 85.7%, respectively.

The multi-target and POC nature of the computational xVFA make it uniquely suited for LD diagnostics, presenting major advantages in terms of time, cost, and performance when compared to (first-tier) EIAs with single antigen-targets as well as standard two-tier testing methods that are rather costly (e.g., >$400/test) and slow (>24 hours for results).

Future work can incorporate dual-mode read-out and operation for measuring IgM and IgG antibodies in a single test. Additionally, the computational framework outlined here can be used for iteratively designing more competitive versions of our xVFA that incorporate statistically more stable, sensitive, and specific capture molecules such as synthetic peptides with epitopes engineered for high capture-affinity and low cross-reactivity[32].

3.5 Appendix

Materials

Borrelia burgdorferi recombinant antigens were purchased from Prospec Inc ((OspC (BOR-004), BmpA (BOR-006), P41(BOR-001), DbpB (BOR-007) and OspA(BOR-013)), Rockland Immunochemicals Inc (Crsasp1 (000-001-C18), ErpD (000-001-C09), P35 (000-001-C12)), and MyBioSource, Inc. (VlsE1 (MBS145939)) The modified-C6 and specific p41 epitope containing peptide (Mod-C6) was obtained from Biopeptides Corp. Anti-Human IgG/IgM (ab99741/ab99770) were purchased from Abcam. Anti-mouse IgG (1036-01) was purchased from SouthernBiotech. Blocker™ Bovine Serum Albumin (BSA) (37525) was purchased from Bio-Rad. Nitrocellulose membranes (0.22μm (11327) and 0.45μm (11036)) were purchased from Sartorius Stedim North America Inc. A vivid plasma separation membrane (grade GX) was purchased from Pall Co., and the sample pad (CF7) as well as the conjugation pad (Grade Standard 14) were sourced from GE Healthcare Biosciences Corp. The absorbent pad (Whatman Grade 707)

was acquired from OpticsPlanet, Inc. The gold colloidal solution (40 nm colloid, 15707-1) was purchased from Ted Pella, Inc. Foam tape (Super-Cushioning Food-Grade Polyethylene Foam Sheets 1/16") was purchased from McMaster-Carr. A summary of the purchased materials can be found in Tables 3.1 and 3.2.

Fabrication of the multiplexed vertical flow assay (xVFA)

Preparation of antibody–AuNPs complexes and conjugate pads

Complexes of mouse anti-human IgM/IgG on AuNPs were achieved by adding 100 μL 0.1M borate buffer (pH 8.4) and 20 μL of antibody (0.5 mg mL^{-1}) to 1 mL gold nanoparticle solution (40 nm, 1 OD) in a sterile Eppendorf tube. The mixture was incubated for one hour at room temperature, then 100 μL of 1% BSA in phosphate-buffered saline (PBS) was added as a blocking buffer. After blocking for 30 minutes at 25 ºC, the mixture was incubated at 4 ℃ for one hour. To remove excess mouse anti-human IgG/IgM, the complexes were centrifuged at 4 ℃ for 15 min at 8000 rpm, and washed 3 times by 1 mL washing buffer (10 mM Tris buffer (pH 7.2)). The supernatant was then re-suspended in 100 μL 0.1 M pH 8.5 borate buffer, containing 0.1% BSA and 1% sucrose. The final concentration of the antibody–AuNPs complexes was determined by optical density measurement at 525 nm using a well-plate reader (Synergy 2 Multi-Mode Microplate Reader, BioTek Instruments, Inc.). Only the complexes which had 2 OD were applied to the conjugate pad. 70 μL of the conjugate solution was then pipetted onto each conjugation pad (1.15 X 1.15 cm), which had previously been blocked by 1 % BSA in 0.1 M borate buffer (pH 8.5). The pads were dried at 37°C for 30 min.

Preparation of multiplexed sensing membrane and functional paper layers

The multiplexed sensing membrane was produced using a 0.22 μm NC membrane and wax printer (Colorqube 8580DN, Xerox). Thirteen spatially isolated immunoreaction spots were defined by wax-printed barriers, allowing for different capture antigens to be spotted on the nitrocellulose membrane. After printing, the sensing membranes were incubated for 30 sec at 120 ^{o}C in an oven to allow the printed wax to melt and diffuse downward into the nitro-cellulose. Each of the 13 isolated sensing spots was then loaded by hand-pipetting 0.8 μL of 0.1 mg mL^{-1} capture-antigen solution (1 mg mL^{-1} for Mod-C6), and allowed to dry for 30 minutes at room temperature. The membrane was then submerged in 1% BSA in PBS solution for 30 min to block non-specific binding, and again dried for 10 min at 37 °C in a convection dry oven. The supporting layer (0.22 μm pore-size NC membrane) and vertical flow diffuser (0.45 μm pore-size) were also patterned with a wax printer, and the BSA blocking procedure was performed for these paper layers following the same procedure as for the multiplexed sensing membrane. The absorbent pad (1.2 × 1.2 cm), foam tape (1.7 × 1.7 cm for outside, 1.2 × 1.2 cm for inside dimensions), and asymmetric membrane (1.2 × 1.2 cm for absorption layer and 1st spreading layer and 1.4 × 1.4 cm for 2nd spreading layer) were all laser-cut (60W Speedy 100 CO_2 laser from Trotec) to achieve precise dimensions.

Assembly of the multiplexed vertical flow assay (xVFA)

A 3-D printed cassette, which opens, closes, and locks through a simple twisting mechanism was custom-developed for housing the vertical stack of paper layers. Prepared paper materials were stacked inside of the cassette as follows; for the top-case; asymmetric membrane (absorption layer), vertical flow diffuser, 1st spreading layer, conjugate pad (only for 2nd top-case), sample pad, asymmetric membrane (2nd spreading layer) and supporting layer, for the bottom case; multiplexed

sensing membrane and five absorbent pads (Figure 3.6). In case of absorption layer, the face with the smaller pore-size was contacted with lower layer, with the stacking reversed for the 1st and 2nd spreading layers. The outer edges of the supporting layer in the top-case and the multiplexed sensor in the bottom case were secured with foam tape to protect the sensing membrane from potential shifting or damage.

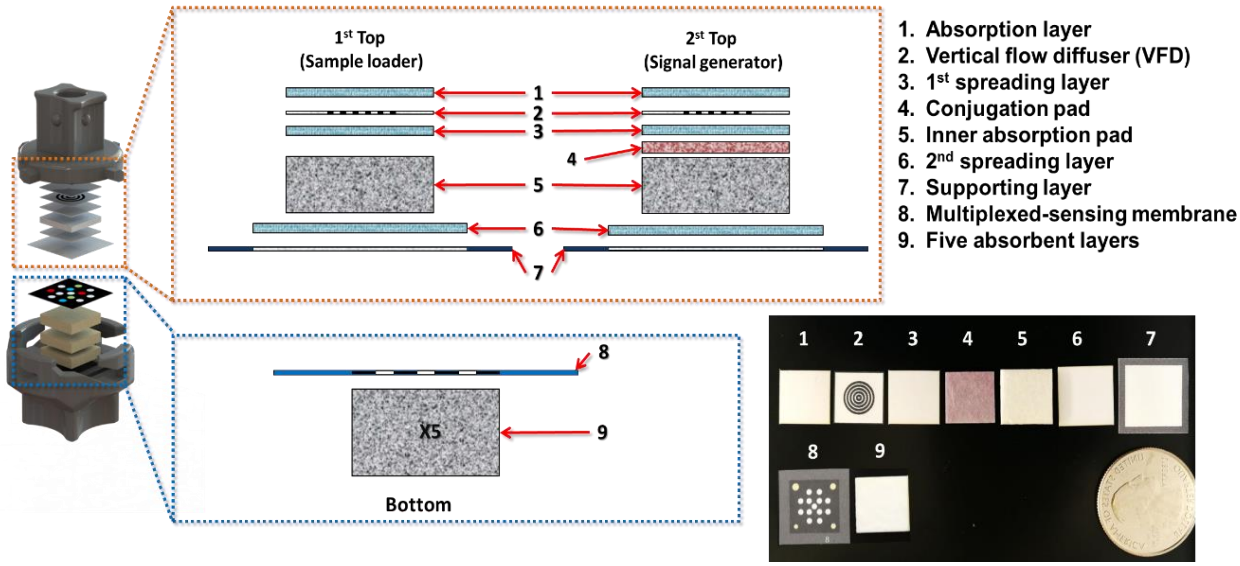


Figure 3.6. Expanded diagram of the multiplexed vertical flow assay (xVFA) detailing the different paper layers contained within the 3-D printed cassette (right). A quarter is show in the bottom right image for scale.

Characterization of the vertical fluid flow in xVFA design

The vertical flow properties of the xVFA were evaluated to better understand the underlying performance of the test as well as to optimize the operational protocol. Specifically, we investigated the signal strength versus the flow rate and the incubation time with a design constraint of 15 min assay time in order to be on-par with standard POC tests and lateral flow assays (LFAs).

Flow rate and signal intensity versus wax-coverage of sensing membrane

The vertical flow rate in the xVFA cassette is limited by the flow-through area (void area) of the multiplexed sensing membrane (i.e. the space absent from the wax-printed barrier) and as a result has a clear effect on the colorimetric signal intensity. To empirically understand this relationship, sensing membranes with four open reaction spots, yet different proportions of wax coverage (60%, 80%, 90%, and 95% of the 2-D membrane area) were tested in the xVFA (see Fig. 3.7A, B). By loading increments of fixed-volumes of the running buffer into the first top-case and recording the time it took for the fluid to absorb completely, the flow rate of the sensors was calculated for xVFA containing membranes with different wax coverage. The signal strength was then evaluated by the (1-R) signals (see Eq. 3.2, Section 3.2) produced by immunoreaction spots functionalized with the positive control proteins (Rabbit anti-Mouse IgG). The ratio of the control protein solution volume to each spot's area, 0.7 $\mu L\ mm^{-2}$, was conserved for the different sensing membranes.

At the beginning of the sample injection, the flow rate is at its fastest before it slows, converging to near constant flow rate that is limited by the flow-through area of the wax-printed sensing membrane. As shown in Figure 3.7B, the 95% wax printed membrane shows the highest signal intensity with the most steady flow rate over time. However, due the slower overall flow rate this sensing membrane design requires around 10 min to fully absorb the running buffer, sample, and washing buffer. In contrast, 60%, and 80% wax-coverage on the sensing membrane can complete absorption within 3 min, but with less uniformity in the sample flow over time. Therefore, due to this time-uniformity trade-off, the sensing membranes with 90% wax-coverage were utilized for the clinical study reported in sections 3.2-3.4 of this chapter.

Flow rate and signal intensity versus loading time

The flow rate and signal intensity (1-R) for the first top-case (sample loader) was recorded over time as described in the previous section (Fig. 3.7 C,D). The flow rate and signal intensity for the second top-case (signal generator) was also determined, however prior to analysis, the xVFA was assembled with the first-top-case and run according the standard protocol outline in the section 3.2 of this chapter. The xVFA cassette was then opened and the sampler top-case was exchanged with the signal generator top-case. The flow rate over time was recorded following the same procedure for the first top-case, with a sensing membrane containing 90% wax-coverage (Fig. 3.7 C-D).

As shown in Figure 3.7D, the sample loader top-case reached a steady flow rate and signal intensity after approximately 6 min, while it took nearly 8 min for the signal generator top-case. Based on these observations, we chose a 6 min and 8 min incubation time for the first and second top-case respectively to help support the flow and signal repeatability as well as rapid assay time (15 min).

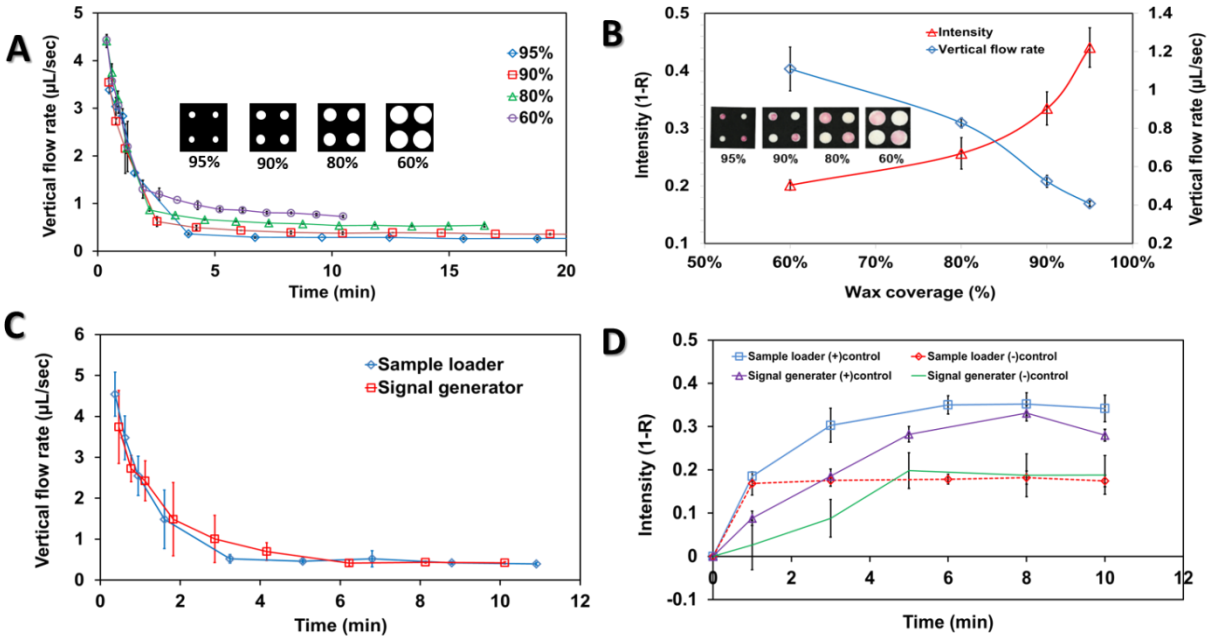


Figure 3.7. Vertical flow rate analysis of the xVFA. (A) The vertical flow rate at the loading inlet versus time shown for sensing membranes with different percentage of wax coverage. **(B)** Colorimetric signal intensity and vertical flow rate versus different percentage of wax coverage. **(C)** Vertical flow rate at the loading inlet verses time for the 1st top-case (sample loader) and 2nd top-case (signal generator) with a sensing membrane containing 90% wax coverage. **(D)** Colorimetric signal intensity versus time of the positive and negative control immunoreaction spots.

xVFA fabrication issues that were observed

In the second sample-pull testing, four IgM sensing membranes exhibited failed positive control spots, due to the accidental incorporation of non-functionalized (blank) sensing membranes during the xVFA assembly. This fabrication error was identified through qualitative observations at the time of testing, and was also self-evident in an analysis of the positive control spots distributions (see Fig. 3.8). As a result, these four xVFA tests were discarded from the testing analysis.

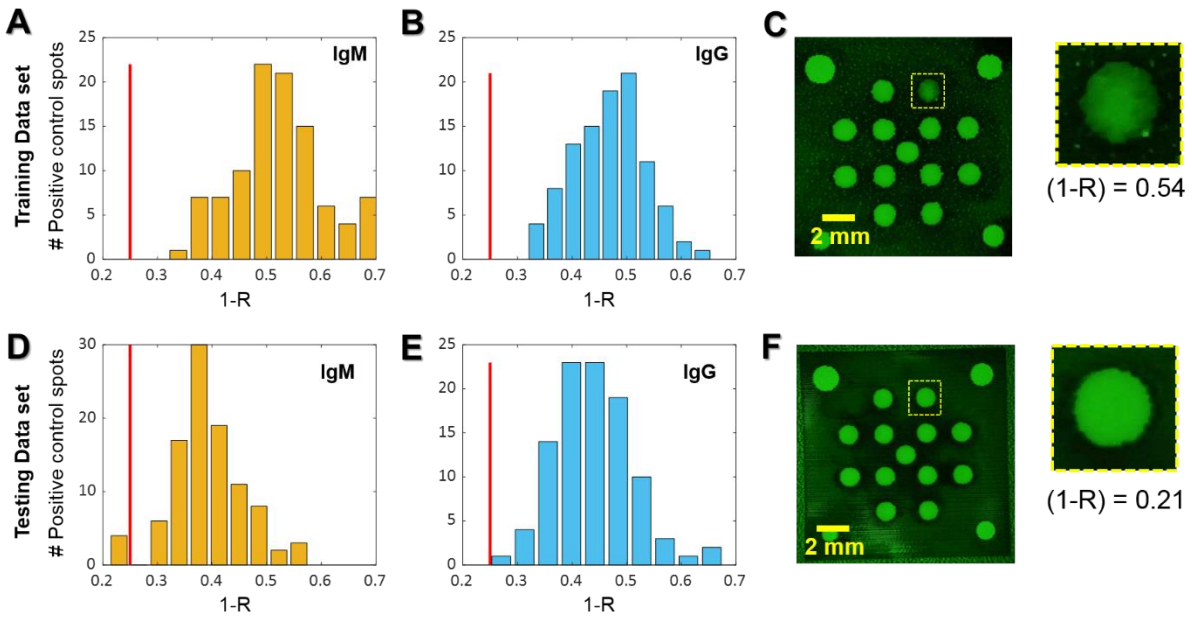


Figure 3.8. Failed xVFA tests due to fabrication error. (A) Histogram of the IgM and **(B)** IgG positive control immunoreaction spot over the training data-set. **(C)** An example image taken with the mobile phone reader of a sensing membrane with a working IgM control spot (right). **(D)** Histogram of the IgM and **(E)** IgG positive control immunoreaction spots over the blind testing data-set, showing 4 outliers (IgM shown in **D**) due to fabrication error. These control spot signals are below the test-failure cutoff denoted by the vertical red line at $(1-R)_{FAILURE} = 0.25$. **(F)** An example image of a sensing membrane with a failed (blank) IgM control spot (right), which does not show any nanoparticle-based binding signal; for comparison also see **C** for an example of a working IgM control spot.

Additionally, in the second sample-pull testing set, three IgM and three IgG sensing membranes had a scaling mismatch between the background image (taken before the assay) and signal image (taken after the assay) possibly due to an expansion of the sensing membrane during the assay operation. This scale mismatch in the registration can lead to misaligned immunoreaction spot sampling, as the background and signal images are sampled with the same mask. Therefore, all sensing membranes which exhibited a scaling-factor greater than 2% as defined by an affine-

mapping (six in total) were automatically re-processed with an affine-transformation for correct registration and spot analysis (see Fig. 3.9).

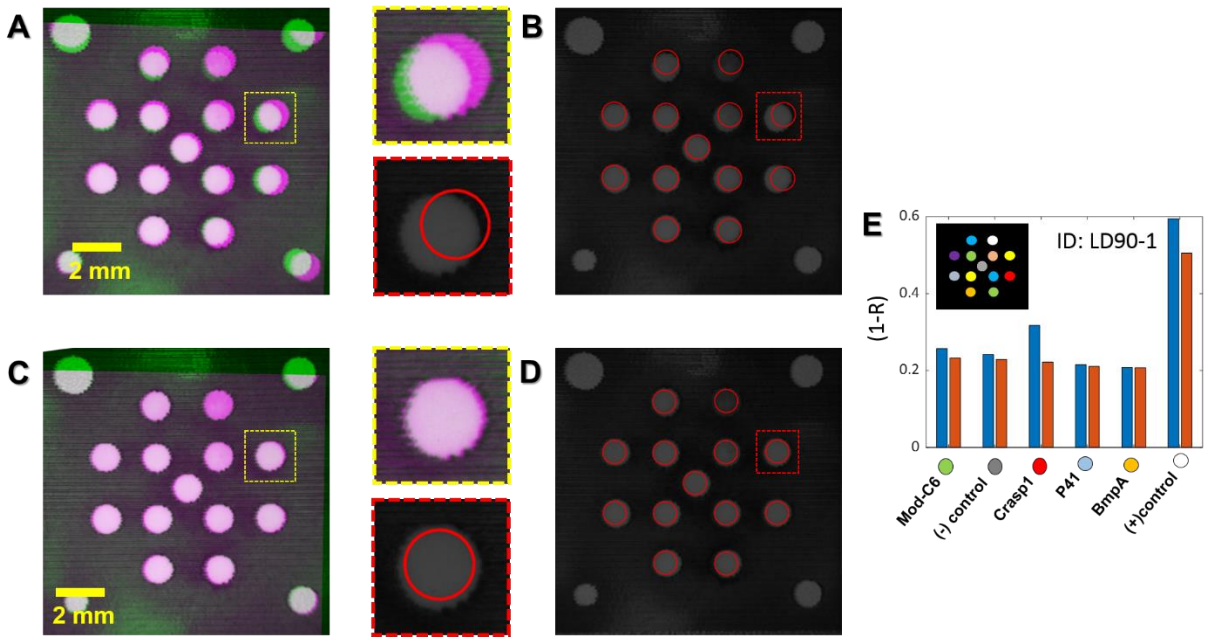


Figure 3.9. Misalignment error due to membrane expansion. (A) Example of failed image registration between the background and signal image using the rigid transformation. Here the background (before image) and signal (after image) are tinted pink and green respectively and overlaid to better illustrate the degree of misalignment. The inset (upper right) shows the misalignment of a single immunoreaction spot outlined by the dotted yellow line. **(B)** The red circles show the fixed-radius masks used for sampling the immunoreaction spots on the signal sensing membrane using the misaligned image. **(C)** Registered background and signal images after applying an affine transformation, which incorporates a scaling factor. **(D)** Correctly-sampled immunoreaction spots in the sensing membrane after the registration correction. **(E)** The (1-R) colorimetric signals of the immunoreaction spots before (blue) and after (red) the image registration correction. For simplicity, only the immunoreaction signals used as inputs to the final neural network are shown.

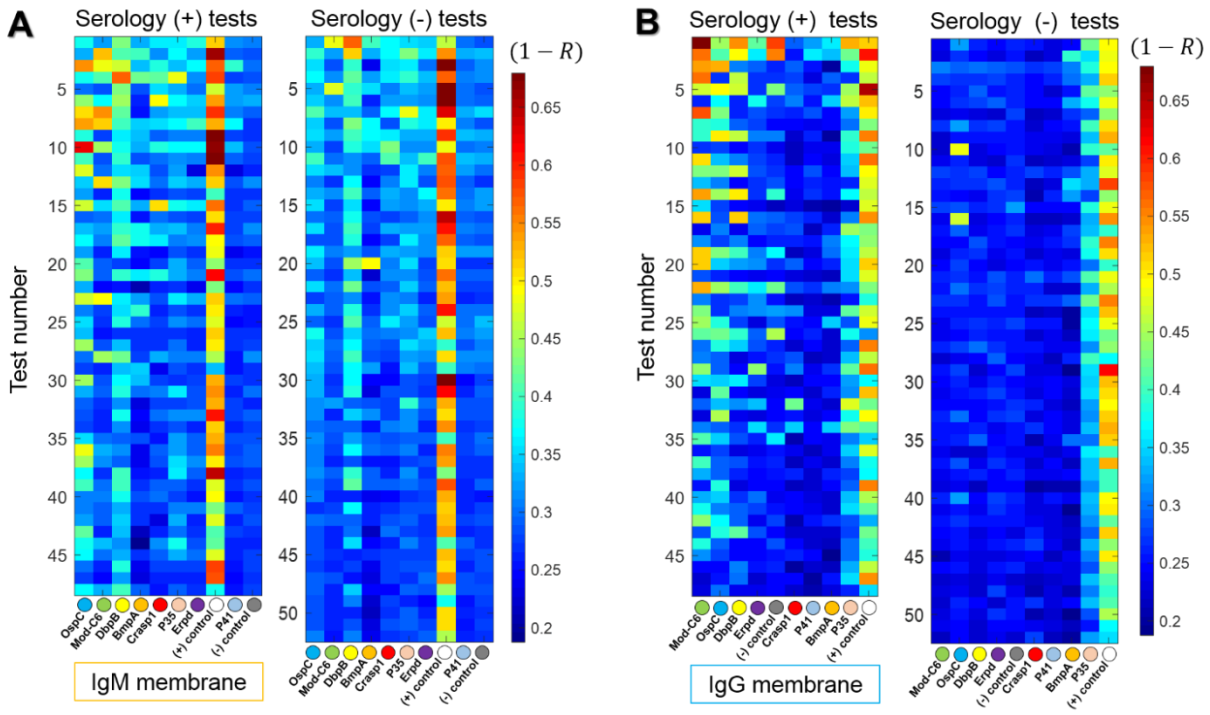


Figure 3.10. Heat-map of the raw colorimetric signals from the training data set. (A) The (1-R) signals from the IgM sensing membranes for the seropositive (left) and negative (right) tests. **(B)** The (1-R) signals for the IgG sensing membranes for the seropositive (left) and negative tests (right). The x-axis and y-axis for each heat-map is in the same order as Figure 3.3B of section 3.3 of this chapter.

Multi-antigen panel pre-selection

Prior to our clinical study, 15 clinical samples (8 cases and 7 endemic controls, also obtained from the Lyme Disease Biobank) were tested in duplicate to screen the following 9 antigens and on synthetic peptide (Mod-C6) for both IgM and IgG antibody detection: OspC, BmpA, P41, ErpD, Crasp1, OspA, DbpB, VlsE, P35 and Mod-C6. The 25-spot multiplexed-sensing membrane was employed for antigen screening with the antigens immobilized into the reaction spot following the same methods in the clinical study (see Fig. 3.11A). Samples loaded into the xVFA undergo complex immunoreactions that depend greatly on the conditions in the nitrocellulose matrix where the capture antigens are immobilized. Such an environment is significantly different from

conventional Enzyme-linked Immunosorbent Assays (ELISA) and Western blotting (WB) therefore, each antigen was assessed within the paper-based sensor before being implemented in the detection panel used for the clinical evaluation.

Figure 3.11 shows the results of antigen pre-screening. The t-score of each detection antigen was ranked, where the t-score was defined by Welch's t-test,

$$t_m = \frac{X_{m(+)} - X_{m(-)}}{\sqrt{\frac{sm_{(+)}{}^2}{N_{(+)}} + \frac{sm_{(-)}{}^2}{N_{(-)}}}} \qquad \text{Eq. 3.5}$$

Here X_m and s_m^2 represent the mean and variance of the $(1 - R_m)$ signal respectively for the seropositive and negative samples as denoted by the (+) and (-) subscripts. N represents the number of seropositive or negative samples. The capture antigens were therefore classified into three regimes: I, II, III (Fig. 3.11B). Regime III shows the worst discriminators, VlsE and OspA, exhibiting a t-score < 1. Despite OspA and VlsE being known to be effective detection antigens for LD, no signals were derived from their immunoreactions within the xVFA system. This could be due to poor antigen immobilization due to weak charge interaction as well as interference caused by the nitrocellulose matrix, as well as due to an absence of antibodies for VlsE and OspA in the samples tested. Therefore, these antigens that fell into Regime III were not included in the panel used for our clinical study. The antigen targets in Regime II were each kept as a single immunoreaction spots on the final 13 spot multiplexed-sensing membrane, while the antigen targets in Regime I, which exhibited the best discriminatory ability (t-score > 2.0), were therefore each given an additional (redundant) immunoreaction spot, in the place of the eliminated antigens from Regime III.

This strategy resulted in the final multi-antigen panel for the clinical study: two-spots in the panel for Regime I (OspC, DbpB and Mod-C6), and single spots for Regime II (Crasp1, Erpd, P41,

BmpA and P35) along with single spots for the positive and negative control proteins which were anti-mouse IgM/IgG and BSA respectively (Fig. 3.3A, Section 3.3).

It is important to note that the 10 antigens employed in this study, though commonly used for LD serology testing, are a subset of the much larger number of target antigens which have been evaluated and developed for LD diagnosis. Future iterations of the multi-antigen detection panel within the xVFA can therefore incorporate novel antigen targets produced through e.g. epitope mapping or other large-scale screening efforts.

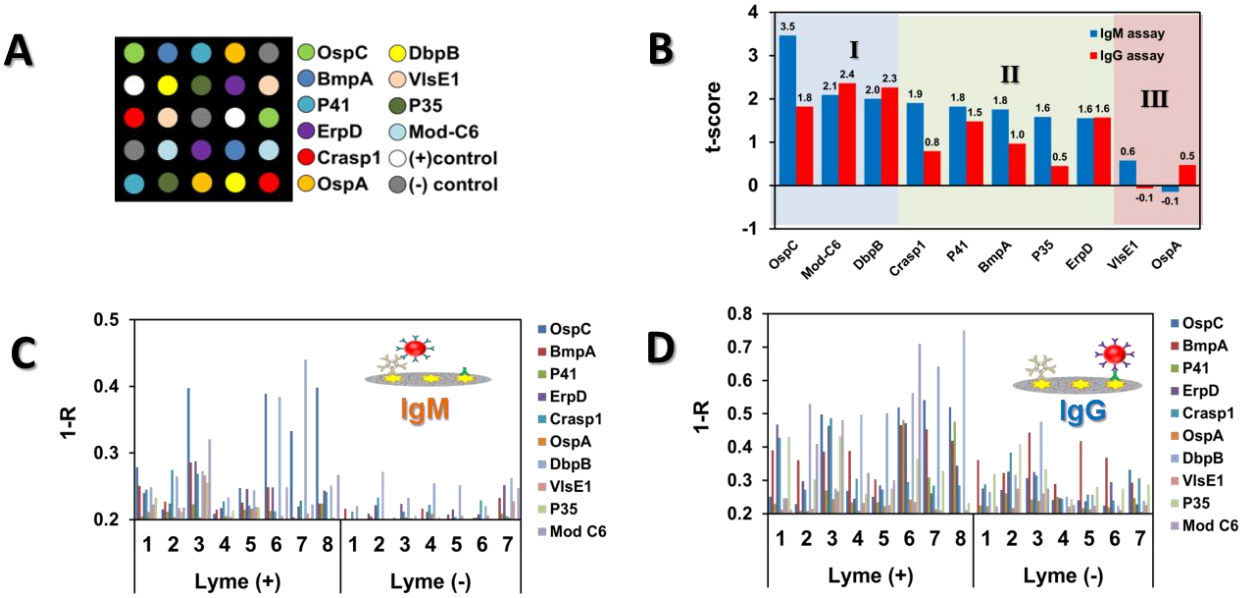


Figure 3.11. Pre-screening antigens for the xVFA sensing membrane. (A) Spot map for the antigen screening process including 10 capture antigens. **(B)** The t-scores (Eq. 3.2, Section 3.2) of the 10 antigens in the pre-screening panel. Three regimes are defined (shaded with different colors); antigens in regime I are given two spots within the testing antigen panel, antigens in regime II are given one spot in the panel, and antigens in regime III are not included in the testing antigen panel. **(C)** The IgM and **(D)** IgG sensing membrane colorimetric signal for each Lyme positive (left) and Lyme negative (right) human serum sample in the pre-screening cohort.

Precision Testing

A precision evaluation was performed using the same assay operation and analysis described above. Two additional seropositive samples and two endemic control samples, were blindly tested with six repeat 10:35measurements by the same operator. To test the inter-operability of the xVFA, two volunteers who had no experience with the xVFA operation were trained for 5 minutes by one of the authors (Dr. Hyou-Arm Joung) before each performing duplicate measurements with one seropositive sample and one endemic control.

The importance of the batch-specific standardization (see Section 3.4) is especially highlighted by this precision testing which similarly shows a drift in the network output without batch-specific standardization. However, by performing the input standardization with the mean and standard deviation of the batch used for precision testing, the overall accuracy during the precision testing results falls in line with the blind-testing performance, and showed no apparent instability between the operators (Table 3.7).

The individual immunoreactions were also examined in terms of their precision (Fig. 3.12, 3.13) as discussed in the section 3.4 of this chapter.

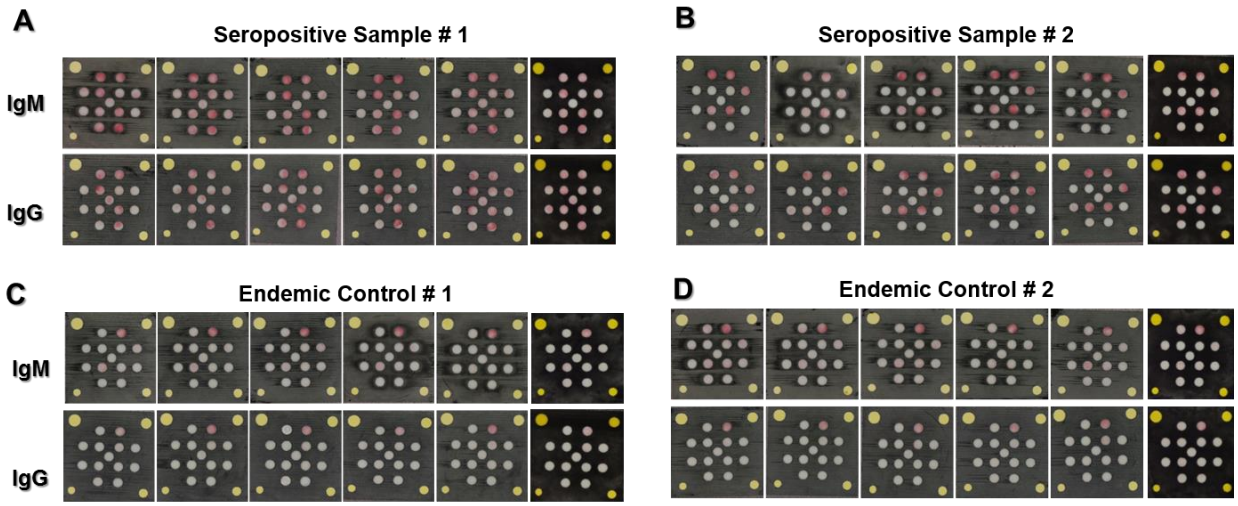
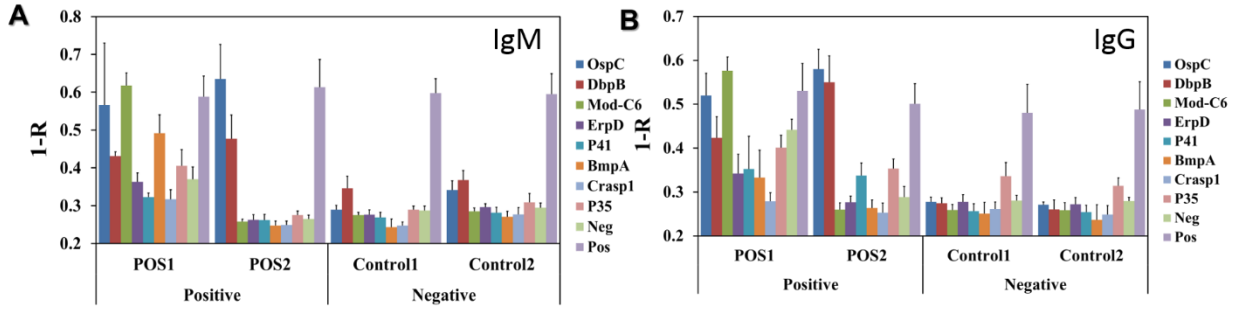


Figure 3.12. Real pictures of sensing membranes during precision study (A)-(D) Pictures (under ambient light conditions) of the IgM and IgG sensing membranes activated by two seropositive samples and two endemic controls, with six repeat measurements for each sample performed by the same operator.



C CV(%)

0% 10% 20%

CV(%)			OspC	DbpB	Mod-C6	ErpD	P41	BmpA	Crasp1	P35	Neg	Pos
IgM	Positive	POS1	31.7%	3.0%	5.9%	7.5%	4.0%	11.0%	9.3%	12.0%	10.4%	10.1%
		POS2	15.5%	14.6%	3.0%	6.4%	7.1%	6.3%	5.5%	4.8%	4.8%	13.1%
	Negative	Control1	4.6%	11.0%	3.3%	5.8%	6.3%	11.6%	4.7%	3.8%	5.0%	7.0%
		Control2	8.5%	7.9%	3.7%	3.7%	6.2%	6.5%	8.0%	9.1%	5.0%	9.9%
IgG	Positive	POS1	10.8%	13.0%	6.0%	14.9%	24.7%	21.9%	8.8%	8.0%	6.2%	13.0%
		POS2	8.5%	12.1%	7.1%	6.2%	9.9%	8.7%	10.9%	7.1%	9.9%	10.2%
	Negative	Control1	4.9%	5.7%	6.2%	7.4%	8.0%	13.0%	7.4%	10.8%	5.4%	15.2%
		Control2	2.9%	10.3%	8.5%	6.9%	7.4%	18.4%	10.1%	7.0%	3.7%	14.5%

D Average CV (%)

	IgM	IgG
(+)	9.3%	10.9%
(−)	6.6%	8.7%

Figure 3.13. Precision analysis of the immunoreactions of the multi-antigen panel. (A) The mean (1-R) signals and standard deviations of the two seropositive (POS1 and POS2) and two seronegative (Control 1 and 2) for the IgM and **(B)** IgG sensing membrane. **(C)** The coefficient of variation (%CV) of each immunoreaction spot for the repeated tests. (**D)** Summary of the %CV over all the immunoreaction spots broken down by the seropositive and seronegative samples as well as the IgM and IgG sensing membrane.

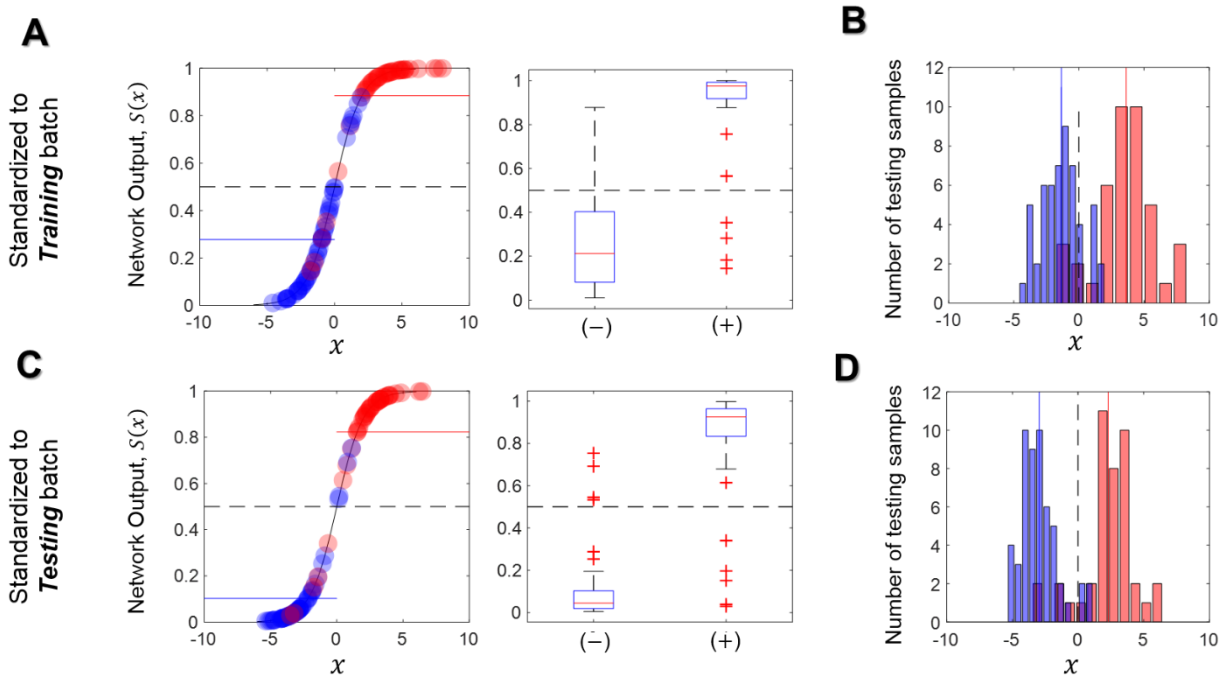


Figure 3.14. The network output from the initial blind testing set comparing different batch-specific standardizations. The network output shown, S(x), is the sigmoid activation function where x represents the output of the final hidden layer of the network before this final nonlinear activation. **(A)** The network output from the blind testing set where the inputs are standardized to the mean and standard deviation over the *training* set. The blue and red markers represent the seronegative and positive samples respectively and the solid blue and red lines (left plot) indicate the mean network output of the seronegative and positive samples, respectively. The boxplot to the right shows the median (solid red line) of the network output of the seronegative and positive samples, with the blue box representing the 25th and 75th percentile, and the black whiskers designating the most extreme data points not considered as outliers. The outliers are shown with the red crosses. **(B)** Histogram of the network output, x, before the nonlinear activation S(x) where the solid blue and red bars represent the mean value for the seronegative and positive samples respectively. **C-D)** A comparison to (A) and (B) of the blind testing set where the inputs are standardized to the mean and standard deviation over the *testing* set.

Supplemental Tables

Table 3.1. Itemized list of the different paper layers included in the xVFA along with their per-test cost. The total material cost (excluding the plastic cassette) is $0.42 per test.

No.	Layers	Product name	Vendor	Thnickness (mm)	Dimension (cm x cm)	Cost/test(c)
1	Absorption layer	Vivid GX™	Pall cooperation	0.3	1.2 x 1.2	1.8
2	Vertical flow diffuser (VFD)	0.45 µm Nitrocellulose Blotting Membrane	Sartorius Corporation	0.1	1.2 x 1.2	5.4
3	1st sample sreading layer	Vivid GX™	Pall cooperation	0.3	1.2 x 1.2	1.8
4	Conjugation pad	Standard 17	GE Healthcare Biosciences Corp	0.37	1.15 x 1.15	1.6
5	Inter pad	CF7	GE Healthcare Biosciences Corp	1.87	1.2 x 1.2	3.5
6	2nd sample sperading layer	Vivid GX™	Pall cooperation	0.3	1.4 x 1.4	2.5
7	Supporting layer	0.45 µm Nitrocellulose Blotting Membrane	Sartorius Corporation	0.1	1.7 x 1.7	10.7
8	Antigen panel	0.22 µm Nitrocellulose Blotting Membrane	Sartorius Corporation	0.1	1.7 x 1.7	5.4
9	Absorbent layer (x5)	Grade 707 Blotting Pad	OpticsPlanet, Inc.	2.6	1.2x 1.2	9.5
10	Foam tape	Food-greade foam tape	MacMaster	1.6	1.7x1.7	1.3
				Cost total/test		42.2

Table 3.2. Itemized list of the different reagents employed in the xVFA along with their per-test cost. The modified-C6 peptide was synthesized at Biopeptides Corp. and was provided without charge as part of a collaboration. For Figure 3.4 in the Section 3.3, the Mod-C6 cost was estimated to be $0.51 per test, which is the average per-test cost of the other 7 antigens included in the testing panel.

Name		Vender	Cat. No.	Quantity	Cost($)	Cost($)/Test
Gold nanoparticles		Ted Pella, Inc.	15707-1	100 mL	274	0.21
Anti-human Ab	Anti-IgG	abcam	ab99770	250ug	300	1.80
	Anti-IgM	abcam	ab99741	250ug	300	
Antigens	OspC	PROSPEC	BOR-004	1000 ug	3600	0.58
	DbpB	PROSPEC	BOR-007	1000 ug	3600	0.58
	Mod-C6					
	ErpD	Rockland Immunochemicals Inc.	000-001-C09	100ug	325	0.28
	Crasp1	Rockland Immunochemicals Inc.	000-001-C18	100ug	325	0.28
	BmpA	PROSPEC	BOR-006	20 ug	120	0.14
	P41	PROSPEC	BOR-001	100ug	150	0.13
	P35	Rockland Immunochemicals Inc.	000-001-C12		325	0.28
2nd antibody(anti-mouse IgG)		SouthernBiotech	1036-01	1mg	45	0.0
					Sub total	4.28

Table 3.3. Clinical testing results of the serum samples which constitute the training data set (i.e. samples from the first sample-pull). All samples were provided by the Lyme Disease Biobank, and collected at a clinical site in East

Hampton. All testing was performed at Stony Brook labs. Abbreviations: **EM**- Erythema Migrans rash, **WB -** Western Blot, **BL**- borderline, **EQV**- Equivocal, **POS**- positive, **NEG**-negative, **IT**- indeterminate, i.e. a WB that contains specific bands but does not meet the positive CDC criteria, **NR**- Non-Reactive. Note that 'Two-tier Positive' corresponds to the standard testing using EIA and Western Blotting in the first and second tier, respectively, whereas 'Serology Positive' encompasses the modified two-tier testing (MTTT) definition described in section 3.2.

			First Tier			Second Tier					
	Sample Label	EM > 5 cm at Enrollment	Whole Cell Lysate ELISA	C-6 Peptide ELISA	VlsE/ PepC10 ELISA	WB IgM	WB IgG	Two-tier Positive ?	Serology Positive?	Primary Method of Confirmation	Final Lyme Diagnosis
1	LD057	Yes	BL	POS	NA	IT	IT	No	Yes	2 pos ELISA, > 5 cm EM	Confirmed Lyme
2	LD083	Yes	RE	POS	NA	IT	IT	No	Yes	2 pos ELISA, > 5cm EM	Confirmed Lyme
3	LD095	Yes	RE	NEG	NA	POS	IT	Yes	Yes	2-tier serology, culture qPCR	Confirmed Lyme
4	LD051	No	RE	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
5	LD052	Yes	NR	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
6	LD053	Yes	NR	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
7	LD062	Yes	RE	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
8	LD063	No	BL	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
9	LD064	Yes	RE	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
10	LD065	Yes	NR	EQV	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
11	LD067	Yes	RE	POS	NA	IT	POS	Yes	Yes	2-tier serology	Confirmed Lyme
12	LD069	Yes	NR	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
13	LD072	Yes	RE	POS	NA	IT	POS	Yes	Yes	2-tier serology	Confirmed Lyme
14	LD073	No	RE	POS	NA	IT	POS	Yes	Yes	2-tier serology	Confirmed Lyme
15	LD074	Yes	NR	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
16	LD076	NA	RE	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
17	LD080	NA	RE	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
18	LD082	Yes	NR	POS	NA	POS	POS	Yes	Yes	2-tier serology	Confirmed Lyme
19	LD084	Yes	NR	EQV	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
20	LD090	No	NR	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
21	LD092	Yes	RE	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
22	LD094	Yes	NR	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
23	LD096	NA	NR	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
24	LD098	Yes	BL	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme
25	LD055	NA	BL	NEG	NA	NEG	IT	No	No	qPCR	Confirmed Lyme
26	LD054	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)
27	LD056	NA	NR	NEG	NA	NEG	IT	No	No	NA	Control (- Serology)
28	LD058	NA	NR	NEG	NA	NEG	IT	No	No	NA	Control (- Serology)
29	LD059	NA	NR	NEG	NA	NEG	IT	No	No	NA	Control (- Serology)
30	LD060	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)
31	LD061	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)
32	LD066	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)
33	LD068	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)
34	LD070	NA	NR	NEG	NA	NEG	IT	No	No	NA	Control (- Serology)
35	LD071	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)
36	LD075	NA	NR	NEG	NA	NEG	IT	No	No	NA	Control (- Serology)
37	LD077	NA	NR	NEG	NA	NEG	IT	No	No	NA	Control (- Serology)
38	LD078	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)
39	LD079	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)
40	LD081	NA	NR	NEG	NA	NEG	IT	No	No	NA	Control (- Serology)
41	LD085	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)
42	LD086	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)
43	LD087	NA	NR	NEG	NA	NEG	IT	No	No	NA	Control (- Serology)
44	LD088	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)
45	LD089	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)
46	LD091	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)
47	LD093	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)
48	LD097	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)
49	LD099	NA	NR	NEG	NA	IT	NEG	No	No	NA	Control (- Serology)
50	LD100	NA	BL	NEG	NA	IT	IT	No	No	NA	Control (- Serology)

Table 3.4. Clinical testing results of the serum samples which constitute the blind testing data-set (i.e. samples from the second sample-pull). All samples were provided by the Lyme Disease Biobank, and collected at clinical sites in East Hampton as well as Wisconsin. All testing for the samples collected from East Hampton was performed at Stony Brook labs and samples collected from Wisconsin were performed at the Mayo clinic. The initial, blinded network predictions are shown in the right most columns, with the grey boxes corresponding to the 4 xVFA tests with failed IgM control spots. Abbreviations: **EM**- Erythema Migrans rash, **WB -** Western Blot, *i.e.* a WB that contains specific bands but does not meet the positive CDC criteria**, BL**- borderline, **EQV**- Equivocal, **POS**- positive, **NEG**- negative, **IT**- indeterminate, **NR**- Non-Reactive. Note that 'Two-tier Positive' corresponds to the standard testing using EIA and Western Blotting in the first and second tier, respectively, whereas 'Serology Positive' encompasses the modified two-tier testing (MTTT) definition described in section 3.2.

			First Tier			Second Tier						Network Predictions			
	Sample Label	EM > 5 cm	Whole Cell Lysate ELISA	C-6 Peptide ELISA	VlsE/ PepC10 ELISA	WB IgM	WB IgG	Two-tier Positive?	Serology Positive?	Primary Method of Confirmation	Final Lyme Diagnosis	NN output (1 of 2)	Lyme Diagnosis (1 of 2)	NN output (2of 2)	Lyme Diagnosis (2 of 2)
1	LD185	Yes	NA	POS	POS	NEG	NEG	No	Yes	2 pos ELISA, > 5 cm EM	Confirmed Lyme	0.898	TRUE	0.964	TRUE
2	LD158	Yes	RE	POS	NA	IT	IT	No	Yes	2 pos ELISA, > 5 cm EM	Confirmed Lyme	0.945	TRUE	0.963	TRUE
3	LD138	Yes	NA	POS	NA	POS	NEG	Yes	Yes	2-tier serology	Confirmed Lyme	0.990	TRUE	0.993	TRUE
4	LD134	NA	NA	NA	POS	POS	NEG	Yes	Yes	2-tier serology	Confirmed Lyme	0.980	TRUE	0.960	TRUE
5	LD136	Yes	RE	POS	NA	NEG	POS	Yes	Yes	2-tier serology	Confirmed Lyme	0.565	TRUE	0.918	TRUE
6	LD145	Yes	RE	POS	NA	IT	POS	Yes	Yes	2-tier serology	Confirmed Lyme	0.352	FALSE	0.282	FALSE
7	LD148	NA	NA	NA	POS	POS	POS	Yes	Yes	2-tier serology	Confirmed Lyme	0.992	TRUE	0.981	TRUE
8	LD149	Yes	RE	POS	NA	POS	POS	Yes	Yes	2-tier serology	Confirmed Lyme	1.000	TRUE	0.995	TRUE
9	LD154	Yes	RE	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme	1.000	TRUE	1.000	TRUE
10	LD160	Yes	NA	POS	NA	NEG	POS	Yes	Yes	2-tier serology	Confirmed Lyme	0.975	TRUE	0.992	TRUE
11	LD162	Yes	NA	POS	POS	POS	NEG	Yes	Yes	2-tier serology	Confirmed Lyme	0.757	TRUE	0.911	TRUE
12	LD163	Yes	NA	POS	POS	POS	NEG	Yes	Yes	2-tier serology	Confirmed Lyme	0.998	TRUE	0.978	TRUE
13	LD164	Yes	NA	POS	POS	POS	NEG	Yes	Yes	2-tier serology	Confirmed Lyme	0.939	TRUE	0.940	TRUE
14	LD167	Yes	NA	POS	POS	POS	NEG	Yes	Yes	2-tier serology	Confirmed Lyme	0.986	TRUE		
15	LD169	No	NA	NEG	POS	POS	NEG	Yes	Yes	2-tier serology	Confirmed Lyme	0.878	TRUE		
16	LD170	Yes	NA	POS	POS	POS	NEG	Yes	Yes	2-tier serology	Confirmed Lyme			0.959	TRUE
17	LD171	Yes	NA	POS	POS	POS	POS	Yes	Yes	2-tier serology	Confirmed Lyme			0.999	TRUE
18	LD177	Yes	NA	POS	POS	POS	NEG	Yes	Yes	2-tier serology	Confirmed Lyme	0.986	TRUE	0.995	TRUE
19	LD184	Yes	NA	POS	POS	POS	POS	Yes	Yes	2-tier serology	Confirmed Lyme	0.994	TRUE	0.996	TRUE
20	LD186	NA	NA	EQV	POS	POS	NEG	Yes	Yes	2-tier serology	Confirmed Lyme	0.953	TRUE	0.928	TRUE
21	LD187	NA	NA	POS	POS	POS	POS	Yes	Yes	2-tier serology	Confirmed Lyme	0.984	TRUE	0.993	TRUE
22	LD188	NA	NA	NEG	EQV	POS	NEG	Yes	Yes	2-tier serology	Confirmed Lyme	0.145	FALSE	0.182	FALSE
23	LD192	Yes	BL	POS	NA	POS	IT	Yes	Yes	2-tier serology	Confirmed Lyme	0.907	TRUE	0.982	TRUE
24	LD168	Yes	NA	NEG	NEG	NEG	NEG	No	No	Seroconversion	Confirmed Lyme	0.498	FALSE	0.334	FALSE
25	LD172	NA	NA	NEG	POS	NEG	NEG	No	No	Seroconversion	Confirmed Lyme	0.190	FALSE	0.072	FALSE
26	LD133	NA	NA	NA	EQV	NEG	NEG	No	No	NA	Control (- Serology)	0.479	FALSE	0.152	FALSE
27	LD137	NA	NA	NEG	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.494	FALSE	0.168	FALSE
28	LD141	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)	0.226	FALSE	0.098	FALSE
29	LD142	NA	NA	NEG	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.329	FALSE	0.383	FALSE
30	LD144	NA	NA	NEG	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.398	FALSE	0.760	TRUE
31	LD147	NA	NA	NA	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.801	TRUE	0.063	FALSE
32	LD150	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)	0.785	TRUE	0.084	FALSE
33	LD151	NA	NR	NEG	NA	NEG	IT	No	No	NA	Control (- Serology)	0.138	FALSE	0.200	FALSE
34	LD152	NA	NA	NEG	NA	NEG	NEG	No	No	NA	Control (- Serology)	0.241	FALSE	0.028	FALSE
35	LD155	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)	0.117	FALSE	0.314	FALSE
36	LD156	NA	NA	NEG	NA	NEG	NEG	No	No	NA	Control (- Serology)	0.288	FALSE	0.081	FALSE
37	LD157	NA	NR	NEG	NA	IT	IT	No	No	NA	Control (- Serology)	0.064	FALSE	0.152	FALSE
38	LD165	NA	NA	NEG	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.286	FALSE	0.403	FALSE
39	LD166	NA	NA	NEG	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.119	FALSE	0.280	FALSE
40	LD173	NA	NA	NEG	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.061	FALSE	0.771	TRUE
41	LD174	NA	NA	NEG	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.415	FALSE	0.284	FALSE
42	LD175	NA	NA	NEG	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.850	TRUE	0.276	FALSE
43	LD179	NA	NA	NEG	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.027	FALSE	0.010	FALSE
44	LD180	NA	NA	NEG	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.100	FALSE	0.103	FALSE
45	LD182	NA	NA	NEG	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.018	FALSE	0.045	FALSE
46	LD183	NA	NA	NEG	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.878	TRUE	0.708	TRUE
47	LD189	NA	NA	NEG	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.029	FALSE	0.031	FALSE
48	LD190	NA	NA	NEG	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.433	FALSE	0.148	FALSE
49	LD191	NA	NA	NEG	NEG	NEG	NEG	No	No	NA	Control (- Serology)	0.027	FALSE	0.061	FALSE
50	LD193	NA	NR	NEG	NA	NEG	IT	No	No	NA	Control (- Serology)	0.476	FALSE	0.225	FALSE

Table 3.5. Comparison between the xVFA and the recently FDA cleared POC Sofia Lyme test from Quidel, which can be used as a first tier test, but is *not* recommended as a replacement of the two-tier testing. **(A)** Initial blind testing results using our xVFA (i.e. using a diagnostic threshold of 0.5) as well as blind testing results of the xVFA when implementing batch-specific standardization and a threshold of 0.66 determined during the training phase. The right most column shows the performance of the standard (CDC recommended) two-tier serology (i.e. with WB in the second tier). All sensitivities and specificities reported here are with respect to the ultimate clinical diagnosis (which

includes diagnosis through qPCR and seroconversion, in addition to the standard and modified two-tier testing methods). **(B)** Blind testing results using a CDC panel of human sera, reported in the Quidel Sofia Lyme package insert (Sofia Lyme FIA | Quidel. Available at: https://www.quidel.com/immunoassays/rapid-lyme-tests/sofia-lyme-fia. (Accessed: 13th September 2019).

A

		xVFA *(threshold = 0.5, Blind testing)*			xVFA *(threshold = 0.66, Batch standardization)*			Standard Two-Tier Serology			
	N*	Pos	Neg	% Agreement with clinical status	Pos	Neg	% Agreement with clinical status	N	Pos	Neg	% Agreement with clinical status
Endemic Controls	50	7	43	86.0%	2	48	96.0%	25	0	25	100%
Early Lyme Positive	46	38	8	82.6%	36	10	78.3%	25	21	4	84.0%

B

		Sofia Lyme IgM			Sofia Lyme IgG		
	N	Pos	Neg	% Agreement with clinical status	Pos	Neg	% Agreement with clinical status
Endemic Controls	190	33	157	82.6%	25	165	86.8%
Early Lyme Positive	60	49	11	81.7%	49	11	81.7%

Table 3.6. Comparison of the initial blind testing xVFA performance (i.e. with a threshold value of 0.5) to blind testing with batch-specific standardization and threshold of 0.66 determined during the training phase. A comparison is also made between the performance compared to two-tier testing serology (including the MTTT method) as well as the ultimate clinical diagnosis (confirmed through Two-Tier testing, PCR, or seroconversion).

			xVFA *(threshold = 0.5, Blind testing)*			xVFA *(threshold = 0.66, Batch standardization)*		
		N	**Pos**	**Neg**	**% Agreement with clinical status**	**Pos**	**Neg**	**% Agreement with clinical status**
Two Tier testing	***Seronegative* controls**	54	7	47	87.0%	2	52	96.3%
	Early Lyme *Seropositive*	42	38	4	90.5%	36	6	85.7%
Two-tier testing + PCR + Seroconversion	**Negative controls**	50	7	43	86.0%	2	48	96.0%
	Early Lyme Positive	46	38	8	82.6%	36	10	78.3%

Table 3.7. Precision testing results of three seropositive samples and three endemic controls repeated 6 times and by different operators. Here the inputs to the network were standardized to the mean and standard deviation of the batch of sensors used for precision testing, and a decision threshold of 0.66 was used as determined during the training phase.

	POC User	$\frac{TP}{TP+FN}$	Network output (mean ± std. dev.)
LD Case #1	Operator A	6/6	0.997 ± 0.003
LD Case #2	Operator A	4/6	0.652 ± 0.445
LD Case #3	Operator A,B,C	6/6	0.947 ± 0.120

	POC User	$\frac{TN}{TN+FP}$	Network output (mean ± std. dev.)
Endemic Control #1	Operator A	6/6	0.139 ± 0.221
Endemic Control #2	Operator A	6/6	0.045 ± 0.088
Endemic Control #3	Operator A,B,C	5/6	0.265 ± 0.423

Table 3.8. Comparison of the performance of different antigen detection panels during the training and testing phase. The number of False Negatives (FN) and False Positives (FP) in the testing batches is displayed in parenthesis.

Training Phase ($N_- = 52,\ N_+ = 48$)			
	Selected Antigen Panel (M = 9)	**Full Antigens (M=20)**	**C6 IgG Only (M =1)**
Sensitivity (FN)	91.7% (4)	72.9%(13)	77.1% (11)
Specificity (FP)	96.2% (2)	92.3%(4)	98.1%(1)
AUC	0.969	0.894	0.850

Testing Phase ($N_- = 54,\ N_+ = 42$)			
	Selected Antigen Panel (M = 9)	**Full Antigens (M=20)**	**C6 IgG Only (M =1)**
Sensitivity (FN)	90.5% (4)	95.2%(2)	73.8% (11)
Specificity (FP)	87.0% (7)	42.6%(31)	98.1%(1)
AUC	0.950	0.908	0.887

References

[1] C. A. Nelson *et al.*, "Incidence of Clinician-Diagnosed Lyme Disease, United States, 2005–2010 - Volume 21, Number 9—September 2015 - Emerging Infectious Diseases journal - CDC."

[2] G. Stanek, G. P. Wormser, J. Gray, and F. Strle, "Lyme borreliosis," *The Lancet*, vol. 379, no. 9814, pp. 461–473, Feb. 2012
.

[3] G. P. Wormser *et al.*, "The Clinical Assessment, Treatment, and Prevention of Lyme Disease, Human Granulocytic Anaplasmosis, and Babesiosis: Clinical Practice Guidelines by the Infectious Diseases Society of America," *Clin Infect Dis*, vol. 43, no. 9, pp. 1089–1134, Nov. 2006.

[4] "Lyme borreliosis - The Lancet." [Online]. Available: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(11)60103-7/fulltext. [Accessed: 29-Aug-2019].

[5] A. C. Steere and V. K. Sikand, "The Presenting Manifestations of Lyme Disease and the Outcomes of Treatment," *New England Journal of Medicine*, vol. 348, no. 24, pp. 2472–2474, Jun. 2003.

[6] G. P. Wormser, "Early Lyme Disease," *New England Journal of Medicine*, vol. 354, no. 26, pp. 2794–2801, Jun. 2006.

[7] E. D. Shapiro, "Lyme Disease," *New England Journal of Medicine*, vol. 370, no. 18, pp. 1724–1731, May 2014.

[8] D. Adukauskienė, A. Čiginskienė, A. Adukauskaitė, D. Pentiokinienė, R. Šlapikas, and I. Čeponienė, "Clinical relevance of high sensitivity C-reactive protein in cardiology," *Medicina*, vol. 52, no. 1, pp. 1–10, Jan. 2016.

[9] A. Moore, C. Nelson, C. Molins, P. Mead, and M. Schriefer, "Current Guidelines, Common Clinical Pitfalls, and Future Directions for Laboratory Diagnosis of Lyme Disease, United States," *Emerging Infect. Dis.*, vol. 22, no. 7, 2016.

[10] S. E. Schutzer *et al.*, "Direct Diagnostic Tests for Lyme Disease," *Clin. Infect. Dis.*, vol. 68, no. 6, pp. 1052–1057, Mar. 2019.

[11] M. Alasel and M. Keusgen, "Promising alternatives for one-tier testing of Lyme borreliosis," *Clin. Chim. Acta*, vol. 479, pp. 148–154, Apr. 2018.

[12] F. Dressler, J. A. Whalen, B. N. Reinhardt, and A. C. Steere, "Western blotting in the serodiagnosis of Lyme disease," *J. Infect. Dis.*, vol. 167, no. 2, pp. 392–400, Feb. 1993.

[13] G. P. Wormser, A. Levin, S. Soman, O. Adenikinju, M. V. Longo, and J. A. Branda, "Comparative cost-effectiveness of two-tiered testing strategies for serodiagnosis of lyme disease with noncutaneous manifestations," *J. Clin. Microbiol.*, vol. 51, no. 12, pp. 4045–4049, Dec. 2013.

[14] E. S. Theel, "The Past, Present, and (Possible) Future of Serologic Testing for Lyme Disease," *J. Clin. Microbiol.*, vol. 54, no. 5, pp. 1191–1196, 2016.

[15] G. P. Wormser *et al.*, "A limitation of 2-stage serological testing for Lyme disease: enzyme immunoassay and immunoblot assay are not independent tests," *Clin. Infect. Dis.*, vol. 30, no. 3, pp. 545–548, Mar. 2000.

[16] P. Mead, "Updated CDC Recommendation for Serologic Diagnosis of Lyme Disease," *MMWR Morb Mortal Wkly Rep*, vol. 68, 2019.

[17] A. R. Marques, "Revisiting the Lyme Disease Serodiagnostic Algorithm: the Momentum Gathers," *J. Clin. Microbiol.*, vol. 56, no. 8, 2018.

[18] A. R. Marques, "Laboratory Diagnosis of Lyme Disease - Advances and Challenges," *Infect Dis Clin North Am*, vol. 29, no. 2, pp. 295–307, Jun. 2015.

[19] M. M. G. Leeflang *et al.*, "The diagnostic accuracy of serological tests for Lyme borreliosis in Europe: a systematic review and meta-analysis," *BMC Infectious Diseases*, vol. 16, p. 140, Mar. 2016.

[20] A. J. Nowalk, R. D. Gilmore, and J. A. Carroll, "Serologic Proteome Analysis of Borrelia burgdorferi Membrane-Associated Proteins," *Infection and Immunity*, vol. 74, no. 7, pp. 3864–3873, Jul. 2006.

[21] L. A. Magnarelli, J. W. Ijdo, S. J. Padula, R. A. Flavell, and E. Fikrig, "Serologic diagnosis of Lyme borreliosis by using enzyme-linked immunosorbent assays with recombinant antigens," *J. Clin. Microbiol.*, vol. 38, no. 5, pp. 1735–1739, May 2000.

[22] P. Lahdenne *et al.*, "Improved serodiagnosis of erythema migrans using novel recombinant borrelial BBK32 antigens," *J. Med. Microbiol.*, vol. 52, no. Pt 7, pp. 563–567, Jul. 2003.

[23] S. M. Engstrom, E. Shoop, and R. C. Johnson, "Immunoblot interpretation criteria for serodiagnosis of early Lyme disease," *J. Clin. Microbiol.*, vol. 33, no. 2, pp. 419–427, Feb. 1995.

[24] A. F. Hinckley *et al.*, "Lyme Disease Testing by Large Commercial Laboratories in the United States," *Clin Infect Dis*, vol. 59, no. 5, pp. 676–681, Sep. 2014.

[25] M. J. C. Gomes-Solecki *et al.*, "A First-Tier Rapid Assay for the Serodiagnosis of Borrelia burgdorferi Infection," *Arch Intern Med*, vol. 161, no. 16, pp. 2015–2020, Sep. 2001.

[26] A. M. López-Marzo and A. Merkoçi, "Paper-based sensors and assays: a success of the engineering design and the convergence of knowledge areas," *Lab Chip*, vol. 16, no. 17, pp. 3150–3176, Aug. 2016.

[27] S. Smith, J. G. Korvink, D. Mager, and K. Land, "The potential of paper-based diagnostics to meet the ASSURED criteria," *RSC Adv.*, vol. 8, no. 59, pp. 34012–34034, Sep. 2018.

[28] "Quidel Receives FDA Clearance, CLIA Waiver for Its Point-of-Care Sofia® 2 Lyme Fluorescent Immunoassay for Use with Sofia® 2 Instrument from Finger-Stick Whole Blood Specimens," *Quidel Corporation*. [Online]. Available: http://ir.quidel.com/news-releases/news-release-details/quidel-receives-fda-clearance-clia-waiver-its-point-care-sofiar. [Accessed: 03-Sep-2019].

[29] F. T. Liang, A. C. Steere, A. R. Marques, B. J. B. Johnson, J. N. Miller, and M. T. Philipp, "Sensitive and Specific Serodiagnosis of Lyme Disease by Enzyme-Linked Immunosorbent Assay with a Peptide Based on an

Immunodominant Conserved Region of Borrelia burgdorferi VlsE," *Journal of Clinical Microbiology*, vol. 37, no. 12, pp. 3990–3996, Dec. 1999.

[30] P. M. Arnaboldi and R. J. Dattwyler, "Cross-Reactive Epitopes in Borrelia burgdorferi p66," *Clin. Vaccine Immunol.*, vol. 22, no. 7, pp. 840–843, Jul. 2015.

[31] A. B. Maulden *et al.*, "Two-Tier Lyme Disease Serology Test Results Can Vary According to the Specific First-Tier Test Used," *J Pediatric Infect Dis Soc*, Feb. 2019.

[32] L. J. Lahey *et al.*, "Development of a Multiantigen Panel for Improved Detection of Borrelia burgdorferi Infection in Early Lyme Disease," *J. Clin. Microbiol.*, vol. 53, no. 12, pp. 3834–3841, Dec. 2015.

[33] A. G. Barbour *et al.*, "A genome-wide proteome array reveals a limited set of immunogens in natural infections of humans and white-footed mice with Borrelia burgdorferi," *Infect. Immun.*, vol. 76, no. 8, pp. 3374–3389, Aug. 2008.

[34] R. B. Porwancher *et al.*, "Multiplex immunoassay for Lyme disease using VlsE1-IgG and pepC10-IgM antibodies: improving test performance through bioinformatics," *Clin. Vaccine Immunol.*, vol. 18, no. 5, pp. 851–859, May 2011.

[35] B. E. Bejnordi *et al.*, "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, Dec. 2017.

[36] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nat Med*, vol. 25, no. 1, pp. 24–29, Jan. 2019.

[37] L. Hu *et al.*, "An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening," *J Natl Cancer Inst*.

[38] J. A. Branda *et al.*, "Advances in Serodiagnostic Testing for Lyme Disease Are at Hand," *Clin. Infect. Dis.*, vol. 66, no. 7, pp. 1133–1139, 19 2018.

[39] C. Bouchard *et al.*, "The increasing risk of Lyme disease in Canada," *Can Vet J*, vol. 56, no. 7, pp. 693–699, Jul. 2015.

[40] A. Hofhuis, M. Harms, C. van den Wijngaard, H. Sprong, and W. van Pelt, "Continuing increase of tick bites and Lyme disease between 1994 and 2009," *Ticks and Tick-borne Diseases*, vol. 6, no. 1, pp. 69–74, Feb. 2015.

[41] "Geographic Distribution and Expansion of Human Lyme Disease, United States," *Emerg Infect Dis*, vol. 21, no. 8, pp. 1455–1457, Aug. 2015.

[42] S. Nayak *et al.*, "Microfluidics-based point-of-care test for serodiagnosis of Lyme Disease," *Scientific Reports*, vol. 6, p. 35069, Oct. 2016.

[43] H.-A. Joung *et al.*, "Paper-based multiplexed vertical flow assay for point-of-care testing," *Lab Chip*, Jan. 2019.

[44] M. J. C. Gomes-Solecki *et al.*, "Recombinant Chimeric Borrelia Proteins for Diagnosis of Lyme Disease," *Journal of Clinical Microbiology*, vol. 38, no. 7, pp. 2530–2535, Jul. 2000.

[45] D. L. DePietropaolo, J. H. Powers, J. M. Gill, and A. Foy, "Diagnosis of Lyme Disease," *AFP*, vol. 72, no. 2, pp. 297–304, Jul. 2005.

[46] Z. Ballard *et al.*, "Deep Learning-Enabled Point-of-Care Sensing Using Multiplexed Paper-Based Sensors," *bioRxiv*, p. 667436, Jun. 2019.

Chapter 4. Computational Sensing Using Low-Cost and Mobile Plasmonic Readers Designed by Machine Learning

This chapter, breaking with chapters 2 and 3 discusses applications of computational sensing in emerging plasmominc sensors. Though the previously discussed VFA platform uses plasmonic nanoparticles for color signal generation through immuno-labels, new photonic crystal structures have enabled more sensitive plasmonic sensors to work via label-free surface chemistry, with the ability to detect mid-sized proteins in the pg/mL range in whole blood. Emerging nano-fabrication techniques have also enabled these sensors to be cost-effectively mass-manufactured onto various types of substrates. To accompany these advances, major improvements in sensor read-out devices must also be achieved to fully realize the broad impact of plasmonic nano-sensors. In this chapter, I propose a machine learning framework which can be used to design low-cost and mobile multi-spectral plasmonic readers that do not use traditionally employed bulky and expensive stabilized light-sources or high-resolution spectrometers. By training a feature selection model over a large set of fabricated plasmonic nano-sensors, we select the optimal set of illumination light-emitting-diodes needed to create a minimum-error refractive index prediction model, which statistically takes into account the varied spectral responses and fabrication-induced variability of a given sensor design. This computational sensing approach was experimentally validated using a modular mobile plasmonic reader. We tested different plasmonic sensors with hexagonal and square periodicity nano-hole arrays, and revealed that the optimal illumination bands differ from those that are ‘intuitively’ selected based on the spectral features of the sensor, e.g., transmission peaks or valleys. This framework provides a universal tool for the plasmonics community to design low-cost and mobile multi-spectral readers, helping the translation of nano-sensing technologies to

various emerging applications such as wearable sensing, personalized medicine, and point-of-care diagnostics. Beyond plasmonics, other types of sensors that operate based on spectral changes can broadly benefit from this approach, including e.g., aptamer-enabled nanoparticle assays and graphene-based sensors, among others.

Part of this chapter has been previously published: Adapted with permission from (Z. S. Ballard, D. Shir, A. Bhardwaj, S. Bazargan, S. Sathianathan, and A. Ozcan, "Computational Sensing Using Low-Cost and Mobile Plasmonic Readers Designed by Machine Learning," *ACS Nano*, vol. 11, no. 2, pp. 2266–2274, Feb. 2017). Copyright (2017) American Chemical Society."

4.1 Introduction

Localized Surface Plasmon Resonance (LSPR) is at the heart of a class of biological and chemical sensors which operate by supporting resonances of collective electron oscillations that respond to changes of refractive index within their near field [1]–[3]. In these types of sensors, plasmonic resonances occur when light interacts with sub-wavelength metal nanostructures and can be interrogated via far-field optics in the visible part of the spectrum. Typically, a characteristic Lorentzian peak or trough in the reflection or transmission spectrum can be tracked in response to changes in the near-field refractive index [3]–[5]. For example, when a target analyte such as a protein or a virus is brought into the near-field of the plasmonic sensor structure, the effective refractive index is altered, and the characteristic peak undergoes a spectral shift. This resonance shift is particularly sensitive to surface binding events, making it a powerful tool to probe biological or chemical interactions near or at the sensor surface. LSPR sensors have already extensively been demonstrated in the literature as effective biological and chemical sensors, used for e.g., measuring DNA hybridization, heavy metal ion concentration, cancer bio-marker

detection, quantification of protein concentration, and even viral load measurement in unprocessed blood samples [6]–[21]. Furthermore, recent advances in nano-fabrication technologies such as colloidal self-assembly, soft lithography, and imprint molding have enabled high-throughput, low-cost, and scalable production of flexible, large-area, plasmonic sensors with a variety of different 2D and 3D nanostructure designs including nano-hole arrays, dome arrays, cross arrays, and many other exotic geometries [8], [22]–[32]. This recent and exciting trend now extends the applications of plasmonic sensors beyond laboratory settings for use as wearable sensors or as disposable point-of-care sensors, and also permits their integration into existing medical equipment such as intravenous tubes, syringes, blood bags, bandages, or medical garments [5], [30], [33]–[37].

However, with the proliferation of these low-cost and flexible LSPR-based sensors, new and innovative designs for the corresponding read-out devices must also be considered. Field portability, low-cost, ease-of-use, and network connectivity are all desired design features for ensuring widespread adoption of these sensing systems [33], [38], [39]. Currently, the most common read-out and quantification scheme for LSPR sensors employs a stable, broad-band light source for illumination and a high resolution optical spectrometer for recording the transmission or reflection spectra [3], [6], [40]–[43]. Alternatively, a tunable light source and a single photodiode can be used to obtain the same spectral information. Although a ‘peak/valley tracking’ based detection approach is quite effective in accurately registering and quantifying the sensor response, it requires optical hardware which can retrieve comprehensive spectral information, over a broad wavelength range of e.g., 400-700 nm. The inclusion of the above mentioned optical components can result in spectral read-out devices which are prohibitively bulky and expensive for many applications, especially in field and resource limited settings. Therefore, in an effort to engineer cost-effective and mobile plasmonic read-out devices, inexpensive optical components

such as LEDs, complementary metal–oxide–semiconductor (CMOS) imagers or single photodiodes along with mobile-phone based designs can be considered. Some recent work has showcased such design considerations [44], [45]. However, in designing these mobile devices, one must select a strategic set of bands to be able to sensitively register the spectral response of the plasmonic sensor design. This LED selection process is crucial to the performance of the plasmonic read-out device, and is influenced by many factors such as the spectral location of the plasmonic resonances, as well as the responsivity of the CMOS imager or the photodiode. Additionally, the optical signal resulting from the selected set of LEDs must contribute to a mathematical sensing model that is tolerant to the inherent fabrication variability of the plasmonic chips, which is unavoidable especially when using some of the emerging large-area, low-cost nano-fabrication techniques [46], [47]. Finally, these LSPR sensors should be 'plug and play' and not need an individualized calibration procedure per sensor chip that is often performed in the literature.

In this work we demonstrate a machine-learning based computational sensing framework which can be used to design the most optimum and yet cost-effective plasmonic read-out device that is suitable for various mobile sensing applications in field and resource-poor settings (see Fig. 4.1). Our computational sensing framework implements a statistical approach to determine the optimal set of illumination LEDs needed to create a minimum error refractive index prediction model by taking into account the varied spectral responses as well as the fabrication variability of the plasmonic sensor design of interest. To demonstrate this approach, 'training data' were taken over a statistically significant number of individual sensors (N>30) for two different plasmonic sensor designs (hexagonal and square periodicity nano-hole arrays, see Fig. 4.2), fabricated using a large-area nano-imprint molding approach, to learn the spectral response of each plasmonic

design due to refractive index changes. This data set was then used to train a machine learning algorithm with pre-defined constraints in order to form a computational model which accurately predicts the bulk refractive index of unknown test samples, without the need for an individualized calibration procedure per test.

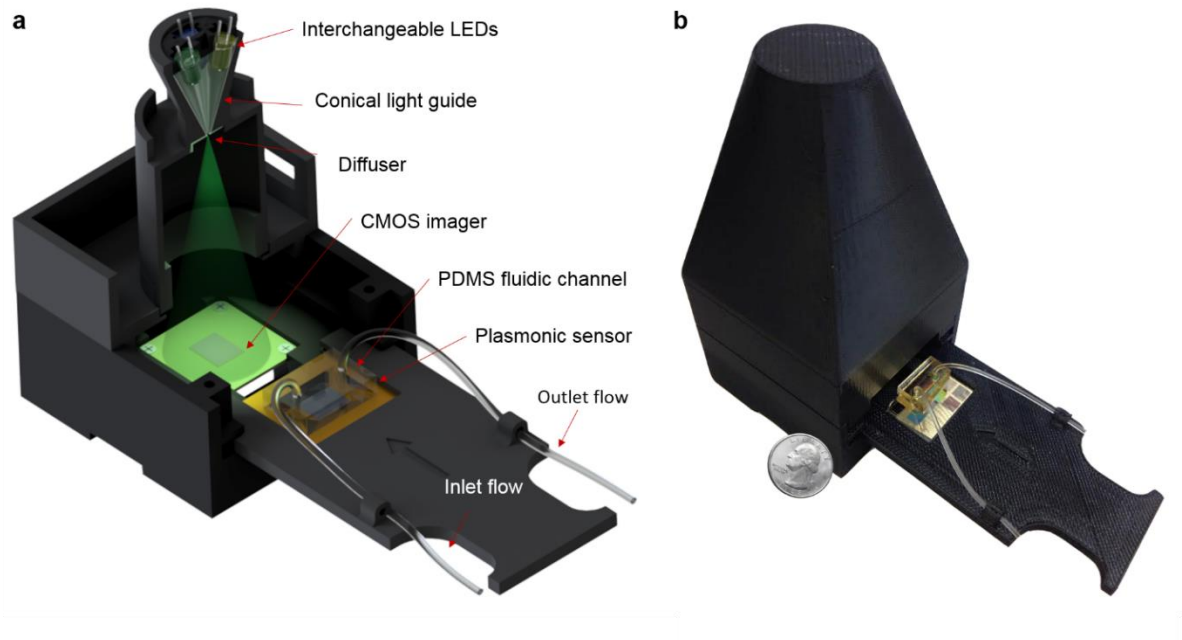


Figure 4.1. Modular design of our mobile, multi-spectral plasmonic reader. (a) Schematic illustration of the components of the reader. (b) 3D-printed prototype of the plasmonic reader used in this work.

This proposed computational sensing framework can be generally used by the plasmonic sensing community to design optimized LSPR read-out devices which aim for cost-effectiveness, mobility, and robustness. Our approach can be broadly applied to any LSPR sensor geometry, and can accommodate any set of practical design constraints. Beyond the scope of LSPR sensors, this computational sensing and design framework can also be employed by any type of biochemical sensor which operates via a spectrum change or shift. Emerging sensing elements such as aptamer-enabled nanoparticle assays, plasmonic interferometers, dynamically tunable plasmonic sensors, and graphene-based materials, with engineered physical properties and responses can similarly be

analyzed by this framework in order to determine the optimal mobile read-out and computational sensing scheme [24], [48]–[54]. Taken together, this proposed framework can be used as a robust engineering tool to design next generation plasmonic read-out and sensing devices for e.g., wearable and embedded sensing systems, personalized medicine applications, and point-of-care diagnostics, among various others.

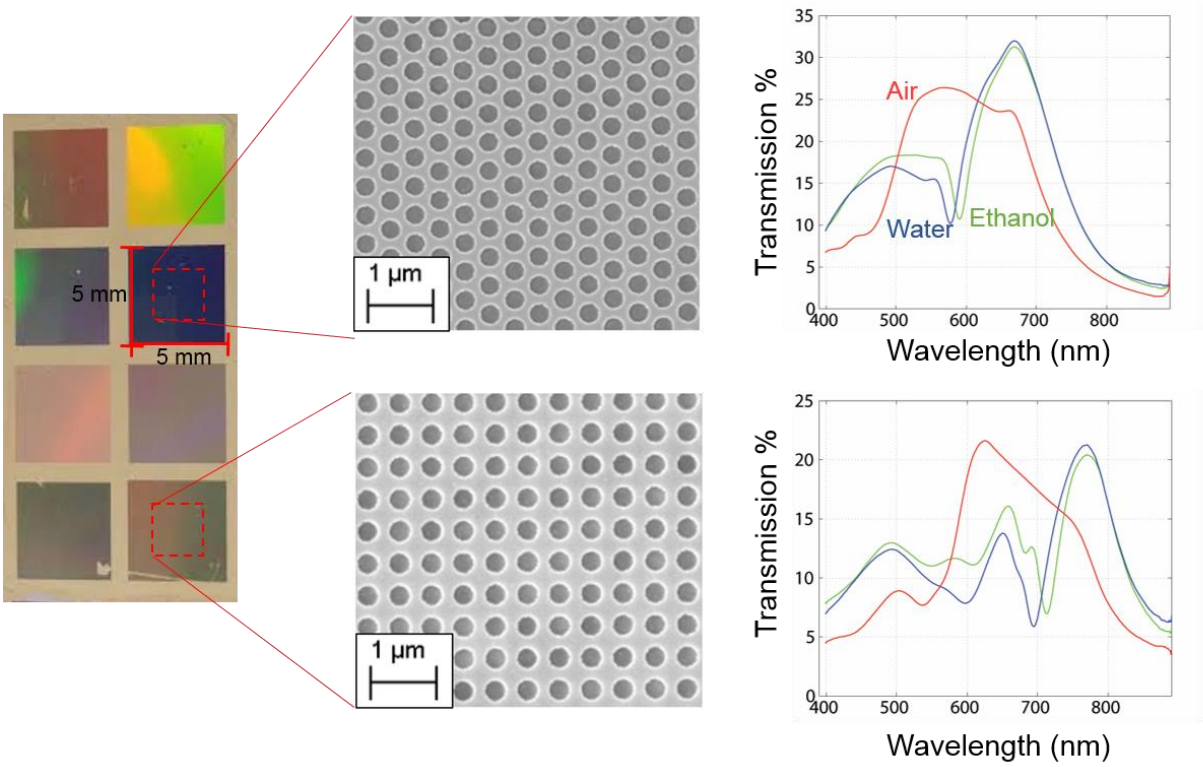


Figure 4.2. Optical and SEM images of the plasmonic sensors consisting of hexagonal and square array of nano-holes and their corresponding transmission spectra in differing refractive index environments (right).

4.2 Results and Discussion

Computational sensing and design framework for multi-spectral mobile plasmonic read-out using machine learning

The machine learning approach used in this work selects the optimal sub-set of LEDs from a pre-defined library, and aims to produce the minimum error refractive index prediction model for a given plasmonic nano-structure design and nano-fabrication method. To generate the initial 'training data' set, our plasmonic sensors, equipped with fluidic channels, were placed on a transmission stage and illuminated by a fiber-coupled broadband stabilized light source (Thor Labs, SLS201) as outlined in Figure 4.3.

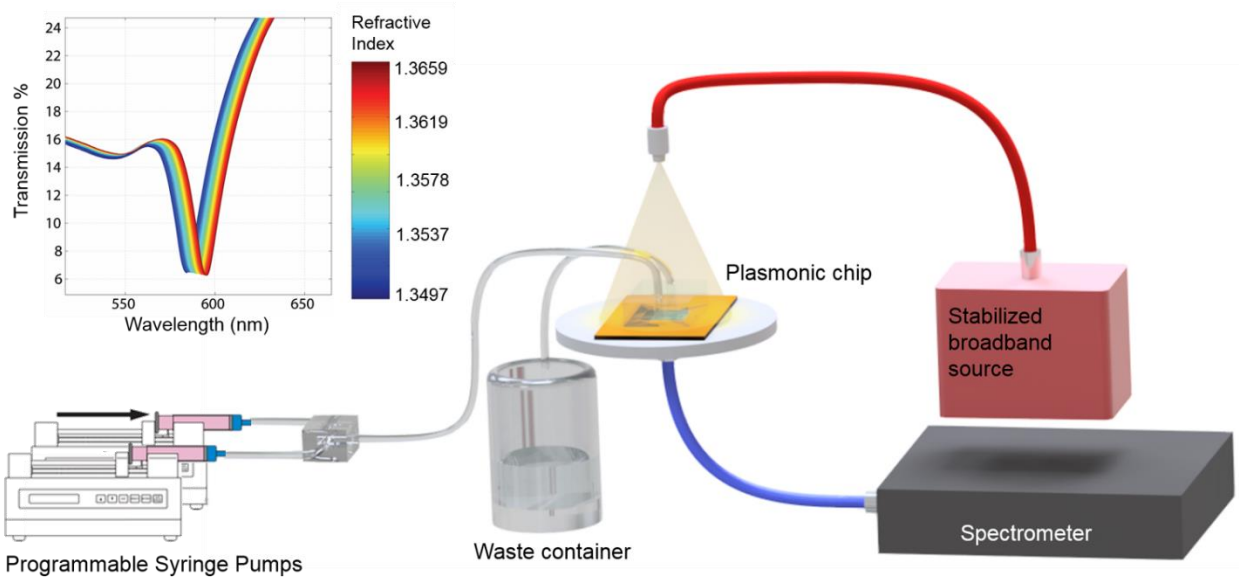


Figure 4.3. Schematic illustration of the fluidics set-up for collecting training data. The inset shows a representative transmission spectrum shift as the surrounding bulk refractive index was increased.

This training process needs to be performed *only once* for a given plasmonic sensor design and nano-fabrication method, and it aims to infer the statistical spectral variations of the sensor design, as a function of bulk refractive index as well as fabrication tolerances and imperfections. After a 'spectral stack' was recorded for every chip in the training set (N = 33 for both the hexagonal and square periodicity nano-hole arrays, Fig. 2), each spectrum was normalized to its 'reference spectrum', which in this case was taken to be the plasmonic transmission spectrum in de-ionized

water (RIU = 1.3325). It is important to note that the resulting normalized '*contrast spectra'*, as shown in Figure 4.4, were normalized to their own characteristic references. A global reference would not properly zero the contrast information from the LEDs for every trained chip and therefore was not used. Next, a custom-built LED library (refer to the Materials and Methods Section for details) was utilized to simulate every possible LED transmission through the sensor. By defining the spectral output of each LED in terms of the peak wavelength and line-width specified in their respective data sheet, the overlap integral of the LED spectra with the plasmonic contrast spectra was numerically calculated. Applying this procedure for each sensor in the training data set yielded a matrix, X_{Lib}, which contained the simulated contrast values of all the LEDs in the pre-defined LED library for each bulk refractive index that was sampled during the training experiments (see Fig. 4.4). Specifically, the number of rows in this matrix is defined by the product of the number of plasmonic nano-sensors trained (N=33) and the number of transmission spectra sampled per sensor during the bulk refractive index modulation, which for our data set was 13, i.e., 33 x 13 = 429 rows exist in X_{Lib}. The number of columns is defined by the number of LED contrast features (i.e., the number of entries in our LED library, denoted with *n* in Fig. 4.4) plus one, where an additional column of ones is concatenated onto the LED contrast feature matrix to provide a constant term in the linear model. *Quite importantly, X_{Lib} contains spectral data from multiple plasmonic sensor chips, and therefore has the appropriate statistical information of how each sensor's spectral response varies due nano-fabrication variability and imperfections*.

This training data matrix, X_{Lib}, was then used as an input into an L1-norm regularization algorithm (i.e., LASSO, least absolute shrinkage and selection operator) which aims to find a regularized least-squares solution between the output of a linear model and the measured 'gold-standard' refractive index values, *y*, [55]–[58], i.e.,

$$b^* = \underset{b \in \mathbb{R}^{n+1}}{argmin} \left\| X_{Lib} b - y \right\|_{\ell_2}^2 + \lambda \|b\|_{\ell_1} \qquad \text{Eq. 4.1}$$

where the vector b is the variable which is solved for, b^* being the optimal solution for computational sensing, corresponding to the $n + 1$ optimal coefficients, which define the linear model. λ is the regularization parameter, which in this manuscript is typically set to be between 10^{-4} and 10^{-7} as discussed later in the manuscript.

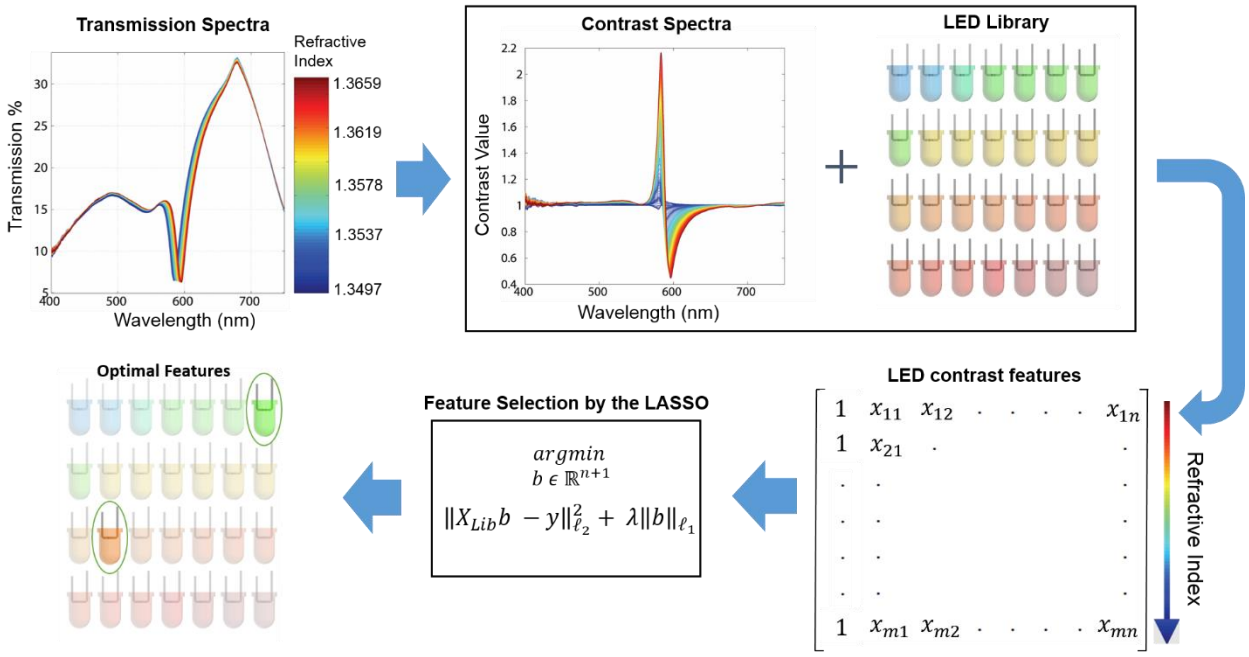


Figure 4.4. Schematic illustration of the machine learning framework.

Based on the above described statistical framework, in order to determine the optimal set of LEDs to be used in our computational mobile plasmonic reader, we used a 'leave-one-sample-out cross validation' (LOOCV) procedure with the training data set (N=33). For each iteration, the 33 samples were partitioned; 32 samples as training data and 1 sample being used a testing sample. The regularization parameter was selected by performing a nested LOOCV within that iteration's set of 32 training samples, revealing what range of parameters yielded the minimum error model,

which for these training sets was typically between 10^{-4} and 10^{-7}. It is also important to note that as the regularization parameter is varied over this range, the mean-squared-error of the nested LOOCV remains within one standard deviation of the minimum error and the features with corresponding non-zero coefficients in the linear models remain constant (see Appendix Fig. 4.7). This analysis indicates that no large bias or over-fitting permeates the linear models with this selection of the regularization parameter.

During this LED selection process and LOOCV procedure, 33 different linear models were generated using the LASSO (Eq. 4.1) and all the non-zero coefficients in each resulting b^* solution were recorded every iteration, revealing which LEDs were statistically favored out of our LED library for forming the optimal linear model. The top four LEDs utilized most in these linear models were then selected for the testing process. A detailed ranking of these LEDs is shown in Appendix Figure 4.8. For the case of the square periodicity nano-hole array (see Fig. 4.2), the fourth ranked LED (λ_{peak} = 735 nm) was replaced by an LED with λ_{peak} = 660 nm so that in the blind validation step we could directly compare our optimal linear model to a model created with LEDs directly to the left and right of the dominant plasmonic spectral feature as discussed later.

After this LED selection process, the optimal set of LEDs that we inferred were installed into our field-portable plasmonic reader device (Figure 4.1), the modular optical hardware of which will be detailed in the next sub-section. To ensure an accurate refractive index prediction model, the exact output spectra of the selected LEDs were also measured and used in our computational sensing step to mitigate any error which might occur due to slight differences between the modeled and actual spectra of the selected LEDs. The final refractive index prediction model was then generated by inputting the entire training data set into the same L1-norm regularized LASSO algorithm, this time with the training data matrix, X_{OptLED}, consisting of 5 columns (4 columns

corresponding to the optimal set of 4 LEDs installed in the portable reader, plus the column of ones as before), i.e.,

$$b^*_{mobile} = \underset{b_{mobile}\epsilon\mathbb{R}^{n+1}}{argmin} \left\| X_{OptLED} b_{mobile} - y \right\|^2_{\ell_2} + \lambda \| b_{mobile} \|_{\ell_1} \qquad \text{Eq. 4.2}$$

where the regularization parameter (λ) was determined to be 1.15 x 10^{-6} and 4.78 x 10^{-7} for the hexagonal and square periodicity implementation of the algorithm, respectively. This selection was based off a secondary LOOCV, which determined the λ needed to achieve the minimum error solution over the training set. This secondary cross-validation step also elucidates the degree to which over-fitting might occur with the final linear model. Parametrically sweeping the regularization parameter over a logarithmic range of 10^{-7} to 10^{0} assures there no over-fitting tendencies with the selection of a small regularization parameter in the range 10^{-7} to 10^{-4}. This is because our set of LED contrast features are highly correlated to the change in bulk refractive index, especially given the optimal set of LEDs which exclude the features that yield high variance across multiple fabricated chips. The optimal solution for our mobile plasmonic reader design, b^*_{mobile}, was then used as the vector of coefficients in the final *computational sensing* step to predict the refractive index (RIU) of the sample using the mobile reader device, i.e.,

$$RIU_{prediction} = X_{test} b^*_{mobile} \qquad \text{Eq 4.3}$$

where each row of X_{test} refers to an independent sensing measurement or multi-spectral test made with our mobile plasmonic reader.

Modular optical design of a low-cost and mobile multi-spectral plasmonic reader

Figure 4.1 shows the LED-based mobile multi-spectral plasmonic reader hardware used to

validate our framework. This modular device is compact (6.5 x 6.5 x 7.5 cm), light weight (< 160 g), and cost effective, utilizing multiple LEDs for illumination and a CMOS imager (Basler dart series, daA1280-54µm USB 3.0) for the capture of the transmitted light from a plasmonic sensor of interest. An Arduino micro-controller was programmed to sequentially power on the LEDs and trigger the camera for image capture. The resulting device was prototyped via 3-D printing, and can be powered either through a USB cable or by battery. This device was specifically designed to accommodate interchangeable LEDs (up to a maximum of 4), consisting of a detachable 3-D printed cap which holds the selected LEDs and allows for convenient substitution of each LED depending on the design of the plasmonic chip. In this work, we used plasmonic nanostructures in the form of hexagonal and square array of holes with periodicity, relief depth, and hole diameter of 500 nm, 300 nm, and 380 nm, respectively, as shown in Figure 4.2. These plasmonic sensors were fabricated using soft imprint lithography, a process that is low-cost, scalable, and high-throughput, as detailed in the Materials and Methods section. During the device operation, the LEDs are turned on sequentially and their output light is guided through a conical light-guide onto a diffuser and then into a 1 mm pinhole. This novel design allows for up to four LEDs to mutually illuminate the LSPR sensor at approximately normal incidence angle. Additionally, this design records multi-pixel optical transmission information over a field of view of >17 mm^2, as opposed to the hyper-spectral yet single-pixel information obtained by a traditionally employed spectrometer. This unique difference in our field portable design allows for a multiplexed sensor read out, simultaneously reading transmission information from multiple surface functionalized regions targeting e.g., a panel of analytes. For the purposes of this study and to illustrate the proof-of-concept of our proposed computational sensing framework, we calculated a single transmission value from each CMOS image for a given illumination LED and bulk refractive index by averaging

the pixel intensity of a centrally located binned region (i.e., 100 x 100 pixels). Lastly, the incorporation of a CMOS imager provides a robust, easy-to-align and yet compact design with simple light-coupling as a result of the in-line lensless imaging-based computational detection scheme.

Blind testing and validation of the machine learning and computational sensing framework using a modular mobile plasmonic reader

For independent validation of our computational sensing framework, our low-cost and field-portable plasmonic reader was used to blindly test a set of LSPR sensors, i.e., the plasmonic sensors being tested had ***not*** previously been used in our training data, preventing any sort of data contamination or artificial overfitting in the final linear model. Eight chips (N = 8) were tested with M=13 refractive index measurements sampled for each plasmonic chip, using the same fluidics mixing set-up illustrated in Figure 4.2. To illustrate the generality of this framework, both the hexagonal and square periodicity nano-hole arrays were tested (see Fig. 2), and the mean error of the RIU predictions resulting from our computational sensing framework (Eq. 4.3) was calculated using,

$$Mean\ Error = \frac{1}{N\,M}\sum_{p=1}^{N}\sum_{i=1}^{M}\left|RIU_{prediction}{}_{i}^{p} - y_i^p\right| \qquad \text{Eq. 4.4}$$

where, $RIU_{prediction}{}_{i}^{p}$ and y_i^p represent the predicted and the gold standard refractive index values, respectively, for a given measurement, i, and a given testing chip, p.

Figure 4.5 compares three distinct linear models for our plasmonic sensors, for both the hexagonal (Fig 4.5 a-c) and square nano-hole array designs (Fig 4.5 d-f). The first linear model (dark blue) uses a single LED feature closest to the dominant plasmonic spectral feature (located

at 590 nm and 700 nm for the hexagonal and square periodicity nano-hole arrays, respectively). The second linear model (light blue) uses a pair of LEDs directly to the right and left of the dominant plasmonic spectral feature. Lastly, the third linear model (green) uses the LEDs chosen through the feature selection process based on our machine learning framework. Previously published LED based field-portable plasmonic reader designs often use the second linear model (i.e., a ratio-metric approach) to measure and quantify sensor responses (45). However, this type of a design does not necessarily provide the most accurate results. For example, in the case of the hexagonal nano-hole array LSPR sensor, a linear model using LEDs that are closest to the main spectral feature, i.e. 595 nm and 527 nm LEDs to the right and to the left of the LSPR peak, respectively, resulted in ***~5 times higher mean error*** compared to a linear model using LEDs (525 nm and 611 nm) selected based on our machine learning approach (see Fig 4.5). This reduction in mean error enabled by our machine learning approach is largely due to the inherent fabrication variability of large-area, low-cost, and scalable nano-fabrication techniques. This fabrication variability can be evidenced by varied microscopic and macroscopic defect rates, differing cured photo-polymer thickness, non-uniform topography, which might cause slight deviations from perfectly normal illumination thus altering the coupling conditions, as well as deformation of the desired cylindrical non-hole array structure. The advantages gained in scalability and cost-effectiveness of nano-fabrication can adversely result in subtle differences in peak locations of the main LSPR resonance from chip to chip (1-2 nm difference in peak location) as well as varying line-widths of the resonance. Therefore, when LEDs are 'intuitively' chosen based on their proximity to the dominant spectral features of the plasmonic sensor, they can yield large variances in their contrast values during the analyte-induced spectral shift. Therefore, *such features should be avoided when designing a robust read-out model, especially when dealing with a large number*

of sensors that will naturally exhibit fabrication-related variations in their spectral response, as detailed earlier. Here we illuminate an important trade-off in which signal strength (i.e., absolute LED contrast) is sacrificed for features which yield lower variance from chip to chip. In other words, more stable features exist from LEDs which only have partial spectral overlap (and thus less contrast) with the dominant spectral features. Our machine learning-based computational sensing approach ultimately selects the optimal compromise in this trade-off for a given plasmonic design and fabrication method.

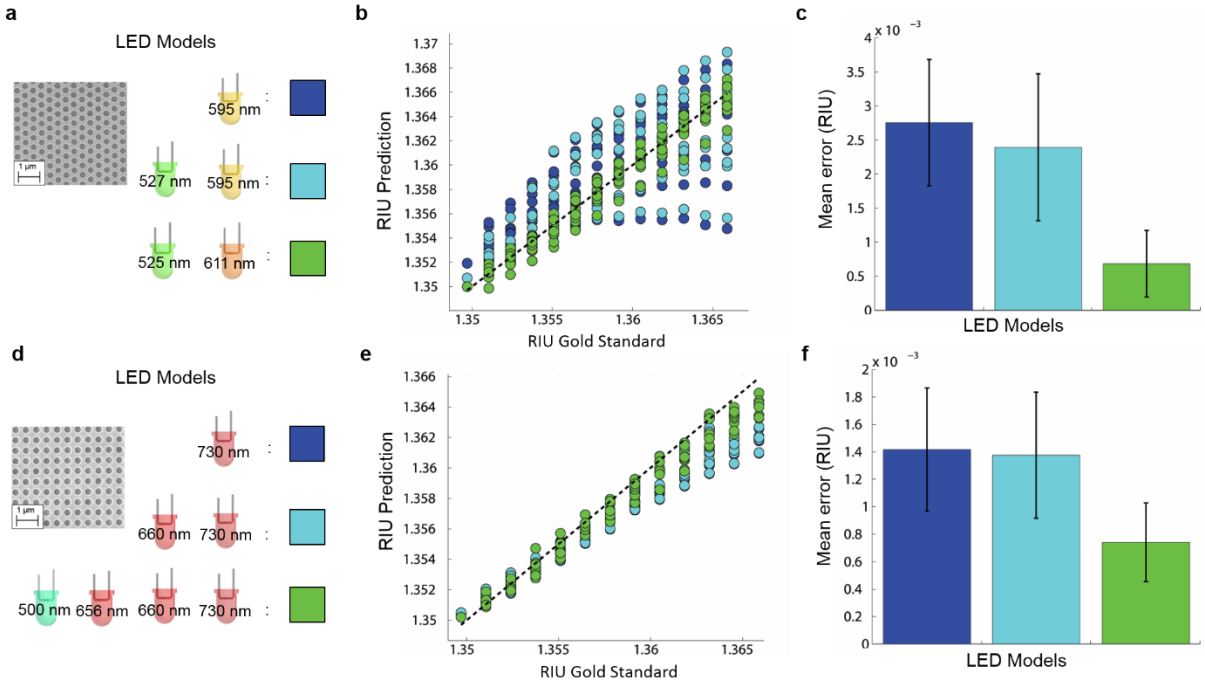


Figure 4.5. Comparison of different LED linear models. LEDs used in linear models (a), (d) for the two plasmonic sensors; hexagonal and square periodicity nano-hole arrays, respectively. (b), (e) Comparison between the refractive index prediction and gold standard values for the three models used in (a) and (d), respectively. Each LED model is color coded for better visualization, and the third linear model (green) uses the LEDs chosen through the feature selection process based on our machine learning framework. Mean error comparison is provided in (c) and (f) for the three linear models. Please refer to the Appendix Information for additional performance comparisons among different models.

In the case of the hexagonal plasmonic sensor, the feature selection property of the L1-norm regularized LASSO algorithm recognized the large variance of the 595 nm LED contrast values and therefore forced this coefficient in the linear model to zero, instead selecting the 611 nm LED as a more stable spectral feature. A similar, albeit less dramatic effect, can be seen with the square periodicity plasmonic sensor. For this testing data, a model which utilized four LED contrast features proved more robust than using the two 'intuitive' LEDs located to the right and left (730 nm and 656 nm, respectfully) of the dominant LSPR spectral feature. Of particular importance is the inclusion of the 500 nm LED into the optimal model which reduced the refractive index prediction error in the testing data by 50%. *This LED would not necessarily have been chosen by intuition because it is almost completely removed from the main LSPR spectral feature.*

In summary, by 'learning' from the training data, which needs to be performed *only once* for a given plasmonic chip design and fabrication method, this machine learning and computational sensing framework identifies the spectral regions with a consistent response, differentiating them from the spectral regions with a varied response, and accordingly adjusts the computational sensing linear model to select the most stable features by divulging the optimal set of spectral bands (LEDs) and their corresponding weights. For completeness, the testing errors corresponding to all the possible linear models which could be made with the 4 selected LEDs and training data sets were also calculated and compared to each other for both of our plasmonic sensor designs (see Appendix Tables 4.1-4.2.) In this comparison, we selected linear models which are solutions of the L-1 norm LASSO regularization, the L-2 norm Tikhonov regularization[59], and a multivariable least squares solution based on QR decomposition, which contains no regularization term[60].

Lastly, in an effort to realistically address the role of future plasmonic read-out devices for various emerging sensing applications it must be emphasized that the plasmonic sensors used in this study were fabricated with techniques which allow for *high-throughput, low-cost, and scalable* sensor production. For example, the vapor deposited 'non-stick' layer on the silicon master (detailed in the Materials and Methods section) lasts for many iterations of the fabrication procedure, only needing to be reapplied after every 30-40 uses as a mold. Although the production of the initial silicon master requires the use of conventional photolithography tools and procedures, each silicon master can be used indefinitely with proper care and treatment, producing hundreds to thousands of soft molds, thereby dramatically reducing the fabrication cost of each quasi-3D nanostructure. Additionally, the subsequent imprint molding process can be performed by individuals with minimal training and equipment, requiring only basic tools and a standard UV lamp. This process can also be highly parallelized, only being limited by the number of silicon masters available for the initial soft mold fabrication. Each soft mold can be used upwards of 20 times without incurring significant defects. The metal deposition is the only step in this procedure which requires a clean-room facility, however, each metal deposition run can produce *thousands of plasmonic sensors*, depending on the tool configuration. For example, with the metal deposition tool (CHA Solution Electron Beam Evaporator) used in this work, approximately 1,400 plasmonic sensors can undergo metal deposition at once, assuming each sensor has an area of 5x5 mm and seven four-inch wafers can be loaded into the evaporator per run (as in our case). The material costs are also minimal, with ~0.4 grams of gold being used per run for a 50 nm coating.

Conclusions

We developed and validated a universal machine learning and computational sensing framework, which brings mobility and cost-effectiveness to plasmonic sensor read-out device design. This computational sensing approach is especially valuable in designing multi-spectral readers where the selection of optimal low-cost illumination bands is critical. This framework is truly a 'black-box', compatible with any arbitrary plasmonic sensor geometry, and any illumination library based off of user-defined design constraints. As a result of this, the presented framework can be extended to broadly benefit any optical sensor which operates based on spectral changes in its transmission or reflection response. Also, because any illumination library can be used, more targeted libraries which include the spectral output from combinations of filters and LEDs can be explored along with the rolling addition of emerging illumination sources. Taken together, we believe this framework can be used by the plasmonic sensing community to design and optimize low-cost mobile readers for quantification of e.g., protein concentration, ion detection, and even whole-virus quantification with minimum error. Furthermore, coupled with the advances in scalable and low-cost plasmonic sensor fabrication techniques, our computational sensing approach holds significant potential to advance emerging applications for wearable sensors, personalized medicine, and point-of-care diagnostics.

4.3 Materials and Methods

Nano-fabrication of plasmonic sensors

The fabrication process of our plasmonic sensors (depicted in Appendix, Fig. 4.6) involved first generating a silicon 'master,' through a one-time photo-lithography process, which contained the desired nanostructures such as our hexagonal and square periodicity nano-hole arrays. Next, a

monolayer of Tridecafluoro, 1, 1,2,2-Trtrahydrooctyl-1-trichlorosilane (Gelest, Inc.) was deposited via vapor deposition to form a 'non-stick' layer for the subsequent nano-imprint molding process. The masters were then used as molds by casting a drop of UV curable polyurethane acrylate (PUA-311RM, Minuta Technology, Inc.) onto the surface. A flexible cellulose-acetate film was used to disperse and flatten the liquid droplet over the silicon master, before being placed under a UV lamp (UV-A, 4W, 800 μW/cm^2, Thermo Fisher) for 2 hours to cure. After UV curing, the cellulose-acetate film was peeled from the master, completing the fabrication process of the 'soft mold'. These soft molds therefore consist of the inverse geometries of that on the silicon master, and must be used for a secondary imprint molding process in order to recover the desired nano-hole array structure. The soft mold was then used to imprint its relief features onto liquid precursor of photo-curable polymer (NOA 81) deposited onto an oxygen plasma cleaned glass slide. After 25 minutes of curing time under the UV lamp, the soft mold was peeled away from the glass backed photo-polymer, completing the fabrication process of the desired quasi-3D nano-hole array. Finally, a bi-layer of 5 nm chromium and 50 nm gold was deposited onto the nanostructures by Electron Beam Evaporation (CHA Solution Electron Beam Evaporator) at a deposition rate of 2 Å/s and 3 Å/s, respectively, to complete the fabrication of the plasmonic sensor. Once the final plasmonic sensor was fabricated, a fluidic channel made of Polydimethysiloxane (PDMS) with a 300 μm height and 3 mm width was placed onto the surface with inlet and outlet holes to allow for unidirectional flow over the plasmonic sensor surface.

Experimental training data

A high resolution spectrometer (Ocean Optics, HR2000+) was used to continuously capture the transmission spectra of the plasmonic sensors as the bulk refractive index in the fluidic channel

was modulated over time. To ensure accurate and consistent bulk refractive index modulation during the training experiments, two programmable syringe pumps were employed (Chemyx, Fusion 100 Infusion). The first syringe pump contained filtered de-ionized water, and the second contained 0.3 g/mL glucose solution in water. The syringe pumps each fed into a T-connector where their contents were combined and mixed through diffusion in a fluidic channel with a length of 60 cm. During the course of these training experiments, the combined flow rate of the two syringe pumps was held constant at 30 µL/min, while the flow rate of the first syringe pump containing water was programmatically decreased and the flow rate of the second syringe pump containing the glucose solution was programmatically increased, ensuring a continuous and consistent gradient of bulk refractive index over the plasmonic sensor surface. These continuous spectral measurements of the transmission formed a 'spectral stack' which describes the individual plasmonic sensor's spectral evolution in response to increasing bulk refractive index. To ensure the bulk refractive index modulation was appropriately being executed, bulk refractive index samples were taken at the experimental time-points and verified with a refractometer (Bausch and Lomb, Abbe refractometer). For each plasmonic nano-structure design, we measured the transmission spectra of N = 33 individual plasmonic sensors, which provided sufficient statistical information on spectral variations due to fabrication tolerances, as already discussed in the Results and Discussion section.

Formation of the LED library

A library of possible illumination LEDs was generated by applying constraints to the available LEDs in the 'LED Indication-Discrete' database on the Digi-Key website (i.e., ~21,000 LEDs)[61]. The on-line database was filtered using the following design constraints: 465 nm <

λ_{peak} < 700 nm, millicandela rating > 10,000 mCd, and a requirement of through-hole mounting type. These design constraints ensured that the necessary plasmonic spectral region was covered, and that each LED had sufficient brightness for image capture. The following surface mount LEDs were also added to our LED library ad-hoc to cover the spectral region to the right of the dominant plasmonic feature of the square periodicity sensor: 656 nm (DigiKey part number: 475-3008-1-ND), 660 nm (DigiKey part number: 1214-1436-1-ND), 730 nm (DigiKey part number: 1214-1440-1-ND), 735 nm (DigiKey part number: 1416-1913-1-ND). Our final LED library consisted of 28 LEDs as depicted with color representation in Figure 4.4.

4.4 Appendix

Supplementary Figures

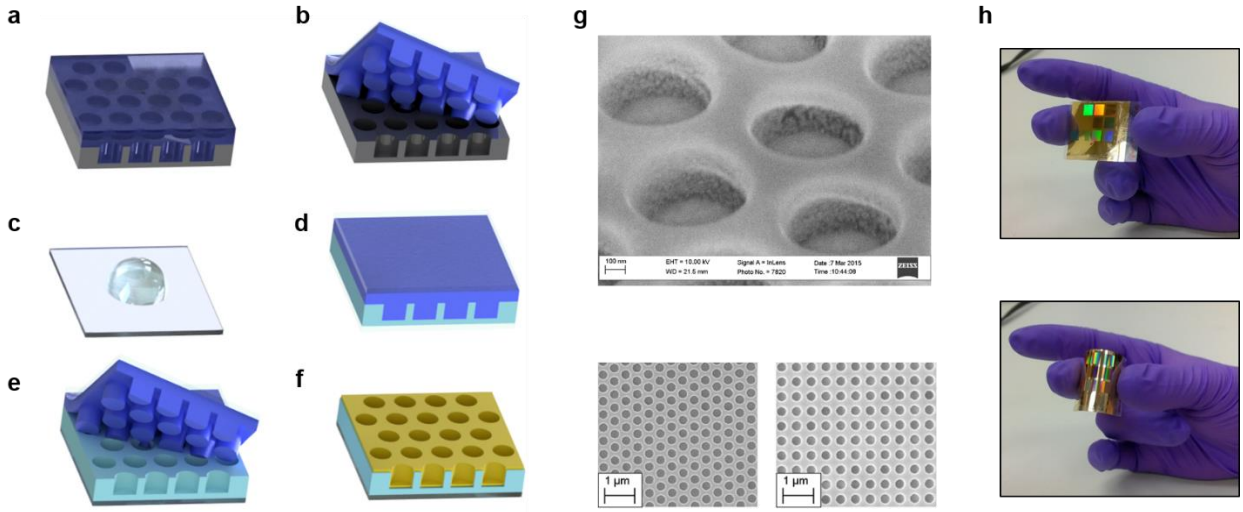


Figure 4.6. The imprint mold fabrication procedure for plasmonic sensor fabrication (a)-(f) starting from the one-time produced silicon master which is molded with a polyurethane acrylate (a) and (b). This initial mold is then peeled from master and used in a secondary imprint process (c)-(e) with a UV curable polymer (NOA 81) before the final gold deposition procedure (f). SEM images (g) show the hexagonal and square periodicity nano-hole array structures with the side-wall profile characteristic of the line-of-sight Electron Beam Evaporator deposition. The final plasmonic sensors (h) fabricated on a flexible mylar backing layer.

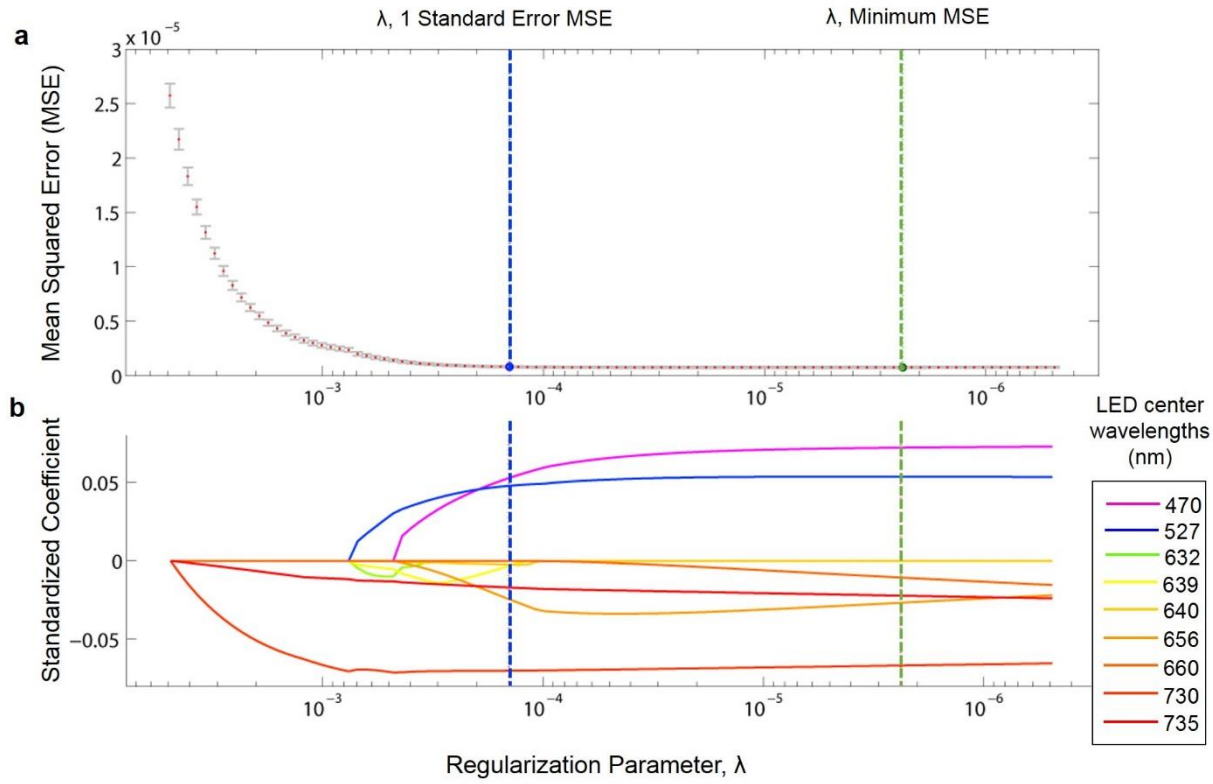


Figure 4.7. Feature selection through L-1 regularization tuning. (a) The mean squared error (MSE) and (b) the standardized coefficients as a function of regularization parameter for a representative set of 32 training samples within the cross validation LED selection procedure. The regularization parameter for the minimum MSE is found by analyzing the error from a nested LOOCV performed on the 32 training samples. This regularization parameter is then selected for the corresponding linear model.

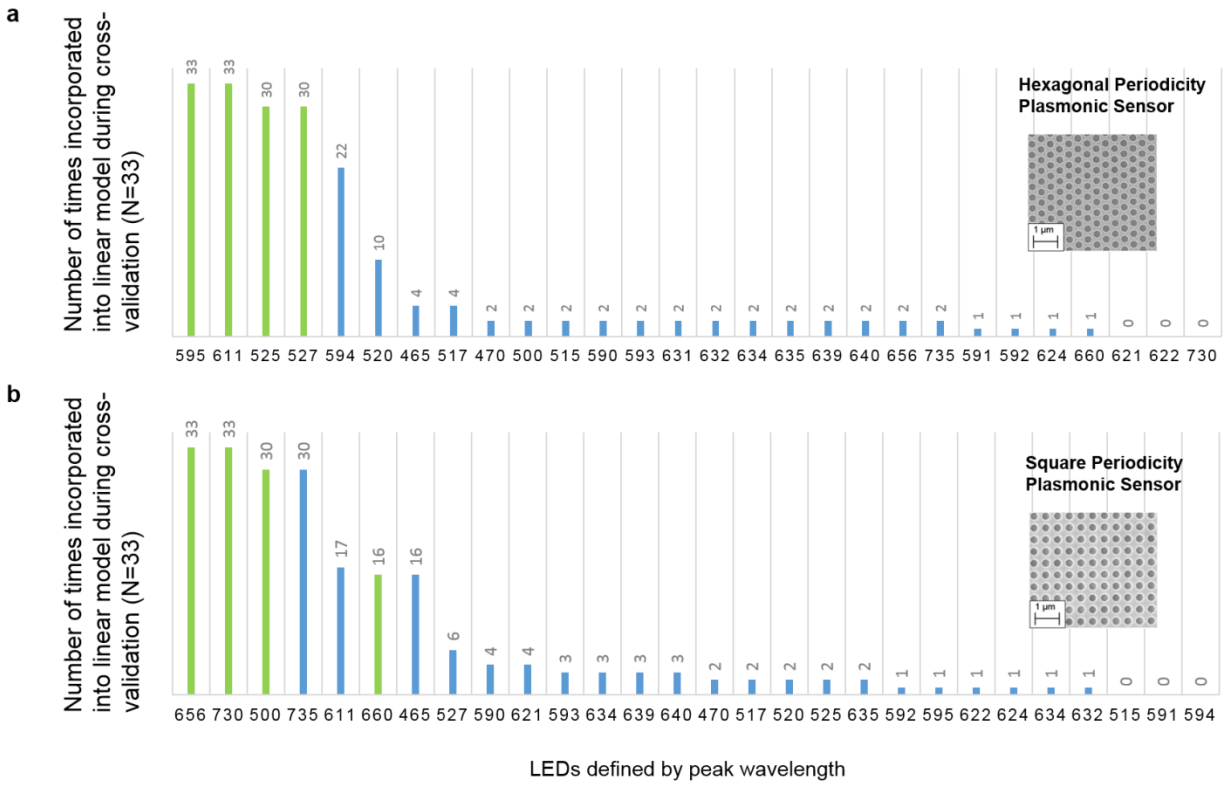


Figure 4.8. Ranking of LEDs in terms of the number of times they are used in linear models during the LASSO cross validation step for (a) the hexagonal and (b) square periodicity nano-hole array plasmonic sensors.

Supplementary Tables

Table 4.1. Hexagonal periodicity plasmonic sensor testing error (RIU).

LEDs used in model	LASSO	Tikhonov Regularization	Least Squares Solution
525	0.0039954	0.0039763	0.0039743
527	0.0041323	0.0041336	0.0041363
595	0.0027801	0.0027736	0.0027566
611	0.00074648	0.00074205	0.00075583
525 +527	N/A	0.0074443	0.010312
525 + 595	0.0023981	0.0024091	0.0023862
525 + 611	0.00068442	0.00068453	0.00068494
527 + 595	0.0023967	0.0023953	0.0024005
527 + 611	0.00070253	0.00070486	0.00070221
595 + 611	N/A	0.0007807	0.00078449
525 + 527 +595	N/A	0.0038764	0.0055691
525 + 527 + 611	N/A	0.00099724	0.0014449
525 + 595 + 611	N/A	0.00067642	0.00067551
527 + 595 +611	N/A	0.00069053	0.00069022
525 + 527 + 595 +611	N/A	0.00096541	0.0014056

Table 4.2. Square periodicity plasmonic sensor testing error (RIU).

LEDs used in model	LASSO	Tikhonov Regularization	Least Squares Solution
500	0.0034	0.003354	0.003369
656	0.003869	0.003871	0.003864
660	0.003903	0.003878	0.003875
730	0.001392	0.001389	0.001407
500 +656	0.002202	0.002189	0.00219
500+660	0.002697	0.002714	0.00272
500+730	0.000972	0.000976	0.000939
656+660	0.003795	0.003784	0.003789
656+730	0.001512	0.00149	0.001484
660+730	0.0014	0.001411	0.00142
500 +656 + 660	0.001812	0.001809	0.001803
500+656+730	0.000971	0.000948	0.00096
500+660+730	0.000781	0.000779	0.000787
656+660+730	N/A	0.001495	0.001481
500+656+660+730	0.000748	0.000749	0.000753

Discussion

Supplementary Tables 4.1-4.2 not only compare the performances of the linear models solved using these three techniques, but also illustrate comprehensively how our 'optimal model,' b^*_{mobile}, compares to other linear models made with all the combinations of the LEDs that could be installed in our modular mobile plasmonic reader. More specifically, the Tikhonov regularization is similar to the LASSO regularization (Eq. 4.1), differing only in the L-2 norm or the Euclidian norm, $\|\cdot\|_{\ell 2}$, i.e.,

$$t^*_{mobile} = \underset{t_{mobile}\epsilon\mathbb{R}^{n+1}}{argmin} \left\|X_{OptLED}t_{mobile} - y\right\|^2_{\ell_2} + \|\Gamma t_{mobile}\|_{\ell_2}, \quad \text{Eq. 4.5}$$

which, similar to Eq. 4.3, can be used for computational sensing of the bulk refractive index:

$$RIU_{prediction} = X_{test}t^*_{mobile}, \quad \text{Eq. 4.6}$$

The Tikhonov matrix, Γ, in Eq. 4.5 is generally employed as the regularization term and can be chosen in some cases to be a multiple of the identity matrix (I), i.e. $\Gamma = \lambda I$, where λ is a scalar term referred to as the regularization parameter. The use of such regularization terms in both the Tikhonov and LASSO solutions allows for a computational sensing model which is tolerant to outliers while also accounting for the statistical variance (e.g., due to fabrication) and average of the inputted features. The LASSO, however, is a very powerful optimization tool, because it forces many of the coefficients in the linear model to *zero*, due to the L-1 norm term. This process is often referred to as 'feature selection,' and in the context of this work discriminates the optimal subset of LEDs from the larger LED library. This feature selection property is precisely why many of the entries in the LASSO column of Supplementary Tables 4.1 and 4.2 are empty. When the LED features are inputted into the LASSO model, it sometimes forces the LED weights to zero, yielding a linear model that does not include all the inputted features (and is therefore not applicable for that specific entry of the comparison table). It is also evident from the testing data reported in Supplementary Tables 4.1-4.2 that the Tikhonov regularization, similar to LASSO, is effective in designing low error computational sensing models. However, it should be noted that the Tikhonov regression cannot

replace the feature selection properties of LASSO, and therefore cannot explicitly be used to determine an optimal subset of features from a larger non-sparse set of features.

References

[1] J. Zhao, X. Zhang, C. R. Yonzon, A. J. Haes, and R. P. Van Duyne, "Localized surface plasmon resonance biosensors," *Nanomed.*, vol. 1, no. 2, pp. 219–228, Aug. 2006.

[2] S. Unser, I. Bruzas, J. He, and L. Sagle, "Localized Surface Plasmon Resonance Biosensing: Current Challenges and Approaches," *Sensors*, vol. 15, no. 7, pp. 15684–15716, Jul. 2015.

[3] M. E. Stewart *et al.*, "Nanostructured Plasmonic Sensors," *Chem. Rev.*, vol. 108, no. 2, pp. 494–521, Feb. 2008.

[4] M. E. Stewart *et al.*, "Quantitative multispectral biosensing and 1D imaging using quasi-3D plasmonic crystals," *Proc. Natl. Acad. Sci.*, vol. 103, no. 46, pp. 17143–17148, Nov. 2006.

[5] C. J. Choi and S. Semancik, "Multi-resonant plasmonic nanodome arrays for label-free biosensing applications," *Nanoscale*, vol. 5, no. 17, pp. 8138–8145, Sep. 2013.

[6] F. Inci *et al.*, "Nanoplasmonic Quantitative Detection of Intact Viruses from Unprocessed Whole Blood," *ACS Nano*, vol. 7, no. 6, pp. 4733–4745, Jun. 2013.

[7] Q. Wei *et al.*, "Detection and Spatial Mapping of Mercury Contamination in Water Samples Using a Smart-Phone," *ACS Nano*, vol. 8, no. 2, pp. 1121–1129, Feb. 2014.

[8] X. D. Hoa, A. G. Kirk, and M. Tabrizian, "Towards integrated and sensitive surface plasmon resonance biosensors: A review of recent progress," *Biosens. Bioelectron.*, vol. 23, no. 2, pp. 151–160, Sep. 2007.

[9] L. Guo, L. Chen, S. Hong, and D.-H. Kim, "Single plasmonic nanoparticles for ultrasensitive DNA sensing: From invisible to visible," *Biosens. Bioelectron.*, vol. 79, pp. 266–272, May 2016.

[10] H. Jang, C. H. Kwak, G. Kim, S. M. Kim, Y. S. Huh, and T.-J. Jeon, "Identification of genetically modified DNA found in Roundup Ready soybean using gold nanoparticles," *Microchim. Acta*, vol. 183, no. 9, pp. 2649–2654, Sep. 2016.

[11] A. J. Wood, S. Basuray, S. Bok, K. Gangopadhyay, S. Gangopadhyay, and S. A. Grant, "Enhanced DNA Detection Through the Incorporation of Nanocones and Cavities Into a Plasmonic Grating Sensor Platform," *IEEE Sens. J.*, vol. 16, no. 10, pp. 3403–3408, May 2016.

[12] C. Huang, J. Ye, S. Wang, T. Stakenborg, and L. Lagae, "Gold nanoring as a sensitive plasmonic biosensor for on-chip DNA detection," *Appl. Phys. Lett.*, vol. 100, no. 17, p. 173114, Apr. 2012.

[13] W. P. Hall, J. Modica, J. Anker, Y. Lin, M. Mrksich, and R. P. Van Duyne, "A Conformation- and Ion-Sensitive Plasmonic Biosensor," *Nano Lett.*, vol. 11, no. 3, pp. 1098–1105, Mar. 2011.

[14] J. Liang *et al.*, "Silver nanoprism etching-based plasmonic ELISA for the high sensitive detection of prostate-specific antigen," *Biosens. Bioelectron.*, vol. 69, pp. 128–134, Jul. 2015.

[15] H.-I. Peng and B. L. Miller, "Recent advancements in optical DNA biosensors: Exploiting the plasmonic effects of metal nanoparticles," *Analyst*, vol. 136, no. 3, pp. 436–447, 2011.

[16] A. A. Yanik *et al.*, "An Optofluidic Nanoplasmonic Biosensor for Direct Detection of Live Viruses from Biological Media," *Nano Lett.*, vol. 10, no. 12, pp. 4962–4969, Dec. 2010.

[17] D.-K. Kim *et al.*, "Plasmonic Properties of the Multispot Copper-Capped Nanoparticle Array Chip and Its Application to Optical Biosensors for Pathogen Detection of Multiplex DNAs," *Anal. Chem.*, vol. 83, no. 16, pp. 6215–6222, Aug. 2011.

[18] G. K. Joshi, S. Deitz-McElyea, M. Johnson, S. Mali, M. Korc, and R. Sardar, "Highly Specific Plasmonic Biosensors for Ultrasensitive MicroRNA Detection in Plasma from Pancreatic Cancer Patients," *Nano Lett.*, vol. 14, no. 12, pp. 6955–6963, Dec. 2014.

[19] H. Minh Hiep *et al.*, "A localized surface plasmon resonance based immunosensor for the detection of casein in milk," *Sci. Technol. Adv. Mater.*, vol. 8, no. 4, pp. 331–338, May 2007.

[20] A. J. Haes, L. Chang, W. L. Klein, and R. P. Van Duyne, "Detection of a Biomarker for Alzheimer's Disease from Synthetic and Clinical Samples Using a Nanoscale Optical Biosensor," *J. Am. Chem. Soc.*, vol. 127, no. 7, pp. 2264–2271, Feb. 2005.

[21] A. J. Haes, W. P. Hall, L. Chang, W. L. Klein, and R. P. Van Duyne, "A Localized Surface Plasmon Resonance Biosensor: First Steps toward an Assay for Alzheimer's Disease," *Nano Lett.*, vol. 4, no. 6, pp. 1029–1034, Jun. 2004.

[22] Y. Chuo, D. Hohertz, C. Landrock, B. Omrane, K. L. Kavanagh, and B. Kaminska, "Large-Area Low-Cost Flexible Plastic Nanohole Arrays for Integrated Bio-Chemical Sensing," *IEEE Sens. J.*, vol. 13, no. 10, pp. 3982–3990, Oct. 2013.

[23] S. Y. Lee, H. C. Jeon, and S.-M. Yang, "Unconventional methods for fabricating nanostructures toward high-fidelity sensors," *J. Mater. Chem.*, vol. 22, no. 13, pp. 5900–5913, 2012.

[24] H. L. Chen, K. C. Hsieh, C. H. Lin, and S. H. Chen, "Using direct nanoimprinting of ferroelectric films to prepare devices exhibiting bi-directionally tunable surface plasmon resonances," *Nanotechnology*, vol. 19, no. 43, p. 435304, Oct. 2008.

[25] L. Polavarapu and L. M. Liz-Marzán, "Towards low-cost flexible substrates for nanoplasmonic sensing," *Phys. Chem. Chem. Phys.*, vol. 15, no. 15, pp. 5288–5300, 2013.

[26] D. Qin, Y. Xia, and G. M. Whitesides, "Soft lithography for micro- and nanoscale patterning," *Nat. Protoc.*, vol. 5, no. 3, pp. 491–502, 2010.

[27] S. Aksu *et al.*, "Flexible Plasmonics on Unconventional and Nonplanar Substrates," *Adv. Mater.*, vol. 23, no. 38, pp. 4422–4430, Oct. 2011.

[28] J. He, T. Kunitake, and A. Nakao, "Facile In Situ Synthesis of Noble Metal Nanoparticles in Porous Cellulose Fibers," *Chem. Mater.*, vol. 15, no. 23, pp. 4401–4406, Nov. 2003.

[29] M. Xue, F. Li, and T. Cao, "Fabrication of ultra-fine nanostructures using edge transfer printing," *Nanoscale*, vol. 4, no. 6, pp. 1939–1947, Mar. 2012.

[30] S.-C. Tseng *et al.*, "Eco-Friendly Plasmonic Sensors: Using the Photothermal Effect to Prepare Metal Nanoparticle-Containing Test Papers for Highly Sensitive Colorimetric Detection," *Anal. Chem.*, vol. 84, no. 11, pp. 5140–5145, Jun. 2012.

[31] L.-L. Qu, D.-W. Li, J.-Q. Xue, W.-L. Zhai, J. S. Fossey, and Y.-T. Long, "Batch fabrication of disposable screen printed SERS arrays," *Lab Chip*, vol. 12, no. 5, pp. 876–881, 2012.

[32] H. Hiramatsu and F. E. Osterloh, "A Simple Large-Scale Synthesis of Nearly Monodisperse Gold and Silver Nanoparticles with Adjustable Sizes and with Exchangeable Surfactants," *Chem. Mater.*, vol. 16, no. 13, pp. 2509–2511, Jun. 2004.

[33] D. Shir, Z. Ballard, and A. Ozcan, "Flexible Plasmonic Sensors," *IEEE J. Sel. Top. Quantum Electron.*, vol. PP, no. 99, pp. 1–1, 2015.

[34] W. W. Yu and I. M. White, "Inkjet-printed paper-based SERS dipsticks and swabs for trace chemical detection," *The Analyst*, vol. 138, no. 4, pp. 1020–1025, 2013.

[35] C. H. Lee, L. Tian, and S. Singamaneni, "Paper-Based SERS Swab for Rapid Trace Detection on Real-World Surfaces," *ACS Appl. Mater. Interfaces*, vol. 2, no. 12, pp. 3429–3435, Dec. 2010.

[36] H.-Y. Wu and B. T. Cunningham, "Point-of-care detection and real-time monitoring of intravenously delivered drugs via tubing with an integrated SERS sensor," *Nanoscale*, vol. 6, no. 10, pp. 5162–5171, Apr. 2014.

[37] C.-J. Heo *et al.*, "Robust plasmonic sensors based on hybrid nanostructures with facile tunability," *J. Mater. Chem.*, vol. 22, no. 28, pp. 13903–13907, Jun. 2012.

[38] E. McLeod, Q. Wei, and A. Ozcan, "Democratization of Nanoscale Imaging and Sensing Tools Using Photonics," *Anal. Chem.*, vol. 87, no. 13, pp. 6434–6445, Jul. 2015.

[39] A. Ozcan, "Mobile phones democratize and cultivate next-generation imaging, diagnostics and measurement tools," *Lab. Chip*, vol. 14, no. 17, pp. 3187–3194, Sep. 2014.

[40] S. Krishnamoorthy, "Nanostructured sensors for biomedical applications — a current perspective," *Curr. Opin. Biotechnol.*, vol. 34, pp. 118–124, Aug. 2015.

[41] W. Knoben, S. H. Brongersma, and M. Crego-Calama, "Plasmonic Au islands on polymer nanopillars," *Nanotechnology*, vol. 22, no. 29, p. 295303, Jul. 2011.

[42] F. Inci *et al.*, "Multitarget, quantitative nanoplasmonic electrical field-enhanced resonating device (NE2RD) for diagnostics," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 32, pp. E4354-4363, Aug. 2015.

[43] A. Sereda, J. Moreau, M. Canva, and E. Maillart, "High performance multi-spectral interrogation for surface plasmon resonance imaging sensors," *Biosens. Bioelectron.*, vol. 54, pp. 175–180, Apr. 2014.

[44] A. E. Cetin *et al.*, "Handheld high-throughput plasmonic biosensor using computational on-chip imaging," *Light Sci. Appl.*, vol. 3, no. 1, p. e122, Jan. 2014.

[45] A. F. Coskun, A. E. Cetin, B. C. Galarreta, D. A. Alvarez, H. Altug, and A. Ozcan, "Lensfree optofluidic plasmonic sensor for real-time and label-free monitoring of molecular binding events over a wide field-of-view," *Sci. Rep.*, vol. 4, Oct. 2014.

[46] Y. Y. Huang *et al.*, "Stamp Collapse in Soft Lithography," *Langmuir*, vol. 21, no. 17, pp. 8058–8068, Aug. 2005.

[47] C.-C. Liang, C.-H. Lin, T.-C. Cheng, J. Shieh, and H.-H. Lin, "Nanoimprinting of Flexible Polycarbonate Sheets with a Flexible Polymer Mold and Application to Superhydrophobic Surfaces," *Adv. Mater. Interfaces*, vol. 2, no. 7, May 2015.

[48] D. B. Farmer, P. Avouris, Y. Li, T. F. Heinz, and S.-J. Han, "Ultrasensitive Plasmonic Detection of Molecules with Graphene," *Acs Photonics*, vol. 3, no. 4, pp. 553–557, Apr. 2016.

[49] J. Kim *et al.*, "Electrical Control of Optical Plasmon Resonance with Graphene," *Nano Lett.*, vol. 12, no. 11, pp. 5598–5602, Nov. 2012.

[50] S. C. B. Gopinath, T. Lakshmipriya, Y. Chen, W.-M. Phang, and U. Hashim, "Aptamer-based 'point-of-care testing,'" *Biotechnol. Adv.*, vol. 34, no. 3, pp. 198–208, Jun. 2016.

[51] J. Feng *et al.*, "Nanoscale Plasmonic Interferometers for Multispectral, High-Throughput Biochemical Sensing," *Nano Lett.*, vol. 12, no. 2, pp. 602–609, Feb. 2012.

[52] Y. Gao, Z. Xin, B. Zeng, Q. Gan, X. Cheng, and F. J. Bartoli, "Plasmonic interferometric sensor arrays for high-performance label-free biomolecular detection," *Lab. Chip*, vol. 13, no. 24, pp. 4755–4764, Nov. 2013.

[53] S. Olcum, A. Kocabas, G. Ertas, A. Atalar, and A. Aydinli, "Tunable surface plasmon resonance on an elastomeric substrate," *Opt. Express*, vol. 17, no. 10, pp. 8542–8547, May 2009.

[54] M. Kahraman, P. Daggumati, O. Kurtulus, E. Seker, and S. Wachsmann-Hogiu, "Fabrication and Characterization of Flexible and Tunable Plasmonic Nanostructures," *Sci. Rep.*, vol. 3, Dec. 2013.

[55] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *J. R. Stat. Soc. Ser. B*, vol. 58, pp. 267–288, 1994.

[56] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005.

[57] E. J. Candès and Y. Plan, "Near-ideal model selection by ℓ1 minimization," *Ann. Stat.*, vol. 37, no. 5A, pp. 2145–2177, Oct. 2009.

[58] P. Zhao and B. Yu, "On model selection consistency of Lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, Nov. 2006.

[59] D. Brady, *Optical Imaging and Spectroscopy*. Wiley, 2009.

[60] Ake Björck, *Numerical Methods for Least Squares Problems*. SIAM, 1996.

[61] "LED Indication - Discrete | Optoelectronics | DigiKey." [Online]. Available: http://www.digikey.com/product-search/en/optoelectronics/led-indication-discrete/524729. [Accessed: 18-Nov-2016].

Chapter 5. Deep-learning spectroscopy

After showing computational sensing approaches to iterative assay development and band selection for spectroscopic-based sensing, I will now discuss a more fundamental biophotonic sensing platform: the spectrometer. This chapter introduces a deep learning based approach to spectroscopy, enabling fundamentally different optical hardware. This computational approach can pose unique advantages for certain applications in terms of performance, cost, and size over the traditional grating and CCD based designs due to its radically different design principles. Furthermore, this approach to spectroscopy opens the doors for computational design of future spectrometers through a user-defined cost function that e.g. optimized around application-specific spectral bands, or perhaps trained to classify different classes of spectral signals.

This work is currently in the stages of being drafted for peer-review: Z. Ballard, C. Brown, A. Goncharov, M. Fordham, A. Clemens, Y. Rivenson, A. Ozcan, "*Deep learning spectroscopy*" (In preparation).

5.1 Introduction

Spectroscopy is a fundamental tool in biophotonics, and is the cornerstone read-out instrumentation for optical sensors which exhibit a spectral response upon interaction with a sensing analyte. Spectrometers are a well-established measurement tool which rely on a grating to separate different wavelengths of light and project this spatial information onto a photosensor (typically a line-CCD). Though simple and effective, this grating based design poses some limitations to the sensing hardware. For example, it requires photodetectors which are specialized in their design (as opposed to the ubiquitous CMOS) resulting in more expensive manufacturing costs. Additionally, the diffraction properties of the grating can be restrictive for compact, high-

resolution designs at longer wavelengths leading to the development of compressive spectroscopy designs [1], [2]. Therefore, leveraging recent trends in deep-learning, I propose an alternative spectroscopic design which works through encoding optical spectra through a spatially multiplexed plasmonic-filter chip, and reconstructing the spectra using the encoded components and a trained neural network.

5.2 Materials and Methods

The proposed deep learning spectrometer is shown in Figure 5.1a. An arbitrary spectrum is coupled into a large core fiber, which outputs the light onto a spectral encoding chip (inset Figure 5.1a.) containing a number of geometrically unique (in terms of periodicity and diameter) nano-hole array tiles (each 100x100 um). The light is then sampled with a monochrome CMOS camera (Basler Dart, daA1280-54um) after being transmitted through the spectral encoding chip, resulting in a raw image which is pixel-binned in order to recover the encoded spectral signal (Figure 5.1b and c). A neural network is then used to reconstruct the input spectra using the pixel binned encoded spectral signals.

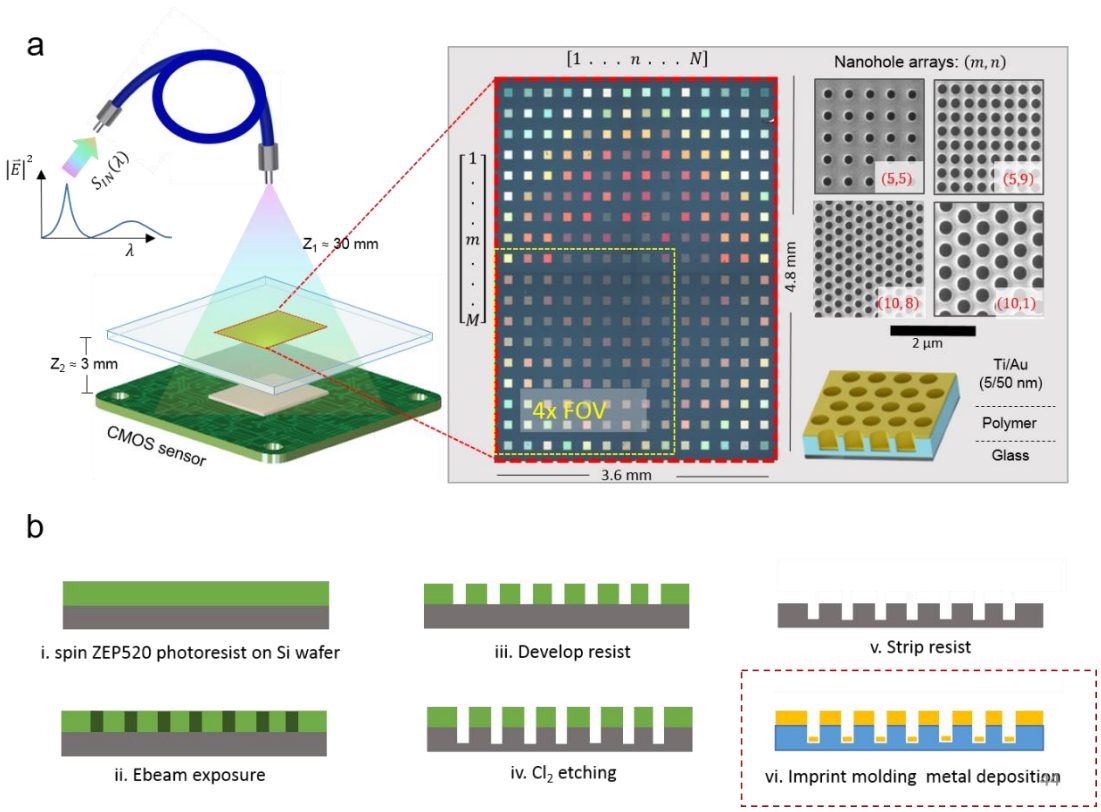


Figure 5.1. Overview of the spectral encoder (a) The hardware set up for deep-learning spectroscopy, including a large-core multimode fiber to sample input spectra, along with the spectral encoding chip (details shown on the inset to the right) and monochrome CMOS sensor for acquisition. The inset shows a stitched bright field image of the spectral encoding chip including taken with a 4x objective in transmission mode. The 4x field-of-view (FOV) is shown with the dotted yellow line. The SEM images to the right show a sampling of the geometrically diverse nanohole arrays created on the spectral encoding chip. **(b)** The silicon wafer (grey) fabrication for the spectral encoder. The step outlined in red can be performed outside of a clean room in high-throughput as a result of the soft lithography process.

The spectral encoder

The spectral encoding step demonstrated in this work is similar to a few compressive-sensing based spectrometers reported in the literature except that here the encoding step is performed completely passively and statically, requiring no dynamic or mechanical scanning components. Furthermore the acquisition step is performed in a snap-shot which leverages the ultra-wide field of view of the on-chip imaging configuration show in Figure 5.1a [1]–[3]. Additionally, the

spectral coder is fabricated by a scalable and high-throughput fabrication technique, i.e. soft-lithography (red-dotted line, Figure 5.1b), which is described in the Appendix of Chapter 4. Therefore, with large-scale patterning through imprint molding and only one metallization step (i.e. sputter coating), the encoder can be made to be ultra-low cost, leaving the limiting cost-factor for this spectrometer to be the CMOS sensor, a ubiquitous technology that benefits from economies of scale.

These added practical benefits of the spectral encoder do however come with a caveat; the spectral encoding chip is a less-than ideal encoder in terms of the linearly independent basis with which it samples input spectra. Specifically the localized plasmon resonance filters used in this work exhibit broad pass-bands with considerable overlap of their transmission functions. To overcome this restraint the deep learning framework is used as an alternative to the well-established compressive sensing frameworks, in order to achieve more accurate spectral reconstructions (Figure 5.1.d).

Training the deep learning spectrometer

Encoded spectral information is acquired over eight, fixed exposure times spanning 100,000 μs to 312.5 μs (decreasing by 50% over each acquired image in the set). These images are then converted to a single High Dynamic Range (HDR) image and pixel-binned with a square mask (60x60 pixels) centered about each encoder tile. Each mask is then further subdivided into a 3 by 3 configuration of 20x20 pixel masks, and each feature space is sampled through pixel-averaging, yielding M = 2,268 input features from 252 encoder tiles each dived into 3x3-feature spaces. These features are then input into a neural network with three fully connected hidden layers, each with 2048 nodes (Figure 5.2). A dropout of 0.2 was determined empirically through its minimum

validation loss compared to 0, 0.1, 0.3, 0.4 and 0.5 dropout. Each layer of the network includes batch normalization, and Rectified Linear Units (ReLU) activation at its output, defined by $f(x) = \max(0, x)$, the input layer contains a standardization to the mean g_m and standard deviation, σ_m, of the training set,

$$g'_m = \frac{g_m - \overline{g_m}}{\sigma_m}, \qquad \text{Eq. 5.1}$$

where g_m represents the encoded signal from the m[th] input feature. The output layer of the network includes 1315 nodes representing each pixel on the ground-truth spectrometer.

The network is trained by a mean squared error cost function,

$$J(\boldsymbol{s}, \boldsymbol{s}') = \frac{1}{N}\sum_{i=1}^{N}\sum_{p=1}^{P}\left(s_{i,p} - s'_{i,p}\right)^2 \qquad \text{Eq. 5.2}$$

where $\boldsymbol{s_i}$ is the ground truth spectral information for the i[th] spectra in the validation set. P represents the number of pixels in the output layer (i.e. P = 1315) as recorded by a commercial spectrometer in the case of $\boldsymbol{s_i}$. $\boldsymbol{s'_i}$ is the predicted spectral output from the network.

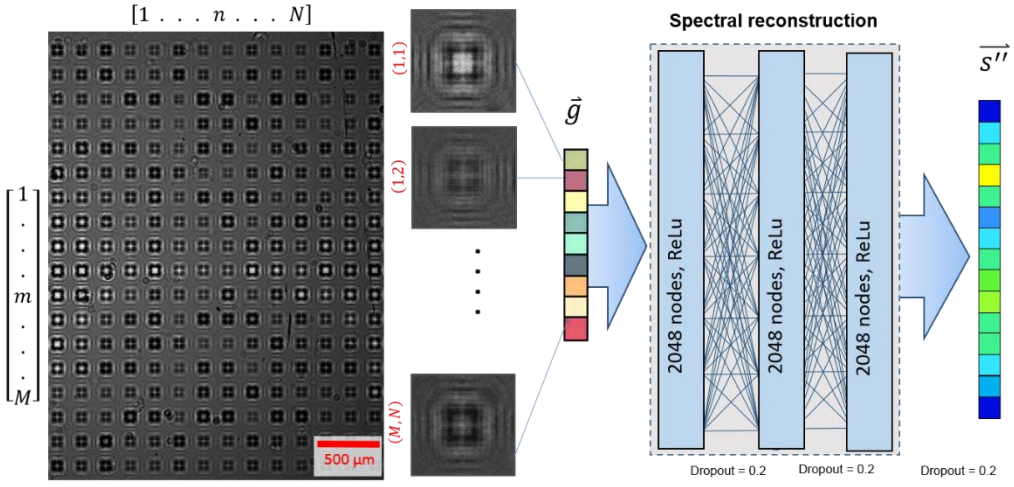


Figure 5.2 Deep-learning processing of encoded spectra (a) Raw image captured by the CMOS sensor of the input spectra after it is transmitted through the spectral encoding chip. The neural network inputs and outputs are shown

and denoted as $\vec{g}$ and $\overrightarrow{s'}$ respectively. **(b)** A representative example of the spectral reconstruction (orange) as compared to the ground truth spectra (blue) measured with a standard 0.2 nm spectrometer.

The deep-learning spectrometer was introduced to over 40,00 randomly generated spectra over 48 hours produced by a super continuum laser (Fianium) with up to eight bands, each 3 nm in bandwidth, spanning 450 to 750 nm (Figure 5.3). Each input spectra is recorded in real-time by a standard spectrometer (ThorLabs, CCS100) alongside the acquired images containing the encoded spectra. The images are then processed as previously describe, and the neural network is then trained over 60 hours using an NVIDIA GeForce FTX1080 Ti GPU with a learning rate of 0.0001 and batch size of 1024.

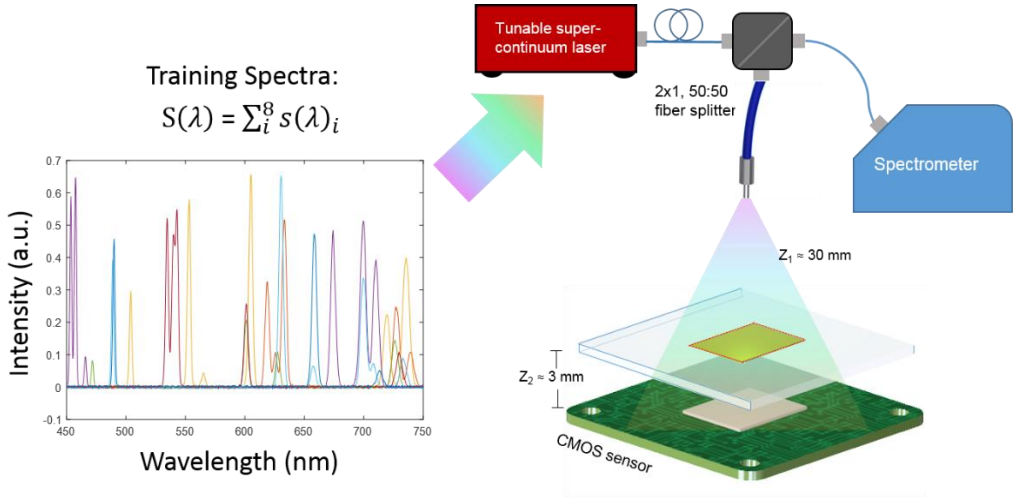


Figure 5.3 The deep learning spectroscopy training procedure. Examples of randomly generated input spectrum (show in different colors, left) used for training the neural network. The randomly generated spectra are produced by a super-continuum laser (shown in red), and the input signal is split into the deep-learning spectrometer as well as a standard spectrometer for recording the ground-truth data.

5.3 Results and future work

The blindly tested spectral reconstructions are shown in Figure 5.4. Overall the reconstructed spectra fit very well to the ground truth data not only in terms of their intensity, but also their peak

localization. Over 10,000 blindly tested input spectra, the average peak localization accuracy was calculated to be ± 0.253 nm, which is within 10% of the resolution of the ground-truth spectrometer (0.23 nm). Additionally a minimum of 3.8 nm was found to be the smallest resolvable distance between neighboring peaks.

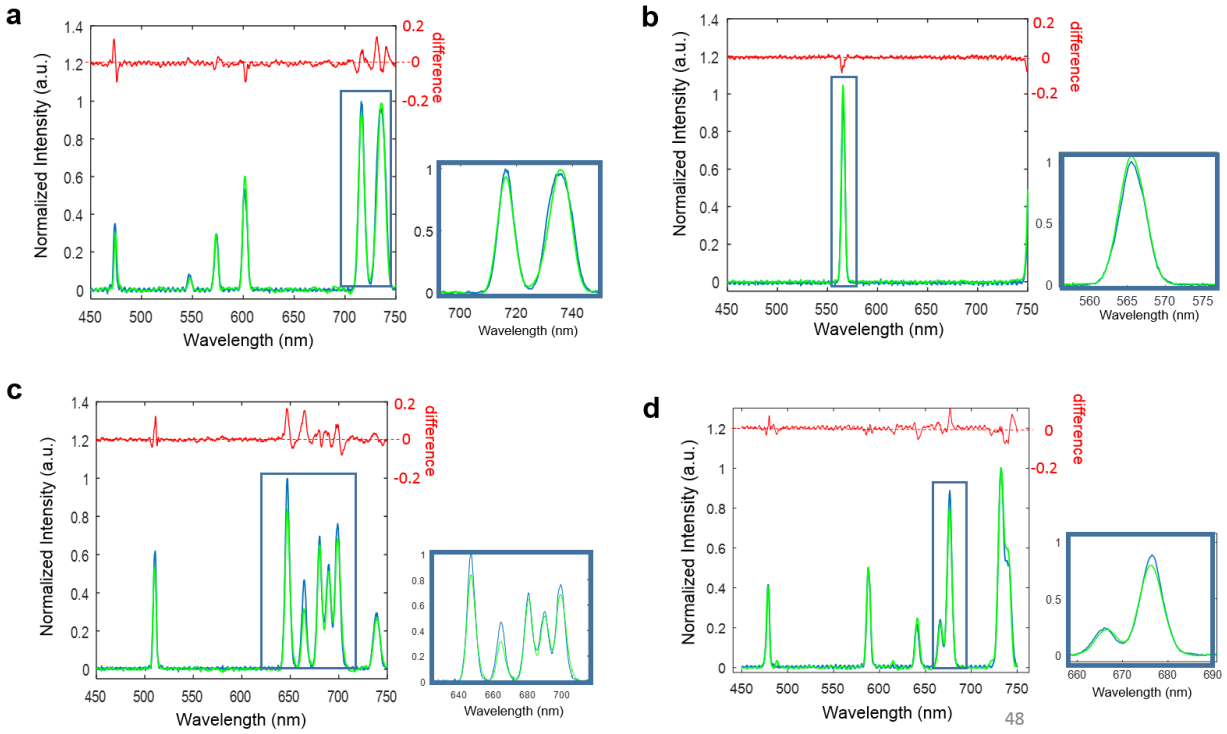


Figure 5.4. Spectral reconstructions via deep learning spectroscopy. (a)-(d) Various randomly generate spectrum where the ground-truth spectrometer is shown in blue and the deep learning reconstructions are overlaid in green. A zoomed-in view of key bands is shown to the right of each graph. The red line shows the difference between the ground-truth and deep learning output.

For future work, the stability over time of the deep learning algorithm must be addressed. Minute changes in temperature or alignment can cause changes in the transmission properties of the plasmonic tiles, as well as sub-pixel shifts between the input aperture, spectral encoder, and CMOS camera. The quality of spectral reconstructions must be robust to these inevitable sources of drift and noise in order to overcome impending practical sensing challenges. Dynamic

registration that occurs after a set ‘calibration period’ can be implemented to computationally correct for any minor changes in alignment. Additionally, on-line learning can be utilized to fine-tune the weights of the network based on a recurring set of fixed input spectra. Such an approach would be similar to current calibration methods used for standard spectrometers in which the line-scan CCD camera must be digitally adjusted against some gold-standard input spectrum in order to obtain subsequent accurate readings.

Additionally, new encoder designs can be considered which more densely pack the plasmonic filter encoding elements, and perhaps lead to novel designs for hyperspectral imaging applications, along with other computationally determined designs configured against a user-defined cost function and training set for targeted applications.

References

[1] Y. August and A. Stern, “Compressive sensing spectrometry based on liquid crystal devices,” *Opt. Lett.*, vol. 38, no. 23, pp. 4996–4999, Dec. 2013.

[2] “High-performance and scalable on-chip digital Fourier transform spectroscopy | Nature Communications.” [Online]. Available: https://www.nature.com/articles/s41467-018-06773-2. [Accessed: 10-Jan-2019].

[3] D. J. Starling, I. Storer, and G. A. Howland, “Compressive sensing spectroscopy with a single pixel camera,” *Appl. Opt.*, vol. 55, no. 19, pp. 5198–5202, Jul. 2016.

Chapter 6. Conclusions

Computationally designed sensing systems will provide various exciting opportunities, however, like any emerging technology, there exist inherent challenges which must be understood and addressed. Specifically, sensing systems designed by statistical learning approaches inevitably share the well-known pitfalls of machine learning. For example, access to large amounts of rigorously vetted, well-characterized, and diverse training data can sometimes be infeasible for a given sensing system. In the field of bio-sensing, for instance, where cost-per-test can be sometimes high, sensing outcome can depend on a myriad of potentially un-accounted and confounding factors such as the choice of vendor, shelf-life of reagents, ambient conditions (e.g., temperature, humidity etc.), and cross contamination, among other factors. Therefore, it becomes a central challenge to ensure that the training data sets are not biased or severely contaminated by noise sources characteristic of only the training set. Such scenarios would lead to overfitting, where learned algorithms can fail to generalize, sometimes *catastrophically* upon the introduction of sensing inputs that deviate only slightly from what has already been explicitly 'learned.' Overfitting can also occur when training data sets are not appropriately diversified in terms of e.g., dynamic range, resolution or sensor-to sensor performance variability, among others. To combat this issue, Chapters 2 and 3 discuss the inclusion of batch-specific information to be input into the network, along with on-line batch-specific standardization to account for statistical drift in the inputs. I have showed that even with modest data sets (~100 tests, i.e. training instances), deep learning networks can lock-on to meaningful signals without overfitting. It is important to note that, to achieve this, the networks implemented in chapter 2 and 3 utilized the maximum amount of drop-out rate, which helps regularize the model during the training process. The training data

sets were also cross-validated across multiple k-fold validation sets, that included sensors made from a diverse set of fabrication batches.

Discussion of some of these challenges may lead the reader to believe that a properly executed learning-based computational sensing approach is prohibitively time and resource intensive for the design phase of a given intelligent sensing system. However, the computational sensor design methods highlighted in this Dissertation can always be implemented in the subsequent iterations of the design, i.e., they need not be implemented as a first step of prototyping. In fact, these statistical learning methods are ideal for iterative and adaptive design strategies as they converge to locally optimal, cost-effective solutions for application specific sensing scenarios, without the need for a complete understanding of noise contributions, physical principles and their complex interactions and modelling. For example, a computational sensor that makes a catastrophic sensing error, such as missing a significant event or analyte, or simply missing a "small data" related outlier event, could be mended by readjusting and optimizing its measurement features and their relative weights once such errors are identified. At the iterative design phase of a computational sensor, one can redefine or adjust the cost function of the sensor design framework to appropriately put weights to avoid (or highly penalize) certain classes of errors that might lead to catastrophic outcomes. The same is true for correcting the failures or errors introduced by new uses of a sensor in a region of the world that it was not initially designed for. Therefore, this Dissertation suggests that data-driven computational sensor design approaches provide a scalable, cost-effective and dynamic framework that can easily be adjusted on the go as new data sets are created. Stated differently, such computational sensors can learn, evolve and become more robust as they are being used more and more.

Taken together, I envision the methodologies discussed in this Dissertation permeating the design phase of biosensing hardware, thereby fundamentally changing and challenging traditional, intuition-driven sensor and read-out designs, in favor of application-targeted and perhaps non-intuitive implementations. Such computational sensors can therefore enable new and widely distributed applications that will benefit from ‘big data’ analytics and IoT solutions to create powerful sensing networks, impacting fields of biomedical diagnostics, environmental sensing, and global health, among others.