Electronic Supplementary Material (ESI) for Analyst

Fluorescent Polymer-Based Post-Translational Differentiation and Subtyping of Breast Cancer Cells

Michael D. Scott,[1] Rinku Dutta,[1] Manas Haldar,[1] Anil Wagh,[1] Thomas R. Gustad,[2] Benedict Law,[1] Daniel L. Friesner[3] and Sanku Mallik[1]

[1]Department of Pharmaceutical Science,
[2]Department of Veterinary and Microbiological Sciences,
[3]Department of Pharmacy Practice,
North Dakota State University, Fargo, ND 58102, USA

Sanku.Mallik@ndsu.edu

Supporting Information

Table of Contents

Polymer Characterization:

Monomer synthesis and characterization has already been reported.[19] **Supplementary Table 1** below shows the characterization by GPC and the concentration of the polymer used for the fluorescent experiments.

Supplementary Table 1 This table shows the results of the GPC that was used to determine the molecular weights and concentration of each polymer used for the fluorescent studies.

Polymer	M_w	M_n	P.I.	Concentration (nM)
Polymer M	117,191	64,577	1.86	31
Polymer V	114,428	78,161	1.46	27

Cell Culture Studies:

MCF-7 is a classical human breast cancer cultured by a 69 year old female.[20] This cell line was established by the Michigan Cancer Foundation in 1973.[20] It is a commonly used breast cancer model that should respond to several traditional drug therapies.[20]

MDA-MB-231 is another human breast cancer cell line which was cultured by a 51 year old female.[21] It distinguishes itself from MCF-7 in that it has a mutant p53 gene.[22] It also differs in that it is a multiple drug resistant breast cancer cell line.[23] Thus, being diagnosed with a breast cancer that is similar to this one means administration of many drugs will not successfully control this cancer. The patient will then suffer through the medications painful side-effects and the total financial cost of therapy will be maximized. Thus more aggressive means of detection coupled with distinguishing between cancer sub-typing, need to be explored.

To demonstrate that our polymers can also distinguish breast cancer from non-breast cancer, we choose to explore the polymer's response to HeLa cells. This cervical cancer cell line was taken from a female.[24] This is one of the most commonly used and oldest cancer cell lines.[25]

HEK-293 is a human embryonic kidney cell line.[26] This is a non-cancerous cell line, but has been reported to secrete MMP-9.[27] This cell line was used to demonstrate our polymers ability to distinguish between cancerous and non-cancerous cells, which secrete MMP-9.

All cell lines were cultured as instructed and maintained at 37˚C in an atmosphere of 5% CO_2 in

humidified air. HEK-293, MCF-7 and HeLa were all grown in MEM media with 10% FBS and 1% antibiotics. MDA-MB-231 was grown in DMEM media with 10% FBS and 1% antibiotics. They were all sub-cultured (each 3-5 days) for a total of three splits in their respected phenol red media. They were then split twice (each 3-5 days) in their respected phenol-free-media. All splits used as needed HBSS and Trypsin-Versene. Upon reaching a confluent state, the cells were then aseptically transferred from the cultured flask into centrifuge tube. Before being allocated for their fluorescent studies cells were then centrifuged at 1500 rpm for 8 minutes to pellet and remove any remaining cells.

Experimental Information Fluorescence Spectroscopy Study:

The fluorescent experiments were performed using a Fluoromax-4 Spectroflourometer by HoribaJobinYvon. Polymers (2 mg) were weighted out and dissolved in 2 mL of 30 mM phosphate buffer (pH = 7.4). They were then diluted to achieve the desired concentration (Polymer M at 31 nM) (Polymer V at 27 nM) in the cuvette. 50 µL of the corresponding conditioned media was added and mixed into the cuvette. The solution in the cuvette was excited at 325 nm, using a 395 cut-off filter. The emission spectra's was recorded between 350 nm and 750 nm. The first peak was noticed (410 nm and 420 nm was the peak emission intensities for the Polymer V and Polymer M polymers respectively. Similarly a second peak developed at 510 nm and 541 nm for Polymer V and Polymer M respectively). The same procedure for unconditioned media was used.

Fluorescent Ratios Determination:

To mathematically eliminate the fluorescent signal contributions from the media we choose to calculate the ratios of the conditioned cell culture media over the corresponding unconditioned media at three different wavelengths. By performing this step we eliminated any potential variation caused by growing the cell lines in different cell culture media (MEM or DMEM). These ratios (**Supplementary Table 2**) were submitted for statistical analysis.

Supplementary Table 2 Table of ratios generated from fluorescent experiments.

Polymer	Cell Line	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8
M	MDA-MB-231$_{420\ nm}$	1.080916	0.940839	0.932647	0.911347	0.881809	0.902869	0.887206	0.905851
	MCF-7$_{420\ nm}$	1.606264	1.62015	1.841565	1.708477	1.654629	1.679794	1.676482	1.611221
	HeLa$_{420\ nm}$	2.03411	2.01602	2.0644	2.078443	2.010565	2.026123	2.010012	2.017582
	HEK-293$_{420\ nm}$	1.734831	1.766465	1.841755	1.788262	1.767768	1.851178	1.846879	1.842947
	MDA-MB-231$_{523\ nm}$	1.119972	1.082309	1.055043	1.038331	1.069302	1.045312	1.050322	1.008249
	MCF-7$_{523\ nm}$	1.07821	1.087998	1.07334	1.103308	1.081348	1.095925	1.083516	1.078523
	HeLa$_{523\ nm}$	1.067634	1.083515	1.065766	1.105408	1.060263	1.07273	1.051424	1.032221
	HEK-293$_{523\ nm}$	1.234752	1.251697	1.234789	1.288396	1.248991	1.264972	1.266787	1.236204
	MDA-MB-231$_{541\ nm}$	1.081602	1.066703	1.089008	1.073585	1.086683	1.053653	1.053267	1.019529
	MCF-7$_{541\ nm}$	1.041337	1.051019	1.081892	1.053521	1.066059	1.076953	1.062204	1.083289
	HeLa$_{541\ nm}$	1.039678	1.031022	1.067872	1.032628	1.02431	1.003782	1.020121	1.018158
	HEK-293$_{541\ nm}$	1.174917	1.181515	1.185666	1.174749	1.186022	1.203861	1.209758	1.197237
V	MDA-MB-231$_{410\ nm}$	0.95345	0.925732	0.911819	0.909788	0.854827	0.747671	0.807543	0.897796
	MCF-7$_{410\ nm}$	1.430649	1.595537	1.542606	1.541282	1.581215	1.683029	1.721094	1.819321
	HeLa$_{410\ nm}$	1.859018	1.863717	1.991359	1.974274	1.688984	1.777685	1.824472	1.897512
	HEK-293$_{410\ nm}$	1.35795	1.462229	1.533244	1.616637	1.229342	1.432922	1.467823	1.454648
	MDA-MB-231$_{510\ nm}$	1.288346	1.222454	1.185932	1.143421	1.064746	1.037664	1.025797	1.012537
	MCF-7$_{510\ nm}$	1.158239	1.119843	1.044218	1.020922	1.153711	1.152986	1.116271	1.112301
	HeLa$_{510\ nm}$	1.443343	1.308815	1.279059	1.280228	1.27433	1.195992	1.216297	1.194899
	HEK-293$_{510\ nm}$	1.205637	1.153849	1.148267	1.147307	1.13584	1.118666	1.100398	1.080768
	MDA-MB-231$_{541\ nm}$	1.283912	1.232823	1.157046	1.141286	1.053587	1.047762	1.007657	1.019795
	MCF-7$_{541\ nm}$	1.171913	1.094961	1.041461	0.989715	1.149254	1.138179	1.113694	1.123459
	HeLa$_{541\ nm}$	1.411381	1.262355	1.277228	1.217731	1.218143	1.192556	1.18488	1.142353
	HEK-293$_{541\ nm}$	1.170907	1.153743	1.135018	1.130831	1.11153	1.10938	1.083695	1.069458

Synopsis - Statistical Analysis:

As noted earlier in the manuscript, the experimental design consists of a set of four cell lines and two fluorescent polymers. The primary objective of the empirical analysis is to select the polymer that most effectively predicts (or discriminates between) the different cell lines. This study uses linear discriminant analysis (LDA) to identify the polymer with the greatest predictive power.[28] In a traditional application of LDA, emissions intensity data from each of the two polymers (the predictor variables) and four cell lines (the dependent variable) are replicated a total of eight times, yielding a total of 32 observations available for analysis.

One complicating feature of the current study is that each of the polymers achieves multiple peak emission intensity ratios. Without identifying the peak emission intensity for each polymer, it is impossible for LDA to accurately and precisely identify the polymer that best discriminate across (or predicts) each the cell lines. In such cases, the polymer identified by LDA may, in fact, be the polymer that gives maximum discrimination, or it may only be superior because it was compared to a polymer whose emission intensities were not measured at their true peak values.

To account for this possibility, a stepwise analysis was employed, the details of which are contained in the following sections of this document. Analysis of each polymer identified three possible emission peak intensities: 420 nm, 520 nm and 541 nm for Polymer M; and 410 nm, 510 nm and 541 nm for Polymer V. We conducted a separate LDA for each polymer to identify the emission intensity that best predicts (or discriminates between) the four cell lines for that polymer. As noted above, each of these analyses is based on 32 observations (eight replicates times four cell lines) and 4 variables (a variable indicating the cell line in question and three variables identifying the emissions intensity for the specified cell line at a given wavelength). Because there is prior information identifying each of these wavelengths as potential candidates for inclusion in the final analysis (as well as a sufficient number of observations), we took the conservative approach of including all three wavelengths in a given LDA procedure, as opposed to using a second stepwise procedure where wavelengths are eliminated using a Wilks' Lambda or F-statistic prior to identifying the discriminant functions.[29] We note in passing that the analysis was replicated using a second set of stepwise LDA procedures, and these replications produced very similar results. Once the optimal emission intensities for each polymer are identified, a third application of LDA was used to compare the optimal wavelengths that were identified in the previous two applications of LDA.

As noted in the supporting documents, the emission intensity giving Polymer M the "best" predictive power occurred at 420 nm. Similarly, LDA identified the 410 nm wavelength as the intensity of choice for Polymer V. **Supplementary Table 3** contains means F-statistics and Wilks' Lambda values for each of the two polymers, disaggregated by cell line type. All F-statistics have associated p-values less than 0.05 indicating significant differences exist across group means for each cell lines. For each cell line, Polymer M (evaluated at 420 nm) exhibits higher mean values and lower Wilks' Lambda values than Polymer V (evaluated at 410 nm). Additionally, for each polymer, the HeLa cell line exhibits the highest mean emission intensity value, while the MDA-MB-231 cell line exhibits the lowest mean values.

Supplementary Table 3 This table shows the tests of emission intensity ratio at 420 nm for Polymer M and 410 nm for Polymer V.

	Polymer M (at 420 nm)	Polymer V (at 410 nm)
MDA-MB-231	0.930	0.876
MCF-7	1.675	1.614
HeLa	2.032	1.860
HEK-293	1.805	1.444
Wilks' Lambda	0.016	0.066
F-Statistic	570.909	131.264
P-Value	< 0.001	< 0.001

LDA extracted two canonical variables, each with its own discriminant function. At the 5% level, chi-square tests indicate that both canonical functions significantly explain the four cell lines. The first canonical function is the more important of the two, as it explains 98.3% of the

variation across cell lines, while the remaining function explains only 1.7% of this variation. Analysis of these canonical functions suggests that Polymer V contributes more to the formation of the second canonical function, while Polymer M contributes relatively more to the first canonical function.

Supplementary Table 4 contains the structure matrix and the cumulative potency indices, which can be used to assess the overall contribution of each polymer (evaluated at the "best" emission intensity) to the ability of LDA to discriminate between (or predict) the four cell lines. The potency indices suggest that Polymer M provides the largest overall contribution to the model's ability to distinguish between the cell lines when compared to Polymer V.

Supplementary Table 4 This table shows the potency index of various functions for Polymer M and Polymer V.

	Polymer M	Polymer V
Function 1	0.968	0.449
Function 2	-0.250	0.894
Potency Index	0.922	0.212

Statistically speaking, these results are intuitive. As noted in **Supplementary Table 3**, the Polymer M has the highest mean emission intensity values. This polymer also is the primary determinant of the first canonical function, which explains the vast majority of the variation in the data.

Supplementary Table 5. Canonical function summary[a] of **Polymer M** and **Polymer V**.

Fct.	Eigen-value	Pct. of Variance Explained	Canonical Correl.	Wilks' Lambda[a]	Chi-Square Statistic	P-Value
1	0.930	98.3	0.992	0.007[b]	138.879	<0.001
2	1.675	1.7	0.732	0.464[c]	21.494	<0.001

[a] Lower values of Wilks' Lambda indicate greater discrimination. Wilks' Lambda and chi-square tests apply sequentially. [b] tests functions 1-3 cumulatively. [c] tests functions 2-3 cumulatively [d] tests functions 3.

Statistical Data Analysis of Polymer M:

LDA has been used extensively in the literature, and the reader is referred to these sources for additional detail on the mechanics of LDA.[28-29] Within this analysis, we assess the LDA results using several standard metrics. Standard F tests and Wilks' Lambda values are used to access mean differences across each of our cell lines and identify the ability of the predictor variables (either the emission intensities for a single polymer or the two polymers evaluated at optimal

emission intensities) to discriminate across the four cell lines. The significance of the canonical correlations discriminant functions are assessed using chi-square tests. An overall "potency index" for each predictor variable (either emission intensity or a given polymer) is used to identify the predictor variables which play the largest role in the entire system of canonical discriminant functions. Higher values for each index signal the overall importance of each predictor variable to the model as a whole. Overall model fit is assessed by examining canonical function plots to identify whether each of the group centroids (one for each of the four cell lines) is sufficiently distinct. A large amount of overlap between the data points of two or more groups indicates poor discrimination across the cell lines, and by extension poor model fit. The model's internal validity is assessed by comparing the percentage of cell line observations that are correctly predicted by the model. All predicted values are computed using both traditional and (leave one out) cross-validation techniques. Models with a high degree of internal validity should correctly predict a high percentage of observations, and display consistency in predicted values across both techniques. All statistical analyses were conducted using the PASW (formerly SPSS) Statistical Package, Version 18.

Supplementary Table 6 contains means, F-statistics and Wilks' Lambda values for each Polymer M's emission intensity, disaggregated by cell line type. All F-statistics have associated p-values less than 0.05, indicating significant differences exist across group means for each cell line. For the MDA-MD-231 cell line, the 541 nm emission intensity appears to be the highest value. For all other cell lines, the highest mean emission intensities appear at 420 nm. Wilks' Lambda values are lowest for 420 nm, followed by 523 nm and 541 nm.

Supplementary Table 6 Tests of equality of group means for Polymer M.

Cell Line	420 nm[a,b]	523 nm[a,b]	541 nm[a,b]
MDA-MB-231	0.930	1.059	1.066
MCF-7	1.675	1.085	1.065
HeLa	2.032	1.067	1.030
HEK	1.805	1.253	1.189
Wilks' Lambda	0.016	0.065	0.072
F-Statistic [3,28]	570.909	134.082	120.903
P-Value	<0.001	<0.001	<0.001

[a] first panel provides group-specific means [b] second panel provides statistics and p-values.

Supplementary Table 7 identifies the number of significant canonical correlations and canonical functions. At the 5% level, two of three canonical functions significantly explain the four cell lines. Of these, the first canonical function is most important, as it explains 80.2% of the variation across cell lines. The remaining functions explain 19.8% and 0.0002%,

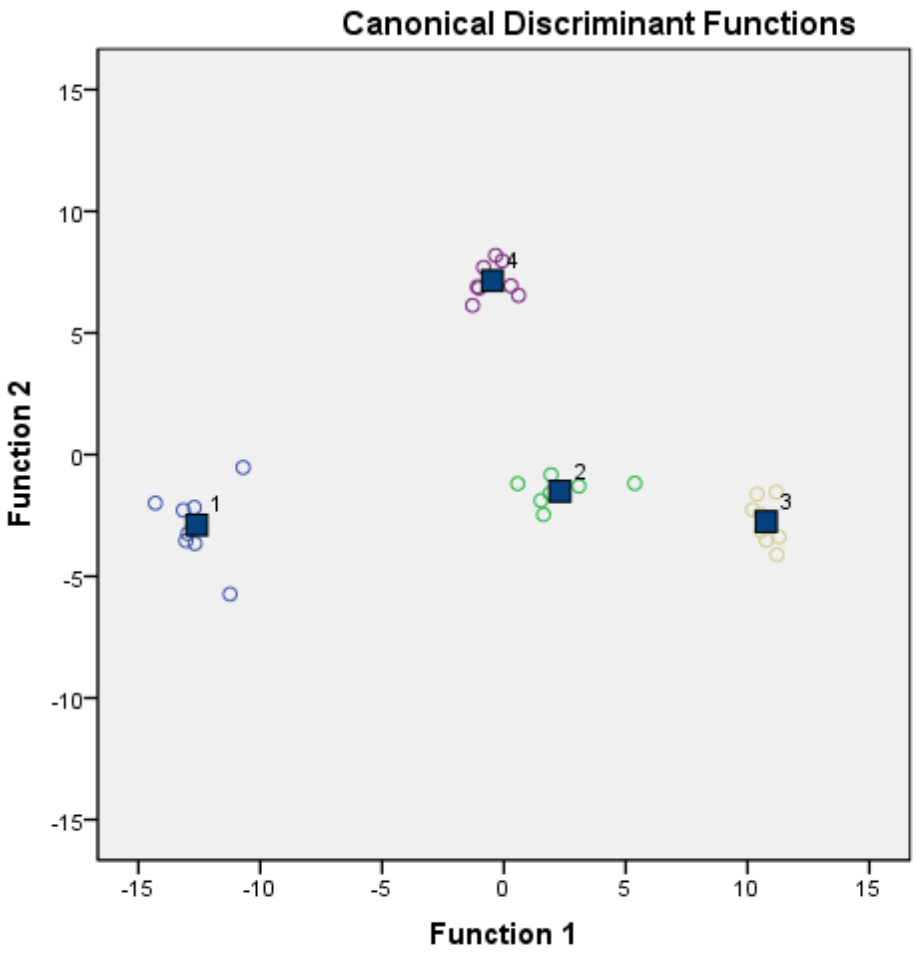
respectively. As in the previous analysis, these results lead us to focus primarily on the first discriminant function.

Supplementary Table 7 Canonical function summary[a] for Polymer M.

Fct.	Eigen-value	Pct. of Variance Explained	Canonical Correl.	Wilks' Lambda[a]	Chi-Square Statistic	P-Value
1	80.085	80.2	0.994	0.001[b]	204.861	<0.001
2	19.806	19.8	0.976	0.047[c]	83.985	<0.001
3	0.019	0.0002	0.136	0.981[d]	0.516	0.473

[a] Lower values for Wilks' Lambda indicate greater discrimination. Wilks' Lambda and chi-square tests apply sequentially. [b] tests functions 1 – 3 cumulatively. [c] tests functions 2 – 3 cumulatively [d] tests function 3.

Supplementary Figure 1 contains a canonical function plot of the first two canonical functions (explaining 99.9% of the variation in the cell lines). Note that each of the cell line is clearly distinguished as a group in the plot. Moreover, traditional and cross-validated discriminant functions each correctly predicted 100.0% of the cell lines, respectively, indicating a high likelihood of interval validity.



Supplementary Figure 1 Polymer M's canonical correlation plot between two largest canonical correlations and each of the four cell lines: MDA-MB-231 (group 1), MCF-7 (group 2), HeLa (group 3) and HEK (group 4).

Supplementary Table 8 contains the standardized discriminant function coefficients, which measure the relative contributions of each emission intensity to a specific discriminant function. For function 1, the 420 nm wavelength exhibits the highest coefficient in absolute value. Additionally, the 523 nm and 541 nm emission intensities carry values which (in absolute magnitude) are much smaller in absolute magnitude than for 420 nm. Concomitantly, the 523 nm exhibits the highest value for the second function, while 541 nm has the largest coefficient for the third (insignificant) canonical function. In both the second and third canonical functions, the coefficient values for the 420 nm variable suggest that the 420 intensities have very little contribution to the second and third canonical discriminant functions. On the other hand, the 523 nm and 541 nm coefficient values for the second and third functions are large in absolute value, implying that these predictors contribute substantially to these functions.

Supplementary Table 8 Standardized canonical discriminant function coefficients for Polymer M.

Predictor	Canonical Function 1	Canonical Function 2	Canonical Function 3
420 nm	1.152	0.050	0.103
523 nm	-0.268	0.641	-0.525
541 nm	-0.495	0.542	0.628

To assess the overall contribution of each emission intensity to the discriminatory power of the LDA, we present **Supplementary Table 9**, which contains the structure matrix and the cumulative potency indices. The potency indices suggest that 420 nm emission intensity provides the largest overall contribution to the model's ability to distinguish between the cell lines.

Supplementary Table 9 Structure matrix and potency index for Polymer M.

Predictor	Canonical Function 1	Canonical Function 2	Canonical Function 3	Potency Index
420 nm	0.832	0.537	0.139	0.612
523 nm	0.010	0.851	-0.525	0.144
541 nm	-0.088	0.789	0.608	0.130

On total, the LDA has a clear and intuitive interpretation. The results in **Supplementary Table 6** suggest that the first canonical function is, by far, the most important discriminant function. **Supplementary Tables 8** and **Supplementary Table 9** jointly suggest that the 420 nm variable contributes the most towards the first canonical discriminant function, while the 523 nm and 541 nm variables contribute relatively more to the second and third canonical discriminant functions. This implies that the 420 nm emission intensity is the "best" determinant of the cell lines for the Polymer M. As with Polymer V analysis, the Wilks' Lambda and F-statistics in **Supplementary Table 6** supports this assertion, as the 420 nm variable exhibits the highest mean values for 3 of the 4 cell lines.

Statistical Data Analysis of Polymer V:

Supplementary Table 10 contains means, F-statistics and Wilks' Lambda values for each Polymer V's emission intensity, disaggregated by cell line type. We note in passing that smaller values for the Wilks' Lambda indicate a greater potential for the given emission intensity to discriminate across cell lines. All F-statistics have associated p-values less than 0.05, indicating significant differences exist across group means for each cell lines. For the MDA-MD-231 cell

line, the 510 nm emission intensity appears to be the highest value. For all other cell lines, the highest mean emission intensities appear at 410 nm. Wilks' Lambda values are lowest for 410 nm, followed by 510 nm and 541 nm.

Supplementary Table 10 Tests of equality of group means for Polymer V.

Cell Line	410 nm[a,b]	510 nm[a,b]	541 nm[a,b]
MDA-MB-231	0.876	1.123	1.118
MCF-7	1.614	1.110	1.103
HeLa	1.860	1.274	1.238
HEK	1.444	1.136	1.121
Wilks' Lambda	0.066	0.514	0.620
F-Statistic [3,28]	131.264	8.821	5.722
P-Value	<0.001	<0.001	0.003

[a] first panel provides group-specific means [b] second panel provides statistics and p-values.

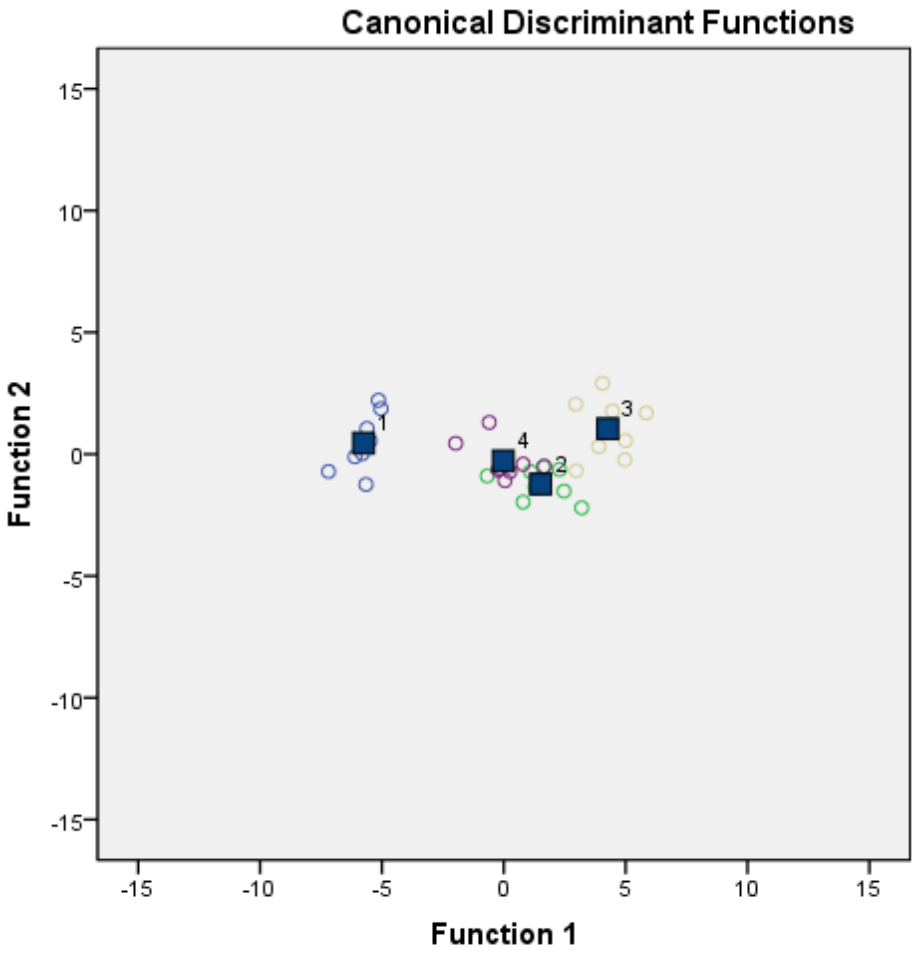
Supplementary Table 11 identifies the number of significant canonical correlations and canonical functions. At the 5% level, two of three canonical functions significantly explain the four cell lines. Of these, the first canonical function is most important, as it explains 94.8% of the variation across cell lines. The remaining functions explain 5.1% and 0.1%, respectively. Based on these results, we focus primarily on the first discriminant function.

Supplementary Table 11 Canonical function summary[a] for Polymer V.

Fct.	Eigen-value	Pct. of Variance Explained	Canonical Correl.	Wilks' Lambda[a]	Chi-Square Statistic	P-Value
1	15.309	94.8	0.969	0.033[b]	93.686	<0.001
2	0.823	5.1	0.672	0.541[c]	16.914	0.002
3	0.015	0.1	0.120	0.986[d]	0.396	0.529

[a] Lower values for Wilks' Lambda indicate greater discrimination. Wilks' Lambda and chi-square tests apply sequentially. [b] tests functions 1 – 3 cumulatively. [c] tests functions 2 – 3 cumulatively [d] tests function 3.

Supplementary Figure 2 contains a canonical function plot of the first two canonical functions (explaining 99.9% of the variation in the cell lines). Note that cell lines 1 (MDA-MB-231) and 4 (HEK-293) are clearly distinguished as a group in the plot, while groups 2 (MCF-7) and 3 (HeLa) overlap slightly. Traditional and cross-validated discriminant functions each correctly predicted 90.6% of the cell lines, respectively, indicating a reasonable (but not perfect) degree of interval validity.



Supplementary Figure 2 Polymer V's canonical correlation plot between two largest canonical correlations and each of the four cell lines: MDA-MB-231 (group 1), MCF-7 (group 2), HeLa (group 3) and HEK-293 (group 4).

Supplementary Table 12 contains the standardized discriminant function coefficients, which measure the relative contributions of each emission intensity to a specific discriminant function. For function 1, the 540 nm wavelength exhibits the highest coefficient in absolute value. However, the 410 nm and 510 nm emission intensities carry values which (in absolute magnitude) are only slightly smaller in absolute magnitude than for 540 nm. Concomitantly, the 510 nm exhibits the highest value for the second function, while 540 nm has the largest coefficient for the third (insignificant) canonical function. In both the second and third canonical functions, the coefficient values for the 410 nm variable suggest that the 410 intensities have very little contribution to the second and third canonical discriminant functions. On the other hand, the 510 nm and 540 nm coefficient values for the second and third functions are large in absolute value, implying that these predictors contribute substantially to these functions.

Supplementary Table 12 Standardized canonical discriminant function coefficients for Polymer V.

Predictor	Canonical Function 1	Canonical Function 2	Canonical Function 3
410 nm	1.008	-0.224	0.036
510 nm	1.193	3.087	-3.048
540 nm	-1.280	-2.284	3.692

To assess the overall contribution of each emission intensity to the discriminatory power of the LDA, we present **Supplementary Table 13**, which contains the structure matrix and the cumulative potency indices. The potency indices suggest that 410 nm emission intensity provides the largest overall contribution to the model's ability to distinguish between the cell lines.

Supplementary Table 13 Structure matrix and potency index for Polymer V.

Predictor	Canonical Function 1	Canonical Function 2	Canonical Function 3	Potency Index
410 nm	0.958	-0.109	0.265	0.871
510 nm	0.161	0.814	0.559	0.059
540 nm	0.123	0.673	0.730	0.038

On total, the LDA has a clear and intuitive interpretation. The results in **Supplementary Table 11** suggest that the first canonical function is, by far, the most important discriminant function. **Supplementary Table 12** and **Supplementary Table 13** jointly suggest that while all three emission intensity variables contribute to the first canonical function, the 510 nm and 540 nm variable contribute relatively more to the second and third canonical discriminant functions, while the 410 nm variable contributes very little to these latter functions. This implies that the 410 nm emission intensity is the "best" determinant of the cell lines for Polymer V. The Wilks' Lambda and F-statistics in **Supplementary Table 10** supports this assertion, as the 410 nm variable exhibits the highest mean values for 3 of the 4 cell lines.

Statistical Data Analysis of the Polymer M and Polymer V at Optimal Emission Intensities:

Supplementary Table 14 contains means, F-statistics and Wilks' Lambda values for each of the two polymers, disaggregated by cell line type. All F-statistics have associated p-values less than 0.05, indicating significant differences exist across group means for each cell lines. For each cell line, Polymer M (evaluated at 420 nm) exhibits higher mean values and lower Wilks' Lambda values than Polymer V (evaluated at 410 nm). Additionally, for each polymer, the HeLa cell line

exhibits the highest mean emission intensity values, while the MDA-MB-231 cell line exhibits the lowest mean values.

Supplementary Table 14 Tests of equality of group means of Polymer M and Polymer V.

Cell Line	Polymer M (at 420 nm) [a,b]	Polymer V (at 410 nm)[a,b]
MDA-MB-231	0.930	0.876
MCF-7	1.675	1.614
HeLa	2.032	1.860
HEK	1.805	1.444
Wilks' Lambda	0.016	0.066
F-Statistic [3,28]	570.909	131.264
P-Value	<0.001	<0.001

[a] first panel provides group-specific means [b] second panel provides statistics and p-values.

Supplementary Table 15 identifies the number of significant canonical correlations and canonical functions. At the 5% level, both canonical functions significantly explain the four cell lines. The first canonical function is most important, as it explains 98.3% of the variation across cell lines, while the remaining function explains 1.7%.

Supplemntary Table 15 Canonical function summary[a] of Polymer M and Polymer V.

Fct.	Eigen-value	Pct. of Variance Explained	Canonical Correl.	Wilks' Lambda[a]	Chi-Square Statistic	P-Value
1	65.177	98.3	0.992	0.007[b]	138.879	<0.001
2	1.155	1.7	0.732	0.464[c]	21.494	<0.001

[a] Lower values for Wilks' Lambda indicate greater discrimination. Wilks' Lambda and chi-square tests apply sequentially. [b] tests functions 1 – 3 cumulatively. [c] tests functions 2 – 3 cumulatively [d] tests function 3.

Figure 3 (in main text) contains a canonical function plot of the first two canonical functions. Note that each of the cell line is clearly distinguished as a group in the plot. Moreover, traditional and cross-validated discriminant functions each correctly predicted 96.9% and 93.8% of the cell lines, respectively, indicating a high likelihood of interval validity.

Supplementary Table 16 contains the standardized discriminant function coefficients, which measure the relative contributions of each emission intensity to a specific discriminant function. For function 1, Polymer M exhibits the highest coefficient in absolute value, while Polymer V exhibits the highest value for the second function.

Supplementary Table 16 Polymer M and Polymer V's standardized canonical discriminant function coefficients.

Predictor	Canonical Function 1	Canonical Function 2
Polymer M	0.914	-0.459
Polymer V	0.256	0.991

To assess the overall contribution of each polymer (evaluated at the "best" emission intensity) to the discriminatory power of the LDA, we present **Supplementary Table 17**, which contains the structure matrix and the cumulative potency indices. The potency indices suggest that Polymer M provides the largest overall contribution to the model's ability to distinguish between the cell lines, when compared to Polymer V.

Supplementary Table 17 Structure matrix and potency index.

Predictor	Canonical Function 1	Canonical Function 2	Potency Index
Polymer M	0.968	-0.250	0.922
Polymer V	0.449	0.894	0.212

Statistically speaking, these results are intuitive. As noted in **Supplementary Table 14**, Polymer M has the highest mean emission intensity values. This polymer also is the primary determinant of the first canonical function (**Supplementary Table 15** and **Supplementary Table 16**), which explains the vast majority of the variation across the four cell lines. Overall, this implies that the Polymer M is the "best" determinant of the cell lines evaluated in this analysis.

Tissue Samples

Supplementary Table 18 Table of fluorescence emission intensity ratios collected from tissue samples at wavelength of 410 nm using Polymer **M**.

	$\text{Breast}_{410\ nm}$	$\text{Kidney}_{410\ nm}$	$\text{Liver}_{410\ nm}$	$\text{MDA-MB-231}_{410\ nm}$
Run 1	0.785669390	0.509715026	0.158913413	0.795036029
Run 2	0.380777096	0.597025017	0.640076579	0.982608696
Run 3	0.368061486	0.550145349	0.669571046	0.903418803
Run 4	0.402304369	0.610594130	0.696765499	0.923404255
Run 5	0.429188670	0.585215606	0.665562914	0.917832168
Run 6	0.456976179	0.571527298	0.739130435	0.876182287
Run 7	0.486243122	0.636891794	0.703175632	0.806089744
Run 8	0.435971223	0.658357771	0.752311757	0.825680272
Run 9	0.429752066	0.609978918	0.673429319	0.865072588
Run 10	0.437997192	0.606060606	0.699296225	0.862184874
Run 11	0.460373549	0.611629576	0.679127726	0.923265306
Run 12	0.469346734	0.612762872	0.611604500	0.876305221
Run 13	0.449317739	0.620928621	0.649523810	0.903908795
Run 14	0.462537463	0.681105302	0.625000000	0.850318471
Run 15	0.495145631	0.608938547	0.762299941	0.941730934
Run 16	0.482901554	0.633823529	0.666464891	0.838920687
Run 17	0.501274860	0.615658363	0.668018540	0.848856209
Run 18	0.492553191	0.664150943	0.671215881	0.823101777
Run 19	0.492483152	0.650916784	0.618471688	0.850686037
Run 20	0.488798371	0.611228070	0.689116055	0.923203964

Supplementary Table 19 Table of fluorescence emission intensity ratios collected from tissue samples at wavelength of 510 nm using Polymer **M**.

	$\text{Breast}_{510\ nm}$	$\text{Kidney}_{510\ nm}$	$\text{Liver}_{510\ nm}$	$\text{MDA-MB-231}_{510\ nm}$
Run 1	1.187765957	1.135219559	1.055590242	1.088529906
Run 2	1.134260235	1.111629079	1.118110619	1.098704484
Run 3	1.117064062	1.088330375	1.111567894	1.092202691
Run 4	1.121580878	1.083873883	1.121064212	1.088481675
Run 5	1.136165336	1.086248690	1.110832407	1.070738589
Run 6	1.123589269	1.080210734	1.093467829	1.052376366
Run 7	1.118583606	1.086108773	1.086423910	1.052229472
Run 8	1.138002138	1.083691641	1.120442777	1.051055740
Run 9	1.136178862	1.083139136	1.139542293	1.036351955
Run 10	1.135041739	1.094378492	1.112627804	1.023951295
Run 11	1.115104884	1.075912459	1.128491319	1.041495762
Run 12	1.125787898	1.088727601	1.152608015	1.038188337
Run 13	1.126156137	1.080004180	1.158794274	1.028433314
Run 14	1.103179224	1.073831066	1.158098003	1.031888907
Run 15	1.094186047	1.085574401	1.173907009	1.052346631
Run 16	1.093902964	1.080188348	1.145699385	1.045747466
Run 17	1.093398765	1.072343553	1.161240872	1.046128945
Run 18	1.099936194	1.059716529	1.137265850	1.051871621
Run 19	1.069069895	1.057461616	1.106760680	1.047046290
Run 20	1.088186422	1.079633721	1.125629829	1.064049156

Supplementary Table 20 Table of fluorescence emission intensity ratios collected from tissue samples at wavelength of 541 nm using Polymer M.

	Breast$_{541\ nm}$	**Kidney$_{541\ nm}$**	**Liver$_{541\ nm}$**	**MDA-MB-231$_{541\ nm}$**
Run 1	1.176621386	1.110407745	1.070169713	1.070888738
Run 2	1.129121537	1.083530246	1.091119234	1.059482699
Run 3	1.144059487	1.057031435	1.080778527	1.065883459
Run 4	1.136163341	1.058535403	1.094519607	1.054648466
Run 5	1.142646404	1.053712015	1.096016936	1.051657653
Run 6	1.127815362	1.050240898	1.092535681	1.032816127
Run 7	1.126496397	1.060984552	1.077684663	1.029308439
Run 8	1.126726607	1.071417589	1.105424474	1.016332609
Run 9	1.129528240	1.056071182	1.110856699	1.025219164
Run 10	1.127446466	1.065449207	1.087512576	1.018451182
Run 11	1.144288113	1.070877878	1.115955198	1.025206357
Run 12	1.113201025	1.053787976	1.111222484	1.005639829
Run 13	1.124159991	1.060504070	1.137141026	1.005017247
Run 14	1.099577064	1.049613885	1.134723496	1.015575719
Run 15	1.107663913	1.054724029	1.131947221	1.030437865
Run 16	1.102834794	1.065399738	1.123383946	1.037202344
Run 17	1.088510266	1.054720606	1.115794779	1.015382372
Run 18	1.104600911	1.046057788	1.104277158	1.040055283
Run 19	1.104238090	1.051978441	1.090225413	1.041292774
Run 20	1.094128419	1.043923016	1.109540107	1.021515599

Statistical Methodology of Tissue Samples

Linear discriminant analysis (LDA) was used to evaluate the four tissues at the three different emission intensity ratios (410 nm, 510 nm and 541 nm). The application of LDA proceeds in a series of steps. First basic descriptive statistics, including F-tests and Wilks' Lambda statistics, are used to assess whether statistically significant joint mean differences exist across each of the three emission intensity ratios. Higher F-test values and lower Wilks' Lambda values indicate that significant joint men differences do exist across the three emission intensity ratios, and by extension imply that the data can be appropriately analyzed using LDA. Next, the eigenvalues of the data matrix are extracted and used to determine the number of significant (but latent) underlying factors that drive the statistical relationships across the four tissues. Eigenvalues that explain a statistically significant percentage of the variation in the data (as indicated by chi-square tests) are retained for further analysis, while insignificant eigenvalues are discarded. The significant eigenvalues are used to estimate standardized canonical discriminant functions, which parameterize the (linear) relationship between the emission intensity ratios and the eigenvalues. Canonical standardized discriminant function coefficients that are larger in absolute value indicate that the corresponding emission intensity ratio aligns more closely with that eigenvalue. The most appropriate emission intensity ratio typically exhibits large standardized discriminant function coefficients for the first (and largest) eigenvalue, which explains the majority of the variation across the four sets of tissues. These discriminant functions can also be used to generate a plot depicting how the LDA model groups tissues across the two most important (or largest) canonical standardized discriminant functions. If the LDA model is valid, it should

produce a graph in which each of the tissue observations are distinctly identified and tightly clustered around the group means (or centroids). As a second best solution, one would expect the MDA-MB-231 tissue to form a group which is distinct from the other (breast, kidney and liver) healthy tissue groups. Another measure of the LDA model's fit is to examine the percentage of observations in the data set to which the LDA model correctly predicts tissue membership. If the LDA model expresses higher internal validity, it should correctly predict tissue membership for a very high percentage of the observations in the data set. To ensure an appropriate estimate of predicted tissue membership, we used cross-validation (leave-one-out) methods to generate predicted values. Finally, to determine which emission intensity ratio is the "best" predictor of the four tissues, we calculated potency indices. Emission intensity ratios with higher potency indices indicate that the emission intensity ratio in question contributes more to the formation of the primary eigenvalues in the data set (relative to the other emission intensities being analyzed) and relatively less to those secondary eigenvalues that do not explain as much of the variation across the four tissues. Thus, we identify the emission intensity ratio with the highest potency index as the "best" predictor of the tissues. As noted earlier in the manuscript, there are four tissues (breast, liver, kidney and MDA-MB-231), each of which was evaluated at three emission intensities. Each tissue-intensity pair was replicated a total of ten times. This provides a working sample of 80 observations (4 cell lines by 10 replications) and 3 variables (emission intensity ratios).

Table 21 contains the mean values, F-statistics and Wilks' Lambda values for each of the tissue. All F-statistics are statistically significant at the 5 percent level, indicating that significant joint differences exist across the tissues for each emission intensity ratio. Wilks' Lambda values for all three emission intensities are also relatively small in magnitude, indicating that the data are amenable to analysis by LDA. It is interesting to note that the 410 nm and 541 nm emission intensity ratios exhibit the lowest Wilks' Lambda values, indicating that they are the most amenable to LDA. For the 410 nm ratio, the cancer tissue exhibits the highest mean emission intensity ratio of the four tissues. For the 541 nm ratio, the cancer tissue exhibits the lowest mean value.

Table 21 Descriptive Analysis of Group Means.

	Emission Intensity Ratio		
Cell Line	**410**	**510**	**541**
Breast Cell Line	0.470	1.118	1.122
Kidney Cell Line	0.612	1.084	1.061
Liver Cell Line	0.652	1.126	1.104
MDA-MB-231 Cell Line	0.877	1.055	1.033
Wilks' Lambda	0.225	0.400	0.213
F-Statistic [3, 76]	87.142	38.066	93.772
P-Value	<0.001	<0.001	<0.001

Table 22 contains a summary of the eigenvalues and canonical correlations extracted by LDA. Three eigenvalues were extracted, each of which explains a significant percentage of the variation in the four tissues. The first eigenvalue explains 89.4 percent of the variation, while the remaining eigenvalues explain 6.7 percent and 3.8 percent, respectively. Thus, while all three eigenvalues express distinct statistical information, the first eigenvalue is the primary eigenvalue of interest.

Table 22 Canonical Function Summary

Function	Eigenvalue	Pct. Variance Explained	Canonical Correlation	Wilks' Lambda	Chi-Square Statistic	P-Value
1	8.259	89.4	0.944	0.049	227.410	<0.001
2	0.623	6.7	0.619	0.455	59.380	<0.001
3	0.353	3.8	0.511	0.739	22.831	<0.001

Table 23 contains the coefficients which determine the three canonical standardized discriminant functions. The 541 nm emission intensity ratio contains the largest coefficient for the first canonical function (0.997), which corresponds to the first eigenvalue. This emission intensity also exhibits the smallest coefficient in absolute value for the third (and least important) canonical function. Concomitantly, the 510 nm emission intensity ratio exhibits the smallest coefficient in absolute value for the first function, and the largest coefficients (in absolute magnitude) for the second and third functions. The 410 nm intensity ratio exhibits moderately sized coefficient values (in absolute value) for the first and second functions, and a relatively small coefficient in absolute value for the third function. Overall, this implies that the 541 nm and 410 nm ratios contribute more to the formation of the first canonical function. The 510 nm ratio contributes more to the formation of the second and third functions.

Table 23 Standardized Discriminant Function Coefficients

Emission Intensity	Function 1	Function 2	Function 3
410	-0.727	0.713	0.314
510	-0.249	-1.665	1.153
541	0.997	1.661	-0.380

Figure 4 (main text) shows the canonical function chart for the first two primary canonical functions. Examining this chart, we see that the MDA-MB-231 cancer tissue observations (group 4, in purple) are clearly distinguished from the other three (non-cancerous) tissue. However, while it is possible to see the groups of the healthy cell lines as distinct groups, the here healthy cell line groups do overlap. As a result, one can conclude that the LDA model does

an acceptable job of distinguishing between cancerous and non-cancerous tissues, but does not fully distinguish between healthy tissues.

Table 24 includes the potency indices (with corresponding structure matrix values) used to identify the "best" emission intensity ratio. Clearly, the 541 nm emission intensity ratio exhibits the highest potency index value, and is the wavelength of choice. The 410 nm emission intensity ratio exhibits a potency index which is only slightly smaller in magnitude than the potency index for the 541 nm. The potency index for the 510 nm ratio is substantially smaller in magnitude than the other two emission intensity ratios, and can be considered as the ratio that provides the smallest amount of discrimination power across the three groups.

Table 24 Structure Matrix and Potency Index

Emission Intensity	Function 1	Function 2	Function 3	Potency Index
410	-0.618	0.508	0.600	0.373
510	0.381	-0.018	0.924	0.162
541	0.647	0.366	0.669	0.401

A final question of interest is the model's internal validity. Cross-validation methods indicate that the LDA model predicts 91.2 percent of the observations correctly. Since this number is relatively close to 100 percent, one can conclude that the model exhibits a relatively high degree of internal validity. Additionally, all of the cancerous tissue observations were correctly predicted by the LDA model. Given that the cancerous tissues in Figure 1 are depicted as a distinct group, while some overlap exists in Figure 1 across the health tissues, it is not surprising that the model would inaccurately predict a small number of healthy tissue observations, especially those observations that are located near areas where the groups overlap.

References:

[19] R. Dutta, M. D. Scott, M. K. Haldar, B. Ganguly, D. K. Srivastava, D. L. Friesner, S. Mallik, *Bioorg Med Chem Lett* **2011**, *21*, 2007-2010.

[20] H. D. Soule, J. Vazguez, A. Long, S. Albert, M. Brennan, *J Natl Cancer Inst* **1973**, *51*, 1409-1416.

[21] S. R. Aspinall, S. Stamp, A. Davison, B. K. Shenton, T. W. Lennard, *J Steroid Biochem Mol Biol* **2004**, *88*, 37-51.

[22] L. Hui, Y. Zheng, Y. Yan, J. Bargonetti, D. A. Foster, *Oncogene* **2006**, *25*, 7305-7310.

[23] J. Chen, L. Lu, Y. Feng, H. Wang, L. Dai, Y. Li, P. Zhang, *Cancer Lett* **2011**, *300*, 48-56.

[24] J. R. Masters, *Nat Rev Cancer* **2002**, *2*, 315-319.

[25] R. Rahbari, T. Sheahan, V. Modes, P. Collier, C. Macfarlane, R. M. Badge, *Biotechniques* **2009**, *46*, 277-284.
[26] F. L. Graham, J. Smiley, W. C. Russell, R. Nairn, *J Gen Virol* **1977**, *36*, 59-74.
[27] M. T. Malik, S. S. Kakar, *Mol Cancer* **2006**, *5*, 61.
[28] aO. R. Miranda, B. Creran, V. M. Rotello, *Curr Opin Chem Biol* **2010**, *14*, 728-736; bU. H. Bunz, V. M. Rotello, *Angew Chem Int Ed Engl* **2010**, *49*, 3268-3279; cA. Bajaj, O. R. Miranda, R. Phillips, I. B. Kim, D. J. Jerry, U. H. Bunz, V. M. Rotello, *J Am Chem Soc* **2010**, *132*, 1018-1022; dE. K. Nyren-Erickson, M. K. Haldar, Y. Gu, S. Y. Qian, D. L. Friesner, S. Mallik, *Anal Chem* **2011**, *83*, 5989-5995.
[29] aR. A. Johnson, D. W. Wichern, *Applied multivariate statistical analysis*, 5th ed., Prentice Hall, Upper Saddle River, N.J., **2002**; bJ. F. Hair, *Multivariate data analysis*, Prentice Hall, Upper Saddle River, N.J., **1998**.