Simpler predictive models provide higher accuracy for ovarian cancer detection

Derrick E. Wood[1,*], Joseph Roy[1,2,3], Bari J. Ballew[1]

[1]Blackjack Biotechnologies, Baltimore, MD, USA
[2]American Society for Engineering Education, Washington, DC, USA
[3]Department of English, University of Texas at San Antonio, San Antonio, TX, USA
[*]Correspondence to: dw@blackjackbiotech.com

Abstract

Ovarian cancer remains a danger to women's health, and accurate screening tests would likely increase survival. Two established protein biomarkers, CA125 and HE4, have been shown to work well in isolation, but achieve even higher accuracy when combined using logistic regression (LR). We show here that this LR-based combination of protein concentrations achieves high accuracy when distinguishing healthy samples from cancer samples (AUC = 0.99) and benign masses from cancer (AUC = 0.86). This approach exhibits superior performance on an external validation cohort compared to a more complex method, which was published with the dataset we use here and tested on the same data. While our method only uses proteins, the more complex method also uses features derived from cell-free DNA (cfDNA). We show that many of that method's cfDNA features are affected by confounding technical variation, which impacts the previously reported results. Our results are in line with the principle that simpler machine learning models will tend to exhibit better generalizability on new data.

Introduction

The American Cancer Society estimates that in 2024, the United States will see 19,680 new cases of ovarian cancer, and 12,740 women will die due to ovarian cancer [1]. Ovarian cancer is often diagnosed at advanced stages, when the prognosis is generally poorer than when found at an earlier stage [2]. For these reasons, accurate assays to detect ovarian cancer earlier are important to increase survival rates. These assays could be used in both a screening setting, where asymptomatic individuals are tested for the presence of cancer, and in a diagnostic context, where an individual with a known ovarian mass is tested to determine the likelihood that the mass is malignant.

For diagnosis of ovarian cancer, two common biomarkers used are the CA125 and HE4 proteins [2]. Elevated concentrations of each protein have been found to be associated with ovarian

cancer [3,4], although HE4 also increases with age [5]. While each protein can act as a biomarker in its own right, researchers have also found ways to combine them to create a more accurate cancer detection algorithm. In 2009, Moore, *et al*. presented a logistic regression (LR) approach, the Risk of Ovarian Malignancy Algorithm (ROMA), which combined the two values to achieve greater diagnostic accuracy than either protein alone [6]. ROMA uses the menopausal status of a patient to decide which LR model to use; in contrast, the Copenhagen Index [7] uses the patient's age along with the CA125 and HE4 serum levels in an LR model to arrive at its predicted probability of cancer.

Other cancer detection approaches have sought to address both screening and diagnostic scenarios. DELFI [8,9] is an algorithm that analyzes fragmentation patterns seen through sequencing of cell-free DNA (cfDNA) along with copy number information to detect cancer. Recently, Medina, *et al*. introduced their method DELFI-Pro [10], a computational pipeline using LR that combines CA125 and HE4 measurements with many of DELFI's cfDNA-derived features. For the detection of ovarian cancer, the authors show DELFI-Pro outperforms both the CA125 and the HE4 protein concentration levels. The authors also reported that the two proteins had the coefficients with the highest magnitude in their screening model, with the two proteins combined representing 68.8% of the total feature importance; features related to copy number changes represented 24.1% of the feature importance.

As part of the DELFI-Pro article's experimental design, the authors noted that they took measures "to reduce the possibility that differences between patients with or without cancer were not due to batch variability." Batch effects are technical variation that correlates with the batches in which samples are processed, rather than some underlying biological or scientific phenomenon [11,12]. These confounding effects can impact downstream analyses, sometimes leading to unrealistically positive results, even with cross validation [13,14]. Batch effects are a well-established concern in laboratory analyses [15], and are known to influence certain aspects of projects that process nucleic acids, such as GC-content bias [16], RNA transcript expression [17], and copy number [18,19].

Related to concerns of including batch-affected features, an important consideration in using machine learning models such as LR is to ensure that the size of the feature set is minimal while still providing adequate performance [20]. As recently stated by Divasón, *et al*. [21]: "A model that has good accuracy and low complexity, *i.e.* a model with fewer input features, will be much more robust to perturbations, faster and easier to understand, and much cheaper to maintain and update." Our own desire to follow the principle of parsimony motivated us to examine the DELFI-Pro data and determine how well LR models using only the measurements of CA125 and HE4 would perform. Such models, with similar architecture to that of the long-established ROMA, might give firmer insight into the added predictive power of DELFI-Pro's cfDNA-derived features. We present the results of this examination here.

Results

Collection of protein concentration data

Briefly, the blood samples from which Medina, *et al*. obtained their data were from two cohorts: a "discovery" cohort of samples from European studies, and a "validation" cohort of samples from United States studies. Protein concentration levels for CA125 and HE4 were quantified using immunoassay measurements from these blood samples, in most cases from plasma, although a minority of measurements were made from serum samples.

In both discovery and validation cohorts, samples were collected from three classes of donors: those with no known ovarian lesions (labeled "healthy" hereafter), those with benign adnexal masses, and those with ovarian cancer. Both cohorts have two overlapping datasets: a "screening dataset" consisting of the healthy samples and the cancer samples; and a "diagnostic dataset" consisting of the samples from benign masses and the cancer samples. The two datasets allow building of classification models in both a screening and a diagnostic context, and evaluation of the predictive power of an approach in both contexts. To facilitate comparison between the models we built and the DELFI-Pro models, we maintained this same division of cohorts and use of classification contexts.

Building a protein-only logistic regression model

We chose to use a logistic regression model architecture for our machine learning approach. Other models are of course possible, but bearing in mind that our input data was only two-dimensional, that ROMA had achieved good performance with such an LR model, and that DELFI-Pro also utilized LR, we felt LR was the appropriate choice here. Our implementation of LR used a regularization penalty to reduce the amount and likelihood of overfitting, much like DELFI-Pro's implementation. To set parameters for our LR model, and determine appropriate data preprocessing steps, we focused our initial analysis on the discovery cohort.

A common and important early step in machine learning with laboratory analytes is to use the zlog transformation [22], which compensates for the fact that the range of values from pathological samples can be an order of magnitude higher than that seen in healthy samples. The zlog transformation both takes the logarithm of the input value so that the range of data is reduced, but also standardizes the result so that healthy data tends to be close to a value of zero. We applied this transformation first to the CA125 and HE4 values, and worked with the zlog values in place of the original protein concentrations moving forward. In this zlog feature space, we can see that CA125 appears to distinguish between healthy and cancer samples well, but that HE4 may be needed to appropriately distinguish between benign masses and cancer samples (Figure 1). The two zlog values are not strongly correlated, with a Pearson's R of 0.57 in the discovery cohort.

Similarly to Medina, *et al*., we evaluated each model's performance against the appropriate discovery data in a 5-fold stratified cross validation which we repeated 10 times, and we recorded the mean score of a sample across all 10 repeats of cross validation. To evaluate the performance of the model, we use the receiver operating characteristic (ROC) area under the curve (AUC) [23]. After cross validation, our screening protein-only model had an AUC of 0.96, equal to the 0.96 reported by DELFI-Pro (Figure 2). In the diagnostic context, our protein-only model had an AUC of 0.86, slightly less than the cross validation AUC of 0.88 reported by DELFI-Pro (Figure 3).

In addition to this cross validation work, we also fit the same models to the full discovery datasets, so that the models could be examined and run on the validation cohorts. Examining the slopes of the models' decision boundaries (Figure 4), we see that the screening model is influenced slightly more by CA125's value than HE4's, with a decision boundary slope of -1.23. In contrast, the diagnostic model's boundary line is far more horizontal with a slope of -0.12, indicating that HE4 influences that model more. These results are congruent with our expectations from reviewing Figure 1.

Validation and performance of the fitted models

While cross validation results are useful, testing on a separate and unrelated dataset is considered to be a more reliable test of a model's generalizability [24]. As did Medina, *et al*., we ran the models fit to the discovery cohorts on the corresponding validation cohorts. While our protein-only models performed very similarly to DELFI-Pro in cross validation, our models substantially outperformed DELFI-Pro on the validation cohorts. Our screening model achieved an AUC of 0.99, above the corresponding DELFI-Pro AUC of 0.93 (Figure 2); and our diagnostic model achieved an AUC of 0.86, above the DELFI-Pro diagnostic validation AUC of 0.81 (Figure 3). We also compared our screening model against results from an unpublished protein-only model created by the DELFI-Pro authors (found in the online code repository associated with the DELFI-Pro publication [25]), where our validation AUC of 0.99 was slightly higher than the unpublished DELFI model's AUC of 0.98 (Supplementary Figure S1).

Batch effects confounded copy number features

Because the DELFI-Pro models exhibited lower accuracy on validation data when compared to the discovery cohorts, we believed that some sort of confounding factor may be affecting the cfDNA features. Reviewing the heatmap published by Medina, *et al.*, we saw that several copy number-related features appeared to be: (a) close to zero in all benign and healthy samples; (b) negative in the vast majority of cancer samples; and (c) uncorrelated with the DELFI-Pro score in cancer samples. (These features are highlighted for clarity in Supplementary Figure S2.) These facts appeared incongruous to us, as a feature that is negative only in cancer samples would intuitively be well-correlated with the probability of a sample belonging to a cancer patient. We then investigated the code and data used to generate the heatmap, and found that the final 85 samples in the dataset had sample IDs that shared the prefix "PGDX", as opposed to the other 394 samples with IDs with a "CGPL" prefix. Seeing this, we generated our own

heatmap, with samples grouped by cancer status and sample prefix. Our heatmap, focused on copy number and protein features, shows that several copy number features have values that are lower in PGDX samples compared to the CGPL samples (Figure 5). These results are consistent with the conclusion we reached upon our reading of the DELFI-Pro heatmap code: because the data of the heatmap was not properly joined with the row labels, the reordering of the row labels did not cause the heatmap's corresponding data rows to be reordered from the initial ordering within the input file. The connection we saw between negative copy number features and cancer samples in the original publication, then, appears to be merely an artifact of batch effects and an unfortunate error in data visualization.

Seeing the visually apparent differences in PGDX and CGPL samples within the copy number features, we sought to quantify the predictive value of the features absent this confounding factor, as well as the impact of the differences. Because the CGPL samples were 4.6 times more numerous than the PGDX samples, and the CGPL samples included all three of our conditions (healthy, benign, and cancer), we focused our analysis first on the CGPL samples. Within the CGPL samples of the discovery cohort, we calculated the AUC for cancer prediction of each copy number feature as well as our two protein features (Figure 6). In a screening context, the AUCs of the two protein features were both above 0.9, indicating strong association with cancer for each individually. None of the 39 copy number features had an AUC above 0.75, indicating that within the CGPL samples, copy number features were individually not strongly associated with cancer. In a diagnostic context, HE4 had an AUC of 0.90, and CA125 had an AUC of 0.79, results which are congruent with the LR coefficients we reported earlier. None of the copy number features had an AUC above 0.75 in the diagnostic context. We then examined only the cancer samples of the discovery cohort, calculating the AUC of predicting the sample prefix (Figure 6). The cancer subset of the discovery cohort contained 52 CGPL samples, and 42 PGDX samples. In this experiment, neither protein had an AUC of more than 0.56. However, 44% (17 of 39) of the copy number features had AUCs above 0.75, with three (corresponding to changes to chromosome arms 11p, 18p, and 19q) having AUCs above 0.9.

All samples labelled with a PGDX prefix in the repository metadata were listed with CGPL prefixes in the original paper's supplementary data, but we are able to match samples across the two namespaces by comparing protein concentration values. Reviewing the metadata associated with the PGDX samples showed that all PGDX samples were processed as part of library batches 1-16, and all CGPL samples were processed as part of library batches 17-38. We also see that the PGDX samples contain no healthy or validation samples (Figure 5, Figure 7). We are unable to determine from the original publication what caused this cohort to be processed with a different sample prefix.

Discussion

With respect to the predictive value of the proteins we examined, the results from Medina, *et al*. are largely consistent with our results here. The concentrations of CA125 and HE4 are valuable biomarkers by themselves, but combining them using a machine learning model provides even greater predictive power. We go further than the DELFI-Pro authors here, however, in examining

and reporting the performance of the two proteins without any cfDNA-associated features. When evaluating a complex model such as DELFI-Pro, this two-protein model would seem to be an appropriate baseline to compare to, and perhaps more appropriate than the comparisons to the two protein concentrations in isolation that we find in the original publication.

When compared to the DELFI-Pro models' performance on the validation cohorts, we see superior performance from our simpler models. We also see concordant performance with a protein-only model available in the DELFI-Pro code repository, giving us further confidence in the results of our parsimonious approach. Given these validation results and the very similar cross validation performance between our models and DELFI-Pro on the discovery cohorts, it appears that the simpler models exhibit improved generalizability compared to the DELFI-Pro models. Our protein-only models would also tend to be easier, faster, and cheaper to perform than the DELFI-Pro approach. This is because a laboratory test using our models would only require the quantification of protein concentrations, rather than also needing to prepare, sequence, and bioinformatically analyze DNA libraries. In light of these results, the additional predictive utility over CA125 and HE4 that would be provided by DELFI-Pro's cfDNA-derived features remains unclear.

In the specific dataset we analyzed here, we see that a substantial amount of the data within the discovery cohort was confounded by a batch effect, and that the DELFI-Pro models performed worse on validation data than on the discovery cohort. Such lower validation performance is likely attributable to the models incorrectly associating technical variation with clinical labels due to the dataset's construction. Because no healthy samples in the dataset were part of the affected batches, the screening model had no way to know that the technical artifacts present only in cancer samples were merely artifacts. Although the samples in the later batches were generally a mixture of the three conditions, the first 16 batches (the "PGDX"-prefixed samples) were nearly all homogenous, with only batch 10 contributing both cancer and benign samples to the dataset (Figure 7). Even if other samples outside of the dataset were processed as part of the batches, the composition of the dataset itself still exhibited a distinct lack of condition diversity within the early batches that, combined with the striking effects on copy number features, can lead to inaccurate models. In this instance, much of the importance of copy number features reported by the DELFI-Pro authors is likely attributable to batch effects rather than biological factors (Figure 6).

Although we do not report evidence of strong batch effects within any of the validation data, we note that the mixing of validation and discovery data within batches (Figure 7) constitutes a form of data leakage [26]. Such data leakage could result in overly optimistic validation results, because machine learning models could effectively learn batch-associated information that would not occur in future batches. To prevent this, we recommend that validation samples are both collected and processed after any samples used to train models. This creates a time separation between the two cohorts that can provide a much stronger assurance that, when the model is deployed, validation results will generalize to clinical performance.

The presence of confounding batch effects on the DELFI-Pro copy number features points to an important concern in clinical diagnostics. Experimental design to mitigate the effects of batch variation is essential, but this must also be paired with robust quality control and data analyses to either correct for batch variation or disqualify samples when extraordinary variation is encountered. The absence of such checks can lead to the creation of screening or diagnostic models that do not generalize to new samples. Such a lack of generalizability in turn can lead to a costly failure to pass a well-designed validation experiment, or worse, elevated rates of inaccurate clinical results if the validation experiment is not well-designed. A method that only examines protein concentrations would require fewer and less expensive quality control checks than a method that also includes DNA sequencing. These considerations do not make cfDNA-based diagnostics unsuitable for clinical tests, of course; they simply require that the additional benefits of cfDNA over other possible approaches be made clear for the specific application. A straightforward way to demonstrate such benefits for something like DELFI-Pro would be to compare DELFI-Pro against a protein-only approach. While the DELFI-Pro publication did compare against each of the two proteins individually, that article omits a comparison against a ROMA-like two-protein model that would have been uncomplicated to implement and report given the data available to the authors. Such a comparison would be illustrative of the additional value of DELFI's cfDNA features in detecting ovarian cancer.

With regard to ROMA, our results suggest that the ROMA approach can be successfully adapted to screening. Even without the precedent of ROMA, the combination of CA125 and HE4 makes sense to improve detection accuracy of ovarian cancer, especially in light of the clear orthogonality of the two proteins' concentrations as seen in Figure 1. Adding more orthogonal biomarkers to an LR model is quite reasonable, as long as the addition of the new biomarkers is shown to provide increased predictive value. As new biomarkers for ovarian cancer are investigated and shown to be orthogonal to both CA125 and HE4, it would certainly be reasonable to examine them in a combined model (be it LR or another machine learning architecture) with CA125 and HE4. It will remain important to compare such a potential improvement against a suitable baseline model, however, and the evidence we presented here strongly supports using a baseline model that combines these two proteins rather than using each protein in isolation. The comparison against such a baseline model will help to ensure that each new feature actually provides a meaningful advance, and that the complexity added by new features is merited by improvement in accuracy.

Methods

Data collection

We collected all data (except for library batch information) used here from the "delfipro2024" GitHub repository [25], which the DELFI-Pro authors made public as supporting material along with the DELFI-Pro publication. We also obtained in that repository scores from the protein-only screening model generated by the DELFI-Pro authors, along with the script ("code/proteins_screening_model.r") to generate that model. Library batch data was obtained

from a supplemental table of the DELFI-Pro publication, and protein values in the repository data were matched against values in another supplemental table to ensure the PGDX samples from the repository data were correctly associated with the library batches.

Data preprocessing

As defined by Al-Mekhlafi, *et al*. [22], the zlog transformation requires the use of a reference range derived from a healthy population, rather than calculating the mean and standard deviation (SD) from our own data. This use of a reference range is done to mitigate against outliers that could influence the mean and SD, and is based upon the 2.5th and 97.5th percentile values of a healthy population.

To perform the zlog transformation, we could not easily find firm population-wide distributions for CA125 or HE4, and so we adopted values of 35 U/mL as the upper reference limit for CA125, and 140 pmol/L as the upper reference limit for HE4, based on the highest limits given by Dochez, *et al*. [2]. For the lower reference limits, we used the recommendation given by Haeckel, *et al.* [27] to set the lower reference limit to 15% of the upper limit (*i.e.*, 5.25 for CA125, and 21 for HE4), because we could not find firm guidance on a lower reference limit for either protein. These reference limits, then, served as our estimates for the 2.5th and 97.5th percentile values for the two proteins.

Logistic regression model training

When building our regression models, we used default values for all parameters except for one. The one exception was the class_weight='balanced' argument, which ensured that even though there were approximately twice as many noncancer samples as cancer samples in the discovery cohort (for either of the screening or diagnostic contexts), the class imbalance would be accounted for in the model's fitting. This use of default values means our regularization penalty was an L2 (ridge regression) penalty term that was weighed equally with the error term of the objective function [28].

Generation of a new feature heatmap

For ease of interpretation, we restricted our analyses to only the chromosomal copy number features and the two protein zlog values. To highlight the impact of the prefix-associated batch effect, we performed an analysis of variance (ANOVA) within the cancer samples, grouped by prefix. Features were sorted by ANOVA *p* values, with lower *p* values on the left. To make the aberrant feature values more apparent, we standardized each feature prior to plotting, subtracting the mean of each feature and dividing by each feature's standard deviation. So that readers can see the ability to detect the batch effect on the raw data with a regular clustering, we provide a heatmap where rows and columns are clustered, and values are not standardized, as Supplementary Figure S3.

Calculation of feature AUC values

When calculating the AUC for individual features' ROC curves (shown in Figure 6), we computed the AUC as *a*. The reported AUC was the maximum of *a* and 1-*a*, to account for the fact that some features would need to be multiplied by -1 to positively correlate with the target variable. This has the impact of making 0.5 be the minimum reportable AUC for these purposes. We used 0.75 as the threshold for declaring a feature "important" here, following the advice of Fan, *et al*. [23] that features below such a threshold would not tend to be clinically useful.

Software analyses

We used the scikit-learn package (v0.23.2) to build our logistic regression models and calculate classification statistics, matplotlib (v3.5.1) and seaborn (v0.11.2) to create plots, and pandas (v1.3.5) and numpy (v1.21.5) for operating on tabular data and doing general arithmetic. All code was written in Python (v3.10.12) using a Jupyter (v6.4.8) notebook.

Data availability

Our Jupyter notebook containing all code we wrote for this project, along with copies of the data we used, are archived with and freely available from Zenodo [29].

References

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin. 2024;74: 12–49.
2. Dochez V, Caillon H, Vaucel E, Dimet J, Winer N, Ducarme G. Biomarkers and algorithms for diagnosis of ovarian cancer: CA125, HE4, RMI and ROMA, a review. J Ovarian Res. 2019;12: 28.
3. Bast RC Jr, Klug TL, St John E, Jenison E, Niloff JM, Lazarus H, et al. A radioimmunoassay using a monoclonal antibody to monitor the course of epithelial ovarian cancer. N Engl J Med. 1983;309: 883–887.
4. Hellström I, Raycraft J, Hayden-Ledbetter M, Ledbetter JA, Schummer M, McIntosh M, et al. The HE4 (WFDC2) Protein Is a Biomarker for Ovarian Carcinoma. Cancer Res. 2003;63: 3695–3700.
5. Moore RG, Miller MC, Eklund EE, Lu KH, Bast RC Jr, Lambert-Messerlian G. Serum levels of the ovarian cancer biomarker HE4 are decreased in pregnancy and increase with age. Am J Obstet Gynecol. 2012;206: 349.e1–7.
6. Moore RG, McMeekin DS, Brown AK, DiSilvestro P, Miller MC, Allard WJ, et al. A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass. Gynecol Oncol. 2009;112: 40–46.
7. Karlsen MA, Høgdall EVS, Christensen IJ, Borgfeldt C, Kalapotharakos G, Zdrazilova-Dubska L, et al. A novel diagnostic index combining HE4, CA125 and age

may improve triage of women with suspected ovarian cancer - An international multicenter study in women with an ovarian mass. Gynecol Oncol. 2015;138: 640–646.
8. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019;570: 385–389.
9. Mathios D, Johansen JS, Cristiano S, Medina JE, Phallen J, Larsen KR, et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. Nat Commun. 2021;12: 5060.
10. Medina JE, Annapragada AV, Lof P, Short S, Bartolomucci AL, Mathios D, et al. Early detection of ovarian cancer using cell-free DNA fragmentomes and protein biomarkers. Cancer Discov. 2025;15: 105–118.
11. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11: 733–739.
12. Anderson-Trocmé L, Farouni R, Bourgey M, Kamatani Y, Higasa K, Seo J-S, et al. Legacy data confound genomics studies. Mol Biol Evol. 2020;37: 2–10.
13. Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. Trends Biotechnol. 2017;35: 498–507.
14. Soneson C, Gerster S, Delorenzi M. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. PLoS One. 2014;9: e100335.
15. Yu Y, Mai Y, Zheng Y, Shi L. Assessing and mitigating batch effects in large-scale omics studies. Genome Biol. 2024;25: 254.
16. Teng M, Irizarry RA. Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-seq data. Genome Res. 2017;27: 1930–1938.
17. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform. 2020;2: lqaa078.
18. Scharpf RB, Ruczinski I, Carvalho B, Doan B, Chakravarti A, Irizarry RA. A multilevel model to address batch effects in copy number estimation using SNP arrays. Biostatistics. 2011;12: 33–50.
19. Cristiano S, McKean D, Carey J, Bracci P, Brennan P, Chou M, et al. Bayesian copy number detection and association in large-scale studies. BMC Cancer. 2020;20: 856.
20. Stoltzfus JC. Logistic regression: a brief primer. Acad Emerg Med. 2011;18: 1099–1104.
21. Divasón J, Ceniceros JF, Sanz-Garcia A, Pernia-Espinoza A, Martinez-de-Pison FJ. PSO-PARSIMONY: A method for finding parsimonious and accurate machine learning models with particle swarm optimization. Application for predicting force–displacement curves in T-stub steel connections. Neurocomputing. 2023;548: 126414.
22. Al-Mekhlafi A, Klawitter S, Klawonn F. Standardization with zlog values improves exploratory data analysis and machine learning for laboratory data. Journal of Laboratory Medicine. 2024;48: 215–222.
23. Fan J, Upadhye S, Worster A. Understanding receiver operating characteristic (ROC) curves. CJEM. 2006;8: 19–20.
24. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021;14: 49–58.

25. Medina JE, Annapragada AV, Lof P, Short S, Bartolomucci AL, Mathios D, et al. delfipro2024 GitHub repository. 2024. Available: https://github.com/cancer-genomics/delfipro2024
26. Kaufman S, Rosset S, Perlich C. Leakage in data mining: formulation, detection, and avoidance. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM; 2011. doi:10.1145/2020408.2020496
27. Haeckel R, Wosniok W, Gurr E, Peil B. Permissible limits for uncertainty of measurement in laboratory medicine. Clin Chem Lab Med. 2015;53: 1161–1171.
28. scikit-learn 0.23.2 documentation, section 1.1, Linear Models. 2020. Available: https://scikit-learn.org/0.23/modules/linear_model.html
29. Wood DE. Software and data to support protein-only detection of ovarian cancer. 2025. Available: https://dx.doi.org/10.5281/zenodo.14838464

Figures

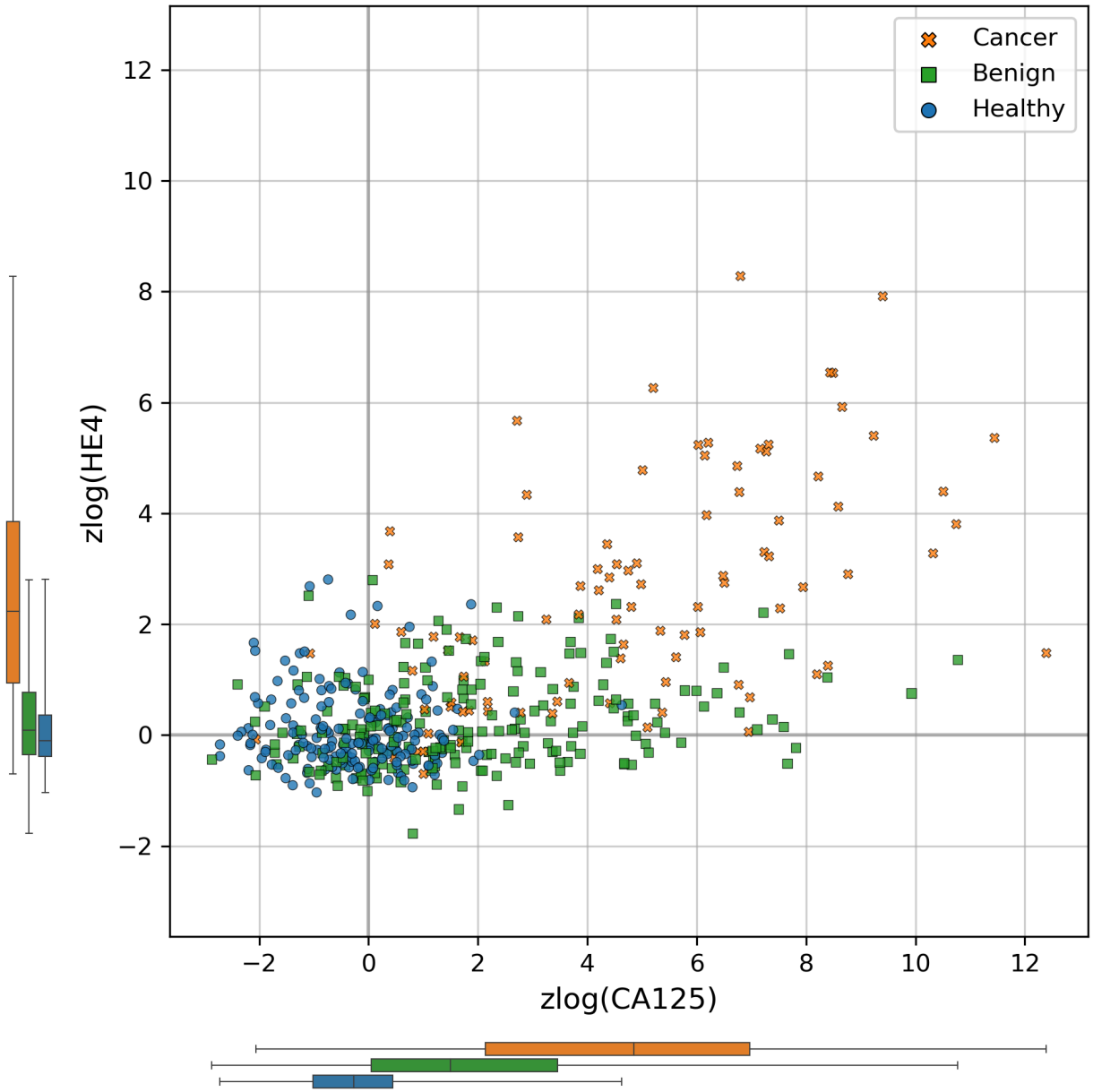


Figure 1. Scatter plot of zlog-transformed protein concentration values in the discovery cohort. Marginal distributions of the two protein levels within the three conditions of interest are shown to the left and under the corresponding axes, with boxplots indicating the interquartile range and whiskers representing the range from the minimum to maximum values.

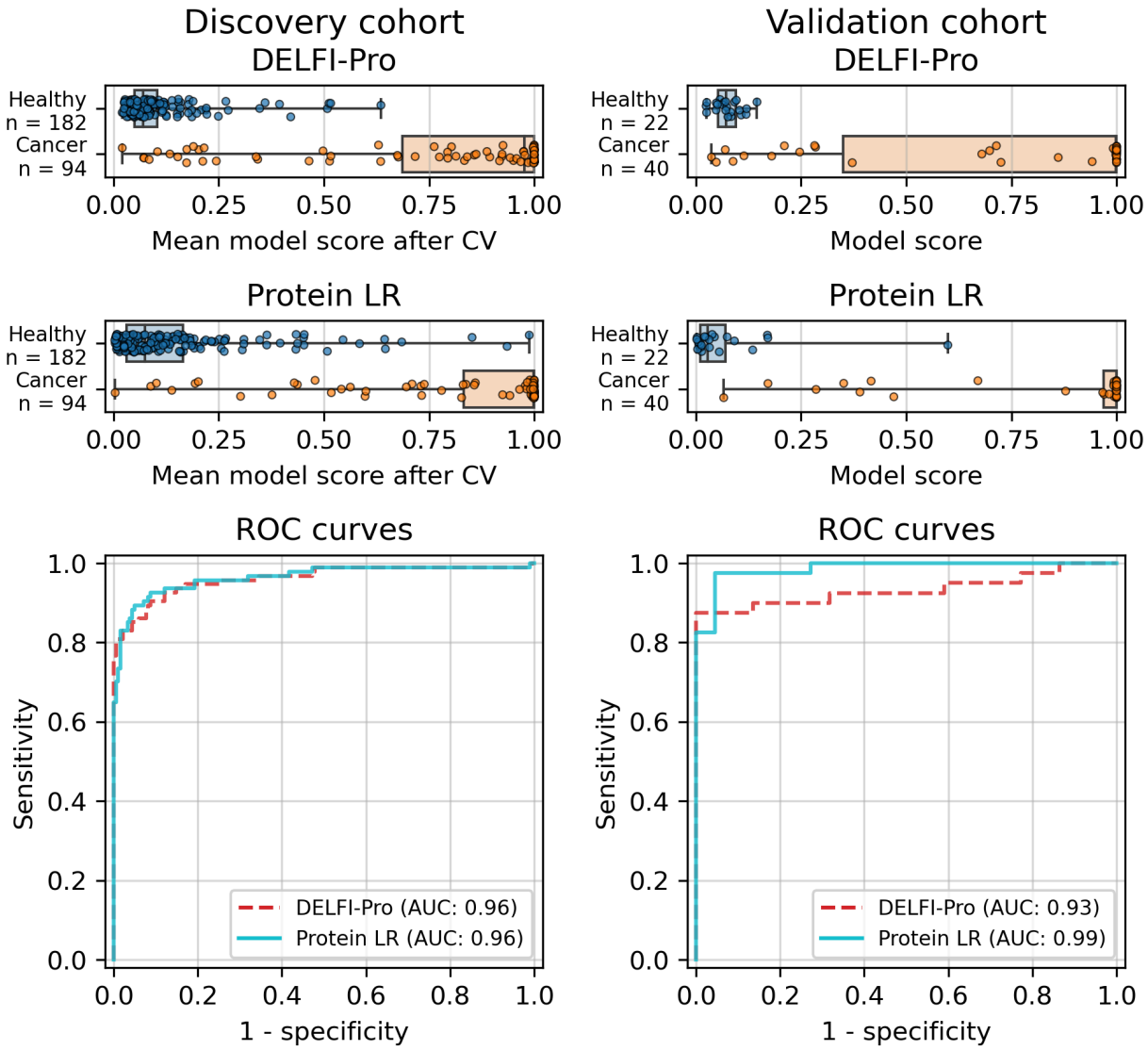


Figure 2. Logistic regression model performance in a screening context. Cross validation ("CV") results for the discovery cohort, and validation results on the validation cohort are shown for both the DELFI-Pro screening model and a two-protein logistic regression ("Protein LR") screening model. Both models seek to distinguish between samples from healthy donors and samples from patients with ovarian cancer. Receiver operating characteristic (ROC) curves demonstrate the performance of the two models at various score thresholds, with the area under the curve (AUC) growing proportionally to model accuracy.

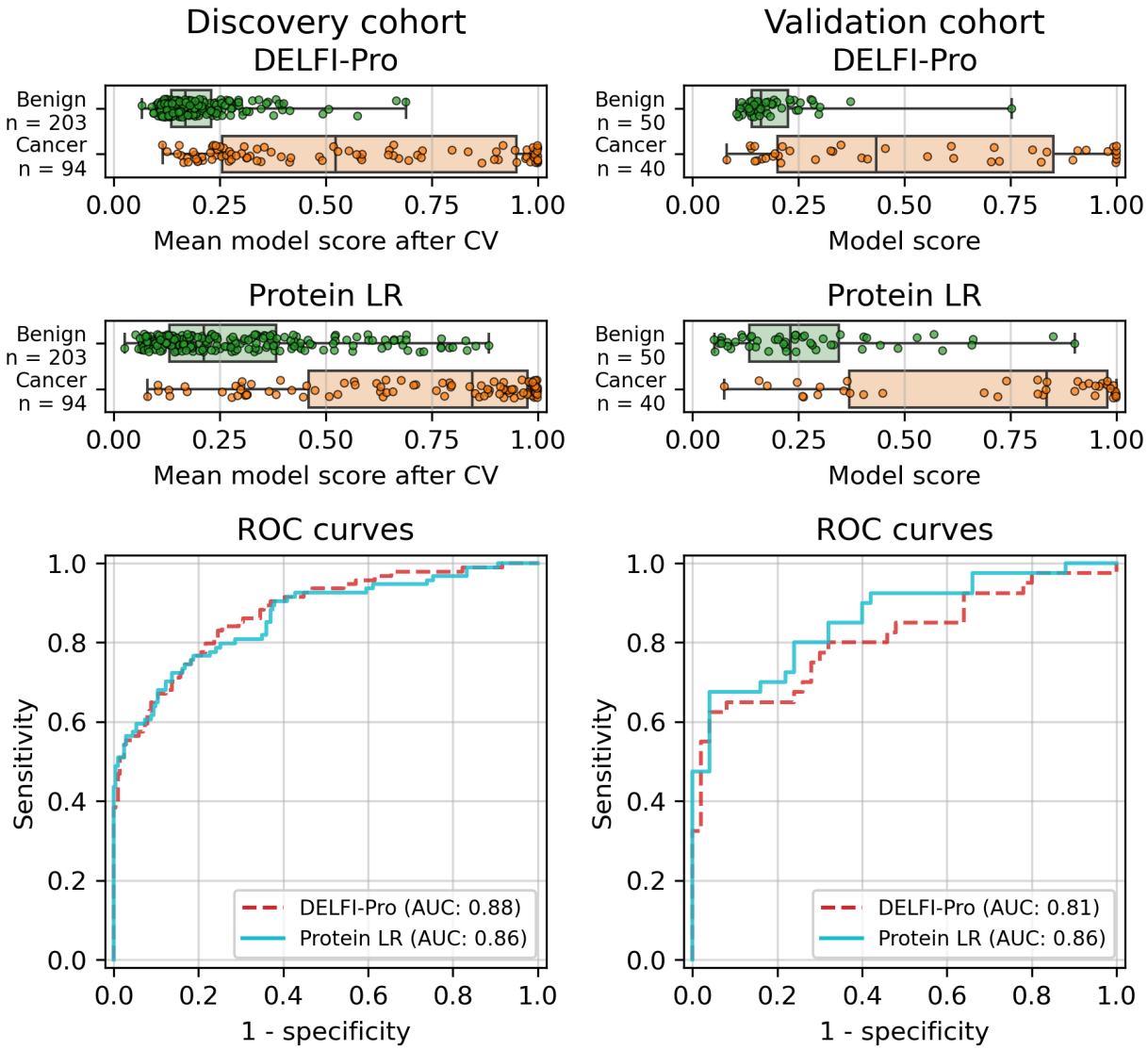


Figure 3. Logistic regression model performance in a diagnostic context. Cross validation ("CV") results for the discovery cohort, and validation results on the validation cohort are shown for both the DELFI-Pro diagnostic model and a two-protein logistic regression ("Protein LR") diagnostic model. Both models seek to distinguish between samples from donors with benign adnexal masses and samples from patients with ovarian cancer. Receiver operating characteristic (ROC) curves demonstrate the performance of the two models at various score thresholds, with the area under the curve (AUC) growing proportionally to model accuracy.

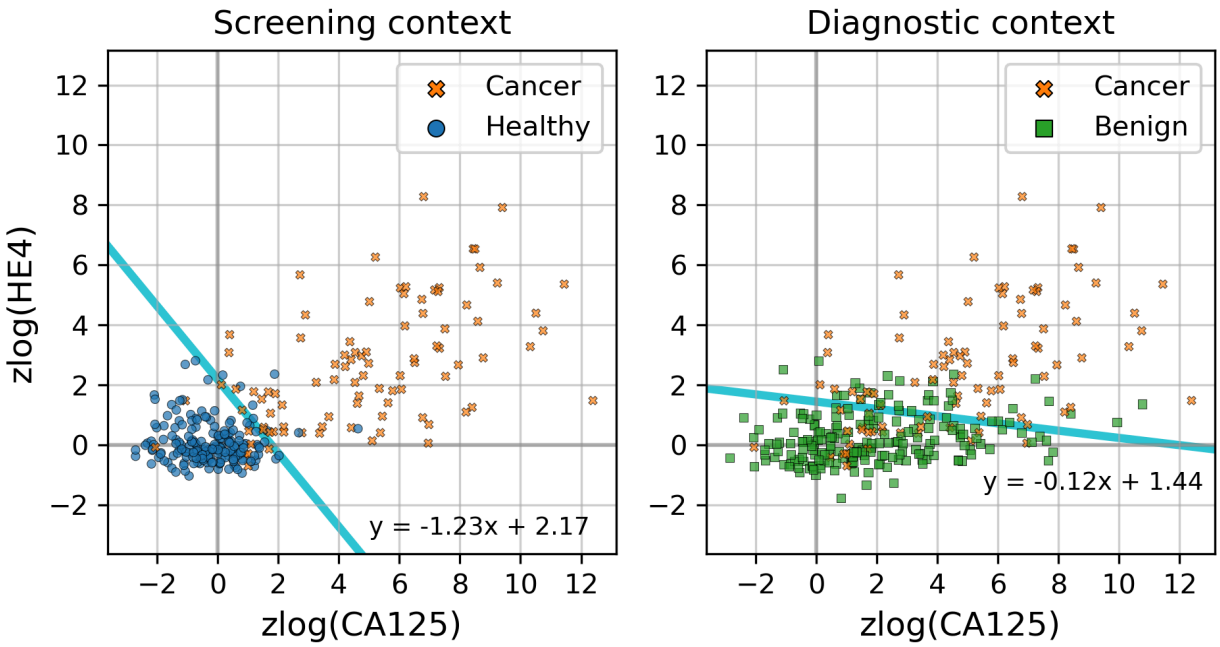


Figure 4. Decision boundaries learned by logistic regression models on the discovery cohort. On the discovery cohort, two protein-only models were fit to the data. In the screening context, the model distinguishes between samples from healthy donors and samples from patients with ovarian cancer. In the diagnostic context, the model distinguishes between samples from donors with benign adnexal masses and samples from patients with ovarian cancer. The decision boundaries for the two models are drawn in cyan, and equations of the decision boundary lines are written near the lines.

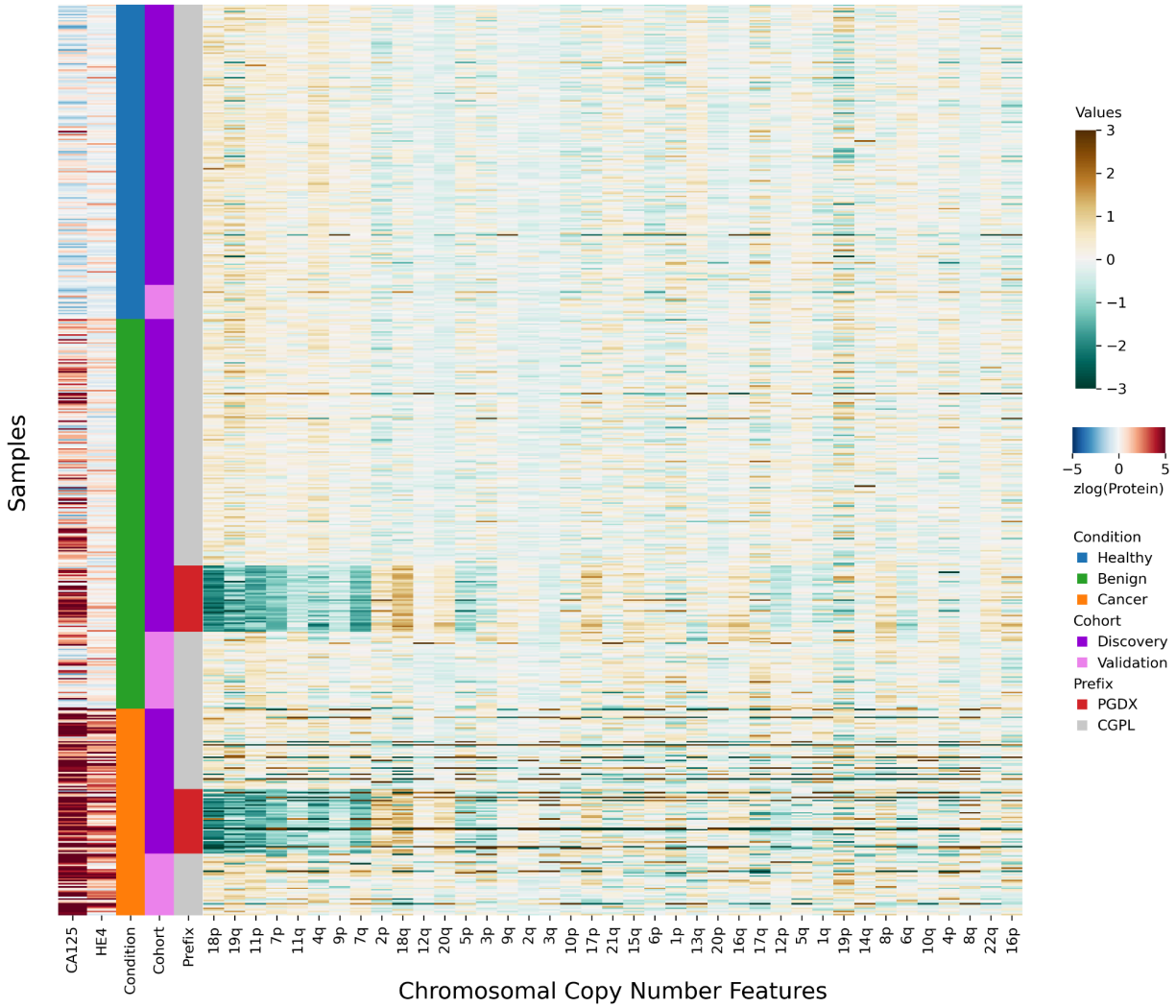


Figure 5. Heatmap demonstrating sample prefix-associated effect on copy number features. The samples from the DELFI-Pro dataset are shown here, with each row representing features from a single sample. For the proteins CA125 and HE4, we display the zlog-transformed values for each sample. Features are sorted such that the leftmost features are more affected by the batch effect. All features had values standardized here prior to display.

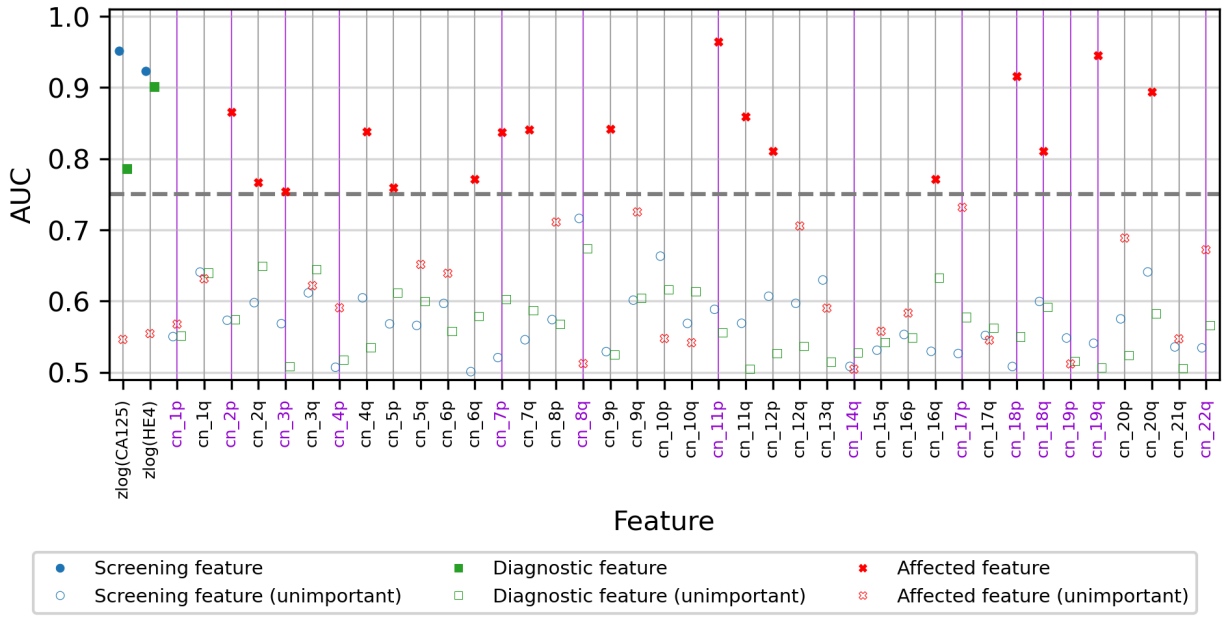


Figure 6. Areas under the curve for features measured in various contexts. For each displayed feature, we analyzed the discovery cohort and computed three receiver operating characteristic (ROC) curves and measured the area under the curve (AUC). Two curves were calculated on CGPL-prefixed samples, measuring a feature's association with cancer in a screening and a diagnostic context. The third curve was calculated on cancer samples only, and measured a feature's association with the two different prefixes (CGPL and PGDX) to determine how affected a feature was by batch confounding. Features with a "cn_" prefix are copy number features, and those with a "zlog" prefix are the zlog-transformed protein concentrations; copy number features listed in the DELFI-Pro results as contributing to the DELFI-Pro screening model performance are highlighted in violet. We use an AUC threshold of 0.75 to separate features that are good predictors from those features that are less important; features described as "unimportant" in the figure are those with AUC values less than 0.75.

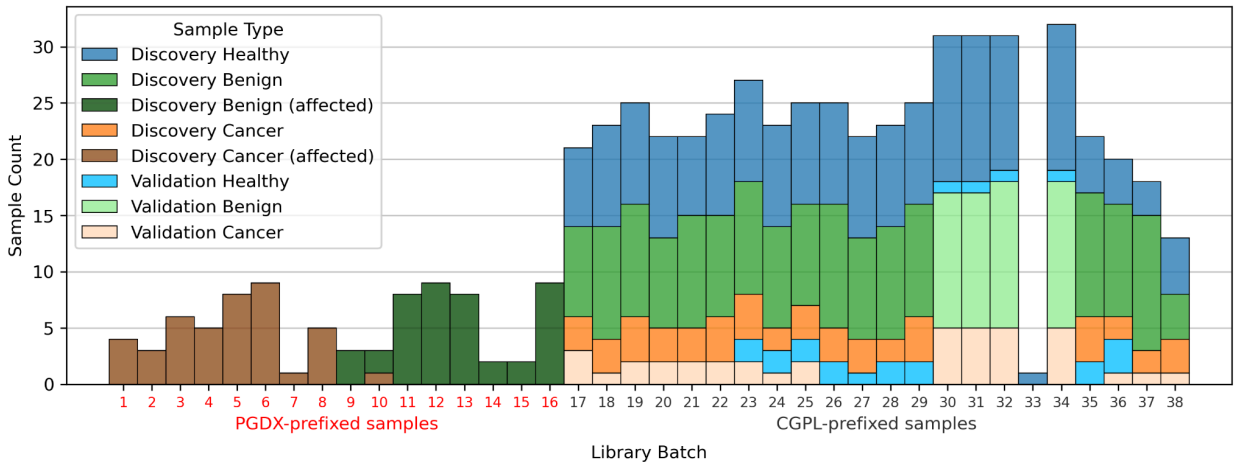
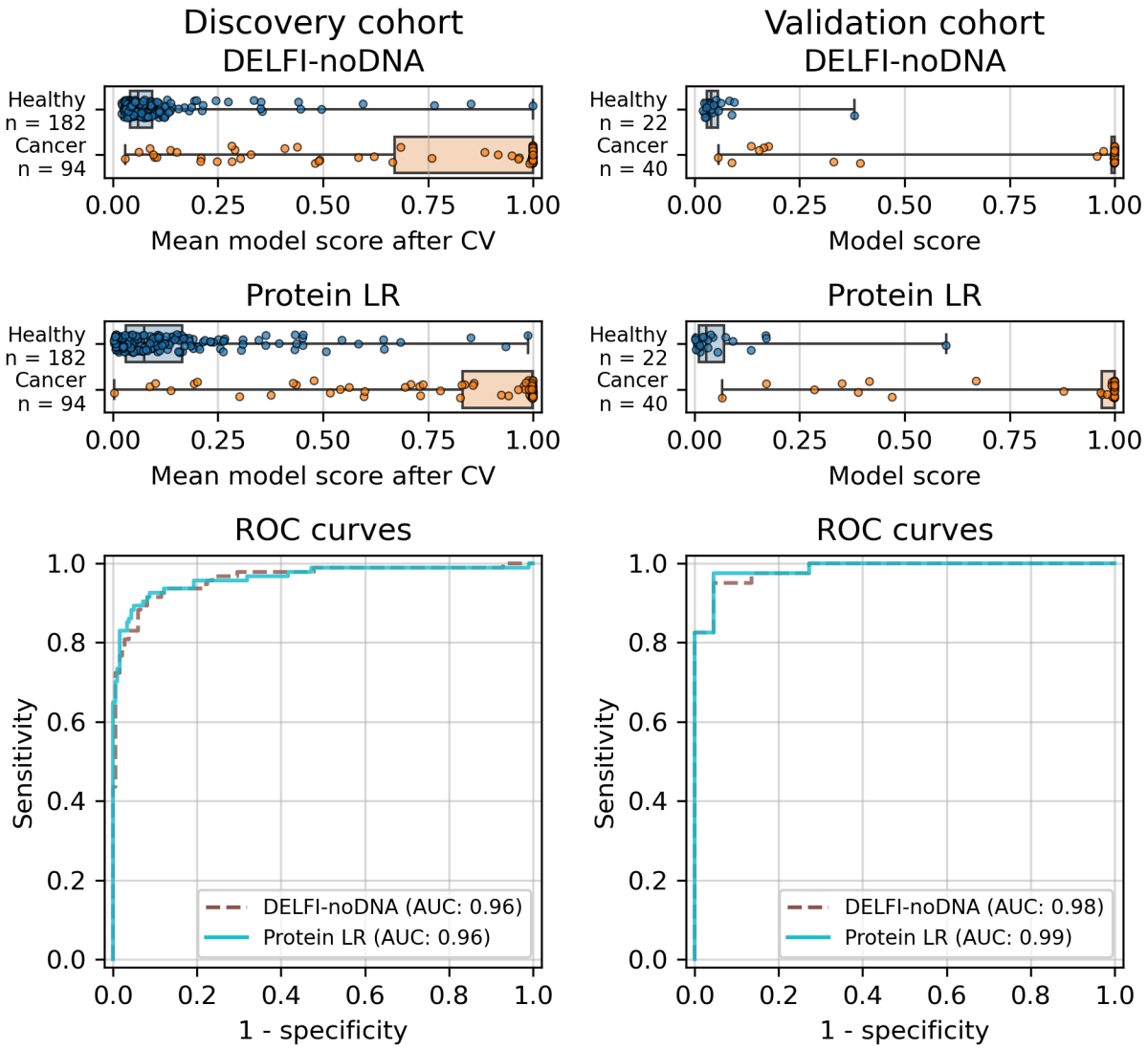
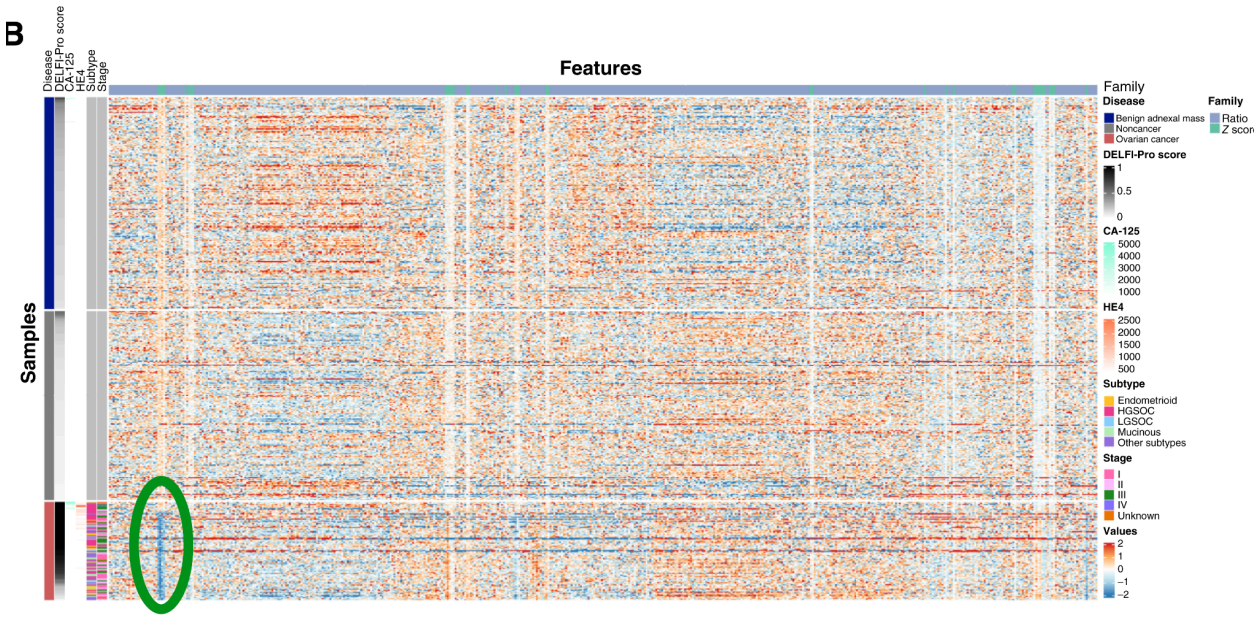


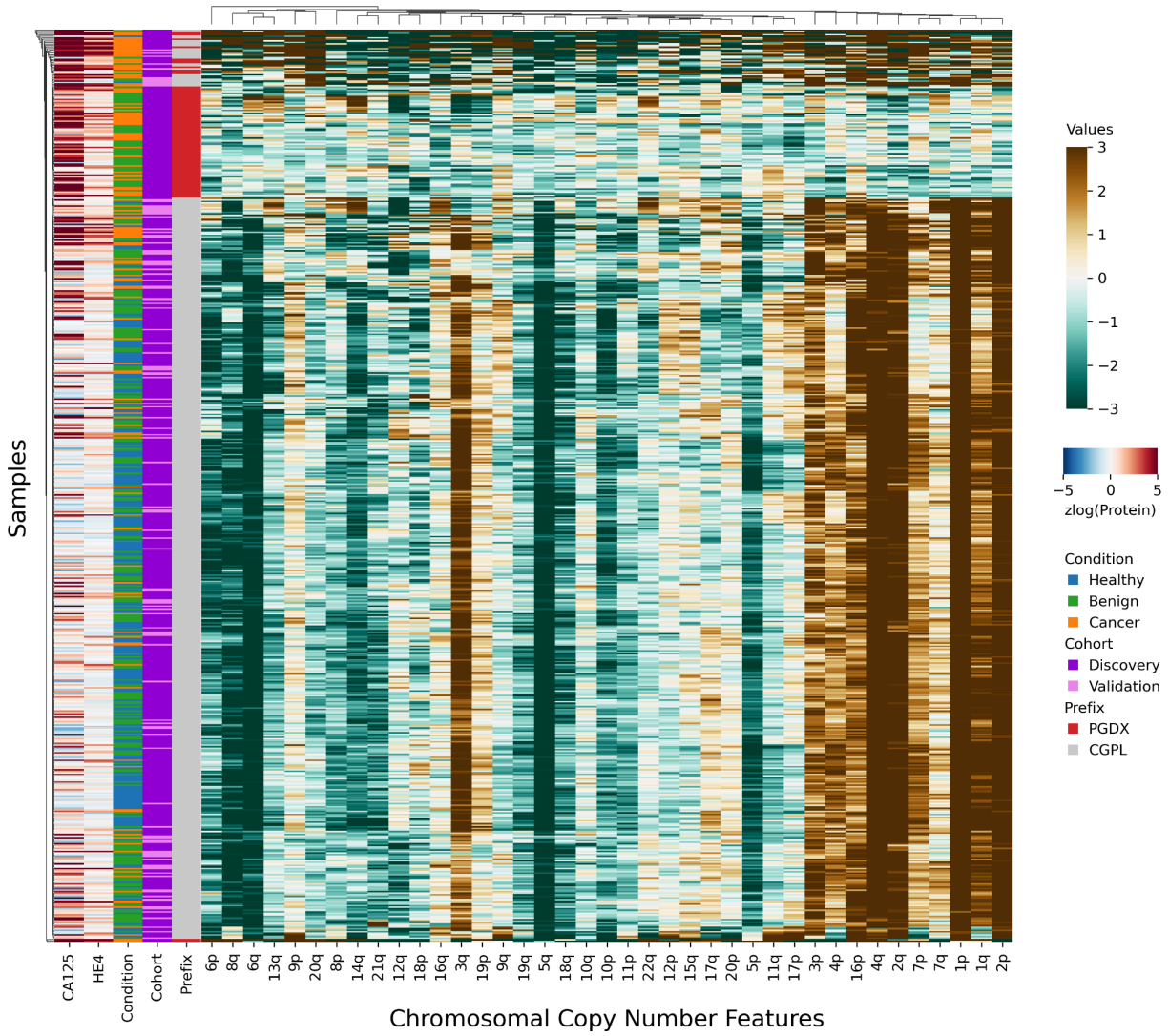
Figure 7. Histogram of various sample types by library batch. All samples used as part of the DELFI-Pro publication's dataset are divided into three conditions (healthy, benign, and cancer), and into two cohorts (discovery and validation). Samples with the "PGDX" sample prefix-associated effect (labelled "affected") are colored separately. The 16 batches that only have affected samples are colored red, with the other batches that only have CGPL-prefixed samples colored gray.



Supplementary Figure S1. Comparison of our protein screening model with a model made available by DELFI-Pro authors. Cross validation ("CV") results for the discovery cohort, and validation results on the validation cohort are shown for both the DELFI-Pro authors' two-protein logistic regression screening model ("DELFI-noDNA") and our two-protein logistic regression ("Protein LR") screening model. Both models seek to distinguish between samples from healthy donors and samples from patients with ovarian cancer. Receiver operating characteristic (ROC) curves demonstrate the performance of the two models at various score thresholds, with the area under the curve (AUC) growing proportionally to model accuracy.



Supplementary Figure S2. Heatmap from DELFI-Pro publication with incongruous region highlighted. This is Figure 2B from the Medina, *et al*. publication, with the addition of a green ellipse highlighting the copy number ("Z score") features that, according to this figure, only had negative values in cancer samples. Such features should be very good predictors for cancer status, but the figure shows that they are not well-correlated with the assigned probability of cancer ("DELFI-Pro score"). After analysis of the code and data used to generate the heatmap, we found that due to an error in constructing this visualization, the row values within the heatmap only reflect the order of the input file, and do not reflect the grouping indicated by the "Disease" and "DELFI-Pro score" columns on the left side.



Supplementary Figure S3. Clustermap demonstrating sample prefix-associated effect on copy number features. The samples from the discovery cohort are shown here, with each row representing features from a single sample.