HHS Public Access

Author manuscript

Acc Chem Res. Author manuscript; available in PMC 2018 March 21.

Published in final edited form as:

Acc Chem Res. 2017 March 21; 50(3): 647–651. doi:10.1021/acs.accounts.7b00009.

The “Ok Molly” Chemistry

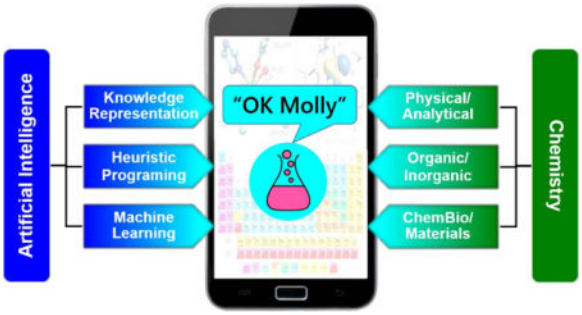
Yi Lu

Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Abstract

One day in the near future, Molly, an artificial intelligence assistant, can answer almost any questions in chemistry and related disciplines. She will have a major impact in the way we perform research, education and public outreach.

TOC image



1. Introduction

“Ok Google” has become a common phrase used by people all over the world to ask questions, such as “why is the sky blue?” and “how to design a treehouse?” Similar voice-activated assistants, such as Siri, Alexa, and Cortana, have also been developed and are changing our everyday lives. Behind these assistants are powerful search engines combined with the most sophisticated artificial intelligence (AI) algorithms developed to date, which include knowledge representation, pattern/image recognition, machine learning, automated reasoning and prediction, in addition to modern voice recognition capabilities. Given the demonstrated potential and initial success of these AI assistants, a conceivable Holy Grail for chemistry is an AI assistant for chemists, i.e., some day in the future, we will be able to wake up in the morning and find answers to any question in the chemical science and engineering fields, such as “why is my catalyst so much better (or worse) than I initially designed?”, “what’s in my materials?”, “how to design a compound that can bind to this target or catalyze this reaction?”, “how to synthesize this compound?”, and even “how to cure cancer?” To differentiate this powerful assistant from the ones mentioned above, let’s call her “Molly”, which is a cute nickname for “Molecules”, a foundation for almost all chemical sciences and related scientific and engineering fields.

Corresponding Author: yi-lu@illinois.edu.

Molly differs from Google in more than the name: currently, “Ok Google” is very good at searching and presenting facts available on the web. Many chemists have already been taking advantage of this capability to search through literature and find the most relevant publications. The Holy Grail for chemists is for Molly to be able to “think” like us; e.g., to be able to analyze information from publications and propose novel chemical strategies for a given problem. However, since solutions to most chemistry problems are not numerical, it is not easy to come up with the best ideas using pure numerical computation. To accomplish these goals, Molly needs much more sophisticated AI, such as heuristic programming and machine learning that includes neural network and deep learning. For example, when chemists want to tackle a problem in initiating a project, we 1) search the literature to find the most relevant publications; 2) read the literature and summarize the state-of-the-art in the field; 3) analyze the literature and combine with personal knowledge and experience to come up with novel ideas to solve the problem; 4) design methods or approaches to implement the ideas; 5) carry out experiments and collect data; 6) analyze the data to assess the success or failure of the ideas; and 7) summarize lessons learned from the project to enrich our knowledge, publish the results and thus contributing to the literature used in Step 1. Limitations or mistakes in any of these steps can derail the project, such as lack of comprehensiveness in the literature search, reading and summary in Steps 1 and 2, or lack of knowledge and experience or personal bias in Steps 3 to 7. Right now, “Ok Google” is very good at Step 1, is beginning to improve at Step 2, but is largely incapable of Steps 3 to 7. To overcome this limitation, we often hold meetings among researchers in the field and recruit experts outside our area of expertise as collaborators. However, geographic restrictions and time limitations mean that we often miss the opportunities to seek advice from or work with the most knowledgeable and experienced collaborators and miss many of the options to solving the problems in science. Molly can help overcome this limitation by searching for and seeking input from chemists and collaborators all over the world. More importantly, through machine learning, Molly can come up with ideas that combine the best intellectual inputs of all chemists and their collaborators without geographic restrictions or personal bias. Furthermore, since chemistry is fundamentally an experimental science, Molly is able to take advantage of powerful chemical instrumentation, including smartphones, or accessories and sensors that can transmit the results to a smartphone, to collect data in order to provide much more in-depth answers than “Ok Google”.

2. “Ok Molly” Chemistry: past achievements, present practices, and future challenges

Molly has a strong foundation built from many years of research in applying AI to chemistry,[1–4] with the DENDRAL project (originally stood for DENDRitic Algorithm)[1] as one prime example. Despite progress made by many pioneers in the field, applications of AI to the frontiers of chemical research are not common in today’s academic and industrial labs. Now is an ideal time to pursue the “Ok Molly” project, because recent advances in data bases, cloud storage and computing, modern heuristics, pruning and deep learning algorithms, and computing power (e.g., the Blue Waters Supercomputer capable of 13 quadrillion calculations per second) have made it possible for computers such as IBM’s Deep Blue to defeat a world chess grandmaster in 1997 and IBM’s Watson to beat the

world's best Jeopardy champions in 2011. Early this year, Google's AlphaGo completed a 4-1 series victory over a world champion of Go by being able to teach itself. Self-driving cars that can train themselves based on feedbacks from camera, radar and sonar to navigate roadways are being tested in cities around the world. Although impressive progress has been made in areas *outside* of chemistry, we need much better integration between AI and chemistry in order to make Molly a reality.

While there are many potential applications of Molly, some are easier to implement than others and can be adopted into the Molly format in the near future, while many others are much more difficult and require further research and development. The first area is likely in theoretical explanation of chemical phenomena using computation, as most of the solutions have already been written into numerical forms, such as computer programs, that can be handled by the powerful computers behind all AI assistants. A primary example is the successful development of density functional theory and the wide application of computer programs implementing the theory. The key to the success of Molly is to make such programs accessible to non-specialists outside theoretical and computational chemistry, including the general public so as to expand the outreach of chemistry to society (see Section 3 below), by setting up servers in the cloud and making them more available. Critical to Molly's success are proper parameterization of chemical phenomena and choice of the best computer programs to provide the answers. If successful, one day in the near future, anyone should be able to ask, "Ok Molly, why is my catalyst so much better (or worse) than I initially designed?"

The second area of application of Molly is in data analysis and interpretation. Examples include the matching of either a mass spectrum or NMR spectrum of an unknown compound to those in the databases, which is being practiced in chemistry right now. A more sophisticated example is collection and analysis of databases of either chemical reactions or metabolic pathways so that Molly can predict reactivity of chemical and biological catalysts. A major issue is that many of these databases for different reactions are collected independent of each other, often written using completely different computer languages. As a result, it is very difficult to use one type of program to read and analyze data across different databases. An even further challenge for Molly is when there is not an exact match. Can it learn from the results in the current databases to deduce spectra and thus identify compounds not yet in the databases? If yes, one day in the near future, anyone should be able to ask, "OK Molly, what is in my material"?

The third and perhaps a more challenging area of the application of Molly is in the design of compounds that have predictable properties such as the ability to bind a target molecule selectively or catalyze a reaction efficiently. Given the tremendous success in designing protein-protein interactions using computer programs,[5] including those taking advantage of participation by the general public via crowd sourcing computational powers (Figure 1a),[6] it is possible in the near future to design compounds that can bind a target molecule, including pharmaceutical drugs that bind and inhibit certain enzymes in metabolic pathways (Figure 1b).[7,8] It is much more difficult, however, to design catalysts that can accelerate certain reactions, because, instead of focusing on the resting states as in designing most binders, catalysts exert most of their influence on transition states, which are often not well

characterized. It is particularly challenging to design functional metalloenzymes or metal-containing catalysts. Unlike non-metalloenzymes where the geometric parameters and bonding characters of active sites containing carbon, oxygen and nitrogen centers are well defined, those of metal-containing active sites, especially in the transition state, vary much more widely and are less well-defined.[9] As a result, few metalloenzymes have been designed that display activities can match those of native enzymes (Figure 1c).[10] This challenge also provides an excellent opportunity for Molly to implement artificial neural network and machine learning to search good examples of native enzymes and proven chemical catalysts and come up with new designs that a group of chemists and their collaborators could not otherwise achieve due to limitations of knowledge. If successful, one day in the near future, anyone should be able to ask, “OK Molly, how to design a compound that can bind to this target or catalyze this reaction?”

An even further challenge for Molly is to propose synthetic routes for compounds of interest. While retro-synthesis has been practiced by organic chemists for decades and some programs have been developed to implement the practices (Figure 2a),[1–4,11–13] it is still difficult to replace manual designs with computer programs, at least in the frontiers of research programs. A major challenge is that there are multiple branches and procedures to synthesize the same target compounds, and it is difficult to evaluate and decide on the most efficient route with the best overall yields and lowest costs. Like a chess game, the first step that is obviously the most efficient may not lead to the overall victory; one needs to take into account the next few steps before making the first move. In addition to overall yields, Molly also needs to take into consideration other factors such as overall costs, chemical safety, and waste generation. One advantage of machine learning is that, like human being, it can learn from not only successful, but also failed experiements.[14] The key barrier to Molly’s success is how to parameterize chemical compounds, particularly their functional groups in three-dimensions, as well as their reactivity (Figure 2b), into digital forms that computers can recognize, and how to combine the chemical information with AI algorithms such as pattern recognition, machine learning, and automated reasoning. If these barriers can be overcome, one day in the near future, anyone should be able to ask, “OK Molly, how to synthesize this compound?”

Most people who have used voice-activated assistants such as “Ok Google” are often impressed by the initial answers to certain simple questions and then frustrated by “I don’t know how to answer that yet” in response to some more sophisticated questions. Molly is no exception. An ultimate challenge for Molly is the ability to answer more sophisticated questions such as “how to cure cancer?” because answering it correctly requires knowing solutions to many components of the problems, such as what causes cancer, how to design compounds or drugs to inhibit the key components that cause cancer, and how to synthesize and then test these compounds in cancer cells in animals and in human, based on knowledge from clinical trials. Just like users of “Ok Google” who believe that, with more time and training, it will get better at answering more sophisticated questions, Molly will, too.

3. “Ok Molly” as a means for more effective chemical education and outreach to the general public

While Molly is designed to mainly help chemists to expand the frontiers of research as described above, she will also have a huge impact in our chemical education and public outreach. Many students think chemistry is difficult to understand and change their majors from chemistry and shun chemistry after graduation as a result. AI algorisms, such as those in machine learning, have the potential to discover best types of materials and best means to deliver the materials to cater individual styles of learning in order make it easier to study and understanding chemistry. Molly will also make chemical education much more accessible to anyone that has a smartphone or a similar device, like a tablet. However, a well-developed Molly who know answers to almost all questions may pose challenges for course instructors to test students in giving homework, quizzes, and exams, as the students can often find answers to the questions by asking Molly. However, just like calculators and Google searches have not made tests in courses obsolete, Molly will not either, as educators will adapt the practices to make instructions more effective. For the same reason, even if Molly can answer all the questions, it will not take away our chemistry jobs, as it will only make us work more efficiently and intelligently, because there will always be challenges to integrate AI with chemistry to make Molly smarter, and there are always new and higher level questions that Molly needs to answer.

Moreover, we chemists often lament the wrong perception of chemistry by the general public who frequently associate chemistry with toxic chemicals. The successful development and implementation of Molly will help change such a perception in demonstrating that chemistry is much more useful for everyday lives and many seemingly difficult chemistry concepts can be explained in plain language by Molly. One way to bring a better understanding of chemistry to the public is to equip Molly with highly sensitive chemical sensors interfaced with smartphones for on-site and real-time detection, wireless transmission and interpretation of the results (Figure 3a), demonstrating how chemistry can make their lives better.[15–17] To make Molly work well in this area, we need to make the sensor attachments as small as possible, and as generally applicable to as many targets as possible. A primary example is to repurpose the widely available and well developed portable meters such as FDA-approved glucose meter attached to smartphone to measure many other non-glucose targets, by transforming selective binding of non-glucose targets by aptamers or antibodies into generation of glucose (Figure 3b).[18,19] The development of sensor arrays is also very important[20] because it is extremely difficult to design a sensor with 100% selectivity towards the target without any interference from other components in the real samples; sensor arrays that mimic our nose or tongue can overcome this limitation. Just like our brains can tell us what is in the material based on smell or taste, the artificial neural network and machine learning capability of Molly can be developed to identify and quantify any materials in front of us.

4. Conclusion and perspectives

Like many Holy Grails, Molly sounds familiar, as many people have been pursing it for a long time. However, to achieve its true mission, we are still far away from it. This commentary attempts to summarize past achievements, present practices, and future challenges for Molly in order to stimulate discussions and further development in this area. Given the progress made in AI behind “Ok Google”, IBM’s Deep Blue, Watson, and Google’s AlphaGo, it is time to integrate them for chemical research, education and public outreach. To achieve the goal, we chemists cannot wait for the AI community to apply the AI algorithms to chemistry problems. Instead, we need to play an active role in recruiting and collaborating with AI researchers to overcome major barriers, some of which are described above. We also need to educate a new generation of chemists who are experts in the interface of AI and chemistry.

Molly does not belong just to us chemists. Since chemistry is the central science, and molecules are the building blocks of most science and engineering, including biology, bioengineering, and materials science and engineering, the development of the platform requires expertise from all disciplines. Its applications will have a huge impact on not only all of science and engineering, but all of society in general.

While the majority of users will employ Molly for good causes, it is inevitable that some people will exploit Molly for malicious purposes, including something as dire as terrorists asking Molly how to make chemical weapons. Additionally, while machine learning can be powerful, it can be problematic if Molly is fed with misinformation on purpose by people with unscrupulous intentions. One example is Tay, a Twitter bot created by Microsoft as an experiment in AI based conversational understanding. In principle, the more Tay chats with people, the smarter it gets. However, it took less than 24 hours for Twitter to corrupt an innocent Tay to make racist and sexist remarks. These negative effects are not reasons for us to stop developing Molly; they are incentives for us to work harder to make Molly not only think like a human being, but also learn to have a moral and ethical standard in carrying out chemical research and education. For example, can Molly be programed to learn to identify false positive results or her own mistakes, and then to employ proof-reading and error removal algorisms to correct these mistakes. If she can, such a Molly sounds like an ultimate Holy Grail.

Acknowledgments

I wish to thank Lu group members for their help with revising this commentary and Kevin Harnden, Julian Reed and Chang Cui for their help with making the figures. The Lu group research is supported by the National Science Foundation (CHE-0552008) and National Institute of Health (GM062211 and MH110975).

References

1. Lindsay, RK.Buchanan, BG.Feigenbaum, EA., Lederberg, J., editors. Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project. McGraw-Hill, Inc; 1980.
2. Pierce, TH., Hohne, BA. Artificial Intelligence Applications in Chemistry. American Chemical Society; Washington, DC: 1986. ACS Symposium Series

3. Hippe, Z. Artificial Intelligence in Chemistry: Structure Elucidation and Simulation of Organic Reactions. Elsevier; 1991.

4. Cartwright, H., editor. Using Artificial Intelligence in Chemistry and Biology: A Practical Guide. CRC Press; 2008.

5. Huang PS, Boyken SE, Baker D. The coming of age of de novo protein design. Nature. 2016; 537:320–327. [PubMed: 27629638]

6. Khatib F, DiMaio F, Foldit Contenders Group; Foldit Void Crushers Group. Cooper S, Kazmierczyk M, Gilski M, Krzywda S, Zabranska H, Pichova H, Thompson J, Popović Z, Jaskolski M, Baker D. Crystal structure of a monomeric retroviral protease solved by protein folding game players. Nature Struct Mol Biol. 2011; 18:1175–1177. [PubMed: 21926992]

7. Subramanian G, Ramsundar B, Pande V, Denny RA. Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. J Chem Inf Model. 2016; 56:1936–1949. [PubMed: 27689393]

8. Pereira JC, Caffarena ER, dos Santos CN. Boosting Docking-Based Virtual Screening with Deep Learning. J Chem Inf Model. 2016; 56:2495–2506. [PubMed: 28024405]

9. Lu Y, Yeung N, Sieracki N, Marshall NM. Design of functional metalloproteins. Nature. 2009; 460:855–862. [PubMed: 19675646]

10. Yu Y, Cui C, Liu X, Petrik ID, Wang J, Lu Y. A Designed Metalloenzyme Achieving the Catalytic Rate of a Native Enzyme. J Am Chem Soc. 2015; 137:11570–11573. [PubMed: 26318313]

11. Corey EJ, Long AK, Rubenstein SD. Computer-assisted analysis in organic synthesis. Science. 1985; 228:408–418. [PubMed: 3838594]

12. Todd MH. Computer-aided organic synthesis. Chem Soc Rev. 2005; 34:247–266. [PubMed: 15726161]

13. Hanessian SM. Man, machine and visual imagery in strategic synthesis planning: computer-perceived precursors for drug candidates. Curr Opin Drug Discov & Devel. 2005; 8:798–819.

14. Raccuglia P, Elbert KC, Adler PDF, Falk C, Wenny MB, Mollo A, Zeller M, Friedler SA, Schrier J, Norquist AJ. Machine-learning-assisted materials discovery using failed experiments. Nature. 2016; 533:73–76. [PubMed: 27147027]

15. Laksanasopin T, Guo TW, Nayak S, Sridhara AA, Xie S, Olowookere OS, Cadinu P, Meng F, Chee NH, Kim J, Chin CD, Munyazesa E, Mugwaneza P, Rai AJ, Mugisha V, Castro AR, Steinmiller D, Linder V, Justman JE, Nsanzimana S, Sia SK. A smartphone dongle for diagnosis of infectious diseases at the point of care. Sci Transl Med. 2015; 7:273re1.doi: 10.1126/scitranslmed.aaa0056

16. D'Ambrosio MV, Bakalar M, Bennuru S, Reber C, Skandarajah A, Nilsson L, Switz N, Kamgno J, Pion S, Boussinesq M, Nutman TB, Fletcher DA. Point-of-care quantification of blood-borne filarial parasites with a mobile phone microscope. Sci TranslMed. 2015; 7:286, 286re4.doi: 10.1126/scitranslmed.aaa3480

17. Ming K, Kim J, Biondi MJ, Syed A, Chen K, Lam A, Ostrowski M, Rebbapragada A, Feld JJ, Chan WCW. Integrated Quantum Dot Barcode Smartphone Optical Device for Wireless Multiplexed Diagnosis of Infected Patients. ACS Nano. 2015; 9:3060. [PubMed: 25661584]

18. Xiang Y, Lu Y. Using personal glucose meters and functional DNA sensors to quantify a variety of analytical targets. Nature Chemistry. 2011; 3:697–703.

19. Lan T, Zhang J-J, Lu Y. Transforming the blood glucosemeter into a general healthcaremeter for in vitro diagnostics in mobile health. Biotech Adv. 2016; 34:31–41.

20. Lim SH, Feng L, Kemling JW, Musto CJ, Suslick KS. An optoelectronic nose for the detection of toxic gases. Nature Chemistry. 2009; 1:562–567.

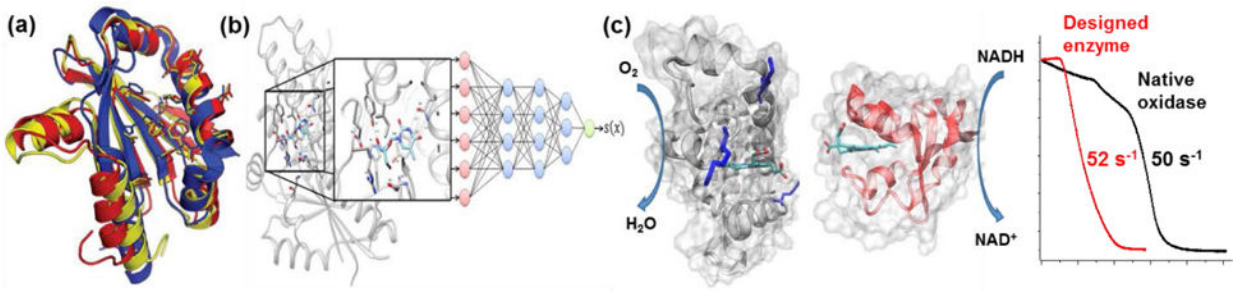


Figure 1.

a) Crystal structures of a monomeric retroviral protease solved by protein folding game players. Starting from a model (red), a group of protein folding game players generated a model (yellow) that was closer to the crystal structure later determined (blue).[6] The figure is produced from Reference 6, with permission from the Nature Publishing Group; b) Employing deep-learning neural network to improve virtual screening of libraries of molecules that dock strongly to their targets;[8] c) A rational designed enzyme with catalytic activity matching that of a native oxidase.[10]

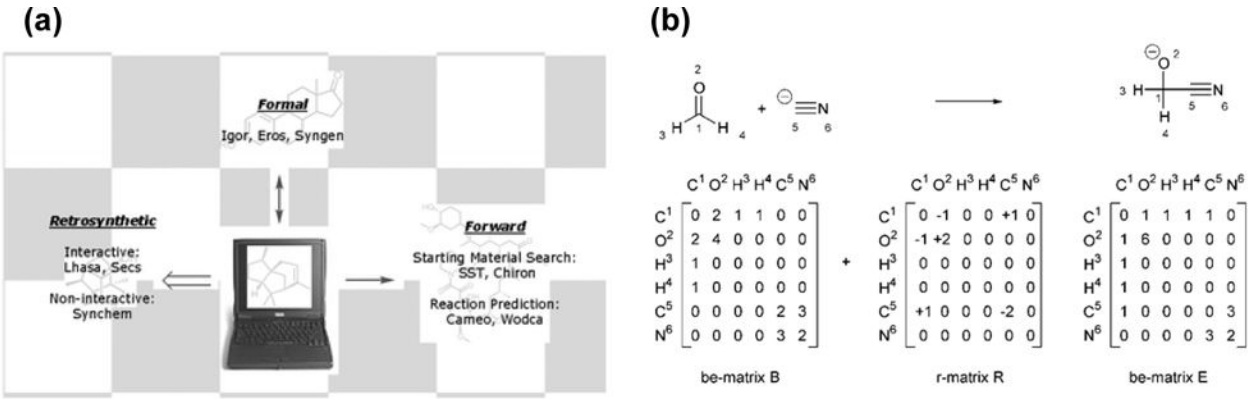


Figure 2.

a) Selected computer-aided organic synthesis programs; b) An example of describing an organic reaction in digital forms by one of these programs. The figures are reproduced from Reference 12, with permission from The Royal Society of Chemistry.

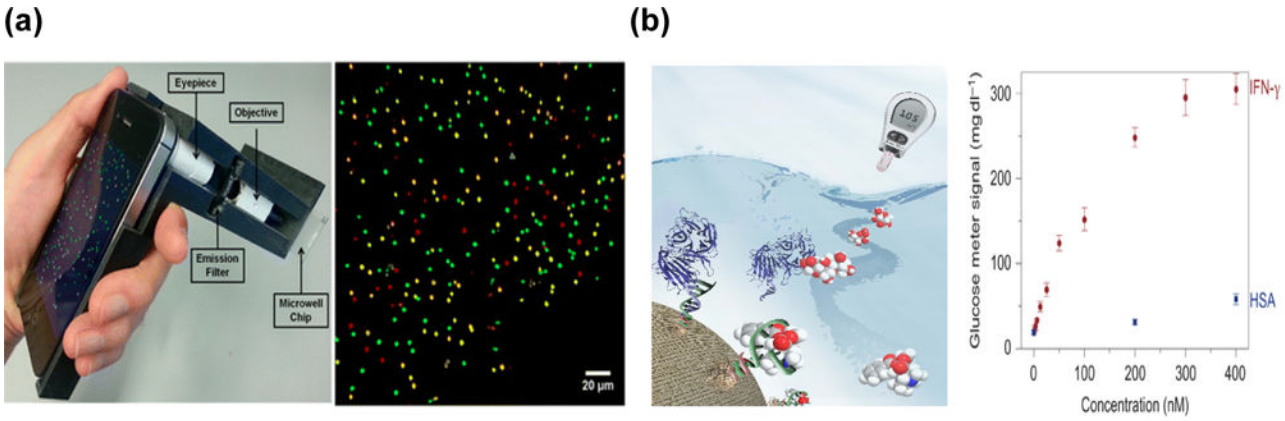


Figure 3.

a) Integrated quantum dot barcode smartphone optical device for wireless multiplexed diagnosis of infected patients.[16] b) Repurposing the widely available and well developed portable glucose meter as a general health monitor for many other non-glucose targets, such as interferon gamma (IFNγ), a biomarker for tuberculosis, with excellent selectivity against other components in the blood, such as human serum albumin (HSA).[17]