2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3503-7149-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMBC53108.2024.10782594

Exploring Biomarker Relationships in Both Type 1 and Type 2 Diabetes Mellitus Through a Bayesian Network Analysis Approach

Yuyang Sun
Department of Engineering
King's College London
London, United Kingdom
yuyang.1.sun@kcl.ac.uk

Jingyu Lei
Division of Surgery and Interventional Science
University College London
London, United Kingdom
jingyu.lei.22@ucl.ac.uk

Panagiotis Kosmas
Department of Engineering
King's College London
London, United Kingdom
panagiotis.kosmas@kcl.ac.uk

***Abstract*—Understanding the complex relationships of biomarkers in diabetes is pivotal for advancing treatment strategies, a pressing need in diabetes research. This study applies Bayesian network structure learning to analyze the Shanghai Type 1 and Type 2 diabetes mellitus datasets, revealing complex relationships among key diabetes-related biomarkers. The constructed Bayesian network presented notable predictive accuracy, particularly for Type 2 diabetes mellitus, with root mean squared error (RMSE) of 18.23 mg/dL, as validated through leave-one-domain experiments and Clarke error grid analysis. This study not only elucidates the intricate dynamics of diabetes through a deeper understanding of biomarker interplay but also underscores the significant potential of integrating data-driven and knowledge-driven methodologies in the realm of personalized diabetes management. Such an approach paves the way for more custom and effective treatment strategies, marking a notable advancement in the field.**

***Index Terms*—Bayesian network, Structure learning, Glucose prediction, Diabetes management.**

I. Introduction

Diabetes mellitus (DM), a chronic metabolic disorder, has emerged as a global health crisis, affecting millions and escalating rapidly in prevalence and presenting significant challenges in diagnosis and management [1]. In addressing these challenges, continuous glucose monitoring (CGM) has been a pivotal development, offering real-time glucose data critical for effective DM management. While various methods exist for predicting glucose levels based on previous glucose trajectories [2]–[5], the analysis of diabetes-related biomarkers and their impact on glucose levels remains an area less explored, with many interrelationships yet to be fully understood [6]–[8]. Addressing this issue can therefore have a positive impact on the accuracy of prediction systems.

Recent advancements in machine learning, particularly in the realm of Bayesian networks (BNs) [9], [10], present novel avenues to unravel the complex interplay between diabetes-related biomarkers and glucose measurements. BNs, known for their proficiency in handling uncertainties and probabilistic relationships, are suitable for modeling the complex interactions inherent in diabetes-related data. By building BNs on Type 1 and Type 2 DMs (T1DM and T2DM) datasets, we analyze the relationships between these biomarkers and glucose levels while considering the complexity and interdependencies of biomarkers.

More specifically, our study leverages publicly available diabetes datasets [11], applying Bayesian network structure learning [12], [13] to conduct a comprehensive analysis of key diabetes-related characteristics, including glycated hemoglobin (HbA1c), glycated albumin (GA), estimated glomerular filtration rate (eGFR), creatinine (CR), etc., and glucose measurements such as FPG and 2HPP. By categorizing the identified arcs as causal, correlated, or independent, the paper also uncovers relationships between these characteristics which can be both data and knowledge-driven, as discussed in our Results section.

II. Methodology

This section outlines the methodology employed in our investigation of the interrelationships among diabetes-related biomarkers using BNs. Our approach encompasses the utilization of publicly available datasets [11] and sophisticated structure learning techniques to elucidate complex dependencies inherent in diabetes data. This methodology not only leverages advanced machine learning techniques but also tailors them specifically to address the unique challenges posed by the multifaceted nature of diabetes-related data.

A. Shanghai Diabetes Mellitus Datasets

Our study leverages the publicly available Shanghai DM datasets [11], which include data on T1DM and T2DM. The ShanghaiT1DM dataset contains records from 12 T1DM patients, and the ShanghaiT2DM dataset includes data from 100 T2DM patients. These datasets record valuable anthropometric and biochemical characteristics alongside glucose measurements, forming the basis of our analysis.

The key features of these datasets include anthropometric characteristics such as age, weight, height, body mass index (BMI), and gender, and biochemical characteristics including HbA1c, GA, total cholesterol (TC), triglycerides (TG), high-density lipoprotein (HDL), low-density lipoprotein (LDL), CR, eGFR, uric acid (UA), and blood urea nitrogen (BUN). Additionally, glucose measurements, such as Fasting Plasma Glucose (FPG) and 2-Hour Postprandial Glucose (2HPP), are incorporated. These characteristics are categorized into five classes, as shown in Table I.

TABLE I
CATEGORIZATION OF DIABETES-RELATED CHARACTERISTICS IN THE SHANGHAI DM DATASETS

Categories	Characteristics
Glycemic biomarkers	HbA1c, GA
Anthropometric biomarkers	Age, Weight, Height, BMI, Gender
Lipid biomarkers	TC, TG, HDL, LDL
Kidney biomarkers	CR, eGFR, UA, BUN
Glucose measurements	FPG, 2HPP

Detailed data specifications and additional features, including medical histories and complications, are available in the original dataset documentation [11]. In our analysis, we focus on characteristics that directly influence blood glucose fluctuations, such as factors related to glucose metabolism, lipid profiling, and kidney function. For data preprocessing, missing values were added using averages from other individual data when less than 20% of data was missing for a given characteristic. Features with more than 20% missing data, such as UA and BUN, were excluded from the Bayesian network structure learning process to maintain data integrity.

B. *Bayesian Network Structure Learning*

Our analysis utilizes BNs, probabilistic graphical models denoted as $\mathcal{B}$ and defined by a tuple $(G, \mathbf{\Theta})$. Here, G represents a directed acyclic graph illustrating dependencies among random variables, while $\mathbf{\Theta}$ encompasses parameters that define the strength of these dependencies. BNs are particularly valued in biomedical research for their interpretability, revealing complex dependencies that are often not straightforwardly causal, especially when latent variables are involved [9], [10], [12], [14].

In this study, we focused on structure learning of BNs to elucidate relationships among diabetes-related characteristics we mentioned before. Specifically, we employed the Tabu search algorithm [15], complemented by Bootstrap resampling [16], [17] to generate two BNs respectively from the ShanghaiT1DM (12 samples) and ShanghaiT2DM (100 samples) datasets. The strength of arcs in generated BNs enables the identification of reliable dependencies by assessing the frequency of arc occurrence, with higher frequency indicating stronger dependencies. The robustness of these BNs was assessed by evaluating the frequency of arc occurrence in 100 Bootstrap iterations, with a strength threshold of 0.85 for arc retention. The resulting 'Biomarkers_Network' on the ShanghaiT2DM dataset combines arcs common to both DM models, arcs exceeding the strength threshold on T1DM, and some artificially added arcs for comprehensive analysis.

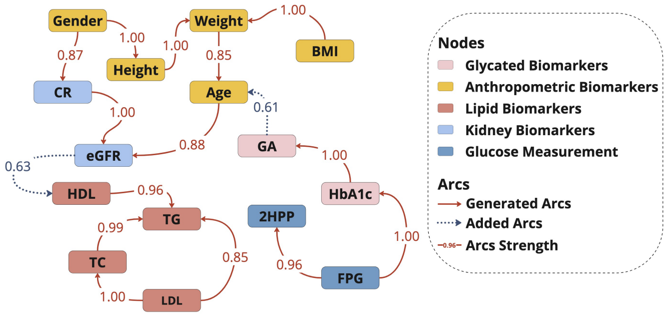

Fig. 1. Bayesian Network Structure for Diabetes Biomarker Analysis on both ShanghaiT1DM and ShanghaiT2DM Datasets.

'Biomarkers_Network', as depicted in Figure 1, visualizes these dependencies through its structure of arcs and nodes. The nodes are categorized into four biomarker types and one measurement category, as tabulated in Table I). The arcs indicate statistical correlations between connected variables. The 'arc strength', depicted at the arcs, quantifies the probability and directionality of these dependencies.

Notably, the main differences in the generated BNs between T1DM and T2DM are mainly due to the sample size., with T2DM's larger dataset revealing more potential arcs. Some arcs, while not meeting the high strength threshold, were artificially introduced in 'Biomarkers_Network' to ensure a full connection of all available nodes to maximize the consideration of all variables for our dependencies analysis and prediction experiments. Further discussion on correlation and causality analysis is presented in the subsequent discussion section.

Performance evaluation of 'Biomarkers_Network' was conducted through leave-one-domain experiments, setting aside one individual's data for testing while using the remaining individuals' data for network training. The prediction accuracy for FPG and 2HPP values was evaluated using mean absolute error (MAE) and root mean squared error (RMSE), evaluated at each individual, and then calculated as the averages. Additionally, a Clarke error grid [18] is introduced to visualize the predicted results of T1DM and T2DM.

III. RESULTS AND DISCUSSION

A. *Prediction Results of Leave-One-Domain Experiment*

The performance of 'Biomarkers_Network' in predicting glucose levels was evaluated using leave-one-domain experiments. These experiments were conducted separately for T1DM and T2DM datasets to predict FPG and 2HPP values, employing MAE and RMSE as evaluation metrics. The raw FPG levels for T1DM range from 117.00 to 262.35 mg/dL, whereas for T2DM they span from 126.00 to 194.40 mg/dL. Regarding the raw 2HPP levels, T1DM exhibits a range of 248.76 to 348.84 mg/dL, compared to 196.16 to 317.88 mg/dL for T2DM. The predicted results, as detailed in Table II, demonstrate the network's capability in predicting glucose levels, with a notably better performance observed in the T2DM dataset compared to T1DM. Notably, the superior performance in the T2DM dataset is attributed to its larger size and the complexity of interactions it captures. The results highlight the potential of Bayesian networks in analyzing and predicting key diabetes metrics, laying the groundwork for further research and application in diabetes management.

TABLE II
LEAVE-ONE-DOMAIN EXPERIMENT RESULTS (MEAN) OF 'BIOMARKERS_NETWORK'.

Datasets	Label	MAE (mg/dL)	RMSE (mg/dL)
T1DM	FPG	30.29	36.16
	2HPP	31.94	41.59
T2DM	FPG	19.22	28.23
	2HPP	28.99	40.12

B. Clarke Error Grid visualization

For visualization, Clarke error grid analysis [18], illustrated in Figure 2, is introduced to provide insights into the model's predictive accuracy. The grid categorizes results into zones reflecting clinical impact: Zone A (clinically acceptable), Zone B (benign errors), and Zones C to E (errors with potential clinical significance). The T1DM predictions are predominantly within Zones A and B, denoting acceptable accuracy. 2HPP predictions show a slightly higher dispersion, suggesting increased variability in postprandial glucose responses. Conversely, T2DM results are more concentrated in Zone A for both FPG and 2HPP, denoting enhanced reliability and clinical applicability. This visual analysis corroborates the numerical findings, where T2DM showed lower MAE and RMSE values compared to T1DM, and underscores the model's clinical viability for diabetes management.

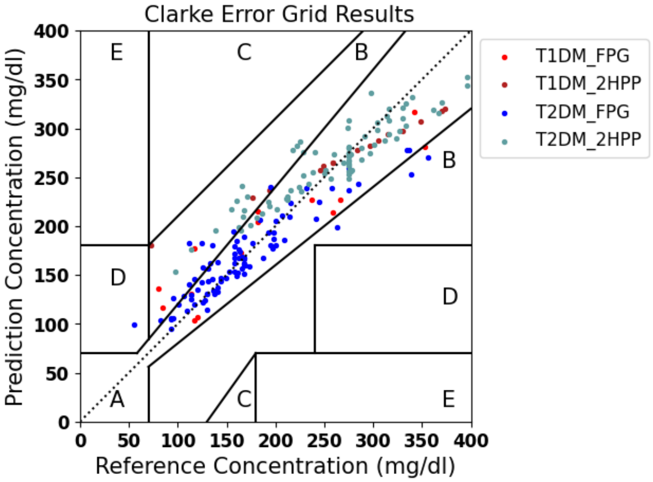

Fig. 2. Clarke Error Grid Analysis for T1DM and T2DM Predictions. FPG and 2HPP predictions for T1DM (red and brown) and T2DM (blue and cyan) are displayed across the grid.

C. Analysis of Relationships in the Bayesian Network

The structure of the Bayesian network, derived through a data-driven approach, reveals intricate relationships among diabetes-related biomarkers. This subsection offers a knowledge-driven approach, utilizing expert knowledge to interpret and validate these relationships.

1) Glucose Measurement and Glycated Metrics Relationships:

- **'FPG' to 'HbA1c' Arc:** The strong arc strength (1.0) observed between FPG and HbA1c in our network aligns with established medical knowledge. FPG is a direct measure of blood glucose after fasting, while HbA1c offers a long-term glycemic index. While high FPG levels over time contribute to increased HbA1c, the relationship isn't strictly causal, considering HbA1c's sensitivity to various other factors beyond fasting glucose levels. This **correlation** is well-supported by several studies [19], [20], highlighting the interdependence of these metrics in diabetes management. This arc in our network reflects this well-established interdependence, crucial for diabetes diagnosis and management.
- **'HbA1c' to 'GA' Arc:** Another notable arc is between HbA1c and GA, with strength values of 0.96 and 1.0 across our models. The positive **correlation** between these two metrics is reinforced by existing literature [21], [22]. GA, unlike HbA1c, offers a shorter-term view of glucose control, making it especially useful in cases where HbA1c results might be unreliable. However, establishing a causal link is challenging due to the distinct periods these metrics reflect.
- **'FPG' to '2HPP' Arc:** The difference in the arc strength between T1DM (0.35) and T2DM (0.96) networks highlights the distinct pathophysiological profiles of these conditions. T2DM's higher **correlation** suggests a stronger link between fasting and postprandial glucose levels, a phenomenon well-documented in diabetes research [23]. This differential relationship underscores the need for distinct management strategies for T1DM and T2DM.

2) Anthropometric Metrics Relationships:

- **'Gender' to 'Height' Arc:** The relationship between gender and height in our network is a clear **causal** example of genetic and hormonal influences on physical characteristics, extensively corroborated by global health data [24]. The presence of this arc is a validation of the network's ability to capture existing fundamental biological relationships by a data-driven approach.
- **Arcs among 'Height', 'Weight', and 'BMI':** The **correlations** among these anthropometric measures are well-established in physiological research [25]. BMI, calculated using height and weight, serves as a key health indicator. The strength of these arcs, while significant, varies due to the dataset sizes, underscoring the importance of considering data variability in such analyses.

3) Lipid Metrics Relationships: These lipid metrics are all critical in assessing cardiovascular risk in diabetes patients. The relationships are categorized as **correlations** due to the complex interplay of metabolic and biochemical factors influencing these metrics [26]–[28]. Notably, the Friedewald Equation [26] demonstrates these dependencies, though pinpointing direct causal links remains a challenge.

4) Kidney Metrics Relationships: The network underscores the relationship between serum CR levels and eGFR, as well as the influence of age on eGFR. These **correlations** align with nephrology research [29]–[33], which emphasizes the importance of these biomarkers in assessing kidney health in diabetes patients. Besides, the MDRD equation [29], [30] provides a quantitative framework, illustrating how CR levels and age are crucial in estimating eGFR, with gender and ethnicity as additional factors.

D. Limitations and Future Directions

Our Bayesian network has provided valuable insights into the relationships among diabetes-related biomarkers. However, it's important to note that deducing causality from such models is inherently challenging. It often necessitates access to more extensive datasets and more refined modeling techniques capable of handling complex biological interactions. Future research directions include expanding the datasets in both size and diversity to enhance model robustness and incorporating patient data from real-world settings. This expansion would allow for the exploration of dynamic models that better capture the temporal fluctuations of diabetes biomarkers. Additionally, the application of these analytical insights to develop decision-support tools or to tailor personalized diabetes management strategies holds great promise. Such advancements have the potential to transform patient care and improve clinical outcomes in diabetes management.

IV. Conclusion

In this study, we applied Bayesian network structure learning to T1DM and T2DM datasets, uncovering the intricate relationships among various diabetes-related biomarkers. This analysis has deepened our understanding of the complex interplay of factors involved in diabetes, thereby enriching the broader knowledge base in this area. The notable predictive accuracy observed, especially within the T2DM dataset, underscores the practical utility of our methodology in diabetes management. Our future work will explore further this methodology of combining knowledge and data-driven approaches in biomedical research.

References

[1] Cho, N., Shaw, J., Karuranga, S., Huang, Y., Rocha Fernandes, J., Ohlrogge, A. & Malanda, B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research And Clinical Practice*. **138** pp. 271-281 (2018)

[2] Li, K., Liu, C., Zhu, T., Herrero, P. & Georgiou, P. GluNet: A deep learning framework for accurate glucose forecasting. *IEEE Journal Of Biomedical And Health Informatics*. **24**, 414-423 (2019)

[3] Mirshekarian, S., Bunescu, R., Marling, C. & Schwartz, F. Using LSTMs to learn physiological models of blood glucose behavior. *2017 39th Annual International Conference Of The IEEE Engineering In Medicine And Biology Society (EMBC)*. pp. 2887-2891 (2017)

[4] Xie, J. & Wang, Q. Benchmarking machine learning algorithms on blood glucose prediction for type I diabetes in comparison with classical time-series models. *IEEE Transactions On Biomedical Engineering*. **67**, 3101-3124 (2020)

[5] Turksoy, K., Samadi, S., Feng, J., Littlejohn, E., Quinn, L. & Cinar, A. Meal detection in patients with type 1 diabetes: a new module for the multivariable adaptive artificial pancreas control system. *IEEE Journal Of Biomedical And Health Informatics*. **20**, 47-54 (2015)

[6] Georga, E., Protopappas, V., Ardigo, D., Marina, M., Zavaroni, I., Polyzos, D. & Fotiadis, D. Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression. *IEEE Journal Of Biomedical And Health Informatics*. **17**, 71-81 (2012)

[7] Katsarou, D., Georga, E., Christou, M., Tigas, S., Papaloukas, C. & Fotiadis, D. Short Term Glucose Prediction in Patients with Type 1 Diabetes Mellitus. *2022 44th Annual International Conference Of The IEEE Engineering In Medicine & Biology Society (EMBC)*. pp. 329-332 (2022)

[8] Sun, Y., Cano-Garcia, H., Kallos, E., O'Brien, F., Akintonde, A., Motei, D., Ancu, O., Mackenzie, R. & Kosmas, P. Random Forest Analysis of Combined Millimeter-wave and Near-infrared Sensing for Non-invasive Glucose Detection. *IEEE Sensors Journal*. (2023)

[9] Pearl, J. Probabilistic reasoning in intelligent systems: networks of plausible inference. (Morgan kaufmann,1988)

[10] Pearl, J. Bayesian networks: A model of self-activated memory for evidential reasoning. *Proceedings Of The 7th Conference Of The Cognitive Science Society, University Of California, Irvine, CA, USA*. pp. 15-17 (1985)

[11] Zhao, Q., Zhu, J., Shen, X., Lin, C., Zhang, Y., Liang, Y., Cao, B., Li, J., Liu, X., Rao, W. & Others Chinese diabetes datasets for data-driven machine learning. *Scientific Data*. **10**, 35 (2023)

[12] Kitson, N., Constantinou, A., Guo, Z., Liu, Y. & Chobtham, K. A survey of Bayesian Network structure learning. *Artificial Intelligence Review*. pp. 1-94 (2023)

[13] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. & Nolan, G. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*. **308**, 523-529 (2005)

[14] Scanagatta, M., Salmerón, A. & Stella, F. A survey on Bayesian network structure learning from data. *Progress In Artificial Intelligence*. **8** pp. 425-439 (2019)

[15] Glover, F. Future paths for integer programming and links to artificial intelligence. *Computers & Operations Research*. **13**, 533-549 (1986)

[16] Breiman, L. Bagging predictors. *Machine Learning*. **24**, 123-140 (1996)

[17] Friedman, N., Goldszmidt, M. & Wyner, A. Data analysis with Bayesian networks: A bootstrap approach. *ArXiv Preprint ArXiv:1301.6695. (2013)*

[18] Clarke, W., Cox, D., Gonder-Frederick, L., Carter, W. & Pohl, S. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care*. **10**, 622-628 (1987)

[19] Nathan, D., Kuenen, J., Borg, R., Zheng, H., Schoenfeld, D., Heine, R. & Group, A. Translating the A1C assay into estimated average glucose values. *Diabetes Care*. **31**, 1473-1478 (2008)

[20] Sherwani, S., Khan, H., Ekhzaimy, A., Masood, A. & Sakharkar, M. Significance of HbA1c test in diagnosis and prognosis of diabetic patients. *Biomarker Insights*. **11** pp. BMI-S38440 (2016)

[21] Peacock, T., Shihabi, Z., Bleyer, A., Dolbare, E., Byers, J., Knovich, M., Calles-Escandon, J., Russell, G. & Freedman, B. Comparison of glycated albumin and hemoglobin A1c levels in diabetic subjects on hemodialysis. *Kidney International*. **73**, 1062-1068 (2008)

[22] Yazdanpanah, S., Rabiee, M., Tahriri, M., Abdolrahim, M., Rajab, A., Jazayeri, H. & Tayebi, L. Evaluation of glycated albumin (GA) and GA/HbA1c ratio for diagnosis of diabetes and glycemic control: A comprehensive review. *Critical Reviews In Clinical Laboratory Sciences*. **54**, 219-232 (2017)

[23] Association, A. Diagnosis and classification of diabetes mellitus. *Diabetes Care*. **33**, S62-S69 (2010)

[24] Rodriguez-Martinez, A., Zhou, B., Sophiea, M., Bentham, J., Paciorek, C., Iurilli, M., Carrillo-Larco, R., Bennett, J., Di Cesare, M., Taddei, C. & Others Height and body-mass index trajectories of school-aged children and adolescents from 1985 to 2019 in 200 countries and territories: a pooled analysis of 2181 population-based studies with 65 million participants. *The Lancet*. **396**, 1511-1524 (2020)

[25] Keys, A., Fidanza, F., Karvonen, M., Kimura, N. & Taylor, H. Indices of relative weight and obesity. *Journal Of Chronic Diseases*. **25**, 329-343 (1972)

[26] Friedewald, W., Levy, R. & Fredrickson, D. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical Chemistry*. **18**, 499-502 (1972)

[27] Sosenko, J., Breslow, J., Miettinen, O. & Gabbay, K. Hyperglycemia and plasma lipid levels: a prospective study of young insulin-dependent diabetic patients. *New England Journal Of Medicine*. **302**, 650-654 (1980)

[28] Walden, C., Knopp, R., Wahl, P., Beach, K. & Strandness Jr, E. Sex differences in the effect of diabetes mellitus on lipoprotein triglyceride and cholesterol concentrations. *New England Journal Of Medicine*. **311**, 953-959 (1984)

[29] Levey, A., Coresh, J., Bolton, K., Culleton, B., Harvey, K., Ikizler, T., Johnson, C., Kausz, A., Kimmel, P., Kusek, J. & Others K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *American Journal Of Kidney Diseases*. **39**, i-ii+ (2002)

[30] Levey, A., Bosch, J., Lewis, J., Greene, T., Rogers, N., Roth, D. & Renal Disease Study Group*, M. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. *Annals Of Internal Medicine*. **130**, 461-470 (1999)

[31] Stevens, P., Levin, A. & Members*, K. Evaluation and management of chronic kidney disease: synopsis of the kidney disease: improving global outcomes 2012 clinical practice guideline. *Annals Of Internal Medicine*. **158**, 825-830 (2013)

[32] O'Hare, A., Choi, A., Bertenthal, D., Bacchetti, P., Garg, A., Kaufman, J., Walter, L., Mehta, K., Steinman, M., Allon, M. & Others Age affects outcomes in chronic kidney disease. *Journal Of The American Society Of Nephrology*. **18**, 2758-2765 (2007)

[33] Delanghe, J., De Slypere, J., De Buyzere, M., Robbrecht, J., Wieme, R. & Vermeulen, A. Normal reference values for creatine, creatinine, and carnitine are lower in vegetarians.. *Clinical Chemistry*. **35**, 1802-1803 (1989)