# Analysis and Purification of Bioactive Natural Products: The AnaPurNa Study

Guido F. Pauli,*[†,‡] Shao-Nong Chen,[†] J. Brent Friesen,[†,§] James B. McAlpine,[†,‡] and Birgit U. Jaki[†,‡]

[†]Department of Medicinal Chemistry and Pharmacognosy and [‡]Institute for Tuberculosis Research, College of Pharmacy, University of Illinois at Chicago, Chicago, Illinois 60612, United States

[§]Physical Sciences Department, Rosary College of Arts and Sciences, Dominican University, River Forest, Illinois 60305, United States

**S** *Supporting Information*

**ABSTRACT:** Based on a meta-analysis of data mined from almost 2000 publications on bioactive natural products (NPs) from >80 000 pages of 13 different journals published in 1998−1999, 2004−2005, and 2009−2010, the aim of this systematic review is to provide both a survey of the status quo and a perspective for analytical methodology used for isolation and purity assessment of bioactive NPs. The study provides numerical measures of the common means of sourcing NPs, the chromatographic methodology employed for NP purification, and the role of spectroscopy and purity assessment in NP characterization. A link is proposed between the observed use of various analytical methodologies, the challenges posed by the complexity of metabolomes, and the inescapable residual complexity of purified NPs and their biological assessment. The data provide inspiration for the development of innovative methods for NP analysis as a means of advancing the role of naturally occurring compounds as a viable source of biologically active agents with relevance for human health and global benefit.

## INTRODUCTION

**Preface.** Natural product (NP) research is a demanding science, requiring an in-depth knowledge of many aspects of organic, analytical, and biological chemistry, including separation science, spectroscopy, biosynthesis, and pharmacology, as well as the biology and taxonomy of the investigated phyla. Nonetheless, most contemporary practitioners would agree that the role of this discipline within biomedical science has declined, as evidenced by its present abandonment by most large pharmaceutical companies in their search for new chemical entities to provide new drug discovery leads. This review takes a comprehensive look at the current practice of NP research, with the aim of pinpointing potential areas where the practitioners might improve the overall efficiency of this type of work. By focusing on NP chemistry as one fundamental aspect of NP research, it might be possible to recognize patterns that impact the bigger picture and identify opportunities that otherwise would go unnoticed. This review intends to stimulate discussion and inspire the development of new approaches to yield more rapid results and a greater number of new chemical entities discovered, and thereby promote the future role of NP research in interdisciplinary programs.

**Role and Sourcing of Natural Products.** A series of excellent articles, coauthored by G. M. Cragg, D. J. Newman, and colleagues,[1−5] has documented the invaluable role of NPs in drug discovery. Underlying evidence came from an extensive meta-analysis of the primary literature of all drugs, in or completing FDA-approved studies within a set time frame and classifying them according to their origin as NPs, inspired by NPs, analogues of these two classes, or from non-NP sources.

These analyses have indicated that a high proportion of new drugs approved in Western countries in recent decades are, in some manner, connected to NPs. As primordial biosynthetic pathways endow Nature's library of chemicals with an evolutionary advantage over man-made chemicals, NP libraries are keyed to Nature's biochemistry and diversity and, thus, continue to be an attractive source[6] for new bioactive agents, for both therapeutic and diagnostic uses. Moreover, the chemical diversity in NPs is tied intrinsically to the complexity of the metabolome contained in the source material.

Ultimately, both the discovery and the resupply of bioactive NPs depend on the availability of preparative-scale analytical methods having the capability of resolving the complex primary and secondary metabolomic mixtures that are typically isolated from the source organism, yielding a purified NP (NP in Figure 1), and eventually providing a well-characterized NP as a single chemical entity (SCE; Figure 1). It should be noted that, in the practice of NP chemistry research, a purified NP does not necessarily represent an SCE, but may only have been purified to the degree necessary, e.g., for structure elucidation or identification. A SCE may be defined as a substance for which all chemical, physical, and biological characteristics can be attributed to a single molecular structure. Accordingly, a NP becomes a SCE only after its singleton character has been demonstrated (high-purity NP). This is in line with practice for SCEs that are used and regulated as drugs: their purity plays a pivotal role in all pharmacopoeias worldwide. This topic has
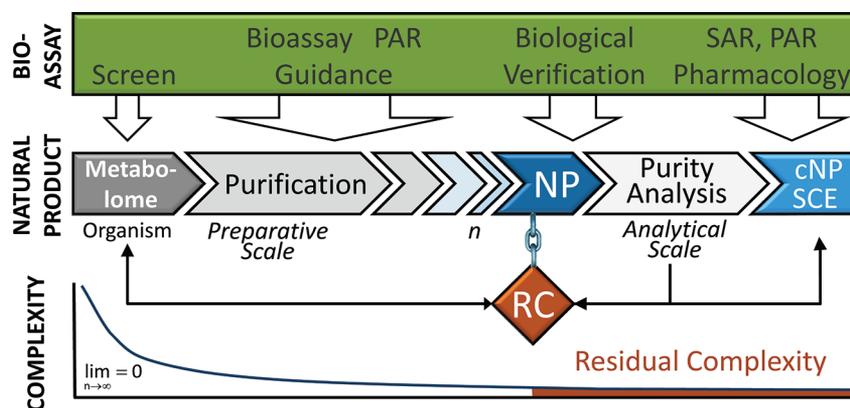
**Figure 1.** Progression of NP purification from a metabolomic mixture. The process involves repeated ($n$-times) preparative- and analytical-scale separation and, depending on the methods and $n$, results in a NP that is linked to varying residual complexity (RC), reflecting both its metabolomic heritage and the purification protocol. Subsequent analytical characterization including purity assessment is required to generate a fully quality controlled NP (cNP) or single chemical entity (SCE, Figure 2). The nearly 2000 publications evaluated employ bioassays to address screening of crude NPs, bioassay-guided fractionation, biological assessment of purified NPs, and detailed pharmacological investigation of, for example, structure−activity and structure−purity relationships (SARs and PARs, respectively; see text and Figure 2).

recently received global public attention when an isosorbide-5-mononitrate preparation containing pyrimethamine as an impurity caused the death of more than 100 patients in Pakistan.[7] This tragedy demonstrates the importance of purity as a parameter for the safety of medicines, but also exemplifies that purity should never be ignored and always be part of quality control of drugs − and NPs.

The majority of pure NPs represent rare chemicals of extremely limited supply. Frequently, particularly in the case of newly reported structures, such compounds are also unique commodities and are only immediately available from a single source, namely, the original investigators, or by re-isolation. Practitioners of NP chemistry can generally observe additional factors that contribute to the exclusivity of NP samples: (i) their consumption in the bioassay systems of contemporary NP research programs; (ii) a general trend to smaller sample sizes, leading to smaller yields; (iii) the frequently unfavorable consistency of small-scale isolation products, and (iv) the practical challenges of handling small samples for distribution, such as precise weighing in the submilligram range.

Considering both commercial and noncommercial/academic sources and supply chains, most pure NP compounds can be traced back ultimately to crude natural materials (extracts) that require various purification steps before being considered "pure". Consequently, "pure" NPs carry a natural signature in the form of a characteristic impurity profile called residual complexity (RC), which originates ultimately from the biosynthetic cocktail(s) of the producing organism(s).[8] As a result of the authors' own experience, the often elaborate purification process potentially adds unwanted "tracer" components to the purified NP, such as sorbents, laboratory pollutants, residual solvents, or other chemicals, which can evade detection by the analytical methods used. These considerations also affect studies with a biological or pharmacological focus that utilize as tools NPs, which might be acquired from outside sources. Most such studies generally consider NPs as "fine chemicals" rather than a material derived from Nature. Exceptions may be compounds obtained by (semi)synthesis, a process typically only accomplished at an advanced discovery stage and for select NPs. Even in these instances, as minor congeners potentially can undergo the same

reaction, carryover of minor components (commonly analogues) through semisynthetic schemes has to be considered.

All of these considerations reveal NPs as being both highly sought after and hard to obtain entities. They also explain why the NP drug discovery process and the biological assessment of NPs to date are almost inevitably tied to preparative-scale analytical methods used for NP purification. The ability to purify a few milligrams of a rare NP from kilograms of a crude extract has been one of the significant skills of scientists trained in NP chemistry, pharmacognosy, and analogous disciplines and represents one of the keys to NP research.

**Approach.** In clinical research, numeric meta-analysis of literature is a well-established tool, allowing recognition of more general trends, and is used frequently to improve clinical practice. While such meta-analyses are rarely done in NP research, they can be very helpful tools for gaining new and more generalized insights. One example of such a report is the study by G. A. Cordell et al.,[9] revealing that only about 3% of some 20 000 known alkaloids have been evaluated biologically in more than five test systems, whereas 36% of alkaloids that were evaluated in 20 or more bioassays are pharmaceutically relevant. The present contribution is based on the meta-analysis of the recent literature with a focus on parameters that reflect the analysis and purification of bioactive NPs (AnaPurNa).

The production of pure NPs of controlled quality (cNPs) involves two main aspects: (a) the actual *purification* process used for NP isolation, i.e., the (semi)preparative-scale analytical method employed; (b) the assessment of the *purity* [or residual complexity (RC)] of the isolated NP, including the analytical method used for purity assessment. The aim of this review is to describe the status quo regarding both aspects, through a comprehensive assessment of the contemporary literature on bioactive NPs. The present report summarizes over a decade of data-mining activity by the authors, which involved manual screening of >80 000 pages of scientific literature during the periods 1998−1999, 2004−2005, and 2009−2010. To date, data have been extracted from nearly 2000 peer-reviewed articles, forming the foundation of this survey of analysis and purification of bioactive natural products (AnaPurNa). The framework of this study was designed at the survey onset and in a prospective fashion. Throughout the study, the literature was examined for a set of predefined parameters, which were

**Table 1. Source Journals, Time Periods, and Coverage[a] of the Survey**

| | period I | | | | period II | | | | period III | | | |
| | 1998 | | 1999 | | 2004 | | 2005 | | 2009 | | 2010 | |
| journal (group)[a] | vol.[c] | no.[c] | vol.[c] | no.[c] | vol.[c] | no.[c] | vol.[c] | no.[c] | vol.[c] | no.[c] | vol.[c] | no.[c] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| sum (year) | | 342 | | 374 | | 377 | | 125 | | 172 | | 260 |
| *Biological Pharmaceutical Bulletin* (A)[b] | 21 | 24 | 22 | 32 | | | 28 | 125 | | | 33 | 85 |
| *Chemical Pharmaceutical Bulletin* (A)[b] | 46 | 25 | 47 | 27 | 52 | 52 | | | 57 | 66 | | |
| *European Journal of Pharmacology* (B)[b] | 341−364 | 10 | 264−386 | 23 | | | | | | | | |
| *Fitoterapia* (B)[b] | 69 | 7 | 70 | 8 | 75 | 19 | | | | | 81 | 88 |
| *Journal of Asian Natural Products Research* (A)[b] | | | | | 6 | 8 | | | 11 | 58 | | |
| *Journal of Ethnopharmacology* (B)[b] | 60−64 | 11 | 65−68 | 11 | | | | | | | | |
| *Journal of Natural Products* (A)[b] | 61 | 83 | 62 | 99 | 67 | 198 | | | 72 | 64 | 73 | 194 |
| *Journal of Pharmacy and Pharmacology* (B)[b] | 50 | 13 | 51 | 11 | | | | | | | | |
| *Journal of Pharmacology and Experimental Therapeutics* (B)[b,e] | 284−287 | 3 | 288−291 | 11 | 309−311 | 8 | | | | | | |
| *Phytochemical Analysis* (B)[b] | 9 | 1 | 10 | 3 | | | | | | | | |
| *Phytochemistry* (A)[b] | 47−49 | 64 | 50−52 | 63 | 65 | 79 | | | 70 | 42 | | |
| *Phytotherapy Research* (B)[b] | 12 | 38 | 13 | 21 | | | | | | | | |
| *Planta Medica* (B)[b] | 64 | 63 | 65 | 65 | 70 | 21[d] | | | | | | |

[a]$n_{tot}$ = 1823. [b]Journals were assigned to two groups, A and B, according to the depth of reported spectroscopic information (see Methods section and main discussion for details). [c]These two columns give the journal volume number and the number of articles that fulfilled the inclusion criteria and were evaluated, respectively. [d]Only 21 articles were assessed, although about 80 articles would have fulfilled the inclusion criteria. [e]Only this journal was assessed for the period 2000−2003 ($n$ = 77; not included in $n_{tot}$).

recorded using predefined scoring and key systems. Articles also had to fulfill certain inclusion and exclusion criteria. Furthermore, a set of 15 questions to be answered was developed at the beginning of the study. These questions are addressed individually in the discussion of the observations made below.

This review is organized as follows: the methodology section describes the data-mining methodology employed as well as the journal and time coverage of the survey. Subsequent sections present the survey results as well as the numerical and statistical measures developed from these data. The next section concentrates on the following aspects: sources of purified NPs; chromatographic methodology used for NP isolation; spectroscopic methods used for NP characterization; and the role of purity and the methods used for the purity analysis of NPs. The final section of the review summarizes the findings from the perspective of potential new approaches to the analysis of NP complexity and the achievement of novelty. It proceeds to point out areas of challenges in chromatography and spectroscopy. Final discussions are devoted to the role of NP integrity including purity and linkages between chemical and biological properties of bioactive NPs. The integration of these aspects potentially could help in advancing the future role of NPs as a viable source of new biologically active agents.

### ■ METHODOLOGY

**Data-Mining Procedures, Journals, and Time Period Coverage.** The source journals ($n$ = 13), intervals monitored (1998/1999 [period I], 2004/2005 [period II], 2009/2010 [period III]), and coverage of evaluated articles ($n_{tot}$ = 1823) are summarized in Table 1. All journals screened are well-established and peer-reviewed and dedicated to or frequently publish studies on bioactive NPs. They are focused on drug discovery and/or pharmacology involving NPs and exhibit a wide range of ISI impact factors (ca. 0.4 to 4.0).

In the initial stage, the survey consisted of a large-volume screening of reports from both years of period I, which involved manual screening of ca. 55 000 journal pages from 12 journals.

Upon compilation and preliminary data evaluation, this led to the selection of six priority journals and the addition of one journal (*Journal of Asian Natural Products Research*) for continuation of the survey in the subsequent periods II and III, which mostly focused on one year of the two-year periods (see Table 1 for details). The seven priority journals were selected due to their much higher information density, i.e., the number of qualifying articles. Accordingly, the number of published pages to be screened was reduced to ca. 15 000 and about 12 000 pages in the periods II and III. Of the journals with a lower prevalence of qualifying reports, one (*Journal of Pharmacology and Experimental Therapeutics*) was included with the seven priority journals and assessed for the entire period 2000−2003 ($n$ = 77; not included in $n_{tot}$) plus one year of period II, in order to provide an example of extended coverage for these journals.

**Inclusion and Exclusion Criteria.** For any given journal volume included in the study, *all* articles were prescreened manually for the following inclusion criteria: they had to report on both bioactivity and chemistry of NPs and provide a substantial experimental description, regardless of how well the NP-related portion was developed. Exclusion criteria were as follows: reports in which bioactivity was clearly a minor aspect of the work; reports in which NP and/or synthetic chemistry was so dominant that the bioactivity portion was insignificant; and reports with ambiguous experimental descriptions of the analytical parameters. By default, only full papers generally were included. However, depending on the journal and its editorial framework, in some instances such as limited coverage of a given volume or year, publications in content-limiting formats (e.g., Notes) that fulfilled the other inclusion criteria and had a sufficient level of detail to address the key study parameters were included in the survey. By ensuring that formatting restrictions did not impact the scores, these added publications contributed to the statistical significance of the survey by increasing the total number of articles evaluated.

**General Methods.** The raw data were collected into tabbed spreadsheets (Microsoft Excel 2010) and analyzed using

mathematical, sorting, and Boolean and other logics as well as conditional formatting functions of the software.

**Prospective Setting of Parameters.** The parameters extracted from the primary literature, as well as the scoring and key system used to record the information in a standardized spreadsheet format, resulted from a preliminary, randomized screening of ca. 100 articles and were defined before starting the main survey. As a means of prospective guidance for the data-mining process, all of the questions addressed under the Observations Made section of this review were also formulated at the onset of the study. While the insight gained during the study led to additional hypotheses that were eventually tested with the complete survey data, the scores initially implemented and keys as well as the basic set of questions were maintained constant during the entire study.

**Collected Basic and Compound Information about the Reported Bioactive NPs.** In addition to the basic header information about each article (author name, journal volume, page), the primary biological activity or target of each report was recorded. Furthermore, for each report, the total number of NPs and the predominant class of compounds were recorded, and the following data were determined: total number of compounds; number of compounds that were isolated by purification from natural material; number of compounds that were synthesized (full or partial); number of compounds that were gifts from colleagues; and the number of new structures.

**Scoring and Key Systems for the Evaluation of Isolation and Spectroscopic Methods.** The experimental section of each report was evaluated for parameters that reflect the methods used for the purification of the NPs and their spectroscopic characterization. As many reported isolation procedures are rather convoluted, the assessment included determination of the longest purification pathway and the highest degree of diversification of the methods employed. The maximum number of isolation steps was counted, excluding extraction and solvent partition procedures. The use of normal-phase silica gel as a primary or secondary purification step, after any partitioning or precipitation steps, was recorded as a binary number. In addition, the diversity of the purification methods was assessed and encoded into binary format as a byte integer, consisting of the following five bits: $0$ = undefined or literature reference only; $2°$ = precipitation or crystallization; $2^1$ = paper or thin-layer chromatography (TLC), including centrifugal TLC; $2^2$ = column liquid chromatography (LC), vacuum LC, and low-pressure LC, and the value $2^1$ was added to encode repetition in the entire scheme; $2^3$ = medium-pressure and high-pressure LC (MLPC, HPLC), and also the value $2^1$ was added to encode repetition in the entire scheme; $2^4$ = countercurrent chromatography. The reported LC techniques applied numerous different solid-phase packings, which were not individually differentiated in the survey and included primarily the following: normal- and reversed-phase (RP-8/18; cyano) silica gel; Sephadex LH-20 (see ref 10 for a comprehensive review); styrene resins (see ref 11 for theory and applications).

Data collected on the spectroscopic characterization of the NPs, for which bioactivities were reported, included the number of compounds for which spectroscopic data were reported, the comprehensive nature of the general physical/spectroscopic data in general, and, separately, the nature of the NMR data utilized in particular. For this study, "depth" is defined by the completeness, detail of interpretation, and comprehensive nature of data. This was assessed and scored as follows: for the general spectroscopic and other analytical data, 1 = highly comprehensive (X-ray and/or very comprehensive 1D and 2D NMR, MS, physical data); 2 = comprehensive (1D and some 2D NMR, MS, physical data); 3 = as for 2 but with apparent gaps; 4 = mainly or fully lacking; 5 = literature reference only, or in cases where no spectroscopic data were reported, or referred to "as previously described", with reference to other literature. Similarly, the depth of NMR data were scored: 1 = highly comprehensive (1D and 2D NMR and/or special experiments such as selective pulse experiments, spectral simulation, connection with molecular modeling studies); 2 = comprehensive; 3 = as for 2 but with gaps; 4 = mainly or fully lacking; 5 = literature reference only or in cases where no spectroscopic data were reported. In judging the completeness of physical data, the reports were considered adequate despite not providing UV data if the compounds had no chromophore, and similarly optical rotation was not expected if the molecules were achiral. For the assessment of reports from the most recent time period, III, the provision of spectroscopic data as Supporting Information (SI) was considered as added comprehensiveness that was linked via cross-references in the main text. Due to workload and practical considerations, however, the SI materials were not screened.

**Scoring and Key System for the Evaluation of Purity Assessment.** Finally, each article in its entirety was mined for information about purity assessment of the bioactive NPs. The information was encoded into binary format as a byte with five bits, as follows: 0 = undefined or obscure method; $2°$ = taken from vendor label; $2^1$ = single spot on TLC; $2^2$ = determined by HPLC; $2^3$ = determined by quantitative $^1$H NMR (qHNMR) through basic integration using the 100% method; and $2^4$ = determined by qHNMR with calibration or by titration, data given for each compound.

**Estimation of Human Error in the Evaluation and Scoring Process.** All authors were involved in the data-mining process, which involved manual page turning of journal hard copies, judgment of inclusion criteria of each article in the screened volumes, and mining of the aforementioned data from each of the 1823 articles. It is realized that the scoring systems involve an element of subjectivity that may lead to deviations in scores assigned by individual assessors. Another potential source of variation was the screening of experimental sections for data about isolation methods, in particular in reports where the purification procedures were lengthy and/or convoluted. While the information was mined with particular care and attention to detail, the extracted data might have deviated slightly in a few instances; for example, the number of isolation steps might be off by one step from the actual experiments. Given the workload of manually screening 80 000 pages of literature information, it was not feasible to perform the entire survey in triplicate and/or by multiple individuals. As no averaging was performed, the data represent the outcome of single assessments. The authors distributed their efforts as assessors across the journals, as this helped by averaging the influence of interindividual subjectivity. Moreover, a limited amount of cross-checking between the authors was also undertaken.

## ■ OBSERVATIONS MADE

Data from this AnaPurNa study are presented in the following paragraphs and are discussed with respect to the questions (Q; numbered) that were formulated initially in the prospective study. As the study evolved over the last 10+ years, additional

aspects for which the survey data could provide insight were added and are included.

**Literature Characteristics.** NP isolation accompanied by bioactivity measurement is considered to be the specialization of a larger discipline. While many journals occasionally publish articles on the isolation of NPs and their bioactivity, only a few journals regularly publish such reports. This study is an in-depth investigation of a handful of journals (limited for practical reasons) that publish articles routinely on the bioactivity of NPs, rather than a broad study of the general literature. This systematic literature review focuses on a selection of journals that contribute heavily to the specialization of NP research, with a particular focus on pharmacognosy and natural products chemistry. Independent of the choice of journals, it is likely that the editorial policies of the selected journals influence the data. This reflects the natural flow of disciplines in science, where areas of specialization form their own communities and at the same time contribute to the greater scientific endeavor in a variety of ways. For example, purified NPs may be incorporated into human clinical trials, which will be published in a medical journal rather than a natural products journal.

*Q1: Is Information on Bioactive Natural Products Concentrated in a Few Journals?* Yes. Seven of the 13 journals (Table 1) represent 79% coverage of qualifying reports. This number increases to 88% when including the proportional numbers of qualifying articles in *Planta Medica* from periods II and III. The articles assessed were almost evenly distributed over the time periods I, II, and III, providing 716, 501, and 597 surveyed reports, respectively.

*Q2: Which Journals Are the Major Sources of NPs Bioactivity Information?* When ranking the journals by number of qualifying articles, about one-third of them were published in the *Journal of Natural Products*. Following in the ranks are *Biological/Chemical and Pharmaceutical Bulletin* (combined), *Planta Medica, Phytochemistry, Fitoterapia,* and the *Journal of Asian Natural Products Research.* The latter was included in the survey for period III, in order to get a perspective on a publication that reflects the outlet of the very productive NPs research community in Asia. A graphical overview of the journals by contributed survey articles is provided in Figure S1, Supporting Information.

**Sources of Purified NPs.** There were four main sources of the NPs: (i) isolation and purification of the NP by the authors as reported in a scientific publication; (ii) purchase of NPs from commercial sources; (iii) receipt of NPs from colleagues who have performed the isolation and purification themselves; and (iv) (semi)synthesis.

*Q3: What Is the Role of Gifts and Synthetic Test Compounds?* The proportion of bioactive NPs that were described as gifts from colleagues has dropped over the survey time period from 1.6% in period I to 0.7% in period II and 0.6% in period III. Looking only at the group A journals, gifts were reported for around 1−2% of all investigated NPs (0.6% in period I, 2.5% in period II, 1.1% in period III) and, thus, contribute to only a very small proportion of the studies. The overall reduction in shared compounds might be a result of the trend toward smaller isolation yields and their consumption in the bioassays, together reducing the availability of the compounds. These observations are also in line with an observed trend toward collaborative research, which indicates that teams involving NPs researchers produce compounds dedicated to biological evaluation. This again may result in the

unavailability of the compounds for subsequent studies. Recently, some journals have implemented requirements for the inclusion of copies of original spectra as Supporting Information, which facilitates structural dereplication by other researchers. At the same time, this new mechanism may contribute to the observed reduction of sharing of the actual compounds among researchers.

The involvement of synthetic NPs has seen a significant decline, by 75%, over the study period: while they contributed a similar proportion of study compounds in period I in all journals (11.9 in journal group A, 7.9% in journal group B), their overall contribution to all study compounds decreased from a relatively high 9.5% in period I to 6.3% in period II and 2.8% in period III in the group A journals and from 11.5% in period I to 6.5% in period II and 3.1% in period III when adding both groups A and B together. This shows that the role of (semi)synthetic chemistry in the surveyed journals has diminished over the observation period.

*Q4: What Is the Role of Purchased Test Compounds? What Role Do Commercial Suppliers Have in Pharmacology-Oriented NPs Research?* Addressing this question eventually required a more elaborate analysis of the data, including differentiation by looking at individual journals and the groups of journals. In this study, it was determined that the vast majority of NPs used in bioactivity studies were isolated and purified by the authors from their natural sources by the protocols described in the experimental section included in the publication. The average proportions of purified NPs across all journals rose from 78% in period I to 93% in period II and 95% in period III. The main reason for both the high proportion and the rise may be that the probability of a major or even breakthrough discovery is lower with a compound that has been extensively investigated due to its unrestricted (commercial) availability. However, dividing all journals into two groups, A and B, according to the overall depth of spectroscopic data (see Methodology section, Table 1, and details below) reveals a different trend: while in group A, ca. 85% of compounds reported in each journal were isolated and characterized, their proportion in group B is only about 55%. This means that in the group B journals ca. 45% of NPs are purchased, gifts, or synthesized compounds. Considering that gifts and synthetic substances are generally minor sources of NPs, this implies that the amount of purchased NPs has increased in the group B journal reports. These interpretations are supported by the analysis of all 12 journals from period I: only 3.5% of NPs (107) reported in the group A journals were from commercial sources, but their proportion in the group B journals was 6 times higher, at 22% (305). For one group B journal (*Journal of Pharmacology and Experimental Therapeutics*) analyzed in period II, the proportion of commercial NPs was 47%. Conversely, in group B the proportion of isolated/characterized compounds was as low as 36% in reports within an individual journal. These observations regarding the sourcing of the investigated NPs are independent of differences in scope and policy of the journals in groups A and B and also of the diverse foci (e.g., chemistry or biology orientation) of individual reports.

The high percentage of bioactive NPs that are isolated and characterized (currently about 95%) also indicates the rarity of purified NPs in that most researchers tend to produce these compounds by themselves rather than obtaining them commercially. This observation is important, because one consequence of this practice is that the authors themselves are

responsible for establishing not only the identity of the NP but its purity as well. In cases where NPs are obtained from commercial sources, the isolation process may well be proprietary; however, the NPs will also carry a specification sheet, certificate of analysis, and/or certificate of origin that includes a purity statement conforming to the standards of the manufacturer.

The percentage of reports on new chemical entities has been remarkably stable over time. In the group A journals, an average of 30.1% of reported NPs were new chemical entities (30.2% in period I, 26.5% in period II, 31.0% in period III), which represents a 4-fold higher incidence than in group B. Since the beginning of the survey, the reports in one journal (*Fitoterapia*) included in the seven priority journals have shown a significant increase in new NPs and today closely match the average of the group A journals (28.3% in period III).

Considering this further, in most cases the principal division of labor in the surveyed reports is the sharing of responsibilities between the NP chemist performing the isolation and the biologist completing the bioassay work. This means that at a certain point the isolated NP(s) are handed off from a NP laboratory to a biology laboratory. In this case, unless activity-guided chromatographic fractionation is conducted, it seems that the investigations will almost always be chemistry driven, so that the chemist will select the most interesting and accessible NPs to isolate. Typically, the bioactivity of the crude extract was reported along with the bioactivities of the final isolation product(s), while the potency of fractions throughout the separation scheme was reported much less frequently. It is important to monitor the activity of NPs (extracts, fractions, purified compounds) through at least three purification steps in order to establish the correlation between chemical *purity* and biological *activity*. Purity–activity relationships (PARs)[12] are quantitative correlations between chemical (purity) and biological (potency) parameters, which indicate whether or not the observed biological activity can be attributed to the main component, assigned as active principle. As such, PARs can be helpful indicators for prioritization. Considering the role of purity in the literature, as observed in this study, this type of information might currently be under-utilized. Notably, PARs can also be established at the level of purified compounds (NPs and cNPs; Figure 1), e.g., by comparing the potencies of the same NP purified from different sources and/or by different purification protocols. Another measure of the importance of the biological component may be how many NPs that produce promising "hits" in the biological assay are investigated further for their biological activity. This interface between NP chemistry and biology is crucial to the ongoing success of this specialization, which seeks to harmonize these two aspects of scientific research.

**Chromatographic Methods for NP Purification.** Today, numerous chromatographic procedures with widely differing characteristics (selectivity, mechanism, resolution, loading capacity, scale-up behavior) are available to the NP researcher. The chromatographic information extracted from the surveyed literature provides insights into the ongoing use of this diverse toolbox. The binary encoding and scoring of characteristics of the purification methods used in surveying all reports is described in the Methodology section and rests on a thorough case-by-case analysis of the experimental section of each report. Considering the correlation between the metabolomic complexity of crude extracts and the residual complexity of purified NPs (Figure 1), the codes and scores were designed to

provide metrics to answer questions about the depth and diversity of isolation procedures as they are used in laboratory practice.

While the number of NPs per individual report varies considerably, the average number of NPs per report has increased slightly over the survey time period as follows: 6.6 ± 7.3 (SD) in period I, 6.1 ± 5.8 in period II, and 7.8 ± 7.7 in period III. An upward trend is also noticed for the proportion of new NPs, which has increased by more than half from 26% in period I to 37% in period II and 41% in period III. It is noteworthy that this observed trend applies primarily to journals in which reports include a comprehensive coverage of the spectroscopic data (group A journals and *Fitoterapia*; see Table 1 and discussion below).

*Q5: What Is the Average Number of Isolation Steps to Yield a "Pure Compound?".* One significant outcome of this literature analysis is the revelation that the average number of steps taken to isolate and purify a natural product is less than three (*n* = 1823). In addition, this number has not changed significantly in the time period covered by the study. In period I, an average of 2.0 isolation steps (SD 1.7) was used to yield a "pure" NP. This increased to 2.4 in period II (SD 1.5) and 2.7 in period III (SD 1.6). Taking into account that 22.9% of reports did not employ any isolation steps (assigned value of 0), the other studies employed an average of three isolation steps. The data fit a Gaussian normal distribution reasonably well (Figure S2, Supporting Information), with a tail toward higher numbers representing the very few studies (*n* = 39, 0.6%) that employed six to 10 isolation steps. One conclusion from this data is that compounds that can be isolated in three steps or less are the predominantly isolated NP. On the other hand, this observation also indicates that compounds present in very small amounts and/or similar to more abundant congeners are currently rarely pursued, likely because they are more arduous to isolate. An interesting example of this is that ginkgolides A, B, C, and J have been reisolated from *Ginkgo biloba* L. (Ginkgoaceae) hundreds of times, while ginkgolides L and M are described in only one publication.[13] Recently, the two new ginkgolide congeners P and Q have been isolated in less than 30 mg quantities from 8 kg of *G. biloba* leaves.[14]

*Q6: How Much Effort Is Required to Isolate a Bioactive NP?* This question includes two aspects: the number of steps is addressed here, and the chromatographic methodology in the following section. About half of all reports (48.8%) either did not perform an isolation or employed only one or two steps to produce the bioactive NP. Isolation efforts included a maximum of three steps in about three-quarters of all reports (76.3%). Publications that described at least four or five isolation steps contributed to 23.7% or 7.2% of studies, respectively, and, in turn, can be considered in-depth isolation studies. There was no clear trend of their prevalence over the survey periods I/II/III, with 21.8/18.1/30.8% and 6.8/5.4/9.0% of ≥4- and ≥5-step studies, respectively.

The effort required to achieve single chemical entity parameters for an isolated NP depends on many factors, including (i) the concentration of the NP in the crude material (the higher, the easier the purification); (ii) the physicochemical characteristic of the compound, in particular solubility (precipitation) and tendency to form crystals (a historically important property of NPs); (iii) the "match" between the selectivity characteristics of the chosen purification methods and the NP (some methods appear to work better than others for certain compound classes or types of source materials;

different standard protocols for marine vs microbial vs plant NPs); and (iv) the nature of the matrix components in the crude NP, which may cause difficulties in the purification process (e.g., polyphenols or chlorophyll in plants, high-polarity overlap with primary metabolites and other polar substances in the case of marine NPs). Accordingly, a one-step isolation procedure might be sufficient to purify a NP that is present at relatively high concentration, i.e., in the 0.2% range (relative to dry weight of the biomass) and above. During this survey, numerous examples of such rapid access to a purified NP were noted in the literature. They involve typically solvent partitioning and just one step of normal-phase silica gel column chromatography, sometimes followed by precipitation or crystallization. Examples well-known to the authors are vitexin from *Vitex agnus-castus* (0.1−0.2% content) and xanthorrhizol from *Curcuma xanthorrhiza* (>0.2% content). As the information about purity in the literature has generally been very scarce (see below), there is very little basis for judgment of the properties of these kinds of materials and its impact on the biological activity. Given the long history of NPs research, it appears to be likely that more elaborate isolation schemes could produce new insights and novel structural and biological information, in particular when performing research on NPs that have previously been (extensively) studied.

*Q7: What Is the Preferred Methodology of Isolation?* On the basis of the entire data set ($n = 1823$), about two-thirds of all studies utilize normal-phase silica gel for the isolation of NPs. The proportion of these studies has increased over the observation period from 57% (I) to 63% (II) and recently 71% (III). Interestingly, studies that use normal-phase silica gel report isolation of crystalline compounds 2−5 times more often than studies that do not use this sorbent. Comparing studies that use normal-phase silica gel with those that do not, the ratio of the average number of crystalline compounds per study was 2.0 in period I, 2.6 in period II, and 4.9 in period III. In the same time interval, the proportion of crystalline isolates has declined from 10.1% to 7.3% and recently 4.9%, respectively. Overall, this may attest to the ability of normal-phase silica gel to concentrate and/or remove unwanted constituents and offer one reason for its steady popularity. Its widely known disadvantages such as irreversible absorption or degradation of desirable constituents are less frequently conveyed for bonded silica gel derivatives. Assessment of the actual impact of these unpredictable properties of silica gel-based stationary LC phases on the outcome of the purification protocol requires dedicated studies. One such example has been reported by Pinel et al.,[15] who directly compared normal-phase silica gel and liquid only based LC (countercurrent separation) for the purification of xanthanolides from *Zanthium macrocarpum* (Asteraceae). One intriguing finding was the ca. 13-fold reduced yield of one particular xanthanolide, xanthatin, when using the solid-phase method. This almost selective removal of a compound from a crude NP might inspire future developments and/or validation of silica gel-based purification methods.

With regard to the generation of crystalline NPs, it is noteworthy that their proportion has dropped from 10.1% to recently 4.9%. This may reflect the trend to smaller starting amounts of biomass and isolation yields based on the capability of modern spectroscopy to obtain structural information from smaller and/or less pure samples. These observations are in line with a conclusion recently made by Meyer and Imming,[16] underscoring the value of practical skills in compound

crystallization for contemporary research programs that involve purification of NPs and other drug leads.

Considering the extremely wide use of normal-phase silica gel, it is not surprising that one- to two-third of studies (63.7/35.5/32.5% in the periods I/II/III, respectively) used gravity-driven column chromatography *exclusively*. While this proportion is declining, the data show that a large proportion of isolation procedures are uniform rather than diverse. Likely the most prevalent isolation methodology consists of normal-phase silica gel, (repeated) gravity-driven column chromatography, and HPLC. This combination was found to also increase in popularity and has most recently been employed by almost one-half of all studies (27.4/36.1/45.6% in the periods I/II/III, respectively). These observations may reflect preferences for fast approaches such as automated flash chromatography and preparative HPLC and/or may also be a sign of the increased availability of such equipment. Although not specifically tracked and encoded in this survey, a general observation is the very frequent use of $C_{18}$ reversed-phase silica gel and Sephadex LH-20 as stationary phases for LC purification of NPs. Both materials are significantly more costly than normal-phase silica gel, which might explain their relatively lesser use, but they have the advantage of being reusable. Reversed-phase silica gel appears to be the second most widely used stationary phase and like normal-phase silica gel is widely employed in (semi)-automated LC applications such as HPLC, MPLC, and vacuum and flash LC (including high-throughput settings[17]).

In the present meta-analysis, NP purification schemes have two primary dimensions: the number of purification steps and the chromatographic methodology used in each step. While together they describe the overall depth of the purification process, a chemically diverse metabolome likely requires a chromatographically diverse purification scheme for the efficient mining of NPs. The binary scores given in this study for the diversity of the purification methods (see Methodology) allowed us to study the relationships between the number of steps and the chromatographic methodology (see scatter plot, Figure S3, Supporting Information). A general observation from the distribution of the purification diversity scores is that an increase in the number of purification steps does not necessarily indicate an increase in chromatographic diversity. The data exhibit the presence of general trends, as follows: two-step procedures mostly consisted of two LC steps (often repeated) or a combination of one LC and one HPLC step. Three- and four-step procedures frequently applied repeated LC and one level of HPLC, although a relatively large number of these purification schemes apply gravity-, vacuum-, or low-pressure-driven LC methods only.

Emerging from the authors' research and interest in countercurrent separation (CS; syn. CCC; see ref 18 for a review), this survey also explored how widespread the use of this methodology is. In all studies and over the entire survey period, countercurrent methods such as HSCCC, CPC, and DCCC are used only sporadically (average 0.9%). In fact, despite recent developments of countercurrent technology, its proportional use in studies on bioactive NPs has actually decreased over the project period, from 1.7% in period I to 0.3% in period III. Even when looking only at in-depth isolation studies (see above), the proportion of countercurrent chromatography use fell from 4.5% in period I to 0.5% most recently. However, the number of reports that employ countercurrent techniques and fractionate NPs in-depth, by using at least three (58/67/50% in periods I/II/III, average

58.3%) or four isolation steps (83/67/50% in periods I/II/III, average 66.7%), is high. This implies that countercurrent methodology is applied primarily in more complex isolation schemes rather than as an alternative to other techniques. These observations reflect the need for specialized countercurrent chromatography equipment, which might not be widely available even to well-equipped laboratories. Another consideration is that, unlike many (semi)automated solid-phase LC methods (e.g., preparative HPLC), countercurrent separation techniques require some time to be optimized. While this may be perceived as being disadvantageous, significant progress has been made recently on key aspects such as solvent system selection, instrument design, and operation modes, and there is a wealth of recent reports on efficient NP purification protocols that employ countercurrent techniques (see ref 18 and references therein).

*Q8: Is There a Preference for Well- and/or Long-Established Techniques?* Following from the observations made in Q7, the uniformity of isolation approaches may also be due to the fact that this systematic literature survey looked at only a limited number of journals. For example, there are dozens of articles published every year featuring the isolation of NPs with countercurrent separation with subsequent analysis of bioactivity. These articles are typically published in chromatography journals rather than NP publications.[19−21] Similar considerations apply for supercritical fluid separations. That having been said, the use of normal-phase silica gel as a chromatographic method of choice is much more entrenched than can be simply explained by the fact that some alternatives are considered to be specialized techniques. The reported use of normal-phase silica gel has actually increased during the time period of this literature survey.

Numerous preparative-scale analytical methodologies are used in minor compound purification in the laboratories of NP researchers. Owing to the complexity, newly developed techniques are often "test driven" in NPs laboratories. Examples are the development of countercurrent chromatography, as pioneered by Y. Ito and co-workers,[22,23] and the advent of HPLC in the 1970s.[24] While a few techniques have established themselves as mainstream, it remains unclear as to what other techniques have to offer and what roles they can play in the future.

**Spectroscopic Methods for NP Characterization.** Once a NP has been isolated from its metabolomic background (Figure 1), characterization of its chemical structure (verification, dereplication, or elucidation) is the next step toward a quality-controlled material (cNP, Figure 1) for biological evaluation. The questions posed were as follows:

*Q9+Q10: What Is the Level of Analytical Detail Provided for the Tested Bioactive NPs? Considering Available Instrumentation and Methods, What Is the Depth of the General Spectroscopic Data?* In order to answer these questions, both the physical data in the experimental sections as well as the tables and descriptions in the main text of the articles were assessed, and the extracted information was coded as previously described under Methodology. The criteria took into account the widespread availability of spectroscopic equipment (NMR, MS, UV, IR, less so CD/ORD). While the depth of spectroscopic analysis per se is scientifically independent, editorial policies and journal format constraints undeniably have an impact on the information finally reported and, very possibly, which experiments are performed. Therefore, to apply equal measures in the entire survey, the same

coding scheme was applied to all publications and across the survey time period.

Compounding the scores for the depth of spectroscopic data for all articles ($n$ = 1908) yielded a numerical average of 2.5 on the discrete scale from 1 to 5 (lower number better; see under Methodology). Thus, on average, the spectroscopic foundation of all reported bioactive NPs ($n_{cpd}$ = 12 570) was between "comprehensive" (2) and "with gaps" (3). The distribution of the scores (S4, Supporting Information) shows "tailing" toward higher scores, as a result of 21.9% of the reports lacking support by spectroscopic characteristics at all (13.3%) or in the same publication (8.6%). In relation to the NPs, spectroscopic data were provided for less than half of all compounds (5510 = 44%), with only minor differences over the 12-year survey period.

When analyzing reports by journal source, the distribution and average depth of spectroscopic information in the 13 surveyed journals were heterogeneous. This is not an unexpected outcome for a number of possible reasons already discussed above. In fact, when evaluating the depth scores of both general spectroscopic and NMR spectroscopic data for the entire survey period, a clear gap was noted between average scores of 2.5 and 3.0, as can be seen in the tables and graphs in S5 and S6 of the Supporting Information, respectively. This led to the classification of the journals into the groups A and B (≤2.5 [five journals] vs ≥3.0 [eight journals], respectively; see also Table 1). The five group A journals showed an average score of 2.1 and covered greater than four-fifths (10 660 = 84.7%) of all studied bioactive NPs. Conversely, reports of less than one-fifth of the compounds (1827 = 14.5%) were in the group B journals, which gave an average score of 3.8. Considering that about one-fifth of the reported NPs lack support by spectroscopic characteristics (see above), the distributions of the spectroscopic depth scores in the two journal groups are almost mirror images of each other (S5, Supporting Information). Analogous observations were made for NMR spectroscopic data, which is usually essential for structure elucidation and compound identification. Of the seven priority journals selected for long-term surveillance over the whole 1998−2010 period, five were group A journals.

*Q11: Considering Available NMR Instrumentation and Methods, What Is the Depth of NMR Spectroscopic Data?* The total average depth score (see Methodology section) for the NMR spectroscopic data (2.7) is almost identical to that of the general spectroscopy (2.5). The two sets of spectroscopic depth measures also show parallel behavior over time and have experienced a steady improvement over the three survey periods: from 3.3 to 2.3 for the NMR and 3.1 to 2.0 for the general spectroscopic data. This can be seen clearly from the score distribution plots provided in S7, Supporting Information. These results indicated that NMR spectroscopy in general and 2D-NMR, in particular, have become the mainstay of structure elucidation. The observation that in the most recent period, III, 49% of all NPs were reported with NMR spectroscopic information categorized as "comprehensive" and an additional 18% as "highly comprehensive" can be interpreted as a sign of strong NMR evidence for the structure of about two-thirds of all bioactive NPs. These encouraging observations, however, do not necessarily indicate that dereplication of two-thirds of all NPs is straightforward. While the scoring system was not designed to specifically address this question, it is the authors' impression that unambiguous dereplication requires (NMR) spectroscopic data sets that scored typically as one in this

survey. While additional studies will be necessary to draw conclusions about the level of detail that is needed and/or practical for structure dereplication and, thus, full reproducibility, the survey indicates that NMR spectroscopy has been playing an increasingly strong role in this regard.

What has changed over time is that the NMR spectra are now included as Supporting Information in most journals, especially in the case of new compounds, due to space constraints. Unfortunately, the depth of the NMR data cannot be assessed in many cases simply because the spectra are not available. One way to assess depth of NMR data is level of detail, such as the completeness of the assignments, the coupling pathways, the coupling constants, and the multiplicity assignments. In addition, not only may structural data be lost but valuable information on the purity of an isolated NP may be disregarded by consigning NMR data to a table or brief listing. Another way to assess the depth of structural information is to consider the number and sophistication of the spectroscopic tests that are reported. For example, 2D NMR techniques generally reveal more structural subtleties than can typically be deduced from 1D $^1$H and $^{13}$C experiments only. This brings up an important point of the sophistication of both the technique and the individual who interprets the data. Two scenarios present themselves: a rather simple technique in the hands of a skilled researcher can reveal structurally accurate conclusions, while a sophisticated technique may be poorly interpreted and even misinterpreted. All in all, the depth of structural information relies on what constitutes an adequate attempt to assign a structure to a given compound. With NMR prediction and simulation techniques becoming more mainstream (see refs 25, 26 and citations therein), it is possible that computational analysis of NMR spectra may be encouraged in the future as supporting or possibly even definitive evidence of a correct spectroscopic interpretation.

**Role of Purity and Methods for NP Purity Assessment.** The role of purity is typically, but not necessarily (see discussion below), assessed last in the NP isolation workflow (Figure 1). Ideally, the purity of quality-controlled NP for biological evaluation is high, making it a single chemical entity. The four initial survey questions regarding purity were addressed as follows.

*Q12: How Frequently Is Information on Compound Purity Reported?* The short answer is that reports occur rather infrequently and at a declining rate. Compounding the information for all journals and sorting by survey period, the topic of purity is only addressed (not necessarily measured) in 4.6−8.4% of the reports (6.3% total average). Over time, purity reporting has been on a decline and was found in only 31 of 597 reports evaluated in the most recent period, III (5.2%). Interestingly, when considering the whole survey period, 4.2% (76) of reports from the group B journals address purity vs 1.4% (26) reports in group A. Assuming a general awareness of purity as a parameter, it is possible that some studies determined purities without publishing this information, but there was no way of determining the abundance of such cases. In summary, purity analysis was reported as being performed for less than 10 in 1000 compounds, and HPLC or more elaborate methodology was used for less than five in 1000.

*Q13: What Is the Role of Labeled Purity?* It is important to differentiate between awareness and actual assessment of purity: an average of 3.6% of all reports included some form of purity analysis, and the rate has been declining over the survey period (4.6% to 3.2% to 2.8% in periods I, II, and III,

respectively). The difference between "purity addressed" and "purity assessed" (6.3% vs 3.6% of all reports) suggests that about 40% of reported purity information is taken from (vendor) labels or derived from undocumented or otherwise obscure methods.
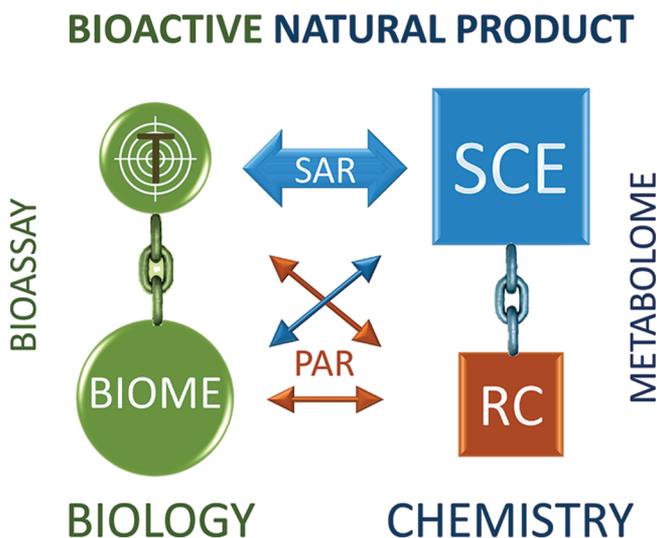


**Figure 2.** Connectivities between bioactive NPs and biological test systems. A "pure" NP ideally represents a single chemical entity (SCE). Its interaction with a defined biological target (T) establishes a definite structure−activity relationship (SAR) and typifies how the majority of bioactivities of NPs are characterized. However, due to variation in purification protocols and their source, NPs are inevitably (→ chain) impure profiles, by virtue of residual complexity (RC) from the source organism's metabolome. Similar considerations apply to the bioassay: whole cell assays in particular entail the entirety of biological targets and processes (biome) with which the NP sample interacts. Interactions between the SCE and/or the RC and the biome can lead to a response and need to be considered when interpreting outcomes. Depending on the proportions of SCE and RC and the interactions with T and the biome, the SARs and purity−activity relationship (PAR) of a NP will interfere, with possibly profound impact on the outcome.

*Q14: Is HLPC the Preferred Method of Purity Determination?* In contrast to the declining awareness of purity, the use rate of HPLC for purity analysis is flat over the survey period (2.2−2.4%): An average of 2.3% of reports, representing about one-third of all reports surveyed, used HPLC for this purpose. Accordingly, HPLC is the most common method of purity measurement when purity is reported. Ostensibly, this is done because an HPLC method was developed as part of the isolation scheme to either prepare the target NP(s) or assess the purity of fractions. HPLC provides the chromatography method, and the detector actually monitors the composition of the column effluent. The UV−vis method of compound detection is used widely in HPLC and other liquid chromatography methods. It can be a highly sensitive method to detect and quantify a target compound and has the potential to be universally applied to NPs that involve a UV−vis HPLC method at some stage of the purification protocol. On the other hand, this method often has severe limitations in detecting sample impurities. More sophisticated methods of LC detection are available,[27] including MS(−MS), ELSD, and corona charged aerosol detector, and used for this purpose. These methods all are limited to varying degrees in that they have
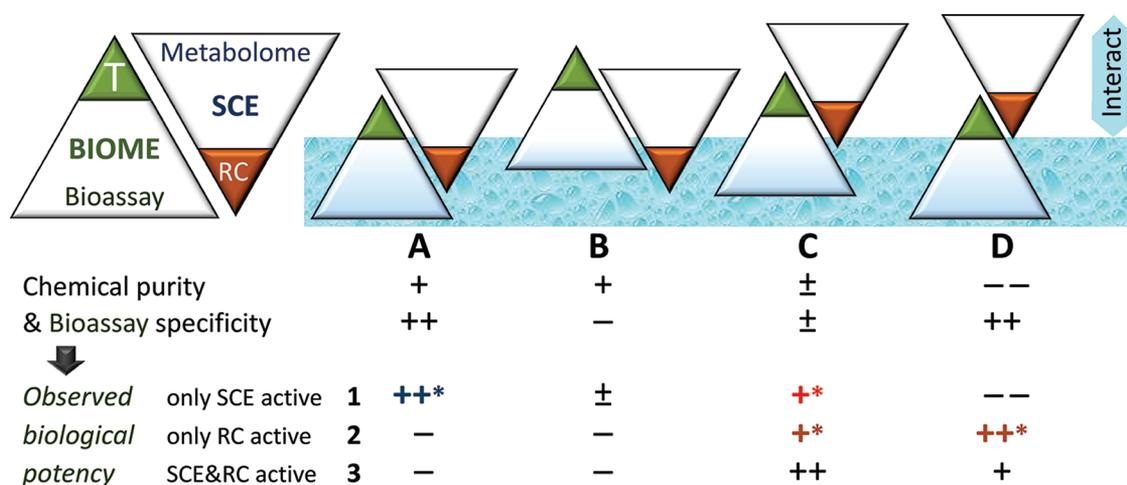
**Figure 3.** Interplay between the purity of NPs and different biological test systems. The outcome of testing "purified" NPs in bioassays that comprise different response elements (biome; Figure 2) can be symbolized by triangles in which proverbial "tips of the iceberg" represent both the well-defined target (T) of the bioassay and the residual complexity (RC; Figure 2) of the NP, respectively. As purity decreases and RC increases, four main scenarios, A−D, can be distinguished: (A) highly pure NP, only the SCE interacts with T; (B) like A, but the SCE interacts with additional biome processes; (C) in a bioassay with reduced specificity, NPs containing some impurities exhibit intermediate bioactivities, which can result from all four interactions depicted in Figure 2; (D) in impure NPs, the RC component dominates the biological response, even if T is highly defined. In addition to A−D, depending on whether only SCE or RC or both components are active principles, three series, 1−3, of biological potency may be observed. Particularly relevant (marked *) for NPs research are (A1) the ideal case: the purification leads to a near pure SCE as bioactive principle; (C1) false potency: the inactive RC "dilutes" the bioactivity of the SCE such that potency is misjudged; (C2 and D2) false assignment of bioactive principle: the purification yields the active principle, but it is contained in the RC component and not represented by the (apparent) SCE.

altered sensitivity between different compounds, and thus cannot be considered universal detectors and require carefully tuned parameter optimization. Relevant to purity assays, but primarily of importance to the metabolomic aspects of NP research, it shall be noted that recent developments in LC (e.g., UHPLC) and hyphenated technology (e.g., LC−MS, LC−NMR) have significantly advanced analytical capabilities for the characterization of both complex and purified NPs (see reviews 28, 29 and references therein).

The survey data clearly show that the purity of NPs used in bioassays is rarely reported. There may be some feeling that if a NP is pure enough to determine its structure by NMR and mass spectrometry, it is sufficiently pure to analyze its bioactivity. As shown in Figures 2 and 3, this may or may not be the case, depending on the balance of NP and its RC component and their interactions with the target and the biome of the test system. Parameters such as purity, specificity, and (residual) complexity are involved in both the chemical and the biological portions of the analyses, and they play equally important roles in the outcome. From the chemical perspective, there can be no question that the purity of a NP is of utmost importance in determining its bioactivity. At best, an inactive impurity will dilute the apparent activity that is usually measured in terms of bioactivity per mass of NP. More impactful is the possibility that the bioactivity of a minor component could mask the true bioactivity, or lack of it, for the target NP. This underscores the importance of following purity−activity relationships[12] as a way of correlating the measured bioactivity and the bioactivity of the target NP.

It should be noted that the aforementioned 3.6% rate of HPLC purity reports only referred to qualifying HPLC statements, whereas these reports did not include details of the analytical methods used such as chromatograms, integrals, and calculations. Elaborate reports of purity were coded separately and were very rare at a total average of 0.9%.

Moreover, the rate of detailed purity assessment has declined to levels of 0.20−0.34% in the most recent two survey periods.

Despite research progress and increasing interest in the methodology, quantitative [$^1$H] NMR (q[H]NMR; see refs 30, 31 for a literature overview) so far has rarely been used for purity assessment,[8] with only 13 reports, or 0.72%, employing qHNMR. Interestingly, the survey did not detect any further use of qHNMR for purity analysis of NPs involved in bioactivity studies in the most recent time period, III. However, virtually all investigators today use $^1$H NMR spectroscopy in the structural elucidation of their NPs and, therefore, have at their fingertips the data to determine the purities of the NPs isolated. Considering that qHNMR methodology is well-established, purity evaluation using the 100% method is straightforward from most existing $^1$H NMR data sets. Calculation and citation of such data are relatively uncomplicated and would add important evidence to the matter of RC highlighted here, as well as to related discussions about bioactive NPs.

*Q15: What Is the Average Purity of Tested Compounds?* Considering the observation that ca. 40% of reported purities are taken from (vendor) labels or undocumented methods and an additional ca. 30% from HPLC statements, the reported purity values have to be interpreted with caution. In order to put the small numbers of purity-tested NPs into perspective, of the total number of bioactive NPs ($n_{cpd}$ = 12 570), an average of seven compounds (SD 7) were included in one report. Given 102 reports on actual purity analysis by HPLC (statement) or better, this potentially affected ca. 700 NPs (ca. 6%). Due to the design of the study, the actual number of analyzed compounds was not recorded in the periods I and II. During the literature assessment it became clear that this number is much lower, likely by a factor of 5 to 10. This is in line with two other observations: (i) the low (0.1−1.1%) proportion of compounds in reports with HPLC or better purity analysis and (ii) that 85% of purity statements contained

one single value rather than a range of purities, indicating that only one or very few NPs were analyzed and/or individual samples were not differentiated.

Based on the evaluation of 102 reports, the following can be said about purity statements: the vast majority of reports (87%) state purities of "95%" or higher, and almost two-thirds (60%) of which report purities of "98%" and above. While a few studies (3%) even report absolute purity ("100%"), the same proportion reports purities below "80%". In general, assigned purity values were mostly given without decimal places.

## ■ SUMMARY AND CONCLUSIONS

**Added Dimensions of Complexity and Potential for New Approaches.** At least three additional dimensions of complexity affect the interpretation of research data on bioactive NPs and will be addressed in the following: (i) the role and relationship of in vivo (here including whole cell- and animal-based) vs in vitro bioassays used to assess bioactivity in the NP purification and characterization workflow (Figure 1); (ii) the depth and diversity of the purification workflow; and (iii) the role of purity and RC. In addition, just as RC is almost inevitable when purifying NPs, biological test systems are seldom singleton but rather residually complex or even very complex entities (e.g., in vivo systems). Potential further dimensions to consider relate to the connectivity between SCE, RC, and bioassay (Figure 2), i.e., the specificity, both qualitative and quantitative, of the bioassay. Finally, both the bioassay and the purified NP may behave like the proverbial *tip of the iceberg*, depending on their individual RC characteristics. As a result, a matrix of scenarios can be conveyed (Figure 3), which reflect the multidimensional interplay of the NP/SCE, its purity, its RC characteristics, the activities of the SCE and the RC component(s), and the specificity of the bioassay. Figure 3 shows that, depending on the combination of these factors, observations from the lack of bioactivity to the presence of strong activity potentially can be explained for NPs that otherwise appear to be identical or at least comparable entities.

**Residual Complexity.** The importance and potential impact of RC on, for example, the efficiency of drug discovery workflows cannot be underestimated. The initial discovery of an inverse correlation of anti-TB activity and the purity of ursolic acid (purity−activity relationships)[12] has led to the routine integration of qHNMR[30,31] purity assessment in the authors' laboratory. A recent report by Fitch et al. describes the comprehensive efforts aimed at establishing solid structure−activity relationships for the "frog" alkaloid epiquinamide.[32] Their studies involved chiral synthesis of three and pharmacological study (nicotinic acetyl choline receptor) of all four stereoisomers of the NP to eventually determine that all of them are inactive. The authors state that "the misleading activity in the natural product material is concluded to be trace contamination by co-occurring epibatidine", a finding that bodes heavily on the relevance of purity−activity relationship analysis for the validation of NP drug leads proposed earlier.[12]

**Relativity of Novelty.** Cragg and Newman et al.,[1−5] backed by extensive NPs research experience, stated that "the potential for the discovery of new chemotypes from plants, comparable to the taxanes and camptothecins, appears to be relatively low".[33] On the other hand, two recent articles by Kinghorn et al. provide convincing evidence, also backed by long-term research, for the relevance of higher plants and other terrestrial organisms as sources for new bioactive lead compounds.[34,35] Emphasizing the exceptional role of camptothecin and Taxol

(paclitaxel), Kinghorn et al. provide a thought-provoking interpretation that counters the other apparently discouraging outlook:[33] the possibility that an informative in vitro screen is not a substitute for a relevant in vivo assay. For the discovery of five anticancer leads in the 1970s and 1980s, there was an early involvement of in vivo testing.[35] However, the insights from the present AnaPurNa study add yet another possible interpretation. As the purification of highly active principles that are minor constituents likely requires more effort, it is conceivable that an increase in the fractionation depth (*n*, Figure 1) and/or diversification of the preparative-scale separation methodology is a viable means of improving the purification process and, thus, potentially can contribute to a discovery being made. Moreover, as the SCE and RC characters of NPs are closely linked (Figure 2) − a correlation that opens multiple possibilities for the interpretation of observed biological activity/potency (Figure 3) − it can seldom be ruled out that bioactivities originate, in full or in part, from the RC portion of the NP (see scenarios C2 and D2 in Figure 3). The assessment of purity and RC (see discussion above) and the establishment of PARs[12] are potential valuable methods for the NPs discovery process. In light of the findings of this study, all these factors represent aspects that could stimulate future research design.

**Challenging Spectroscopy in Structure Elucidation.** It is widely recognized that structure elucidation, unless supported by direct atomic evidence from X−ray diffraction analysis, is largely based on indirect spectroscopic evidence, primarily from NMR, MS, IR, UV, and CD/ORD methods. As a result, elucidation is an asymptotic process and can be compared with a balancing act between the interpretation of the spectroscopic data and the possible structural variations that can potentially be aligned with it. Accordingly, the non-X-ray approach to structure determination includes an element of uncertainty ("residual doubt"), which depends on the depth of the analysis in terms of the choice of type and number of spectroscopic experiments, but also on how well the chemical space is probed for alternative structures (e.g., isomers, compounds with heteroatoms beyond N and O) that potentially fit the spectroscopic information.

There are clear indications in the literature that in-depth studies lead to higher confidence in the assigned structure (reduced "residual doubt") and frequently lead to reassignment and revision of structures. One such example is the case of hypurticin, a 2*H*-pyran-2-one, which was recently reassigned to contain a 3′,5′,6′- rather than a 3′,4′,6′-tri-OAc side chain.[36] Mendoza-Espinosa et al. make a convincing case by employing detailed density functional theory and ^{1}H NMR analysis, including ^{1}H NMR spectral simulation and the synthesis of an analogue. A similar approach in the authors' laboratory also employed ^{1}H NMR spectral simulation and involved a detailed analysis of the higher order *J* coupling patterns of the sugar moiety. This led to the identification of the first cyanogenic glycoside with *β*-allose rather than *β*-glucose attached to the cyanogenic methine carbon, which has broad implications for the enzymology of cyanogenesis.[37] During the present survey, the authors frequently encountered articles in which the reported ^{1}H NMR spectroscopic data, in particular the interpretation of coupling patterns and the deduction of coupling constants (*J*), were lacking or ambiguous and, thus, would not allow the distinction of diastereoisomers. In many cases, even though other spectroscopic evidence was provided, this did not provide the distinction of the given structure from potential (stereo)isomeric alternatives. The most frequently

encountered examples are epimeric hexopyranoses, which in the $^1$H NMR domain require a relatively tedious analysis of their $^1$H NMR "multiplets between 3.2 and 4.5 ppm" to sort out the correct $J$ couplings and chemical shifts. The aforementioned cases of hypurticin[36] and $\beta$-D-allopyranosy-loxy-2-phenylacetonitrile,[37] among many others, demonstrate that taking on the challenge of "residual doubt" can lead to significant discovery.

**Challenging Preparative Chromatography.** During the writing of this review, the authors became aware of an excellent book chapter by A. D. Wright, which includes a meta-analysis of the literature regarding the isolation of marine NPs.[38] The author analyzed 115 reports during 1995, published in *Journal of the American Chemical Society*, *Journal of Natural Products*, *The Journal of Organic Chemistry*, and *Tetrahedron* and recorded parameters similar to the present survey. One aspect of the study was the differentiation of stationary phases and specific chromatographic methods. The analysis made indicated that "average" isolation methods in marine NP chemistry might be different from those used for NPs from terrestrial organisms. Distinct differences seem to exist with regard to the use of silica gel (6% of studies analyzed in ref 38 compared with 57−71% in the above discussion) or CCC (7% of studies analyzed in ref 38, 0.3−1.7% in the present survey, depending on time period). While the surveys cannot be directly compared, future meta-analyses of the NP literature, including the continuation of the present AnaPurNa study, can likely benefit from extended parameter sets that enable addressing additional aspects in the contemporary methods used to purify and analyze bioactive NPs.

During the extensive literature study performed, the present authors observed that only very few publications include new preparative-scale analytical methods in their NP isolation schemes. It seems that innovative chromatographic methods would enhance the systematic exploitation of NP diversity. This is exactly the chemical space that is the focus of the application of metabolomics to the study of NPs. In effect, metabolomics absolutely requires an organized investigation of the thousands, and possibility ten of thousands, of chemical entities that a single organism may contain. A recent editorial in *Phytochemical Analysis* points out that, although many articles include the term "metabolomics" in their title, the content of the publications concerned tend to reflect only standard methodology and reporting.[39]

**Integrity and Reproducibility.** Ultimately, the depth (see Methodology section) of the spectroscopic data of a quality-controlled NP (Figure 1) and the thoroughness of their reporting represent important parameters of the integrity in NP research. This applies from two angles: from the perspective of a novel structure and its discoverer(s), as mentioned above, it effects an ambiguity to the structure elucidation and the amount of "residual doubt". Recently, the term "NP integrity" has been coined by the National Center for Complementary and Alternative Medicine of the U.S. National Institutes of Health in relation to research on biologically active agents used in complementary and alternative medicine, particularly including botanicals and other dietary supplements.[40] These guidelines are intended to ensure reproducibility of preclinical, translational, and clinical studies with NP agents, which are known to exhibit much larger variation in constitution than other common intervention materials, such as SCE-based drugs. From the perspective of reproducibility, research involving (re)isolation, characterization, and/or other oper-

ations with previously published NPs performed by the same or other scientists, the integrity of spectroscopic information influences the degree of certainty with which structures can be dereplicated and/or distinguished from close congeners.[41] This makes the depth of spectroscopic information, as assessed in this survey, an element of NP integrity and a factor in reproducibility.

Purity is another factor of integrity. Increasing considerations of purity as a standard or required physicochemical parameter might be (mis)interpreted as a quest for ever-increasing purities. High purity certainly has its merit, because it allows the NP to approach SCE status and simplifies interpretation and understanding of a biological outcome (Figures 2 and 3). At the same time, requiring NPs to be highly pure often imposes overly proportional or even unrealistic efforts and costs on the research. Depending on the biological application and research aim, a certain degree of RC can also be of value, as residually complex NPs more closely reflect the natural character of a NP. Provided that RC and purity of a NP are known and documented, even less pure materials can be potentially useful and/or are suitable as unique research tools for biological studies, as long as their greater chemical complexity is considered during the interpretation of the results (Figure 3). Examples of relevant biological topics that can benefit from the availability of such materials include additive, synergistic, and antagonistic action and their use as markers in standardization of biological agents such as botanicals.

While even very thorough analysis may not solve the challenge of reproducing identical NPs with identical RC patterns, biological studies with such highly characterized NPs (cNPs) at least can be compared on a more rational basis. The suitability of NP and cNP materials, e.g., attaining certain potency levels or confirmation of activity at a given target, can usually be assessed only *after* their purification. Accordingly, studies aimed at the biological evaluation of NPs can greatly benefit from assessing the status of any NP sourced from the pipeline from crude NP to SCE/cNP (Figure 1) and from publically disseminating this information as equally valuable along with the chemical and biological data.

## ■ ASSOCIATED CONTENT

**Ⓢ Supporting Information**

This material includes graphical representations of survey data and is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

**Corresponding Author**
*Tel: +1 (312) 355-1949. Fax: +1 (312) 355-2693. E-mail: gfp@uic.edu.

**Notes**

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

## ■ DEDICATION

Dedicated to Dr. David C. Lankin, University of Illinois at Chicago (UIC), on the occasion of his 70th birthday and in recognition of his commitment and most valuable contributions to pharmacognosy research and education at UIC, as well as his collaborative spirit.

## ■ REFERENCES

(1) Cragg, G. M.; Grothaus, P. G.; Newman, D. J. *Chem. Rev.* **2009**, *109*, 3012−3043.

(2) Newman, D. J.; Cragg, G. M. *J. Nat. Prod.* **2007**, *70*, 461−477.

(3) Newman, D. J.; Cragg, G. M.; Snader, K. M. *J. Nat. Prod.* **2003**, *66*, 1022−1037.

(4) Cragg, G. M.; Newman, D. J.; Snader, K. M. *J. Nat. Prod.* **1997**, *60*, 52−60.

(5) Newman, D. J.; Cragg, G. M. *J. Nat. Prod.* **2012**, *75*, 311−335.

(6) Wagenaar, M. M. *Molecules* **2008**, *13*, 1406−1426.

(7) Nishtar, S. *Lancet* **2012**, *379*, 1084−1085.

(8) Chen, S.-N.; Lankin, D.; Chadwick, L. R.; Jaki, B. U.; Pauli, G. F. *Planta Med.* **2009**, *75*, 757−762.

(9) Cordell, G. A.; Quinn-Beattie, M. L.; Farnsworth, N. R. *Phytother. Res.* **2001**, *15*, 183−205.

(10) Henke, H. *Preparative Gel Chromatography on Sephadex LH-20*; Hüthig: Heidelberg, 1995.

(11) Sarker, S. D.; Nahar, L. *Natural Products Isolation*, 3rd ed.; Springer: New York, 2012.

(12) Jaki, B. U.; Franzblau, S. G.; Chadwick, L.; Lankin, D. C.; Wang, Y.; Zhang, F.; Pauli, G. F. *J. Nat. Prod.* **2008**, *71*, 1742−1748.

(13) Wang, Y.; Sheng, L. S.; Lou, F. C. *Yao Xue Xue Bao* **2001**, *36*, 606−608.

(14) Liao, H. J.; Zheng, Y. F.; Li, H. Y.; Peng, G. P. *Planta Med.* **2011**, *77*, 1818−1821.

(15) Pinel, B.; Audo, G.; Mallet, S.; Lavault, M.; De La Poype, F.; Seraphin, D.; Richomme, P. *J. Chromatogr. A* **2007**, *1151*, 14−19.

(16) Meyer, A.; Imming, P. *J. Nat. Prod.* **2011**, *74*, 2482−2487.

(17) Eldridge, G. R.; Vervoort, H. C.; Lee, C. M.; Cremin, P. A.; Williams, C. T.; Hart, S. M.; Goering, M. G.; O'Neil-Johnson, M.; Zeng, L. *Anal. Chem.* **2002**, *74*, 3963−3971.

(18) Pauli, G. F.; Pro, S. M.; Friesen, J. B. *J. Nat. Prod.* **2008**, *71*, 1489−1508.

(19) Zhang, Y.; Shi, S.; Wang, Y.; Huang, K. *J. Chromatogr. B* **2011**, *879*, 191−196.

(20) de Beer, D.; Joubert, E.; Malherbe, C. J.; Brand, J. D. *J. Chromatogr. A* **2011**, *1218*, 6179−6186.

(21) He, S.; Lu, Y.; Jiang, L.; Wu, B.; Zhang, F.; Pan, Y. *J. Sep. Sci.* **2009**, *32*, 2339−2345.

(22) Ito, Y.; Bowman, R. L. *Science* **1970**, *167*, 281−283.

(23) Ito, Y. *J. Chromatogr. A* **2005**, *1065*, 145−168.

(24) Henry, R. A. *The Early Days of HPLC at DuPont*; LC-GC North America. Published by Chromatographyonline.com on February 1, 2009 (accessed on 1/14/2011).

(25) Kuhn, S.; Egert, B.; Neumann, S.; Steinbeck, C. *BMC Bioinf.* **2008**, *9*, 400.

(26) Soininen, P.; Haarala, J.; Vepsaelaeinen, J.; Niemitz, M.; Laatikainen, R. *Anal. Chim. Acta* **2005**, *542*, 178−185.

(27) Wolfender, J. L. *Planta Med.* **2009**, *75*, 719−734.

(28) Eugster, P. J.; Guillarme, D.; Rudaz, S.; Veuthey, J. L.; Carrupt, P. A.; Wolfender, J. L. *J. AOAC Int.* **2011**, *94*, 51−70.

(29) Wolfender, J.-L.; Marti, G.; Queiroz, F. *Curr. Org. Chem.* **2010**, *14*, 1808−1832.

(30) Pauli, G. F.; Jaki, B. U.; Lankin, D. C. *J. Nat. Prod.* **2005**, *68*, 133−149.

(31) Pauli, G. F.; Jaki, B. U.; Gödecke, T.; Lankin, D. C. *J. Nat. Prod.* **2012**, *75*, 834−851.

(32) Fitch, R. W.; Sturgeon, G. D.; Patel, S. R.; Spande, T. F.; Garraffo, H. M.; Daly, J. W.; Blaauw, R. H. *J. Nat. Prod.* **2009**, *72*, 243−247.

(33) Newman, D. J.; Cragg, G. M. In *Anticancer Agents from Natural Sources*; Cragg, G. M.; Kingston, D. G. I.; Newman, D. J., Eds.; CRC Press/Taylor & Francis: Boca Raton, FL, 2005; pp 553−571.

(34) Kinghorn, A. D.; Chai, H. B.; Sung, C. K.; Keller, W. J. *Fitoterapia* **2011**, *82*, 71−79.

(35) Kinghorn, A. D.; Pan, L.; Fletcher, J. N.; Chai, H. *J. Nat. Prod.* **2011**, *74*, 1539−1555.

(36) Mendoza-Espinoza, J. A.; López-Vallejo, F.; Fragoso-Serrano, M.; Pereda-Miranda, R.; Cerda-García-Rojas, C. M. *J. Nat. Prod.* **2009**, *72*, 700−708.

(37) Seigler, D. S.; Pauli, G. F.; Nahrstedt, A.; Leen, R. *Phytochemistry* **2002**, *60*, 873−882.

(38) Wright, A. D. In *Natural Products Isolation*; Cannell, R. J. P., Ed.; Humana Press: Totowa, NJ, 1998; pp 365−408.

(39) Verpoorte, R.; Choi, Y. H.; Kim, H. K. *Phytochem. Anal.* **2010**, *21*, 2−3.

(40) The term "product integrity" was established by NCCAM, NIH, in 2008, replacing the term "product quality" of the intial guidance documents (NOT-AT-05-003 and NOT-AT-05-004), and led to the current NCCAM Policy: Natural Product Integrity (http://nccam.nih.gov/research/policies/bioactive.htm). In this document, the term "product integrity ... refers to the entirety and completeness of information about a product that assures it will meet NCCAM Policy requirements".

(41) Molina-Salinas, G. M.; Rivas-Galindo, V. M.; Said-Fernández, S.; Lankin, D. C.; Muñoz, M. A.; Joseph-Nathan, P.; Pauli, G. F.; Waksman, N. *J. Nat. Prod.* **2011**, *74*, 1842−850.