HHS Public Access

Author manuscript

Anal Chem. Author manuscript; available in PMC 2022 June 14.

Published in final edited form as:

Anal Chem. 2022 January 18; 94(2): 1195–1202. doi:10.1021/acs.analchem.1c04379.

Algorithmically Guided Optical Nanosensor Selector (AGONS): Guiding Data Acquisition, Processing, and Discrimination for Biological Sampling

Christopher W. Smith[†,‡], **Mustafa Salih Hizir**[†], **Nidhi Nandu**[†], **Mehmet V. Yigit**[*,†,‡]

[†]Department of Chemistry, University at Albany, State University of New York, 1400 Washington Avenue, Albany, New York 12222, United States.

[‡]The RNA Institute, University at Albany, State University of New York,1400 Washington Avenue, Albany, New York 12222, United States.

Abstract

Here we report a biomarker-free detection of various biological targets through a programmed machine learning algorithm and an automated computational selection process termed AGONS. The optical data processed/used by algorithms are obtained through a nanosensor array selected for a library of nanosensors through AGONS. The nanosensors are assembled using two-dimensional nanoparticles (2D-nps) and fluorescently labeled single-stranded DNAs (F-ssDNAs) with random sequences. Both 2D-np and F-ssDNA components are cost-efficient and easy to synthesize; allowing for scaled-up data collection essential for machine learning modeling. The nanosensor library was subjected to various target groups including proteins, breast cancer cells, and let-7 miRNA mimics. We have demonstrated that AGONS could select the most essential nanosensors while achieving 100% predictive accuracy in all cases. With this approach, we demonstrate that machine learning can guide the design of nanosensor arrays with greater predictive accuracy while minimizing manpower, material cost, computational resources, instrumentation usage, and time. The biomarker-free detection attribute makes this approach readily available for biological targets without any detectable biomarker. We believe that AGONS can guide optical nanosensor array setups, opening broader opportunities through a biomarker-free detection approach for most challenging biological targets.

Graphical Abstract

[*]Corresponding Author Mehmet Yigit: myigit@albany.edu, Tel: 518-442-3002.

Author Contributions

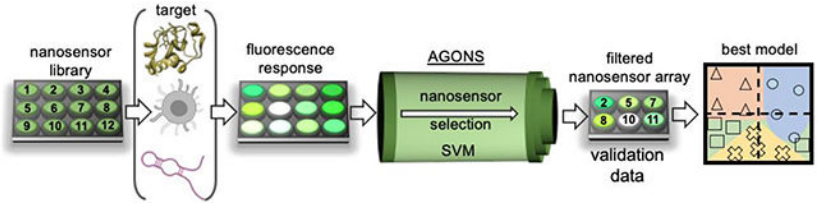
MVY and CWS conceived the study. MVY designed the experiments. MSH collected the protein, breast cancer cell line and miRNA mimic datasets and performed PLSDA modeling. CWS collected the FBS datasets. CWS designed, programmed, and performed modeling with AGONS. NN helped with assay development and contributed intellectually. MVY and CWS wrote the paper.

Present Addresses

The current affiliation of MSH is Bursa Technical University, Bursa/Turkey.

Supporting Information

The experimental details and additional data are provided in the Supporting Information. The Supporting Information is available free of charge on the ACS Publications website.



Keywords

Machine Learning; feature extraction; nanoparticles; fluorescence; aptamer

Today, artificial intelligence is becoming part of our daily lives. Artificial intelligence, simply, is a computational process by which a machine "thinks" through complex tasks and learns from continuous data input and analysis cycles.[1] Artificial intelligence is a powerful tool for decoding complex data matrices. It is anticipated that soon, most of the information from many scientific advances and industrial processes will be paired with artificial intelligence to resolve the complexities of data. The information processed/learned through artificial intelligence is used in data-driven decision-making.

A subset of artificial intelligence, termed machine learning, is widely applied to analyze and understand complex patterns in data to gain new insights and return accurate predictions.[2-4] Decision-making power of machine learning has been evaluated against some current challenges including hurricane forecasting,[5] computer vision plant disease diagnosis,[6] speeding chip production during the ongoing global chip shortage crisis,[7], and more. In chemistry, researchers have evaluated machine learning for various bioengineering processes and biochemical analyses. In addition to using machine learning for the analysis of existing biochemical data, it can be paired with nanotechnology for new experimental data collection leading to advancements in research. Though there are efforts to pair nanotechnology with machine learning, current results achieve only a glimpse of the potential to come.[3,8,9]

The advantage of using machine learning in nanotechnology-driven data acquisition is its inherent automated model building and pattern-seeking capabilities.[1,8] Specifically, processing the data obtained through nanotechnology platforms and analysis by machine learning have led to advancements in precision medicine and agriculture.[1,8] For example, the combination of machine learning and nanotechnology has made an impact on cancer diagnostics and drug design.[9-11] Zhan *et al.* (2020) demonstrated the prediction of the protein corona assemblies on nanoparticle formulations using machine learning algorithms.[11] Recently, Reker *et al.* (2021) evaluated machine learning for self-assembling nano-drug designs.[9] Alafeef, Srivasta, and Pan (2020) developed a neural network architecture to discriminate breast cancer stages using carbon dots.[10] The aforementioned scientific initiatives, along with many others, suggest that machine learning approaches can translate complex experimental data into meaningful and valuable information.

Nanoparticle formulations can provide bulk and high-dimensional data through their many attributes including material composition, size, charge, surface modifications, and physiochemical properties.[9,12,13] Though plentiful qualitative and quantitative information

are stored in nanoparticles, a major prerequisite to machine learning implementation is unleashing this big data, and acquiring it for building a reliable prediction model.[2,3,9] However, the acquisition of the large datasets from nanoparticle formulations can be hindered due to many of the multi-step, complicated, low-yielding synthetic procedures and expensive modifications.[14,15] Consequently, limited data can create a poor quality model that may not be able to generalize properly.[15,16], Especially in sensing applications, limitations by insufficient data, can raise poor quality pattern recognition, causing model overfitting and incorrect predictions.[2,4,17]

To overcome the aforementioned challenges, the employment of easy-to-produce unmodified nanoparticles is well suited for ML applications. Though there are numerous unmodified nanoparticle formulation possibilities, the two-dimensional nanoparticles (2D-nps) are easy and cost-efficient to produce in bulk-scale from their naturally existing raw precursors. 2D-nps exhibit highly useful optical properties that have been implemented for fluorometric analysis of biomacromolecules and biosystems.[18-22] Reproducible data acquisition with these materials makes them robust platforms for molecular diagnostics. Owing to all these features, the 2D-nps are ideal candidates for generating massive and reliable data for ML applications.[23,24]

Previously, we have utilized the surface adsorption properties of 2D-nps: nanographene oxide (nGO), MoS_2, and WS_2 on FAM-labeled single-stranded DNA (F-ssDNA) to build a single nanosensor for the detection of various biological elements. Subsequently, we assembled a nanosensor array, composed of several 2D-np/F-ssDNA nanosensors, for various biosensing studies, (Scheme 1a-b).[18-21] We have demonstrated that such nanosensor array design is highly powerful for the identification of multiple targets through a biomarker-free approach.

Simply, a single nanosensor component of an array is assembled by the non-covalent adsorption of F-ssDNA on a 2D-np surface, (Scheme S1a). Upon adsorption, the fluorescence signal of the F-ssDNA is quenched. When introduced, a target biomolecule of interest fractionally displaces the F-ssDNA from the 2D-np surface generating a partial fluorescence recovery, (Scheme S1b).[18] The nanosensor array is put together using various combinations of nanosensors prepared from different F-ssDNAs and different 2D-np types, (Scheme 1a-b).[18-21] The F-ssDNAs vary by sequence and length, while 2D-nps are selected from nGO, MoS_2, and WS_2. Therefore, each nanosensor is composed of a different 2Dnp or F-ssDNA combination. Consequently, each nanosensor displays a unique release and fluorescence signal intensity in the presence of the target and when combined in an array format, provides a unique optical “fingerprint pattern”, (Scheme 1c).[18-21] In contrast to the bind-and-release approach in target-specific sensing designs, this methodology relies on displacement of F-ssDNA through its competition with a target molecule.

This combinatorial nanosensor array has been incorporated with partial least squares discriminant analysis (PLSDA) statistical modeling, (Scheme 1a-c). We have demonstrated that PLSDA was powerful enough to discriminate bacteria, breast-cancer cells, proteins, and lethal-7 (let-7) miRNA mimics.[18-21] However, a drawback to using PLSDA modeling was that it was limited to solving complex patterns within the data, tends to overfit, and

performed at predictive accuracies of only 80-90%.[18-21,25] Furthermore, it was challenging to select only the essential combination of nanosensors for further data acquisition and predictive modeling.[18-21]

Here, we report a highly powerful in-lab-built machine learning approach that analyzed our breast cancer, protein, and let-7 data and predicted at 100% accuracy in all cases.[18,19] The approach, termed *Algorithmically Guided Optical Nanosensor Selector* algorithm (AGONS), selects only the essential nanosensors in the array while providing the highest predictive accuracy, (Scheme1 d-f). In this approach, a library of nanosensors is screened, the nonessential ones are filtered out and the automated classification/discrimination steps are performed through the data obtained from an array of the essential nanosensors. Such selection procedure saves time, manpower, computational resources, material, instrumentation cost, and opens a wider implementation space for array-based diagnostics. To demonstrate the capability of AGONS, we revisited our breast cancer cell-line, protein, and let-7 datasets.[18,19] In all three datasets, AGONS outperformed PLSDA methods without any additional data collection procedures.[18,19]

Results and Discussion

Utilizing AGONS allows for the selection of the essential nanosensors in the array, data transformation, reduction of the data dimensions, and classification by an estimator algorithm, (Scheme 2, Methods in Supp. Data). The goal of incorporating AGONS in the nanosensor array is to minimize the dimensionality of the nanosensor array for improved data acquisition and achieve higher predictive accuracy. Typically, constructing a machine learning model is done by splitting a dataset into two where the first split is the training data of 60-80% and the second split is the validation set of 10-20% data.[2,3] Then a third dataset blinded from both the training and validation steps is used to make a final assessment of the model.[2] We divided our dataset at a ratio of 6:1:3 (~60% training, ~10% validation, and ~30% testing). Each dataset contained minor class imbalances and to compensate, each split from the dataset was stratified. Stratification retains an equal proportion of representative class labels as possible per respective split percentage.[26] After the data was split, the training data was inserted into AGONS for modeling, (Scheme 2a).

The AGONS architecture was programmed for each step to be retained within a pipeline. Pipelines allow for flexible automation of configuring machine learning tasks by extracting important information from complex patterns.[27,28] Stepwise, the training data composed of a change in fluorescence, ΔF (F_{final}-$F_{initial}$), of various nanosensors is inserted into the AGONS pipeline, (Scheme 2a). The first step of the pipeline ranks the nanosensors by an analysis of variance (ANOVA) F-value for univariate selection. Then, the data of only the essential nanosensors against a particular target group is transformed through mean-centering or normalization methods to remove outlier effects. The transformed data is then dimensionally reduced through principal component analysis (PCA), selecting for dimensions that retain at least 95% cumulative variance, (Figure S1). The last step uses an estimator algorithm, support vector machines (SVM), to classify and form decision regions by a hyperplane separation. The SVM estimators are advantageous in both classification time and performance with limited data.[21,29] A randomized search cross-validator was

implemented to observe 1,000 different hyperparameter combinations at 5-fold cross-validation of all steps during training, (Scheme 2a).

After the initial training, the model selects the hyperparameter combinations that provide at least 70% training accuracy, (Scheme 2b). The hyperparameter combinations are then assessed against the validation dataset for calibration. The top-performing parameters are selected for final modeling. Then, the model's overall predictive quality was assessed, a blinded test set from training and validation datasets is inserted, (Scheme 2c). The prediction on the test set was analyzed for its precision, recall, and accuracy by F1 score, (Supporting Data, Methods).

Fluorescence Pattern Discrimination Between Proteins with AGONS.

Initially, we tested a nanosensor array against 5 proteins, alkaline phosphatase (ALP), bovine serum albumin (BSA), β-galactosidase (β-GAL), lipase (LIP), or protease (PRO).[18] The sensor array was composed of 12 nanosensors assembled using three 2D-nps (nGO, MoS_2, and WS_2) and four F-ssDNA sequences (A_{23}, C_{23}, T_{23}, and S_{22}), Table S1. The nanosensor array was tested for the identification of protein unknowns and their concentration. In this setup, each nanosensor component of the array is individually tested with each protein.

Upon testing, a fraction of F-ssDNA is released from the 2D-np surface through the interactions between proteins and the nanosensors. Again, the F-ssDNA displacement event results in a differential fluorescence recovery (ΔF) pattern which depends on the identity of the protein, the type of the 2D-np, and the F-ssDNA in the nanosensor formulation, (Figure 1a, heatmap). Nevertheless, the total data used was composed of the ΔF per nanosensor at 120 measurements using proteins concentration of which was normalized to $Abs_{280} = 0.1$ (Absorbance value equivalent to 0.1 a.u. at 280nm). The dataset was split at a ratio of 6:1:3, (Table S2). The training data was inserted into the pipeline and the validation data was calibrated for the selection of the best model which required only six nanosensors instead of twelve in our earlier study, (Figure 1b-c, Table S2).

The best model separated the data over four PCs accounting for more than 95% total variance, (Figure 1d and S1). When tested on the blinded dataset, the model was able to accurately predict 100% of 36 protein measurements with a minimum ~80% probabilistic certainty per class, (Figure 1e, Table S3). Compared to our previous report that used PLSDA with 93% predictive accuracy,[18] using a model composed of a pipeline with SVM reached 100% accuracy with an array of 50% fewer nanosensors. We verified that the SKF cross-validation was not biased in its splitting by comparing it to an unbiased cross-validator, leave-one-out cross-validation (LOOCV), (Supporting Data, Methods).[15,29] LOOCV performed just as well as SKF, however, LOOCV is computationally more expensive, (Table S3). Thus, we chose to use SKF for further modeling.

Visualizing the PCA pattern and prediction results from training with 60% of the data, we decided to flip the ratio of the data splitting to 3:1:6. Instead, when trained on fewer measurements, model parameterization required a sensor array of only five nanosensors (Figure S2, Table S2). Remarkably, the model was able to predict 100% out of 72

measurements with at least ~60% certainty per class (Figure 1f, Table S3). With such a small amount of data, the model was still robust with its training and generalized enough for prediction at higher blinded measurements. Combining the fluorescence response pattern from the nanosensors with AGONS provided a “one-size fits all” approach to better differentiate more complex targets.

We further tested the combined capabilities of the nanosensors array and AGONS to discriminate proteins in a complex biological matrix using 10% fetal bovine serum (FBS). Solutions of FBS with and without 1 μM ALP were prepared for testing. Using an array of 12 nanosensors, a dataset of 100 measurements was collected (Supporting Data, Methods, Table S1 and S9). An increase in fluorescence intensity was observed over 2hr kinetic study during testing as expected (Figure S3). The data ($\Delta F = F_{final}-F_{initial}$) was split at a 6:1:3 ratio and AGONS discovered that a minimum of three nanosensors was necessary to differentiate FBS from FBS with ALP (Figure S4, Table S2).

Assessing on 30 unknown samples, AGONS predicted at 100% classification accuracy (Figure S4, Table S4, and S8). The high predictive performance of AGONS in complex matrices is a demonstration towards integration of AGONS-assisted nanosensor array in serum-based molecular diagnostics.

Fluorescence Pattern Discrimination Between Breast Cancer Cell-Lines.

We then applied this automated approach of nanosensor selection and classification for breast cancer analysis (Figure 2a).[18] using subtypes luminal A (MCF-7) or triple-negative (MDA-MB-231 and BT-20).[18] The differences in cellular features resulted in a fluorescent “finger-print” dataset using a nanosensor array composed of the same twelve nanosensors used in protein studies, (Figure 2a). The dataset is composed of 72 measurements observing the fluorescence “finger-prints” obtained using ~1,000 counted cells per measurement. The data was split into three as 6:1:3 for training, validation, and testing, respectively. Initially, the training data was used to train the model, cross-validating, and filtering for at minimum 70% accurate parameter combinations. Once filtered, each model was then assessed.

The top-performing model required a dimension of a sensor array with only three nanosensors and accounted for nearly 100% of the total variance at only 2 PCs (Figure 2b-c, Table S2). When challenged against the blinded dataset, the model performed at 100% accuracy for prediction on 22 measurements, (Figure 2c-d, Table S5). The developed model utilized the sigmoid kernel of SVM for hyperplane separation between all three cell lines and could visualize the decision regions within a 2-dimensional plot, (Figure 2d, Table S2). In com (c) Visualization of PCA showing SVM decision regions on predicted blinded data at 6:1:3 split ratio for (d) 100% accuracy.

Comparison to our previous report with 90% accuracy,[18] the model for this study was able to predict with 100% accuracy. Furthermore, when we flip the ratio of training a model to 3:1:6 we found that the best performing model was achieved using an array of three nanosensors instead of twelve which extensively decreases the cost, measurement numbers by 75%, and labor for data collection, (Figure S5, Table S2).

We further assessed if our models' accuracy depended on the training sample size. We analyzed the data with fewer data training samples and observed that we could still maintain 100% predictive accuracy, (Figure S6a-b, Table S2). The observed probabilistic certainties for each class were above 80%, showing that the model was not “confused” between cell line fluorescence patterns, (Table S5). However, as we are assessing only on a pool of three targets, it may be very simple for our algorithm to be highly predictive. Nonetheless, the performative results of AGONS and the nanosensor library paves a road for future studies to address issues such as cell-line authentication, quality control, or contamination, an ongoing effort in our lab.

AGONS-Guided Discrimination of let-7 mimics.

A favorable feature of using 2D-nps with F-ssDNA is the option of programming the DNAs to recognize specific target attributes. The recognition power of the nanosensor can be modulated against nucleic acid targets by altering sequence compositions of F-ssDNAs for hybridization power. Previously, we reported a semi-specific system that utilized five F-ssDNA probes (P1-P5) to detect nine lethal-7 (let-7) targets (a-j), (Table S6).[19] The nine let-7 targets were standardized at 0.1 a.u. at 260nm, around ~50 nM. Using PLSDA, only ~80% accuracy was achieved to discriminate between all nine highly homologous let-7 targets.[19] We decided to test AGONS against the nine let-7 targets by splitting the 198 measurements at a 6:1:3 ratio.

In our previous report, a sensor array of 15 nanosensors was used to provide a complex ΔF signal pattern against nine let-7 targets (Figure 3a).[19] We assessed the feature selection step of AGONS to resolve a 15 nanosensor complexity and found that as little as four nanosensors were sufficient to construct the best model for prediction (Figure 3b). After the data was mean-centered, it was selected for best-unsupervised separation at four PCs for both modeling and visualized on three PCs (Figure 3c). When tested against the blinded test set, the model predicted at 100% accuracy on 60 measurements at minimum probability scores of ~60% or higher per class, (Figure 3d, Table S7). Our current use of AGONS allowed for 100% predictive accuracy as compared to our previous report which was ~79% using PLSDA.[19]

Here, AGONS was incredibly efficient at separating nine similar target patterns. In this case, the best model was achieved using only an array of four nanosensors rather than fifteen (reducing the sensor array size by almost four-folds) in our previous publication using the same experimental dataset. Finally, we challenged the pipeline to predict 119 blinded measurements when trained with only 30% of the data (split ratio 3:1:6) (Table S2). When testing the model on 117 samples, 98% were predicted accurately (Figure S7, Table S2). Though a high predictive accuracy was achieved, the probabilistic certainty significantly dropped per target class when comparing split ratios 6:1:3 and 3:1:6, which is expected and stresses the importance of training data size (Table S7). Thus, such observation also demonstrates the importance of why this method in the future can improve sensor quality. Incorporating feature selection techniques allow for reduced data acquisition cost while achieving improved model performance by a larger training sample pool. Overall, between all datasets, AGONS proved to be both highly precise and sensitive (Table S8).

Conclusion

Previously, we assembled nanosensor arrays to make breast cancer cells, protein, and let-7 mimic predictions through PLSDA using the experimental fluorescent “finger-print” dataset. Though the results were highly promising, we hypothesized that the prediction accuracy can be significantly advanced when the dataset was analyzed through machine learning. Here, we outperformed our previous methodology by constructing a novel ML algorithm, termed AGONS, that can select an array of most essential nanosensors with much stronger prediction accuracy. The modeling was able to predict at 100% accuracy on blinded samples when trained using 60% and tested on 30% of the data. This method was very efficient at utilizing smaller amounts of data to train (30%) and learning patterns to predict even larger amounts of blinded data (60%). Even with a higher number of classes in let-7 prediction cases, the machine learning pipeline was able to discriminate between all targets with 100% efficiency. Overall, we report that the ML modeling shown through AGONS, and optical data obtained through our nanosensor array platform can be employed for the discrimination and identification of a variety of biological markers. The AGONS algorithm can be applied to other optical sensor array designs and can result in greater predictive accuracies. We believe this platform is highly valuable for medicinal or agricultural automated diagnosis/quality control applications.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENT

This work is supported, in part, by the USDA National Institute of Food and Agriculture, AFRI project (Grants 2018-67021-27973 and 2017-07822), and National Institutes of Health (Grant 1R15GM12811501).

REFERENCES

(1). Adir O; Poley M; Chen G; Froim S; Krinsky N; Shklover J; Shainsky-Roitman J; Lammers T; Schroeder A Integrating Artificial Intelligence and Nanotechnology for Precision Cancer Medicine. Adv. Mater 2020, 32 (13), 1901989.

(2). Mater AC; Coote ML Deep Learning in Chemistry. J. Chem. Inf. Model 2019, 59 (6), 2545–2559. [PubMed: 31194543]

(3). Wang AYT; Murdock RJ; Kauwe SK; Oliynyk AO; Gurlo A; Brgoch J; Persson KA; Persson KA; Sparks TD Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices. Chem. Mater 2020, 32 (12), 4954–4965.

(4). Moosavi SM; Jablonka KM; Smit B The Role of Machine Learning in the Understanding and Design of Materials. J. Am. Chem. Soc 2020, 142 (48), 20273–20287.

(5). Sun X; Xie L; Shah SU; Shen X A Machine Learning Based Ensemble Forecasting Optimization Algorithm for Preseason Prediction of Atlantic Hurricane Activity. Atmosphere. 2021.

(6). Hughes DP; Salathe M An Open Access Repository of Images on Plant Health to Enable the Development of Mobile Disease Diagnostics. arXiv 2015 (arXiv:1511.08060).

(7). Mirhoseini A; Goldie A; Yazgan M; Jiang JW; Songhori E; Wang S; Lee YJ; Johnson E; Pathak O; Nazi A; Pak J; Tong A; Srinivasa K; Hang W; Tuncer E; Le QV; Laudon J; Ho R; Carpenter R; Dean J A Graph Placement Methodology for Fast Chip Design. Nature 2021, 594 (7862), 207–212. [PubMed: 34108699]

(8). Zhang P; Guo Z; Ullah S; Melagraki G; Afantitis A; Lynch I Nanotechnology and Artificial Intelligence to Enable Sustainable and Precision Agriculture. Nat. Plants 2021, 7 (7), 864–876. [PubMed: 34168318]

(9). Reker D; Rybakova Y; Kirtane AR; Cao R; Yang JW; Navamajiti N; Gardner A; Zhang RM; Esfandiary T; L'Heureux J; von Erlach T; Smekalova EM; Leboeuf D; Hess K; Lopes A; Rogner J; Collins J; Tamang SM; Ishida K; Chamberlain P; Yun DS; Lytton-Jean A; Soule CK; Cheah JH; Hayward AM; Langer R; Traverso G Computationally Guided High-Throughput Design of Self-Assembling Drug Nanoparticles. Nat. Nanotechnol 2021, 16 (6), 725–733. [PubMed: 33767382]

(10). Alafeef M; Srivastava I; Pan D Machine Learning for Precision Breast Cancer Diagnosis and Prediction of the Nanoparticle Cellular Internalization. ACS Sensors 2020, 5 (6), 1689–1698. [PubMed: 32466640]

(11). Ban Z; Yuan P; Yu F; Peng T; Zhou Q; Hu X Machine Learning Predicts the Functional Composition of the Protein Corona and the Cellular Recognition of Nanoparticles. Proc. Natl. Acad. Sci. U. S. A 2020, 117 (19), 10492–10499. [PubMed: 32332167]

(12). Sun B; Barnard AS Visualising Multi-Dimensional Structure/Property Relationships with Machine Learning. JPhys Mater. 2019, 2 (3), 034003.

(13). Brown KA; Brittman S; Maccaferri N; Jariwala D; Celano U Machine Learning in Nanoscience: Big Data at Small Scales. Nano Lett. 2020, 20 (1), 2–10. [PubMed: 31804080]

(14). Furxhi I; Murphy F; Mullins M; Arvanitis A; Poland CA Practices and Trends of Machine Learning Application in Nanotoxicology. Nanomaterials. 2020, 10(1), 116.

(15). Krstajic D; Buturovic LJ; Leahy DE; Thomas S Cross-Validation Pitfalls When Selecting and Assessing Regression and Classification Models. J. Cheminform 2014, 6 (1), 10. [PubMed: 24678909]

(16). Liu AL; Venkatesh R; McBride M; Reichmanis E; Meredith JC; Grover MA Small Data Machine Learning: Classification and Prediction of Poly(Ethylene Terephthalate) Stabilizers Using Molecular Descriptors. ACS Appl. Polym. Mater 2020, 2 (12), 5592–5601.

(17). Shiba K; Tamura R; Sugiyama T; Kameyama Y; Koda K; Sakon E; Minami K; Ngo HT; Imamura G; Tsuda K; Yoshikawa G Functional Nanoparticles-Coated Nanomechanical Sensor Arrays for Machine Learning-Based Quantitative Odor Analysis. ACS Sensors 2018, 3 (8), 1592–1600. [PubMed: 30110149]

(18). Hizir MS; Robertson NM; Balcioglu M; Alp E; Rana M; Yigit MV Universal Sensor Array for Highly Selective System Identification Using Two-Dimensional Nanoparticles. Chem. Sci 2017, 8 (8), 5735–5745. [PubMed: 28989614]

(19). Hizir MS; Nandu N; Yigit MV Homologous MiRNA Analyses Using a Combinatorial Nanosensor Array with Two-Dimensional Nanoparticles. Anal. Chem 2018, 90 (10), 6300–6306. [PubMed: 29677441]

(20). Nandu N; Hizir MS; Yigit MV Systematic Investigation of Two-Dimensional DNA Nanoassemblies for Construction of a Nonspecific Sensor Array. Langmuir 2018, 34 (49), 14983–14992. [PubMed: 29739192]

(21). Nandu N; Smith CW; Uyar TB; Chen Y-S; Kachwala MJ; He M; Yigit MV Machine-Learning Single-Stranded DNA Nanoparticles for Bacterial Analysis. ACS Appl. Nano Mater 2020, 3 (12), 11709–11714. [PubMed: 34095773]

(22). Lu C; Liu Y; Ying Y; Liu J Comparison of MoS2, WS2, and Graphene Oxide for DNA Adsorption and Sensing. Langmuir 2017, 33 (2), 630–637. [PubMed: 28025885]

(23). Voiry D; Mohite A; Chhowalla M Phase Engineering of Transition Metal Dichalcogenides. Chem. Soc. Rev 2015, 44 (9), 2702–2712. [PubMed: 25891172]

(24). Voiry D; Yang J; Kupferberg J; Fullon R; Lee C; Jeong HY; Shin HS; Chhowalla M High-Quality Graphene via Microwave Reduction of Solution-Exfoliated Graphene Oxide. Science. 2016, 353 (6306), 1413–1416. [PubMed: 27708034]

(25). Ruiz-Perez D; Guan H; Madhivanan P; Mathee K; Narasimhan G So You Think You Can PLS-DA? BMC Bioinformatics 2020, 21 (1), 1–10. [PubMed: 31898485]

(26). Pedregosa F; Varoquaux G; Gramfort A; Michel V; Thirion B; Grisel O; Blondel M; Prettenhofer P; Weiss R; Dubourg V; Vanderplas J; Passos A; Cournapeau D; Brucher M; Perrot M;

Duchesnay É Scikit-Learn: Machine Learning in Python. J. Mach. Learn. Res 2011, 12 (85), 2825–2830.

(27). Martinez-Vernon AS; Covington JA; Arasaradnam RP; Esfahani S; O'Connell N; Kyrou I; Savage RS An Improved Machine Learning Pipeline for Urinary Volatiles Disease Detection: Diagnosing Diabetes. PLoS One 2018, 13 (9), e0204425. [PubMed: 30261000]

(28). Mohr F; Wever M; Hüllermeier E ML-Plan: Automated Machine Learning via Hierarchical Planning. Mach. Learn 2018, 107 (8), 1495–1515.

(29). Vabalas A; Gowen E; Poliakoff E; Casson AJ Machine Learning Algorithm Validation with a Limited Sample Size. PLoS One 2019, 14 (11), 1–20.

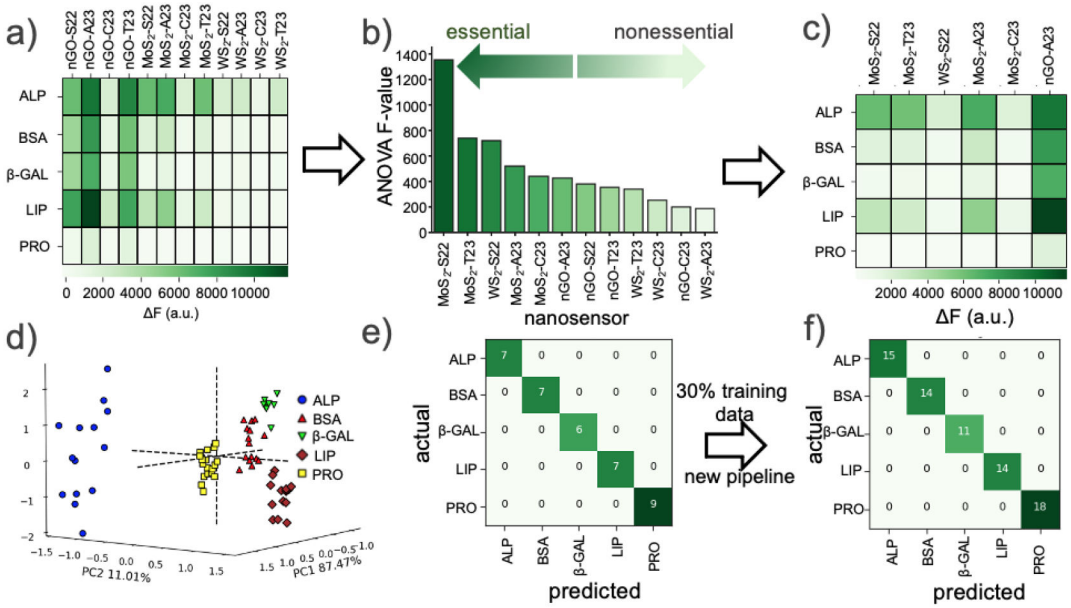


Figure 1:

(a) Observed averaged ΔF patterns by an array of nanosensors and proteins. (b) AGONS ranking of nanosensor importance, and (c) selected for a minimum of six nanosensors. (d) PCA reduction separates five proteins during training. (e) Predictive performance on the blinded dataset with 6:1:3 data split is 100% accurate. (f) Predictive performance on a blinded dataset with 3:1:6 data split is 100% accurate.

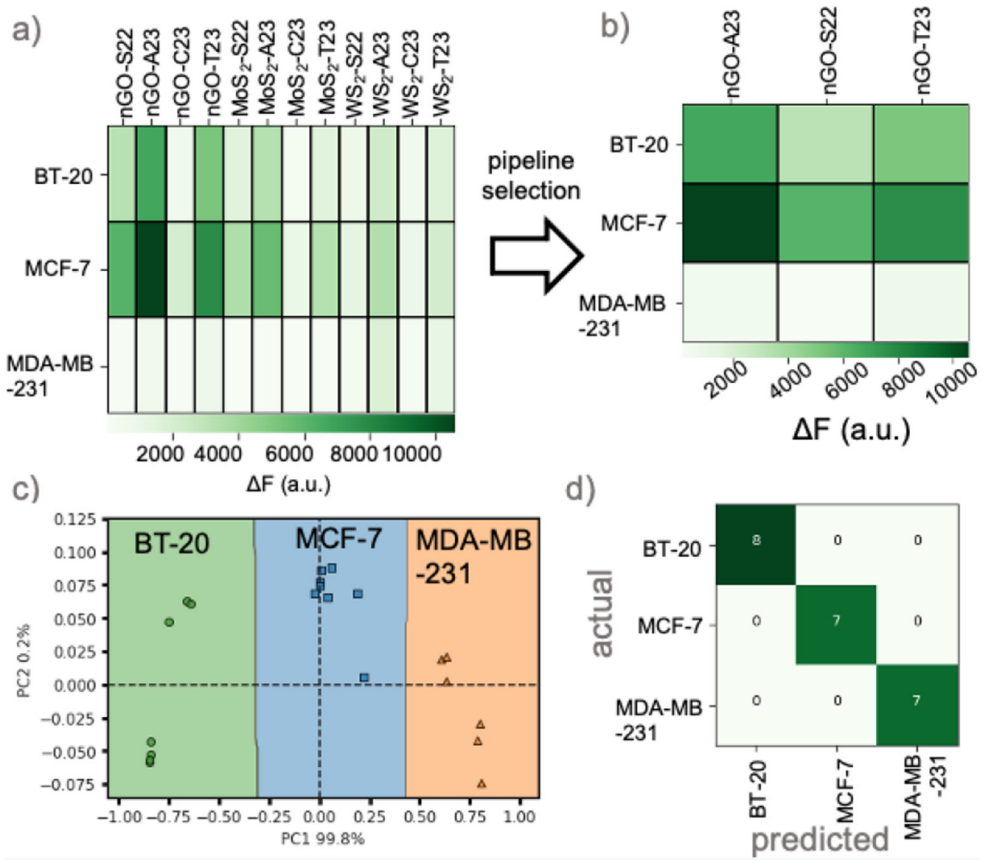


Figure 2:

(a) Observed averaged ΔF pattern by nanosensor library and breast-cancer cell lines. (b) AGONS selection for an array of only three nanosensors post validation for best model parameters.

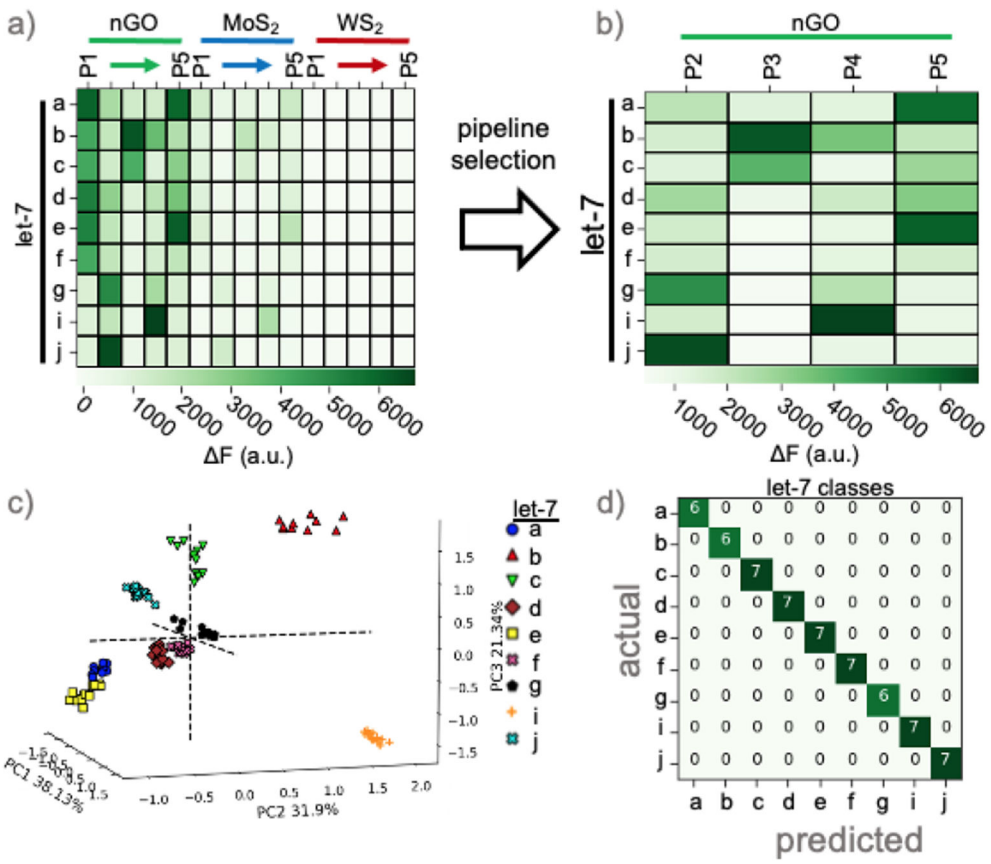
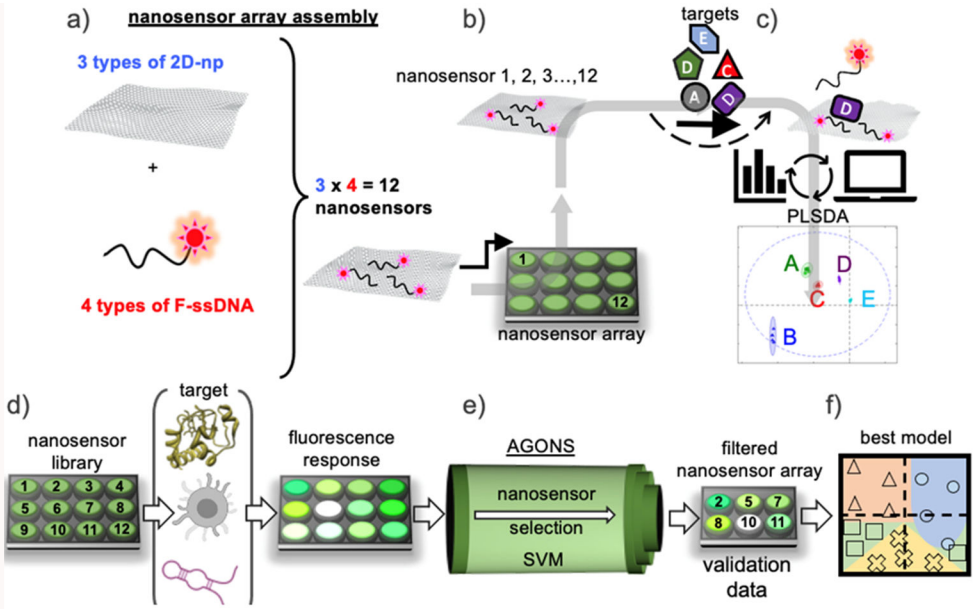


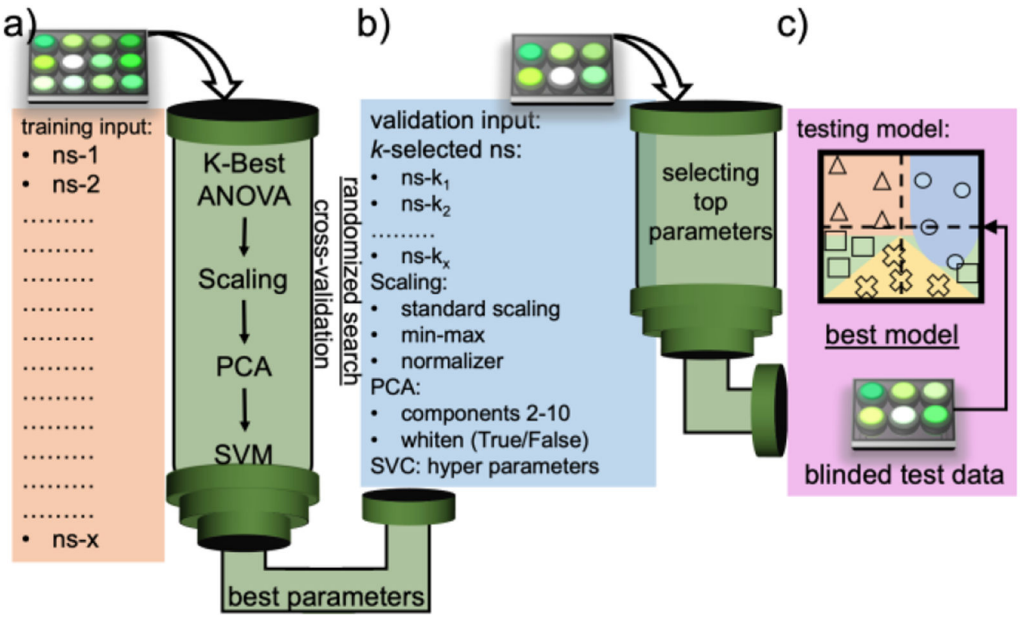
Figure 3:

(a) Observed averaged ΔF pattern by an array of 15 nanosensors and let-7 mimics. (b) AGONS selected an array of only four nanosensors post validation for best model parameters. (c) Visualization of three-dimensional PCA shows separation for nine let-7 mimics with (d) 100% predictive accuracy at data split ratio of 6:1:3.



Scheme 1: Workflows for PLSDA and AGONS Modeling of Fluorescence Responses from Nanosensor Arrays [a]

[a](a) PLSDA workflow of nanosensor array setup. (b) When nanosensor array is introduced to (c) targets, a fluorescence recovery pattern is observed and modeled through PLSDA. (d) When a library of nanosensors is introduced to proteins, breast cancer cells or miRNA mimics and fluorescence response is observed. (e) AGONS uses the fluorescence response pattern to filter the array size with validation data to find (f) the best model.



Scheme 2: Illustration of AGONS Modeling Parameters at Training, Validation and Testing Steps[b]

[b](a) Training initiates by inserting the ΔF values from the nanosensor array into the AGONS pipeline. (b) After initial parameter screening, filtered parameters are used on the validation data for model optimization. (c) The best model with the least number of necessary nanosensors are selected and used to test samples blinded from modeling.