**Machine Perception Nanosensor Array Captures a Disease Fingerprint of Ovarian Cancer**

Mijin Kim[1], Chen Chen[1,2,3], Peng Wang[4], Joseph J. Mulvey[5], Yoona Yang[6], Christopher Wun[7], Merav Antman-Passig[1], Hong-Bin Luo[4], Sun Cho[1], Kara Long-Roche[1], Lakshmi V. Ramanathan[1], Anand Jagota[6], Ming Zheng[8], YuHuang Wang[4], and Daniel A. Heller[1,2]*

[1]Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[2]Weill Cornell Medicine, Cornell University, New York, NY 10065, USA
[3]Tri-Institutional PhD Program in Chemical Biology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[4]University of Maryland, College Park, MD 20742, USA
[5]Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, NY 10461, USA
[6]Lehigh University, Bethlehem, PA 18015, USA
[7]Hunter College High School, New York, NY 10128, USA
[8]National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

*Corresponding author: hellerd@mskcc.org

**Serum biomarkers are widely used as diagnostic indicators, but many are not specific and/or sensitive enough for screening purposes. Ovarian cancer, for example, is challenging to diagnose, in part because the biomarker levels are elevated in other conditions. Here, we acquired a "disease fingerprint" from patient serum by collecting large data sets of physicochemical interactions to a sensor array composed of organic color center-modified carbon nanotubes. Array responses from 269 patients were used to train and validate machine learning models to differentiate ovarian cancer from other diseases and healthy individuals. This strategy yielded 87% sensitivity at 98% specificity versus 84% via the multimodal test using the biomarker cancer antigen 125 and transvaginal ultrasonography. Detection could not be recapitulated by known protein biomarkers, suggesting that heretofore unidentified biomarkers in the serum milieu are responsible for the sensor response.**

Ovarian cancer, the second most common gynecologic malignancy worldwide, is responsible for over 184,000 deaths each year.[1] If there is no sign that cancer has spread outside of the ovaries, five-year survival rates are over 90%.[2] However, 59% of cases are diagnosed after they have metastasized to distant sites, for which the 5-year survival drops to only 29%.[2] The earlier detection of ovarian cancer and timely measurements of disease progression and recurrence would markedly improve outcomes.

Conventionally, serum biomarker measurements, such as cancer antigen 125 (CA125) combined with transvaginal ultrasonography, have been suggested for use as a screening tool for the detection of ovarian cancer.[3,4] Recent reports have found that these methods do not result in early-stage detection and confer little survival benefit[5,6] in part due to the challenge of improving sensitivity while maintaining high specificity. Other complementary serum biomarkers such as human epidermis protein 4 (HE4), chitinase-3-like protein 1 (YKL40), and mesothelin, or panels of biomarkers have been reported to result in higher sensitivity over CA125-based screening.[7-9] However, the improvement in discriminatory power for ovarian cancer diagnosis is still under debate.[8,10] Currently, no screening strategy has been shown to reduce mortality, and screening strategies are associated with a high rate of false-positive results and a risk of harm from invasive testing.

Major factors limiting precise diagnosis using serum biomarkers include patient heterogeneity[11,12] and low specificity resulting from few established molecular markers[13-17]. Known biomarkers may not represent the complete disease state or may be present in many other diseases. Thus, accurate detection of analytes does not always confer high sensitivity and specificity for a disease.[8] Many serum biomarkers in clinical practice thus only provide incremental value for treatment options, and often do not reduce the screening cost for patients.[18]

To seek an alternative approach to overcome diagnostic challenges, we investigated a perception-based strategy. Nature has evolved perception to identify and interpret multidimensional stimuli against target heterogeneity. Perception achieves target identification by using a number of sensory inputs wherein each encodes certain features of the target and analyzing these inputs against a pre-learned target pattern library. For instance, the perception of smell uses an array of non-specific olfactory receptors, whose pattern of responses is processed by the neural network in our brain to identify an odor.[19] Olfactory receptors are

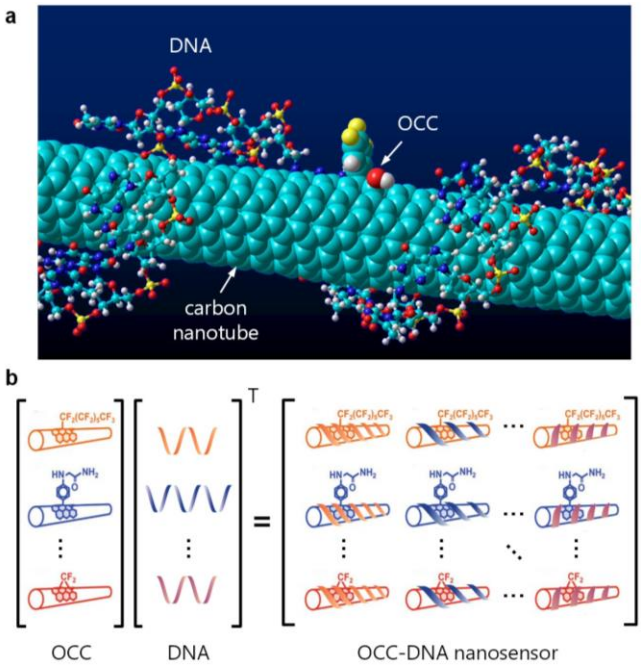


**Fig. 1 | OCC-DNA nanosensor array. a**, Molecular model of an OCC-DNA nanosensor element. Shown is a ss(GT)$_{15}$ DNA-wrapped (6,5)-SWCNT with 3,4,5-trifluoroaryl OCC. **b**, Construction of an OCC-DNA nanosensor array from OCC and ssDNA components.

relatively small in number (100–200), yet through perception, they enable recognition of many different odors, far exceeding what is possible with one-to-one recognition. For these odors, although each signal produces relatively little predictive value, the full array of responses processed as a whole nevertheless lead to accurate identification.

Perception-based approaches have been used to classify various disease conditions based on different patterns in methylation of DNA sequencing[20], volatile organic compounds using electronic noses[21], small metabolites using mass spectrometry[22], and image analysis of pathology, computerized tomography scan, and magnetic resonance imaging[23,24]. Machine learning processes recognize disease-specific patterns that are too subtle or complex to be detected by human eyes or conventional analytical methods and aid in the construction of robust diagnostic models.[24,25] Despite efforts to develop a generalizable platform of perception-based diagnostic screening using pathology or radioimaging data, challenges remain in the identification of effective disease markers to achieve high sensitivity and selectivity and practical feasibility in the clinic.

Semiconducting single-walled carbon nanotubes (SWCNTs) exhibit intrinsic near-infrared fluorescence[26] with environmental responsivity down to the single-molecule level[27]. The emission of SWCNTs ($E_{11}$) is sensitive to dielectric environment[28,29], redox perturbations[30], and electrostatic charge[31,32]. Non-covalent encapsulation with polymers, including short oligonucleotides, facilitates aqueous suspension and confers molecular selectivity to their optical responses via 1) contributing to a molecular masking effect that defines the shape and size of the exposed surface of SWCNTs[33-35] and 2) modulating their optical bandgaps[36].

Organic color centers (OCCs) are molecularly tunable quantum defects on SWCNTs which are produced by covalent functionalization of a SWCNT.[37] OCCs efficiently harvest mobile excitons through the SWCNT antenna, producing distinct fluorescence bands ($E_{11}^-$) at longer wavelengths from the $E_{11}$ band. The $E_{11}^-$ fluorescence introduces new biochemical sensitivities to SWCNTs determined by the chemical nature of the defect, making OCCs the molecular focal points for local environmental responses.[38]

Herein, we present a nanosensor array and a computational model that resulted in the perception-based detection of ovarian cancer from patient serum samples. To transduce broad types of physicochemical properties of a biofluid, we designed nanosensor arrays using OCC-functionalized, ssDNA encapsulated SWCNTs (OCC-DNAs, **Fig 1**). The emission of the OCC-DNA nanosensors exhibited diverse responses to serum samples collected from patients with high-grade serous ovarian carcinoma (HGSOC), other non-HGSOC diseases (including patients in remission, other gynecologic processes such as endometriosis and low-grade ovarian carcinoma, non-gynecologic cancers, and other conditions), and healthy individuals, but the optical responses did not provide substantial predictive value to differentiate these patients using conventional statistical analyses. We thus trained several machine learning models to classify the three categories of patients using the OCC-DNA sensor array responses. Support vector machine models resulted in striking sensitivity and specificity of HGSOC detection with an accuracy approaching 95%–significantly better than conventional serum biomarker-based identification. Potential interferents, such as drug treatments, were accounted for. The sensors were then used to assess the degree of predictive value conferred by known ovarian cancer serum biomarkers, including CA125, HE4, and YKL40. Support vector regression models showed that the sensor elements responded quantitatively to these markers, but they did not account for all of the predictive value, suggesting that unknown biomarkers play an important role in the differentiation of HGSOC by the sensors.

## Results

We synthesized an array of OCC-DNA nanosensors by introducing several $sp^3$ defects to the (6,5) SWCNT via diazonium chemistry[39] and encapsulating them with a library of ssDNA to solubilize the nanosensors in biofluids. The ssDNA sequences were chosen based on the recognition sequences of DNA that form specific wrapping patterns on the SWCNT surface[40] to result in diverse, highly-defined surface morphologies to confer disparate sensitivities to the local environment.[33,34] Ten different OCC-DNA nanosensors were successfully synthesized from the combinations of three OCCs and four DNA sequences (**Table 1**). Each OCC-DNA nanosensor featured a pair of emission peaks depending on the chemical nature of the OCC and DNA sequence. We used 575 nm excitation to selectively excite (6,5)-SWCNT (**Fig 2a** and Fig S1), resulting in emission at ~1000 nm from the (6,5) nanotube species $E_{11}$ band and a peak falling between 1110 to 1170 nm, depending on the aryl functional group. The latter is denoted as the $E_{11}^-$ band, or "OCC peak.".

To determine a minimal set of OCC-DNA combinations that provide the most diverse responses from the patient samples, we measured the fluorescence spectral responses of the OCC-DNAs to serum samples from HGSOC patients and healthy individuals. Four serum samples of the two conditions were incubated with ten different OCC-DNAs for 2 hours, and the fluorescence spectra of the OCC-DNA complexes were acquired. For each OCC-DNA nanosensor, we analyzed four different spectral features of the OCC-DNA nanosensors that were modulated in response to interactions with analytes in serum: $E_{11}$ and $E_{11}^-$ intensity (*int* and *int**) and wavelength (*wl* and *wl**). From these data, we identified the sensors that gave statistically significant differences in response to healthy versus cancer groups in parametric t-tests (quantified by p-value, **Fig 2b** and Fig S2). Our hypothesis was that OCC-DNAs which perform well independently would make good choices when used in combination. Six OCC-DNA nanosensors exhibiting $E_{11}$ or $E_{11}^-$ peak wavelengths with statistically significant differences between HGSOC and healthy groups (p-values < 0.10) were selected for the sensor array used in the subsequent parts of this study (highlighted in **Table 1**, Fig S2). The selection/reduction of features improves the training speed and model performance by eliminating redundant features in the data set.

We initially exposed the OCC-DNA sensor array to 215 patient serum samples and constructed a data set comprised of the spectral feature changes caused by the serum environment. Specifically, the size of the data matrix was $N_{sa} \times (N_f \times N_{OCC\text{-}DNA})$, where $N_{sa}$ is the number of serum samples, $N_f$ is the number of features per OCC-DNA, and $N_{OCC\text{-}DNA}$ is the number of different OCC-DNA complexes in the array. The set of serum samples was collected from 49 HGSOC, 51 other gynecologic diseases (such as endometriosis and low-grade ovarian carcinoma), 29 non-gynecologic cancer, 25 cancer patients in remission, including 7 HGSOC, and 61 healthy donors (Table S1). The fluorescence spectra were collected at three time points during incubation: 2 hours, 24 hours, and 72 hours.

**Table 1 | OCC-DNA nanosensor elements.** Left: Chemical diversity of OCCs with varying terminating moieties on the aryl functional group. Center: Special oligonucleotide sequences that form molecular masks on CNTs. Right: Synthesized OCC-DNA nanosensors. A sensor array comprised of multiple OCC-DNA nanosensors and was used for the machine learning training (highlighted in red). $NEt_2$, 3F, and F-$CO_2H$ represents 4-*N,N*-diethylamino, 3,4,5-trifluoro, and 3-fluoro-4-carboxy aryl organic color centers, respectively.

| Terminating group of aryl OCC | ssDNA sequence | OCC-DNA nanosensor |
|---|---|---|
| -4-N($C_2H_5$)$_2$ | $CTTC_3TTC$ | $NEt_2$*$CTTC_3TTC$ |
| | $(TAT)_4$ | $NEt_2$*$(TAT)_4$ |
| | $(GT)_{15}$ | $NEt_2$*$(GT)_{15}$ |
| -3,4,5-$F_3$ | $CTTC_3TTC$ | 3F*$CTTC_3TTC$ |
| | $(TAT)_4$ | 3F*$(TAT)_4$ |
| | $(AT)_{15}$ | 3F*$(AT)_{15}$ |
| | $(GT)_{15}$ | 3F*$(GT)_{15}$ |
| -3-F-4-$CO_2H$ | $CTTC_3TTC$ | F-$CO_2H$*$CTTC_3TTC$ |
| | $(AT)_{15}$ | F-$CO_2H$ *$(AT)_{15}$ |
| | $(GT)_{15}$ | F-$CO_2H$ *$(GT)_{15}$ |

To reduce the inconsistency in spectral measurements, the averaged sensor response of triplicate was used for the data analysis. We note that the variation of each measurement from the averaged triplicates was small for all the OCC-DNA peaks (Fig S3). The variations in *dwl* and *dwl** showed narrow Gaussian distributions with the standard deviations ranged from 3.72–5.37%. The maximum variation in the same sample was less than 15% (<0.3 nm). The analysis confirmed that our measurement can reliably identify the small spectral shifts. This is likely because OCC-DNAs exhibit relatively narrow bandwidths (35–80 meV), and thus, small spectral shifts are significantly easier to be resolved as compared to conventional fluorophores (>100 meV).

All four spectroscopic variables, *int*, *int**, *wl*, *wl**, measured from the OCC-DNA nanosensor array, exhibited statistically significant differentiation between HGSOC and healthy groups, but the data did not delineate a clear difference between HGSOC and other disease conditions (**Fig 2c**, Fig S4). Principal component analysis (PCA) was performed on the spectroscopic data ($N_f$ = 4) upon a 2-hour incubation from all combinations of OCC-DNA sensors ($N_{OCC\text{-}DNA}$ = 6). The first two principal components accounted for 87.5% of the total variance (principal component loadings listed in Table S2). Similar to **Fig 2c**, healthy samples showed the differentiable signatures from the disease samples, denoted by their segregation into separate regions in the PCA plot (**Fig 2d**), but HGSOC could not be separated from other disease conditions.

To differentiate HGSOC from other conditions, we next trained machine learning models using the sensor responses and clinical diagnostic results (**Fig 3**, Fig S5). The algorithms were used for binary classification of sensor responses: HGSOC vs. other diseases + healthy (the differentiation of HGSOC from all other samples). The set of features chosen for the classification task were the spectroscopic variables *dint*, *dint**, *dwl*, and *dwl** collected from the OCC-DNA sensor array. For robustness, we investigated five standard machine learning algorithms with nested levels of optimization processes: model hyperparameters, model choice, and multilevel validation. We tested supervised machine learning algorithms: logistic

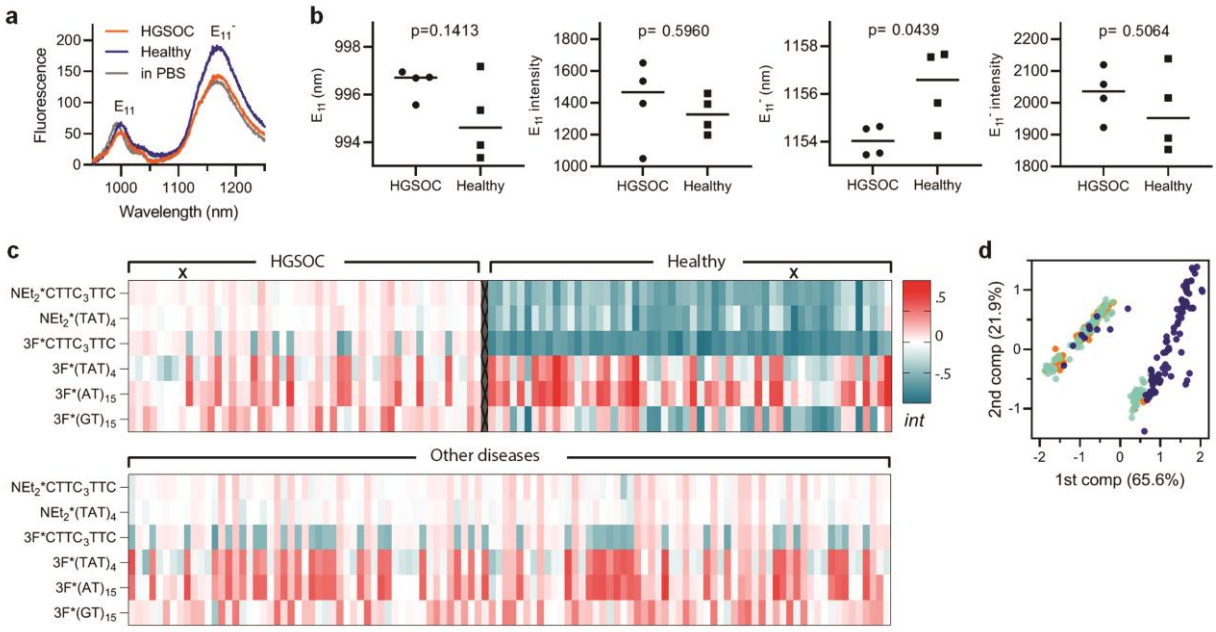


**Fig. 2 | Spectroscopic responses of OCC-DNA sensors to patient serum samples. a**, Representative fluorescence spectra of the ss$(GT)_{15}$ wrapped 3,4,5-trifluoroaryl OCC sensor, 3F*$(GT)_{15}$, in PBS (gray), 20 v/v% serum from an HGSOC patient (orange) and serum from a healthy individual (blue). **b,** Spectral responses of the 3F*$(GT)_{15}$ sensor to cancer and healthy individuals' serum samples. Four spectral parameters – intensity and wavelength of the $E_{11}$ and $E_{11}^-$ peaks (*int, int*, wl,* and *wl**) were extracted from fluorescence spectra of four serum samples for each group. Data points represent the mean value of the spectroscopic variables. Each sample was measured in triplicate. Horizontal lines denote the median. **c,** $E_{11}$ intensity change (*dint*) of each OCC-DNA sensor in response to 215 serum samples from HGSOC and other disease patients, as well as healthy individuals at 2-hour incubation. **d,** Principal component analysis (PCA) of sensor responses to HGSOC (orange), other diseases (light blue), and healthy samples (blue).

regression, decision tree, artificial neural networks, random forest, and support vector machine (SVM), while tuning models' hyperparameters with Bayesian optimization[41]. The averaged F-score in 10-fold cross-validation was used to assess the model performance (see Methods).

We first examined the machine learning algorithm that most accurately classifies HGSOC (**Fig 3a**). We compared the averaged F-scores of the machine learning algorithms using OCC-DNA combinations within the sensor array. We assessed combinations of OCC-DNA nanosensor responses, up to six at a time, out of the six originally-selected OCC-DNAs ($1 \leq N_{OCC\text{-}DNA} \leq 6$), for 63 total possible combinations for each incubation duration (See Table S3). We found that SVM resulted in the best F-scores among the five machine learning algorithms that we tested (Fig S5). Thus, we used SVM models for subsequent optimizations of the HGSOC classifier.

For our second optimization, we compared the differences in model performance using sensor responses measured under different durations of incubation with the serum samples ($N_f = 3\times63$). In all the tested machine learning algorithms, there were no statistically significant differences between incubation times (Fig S5). We found that combining data sets obtained over multiple incubation durations can improve the model performance, but the performance was only marginally better than using 2 hours of serum incubation (**Fig 3a,** Table S3). Thus, we used the 2-hour data set for subsequent model development for simplicity.

Thirdly, we examined which spectroscopic variables in the set of feature vectors optimize F-scores. We compared three combinations of spectral variables, involving the $E_{11}^-$ to $E_{11}$ intensity ratio (*Δint)*, the wavelength difference between $E_{11}^-$ and $E_{11}$ peaks (*Δwl*), *dwl*, *dwl**, *dint*, and *dint**, and combinations thereof ($N_f$ = (2, 4, or 6)×63, **Fig 3b**). The SVM models trained with the data set of 2 variables, *Δint* and *Δwl*, resulted in lower F-scores. We found no statistically significant difference between 4 and 6 variables in the F-scores of the optimized SVM models potentially because *Δint* and *Δwl* are derivative of the others. Thus, we used the 4 variables for further investigations.

We then investigated the impact of the number of different OCC-DNA sensors in the array on the F-score ($1 \leq N_{OCC\text{-}DNA} \leq 6$, **Fig 3c**). When more OCC-DNA elements were added to the sensor array, the F-scores tended to increase systematically. The trend was the same regardless of which machine learning algorithm was used (Fig S5). The best SVM model was trained by the spectral response of five OCC-DNAs: $NEt_2$*$CTTC_3TTC$, $NEt_2$*$(TAT)_4$, 3F*$(TAT)_4$, 3F*$(AT)_{15}$, and 3F*$(GT)_{15}$. The averaged cross-validation score of the SVM model was 93.9% sensitivity, 95.2% specificity, and an F-score of 0.945 differentiating HGSOC from all other disease + healthy samples. Small variances in F-score and sensitivity (<0.1) in cross-validations suggest that the optimized models are generalizable within the sample set.

Lastly, we examined if tuning the hyperparameters to maximize the $F_\beta$ score can improve sensitivity at a high specificity (**Fig 3d**). The $F_\beta$ score is the weighted harmonic mean of PPV and sensitivity and $\beta$ is chosen such that sensitivity is considered $\beta$-times as important as PPV. At decreasing $\beta$ from 3 to 0.2, sensitivity at 98% specificity systematically increased, although the improvement was statistically insignificant. The best performing prediction model was the sensor array combination of 4-$N(C_2H_5)_2$*$CT_2C_3T_2C$, 4-$N(C_2H_5)_2$*$(TAT)_4$, 3,4,5-$F_3$*$(TAT)_4$, 3,4,5-$F_3$*$(AT)_{15}$, and 3,4,5-$F_3$*$(GT)_{15}$, and yielded 87% sensitivity at 98% specificity when PPV and sensitivity were equally weighted ($\beta$ =1).

To further assess the robustness of the sensor array and algorithm, we synthesized a new batch of OCC-DNAs under the same condition and collected the sensor array response to an independent test set of 54 patient samples ($N_{sa}$ = 54). To evaluate the model performance in various medical conditions, the test set was sampled from different patients, comprised of 7 HGSOC, 5 other gynecologic diseases, 32 non-gynecologic diseases, and 10 heathy patients. With this new sample set, we observed 100% sensitivity at 98% specificity and an F-score of 0.978 using the SVM model differentiating HGSOC. These values are consistent with the cross-validation scores and gave a similar receiver operating characteristic (ROC) curve (**Fig 3e**), indicating that the model did not overfit the data.

The risk of bias in the study was evaluated based on Prediction model Risk Of Bias Assessment, PROBLAST[42] (**Appendix 1** in Supplemental Information). The risk of bias scored low in terms of predictors, outcomes, and analysis. In participants, the tool resulted in the finding of no systemic differences between training and cross-validation sets. However, the limited medical record of healthy donors and the enriched fraction of breast cancers in the non-HGSOC group of the test set may introduce systematic bias in participant selection and the validation of machine learning models, respectively. For clinical translation of the technology, these risk of bias must be taken into account.

We also endeavored to account for chemical interferents and background chronic conditions that could confer a bias in the sensor response. From a patient chart review, we identified chronic diseases and most-common medications administered to the patients (Fig S6). We found that 75% of HGSOC and 68% of other disease patients suffered from at least one chronic condition, and the relative abundance between these disease groups was similar. Regarding the medications, we statistically assessed the contribution of each interferent to the sensor results using a multivariate regression model (Table S5). The regression model determined a linear correlation between the sensor response and medication, using estimated parameters and errors. The adjusted $R^2$ of the regression model ranged from -0.045 to 0.233, indicating weak linear correlations of each sensor response to medications. We confirmed that the sensor array platform accurately classified the disease status regardless of medication and chronic conditions, evidenced by high F-scores of the HGSOC prediction models (Table S3). The analysis suggested no indications that such interferences reduced specificity in HGSOC detection by the sensor platform.

To test the utility of the SVM model relative to conventional diagnostic methods, we compared conventional biomarker-based HGSOC detection and histology results to the F-score predicted by the SVM model. We measured known biomarkers in the patient serum samples, including CA125, HE4, and YKL40, creatinine, and bilirubin by immunoassays (see Methods). We assessed the diagnostic accuracy of serum HGSOC biomarkers in these patients (**Fig 3f**). Although the differences in serum CA125, HE4, and YKL40 levels, with respect to the clinical references, were statistically significant between HGSOC, healthy, and other (non-HGSOC) diseases (Fig S7), false-positive rates were high. For example, CA125-based screening with 50 U/mL cutoff resulted in 65.3% sensitivity, 88.3% specificity, and an F-score of 0.621 in our patient sample set. The logistic regression of additional biomarkers marginally improved the HGSOC prediction (**Fig 3f**). PCA plots of HGSOC biomarkers CA125, HE4, and YKL40 showed that these markers failed to differentiate HGSOC from other diseases while the healthy individuals' samples clustered together (**Fig 3g**). Clinical trials using these biomarkers showed similar results.[8,10] These results confirmed that our

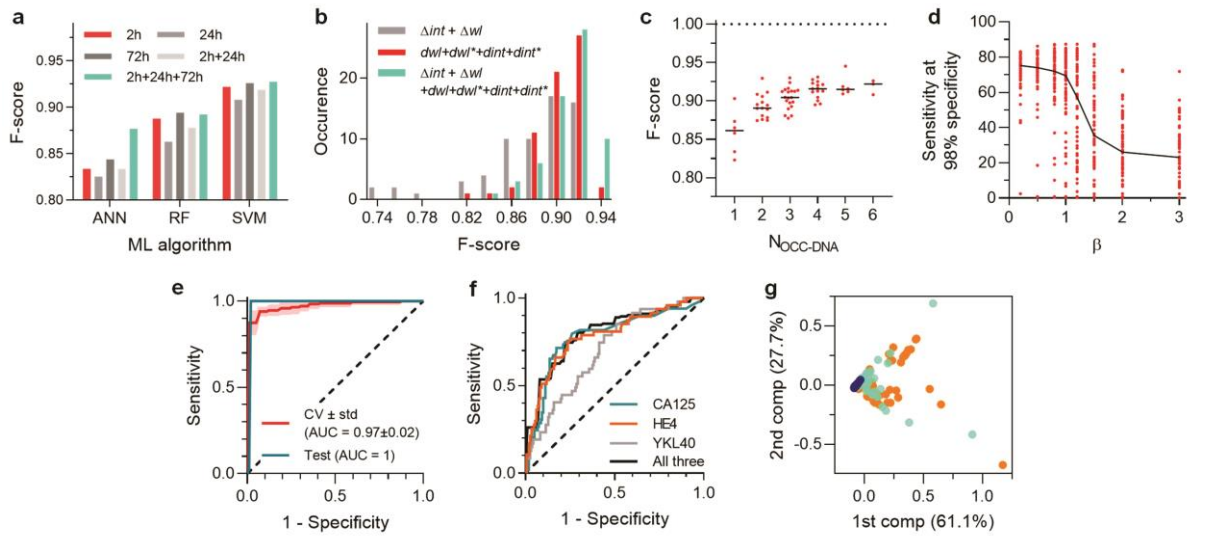


**Fig. 3 | Optimization of machine learning algorithms for HGSOC classification. a**, Comparison of F-scores of HGSOC identification with artificial neural network (ANN), random forest (RF), and SVM models, using sensor data collected with different serum incubation times. **b,** Distribution of F-scores obtained using data with different numbers of spectral variables: 2 variables (*Δwl* + *Δint*) vs. 4 variables (*dwl* + *dwl** + *dint* + *dint**) vs. 6 variables (*dwl* + *dwl** + *dint* + *dint** + *Δwl* + *Δint*). **c,** F-scores obtained with different numbers of OCC-DNA nanosensor types, via SVM. **d,** Sensitivity at 98% specificity obtained with varying β in $F_\beta$ scoring via SVM. The line connects the median of sensitivities for the optimized nanosensor arrays. **e,** Best ROC curves for binary classification of HGSOC, showing both cross-validated training set (CV) and test/validation set (Test). The shaded area is the standard deviation of 10-fold validation. **f,** ROC curves of HGSOC classification using individual serum biomarkers (CA125: blue, HE4: orange, YKL40: gray) and logistic regression of their combination (black). **g,** PCA plot of three disease states, HGSOC (orange), other diseases (light blue), and healthy patients (blue), calculated using conventional serum measurements of CA125, HE4, and YKL40 levels from 215 patient sera.

perception/sensor-based platform significantly outperformed established serum biomarker-based classification, and the accuracy was much closer to diagnosis by the physician (using pathology, imaging, etc.).

To investigate the molecular basis for the sensor-based HGSOC fingerprint, we investigated the sensor response to serum biomarkers (**Fig 4**). We measured the spectral response of the OCC-DNA nanosensors upon single analyte titration with bilirubin, creatinine, and HGSOC serum biomarkers, including CA125, HE4, YKL40, and mesothelin in 20% fetal bovine serum (**Fig 4a-d**, Fig S8). We found that several OCC-DNA spectral responses correlated with CA125, HE4, YKL40, and bilirubin concentrations while mesothelin and creatinine showed no quantitative correlations with the sensor responses. We surmise that, because of this correlation, the inclusion of biomarker-dependent spectroscopic variables in the training data set improved F-scores for HGSOC identification. We then assessed the relative contribution of each spectral parameter to the model performance by an ablation study—individually dropping each spectroscopic variable from the analysis (Fig S9). On analysis of feature importance, we identified that 3F*$(GT)_{15}$ and 3F*$(TAT)_4$ were the most important OCC-DNA nanosensors. We also found that the same feature in different sensor arrays can improve or reduce the prediction scores (Fig S9). For instance, the $E_{11}^-$ intensity (*dint**) of $NEt_2$*CTT has the highest positive feature importance (improved F-score by 0.067) in the sensor array of 4-$N(C_2H_5)_2$*$CT_2C_3T_2C$ while the same feature reduced the F-score by 0.018 in the sensor array combination of 4-$N(C_2H_5)_2$*$CT_2C_3T_2C$ and 3,4,5-$F_3$*$(GT)_{15}$. Overall, the biomarker-dependent features scored highly, indicating that such features improved the SVM model performance (**Fig 4e**). The observations confirmed that 1) OCC-DNA fluorescence transduces broad types of subtle differences in physicochemical properties of physisorbed molecules and 2) known serum biomarkers make up part of the disease fingerprint. However, the use of biomarker-dependent features exclusively did not result in optimal F-scores. The inclusion of certain features that showed no quantitative correlation with known biomarkers improved the model performance. These experiments suggest that the OCC-DNA nanosensor array results may be due, at least in part, to the transduction of heretofore unidentified biomarkers.

To further investigate the correlation between serum biomarker levels and the response of the nanosensor array, we assessed whether the sensor array responses could be used to train an SVM model to identify abnormal levels of known biomarkers in the patient samples. First, we trained an SVM classification model to detect elevated CA125 by dividing the patient sera into groups based on the threshold for suspicion of malignancy; normal (0–50 U/mL) vs. high (>50 U/mL) CA125. The CA125 training resulted in high F-scores (>0.92) for all possible sensor array combinations (**Fig 4f, g**, Table S4). We similarly assessed HE4 and YKL40 with respect to their clinical references of 150 pM for HE4 and 1650 pM for YKL40, and we developed binary classification models to differentiate abnormal levels using the SVM algorithm. Both HE4 and YKL40 classification resulted in high F-scores (0.89–0.98 and 0.81–0.93, respectively) for the detection of abnormal biomarker levels.

We additionally investigated whether support vector regression (SVR) models can quantitatively predict serum biomarker levels using the sensor array (**Fig 4h**). The best CA125 regression model, using three OCC-DNAs, $NEt_2$*$(TAT)_4$, 3F*$(TAT)_4$, and 3F*$(AT)_{15}$ resulted in an average R-squared ($r^2$) value of 0.719 (**Fig 4i**). We note that the prediction error in the normal concentration range (<50 U/mL) was larger than in the high concentration range. This can be attributed to the fact that the detection limit in the single titration experiment was close to the clinical reference of CA125. The SVR models of HE4 and YKL40 were also constructed, resulting in $r^2$ values of 0.55 and 0.56, respectively. The SVR models suggest that the known biomarkers influence the sensor responses, but the models were not sensitive enough to reliably predict the exact biomarker levels.

We assessed the contribution of each spectral parameter to the biomarker classification and regression models (Fig S10). Most of the spectral parameters had positive relative importance on average, indicating that including such features improved the positive predictive value and sensitivity of the biomarker identification. A positive correlation of the feature importance to F-score (for binary classification) and $r^2$ (for regression) was stronger for the biomarker-dependent variables that were identified in the single-analyte experiments (**Fig 4a-d**). Regarding bilirubin, however, although OCC-DNA fluorescence responses quantitatively correlated to its concentration over biologically relevant ranges (Fig S8), we failed to optimize a good SVR model for detection due to the small variance of the biomarker levels within the patient samples. The SVR model performance for serum creatinine was poor due to a lack of quantitative correlation between sensor response and creatinine concentrations in the single-titrant experiment (Fig S8).

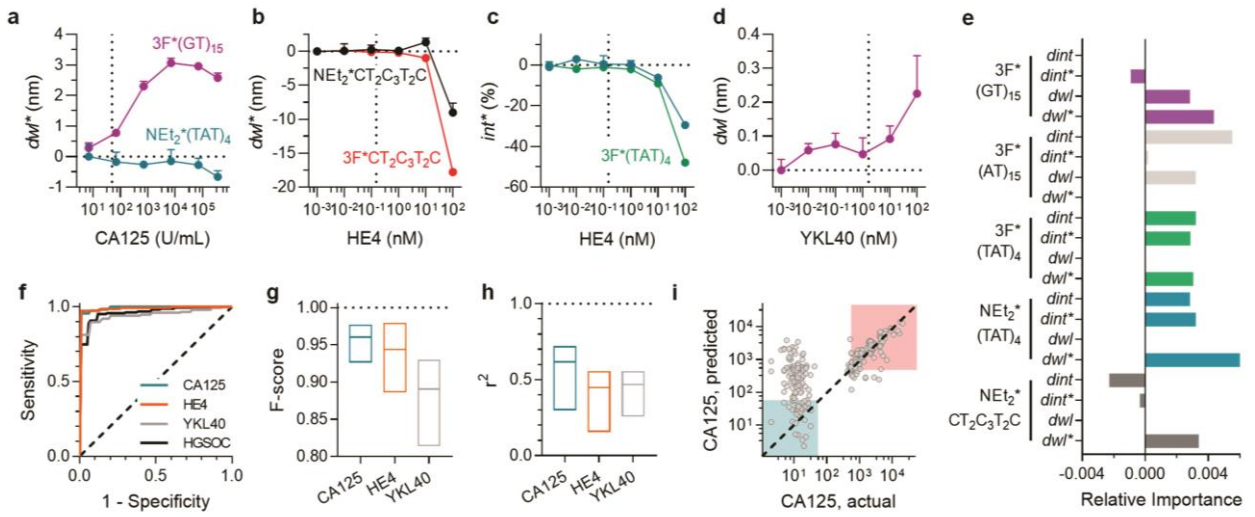


**Fig. 4 | Known serum biomarkers make up part of the disease fingerprint in the nanosensor array platform. a**–**d**, Representative spectral response of OCC-DNA in 20% FBS at increasing concentration of (**a**) CA125, (**b**,**c**) HE4, and (**d**) YKL40. **e,** Feature importance analysis of the binary SVM model. **f,** ROC curves of binary biomarker classification (normal vs. above clinical reference) using SVM of the OCC-DNA sensor responses. **g,** F-score ranges of SVM classifications of HGSOC biomarkers or disease state. Lines in each box indicates the median. **h,** R-squared ranges of biomarker SVR. **i,** Serum CA125 levels predicted by SVR against immunoassay results. The prediction models were trained by the fluorescence response of $NEt_2$*$(TAT)_4$, 3F*$(TAT)_4$, and 3F*$(AT)_{15}$. The highlighted squares classify normal (<50 U/mL, blue) and high CA125 (>250 U/mL, red) groups.

## Discussion

We constructed a nanosensor array platform, comprised of OCC-DNA elements and coupled with machine learning algorithms, to investigate the potential to identify HGSOC in patient sera. The array was comprised of multiple OCC moieties and DNA sequences, which together offer a rich design space for modulating the morphology and chemistry of the exposed nanotube surface. Our DNA sequence selection was based on the recognition sequences that form specific wrapping patterns on the nanotube surface. These sequences were originally selected to isolate individual (n,m) species/chiralities of nanotubes.[40] We reasoned that the recognition sequences of DNAs would confer the greatest diversity of interactions with the serum milieu, which is important to establish an OCC-DNA library for screening disease-specific sensor responses. We based this rationale on the findings that ssDNA encapsulates CNTs via π-π stacking interactions, and certain DNA sequences can behave like a “molecular mask” that defines the shape and size of the exposed surface.[33] Their characteristic surface structures are responsible for diverse physicochemical properties of the OCC-DNAs,[34] leading to different protein corona compositions[43,44]. Different morphologies determined by OCCs and DNA thereby contribute to the selectivity of the nanotube surfaces to the serum milieu. The fluorescence modulation of SWCNTs is caused by several mechanisms including Fermi level shifting through modulation of the immediate redox environment and exciton disruption in response to binding events, which change SWCNT intensity, and solvatochromic (wavelength) shifting due to perturbation of the local dielectric environment, including shifts due to modulation of the local electrostatic environment.[31,45] OCC fluorescence, on the other hand, is molecularly specific and extremely sensitive to the local chemical environment of the atomic defect sites.[38,46] Interactions between HGSOC serum biomarkers and OCC-DNA hybrids elicited additional, diverse spectral responses of the sensor array that enabled sufficient differentiation of signals from other sera.

The sensor platform was used to identify HGSOC with high positive and negative predictive values. Model performance of the sensor technology exceeds the results of the current best clinical screening test using longitudinal CA125 and second-line transvaginal ultrasonography[5] (87% vs. 84% clinical sensitivities at 98% specificity). However, considering the fact that specimens obtained from symptomatic individuals at diagnosis were used for the development and assessment of the technology, prediction outcomes may differ in clinical screening settings in which specimens are obtained in asymptomatic individuals before clinical

diagnosis. Further studies in a high-risk cohort, such as BRCA mutation carriers undergoing risk-reducing surgery, are warranted to evaluate the ability of the model to identify pre-invasive and early-invasive disease.

This sensor technology platform exhibits several unique potential advantages for clinical applications. First, this method could be rapidly adapted to the detection of many diseases/conditions. The array could be used to train an algorithm to recognize nearly any disease when given enough data from the sensor responses to the appropriate patient serum samples. Second, this technology could supplement or replace the use of known biomarkers when there are issues with selectivity in conventional multi-analyte tests. Due to the potential to iteratively modify the sensor array and machine learning algorithms and to additively augment training set size, the selectivity may be increasingly optimized. Third, this sensor platform can be used in a high-throughput fashion to facilitate the screening of large populations. Fourth, because the technology does not rely on antibody-based molecular recognition elements, the sensors could be more robust than existing methods and are not as reliant on cold chain logistics, enabling use in resource-limited settings and in technologies such as point-of-care and wearable/implantable devices[47]. Lastly, the sensor technology also has the potential to be developed as an inexpensive and rapid screening tool to result in a single, easy-to-interpret test result in primary care settings. The materials needed for the sensor cost approximately $5 per sample because of the lack of need for cold chain storage, the small amount of OCC-DNAs needed for screening (<5 nanograms). The cost of the sensor measurement would also diminish if measured via high-throughput instruments, and the potential for the use of very low sample volumes is substantial.

This work employed machine perception to detect disease fingerprints using an array of optical nanosensors. The study carefully investigated the attributes and molecular mechanism that resulted in the striking accuracy of the machine learning-aided nanosensor array. It is important to note that the best-performing HGSOC prediction model (**Fig 3d**) included the spectroscopic variables that were not sensitive to the known biomarkers and their relative importance was much significant than the biomarker-related variables (Fig S9). This suggests that there exist potential biomarkers or combinations thereof that are either unknown or not part of conventional screening approaches but were captured by the OCC-DNA sensor array. Information detailing which biomarkers and molecular interactions primarily result in the disease fingerprint is unknown and largely cannot be determined by current machine learning methods.[48] We believe that it may be possible, with extensive investigations, to use quantitative proteomics aided by the nanosensor array as a discovery tool.[49-51] Such investigation could potentially be used to facilitate biomarker discovery efforts[52] and uncover new information related to disease pathophysiology.

## Methods

Large scale synthesis of OCC-DNAs. Raw SWCNT material, CoMoCAT SG65 and SG65i (Sigma-Aldrich) was used for the large-scale preparation of OCC-SWCNTs. The SWCNTs were dissolved in chlorosulfonic acid (Sigma-Aldrich, 99.9%) at a concentration of ~4 mg/mL with magnetic stirring, followed by the addition of an aniline derivative at different molar ratios relative to the SWCNT carbon, and equimolar amounts of sodium nitrite (Sigma Aldrich, ≥97.0%). The aniline derivatives tested for these experiments include 4-amino-2-fluorobenzoic acid (Sigma-Aldrich, 97%), 3,4,5-trifluoroaniline (Sigma-Aldrich, 98%), and N,N-diethyl-p-phenylenediamine (Sigma-Aldrich, 97%). The SWCNT-superacid mixture was then added drop-by-drop into Nanopure water with vigorous stirring (**Safety Note**: the neutralization process is aggressive; a significant amount of heat and acidic smog can be generated. Personal protective equipment, including goggles/facial mask, lab coats, and acid-resistant gloves, are necessary. The neutralization must be performed in a fume hood). The resulting OCC-SWCNTs instantly precipitate out from the solution. The precipitates were then filtered on an anodic aluminum oxide filtration membrane with a pore size of 0.02 μm (Whatman® Anodisc inorganic filter membrane), thoroughly rinsed with Nanopure water, and then dried in a vacuum oven.

The OCC-SWCNTs were stabilized by 3.5 mg/mL ssDNA in phosphate buffered saline (PBS). The OCC-SWCNT were individually dispersed by ultrasonication at 6W for 60 min using a probe-tip sonicator (Sonics & Materials, Inc) at 4°C for 1 hour. The DNA to SWCNT mass ratio is 5 to 1. Then the OCC-DNA solutions were centrifuged at 100,000 g and 4 °C for 30 min. The 80% supernatant was dialyzed against PBS for 36 hours to remove free DNA (Spectra-Por, Float-A-Lyzer, MWCO = 1MDa). The absorption spectra of the dialyzed solutions were collected with a UV-Vis-NIR spectrophotometer (Jasco, Tokyo, Japan). After

subtracting background, the optical density at (6,5) $E_{11}$ (~1000 nm) was used to estimate the relative OCC-DNA concentration[53] (Fig S11).

OCC-DNA and serum/recombinant protein handling. The OCC-DNA concentration was adjusted to 0.325 mg/L in PBS. We introduced 20 µL of a patient serum sample to 80 µL of OCC-DNAs in a 96-well plate (Corning) to make the OCC-DNA concentration of 0.26 mg/L in each well. OCC-DNAs in 100 µL PBS (0.26 mg/L) was also prepared to compare the relative changes in sensor response in serum for feature vector construction (See Data preprocessing in Methods). The OCC-DNA was incubated at room temperature for 2 hours and in a cold room (4 $^{o}$C) after the spectral acquisition at 2-hour time point. Data were taken at three time points during incubation: 2 hours, 24 hours, and 72 hours.
To test sensor sensitivity to serum biomarkers, OCC-DNA complexes were added to a 96-well plate at a concentration of 0.26 mg/L in a 100-µl total volume of 20% FBS (Gibco). In triplicate, the following were added into wells at biologically relevant concentrations: 0–352000 U/mL recombinant human CA125/MUC16 (R&D Systems), 0–100 nM recombinant human HE4 (RayBiotech), 0–100 nM recombinant human YKL40 (R&D Systems), 0–50 nM recombinant human mesothelin (BioLegend), 0–1000 µM creatinine (Fisher Scientific, ≥98%, anhydrous) or 0–200 µM bilirubin (Fisher Scientific, ≥97%). Experiments were performed with the same time points as above. All experiments were performed in triplicate.

High-throughput near-infrared spectroscopy. Fluorescence emission spectra of OCC-DNAs were acquired using a home-built near-infrared fluorescence spectroscopy apparatus consisting of a tunable white light laser source, inverted microscope, and InGaAs NIR detector. The SuperK EXTREME supercontinuum white-light laser source (NKT Photonics) was used with a VARIA variable bandpass filter accessory, capable of tuning the output 500−825 nm, set to a bandwidth of 20 nm centered at 575 nm. The light path was shaped and fed into the back of an inverted IX-71 microscope (Olympus), where it passed through a 20× NIR objective (Olympus) and illuminated the samples in a 96-well plate. Emission from the OCC-DNAs was collected through the 20× objective and passed through a dichroic mirror (875 nm cutoff, Semrock). The light was f/# matched to the spectrometer using several lenses and injected into a Shamrock 303i spectrograph (Andor, Oxford Instruments) with a slit width of 100 µm, which dispersed the emission using a 86 g/mm grating with 1.35 µm blaze wavelength. The spectral range was 723−1694 nm with a resolution of 1.89 nm. The light was collected by an iDus 1.7 µm InGaAs (Andor, Oxford Instruments) with an exposure time of 10 seconds. An HL-3-CAL-EXT halogen calibration light source (Ocean Optics) was used to correct for wavelength-dependent features in the emission intensity arising from the spectrometer, detector, and other optics. A Hg/Ne pencil-style calibration lamp (Newport) was used to calibrate the spectrometer wavelength. Background subtraction was conducted using a well in a 96-well plate filled with PBS or 20% FBS, depending on the experiment. Following acquisition, the data were processed with custom code written in Matlab that applied the aforementioned spectral corrections and background subtraction and was used to fit the data with Lorentzian functions.

Serum sample set. 269 waste samples were collected from female patients diagnosed with ovarian and other cancers under a Memorial Sloan Kettering Cancer Center Institutional Review Board approved protocol. From this sample set, 56 specimens were collected from patients diagnosed with high-grade serous ovarian cancer, 71 specimens from healthy donors, 56 with other gynecologic diseases, 61 with non-gynecologic diseases, and 25 in remission. There was no statistically significant difference in age distribution for each group. Diagnoses were identified from a chart review of each patient; all diagnoses included histology and were confirmed by gynecologic oncology attending physician. Patient demographics, diagnosis, and biomarker levels are available in Table S1.

Serum assays. Serum concentrations of CA125 and HE4 were determined on the Abbott Architect i2000 analyzer (Abbott Diagnostics, Abbott Park, IL, USA) using a chemiluminescent microparticle immunoassay. YKL40 was analyzed using a singleplex immunoassay on the Protein Simple Ella system. The Abbott C8000 analyzer was used to determine the concentrations of creatinine by quantitating the formation of creatinine picrate in alkaline conditions, and bilirubin was analyzed by the formation of azobilirubin using the diazo reagent under specified conditions.

Data preprocessing. Quantities representing the sensor response to patient serum were acquired by the Lorentzian fitting of OCC-DNA fluorescence spectra: $E_{11}$ intensity, $E_{11}^-$ intensity, $E_{11}$ wavelength, and $E_{11}^-$ wavelength. The average value of triplicates was used as feature data for machine learning processes. Feature values were defined as a difference in sensor response acquired from patient serum and PBS. Specifically, the $E_{11}$ peak position feature, *dwl*, was defined as the wavelength difference between the $E_{11}$ peak in the patient sample, *wl*, and PBS, $wl_0$, $dwl = wl - wl_0$. The $E_{11}$ peak intensity feature, *dint,* was normalized as $dint = int/int_0$, where *int* and $int_0$are the $E_{11}$ peak intensity in serum and PBS, respectively. Similarly, we defined $E_{11}^-$ peak related features, *dwl** and *dint**, indicating the relative $E_{11}^-$ peak position and intensity. We additionally considered the relative change in $E_{11}^-$ to $E_{11}$ intensity, $\Delta int = (int^*/int)(int_0^*/int_0)^{-1} - 1$, and the wavelength difference between two peaks, $\Delta wl = dwl^* - dwl$ to check if the addition of these features would create a larger variance in HGSOC prediction.

We normalized each feature vector to be in the range of [-1, 1] to balance the feature contribution to the model. The imbalance in the size of each group was corrected by upscaling minority species (SMOTE: Synthetic Minority Oversampling Technique)[54] so that the prediction models were not biased by groups with a larger sample size. For the biomarker prediction models, we divided the data into normal versus high biomarker level groups based on the clinical references (CA125: 50 U/mL, HE4: 150 pM, YKL40: 1650 pM) and corrected the group size using SMOTE.

Model training and performance assessment. Using algorithms implemented in *Scikit-Learn*[55], we created models based on Decision Tree, Logistic Regression, Artificial Neural Networks, Random Forest, and Support Vector Machine (SVM) for binary classification. Hyperparameters for each model were optimized using Bayesian Optimization, implemented in the HyperOpt library.[41] The loss function to minimize in the hyperparameter optimization was set to (*1 – F-score*). F-score (or $F_1$-score) is a measure of accuracy in binary classification and calculated from the harmonic mean of the positive predictive value (PPV) and sensitivity: $2/(\text{sensitivity}^{-1} + \text{PPV}^{-1})$. To rule out the possible overfitting in machine learning process, model performance was evaluated using ten-fold cross-validation. In the cross-validation process, stratified shuffle split validation was used to randomly partition the data set into ten subsamples. In each partition, nine of the ten subsamples were used to train the model, while a single subsample was used to test the trained model. The average F-score of the ten-fold cross-validation was used to assess model performance. The trained models were then tested with an independent set of patient sera (N=54), sampled from different patients (test set), as external validation. Support vector regression (SVR) was used to construct the regression models of HGSOC serum biomarkers with 10-fold cross-validation. The loss function in the hyperparameter optimization was $(1 - r^2)$. For SVM and SVR, a radial basis function kernel was used and the hyperparameter optimization was performed for the regularization parameter (cost) and the kernel coefficient (gamma) with the maximum iteration of 1000. The hyperparameter space of each machine learning algorithm for model optimization is shown in Table S6.

### Data availability

All data is available in the manuscript or the supplementary materials.

### Code availability

Python and Matlab codes for the machine learning and data analysis in this article is available upon request, by contacting the corresponding author (hellerd@mskcc.org).

## References


1. Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394-424 (2018).
2. Siegel, R.L., Miller, K.D. & Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **70**, 7-30 (2020).

3. Menon, U. et al. Risk Algorithm Using Serial Biomarker Measurements Doubles the Number of Screen-Detected Cancers Compared With a Single-Threshold Rule in the United Kingdom Collaborative Trial of Ovarian Cancer Screening. *J. Clin. Oncol.* **33**, 2062-2071 (2015).
4. Blyuss, O. et al. Comparison of Longitudinal CA125 Algorithms as a First-Line Screen for Ovarian Cancer in the General Population. *Clin. Cancer Res.* **24**, 4726 (2018).
5. Jacobs, I.J. et al. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. *Lancet* **387**, 945-956 (2016).
6. Menon, U. et al. Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. *Lancet* **397**, 2182-2193 (2021).
7. Dupont, J. et al. Early Detection and Prognosis of Ovarian Cancer Using Serum YKL-40. *J. Clin. Oncol.* **22**, 3330-3339 (2004).
8. Cramer, D.W. et al. Ovarian Cancer Biomarker Performance in Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial Specimens. *Cancer Prev. Res.* **4**, 365 (2011).
9. Han, C. et al. A novel multiple biomarker panel for the early detection of high-grade serous ovarian carcinoma. *Gynecol. Oncol.* **149**, 585-591 (2018).
10. Moore, L.E. et al. Proteomic biomarkers in combination with CA 125 for detection of epithelial ovarian cancer using prediagnostic serum samples from the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. *Cancer* **118**, 91-100 (2012).
11. Johnson, C.C. et al. The epidemiology of CA-125 in women without evidence of ovarian cancer in the Prostate, Lung, Colorectal and Ovarian Cancer (PLCO) Screening Trial. *Gynecol. Oncol.* **110**, 383-389 (2008).
12. Bast, R.C. et al. A Radioimmunoassay Using a Monoclonal Antibody to Monitor the Course of Epithelial Ovarian Cancer. *N. Engl. J. Med.* **309**, 883-887 (1983).
13. Miralles, C. et al. Cancer antigen 125 associated with multiple benign and malignant pathologies. *Ann. Surg. Oncol.* **10**, 150-154 (2003).
14. Buamah, P. Benign conditions associated with raised serum CA-125 concentration. *J. Surg. Oncol.* **75**, 264-265 (2000).
15. Linda, H. et al. Human epididymis protein 4 (HE4) in benign and malignant diseases. *Clin. Chem. Lab. Med.* **50**, 2181-2188 (2012).
16. Moss, E.L., Hollingworth, J. & Reynolds, T.M. The role of CA125 in clinical practice. *J. Clin. Pathol.* **58**, 308 (2005).
17. Park, Y., Lee, J.-H., Hong, D.J., Lee, E.Y. & Kim, H.-S. Diagnostic performances of HE4 and CA125 for the detection of ovarian cancer from patients with various gynecologic and non-gynecologic diseases. *Clin. Biochem.* **44**, 884-888 (2011).
18. Diamandis, E.P. The failure of protein cancer biomarkers to reach the clinic: why, and what can be done to address the problem? *BMC Med.* **10**, 87 (2012).
19. Su, C.-Y., Menuz, K. & Carlson, J.R. Olfactory perception: Receptors, cells, and circuits. *Cell* **139**, 45-59 (2009).
20. Liu, M.C. et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* **31**, 745-759 (2020).
21. Hao, Y. et al. in SENSORS, 2003 IEEE, Vol. 2 1333-1337 Vol.1332 (2003).
22. Zhang, J. et al. Nondestructive tissue analysis for ex vivo and in vivo cancer diagnosis using a handheld mass spectrometry system. *Sci. Transl. Med.* **9**, eaan3968 (2017).
23. Yu, K.-H. et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* **7**, 12474 (2016).
24. Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703-715 (2019).
25. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* **1**, 18 (2018).
26. Bachilo, S.M. et al. Structure-assigned optical spectra of single-walled carbon nanotubes. *Science* **298**, 2361 (2002).
27. Cognet, L. et al. Stepwise quenching of exciton fluorescence in carbon nanotubes by single-molecule reactions. *Science* **316**, 1465 (2007).

28. Heller, D.A. et al. Optical detection of DNA conformational polymorphism on single-walled carbon nanotubes. *Science* **311**, 508 (2006).
29. Jena, P.V. et al. A carbon nanotube optical reporter maps endolysosomal lipid flux. *ACS Nano* **11**, 10689-10703 (2017).
30. Heller, D.A. et al. Multimodal optical sensing and analyte specificity using single-walled carbon nanotubes. *Nat. Nanotech.* **4**, 114-120 (2009).
31. Roxbury, D., Jena, P.V., Shamay, Y., Horoszko, C.P. & Heller, D.A. Cell membrane proteins modulate the carbon nanotube optical bandgap via surface charge accumulation. *ACS Nano* **10**, 499-506 (2016).
32. Williams, R.M. et al. Noninvasive ovarian cancer biomarker detection via an optical nanosensor implant. *Sci. Adv.* **4**, eaaq1090 (2018).
33. Roxbury, D., Mittal, J. & Jagota, A. Molecular-basis of single-walled carbon nanotube recognition by single-stranded DNA. *Nano Lett.* **12**, 1464-1469 (2012).
34. Roxbury, D., Jagota, A. & Mittal, J. Structural characteristics of oligomeric DNA strands adsorbed onto single-walled carbon nanotubes. *J. Phys. Chem. B* **117**, 132-140 (2013).
35. Roxbury, D., Tu, X., Zheng, M. & Jagota, A. Recognition ability of DNA for carbon nanotubes correlates with their binding affinity. *Langmuir* **27**, 8282-8293 (2011).
36. Horoszko, C.P., Jena, P.V., Roxbury, D., Rotkin, S.V. & Heller, D.A. Optical voltammetry of polymer-encapsulated single-walled carbon nanotubes. *J. Phys. Chem. C* **123**, 24200-24208 (2019).
37. Brozena, A.H., Kim, M., Powell, L.R. & Wang, Y. Controlling the optical properties of carbon nanotubes with organic colour-centre quantum defects. *Nat. Rev. Chem.* **3**, 375-392 (2019).
38. Kwon, H. et al. Optical probing of local pH and temperature in complex fluids with covalently functionalized, semiconducting carbon nanotubes. *J. Phys. Chem. C* **119**, 3733-3739 (2015).
39. Luo, H.-B. et al. One-pot, large-scale synthesis of organic color center-tailored semiconducting carbon nanotubes. *ACS Nano* **13**, 8417-8424 (2019).
40. Ao, G., Streit, J.K., Fagan, J.A. & Zheng, M. Differentiating left- and right-handed carbon nanotubes by DNA. *J. Am. Chem. Soc.* **138**, 16677-16685 (2016).
41. Shahriari, B., Swersky, K., Wang, Z., Adams, R.P. & De Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE* **104**, 148-175 (2016).
42. Wolff, R.F. et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann. Intern. Med.* **170**, 51-58 (2019).
43. Pinals, R.L., Yang, D., Lui, A., Cao, W. & Landry, M.P. Corona Exchange Dynamics on Carbon Nanotubes by Multiplexed Fluorescence Monitoring. *J. Am. Chem. Soc.* **142**, 1254-1264 (2020).
44. Tenzer, S. et al. Rapid formation of plasma protein corona critically affects nanoparticle pathophysiology. *Nat. Nanotech.* **8**, 772-781 (2013).
45. Heller, D.A. et al. Peptide secondary structure modulates single-walled carbon nanotube fluorescence as a chaperone sensor for nitroaromatics. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 8544 (2011).
46. Wu, X., Kim, M., Qu, H. & Wang, Y. Single-defect spectroscopy in the shortwave infrared. *Nat. Commun.* **10**, 2672 (2019).
47. Lee, M.A. et al. Can Fish and Cell Phones Teach Us about Our Health? *ACS Sens.* **4**, 2566-2570 (2019).
48. Zednik, C. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philos. Technol.* **34**, 265-288 (2021).
49. Docter, D. et al. Quantitative profiling of the protein coronas that form around nanoparticles. *Nat. Protoc.* **9**, 2030-2044 (2014).
50. Pinals, R.L. et al. Quantitative Protein Corona Composition and Dynamics on Carbon Nanotubes in Biological Environments. *Angew. Chem. Int. Ed.* **59**, 23668-23677 (2020).
51. Lai, Z.W., Yan, Y., Caruso, F. & Nice, E.C. Emerging Techniques in Proteomics for Probing Nano–Bio Interactions. *ACS Nano* **6**, 10438-10448 (2012).
52. Hadjidemetriou, M. et al. Nano-scavengers for blood biomarker discovery in ovarian carcinoma. *Nano Today* **34**, 100901 (2020).
53. Zheng, M. & Diner, B.A. Solution redox chemistry of carbon nanotubes. *J. Am. Chem. Soc.* **126**, 15490-15494 (2004).

54. Chawla, N.V., Bowyer, K.W., Hall, L.O. & Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **16**, 321-357 (2002).
55. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825-2830 (2011).

**Acknowledgements**

We thank B. Kwon, S. Chatterjee, A. Chatterjee, M. Fleisher, B.D. Davison, S. David, and N. Osiroff for helpful discussions.
**Funding:** This work was supported in part by the NCI (R01CA215719), the Cancer Center Support Grant (P30-CA008748), the National Science Foundation CAREER Award (1752506), the Honorable Tina Brozman Foundation for Ovarian Cancer Research, the Tina Brozman Ovarian Cancer Research Consortium 2.0, the Kelly Auletta Fund for Ovarian Cancer Research, the Commonwealth Foundation for Cancer Research, the American Cancer Society Research Scholar Grant (GC230452), the Pershing Square Sohn Cancer Research Alliance, the Expect Miracles Foundation – Financial Services Against Cancer, the Anna Fuller Fund, MSK's Cycle for Survival Equinox Innovation Award in Rare Cancers, the Experimental Therapeutics Center, Mr. William H. Goodwin and Mrs. Alice Goodwin and the Commonwealth Foundation for Cancer Research. M.K. was supported by the Marie-Josée Kravis Women in Science Endeavor Postdoctoral Fellowship. Y.H.W. gratefully acknowledges support from the National Science Foundation (CHE-1904488). H.B.L. acknowledges the support provided by the China Scholarships Council (CSC No. 201708320366) during his visit to the University of Maryland. P.W. gratefully acknowledges the Millard and Lee Alexander Fellowship from the University of Maryland. M.Z.'s work is NIST internal funding. Y.Y. was supported by a Dean's Fellowship at Lehigh University. A.J. would like to acknowledge his contribution as part of the NHI initiative at Lehigh University.

**Author contributions**

M.K. and D.A.H. designed experiments and analysed the data. M.K., D.A.H. Y.H.W., M.Z., and A.J. conceived and supervised the research. M.K., P.W., and H.L. synthesized the sensor materials. M.K., C.C., and M.A.P. performed the screening experiments. M.K., Y.Y., and C.W. performed machine learning. S.C. and L.V.R. obtained and processed the patient samples. J.M. reviewed the patient charts. J.M., L.V.R. and K.L.R. provided clinical direction to the study. M.K. and D.A.H. wrote the manuscript. Y.H.W., M.Z., A.J., and J.M. edited the manuscript.

**Competing interests**
D.A.H. is a cofounder and officer with equity interest of Goldilocks Therapeutics Inc., LipidSense Inc., and Nirova Biosense Inc., as well as a member of the scientific advisory board of Concarlo Holdings LLC, Nanorobotics Inc., and Mediphage Bioceuticals, Inc.