Plausibly deniable chemical encryption: hiding a molecule in a haystack

Wojciech Nogaś[1*]

[1] *Chemical and Biological Research Center, University of Warsaw, Department of Chemistry, Żwirki i Wigury Avenue 101, 02-089 Warsaw, Poland*
[*] *Corresponding author e-mail:* wnogas@chem.uw.edu.pl

Abstract: we propose an outline of plausibly deniable chemical encryption algorithm, a technique aimed at increasing the cost of a small organic molecule identification in a sample by means of chemical analysis through mixing it with a rationally designed randomized mixture of analytical interferents, in a remote analogy to other domains of cryptography. The algorithm is then applied in a proof-of-concept demonstration example.
Keywords: chemical, analysis, identification, small organic molecules, cryptography.

Introduction.

Last decades saw a wide adoption of cryptographic techniques, bringing about security to daily financial operations and communications. Innovations such as cryptocurrencies and blockchain were introduced, opening new prospects in global finance and other fields[1,2]. However, cryptography has not been a very inspiring field for chemists and there are scarce data on relating cryptographic techniques to chemical reality[3]. Science is generally meant to be publicly available and only rarely there is a need for keeping the results secret. An application already considered is sharing information on proprietary pharmaceuticals' properties without sharing their structures[4]. Most of the effort in chemical cryptography has been devoted to molecules designed for storing information and exposing it under certain chemical or physical stimuli (or a specific sequence of stimuli – a decryption key), although the authors summarize them as proof-of-concept works of limited practicality. The main obstacle is unpredictability and cost of the molecules of sufficient information storage capacity[5]. A very interesting idea was the implementation of decision making algorithms in molecules and molecular logic gates, inherently capable of storing a logic state[6,7]. A notable cryptosystem involving m-SMS compound was published in Nature in 2016. It is a molecule containing several pH-, ion-, vicinal alcohol- and solvent-sensitive fragments whose interactions give rise to an enormous variability of the molecule's fluorescence spectrum in response to external conditions (the cryptographic key). Then it was used to encipher a message in a series of proof-of-concept experiments[8]. Apart from storing and extracting information in and from molecules, spontaneous crystallization process was recently used as a source of randomness, which is essential for reliable encryption algorithms[9].
While there are many chemical systems designed for information storage, we believe there are problems that have not been addressed at all. Inspired by this, we envisioned another scientific question to tackle. As chemical structure itself contains certain information, we considered whether it is possible to apply cryptographic principles to protect the structures of small organic molecules against attempts to elucidate them by means of chemical analysis. We believe this may be a valuable contribution with potential uses. Just to name a simple one, we imagine Bob to be a chemical manufacturer who sells a valuable chemical product V to Alice. If its structure is disclosed or reverse-engineered by means of chemical analysis, then Oscar may produce it and sell it to Alice too, thus undermining Bob's business. In order to prevent Oscar from doing so, Bob may patent the chemical or decide to keep it secret, at a risk of finding it reverse engineered. In this latter case, chemical encryption may be at least an interesting option to think about in order to foil reverse engineering attempts. That problem prompted us to figure out a general way of protecting chemicals against the methods of chemical analysis in a way that does not destroy them or compromise their

desired property: for a theoretical example, an agricultural company selling a pesticide that should not be rendered inactive upon encryption. To the best of our knowledge, there is no such method already published. We formulate the problem as equivalent to plausibly deniable encryption[10], which is a combination of encryption and steganography, so that the information is not only encrypted, but also it is possible to convincingly deny its presence there. Translating that to chemistry, we are wishing to design a sample containing an encrypted chemical, in which it is impossible to discern if the encrypted chemical is indeed present using a set of analytical chemical procedures. This may be in other words expressed as the inability to test some hypotheses concerning the structure of the encrypted chemical present in the mixture.
To attain the desired result, we drew our conclusions from the facts, that all physico-chemical analytical procedures have a limitation in their resolving capability and that it is much easier to make a mixture of multiple chemicals than to separate the mixture into individual components. Therefore it is theoretically possible to design a sophisticated, difficult to separate, mixture of analytical interferents yielding ambiguous results when analyzed.

Derivation of chemical plausibly deniable encryption algorithm.

We intend to derive a general algorithm for the preparation of a mixture, which, upon addition to a chemical to be encrypted, will make it unidentifiable in the mixture without its destruction. First a general model of chemical analysis is derived. Qualitative analytical chemistry uses analytical procedures to ascertain the presence of certain chemicals in complex mixtures. Let's define $f_0()$ as an analytical procedure, for example ^{1}H NMR spectrum measurement. A set of $\{f_0(), f_1(), \ldots, f_j()\}$ may be combined to yield another analytical procedure $f_m()$. This is equivalent to applying a set of consecutive chemical procedures to the sample, such as column chromatography followed by HPLC separation followed by NMR measurement etc. To interpret the result of an analytical procedure $f_m()$, one uses a reciprocal function $f_m^{-1}()$ which is the attribution of analytical results to chemicals. For example, for a pH measurement (analytical procedure) of an aqueous solution a value below 7 (its result) can be attributed to a set of all water soluble and dissociating acidic chemicals. Logically, a qualitative analytical chemist intends to combine multiple analytical procedures $\{f_1(), f_2(), \ldots, f_j()\}$ into $f_m()$ so that its results may be attributed to one chemical only to ensure an unambiguous identification.

Then we define m_{i+1} as the moles of the substance to be hidden in a mixture M, which contains the chemical m_{i+1} among other chemicals $m_1, \ldots, m_i$:

$$M = \{m_1, m_2, \ldots, m_i, m_{i+1}\}.$$

A chemical m_{i+1} is considered identified in the mixture M, if there is an analytical procedure $f_m()$, whose result, when applied to M, is attributable only to one compound m_{i+1}:

$$f_m^{-1}(f_m(M)) = m_{i+1}.$$

When a new organic chemical is obtained, its proposed structure is usually confirmed by a set of typical analytical procedures: NMR spectra measurement, elemental analyses, High Resolution Mass Spectrometry, IR spectrum. This serves then as a guide to confirm the presence of this compound in other mixtures.
In order to plausibly deny the presence of the chemical m_{i+1} in the mixture M, ideally, there should be no analytical procedure f_m that, when applied to M, yields results attributable to m_{i+1} only. This is in general nearly impossible due to a theoretically unlimited availability of analytical procedures (which includes also purification methods) that can be applied consecutively. However, due to the relatively high cost of nonstandard chemical analysis, the potential attacker's capabilities are usually limited.

Let's define the conditions that have to be satisfied by the mixture M to prevent the analytical procedure $f_m(M)$ from resolving the proper signal of m_{i+1} and the signals of the interferents' m_1, m_2, …, m_i, thus rendering the result of $f_m(M)$ inconclusive, by its result being explainable both by the presence of m_{i+1} and the matrix signals of m_1, m_2, …, m_i at the same time. Thus the presence of m_{i+1} in M can be plausibly denied. All analytical procedures, including separation techniques, are characterized by their resolution α. From that observation we immediately conclude, that analytical signals of one or more of m_1, m_2, …, m_i should differ by less than α from the analytical signal of m_{i+1} to successfully protect against an analytical procedure $f_m()$. Therefore the condition for $M/\{m_{i+1}\} = \{m_1, m_2, \ldots, m_i\}$ is:

$$f_m(M)-f_m(M\backslash\{m_{i+1}\}) < \alpha.$$

The mixture M is naturally obtained by adding the chemical m_{i+1} to a rationally designed interferent mixture $M/\{m_{i+1}\}$:

$$M\backslash\{m_{i+1}\} = \{m_1, m_2, \ldots, m_i\}.$$

Thus the problem of chemically encrypting the chemical m_{i+1} is reduced to finding a composition of the mixture $M/\{m_{i+1}\}$, so that the analytical methods we intend to protect against will not yield sufficiently resolved signals to enable the identification of the chemical m_{i+1} in the mixture M.

This may be achieved by adding chemicals whose analytical results will be similar to those of m_{i+1}, effectively overlapping with the signals of m_{i+1} and rendering the analytical result unreliable and/or targeting the principle of operation of the analytical method, for if the sample destroys the test, its results will be inconclusive as well. A theoretical example of the latter would be the addition of protein-denaturing agents to the m_{i+1} chemical to protect it against enzymatic analytical methods.

For practical reasons, the interferent mixture $M\backslash\{m_{i+1}\}$ should satisfy additional conditions.
1° if $M/\{m_{i+1}\}$ is the same in more than one encrypted samples, then the security may be compromised by a comparison attack. To avoid it, the $M/\{m_{i+1}\}$ composition (both molar fractions and chemicals) should be random in each instance.
2° depending on the application, in order to avoid the loss of m_{i+1}, no component of $M\backslash\{m_{i+1}\}$ should destructively react with m_{i+1}. Reactions may be used to add layers of security, but they have to be reversible, like acid-base reactions. However, this algorithm may be used as an irreversible one, by allowing destructive interferents and these reaction products should not allow for the identification of the parent encrypted compound. Naturally, in a complex mixture of chemicals, unpredictable reactions are always possible, so it should be tested in every practical application if the encrypted chemical has been successfully preserved upon encryption.

Experimental.

The general algorithm was applied in a practical example of encrypting a chemical against NMR techniques and GC separation. Assume the chemical to be encrypted is salicylic acid. We are not having in mind any of its properties that have to be intact during the encryption, so the experiment will be focused on the demonstration of the application of the aforementioned encryption algorithm. As an encryption method must be generally complemented with an decryption method, work on chemical decryption algorithms is under way.
Salicylic acid is a prototypical small organic molecule featuring a structure of substituted aromatic ring with a distinctive pattern of ^{1}H and ^{13}C NMR signals and characteristic Brønsted acidity of carboxylic and phenolic groups. Following that, we want to create a mixture containing *o*-hydroxybenzoic acid, whose signals in NMR spectroscopy and GC chromatograms are rendered indiscernible by overlapping with other components' signals and shifted owing to acid-base

interactions (i. e. salicylic acid deprotonation and protonation of the amines) with bases present in the interferent mixture.

The interferent mixture composition is randomly selected from a database of small aromatic compounds (SI 2.) being a substituted single phenyl ring to ensure structural similarity with the compound to be encrypted. All of these interferents exhibit NMR signals in the aromatic range of 7-8 ppm in ^{1}H NMR and 120-170 ppm in ^{13}C NMR, so we expect this mixture will ensure sufficient overlap with salicylic acid signals. In this case we want to avoid the destruction of the encrypted compound, so only acid-base reacting interferents are allowed, i. e. aromatic amines. Attempting to retrieve it would be a non-trivial task of designing a mixture-specific decryption procedure. We will consider the algorithm tested successfully, if the encrypted sample's NMR spectra and GC chromatogram does not exhibit signals allowing for salicylic acid identification.
List of chemicals available as interferents, selected interferent mixture compositions and the JavaScript code used to generate them are attached below (SI 1.).

The analytical procedure is a combination of ^{1}H and ^{13}C NMR spectra interpreted by a human. If applied on pure unencrypted sample of salicylic acid, it yields specific results allowing for easy identification of the chemical.

Interferent preparation.

Script randomInterferentGenerator.js was run to choose a random interferent composition (SI 2.). This mixture is designed to contain a random amount of every component in the range 0.01 – 0.06 mmol. Then 10 mg/ml solutions of all interferents in DCM were prepared and according to the script generated preparation procedure (SI 3.), appropriate amounts of these solutions were combined, DCM was evaporated and the red oily mixture was dissolved in 10 ml of DCM, yielding the interferent mixture $M\backslash\{m_{i+1}\}$.

3 ml of the interferent solution was added to salicylic acid solution ($\{m_{i+1}\}$) (2.26 mg, 10 mg/ml, 0.226 ml) and 100 µl of the mixture M was taken for GC measurement. Another 3 ml of the interferent solution was evaporated, dissolved in 1 ml of CD_2Cl_2 and ^{1}H, ^{13}C NMR spectra were measured. Pure salicylic acid spectra and GC chromatogram were measured for comparison.

Spectrum analysis is performed by listing all signals with MestreNova Automatic Peak Picking function. Salicylic acid is considered identified in the sample, if the sample's spectrum contains the signals of pure salicylic acid matching their position, multiplicity and coupling constants in the case of ^{1}H NMR.

Results.

The measured NMR spectra of encrypted salicylic acid samples are provided below (Fig. 1., Fig. 2. SI 4. *Salicylic acid*). For comparison, pure salicylic acid spectra are also recorded (Fig. 3., Fig. 4. and SI 4. *Encrypted sample* section).

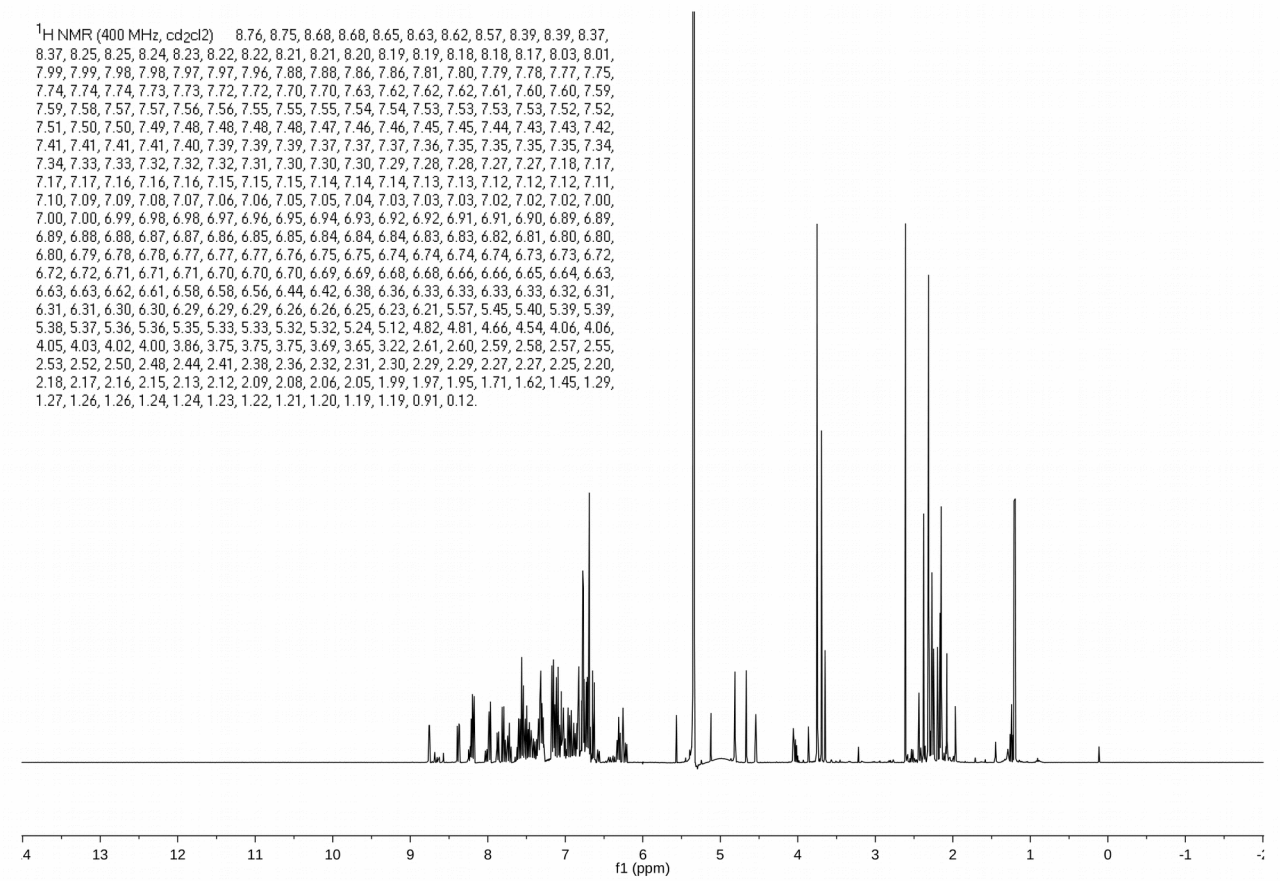

Fig. 1: ^{1}H NMR spectrum of encrypted salicylic acid sample.

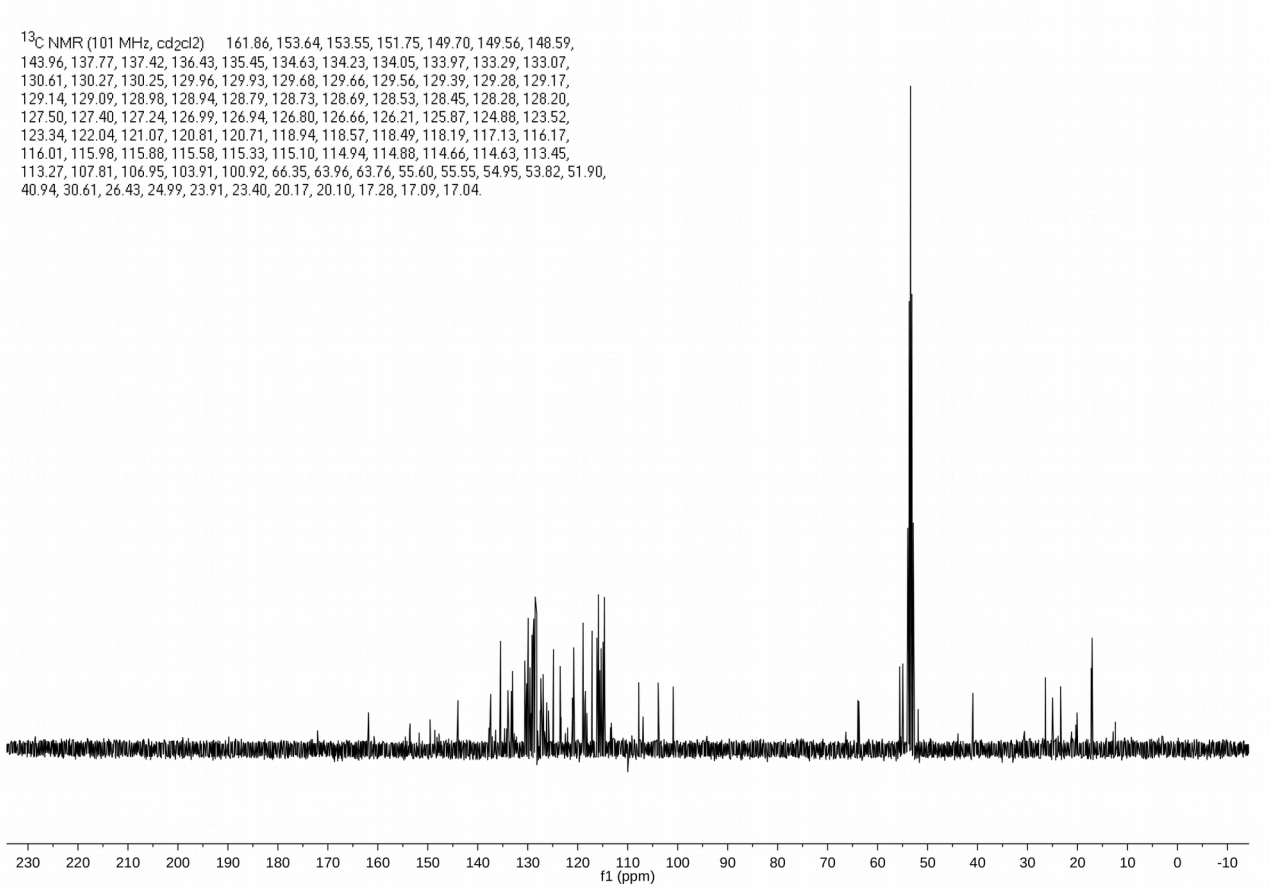

Fig. 2: ^{13}C NMR spectrum of encrypted salicylic acid sample.

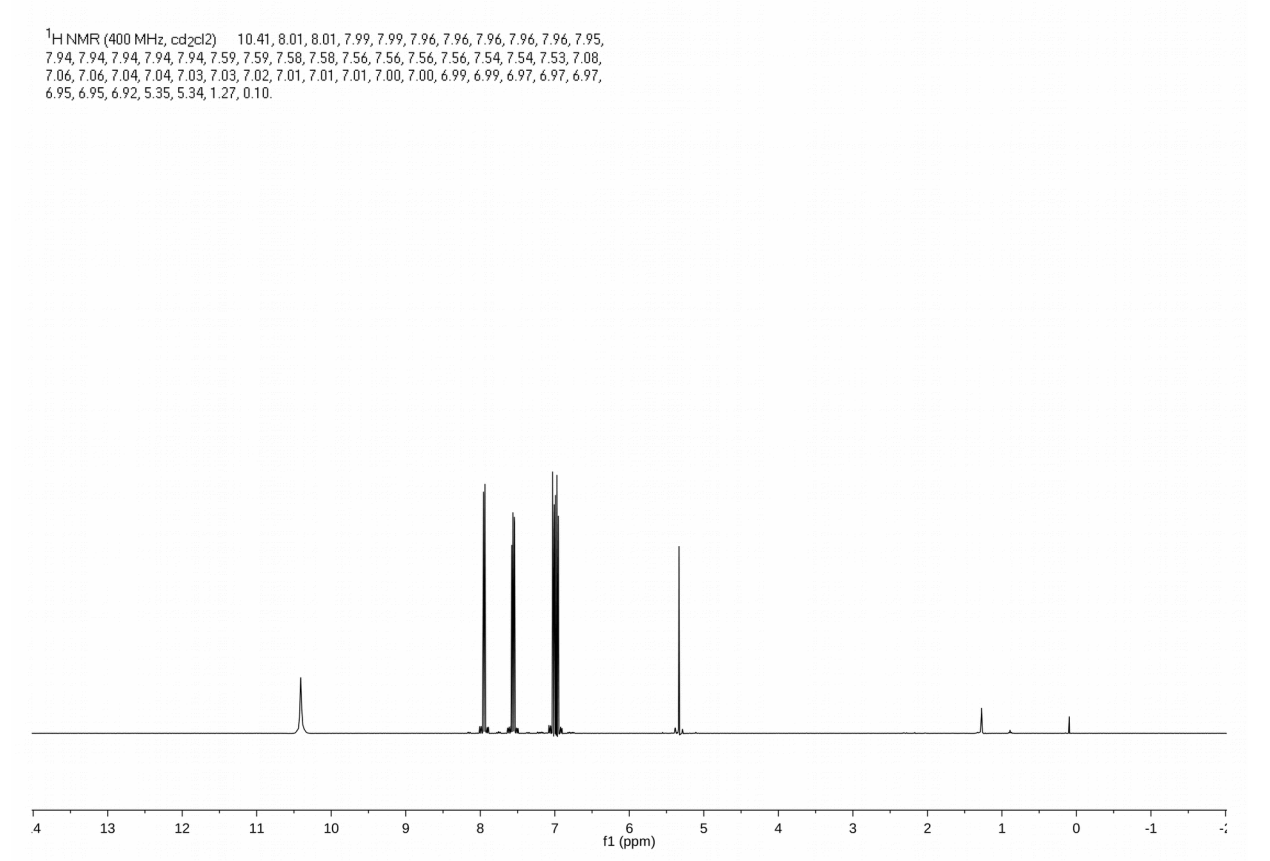

Fig. 3: ^{1}H NMR spectrum of pure salicylic acid sample.

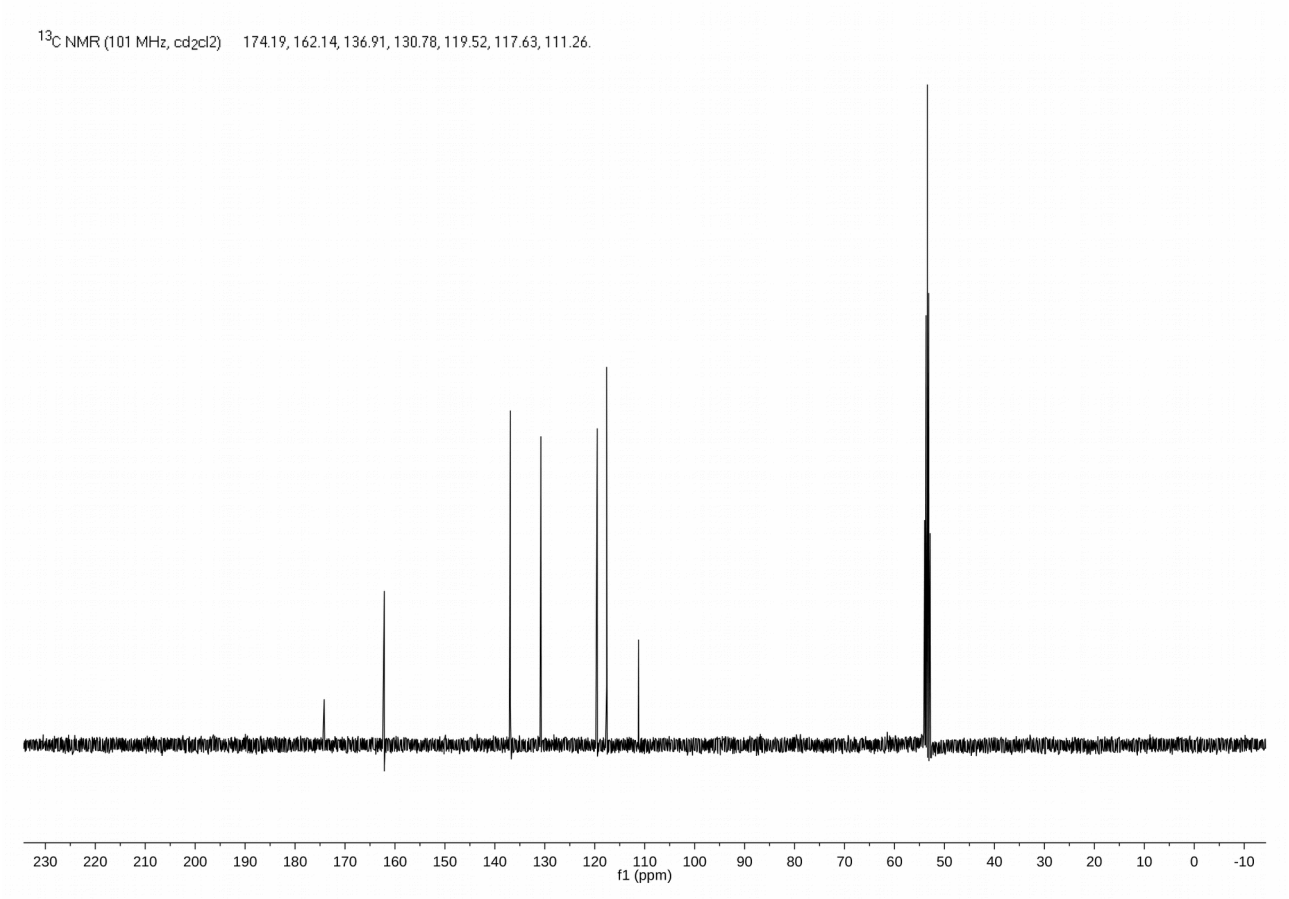

Fig. 4: ^{13}C NMR spectrum of pure salicylic acid sample.

^{1}H NMR spectrum of encrypted salicylic acid is given in Fig. 1 and SI 4. *Salicylic acid* section. Aromatic proton signals of pure salicylic acid comprise of two doublets and two triplets (Fig. 3. and 5.). When the pure and encrypted spectra are compared (Fig. 5.), the pure salicylic acid signals are not distinguishable in the encrypted sample's spectrum. In the ^{13}C NMR spectrum of the encrypted sample (Fig. 2. and *SI4. Encrypted sample* section), the salicylic acid signals (Fig. 4. and SI 4. *Salicylic acid* section) are also indiscernible, as the set of all ^{13}C NMR signals of the encrypted sample does not contain all of the pure salicylic acid signals, despite the acid is present in the sample.

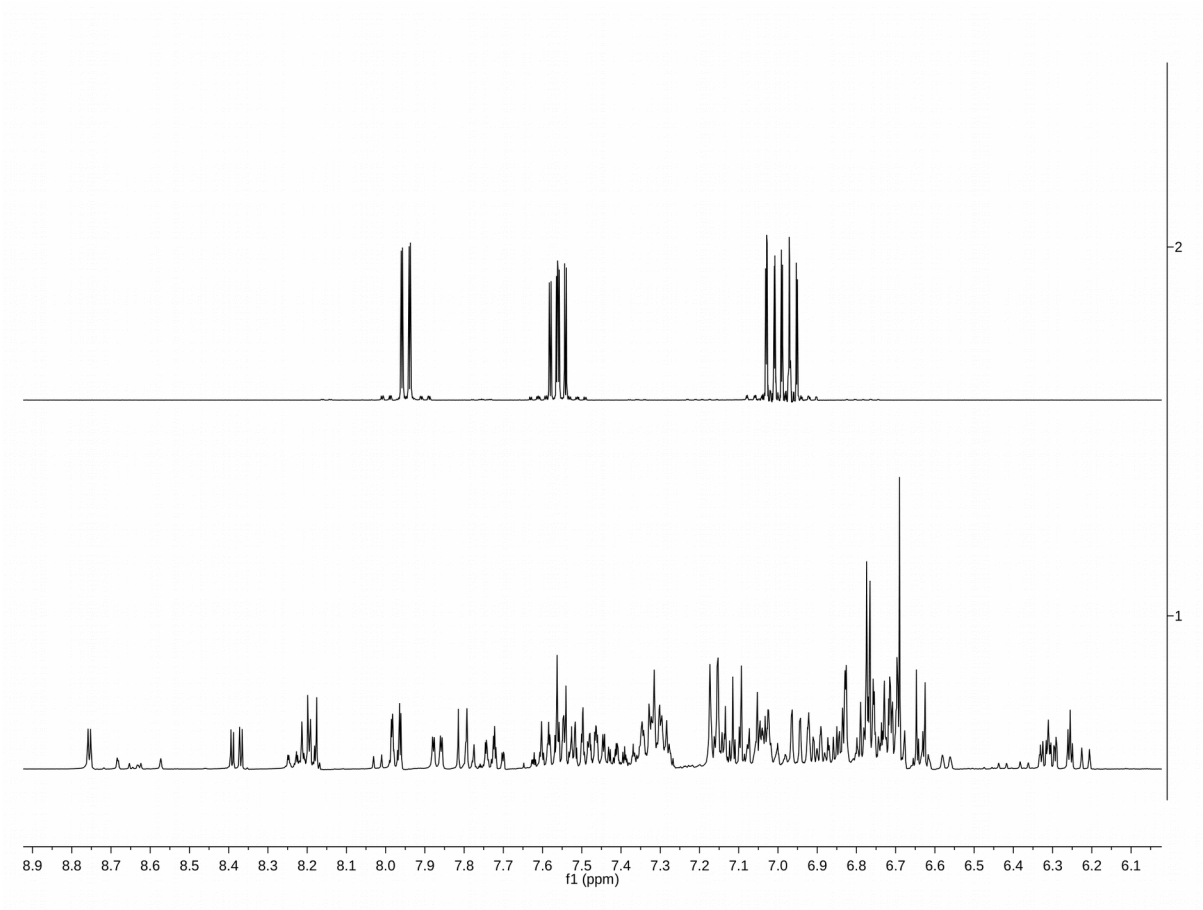

Fig. 5: Comparison of aromatic 7 - 8 ppm range in ^{1}H NMR spectra of (up) pure salicylic acid sample and (down) encrypted salicylic acid sample.

Having successfully obscured the salicylic acid signals in NMR spectra with the interferent mixture, we considered whether the characteristic signal of salicylic acid in GC chromatogram becomes indistinguishable upon encryption (Fig. 6.). While pure salicylic acid exhibited a signal at 17.6 min eluting time, it was absent in the encrypted salicylic acid sample. In fact, the chromatograms of interferent and interferent with the salicylic acid bear no visible differences, suggesting that the acid was rendered involatile through salt formation with the amines present in the mixture.

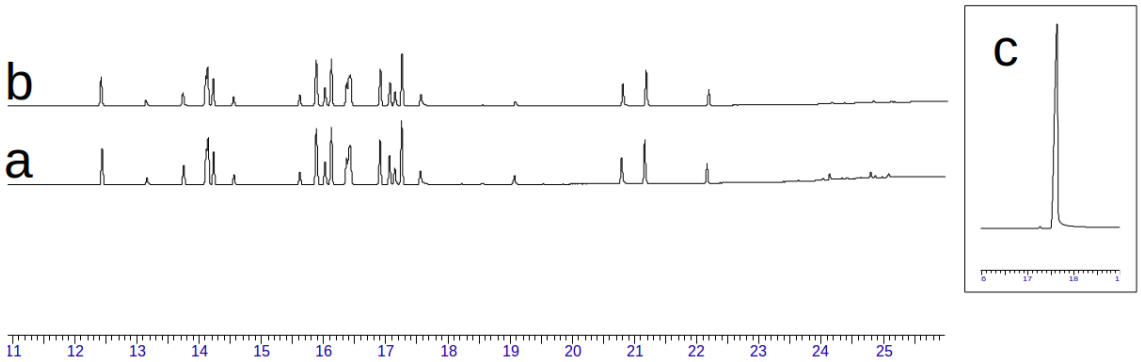

Fig. 6: GC chromatograms of a) interferent, b) interferent with encrypted salicylic acid and c) salicylic acid.

Discussion.

Successfully, the interferent mixture rendered the salicylic acid signals in ^{1}H and ^{13}C NMR irrecognizable. This was due to overlap with the signals of other constituents of the mixture and due to acid-base reactions with the amines in the interferent, shifting the signals of the acid and of the amines involved in a way that is unpredictable without knowing the exact composition of the interferent.
As interferent aromatic signals in ^{13}C NMR tend to be scattered over a wider range of chemical shift scale, we expected much less efficient overlap. The efficiency of protection against this technique, according to the algorithm, requires only including a higher number of interferents or more precise interferent choice to ensure overlap with salicylic acid signals over the broad range of ^{13}C NMR signals.

Encrypted salicylic acid did not form a signal in GC, probably because of the formation of nonvolatile salts with amines present in the mixture.

In conclusion, the interferent mixture designed according to the aforementioned algorithm has been successful in protecting the chemical against its identification in the mixture using ^{1}H and ^{13}C NMR techniques and prevented it from being observed in GC chromatogram. Of course, now a specific procedure may be devised in order to break the the encryption. However, in this way, the overall cost of the identification of salicylic acid in the mixture will be higher, which is the point of chemical encryption in order to deter attacks.

The amount of interferents in the encrypting mixture was rather low (28 constituents), but may be increased indefinitely given the abundance of different organic structures in chemical space. However, the main cost of better encryption would be the increase of the total mass of the interferent mixture, since the encrypted component is diluted as the amount of interferents is increased. Optimal encryption would require designing optimized interferent mixtures which are maximizing the signal overlap in as many analytical techniques (including separation methods) as possible and minimizing the necessary amount of interferents at the same time.

For the algorithm to be useful against other typical techniques, interferents against MS, IR, UV-VIS spectroscopies etc. need to be included. This is a matter of choosing the right interferents fulfilling the conditions mentioned above in *Derivation of chemical plausibly deniable encryption algorithm* section to ensure successful overlap of signals that renders these methods' results ambiguous.
On the other hand, a much simpler protection is enough against chemical specific tests, like immunological or enzymatic ones (protein denaturants as interferents here).

It is worth noting that cryptographic security principles explicitly prohibit repeated use of the same interferent mixture composition as in this case the interferent signals may be easily subtracted, compromising the security. Therefore pure interferent mixture NMR spectra were not measured as a control experiment.

It may be argued, that eventually all complex mixtures may be resolved into individual components by a skilled chemist using appropriate equipment. An equivalent problem is pervasive in cryptography, where growing computational power is able to break older ciphers. For example the commonly used PGP cryptosystem, deemed unbreakable with todays' computers, may be compromised in the future with quantum computing[11]. However, security level is generally adjusted to the expected attacker's capabilities and determination, which is never infinite. Therefore we believe that chemical encryption will exhibit the same feature and the complexity of interferent mixtures will need to increase to match the analytical techniques' resolving capability.

Conclusions.

We derived and tested an algorithm designed to protect a model salicylic acid against chemical analysis by applying rationally designed interferent mixtures to it, that is difficult to separate and renders ambiguous signals when tested with ^{1}H an ^{13}C NMR techniques. While this is only a demonstration of a general algorithm and effort has to be devoted to protect against other analytical techniques, we believe, that the plausibly deniable chemical cryptography may be ultimately inspiring for analytical chemists and interesting in secret protection in science and industry. As the practical aim of chemical cryptography is to make chemical analysis of a sample costly enough to deter the attacker, the complexity of interferent mixtures may be increased, as it is extremely easier to make a mixture of chemicals than to fractionate the mixture into individual components, analogously to traditional encryption algorithms being continuously developed to match the surge in computational power of computers. Yet, while a computer can perform millions of brute force attacks per second, no one can perform that frequent chromatographic separations of a complex interferent mixture.

We believe that, while this technique requires further development, it may bring about interesting implications concerning the security of chemicals, especially control over how chemical information is protected against unauthorized access.

Acknowledgements.

We are grateful to University of Warsaw, Faculty of Chemistry seed funding program. We are also grateful to Michał Dąbrowski for precious advice.

References.

[1] – “A systematic literature review of blockchain-based applications: Current status, classification and open issues”, F. Casino, T. K. Dasaklis, C. Patsakis, *Telematics and Informatics*, **2019**, 36, 55-81.
[2] – “A Review on Blockchain Technology and Blockchain Projects Fostering Open Science”, S. leible, S. Schlager, M. Schubotz, B. Gipp, *Front. Blockchain*, **2019**, 2:16.
[3] - http://blog.rguha.net/?p=1515, access 22.IV.2020.
[4] - “Share and share alike”, D. Bradley, *Nat. Rev. Drug. Discov.*, **2005**, 4, 180.
[5] - “Molecules for security measures: from keypad locks to advanced communication protocols”, J. Andreasson, U. Pischel, *Chem. Soc. Rev.*, **2018**, 47, 2266-2279.
[6] - “Molecules with a sense of logic: a progress report”, J. Andreasson, U. Pischel, *Chem. Soc. Rev.*, **2015**, 44, 1053-1069.
[7] - “Molecules That Make Decisions”, A. Credi, *Angew. Chem. Int. Ed.*, **2007**, 46, 5472-5475.
[8] - “Message in a molecule”, T. Sarkar, *Nat. Commun.*, **2016**, 7:11374.
[9] - “A Crystallization Robot for Generating True Random Numbers Based on Stochastic Chemical Processes”, E. C. Lee, J. M. Parrilla-Gutierrez, A. Henson, E. K. Brechin, L. Cronin, *Matter*, **2020**, 2, 649-657.
[10] - “Defeating Encrypted and Deniable File Systems: TrueCrypt v5.1a and the Case of the Tattling OS and Applications”, A. Czeskis, D. J. St. Hilaire, K. Koscher, S. D. Gribble, T. Kohno, B. Schneier, *HOTSEC’08: Proceedings of the 3rd conference on Hot topics in security*, **2008**, 7, 1-7.
[11] - ”Introduction to post-quantum cryptography”, D. J. Bernstein, in: Post-Quantum Cryptography, **2009**, Springer, Berlin, Heidelberg.