Data-driven pipeline for efficient development of bioingredients assisted by Omics, Bioinformatics and AI technologies

Parizzi, Lucas Pedersen[1]; Gomes, Gustavo da Silveira[1]; Ferrari, Cintia[1]; Dias, Ana Luisa Abrahão[1]; Lago Nold, Juliana Carvalhães[1]; Paludetti, Mayara Fregonezi[1]; Silva, Amaro Emiliano Trindade[1]; Reigada, Juliana Beltrame[1]; **Zimbardi, Daniela[1*]**

[1]Natura Cosméticos S/A, SP, Brazil

*Daniela Zimbardi, Rodovia Anhanguera, s/no. Km. 30,5 Polvilho, CEP: 07790–190, Cajamar – SP, danielazimbardi@natura.net

Abstract

Exploring novel applications for natural products is a challenging task. Plant derived products are complex mixtures of chemicals, oftentimes little known and subject to variabilities such as seasonal, geographical, and metabolic. We present a data-driven process that has been applied in the discovery of cosmetic applications, chemical characterization and molecular safety screening of ingredients derived mostly from Amazon rainforest biodiversity species, here exemplified by the development of Tucumã, also called Tucuma-do-Pará, (*Astrocaryum vulgare*) pulp oil and seed butter and their application with anti-aging benefits. These findings are supported by OMICs (transcriptomics and metabolomics), bioinformatics and Artificial Intelligence (AI) models.

Keywords: Bioinformatics; OMICs; Artificial Intelligence; Amazon Biodiversity

Introduction

Plant-derived natural products are a vast and diverse source of molecules, being widely explored since ancient times, from traditional medicine, rituals and culinary to recent cutting-edge drug discovery initiatives [1, 2, 3, 4]. Although natural products offer a wealth of potential benefits across various fields, often harboring unique biological properties, their chemical composition and biological interactions are very complex and can vary significantly due to geographic location, season, growing conditions, harvesting time and extraction methods, presenting significant challenges for research and development. Despite these difficulties, the development of food, drugs and cosmetics from plant sources is fostered by consumer's demands for natural, safe and high-quality products and enabled by advances in analytical techniques, material processing and computational tools and capabilities. The integration of metabolomics, transcriptomics, Quantitative Structure-Activity Relationship (QSAR) modeling, bioinformatics and AI tools present a comprehensive framework for screening samples, finding novel cosmetic applications and assessing the biological activity and safety of plant-derived products.

Metabolomics is a rapid evolving field and offers new opportunities to understand plant extracts. The untargeted metabolomics approach analyzes all the metabolites in a sample and when coupled with high-resolution analytical techniques such as liquid chromatography (LC) and mass spectrometry (MS) generates comprehensive metabolite profiles with high sensitivity and specificity. In particular ultra-high pressure LC (UHPLC) with tandem MS (MS/MS) has become a standard method for plant metabolomics studies due to the capacity to rapidly separate and detect a wide range of small molecules, including key groups of plant secondary metabolites [5, 6], and identifying new metabolites at very low concentrations [7,8]. However, the vast amount of data generated in untargeted metabolomics can be challenging to analyze and interpret, and precise annotation is still a challenge [6, 9]. For this reason, high mass and spectral accuracies are key [10], and can only be achieved by using high resolution mass spectrometers, such as time-of-flight (TOF) and Fourier transform (FT) MS, including FT ion cyclotron resonance (FT-

ICR) and Orbitrap technology, and computational processes are required, such as machine learning [11, 12] and molecular networking [13, 14, 15, 16].

Molecular networking is a useful tool for analyzing MS/MS-based metabolomic data and categorizing compounds with similar molecular structures into clusters, allowing a deeper exploration of the chemical space of the sample and of unannotated structures [17]. It has been extensively used to identify new compounds in complex plant samples [18, 19], compare the composition of different parts of plants [13] and study plant interaction with other organisms [20], being a crucial step in natural product discovery pipelines. The annotation of the metabolites and clusters links their chemical structures to potential biological activities and can also help to identify undesirable cultivation or process related variabilities.

QSAR models utilize the chemical structures of identified metabolites to predict their biological activities, such as permeability, sensitization [21], cytotoxicity [22] and pharmacological effects [23, 24]. Despite its potential, existing QSAR models for predicting the toxicity of compounds exhibit limitations when applied to natural products, mainly due to a distinct chemical space. To overcome this limitation, developing new QSAR models that encompass a wider range of natural products within their chemical space is crucial, as it can significantly improve the accuracy of predictions [25]. By screening the metabolome, QSAR models are employed to identify chemical structures with alerts, guiding the prioritization of compounds for further investigation and determining the necessity for safety tests.

To provide a more thorough picture of how the plant-derived products interact with skin cells and induce desired effects, both untargeted and targeted transcriptomics experiments can be used. While the untargeted approach, often employing RNA sequencing (RNAseq), inspects a higher number of genes and allows the identification of unexpected pathways and interactions, the targeted transcriptomics, using techniques like quantitative PCR, focus on a predefined set of genes known to be involved in functions of interest, in the case of cosmetics, skin functions like inflammation, pigmentation and elasticity, providing a in-depth analysis of key pathways and modulation of their activity. Expression data analysis methods like Gene Set Enrichment (GSEA) [26] and Over-representation (ORA) are crucial tools to interpret the generated expression profiles in terms of impacted (enriched) pathways or gene sets rather than individual genes. GSEA identifies changes in the expression of functionally related gene sets and helps to understand broader biological processes in terms of gene set coordination. ORA analyzes the enrichment of specific gene categories, allowing to pinpoint key pathways or cellular processes most significantly impacted by the treatments, regardless the direction, just considering differentially expressed genes (DEGs).

Integrated into a robust pipeline, these computational tools allow the identification of valuable bioactive compounds and mixtures within vast and diverse natural resources, particularly those found in rich and often endangered ecosystems, promoting their sustainable use.

Brazil hosts the largest plant biodiversity of the planet, with 39,285 plant species registered, 34% of which are in the Amazonian domain [27]. Native to the eastern Amazon region of Brazil, *Astrocaryum vulgare*, popularly known as Tucumã-do-Pará, is a multi-stemmed palm tree that can reach 30 meters height, with thorns covering its upper half and the entire length of its feather-shaped leaves. The plant is classified as a pioneer species, often among the first to colonize disturbed or degraded areas, exhibiting resilience to fires and adaptability to nutrient-poor soils. The fruits are edible, rich in lipids and carotenoids, used as animal feed and in a variety of food preparations, consumed both raw and processed, while the seeds are used for crafting ornamental objects [28].

Tucumã-do-Pará pulp oil contains 74.4% and 25.6% of unsaturated and saturated fatty acids respectively, mainly oleic (47%), linoleic (26%), palmitic (14%), and stearic (10%) acids [29, 30, 31, 32, 33], flavonoids [34], polysaccharides [31], high amounts of α-tocopherol [35] and carotenoids with high antioxidant activity such as β-carotene and α-carotene [30, 36]. Studies have also shown that Tucumã-of-Pará fruits have antioxidant, anti-inflammatory and anti-hyperglycemic activity, mainly related to the carotenoids composition [33, 37, 38]. The seed butter contains 87.3% and 12.6% of saturated and unsaturated fatty acids respectively, mainly lauric (50%), myristic (24%), palmitic (6%), oleic (8%), and linoleic acids (4%) and high tocopherol and tocotrienol total contents (12 – 18%) [29][39].

This study demonstrates the application of the *in-silico* pipeline to develop Tucumã as a cosmetic bioingredient, investigating in more details its chemical composition, safety and biological effects on skin. By exploring both pulp oil and seed butter, we aim to achieve sustainable use of this valuable biodiversity resource.

Materials and Methods

Plant Material

The fruits were obtained from the Bragantina region in the northeast of Pará during the harvest season. After separation, proprietary drying, pressing and filtration protocols were applied to obtain the crude oil and butter, respectively from the fruit's pulp and kernel. The genetic resources used in this study were accessed under the authorization of the Brazilian National System for the Management of Genetic Heritage and Associated Traditional Knowledge (SisGen permit no A083E9A).

Sample preparation and UHPLC-MS/MS Analysis

A total of 6 samples were analyzed by metabolomics, four crude butter and two crude oils. Samples were solubilized in tetrahydrofuran (THF) to a final concentration of 10mg/ml. LC-MS analyses were performed with a Thermo Scientific 3000 RRLC UHPLC system coupled with the Thermo Scientific Q Exactive hybrid quadrupole-Orbitrap mass spectrometer, using a Acquity BEH 1.7 µm C8 reversed-phase UHPLC column (2.1 x 100 mm), at 40°C, and a gradient pump system of water acidified with formic acid (0.1%, v/v) (A) and acetonitrile (B). The mass spectra were acquired in both positive and negative ionization modes (HESI ionization source), considering a mass range of 100–1,500 Da, using data-dependent acquisition (DDA) mode.

Feature Annotation and Feature-Based Molecular Networking

Raw data from the UHPLC-MS/MS analysis was collected in Thermo Scientific format (.raw) and converted to an open format (.mzXML) using MSConvert software [40]. Data preprocessing, feature detection and sample alignment was carried out using MZMine3 software [41, 42]. The results for the aggregate of samples were exported as a feature quantification table (.csv), and tandem MS spectra in mascot generic format (.mgf).

Feature Annotation process using MS/MS spectral matching and Feature-Based Molecular Networking (FBMN) was conducted using the Global Social Natural Products (GNPS) [43] web platform. Spectral match thresholds used were 0.7 modified cosine score and 6 minimum fragment peaks match and 0.1 m/z threshold (for both spectral match and network modeling). Feature-Based Molecular Networking results were combined with sample metadata, library and quantification results using R scripts to supply node information and enhance network interpretation. Additional information, such as chemical class taxonomy was annotated through the ClassyFire tool [44].

In silico safety screening

The compounds annotated in previous steps were subjected to *in silico* molecular safety screening using both public methods (e.g. QSAR Toolbox, VEGA, Pred-Skin) and a proprietary AI ensemble model based on QSAR, guiding subsequent *in vitro* and clinical safety trials. The proprietary model was developed for mutagenicity prediction and was trained with more than 11 thousand compounds with Ames test data. The ensemble model presented a diverse chemical space, which included a wide range of natural products, achieving a high balanced accuracy for this type of chemical. Further details about this proprietary model were previously described [25].

RT-qPCR

The cosmetics potential of Tucumã pulp oil and seed butter were investigated using transcriptomics studies on skin models (in-vitro 2D or 3D-printed and ex-vivo) treated with formulations containing Tucumã oil and butter, followed by Gene Set Enrichment (GSEA), Over-representation (ORA), Clustering and Pathways analysis. The total RNA was extracted using the Mini Kit Purelink Purific RNA 50 Preps (Invitrogen) [45] according to the manufacturer's instructions. The cDNA was synthesized utilizing a High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems) and a real time RT-PCR analysis was performed according to the TaqMan protocol (Applied Biosystems) using TaqMan Array Fast, 96-well Plate custom targeting genes templates and Thermal Cycler StepOne (Thermofisher). The custom templates include targets mined through bioinformatics searches and manually curated genes, functionally related to specific cosmetic functions such as elasticity, antioxidation and skin aging.

ExpressionSuite software (v1.3) and R language libraries such as PCRedux [46] were used to calculate amplification curve parameters, which are used to estimate quality metrics, such as efficiency and could also be used to generate alternate efficiency-corrected quantification measures.

Normalization strategy was conducted by using control (reference) gene(s). Control gene was selected using the GeNorm algorithm, implemented in R language [47].

Relative quantification (RQ) was considered for quantification, estimated by the 2^(-Delta Delta C(T)) method [48].

Gene targets selection and potential cosmetic use screening

For the gene selection process, available healthy skin mRNA data from public repositories, such as Gene Expression Omnibus (GEO) [49] was used to find most relevant genes related to relevant skin processes, such as skin aging, contrasting obtained samples by their age. This was also combined with knowledge databases data, such as Gene Ontology (GO) [50, 51] The Gene Ontology Consortium, 2023], Reactome [52], and literature review to curate and point skin-related candidate genes for the custom templates. This latter process has been periodically refined, to update the template, or to feed a separate database with relevant potential candidates to be functionally related with skin mapped processes, mostly using data mining strategies from knowledge bases and most recent published studies.

These skin related functions are grouped in gene sets according to the findings. The gene sets are distinct from biological pathways, as they do not imply a direct causal relationship between the genes. This organization of gene sets functions as a simplified ontology of skin functions, encompassing various functions relevant to cosmetics such as skin aging, differentiation, barrier function, extracellular matrix (such as collagens, elastin and other protein synthesis or degradation), wound healing, pigmentation, inflammation, and endogen antioxidant activity. Another 90-gene custom template was developed to access toxicity related gene sets, such as skin sensitization.

The experiments using these templates may vary on the studied skin model (mostly *in vitro* and *ex vivo*), and might generate slightly heterogeneity among samples, limiting the interpretable output from the bioinformatics analyses, according to the biological model cell types.

Over the past seven years, these experiments have generated a substantial database of gene expression patterns, encompassing a large library of unique ingredients derived from biodiversity (including plant extracts, oils, and butters) across over 150 samples. This database enables researchers to conduct upstream analysis both on individual samples and by comparing expression patterns across the entire ingredient portfolio, assessing similarities and differences.

Protein quantification

Protein expression assay used *ex vivo* human healthy female skin fragments from blepharoplasty patients. The pooled samples include three different donors and triplicates for each donor, for each tested sample (control sample, plus oil and butter, separately). Skin was kept in culture and treated with the samples at a 2mg/cm² for a 72-hour period.

Protein was quantified using an enzyme-linked immunosorbent assay kit (ELISA) commercially available (R&D Systems, USA). After collecting the samples, they were incubated overnight with specific primary antibodies. Subsequently, they were incubated for 2 hours with specific secondary antibodies. The quantification analysis was performed in a microplate reader at 450 nm. The protein levels were expressed in pg/mL and calculated from reference values obtained with a standard curve, constructed with known concentrations of the cytokine recombinant. The results were finally expressed in a relative quantification against control samples.

Results

RT-qPCR experiments analyzed through GSEA showed Tucumã samples had potential for anti-aging applications. This potential is evidenced by the enrichment of gene sets related to antioxidant activity (against oxidative stress), extracellular matrix components (ECM), inflammatory processes, and wound healing, including cell proliferation and growth factors. Figure 1 illustrates the relative potential for skin application of Tucumã ingredients, grouped by macro functions, among 154 samples, using principal component analysis:

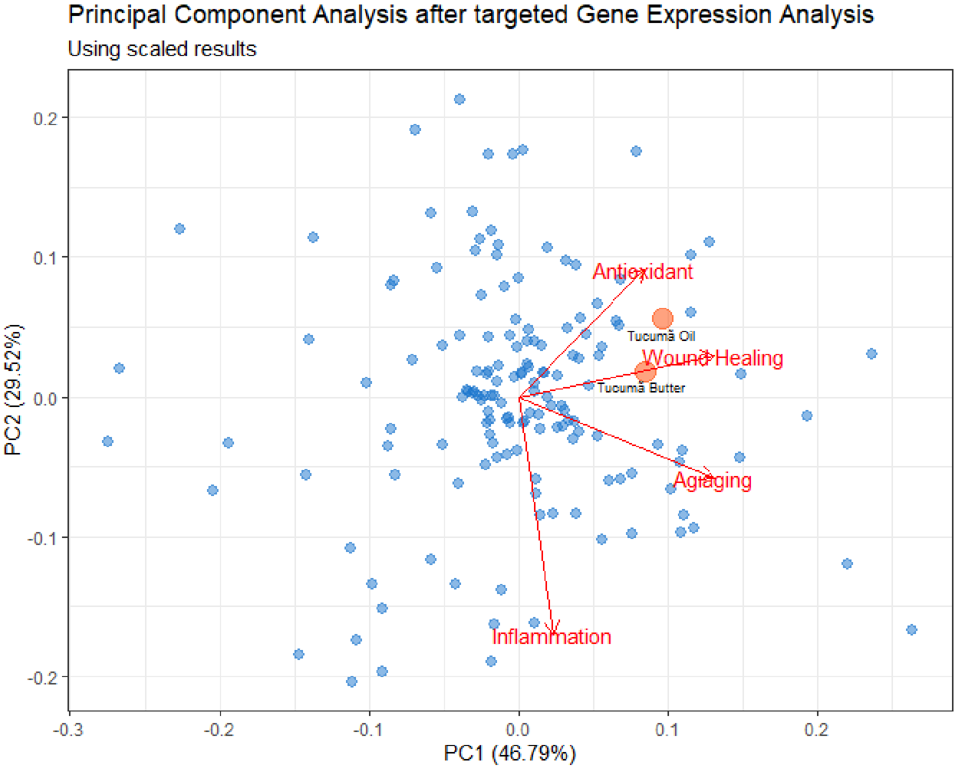

Figure 1: Principal Component Analysis from previously selected enrichment results.

The PCA loadings represent four macro categories for which custom templates were designed. Original variables were calculated as scores using enrichment results for custom categories (scaled GSEA scores from custom gene sets), aggregating specific skin processes into broader related gene sets. For example, ECM genes and custom functions are distributed among "Antiaging" and "Wound Healing", based on function, and cell proliferation is represented by the "Wound Healing" macro category. The original variables also represent positive modulation towards desirable cosmetic effects.

These results highlight Tucumã as a promising ingredient for this combination of cosmetic functions. When comparing from a macro perspective, samples with similar bioactivity may differ in terms of specific gene sets enrichment, despite their position in the PCA plot.

Notably, results from ECM genes and antioxidative activity drew attention to hyaluronic acid synthesis and degradation as promising targets for further investigation.

Efficacy study from *ex vivo* protein expression for HAS (Hyaluronic acid synthase, on butter), and HYAL (Hyaluronidase, on oil) was using ELISA quantification and showed concordant results in terms of regulation when comparing to gene expression, on Tucumã butter (hyaluronic acid synthesis), by an average of 77% of protein upregulation (treatment/control ratio of 1.77), and Tucumã oil (hyaluronidase synthesis), by an average of 44% downregulation (treatment/control ratio of 0.56, or a 78% fold change when comparing control/treatment). These results lead to a potential synergistic combination of the ingredients in terms of functionality. Antioxidant activity measured via SOD2 (Superoxide dismutase 2) protein expression, revealed an average 22,6% fold-change for Tucumã oil and 15,2% for Tucumã butter. Further clinical studies on formulations containing the ingredient combination confirmed the observed efficacy.

Metabolomics workflow, using ultra-performance liquid chromatography (UPLC) hyphenated with triple quadrupole time of flight (qTOF) mass spectrometry was carried out to identify distinct molecule clusters (mainly compound chemical classes) and annotated compounds, with the goal of elucidating the chemical diversity of the samples. Further studies could also establish links between specific compounds and their bioactivity, identifying those unique to or shared between samples, and also track chemical diversity within samples as a quality control measure, accounting for variations arising from seasonal and regional differences regarding the raw plant. In positive ionization mode, 1753 features (metabolites) were detected in the samples and 116 features were given annotation (6,6%) from spectral libraries matching, with specified parameters, with a total of 57 unique compounds annotated. Structurally related molecules form *clusters* in the FBMN analysis, which can be further interpreted in terms of chemical class majority and abundance in samples. From the whole features set, 577 nodes formed clusters (>= 2 nodes), while the remaining are considered singletons (isolated nodes). Figure 2 illustrates the resulting positive mode FBMN without singletons, displaying the corresponding butter/oil ratio for each node:

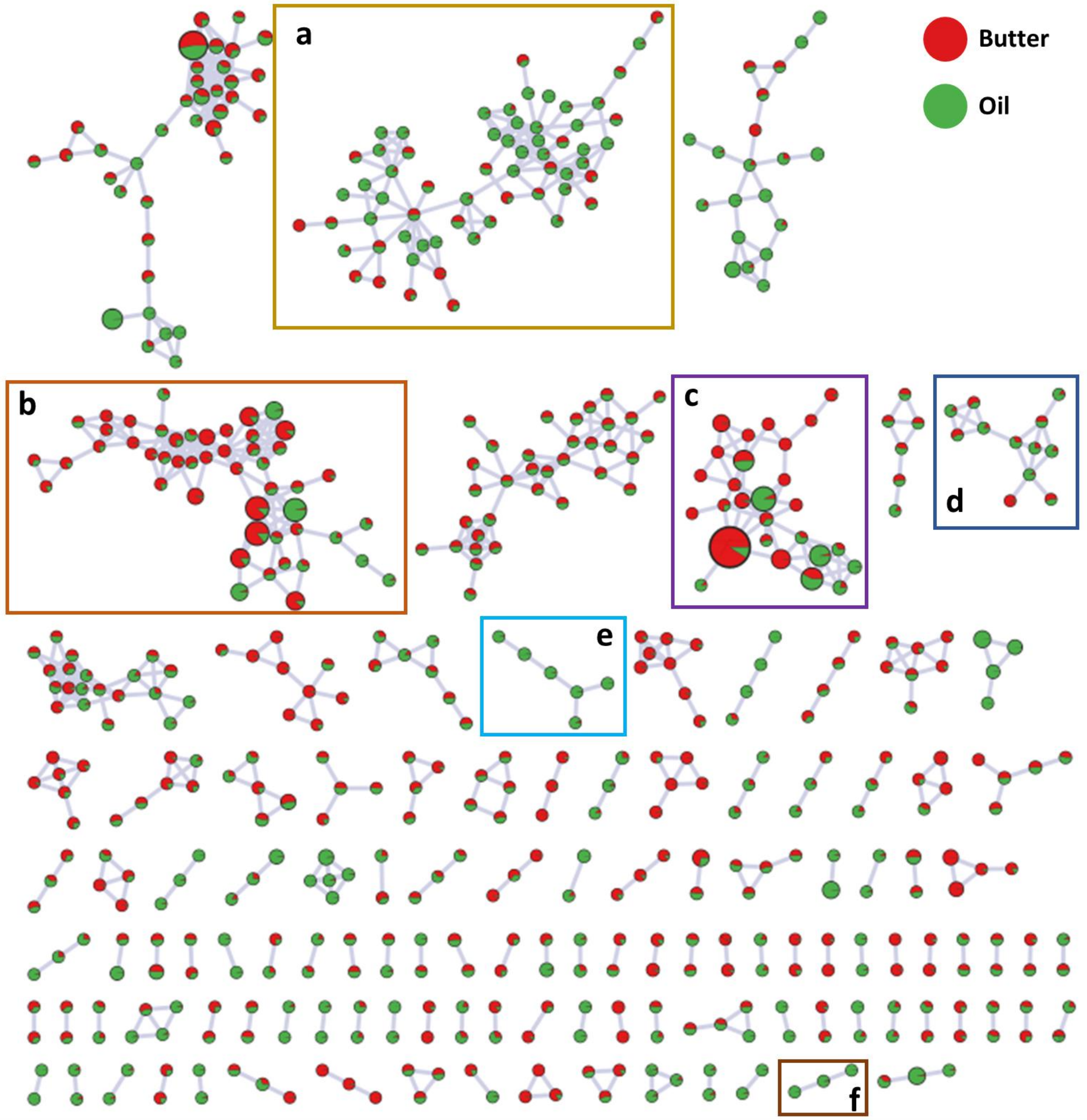

Figure 2: Molecular Networking showing oil/butter ratio on each node.The boxes a, b, c, d, e are the highlighted clusters, described above. Size is scaled from lower to higher intensity values.

Chemical annotation showed that tucumã butter and oil fractions share a similar chemical composition, with the majority of the obtained hits being lipids and lipid-like molecules (55%), including fatty acids, prenol lipids and glycerolipids. Additionally, sample secondary metabolites include phytosterols, terpenoids, sesquiterpenoids, amines and retinoids.

Singular signatures were detected in the metabolomic analysis (Figure 2). Oil samples, for instance, contain several unique sesquiterpenoids, including nootkatone, a human-safe mosquito repellent (Figure 2a). In contrast, several monoradylglycerols and diradylglycerols glycerolipids are enriched or exclusively found on butter samples (Figure 2b), including monolaurin and monomyristin, known for their antimicrobial activity and application in cosmetics as emulsifier and surfactant, respectively. Saturated triacylglycerols such as lauryl-myristyl-palmitate are prevalent in both samples, while capyl-lauryl-myristate is enriched in butter and both butter and oils samples present a singular repertoire of unsaturated fats containing linoleic acid and derivatives (Figure 2c and 2d). Monoterpenes, and prenol lipids such as sesquiterpenoids, α-Tocopheryl acetate (vitamin E acetate) and vitamin K1 were also annotated in the samples, with retinal, alitretinoin and related compounds being annotated as tucumã oil signatures (Figure 2e and 2f)

Annotated features from FBMN and Spectral matches from metabolomics were curated and subject to *in silico* toxicity assessment. This step is used to evaluate the necessity of further *in vitro* and clinical trials for multiple endpoints. A proprietary AI ensemble was used for predicting the mutagenicity of 57 compounds obtained from chemical characterization. Most of the compounds resulted in predictions within the applicability domains of all three QSAR models that compose the ensemble model. Only 8 compounds were evaluated as outside of applicability domain, representing approximately 14% of the chemicals analyzed. Among the chemicals predicted as non-mutagenic and within the applicability domain, the majority of predictions presented a reliability score >70%, while only three compounds showed low reliability scores. Only one chemical was predicted as mutagenic and with a low reliability score (<70%). These results indicated the need for *in vitro* testing to investigate the mutagenic potential of pulp oil and seed butter. The Ames test was conducted based on OECD TG 471 protocol and the results showed that the two samples were not mutagenic.

The skin sensitization and irritation were evaluated using a similar weight of evidence approach (in silico, in vitro and clinical trials), considering public computational tools. The results indicated that the pulp oil and seed butter are not skin sensitizers or irritants.

Discussion

Vegetable oils are important plant-derived ingredients commonly used as emollients in cosmetic formulations owing to their unique chemical composition, rich in lipids and antioxidants. These components contribute to improve skin barrier function, reduce transepidermal water loss (TEWL) and inflammation, and improve sensory attributes such as softness [53].

Due to the extensive chemical diversity found in plant oily ingredients, a high number of pharmacological activities can be found, according to lipidic composition and fatty acid (FA) profile. These ingredients are mainly constituted by triglycerides (or triacylglycerols, TGs) with a variety of FA chains and small fractions of free FAs. Among lipid constituents, unsaturated fatty acids, particularly in their free form, exhibit remarkable cosmetic properties. Carotenoids, retinoids and tocopherols are often associated with antioxidant bioactivity; linoleic acid and α-linoleic acid can be used to maintain barrier integrity, incorporating into cell membranes and regenerating damaged lipid barrier of epidermis, contributing to transepidermal water loss (TEWL) reduction; unsaturated FA can reduce inflammation and are associated to healing effects (for example: linoleic acid, linolenic acid, oleic acid). Free FAs can also act as permeability enhancers for different compounds present in these oils, for example oleic acid [53, 54 e 55]. Despite the prominence and abundance of lipid molecules and FA in these ingredients, other phytochemical constituents also contribute and may play major roles in terms of bioactivity. The described *in silico* pipeline provides a powerful exploration tool for bioingredients, enabling researchers to study the intricate compositions, efficiently screen vast libraries and highlight unique properties and potential cosmetic benefits for innovative, effective and safe products.

In the case of Tucumã development, the antioxidant constituents, such as tocopherol and carotenoids, together with experimental reports and major FA content of pulp oil (47% oleic, 26% linoleic, 14% palmitic, 10% stearic acids) and seed butter (50% lauric, 24% myristic acids) pointed to possible cosmetic uses. The metabolomics and molecular networking analysis identified minor components, 118 clusters, 116 annotations, 57 of them unique, and a comparative picture of shared and unique clusters.

A blend from Tucumã oil and butter could provide potent benefits from its mixed composition. The chemical profile, described by a balanced lipid content, can be beneficial for skin hydration, anti-inflammatory activities, and also possibly enhance the delivery of metabolites, such as a present Vitamin E form through oleic acid. Coupled with gene expression results, viewed from a gene set with cosmetic interest perspective, it gave deeper insights into potential mechanisms of action, with ECM, wound healing, antioxidation, anti-inflammatory and hyaluronic acid emerging as key targets to further investigation. Hyaluronic Acid is a biopolymer of high cosmetic interest, being a major constituent in ECM and playing an important role in skin hydration through water retention. It is also part of many other processes with space filling and wound healing capacity and regulating several aspects of tissue repair through activation of inflammatory cells and cell migration [56].

Conclusion.

This data-assisted pipeline provides a proven, animal test free and cost-effective screening process to discover cosmetic applications of novel ingredients. Coupling Omics, Bioinformatics and AI technologies can help researchers overcome some of the challenges of studying complex and partially characterized natural products and unravel their effects on skin metabolism, to develop safe, potent and sustainable cosmetic products. These methods speed up and assist the bioprospecting process, providing a general overview of potential uses for new, and even understudied ingredients from biodiversity, leading to more subsequent studies to unveil details for any target characteristic or function.

This workflow is a constant work-in-progress task and continuous acquisition of data through the pipeline usage, combined with development of new AI models and integration of multi-OMICs experiments could lead to improved levels of extracts characterization and discovery of novel properties.

Acknowledgments.

The funding of the present work was met by the Natura Research & Development Department.

Conflict of Interest Statement.

None.

References.

1. Giannenas, I., et al. (2020). The history of herbs, medicinal and aromatic plants, and their extracts: Past, current situation and future perspectives. *In Feed additives* (pp. 1-18). Academic Press.
2. Newman, D. J., & Cragg, G. M. (2020). Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. Journal of natural products, 83(3), 770-803.
3. Nasim, N., et al. (2022). Plant-derived natural products for drug discovery: current approaches and prospects. *The Nucleus*, 65(3), 399-411.
4. de Matos, R. C., et al. (2024). Evidence for the efficacy of anti-inflammatory plants used in Brazilian traditional medicine with ethnopharmacological relevance. *Journal of Ethnopharmacology*, 118137.

5. Theodoridis, G., et al (2011). Mass spectrometry-based holistic analytical approaches for metabolite profiling in systems biology studies. *Mass spectrometry reviews*, *30*(5), 884-906

6. Glauser, G., et al (2013). Ultra-high pressure liquid chromatography–mass spectrometry for plant metabolomics: A systematic comparison of high-resolution quadrupole-time-of-flight and single stage Orbitrap mass spectrometers. *Journal of Chromatography A*, 1292, 151-159.

7. Salem, M. A., et al (2020). Using an UPLC/MS-based untargeted metabolomics approach for assessing the antioxidant capacity and anti-aging potential of selected herbs. RSC advances, 10(52), 31511-31524.

8. Lv, J., et al (2024). Untargeted Metabolomics Based on PLC-Q-Exactive-Orbitrap-MS/MS Revealed the Differences and Correlations between Different Parts of the Root of Paeonia lactiflora Pall. Molecules, 29(5), 992.

9. Shen, S., et al (2023). Metabolomics-centered mining of plant metabolic diversity and function: Past decade and future perspectives. Molecular Plant, 16(1), 43-63.

10. Fiehn, O., et al (2000). Metabolite profiling for plant functional genomics. Nature biotechnology, 18(11), 1157-1161.

11. Wang, Z., et al (2023). Comparative antioxidant activity and untargeted metabolomic analyses of cherry extracts of two Chinese cherry species based on UPLC-QTOF/MS and machine learning algorithms. Food Research International, 171, 113059.

12. Sirocchi, C., et al (2024). Exploring machine learning for untargeted metabolomics using molecular fingerprints. Computer Methods and Programs in Biomedicine, 250, 108163.

13. Li, X., et al (2022). UHPLC-Q-Exactive orbitrap MS/MS-Based untargeted metabolomics and molecular networking reveal the differential chemical constituents of the bulbs and flowers of Fritillaria thunbergii. Molecules, 27(20), 6944.

14. Wang, X., et al (2023). A Structure-Guided Molecular Network Strategy for Global

Untargeted Metabolomics Data Annotation. Analytical Chemistry, 95(31), 11603-11612.

15. Chitiva, L. C., et al (2023). Untargeted metabolomics approach and molecular networking analysis reveal changes in chemical composition under the influence of altitudinal variation in bamboo species. Frontiers in Molecular Biosciences, 10, 1192088.

16. Hegazi, N. M., et al (2024). Untargeted metabolomics-based molecular networking for chemical characterization of selected Apiaceae fruit extracts in relation to their antioxidant and anti-cellulite potentials. Fitoterapia, 173, 105782.

17. Nothias, L. F., at al (2020). Feature-based molecular networking in the GNPS analysis environment. Nature methods, 17(9), 905-908.

18. Zhao, X., et al (2022). Combination of untargeted metabolomics approach and molecular networking analysis to identify unique natural components in wild Morchella sp. by UPLC-Q-TOF-MS. Food Chemistry, 366, 130642.

19. Sheng, Y., et al (2024). IMN4NPD: An Integrated Molecular Networking Workflow for Natural Product Dereplication. Analytical Chemistry.

20. Kum, E., & İnce, E. (2024). Metabolomics approach to explore bioactive natural products derived from plant-root-associated Streptomyces. Applied Biochemistry and Biotechnology, 1-14.

21. Alves, V. M., et al (2015). Predicting chemically-induced skin reactions. Part I: QSAR models of skin sensitization and their application to identify potentially hazardous compounds. Toxicology and applied pharmacology, 284(2), 262-272.

22. Furuhama, A., et al (2023). Evaluation of QSAR models for predicting mutagenicity: outcome of the Second Ames/QSAR international challenge project. SAR and QSAR in Environmental Research, 34(12), 983-1001.

23. Zhao, Y., et al (2023). QSAR in natural non-peptidic food-related compounds: current status and future perspective. Trends in Food Science & Technology, 104165.

24. Boulaamane, Y., et al (2023). Exploring natural products as multi-target-directed drugs

for Parkinson's disease: an in-silico approach integrating QSAR, pharmacophore modeling, and molecular dynamics simulations. Journal of Biomolecular Structure and Dynamics, 1-18.

25. Paludetti, M. F. et al (2023) Development of an Artificial Intelligence Model for Predicting the Mutagenicity of Natural Products [Poster presentation]. I Brazilian Congress of Alternative Methods to the Use of Animals in Research and Education, Rio de Janeiro, Brazil.

26. Subramanian, A., et al (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences (PNAS), v102, n43.

27. Jardim Botânico do Rio de Janeiro (2024). Flora e Funga do Brasil. Available at: <http://floradobrasil.jbrj.gov.br/>. Access on 5 Jun 2024.

28. Museu de História Natural e Jardim Botânico, Universidade Federal de Minas Gerais (2024) Banco de Dados e Amostras de Plantas Aromáticas, Medicinais e Tóxicas - DATAPLAMT. Available at: <https://www.dataplamt.org.br/>. Access on 5 Jun 2024.

29. Bora, P. S., et al (2001). Characterisation of the oil and protein fractions of tucuma (Astrocaryum vulgare Mart.) Fruit pulp and seed Kernel caracterización de las fracciones protéicas y lipídicas de pulpa y semillas de tucuma (Astrocaryum vulgare Mart.) Caracterización das fraccións protéicas e lipídicas da pulpa e semillas de tucuma (Astrocaryum Vulgare Mart.). CYTA-Journal of Food, 3(2), 111-116.

30. Santos, M. F. G., et al (2015). Carotenoid composition in oils obtained from palm fruits from the Brazilian Amazon.

31. Santos, M. D. F. G. D., et al (2017). Quality characteristis of fruits and oils of palms native to the Brazilian Amazon. Revista Brasileira de Fruticultura, 39, e-305.

32. Pardauil, J. J., et al (2017). Characterization, thermal properties and phase transitions of amazonian vegetable oils. Journal of Thermal Analysis and Calorimetry, 127, 1221-

1229.

33. Baldissera, M. D., et al (2017). Antihyperglycemic, antioxidant activities of tucumã oil (Astrocaryum vulgare) in alloxan-induced diabetic mice, and identification of fatty acid profile by gas chromatograph: New natural source to treat hyperglycemia. Chemico-biological interactions, 270, 51-58.

34. Dos Santos, M. D. F. G., et al (2015). Amazonian native palm fruits as sources of antioxidant bioactive compounds. Antioxidants, 4(3), 591-602.

35. Rodrigues, A. M., et al (2010). Fatty acid profiles and tocopherol contents of buriti (Mauritia flexuosa), patawa (Oenocarpus bataua), tucuma (Astrocaryum vulgare), mari (Poraqueiba paraensis) and inaja (Maximiliana maripa) fruits. Journal of the Brazilian Chemical Society, 21, 2000-2004.

36. Bony, E., et al (2012). Chemical composition and anti-inflammatory properties of the unsaponifiable fraction from awara (Astrocaryum vulgare M.) pulp oil in activated J774 macrophages and in a mice model of endotoxic shock. Plant foods for human nutrition, 67, 384-392.

37. Baldissera, M. D., et al (2018). Tucumã oil (Astrocaryum vulgare) ameliorates hepatic antioxidant defense system in alloxan-induced diabetic mice. Journal of food biochemistry, 42(2), e12468.

38. Gualberto, L. D. S. (2022). Obtenção e caracterização dos óleos obtidos dos frutos Tucumã (astrocaryum vulgare), Pupunha (Bactris gasipaes) e Bacupari (Garcinia gardneriana).

39. Bereau, D., Benjelloun-Mlayah, B., Banoub, J., & Bravo, R. (2003). FA and unsaponifiable composition of five Amazonian palm kernel oils. Journal of the American Oil Chemists' Society, 80, 49-53.

40. Chambers, M. C., et al (2012). A cross-platform toolkit for mass spectrometry and proteomics. Nature biotechnology, 30(10), 918-920.

41. Schmid, R., et al (2023). Integrative analysis of multimodal mass spectrometry data in MZmine 3. Nature biotechnology, 41(4), 447-449.

42. Heuckeroth, S., et al. (2024). Reproducible mass spectrometry data processing and compound annotation in MZmine 3. Nature protocols, 1-45.

43. Wang, M., et al (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nature biotechnology, 34(8), 828-837.

44. Djoumbou Feunang, Y., et al (2016). ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. Journal of cheminformatics, 8, 1-20.

45. Rump, L. V., et al (2010). Comparison of commercial RNA extraction kits for preparation of DNA-free total RNA from Salmonella cells. BMC Research Notes, 3, 1-5.

46. Burdukiewicz, M., et al (2021). PCRedux: A Data Mining and Machine Learning Toolkit for qPCR Experiments. bioRxiv, 2021-03.

47. Vandesompele, J., et al (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome biology, 3, 1-12.

48. Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method. Methods, 25(4), 402-408.

49. Edgar, R., et al (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research, 30(1), 207-210.

50. Ashburner, M., et al (2000). Gene ontology: tool for the unification of biology. Nature genetics, 25(1), 25-29.

51. Gene Ontology Consortium. (2023). The gene ontology knowledge base in 2023. Genetics, 224(1).

52. Milacic, M., et al (2024). The reactome pathway knowledgebase 2024. Nucleic acids research, 52(D1), D672-D678.

53. Ahmad, A., & Ahsan, H. (2020). Lipid-based formulations in cosmeceuticals and

biopharmaceuticals. Biomedical Dermatology, 4, 1-10.

54. Park, J., et al (2021). Bioactive lipids and their derivatives in biomedical applications. *Biomolecules & Therapeutics*, 29(5), 465.

55. Fernandes, A., et al (2023). A systematic review of natural products for skin applications: Targeting inflammation, wound healing, and photo-aging. *Phytomedicine*, 154824.

56. Papakonstantinou, E., et al (2012). Hyaluronic acid: A key molecule in skin aging. *Dermato-endocrinology*, 4(3), 253-258.