Present-Day Practice of Non-Target Chemical Analysis

B. L. Milman* a, b, * and I. K. Zhurkovichb

a Institute of Experimental Medicine, St. Petersburg, 197376 Russia
b Clinical Research Center of Toxicology, Federal Medical-Biological Agency of Russia, St. Petersburg, 192019 Russia
*e-mail: bormilman@yandex.ru

Received June 19, 2021; revised August 8, 2021; accepted August 9, 2021

Abstract — We review the main techniques, procedures, and information products used in non-target analysis (NTA) to reveal the composition of substances. Sampling and sample preparation methods are preferable that ensure the extraction of analytes from test samples in a wide range of analyte properties with the most negligible loss. The necessary techniques of analysis are versions of chromatography–high-resolution tandem mass spectrometry (HRMS), yielding individual characteristics of analytes (mass spectra, retention properties) to accurately identify them. The prioritization of the analytical strategy discards unnecessary measurements and thereby increases the performance of the NTA. Chemical databases, collections of reference mass spectra and retention characteristics, algorithms, and software for processing HRMS data are indispensable in NTA.

Keywords: non-target analysis, mass spectrometry, chromatography, extraction, informatics, databases

DOI: 10.1134/S1061934822050070

Non-target chemical analysis is the determination of components of test samples unknown to the analyst (“known unknowns” and “unknown unknowns,” Table 1). In general, the detection of specific analytes belonging to a set of several million common individual compounds is most probable [1, 2]. The non-target chemical analysis takes an increasingly important place in modern scientific research of chemists and biochemists and the practical activities of technologists and engineers. Three factors determine this trend. The first reflects the growing need in such analytical determinations associated with the emerging environmental pollutants, more profound control of food quality, a consistent increase in attention to human health, etc. The two other factors enable the implementation of such analyses. Here, we should point out the current high level of the development of analytical methodology, resulting from the development of new types and models of chromatographs and mass spectrometers and new options for extracting compounds from test samples. Finally, we should note the rapid development of computer science, accompanied by improved characteristics of computers and their networks; the emergence of new databases; the creation of algorithms and corresponding software that enable the efficient manipulation with large volumes of the obtained data and reference information. These factors caused a sharp increase in the number of publications in the field of NTA: more than half of scientific articles have been published in the last 5 years (Fig. 1).

The NTA methodology, used in almost all fields of chemical analytics, was considered in numerous reviews often associated with individual types of test samples (Table 2). The techniques and methods of NTA are constantly being improved, and it makes sense to capture the current level of its general deve-
opment, which is typical for most objects of analysis. Such general characteristics of the NTA are considered in this article as a concise review. Along with the techniques of analysis, sample preparation, and processing of information, the issues of the performance of NTA and the level of its errors are discussed. However, there is no way to estimate such errors reliably in most cases. The correctness of the NTA results can often be judged if only modern good practice has been implemented. This means that the corresponding work includes the main necessary stages of analysis, instruments, software, and databases, which are briefly considered in this review.

The review deals with low-molecular weight compounds; the most significant publications of recent years, containing references to previous studies, are predominantly cited. We note new guidelines for carrying out NTA in specific scientific fields [16, 18, 21]; these publications are of particular significance in starting work in this analytics field.

**SUMMARY OF NON-TARGET ANALYSIS**

In the broadest sense, NTA includes the determination of all components of a sample the composition of which is unknown to the analyst prior to the experiment. Prioritization (see below) may correct the number and nature of the compounds to be determined. The NTA stages formally coincide for most test samples (Fig. 2). The result of the analysis covers compounds found and identified in the sample. Quantitative determinations are not usually associated with NTA but may follow the identification of analytes. Absolute quantitation requires suitable analytical methods and reference materials. Semiquantitative determination is possible in using compounds with a similar structure as standards or when assessing the relative sensitivity of techniques to different compounds [12].

The main techniques of NTA are various variants of chromatography–mass spectrometry (CMS). A combination of gas chromatography and electron ionization mass spectrometry (GC–EI-MS) using one quadrupole mass analyzer is indispensable for volatile compounds. In the determination of nonvolatile analytes, including most biologically important compounds, high-performance or ultra-performance liquid chromatography (HPLC/UPLC) is used in combination with high-resolution tandem mass spectrometry (HRMS³); electrospray is used as an
ionization device (electrospray ionization, ESI). Other versions of mass spectrometry are also helpful but, as a rule, less advantageous. Recently, a valuable addition to CMS, ion mobility spectrometry, has been developed [22]. In the detection and structure elucidation of new compounds not described in the scientific publications (“unknown unknowns”), it is advisable to separate them from the test samples and use NMR spectroscopy for identification (in addition to MS).

Non-target chemical analysis is predominantly concerned with organic (bio-organic) compounds; this subject of NTA is mainly considered in the review. Nevertheless, it may also include the determination of elements if the molecular form of their presence is considered (speciation analysis); in this case, a combination of HPLC and inductively coupled plasma mass spectrometry (ICP–MS) is used [23]. Other types of “inorganic” NTA are significant, rather, in a historical sense.

**PERFORMANCE OF NON-TARGET ANALYSIS**

We propose to define the indicator in the title of the section as the proportion of compounds detected and/or correctly identified by NTA. For screening, the performance of NTA can be expressed by the value of $P_s$ and the following equation:

**Table 2. Main fields of non-targeted chemical analysis and the corresponding objects of analysis**

| Field                                | Object/matrix               |
|--------------------------------------|-----------------------------|
|                                      | air, water, soil, sediments, waste, garbage | food and beverages | urine, blood, tissues, and other biomatrices | samples of plant and animal origin |
| Environmental objects [3–9]          | ++                          | +                | +              | +                        |
| Food, food raw materials, beverages [10–12] | +              | ++              | +              | +                        |
| Pharmaceuticals, drugs, doping [13–15]* | +            | +                | ++             | +                        |
| Toxicology [13–15]*                  | +                          | +                | ++             | +                        |
| Metabolomics [16–18]                 | +                          | +                | ++             | +                        |
| Natural compounds [19, 20]           | +                          | +                | ++             | ++                       |

++, the main matrix; +, more rare analysis.
* Combined in forensic medicine.

![Fig. 2. Schematic representation of NTA. (a) Sample; analytes (c) are separated by extraction (b) and injected into the chromatograph–mass spectrometer (d). The operation of the instrument is controlled by a computer (e) equipped with numerous programs (f) and linked to various databases (g). Analytical standards (h) are provided for the final identification of analytes.](image-url)
The screening result consists of the detected components of the test sample and of a preliminary conclusion about their structures. A “full-fledged” NTA, required to identify the components of the sample as completely and reliably as possible, is characterized by the $P_{N\text{TA}}$ value,

$$ P_s(\%) = 100 \times \frac{\text{Number of detected compounds}}{\text{Number of sample components}}. \quad (1) $$

In general, it is difficult to estimate the $P_s$ and $P_{N\text{TA}}$ values, because the composition of the test sample and the number of the components (the denominator of the fractions) are unknown. The numerator of these fractions is easier to calculate, although the reliability of the corresponding identification may be unclear (see below). $P_{N\text{TA}}$ can be estimated in a model situation, for example, using special artificial mixtures of compounds of interest for the field of analysis under consideration [24, 25] and estimating the number of the identified and unidentified compounds. The value of $P_{N\text{TA}}$, in this case, is the indicator of true positive results [26, 27], which in test experiments with mixtures produced by the U.S. Environmental Protection Agency did not exceed 65% for methods based on HPLC–HRMS² [24].

**PRIORITIZATION**

With the NTA priorities (Table 1) formulated, a search is made for the relevant compounds of the greatest value and specific properties/features. The number of the determined compounds is reduced, which decreases the denominator of Eqs. (1) and (2) and, consequently, increases the value of $P$.

The simplest NTA is implemented if the presence of specific compounds in the sample is expected (suspect analysis, Table 1) [4, 5, 7]. Such an analysis can be close to the target, mainly because some modern methods include the determination of hundreds of analytes [28]. The suspects include (a) compounds present in the lists of priority environmental pollutants, for example, the NORMAN database [29], (b) other toxic compounds including those for which dangerous properties were predicted, (c) substances frequently found in specific matrices, and (d) compounds related to the priority ones, similar in physicochemical properties, including their transformation products (for example, metabolites).

The difference between suspect analysis and “real” NTA is that, in the former case, the number of candidates for identification (possible analytes) is relatively small (corresponding to the selected priorities). These often include common compounds with their representative reference chromatographic and mass spectrometric information, which facilitates their identification. For less popular or less expected compounds, the volume of reference data is generally not so significant or fragmentary, so searching the corresponding information requires access to the most extensive databases.

The other options for prioritization may include effect-directed analysis (Table 1). In this case, the sample part is controlled (including separation by fractionation) and analyzed that exhibits biological activity. Another example is the preferred determination of organochlorine and/or organobromine compounds, many of which are hazardous substances. Their presence is easily detected by a characteristic isotope pattern in the mass spectra. Finally, a trivial priority should be noted—the main components of samples with the most intense signals in the chromatograms and mass spectra [8]. A related methodology in searching for new biologically active compounds is called “dereplication.” Here, the main components of samples (as a rule, known compounds) were identified first, and then they were excluded from consideration [19].

Prioritization can cover both analytes and also test samples themselves. Some samples of common origin can be skipped, and only those that demonstrate (a) intense analytical signals in the expected sections of chromatograms and mass spectra and (b) an increase in the intensities of such signals in a series of samples can be analyzed in detail.

**PREANALYTICAL PROCEDURES**

The quality of these procedures, that is, sampling, transportation, and storage of samples, must exclude any loss of the desired analytes (leading to false-negative results (FN), Table 3) or the contamination of a sample with foreign substances (false-positive results, FP). The necessary standard requirements for such procedures are formulated [16, 18, 21, 30], although it is difficult to assert that they are met in all current studies.
The sampling technique affects the representation of different analytes in the sampled substance and the occurrence of false results (Table 3). In a passive sampling of air or water, the ratio between the analytes may depend on the type of the adsorbent. In single water samples taken by a grab, there is no such discrimination between the components of mixtures [7].

Foreign compounds (resulting in false-positive determinations) in the test sample are relatively easy to detect in analyzing blank samples (prepared from matrices, materials, solvents, reagents, etc. [16]). The identification of false negatives is more complicated. It is necessary to establish the loss/decomposition/transformation of analytes before or during analysis using internal standards, which is challenging in NTA because analytes are unknown. It is recommended to use additives in the test samples, that is, one compound from each class (group) of expected substances [21] or a group representing a particular range of physicochemical characteristics of analytes, for example, the \( n \)-octanol–water distribution coefficient \( K_{ow} \) [8]. Similarly, test mixtures are prepared to ensure the quality of the analysis [18].

| Procedure, method | False-positive* | False-negative |
|-------------------|-----------------|----------------|
| Preanalytical procedures | Poor adsorption during sampling Transformation of analytes | |
| Sample preparation | Loss of analytes, inefficient extraction, incomplete targeted chemical transformations (derivatization, deconjugation) | |
| Chromatography | Matching retention characteristics of different compounds | Poor separation, weak signals, variation in retention characteristics, evaporation of volatile analytes in HPLC/UPLC, inability to determine nonvolatile analytes in GC, no reference data available |
| Mass spectrometry | Coincidence of the mass spectra of different compounds | Poor ionization, matrix effects, multiple forms of ions containing analyte molecules, weak signals, errors in ion masses, poor reproducibility of mass spectra, overlapping spectra of different analytes, lack of reference data |

* General sources of false-positive results: components of the materials and excipients used; analyte transformation products, the possibility of which cannot be ruled out at all stages of NTA.

The sampling technique affects the representation of different analytes in the sampled substance and the occurrence of false results (Table 3). In a passive sampling of air or water, the ratio between the analytes may depend on the type of the adsorbent. In single water samples taken by a grab, there is no such discrimination between the components of mixtures [7].

Foreign compounds (resulting in false-positive determinations) in the test sample are relatively easy to detect in analyzing blank samples (prepared from matrices, materials, solvents, reagents, etc. [16]). The identification of false negatives is more complicated. It is necessary to establish the loss/decomposition/transformation (including biotransformation) of analytes before or during analysis using internal standards, which is challenging in NTA because analytes are unknown. It is recommended to use additives in the test samples, that is, one compound from each class (group) of expected substances [21] or a group representing a particular range of physicochemical characteristics of analytes, for example, the \( n \)-octanol–water distribution coefficient \( K_{ow} \) [8]. Similarly, test mixtures are prepared to ensure the quality of the analysis [18].

**SAMPLE PREPARATION**

This necessary procedure is the separation of analytes from the collected samples, accompanied by the preconcentration of the former substances. Most analyzes of liquid (aqueous) samples use liquid—liquid extraction or solid-phase extraction (SPE). Different analytes have nonidentical physicochemical properties; therefore, varied selectivity of these procedures and further losses of analytes are inevitable. Chemical reactions (decomposition, dehydration, oxidation, polymerization) unwittingly accompanying sample preparation procedures lead in some cases to false results. The incompleteness of the chemical procedures included in the sample preparation process also contributes to them (Table 3) [16].

Water—methanol and water—acetonitrile mixtures were used as extractants for blood, blood plasma, and other biological matrices, ensuring close recoveries of many metabolites. By changing the composition of extractants, one could alternately extract and then determine the components of biological matrices with different polarities [18]. The selection of ternary systems as extractants for blood plasma was also promising; for example, using an acetonitrile—isopropanol—water mixture for extracting polar and medium-polar analytes in the range of 25 orders of magnitude of their \( K_{ow} \) values [17]. In biological matrices, rapid transformations (enzymatic reactions) are inevitable, which in one way or another affect the recovery of the analyte [18]. These processes are stopped by freezing samples or by adding cold solvents.

Homogenization, another procedure implemented at the beginning of the preparation of solid samples, can affect the analysis results if a representative sample is not formed [21]. In these cases, samples are ground in special mills, with special measures taken to suppress the activity of enzymes to the components of samples of plant and animal origin. Analytes are extracted from such ground samples, for example, using the popular QuEChERS procedure [11, 12].

In some cases, at relatively high concentrations of analytes, a direct analysis (direct injection into a chromatograph) of liquid samples, such as wastewater and urine, can be performed without significant loss of many analytes. In such cases, the preliminary dilution of the initial samples (dilute-and-shoot methods) is often practiced [8]. Nevertheless, false-negative results remain very likely as well. A comparison of such procedures and SPE demonstrated that in the latter case, a more significant number of polar compounds
could be detected in water than when relatively large volumes (0.5–1 mL) of water sample are directly injected into a liquid chromatograph [9].

**CHROMATOGRAPHY**

As in target analysis, chromatography is the primary separation method in NTA, although the use of capillary electrophoresis [31] and ion mobility spectrometry (in addition to liquid chromatography) [32] can be successful in some situations. The incomplete chromatographic separation of analytes distorts the retention parameters and mass spectra, leading to errors in identifying major analytes and the loss of minor components of the test mixtures (Table 3). Complex samples containing tens, hundreds, thousands of components are the most difficult to analyze because the corresponding chromatograms contain many peaks, a significant part of which overlap to some extent (Fig. 3). Therefore, the deconvolution procedure is implemented in chromatography–mass spectrometry, that is, separation into individual chromatograms corresponding to individual components of mixtures, using their mass spectrometric signals (Fig. 4; for software, see [16]).

Two main types of chromatography—gas and liquid—are used to determine volatile and nonvolatile compounds, respectively.

In GC, nonvolatile compounds are somehow lost. Derivatization (silylation and other reactions) can improve the chances for their determination, although incomplete derivatization or evaporation of the most volatile derivatives can lead to false-negative results [16]. Two-dimensional gas chromatography (GC–GC) results in a more efficient separation of mixtures and solves some basic analytical problems of qualitative analysis II [27], namely, identification (characterization, authentication, etc.) of the test samples themselves [35, 36]. The conventional combination of GC–MS1 has been supplemented in recent years by GC–GC with more advanced versions of mass spectrometry: HRMS and MS2 [35, 36].

Reversed-phase (RP) liquid chromatography is conventionally used to determine nonvolatile compounds, including a version that ensures better separation (RP–UPLC). In recent years, hydrophilic interaction liquid chromatography (HILIC) has gained popularity, demonstrating better separation characteristics in the case of highly polar analytes compared to RP–HPLC/UPLC, where these analytes are not retained by the column [17, 18]. Nevertheless, HILIC is characterized by more significant variations in retention times and a more frequent manifestation of matrix effects, demonstrating the ionization suppression of individual analytes in a mass spectrometer, which is typical for CMS [18]. This can lead to false-negative results due to the poor prediction of the retention times or elution order of analytes and insignificant mass spectrometric signals, respectively.
MASS SPECTROMETRY

A mass spectrometer is the primary instrument for identifying analytes in NTA [26, 27]. Determining “known unknowns” volatile compounds is a relatively simple task solved by comparing experimental EI mass spectra with corresponding reference/library spectra. The spectra of this type are pretty well reproduced and recorded for the vast majority of common volatile compounds (see below).

For nonvolatile compounds, the greatest possibilities of NTA are ensured by high-resolution mass spectrometry and tandem mass spectrometry, primarily in the V/UPLC–ESI-HRMS2 combination. In this tandem, the first mass spectrometer is a quadrupole mass analyzer or an ion trap, and the second is a time-of-flight mass analyzer or a Orbitrap ion trap. Between two mass analyzers, there is a cell of collisions of ions (precursor ions) with a gas target, where fragment ions are formed necessary for identifying analytes.

Note that the highest mass resolution is achieved in ion cyclotron resonance instruments [37], which are inaccessible to most chemical laboratories due to their high cost. Orbitrap technique and methodology have recently been referred to as “high-resolution accurate mass spectrometry” [38]. The use of mass spectrometers with the usual “unity” resolution can, to some extent, lead to success [39] in solving less complex NTA problems. Various analytical methods are based on a combination of different variants of chromatography (RP or HILIC, HPLC or UPLC) and mass spectrometry (different mass analyzers, positive or negative ions, etc.).

Those high-resolution instruments ensure the accuracy of measuring the masses (m/z values) of precursor ions at the parts-per-million level, while that of fragment ions is somewhat worse. Precursor ions are sampled from the full set of their peaks in the MS1 spectra (data-independent acquisition) or can be preset for the expected compounds (data-dependent acquisition). The superposition of some mass peaks of precursor ions may require their deconvolution and separation of MS2 spectra [40]. One should keep in mind that, in the ESI process, analyte molecules form both protonated molecules of the main isotopic forms and adduct ions, including cationized molecules and charged particles of other isotopic forms (“isotopologues”). The exclusion of the latter ions from consideration (“componentization” [8]) significantly simplifies the processing of chromatography–mass-spectrometric data. For identification, one often resorts to comparing experimental spectra with reference ones. Instruments incorporating conventional ion traps result in MS2 spectra less comparable with spectra recorded using other types of tandem mass spectrometers [27, 41].

The nature of false results, which may be a consequence of the use of MS, is reflected in Table 3 (see also [18]).
The trueness and reliability of identification are determined by the compliance of the obtained data with particular criteria. In target analysis, these criteria based on chromatographic and mass spectrometric data are well defined [26, 27, 42]. In NTA, the situation is much less clear.

The most reliable identification with obtaining true-positive results is achieved by (a) the matching of chromatographic and mass peaks in the co-analysis of a sample containing the analyte and the corresponding analytical standard [26, 27, 42, 43]. (b) The identification based on the similarity of the experimental and reference mass spectra and chromatographic retention parameters is in second place in terms of reliability. The similarity can be expressed by its conventional index (see below) or by the degree of closeness of the intensities of several main peaks of the same (differing only within the error) masses. The trueness of identifying volatile compounds from the libraries of reference EI mass spectra is generally ~80%; for nonvolatile analytes and ESI-MS mass spectra, the situation is more uncertain, the proportion of true-positive results varies widely [26, 27, 41].

The second identification method (b) can be as reliable as the first option (a) but under certain conditions: the values of the indicated quantities must be obtained under similar experimental conditions (the same types and models of instruments, similar modes of data acquisition), and these values must be unique for the identified compound. Other methods of identification (interpretation of data, comparison with predicted spectra and chromatographic characteristics [27, 43, 44]) are less reliable but, apparently, applicable to the selection of candidates for identification.

If several such candidate compounds have common substructures, the identification at this stage is called group identification [26, 27]. Such a group of compounds can be identified by “molecular networks,” that is, graphs of structurally similar compounds constructed from similar mass spectra [45, 46].

The term “identification level” is popular in English publications (Table 1). Its meaning is interpreted in detail in [3] and clarified in Table 4 with our comments. The identification level correlates with the confidence of identification (proportion of true-positive results).

There are particular quantitative indicators of reliability [26, 27, 43]. These include α- and β-criteria for accepting statistical hypotheses in considering identification as a procedure for testing them. The concept of identification points, matching values of the measured mass spectrometric and chromatographic quantities, taking into account their different significance, is also popular. The indices of the similarity of mass spectra, such as the point function or the probability of their matching, can also be considered as a particular measure of reliability. Recently, a general reliability scale has been proposed that considers identification levels, the degree of matching of retention characteristics, and the number of identification points [47].

### INFORMATICS

Several types of information products are indispensable for carrying out NTA.

**Mass spectral libraries** are primary sources of reference information. The largest of them are presented in Table 5. The situation is relatively good with the EI-
MS spectra of volatile compounds: the libraries contain most of the known and most important compounds of this class. Libraries of MS2 spectra and even HRMS2 spectra, predominantly related to nonvolatile analytes, that is, most biologically essential compounds, began to be created much later. The mass spectra of many compounds are not available in the libraries, and the MS2 spectra as a whole are not reproduced well enough; they depend on the type of tandem mass spectrometer and the collision energy leading to the fragmentation of precursor ions [26, 27, 41]. The need to upgrade and improve the quality of tandem mass spectra libraries is widely recognized [41, 43].

Retention parameter databases. The NIST database includes 447285 gas chromatographic retention indices (RI) for 139693 compounds [49]. In HPLC/UPLC, the concept of retention indices is less applicable than in GC, but related works and methods for estimating these quantities are emerging [55].

Chemical databases. The largest databases containing information about chemical compounds are listed in Table 6. The information contained in them supplementing the experimental CMS data is called a priori or meta information. It is practical and even necessary in selecting and ranking compounds, candidates for identification.

In using these databases, three circumstances should be considered. First, the ChemSpider database allows searching for molecular formulas using the experimentally found masses of molecular ions (MS) or protonated (cationized) molecules (MS2). Second, these databases enable estimation of the popularity/abundance of chemical compounds by the number of sources of information about them, the amount of such meta information, etc. [1, 2, 27]. In testing identification hypotheses [26, 27], starting with the most prevalent compounds (ceteris paribus) makes sense. Third, the information available in the database on the preparation and properties of chemical compounds and their presence in various objects is also helpful in selecting candidates for identification.

**Table 5. Main libraries of mass spectra of low-molecular weight compounds**

| Database                                           | Amount of data | Comments                                      |
|----------------------------------------------------|----------------|-----------------------------------------------|
| Wiley Registry 12th, EI-MS1 [48]*                  | 817290         | 668435 Volatile compounds                    |
| NIST 20, EI-MS1 [49]**                             | 350643         | 306869 Volatile compounds                    |
| NIST 20, MS2 [49]***                               | 1320389        | 31000 MS2, metabolites predominate           |
| METLIN [50]                                        | >500000        |                                               |
| MassBank of North America (MONA) [51]              | 691548         | 226303 MS2 spectra of biologically active compounds predominate. In silico mass spectra are present |
| mzCloud [53]                                       | 8 065 135      | 19 109 MSn, pharmaceuticals, metasimulators  |
| The Global Natural Product Social Molecular Networking (GNPS) [45] | 221083         | 18 163 MS2, natural compounds. Data of 2016 |
| MassBank [52]                                      | 86 576         | 16 537 MS1, MSn, volatile and nonvolatile compounds |
| Human Metabolome Database (HMDB) [54]              |                |                                               |
| — EI-MS1                                           | 1763           | 1220 Metabolites; in silico mass spectra predominate. Data of 2018 |
| — MS2                                              | 279 972        | 1249                                          |

* May be supplemented by specialized smaller libraries.
** 111 788 compounds were characterized by mass spectra and retention indices.
*** Including high-resolution mass spectra of 27 840 compounds.
calculations [59]. Nevertheless, methods for predicting mass spectra are rapidly developing, and one can expect an improvement in their efficiency. Even now, this methodology can be used in NTA for (a) calculating the mass spectra of candidates for identification, which are selected by the mass of precursor ions and search in chemical databases, and (b) discarding the most dissimilar in silico spectra.

Despite an extensive collection of experimental GC retention indices (see above) is available, some studies are underway to predict these indicators as well, at least to test the effectiveness of computational methods. The application of the machine learning methodology enables calculating their values with satisfactory accuracy [60]. Similar prediction algorithms have been used for calculating HPLC retention indices; the obtained data did not have substantial independent significance for identification but still improved its results based on the prediction of mass spectra [55].

In the case of liquid chromatography (LC), relative retention times are more often predicted rather than retention indices [61–63]. Some prediction results are rather satisfactory. Predictions of relative retention indices for 80000 compounds included in the METLIN database (Table 5) show that in 70% of cases, the corresponding analytes are among the three most probable candidates for identification [62].

**Various software.** Chromatography—mass spectrometry data processing programs are indispensable in analyzing complex samples that yield numerous chromatographic peaks. Good NTA practice implies carrying out the following procedures in automatic mode [4, 6, 8, 16, 40, 64, 65]:

- Deconvolution of chromatographic peaks with separation of signals of individual components and their mass spectra;
- Filtration of mass peaks to remove background, weak peaks, and outliers;
- Annotations of peaks in mass spectra: assigning mass values and even chemical structures of corresponding ions derived from exact masses ($m/z$ values) and isotopic pattern, to peaks;
- Comparison of mass spectra and retention characteristics with the corresponding reference data; evaluation of their similarity;
- Formation of in-house libraries of mass spectra;
- Mutual adjustment of different chromatograms in retention times and/or masses of ions of reference compounds for comparing different samples.

Commercially produced equipment is equipped with appropriate software; programs are also supplied by other companies or organizations [16, 64].

Algorithms and software for multivariate statistical analysis (chemometry) should also be mentioned, which help group and classify studied samples, for example, food samples, based on NTA results [10, 12, 66].

**CONCLUSIONS**

Numerous innovations in analytical instrumentation and computer science have made it possible to determine many dozens, hundreds, and even thousands of organic (bioorganic) compounds simultane-

---

**Table 6. The largest chemical databases**

| Database                              | Number of compounds | Free access | Comments                                                                 |
|--------------------------------------|--------------------|-------------|--------------------------------------------------------------------------|
| Chemical Abstracts Service (CAS) [56] | 182 million substances, including individual compounds* and mixtures** and 68 million biopolymers (peptides, proteins, etc.) | —           | Complete bibliographic information on chemistry and related sciences and related factual data |
| PubChem [57]                          | 110 million individual compounds*, 271 million substances*,** | +           | Biologically active compounds, properties, biological activity, toxicity, etc. Compilation from 795 sources |
| ChemSpider [1, 58]                    | ≥80–90 million structures (individual compounds*) | +           | The same as above, shorter information about substances, but with more search options, including ion masses. Compilation from 279 sources |

* Single-component substances according to EU terminology [28].
** Multicomponent substances and substances of unknown or variable composition according to EU terminology [28].
ously, including those unknown to the analyst before the experiment, in the most complex matrices (biological and medical objects, food, environmental objects, etc.). Among analytical instruments, high-resolution tandem mass spectrometers coupled with chromatographs, which have entered analytical laboratories in a significant number in the last 10–15 years, have become paramount. Simultaneously, significant progress has been made in informatics, which has led to the emergence of large databases and new software for processing CMS data. Advances in instrumentation and computer science materialized into an explosive growth of works in the field of NTA.

Two aspects of these publications should be highlighted. Many relevant subject reviews were devoted to particular methods and/or objects of analysis. A review of these works makes it possible to outline the general practice of NTA, to identify its standard techniques, the use of specific methods of sample preparation, analysis, extraction, and processing of information. The other part of the discussed scientific publications is that they reflect a new, fairly complete analysis of a variety of objects in all their geographical, biological, natural, industrial, etc. diversity. Such studies using the NTA methodology are ongoing, and new evidence is expected regarding the previously unknown composition of substances. The new data obtained in recent years and those expected, probably, need a broad generalization, which is of interest to various chemists.

CONFLICT OF INTEREST
The author declares that they have no conflicts of interest.

OPEN ACCESS
This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit ttp://creativecommons.org/licenses/by/4.0/.

REFERENCES
1. Milman, B.L. and Zhurkovich, I.K., Molecules, 2021, vol. 26, no. 8, p. 2394.
https://doi.org/10.3390/molecules26082394

2. Mil’man, B.L. and Zhurkovich, I.K., Analitika, 2020, vol. 10, no. 6, p. 464.
https://doi.org/10.22184/2227-572X.2020.10.6.464.469

3. Schymanski, E.L., Singer, H.P., Slobodnik, J., Ipolyi, I.M., Oswald, P., Krauss, M., Schulze, T., Heglund, P., Letzel, T., Grosse, S., Thomaidis, N.S., Bletsou, A., Zwiener, C., Ibanez, M., Portolays, T., De Boer, R., Reid, M.J., Onghena, M., Kunkel, U., Schulz, W., Guillou, A., Noyon, N., Leroy, G., Bados, P., Bogiali, S., Stipančič, D., Rostkowski, P., and Hollender, J., Anal. Bioanal. Chem., 2015, vol. 407, no. 21, p. 6237.
https://doi.org/10.1007/s00216-015-8681-7

4. Hollender, J., Schymanski, E.L., Singer, H.P., and Ferguson, P.L., Environ. Sci. Technol., 2017, vol. 51, no. 20, p. 11505.
https://doi.org/10.1021/acs.est.7b02184

5. Ccanccapa-Cartagena, A., Pico, Y., Ortiz, X., and Reiner, E.J., Sci. Total Environ., 2019, vol. 687, p. 355.
https://doi.org/10.1016/j.scitotenv.2019.06.057

6. Ljoncheva, M., Stepišnik, T., Džeroski, S., and Kosjek, T., Trends Environ. Anal. Chem., 2020, vol. 28, e00099.
https://doi.org/10.1016/j.teac.2020.e00099

7. Menger, F., Gago-Ferrero, P., Wiberg, K., and Ahrens, L., Trends Environ. Anal. Chem., 2020, vol. 28, e00102.
https://doi.org/10.1016/j.teac.2020.e00102

8. Schulze, B., Jeon, Y., Kasierzok, S., Heffernan, A.L., Dewapriya, P., O’Brien, J., Ramos, M.J.G., Gorji, S.G., Mueller, J.F., Thomas, K.V., and Samanipour, S., TrAC, Trends Anal. Chem., 2020, vol. 133, 116063.
https://doi.org/10.1016/j.trac.2020.116063

9. Kutlucinar, K.G. and Hann, S., Electrophoresis, 2021, vol. 42, no. 4, p. 490.
https://doi.org/10.1002/elps.202000256

10. Riedl, J., Esslinger, S., and Faulh-Hassen, C., Anal. Chim. Acta, 2015, vol. 885, p. 17.
https://doi.org/10.1016/j.aca.2015.06.003

11. Knolhoff, A.M. and Croley, T.R., J. Chromatogr. A, 2016, vol. 1428, p. 86.
https://doi.org/10.1016/j.chroma.2015.08.059

12. Fisher, C.M., Croley, T.R., and Knolhoff, A.M., TrAC, Trends Anal. Chem., 2021, vol. 136.
https://doi.org/10.1016/j.trac.2021.116188

13. Chen, C., Wohlfarth, A., Xu, H., Su, D., Wang, X., Jiang, H., Feng, Y., and Zhu, M., Anal. Chim. Acta, 2016, vol. 944, p. 37.
https://doi.org/10.1016/j.aca.2016.09.034

14. Oberacher, H. and Arnhard, K., TrAC, Trends Anal. Chem., 2016, vol. 84, p. 94.
https://doi.org/10.1016/j.trac.2015.12.019

15. Mollerup, C.B., Dalsgaard, P.W., Mardal, M., and Linnet, K., Drug Test. Anal., 2017, vol. 9, no. 7, p. 1052.
https://doi.org/10.1002/dta.2120

16. Mastrangelo, A., Ferrarini, A., Rey-Stolle, F., Garcia, A., and Barbas, C., Anal. Chim. Acta, 2015, vol. 900, p. 21.
https://doi.org/10.1016/j.aca.2015.10.001
51. MONA – MassBank of North America. https://mo-na.fiehnlab.ucdavis.edu/spectra/statistics?tab=0. Accessed June 7, 2021.
52. MassBank. https://massbank.eu/MassBank/Contents. Accessed June 7, 2021.
53. mzCloud. http://www.mzcloud.org. Accessed June 7, 2021.
54. The human metabolome database (HMDB). https://hmdb.ca. Accessed June 7, 2021.
55. Samaraweera, M.A., Hall, L.M., Hill, D.W., and Grant, D.F., Anal. Chem., 2018, vol. 90, no. 21, p. 12752. https://doi.org/10.1021/acs.analchem.8b03118
56. CAS. http://www.cas.org/about/cas-content. Accessed June 7, 2021.
57. PubChem. https://pubchem.ncbi.nlm.nih.gov. Accessed June 7, 2021.
58. ChemSpider. http://www.chemspider.com. Accessed June 7, 2021.
59. Krettler, C.A. and Thallinger, G.G., Briefings Bioinf., 2021, vol. 22, no. 6, bbab073. https://doi.org/10.1093/bib/bbab073
60. Matyushin, D.D. and Buryak, A.K., IEEE Access, 2020, vol. 8, p. 223140. https://doi.org/10.1109/ACCESS.2020.3045047
61. McEachran, A.D., Mansouri, K., Newton, S.R., Beverly, B.E., Sobus, J.R., and Williams, A.J., Talanta, 2018, vol. 182, p. 371. https://doi.org/10.1016/j.talanta.2018.01.022
62. Domingo-Almenara, X., Guijas, C., Billings, E., Montenegro-Burke, J.R., Uritboonthai, W., Aisporna, A.E., Chen, E., Benton, H.P., and Siuzdak, G., Nat. Commun., 2019, vol. 10, 5811. https://doi.org/10.1038/s41467-019-13680-7
63. Witting, M. and Bocker, S., J. Sep. Sci., 2020, vol. 43, nos. 9–10, p. 1746. https://doi.org/10.1002/jssc.202000060
64. Kind, T., Tsugawa, H., Cajka, T., Ma, Y., Lai, Z., Mehta, S.S., Wohlgemuth, G., Barupal, D.K., Showalter, M.R., Arita, M., and Fiehn, O., Mass Spectrom. Rev., 2018, vol. 37, no. 4, p. 513. https://doi.org/10.1002/mas.21535
65. Helmus, R., Ter, LaakT.L., Van Wezel, A.P., De Voogt, P., and Schymanski, E.L., J. Cheminf., 2021, vol. 13, no. 1, 1. https://doi.org/10.1186/s13321-020-00477-w
66. Cavanna, D., Righetti, L., Elliott, C., and Suman, M., Trends Food Sci. Technol., 2018, vol. 80, p. 223. https://doi.org/10.1016/j.tifs.2018.08.007

Translated by O. Zhukova