ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Cheminformatics Exploration of Structural Physicochemical Properties, Molecular Fingerprinting, and Diversity of the Chemical Space of Compounds from Betel Nut (*Areca catechu* L.)

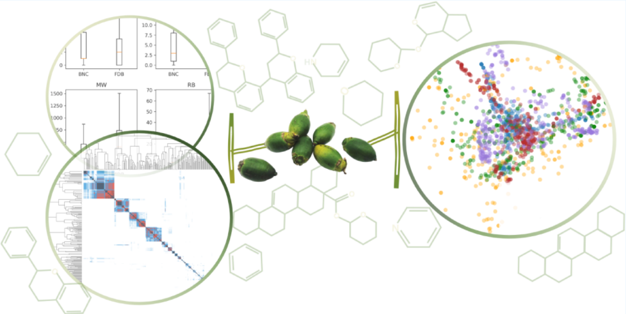
Yubing Li, Xinyue Wang, Haixuan Sun, Hongxin Wang, and Chaoyang Ma*

Cite This: *ACS Omega* 2025, 10, 1551−1561

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In this work, the characterization and diversity of 347 compounds from betel nut (*Areca catechu* L.) were analyzed for the first time. The dataset of compounds from betel nut (BNC) was compared to compounds from food. They were analyzed in terms of physicochemical properties, scaffold diversity, molecular fingerprints, and global diversity. Approximately 48% of compounds in the BNC confirm Lipinski's and Pfizer's rules. The pharmacological and toxicological properties of edible betel nut were evaluated based on their composition. This work applied the research methods of cheminformatics to food science, and it provided theoretical support and data for betel nut pharmacological research, development of betel nut-related novel medication, and healthy products.



1. INTRODUCTION

Betel nut (*Areca catechu* L.) is an evergreen tree widely cultivated in tropical and subtropical regions, and its fruit is highly valued for its unique stimulant properties and a wide range of medicinal uses. Although chewing habits vary among nations and regions, betel nut fruit and its products have been consumed in Asia and the Pacific region for a very long time. This traditional food has been chewed for over 10 000 years and is the most widely used psychoactive substance outside of alcohol, tobacco, and caffeine.[1] Edible betel nut is distinctly different from the medicinal betel nut. Chewing edible betel nuts may greatly boost anxious excitement, reduce weariness, and renew the mind.[2] The World Health Organization designated betel nut as a Class I carcinogen back in 2004. With the rise in health consciousness in recent years, there has been a lot of discussion about the link between excessive betel nut use and several health issues (such as gastrointestinal disorders, oral cancer, etc.).[3]

Whether it is used as medicine to kill worms and alleviate food stagnation or as food to relieve fatigue, its active function is highly related to the unique composition of betel nut. Studies have shown that the active ingredients in betel nut mainly include alkaloids (e.g., arecoline, arecaidine, guvacoline, and guvacine), tannins, flavonoids, fatty acids, etc.[4−6] These chemical components not only give betel nut its distinct pharmacological actions and flavor profiles but also serve as the foundation for research into its health benefits and possible medical properties. Currently, knowledge of these chemical constituents in terms of physicochemical properties and biological activities is rather restricted. However, the diversity and complexity of the chemical constituents of betel nut make it challenging to systematically study its chemical structure and biological activities through conventional experimental methods.

As a combination of chemistry and informatics, cheminformatics involves effectively managing, analyzing, and interpreting chemical data through computer technology and information processing. Cheminformatics enables the construction of chemical databases, molecular similarity searches, combinatorial library design, molecular cluster analysis, structure−activity relationships, and chemical space exploration, among others. This facilitates a better understanding of the physicochemical properties, biological activity, and structural diversity of chemical substances, thereby allowing for their effective design and application.

Cheminformatics has been instrumental in the construction of molecular structural fingerprint libraries, as evidenced by various successful studies. Olmedo et al. employed a cheminformatics approach to characterize 354 natural products from Panama, elucidating the profound structural complexity inherent in natural products and validating their potential as a reservoir for compounds in virtual screening campaigns.[7] Similarly, Avellaneda-Tamayo et al. conducted a comprehen-

Received: October 14, 2024
Revised: December 16, 2024
Accepted: December 19, 2024
Published: December 30, 2024

https://doi.org/10.1021/acsomega.4c09386

sive analysis of the physicochemical properties and structural descriptors of food chemicals, revealing that these components exhibit reduced scaffold and fingerprint-based diversity while presenting heightened structural complexity.[8] By employing cheminformatics methodologies, we enable the targeted screening and precise identification of bioactive compounds within a curated compound collection. Our study allows for the acquisition of more refined data by transitioning from broad, multiplant studies to more focused, single-plant investigations. This strategic shift facilitates a more profound understanding of the intricate pharmacological effects of *Areca catechu* L., which is essential for the development of targeted therapeutics and the elucidation of the plant's medicinal properties within specific therapeutic domains.

In this study, cheminformatics tools were utilized to establish a molecular structural fingerprint library for the compounds isolated from betel nut. The study aimed to visualize and analyze the physicochemical properties, scaffold diversity, molecular fingerprints, and global diversity of phytochemicals within the library. This work not only improves understanding of the betel nut's chemical properties and possible bioactivities but also provides theoretical support and data for betel nut pharmacological research and the development of betel nut-related novel medication and healthy products.

2. METHODS

2.1. Data Collection and Preparation. Betel nut is both a traditional Chinese medicine and a chewing addiction; therefore, data on its composition was collected from available natural product libraries and Chinese medicine composition databases. The main databases utilized were TCM-ID (https://bidd.group/TCMID), Hit 2.0 (http://hit2.badd-cao.net), TCMSP (https://old.tcmsp-e.com), BATMAN-TCM (http://bionet.ncpsb.org.cn/batman-tcm), SymMap (http://www.symmap.org), NAPSS (https://bidd.group/NPASS/index.php), ETCM (http://www.tcmip.cn/ETCM), TCMSI (https://tcm.scbdd.com), TCMIO (http://tcmio.xielab.net), and IMPPAT (https://cb.imsc.res.in/imppat), where the search term "*Areca catechu*" was used in the natural product database and "Da fu pi", "Da fu mao", "Jiao bing lang", and "Bing lang" were used in the traditional Chinese medicine ingredient database.

Inorganic compounds and mixtures were eliminated from the dataset, and salts were transformed into the appropriate acids or bases. A total of 346 nonrepetitive compounds were collected, together with their SMILES, and the .sdf files for each compound were downloaded separately from PubChem. Meanwhile, a dataset of 18 556 food-derived chemicals from FooDB (https://foodb.ca) was utilized for comparison.

2.2. Physicochemical Properties Analysis. The number of hydrogen bond donors (HBD), the number of hydrogen bond acceptors (HBA), the octanol/water partition coefficient (SlogP), the molecular weight (MW), the number of rotatable bonds (RB), and the topological polar surface area (TPSA) were calculated for all molecules in the dataset. These six most important molecular features were calculated by RDKit and further statistically analyzed and visualized.[9]

2.3. Scaffold Analysis. Murcko scaffold analysis was conducted to identify the core structural features of the compound, removing the side chains of the molecules and retaining only the core ring structure and connecting linkers, thus allowing compounds with similar backbones to be identified and compared.[10,11] Murcko scaffolds were identified and drawn by RDKit.[9]

2.4. Molecular Fingerprints. The Morgan fingerprints,[12] the RDK fingerprints,[9] the MACCS structural keys,[13] the Topological Torsion fingerprints[14] and the Atom Pair fingerprints[15] were generated by RDKit[9] with the following calculated parameters (Table 1):

Table 1. List of Molecular Fingerprints Evaluated in this Study

Name	Category	Size	Source	Parameter
Topological Torsion[14]	Path	4096	RDKit	targetSize = 4
Morgan[12]	Circular	2048	RDKit	Radius = 2
MACCS[13]	Substructure	166	RDKit	N.A
Atom Pairs[15]	Path	4096	RDKit	N.A
RDK[9]	Path	2048	RDKit	Depth = 7

2.5. Molecular Fingerprint Similarity. The similarity of the generated molecular fingerprints was calculated using RDKit,[9] and the similarity metric used for all five fingerprints was the Tanimoto similarity. For molecules expressed using bit-vector molecular fingerprints, this similarity coefficient follows the following definition:

$$\mathrm{Tc}(\mathrm{A},\,\mathrm{B}) = \frac{c}{a + b - c}$$

where the Tanimoto coefficient of similarity (Tc) between molecules A and B is a function of the number of features present in molecules A and B, respectively (i.e., *a* and *b*), and the number of features common between molecules A and B (*c*). Thus, depending on the type of fingerprint generated, the specific structural features of the molecular fingerprints differ, resulting in slightly different calculated Tc values.[16]

2.6. Global Diversity Analysis. The global diversity of the dataset was assessed using a Consensus Diversity Plot (CDP), which simultaneously represents four diversity criteria in two dimensions: structure based on pairwise molecular fingerprint similarity as described in Subsection 2.5, scaffolding based on Murcko scaffolds calculated as described in Subsection 2.3, physicochemical properties based on the six attributes described in Subsection 2.2, and dataset size based on the number of all compounds.[17] The structural diversity of each dataset is represented on the *X*-axis and is defined as the median of the Tanimoto coefficient based on the Morgan fingerprint. The scaffold diversity for each dataset is represented on the *Y*-axis and is defined as the area under the corresponding scaffold recovery curve.[18] The Euclidean distance was used to measure diversity based on physicochemical features for six attributes (SlogP, TPSA, AMW, RB, HBD, and HBA). The relative quantity of substances in the collection is represented by data points of varying sizes.[19]

2.7. Chemical Space Visualization. A two-dimensional representation of chemical space was generated by applying the PCA approach to visualize and express the derived molecular characteristics based on various molecular fingerprints. Matplotlib was used to create all of the graphics in the study.[20]

3. RESULTS AND DISCUSSION

3.1. Overview of the Dataset. For 347 components drawn from 10 datasets, the names "*Areca catechu*," "Da fu pi," "Da fu mao," "Jiao bing lang," and "Bing lang" were used. The selected databases encompassed the herbal ingredients data-

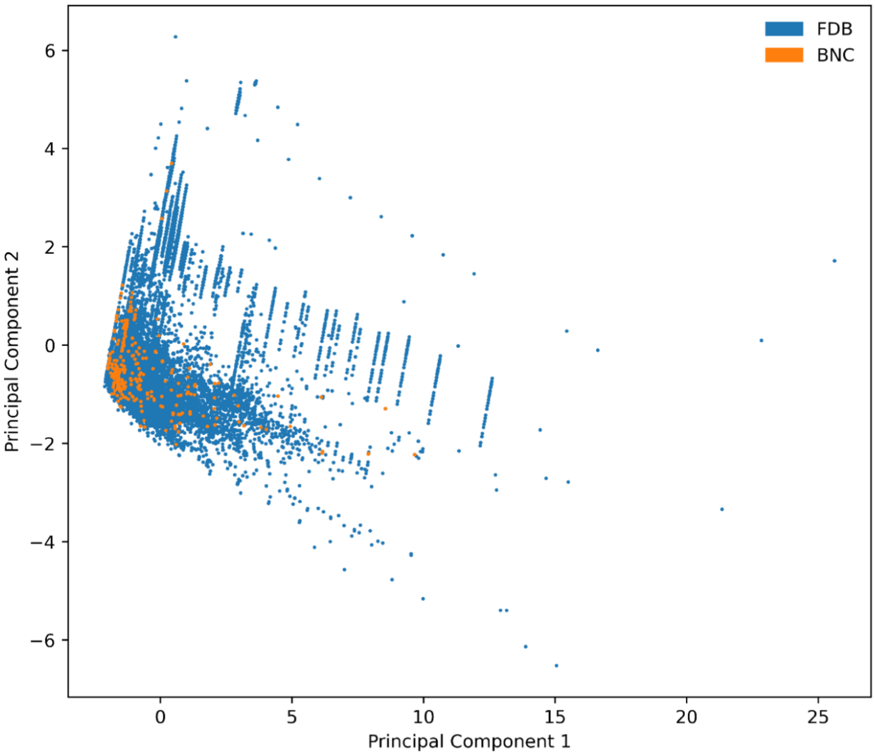


Figure 1. PCA of the BNC and the FDB for the six fundamental physicochemical parameters of them.

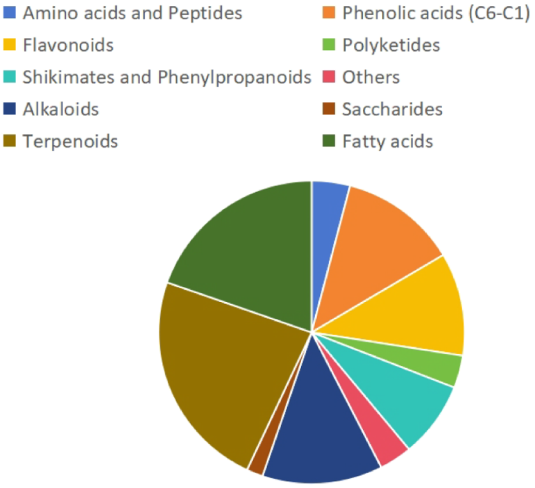


Figure 2. Classification of the compounds in BNC.

base, natural products database, and Indian medicinal plants database. As illustrated in Figure 1, the visualization of the self-constructed dataset and the FooDB database was defined based on the six fundamental physicochemical parameters (HBD, HBA, SlogP, MW, RB, and TPSA). Orange data points represent the betel nut composition dataset (BNC), and blue data points represent the FooDB dataset (FDB). The results show that the BNC can be well covered by the FDB and that compounds from them have similar physicochemical parameters.

All the compounds in BNC were categorized based on NPClassifier, and the results (Figure 2) showed that BNC includes a high concentration of terpenoids and fatty acids (23% and 20%, respectively), followed by alkaloids, phenolic acids, flavonoids, as well as shikimates and phenylpropanoids, which contribute to 13%, 12%, 11%, and 8% of the total number of compounds. The categories of compounds in the database coincide with the results of current studies on the composition of betel nut. Many studies have found polyphenols, flavonoids, triterpenoids, fatty acids, and alkaloids from betel nut,[21] among which polyphenols, flavonoids, and alkaloids are the most studied. These components not only have a high proportion in betel nut but are also the main bearers of physiological activity and toxicity.[22] Although fatty acid compounds also occupy a high proportion in betel nut, research on them is relatively limited. Existing studies indicate that the fatty acids and fatty alcohols with different alkyl chain lengths in betel nut have certain insecticidal activity, which is mainly related to the balance between hydrophilicity and hydrophobicity of fatty compounds.[23]

3.2. Physicochemical Properties Analysis. The six physicochemical properties of the betel nut constituent dataset (BNC) and the reference dataset (FDB) were computationally analyzed using RDKit, resulting in Figure 3. The statistical data of them are shown in Table 2.

The molecular weight distribution indicated that the average molecular weight of the betel nut compounds was 302.28 Da. Approximately 94% of the betel nut compounds had molecular weights less than 500 Da, compared to 83% of the compounds in FDB. Most of the compounds in BNC are distributed between 300 and 400 Da. The SlogP distribution of betel nut compounds spans from −6 to 18. Most of the calculated HBD values are clustered in the range of 0−5, and the HBA values are clustered in the range of 0−10, which is similar to the compounds in the FDB.

By calculating the physicochemical properties of the compounds, candidate compounds can be classified into lead-like molecules,[24] drug-like molecules,[25] and known drugs. Among them, lead-like molecules require that the

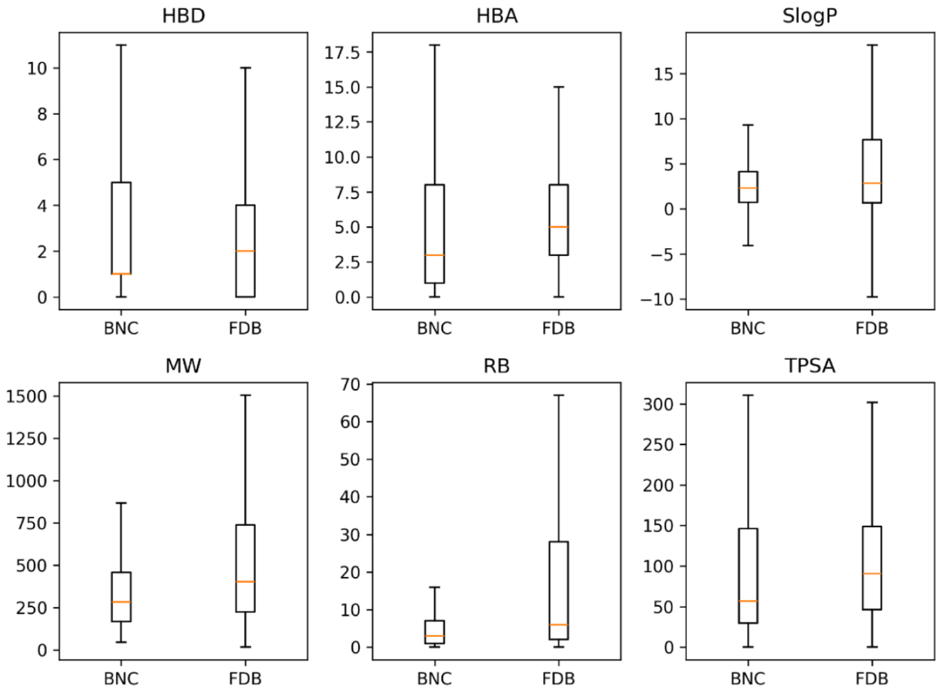


Figure 3. Box plots of the distribution of six physicochemical properties of BNC and FDB.

Table 2. Statistical Distribution of Chemical Descriptors of the Compounds in the BNC and FDB

		HBD	HBA	SlogP	MW	RB	TPSA
Count	ACD	346	346	346	346	346	346
	FDB	18556	18556	18556	18556	18556	18556
Mean	ACD	3.77	6.39	2.56	372.42	5.29	114.01
	FDB	3.51	7.38	4.17	503.83	14.74	127.55
Std	ACD	5.33	8.61	3.06	302.28	6.24	145.69
	FDB	4.96	7.44	5.55	363.58	17.21	134.59
Min	ACD	0	0	−5.3956	46.069	0	0
	FDB	0	0	−30.8741	16.043	0	0
Q1	ACD	1	1	0.685	167.703	1	29.54
	FDB	0	3	0.679075	224.3	2	46.53
Q2	ACD	1	3	2.2963	283.393	3	56.79
	FDB	2	5	2.83632	404.393	6	90.9
Q3	ACD	5	8	4.1261	456.711	7	146.09
	FDB	4	8	7.670875	738	28	148.82
Max	ACD	29	52	18.7691	1871.282	53	877.36
	FDB	73	104	33.8283	4628.234	148	2093.55

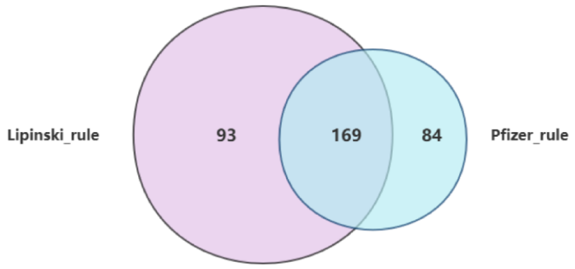


Figure 4. Distribution of compounds from betel nuts based on two rules.

compounds must be low-complexity small molecule compounds with low molecular weight and lipophilicity. Lipinski's Rule (Rof) is a set of five fundamental guidelines frequently used for screening compounds for drug-like molecules.[25] Compounds that follow the Rof have superior pharmacokinetic features and are more likely to have high bioavailability during metabolism. Meanwhile, existing studies have explored the known drug chemical space (KDS), which includes all small-molecule organic compounds that have been evaluated in human clinical trials and subsequently used for therapeutic purposes.[26] The particular molecular descriptors for these three categories are as follows: (1) lead-like: MW $\leq$ 300, logP $\leq$ 3, HBD $\leq$ 3, HBA $\leq$ 3, TPSA $\leq$ 60 Å^2, RB $\leq$ 3; (2) drug-like: MW $\leq$ 500, logP $\leq$ 5, HBD $\leq$ 5, HBA $\leq$ 10, RB $\leq$ 10; (3) KDS: MW $\leq$ 800, logP $\leq$ 6.5, HBD $\leq$ 7, HBA $\leq$ 15, TPSA $\leq$ 180 Å^2, RB $\leq$ 17.

Betel nut has long been a source of herbal medicine, and the physicochemical properties of the compounds in BNC provide the basis for their pharmacological properties. As shown in Figure 4, approximately 76% of the components in the BNC followed Rof, most of which were polyphenols, flavonoids, and terpenoids, which greatly corroborates the potential drug

Table 3. Scaffold Diversity Summary of the Two Databases

Datasets	M	N	NSING	FNM	FNSING	AUC	F_{50}
BNC	346	134	81	0.3862	0.2335	0.7635	0.0970
FDB	18556	3686	2436	0.1986	0.1313	0.8791	0.0046

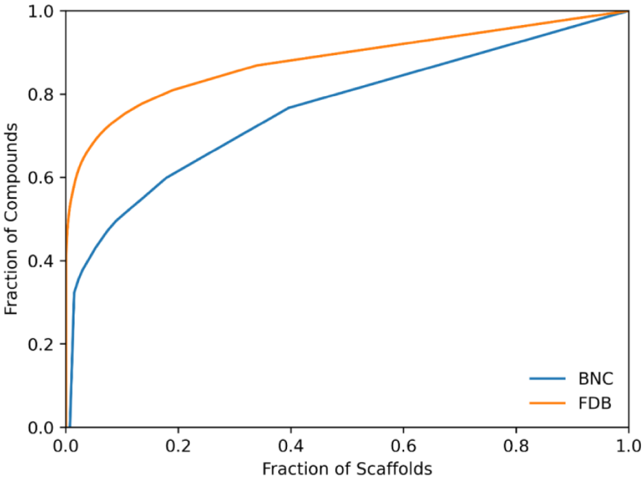


Figure 5. Cyclic system retrieval (CSR) curves for BNC and FDB.

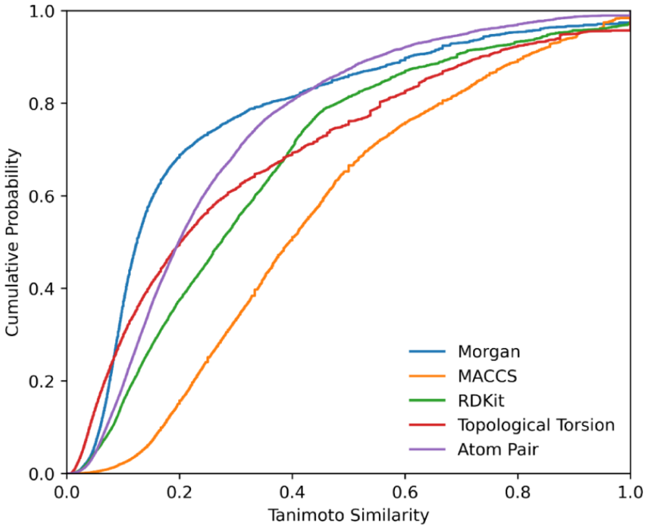


Figure 7. Cumulative distribution function (CDF) plotted with a Tanimoto similarity for different molecular fingerprints.

development possibilities of compounds from betel nut sources. However, most of the current studies on the medicinal properties of betel nut are still based on mixed betel nut extracts, which mainly exhibit anti-inflammatory,[27] antibacterial, and antiparasitic effects,[28] whereas fewer related drug developments start from betel nut compound monomers. The main starting point for research is betel nut alkaloids, which primarily affect the central nervous system of humans.[2] A study found that zebra fish are highly sensitive to screening betel nut alkaloids and related compounds and that novel anxiolytics can be developed using arecoline.[29]

In addition, for the safety of compounds, Pfizer's Rule determines the potential toxicity of compounds. The rule is that when a compound satisfies logP > 3 and TPSA < 75, the compound is considered to be potentially toxic.[30] Approximately 73% of the compounds in the BNC satisfy Pfizer's Rule, and 169 compounds among them satisfy both Lipinski's Rule and Pfizer's Rule. The result can also reflect to some extent that the overchewing of betel nut is probably harmful to humans. It is different from drug-like properties in that most of the potentially toxic compounds are alkaloids.[31] According to existing studies, alkaloids in betel nut are considered to be the main components responsible for diseases such as oral submucous fibrosis and laryngeal cancer.[32] Therefore it is

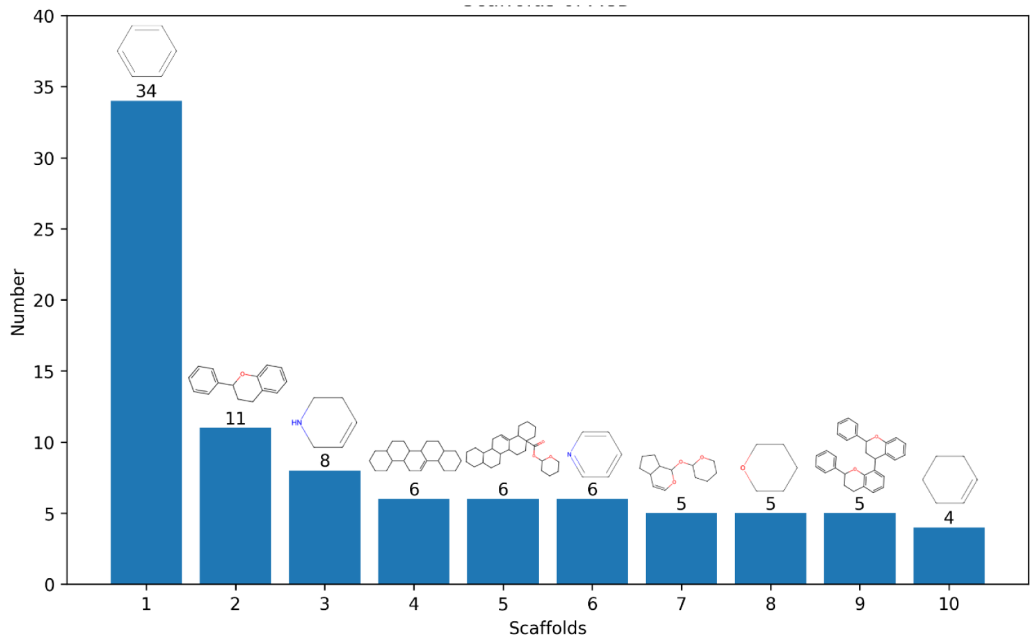


Figure 6. Frequency of the 10 scaffolds in the BNC.

Table 4. Statistical Distribution of Tanimoto Similarity Calculated by Different Molecular Fingerprints

Fingerprint	Average	Q1	Q2	Q3	Max	Min	Std
Morgan	0.1066	0.0526	0.0811	0.1159	1.0	0.0	0.1144
MACCS	0.3001	0.1556	0.25	0.4	1.0	0.0	0.2058
RDKit	0.1615	0.0506	0.1047	0.2145	1.0	0.0	0.1633
Topological Torsion	0.0724	0.0	0.0303	0.0759	1.0	0.0	0.1305
Atom Pair	0.1393	0.0565	0.1022	0.1769	1.0	0.0	0.1301

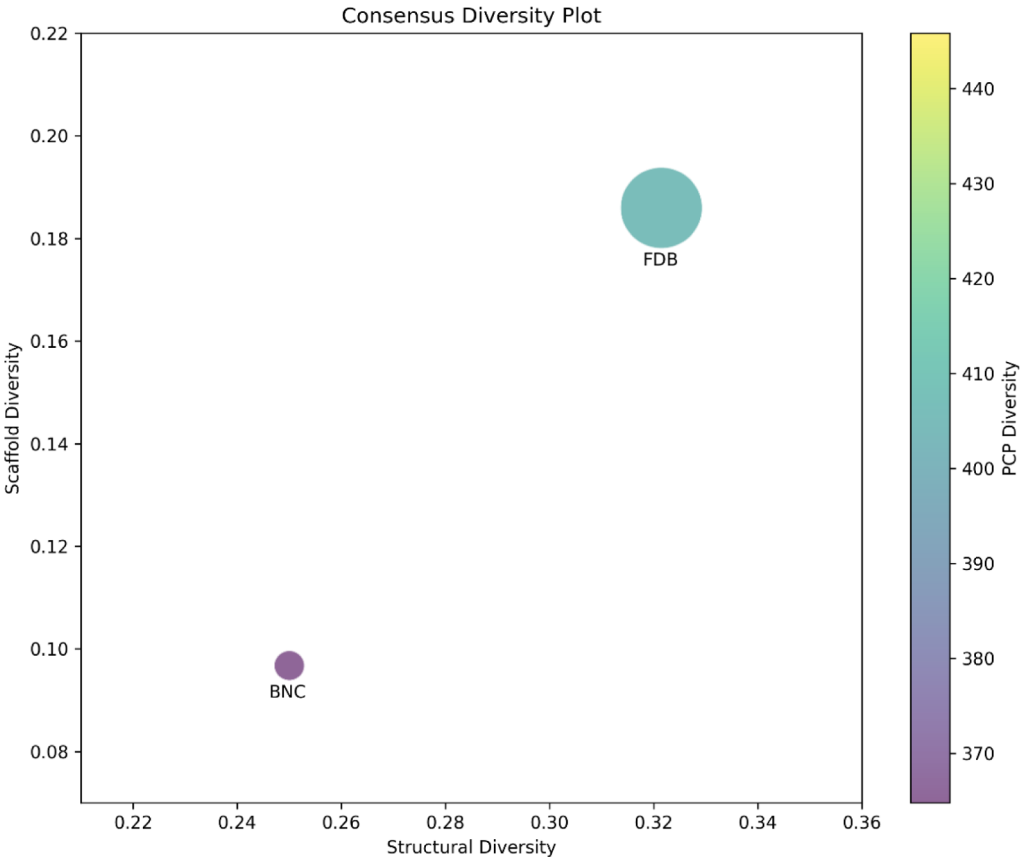


Figure 8. Consensus Diversity Plot of the BNC and FDB.

necessary for drug development using arecoline to avoid their toxic effects as much as possible while making full use of their neurological activity.

3.3. Scaffold Analysis. Compound scaffolds were used to describe the central or core structures of the molecules. The total compounds (M), the unique scaffolds (N), the number of chemotypes containing only one compound (NSING), the chemotype fraction (FNM), the fraction of single chemotypes (FNSING), the area under the curve (AUC), and the chemotype fraction containing 50% of the dataset (F_{50}) were computed for BNC and FDB, respectively. The results are shown in Table 3.

Overall, a total of 134 unique scaffolds were identified for 346 compounds in BNC and 3686 unique scaffolds for 18 556 compounds in FDB, and this number is closely related to the size of the dataset, while the scaffold diversity in BNC is not as rich as that in FDB.

Also, to further illustrate the scaffold diversity of the BNC compounds, it was quantified using a CSR curve with the FDB as a reference. The CSR curve shows the distribution of scaffolds across the compound set. The area under the curve (AUC) and the fraction of scaffolds required to capture half of the compounds (F_{50}) provide a more direct response to the scaffold diversity of the compound set. With the AUC approaching approximately 0.5, the larger the value of F_{50}, the greater the scaffold diversity of the dataset. As shown in Figure 5, the AUCs for BNC and FDB were 0.7635 and 0.8791, respectively, and F_{50} values were 0.0970 and 0.0046, suggesting that the scaffold diversity of BNC was greater than that of FDB. Although the amount of data in BNC is much smaller than that in FDB, BNC still has a diverse range of compounds sourced in food. According to the CSR curves, in the low fraction of scaffolds segment, FDB has a higher slope of the curve, indicating that the low fraction of scaffolds can cover more compounds in the FDB dataset. In the latter half of the curve, the CSR curve of BNC is not as smooth as that of FDB, indicating that the chemical space of BNC can still be further improved, which can be corroborated by the actual research situation. Most studies on the composition of betel nut focused on phenols and alkaloids. At the same time, there are fewer studies on acyclic structures, such as fatty acids.

Figure 6 illustrates the top 10 scaffolds in terms of number in the BNC. One of the most common scaffolds is benzene, with a frequency of 34; it is also the most common scaffold in

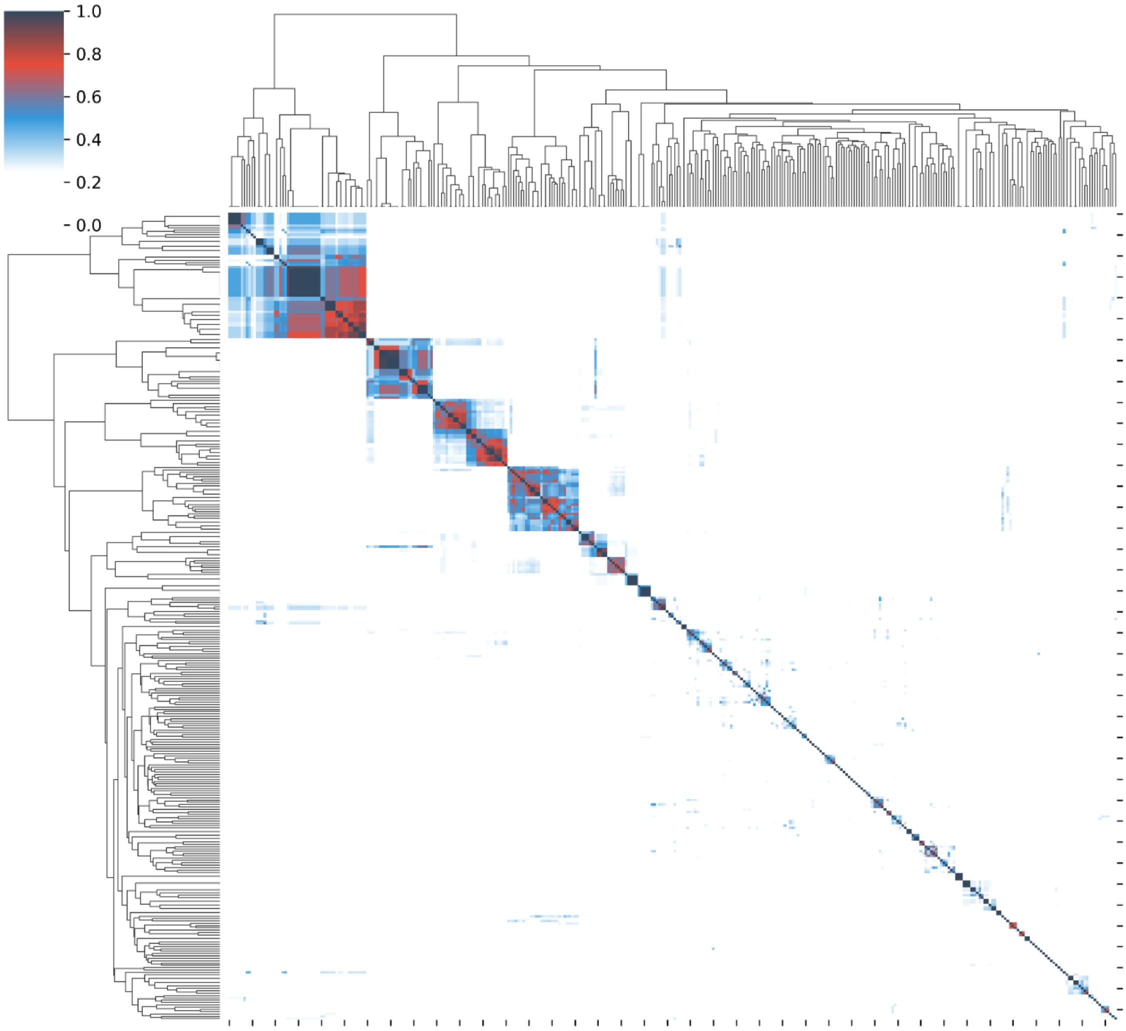


Figure 9. Hierarchical structural clustering of the compounds in the betel nut.

the chemical datasets for drug discovery.[11,33] The representative compounds with benzene are organic acids such as gallic acid and protocatechuic acid, as well as simple phenols such as eugenol, which are potential sources of natural antioxidants.[34] In addition, flavan also covers a high number of compounds in the BNC with a frequency of 11, representing compounds such as epicatechin and lignans. According to relevant studies, betel nut contains such flavan in different parts of the plant (flowers, fruit shells, and seeds). Still, there are significant differences in their respective major compounds.[35] All of them can scavenge free radicals, inhibit angiotensin-converting enzymes, and prevent hypertension, hyperglycemia, etc.[5] Then, pyridine, which occurs with a frequency of 8 in BNC compounds, is mainly a series of pyridine alkaloids, such as arecoline, arecaidine, etc., with insecticidal activity.[36]

3.4. Fingerprint-Based Structural Diversity. This paper calculated the molecular fingerprints of all compounds in the BNC. The MACCS keys, Morgan, RDK, Topological Torsion, and Atom Pair fingerprints were calculated using RDKIT as described in Subsections 2.4 and 2.5 and represented by cumulative distribution function (CDF) plots with Tanimoto similarity.[37] The CDF plots based on different molecular fingerprints are shown in Figure 7. The statistics of molecular similarity among the fingerprint profiles are shown in Table 4. The results show that all the molecular similarities are in the range of 0−1, but the molecular similarities calculated based on different molecular fingerprints are slightly different.[38] The highest average similarity was calculated based on the MACCS keys, which is 0.3001, and the lowest average similarity was calculated based on the Topological Torsion fingerprint, which is 0.0724, while the average similarity calculated based on the remaining two fingerprints were 0.1066 (Morgan fingerprint) and 0.1393 (Atom Pair fingerprint), respectively.

Except for the MACCS keys based on compound substructures, the molecular similarities calculated for the remaining fingerprints were less than 0.2, indicating that the collected betel compounds are somewhat similar in structure, which is consistent with the reliance of MACCS on predefined fragments of the compounds. There are many informative substructures of the natural products that are not defined for small molecules due to the selection of fragments for MACCS keys encoded, resulting in more similar vectors overall.[38]

3.5. Global Diversity Analysis. The diversity of the molecular set varies due to differences in molecular representations, and it is highly correlated with the methods used to quantify diversity.[39] To reduce the dependence of molecular diversity on molecular representations, multiple representations are combined using a consensus diversity plot (CDP).[19] As shown in Figure 8, four diversity metrics for the BNC and FDB were displayed in the graphs, namely, molecular

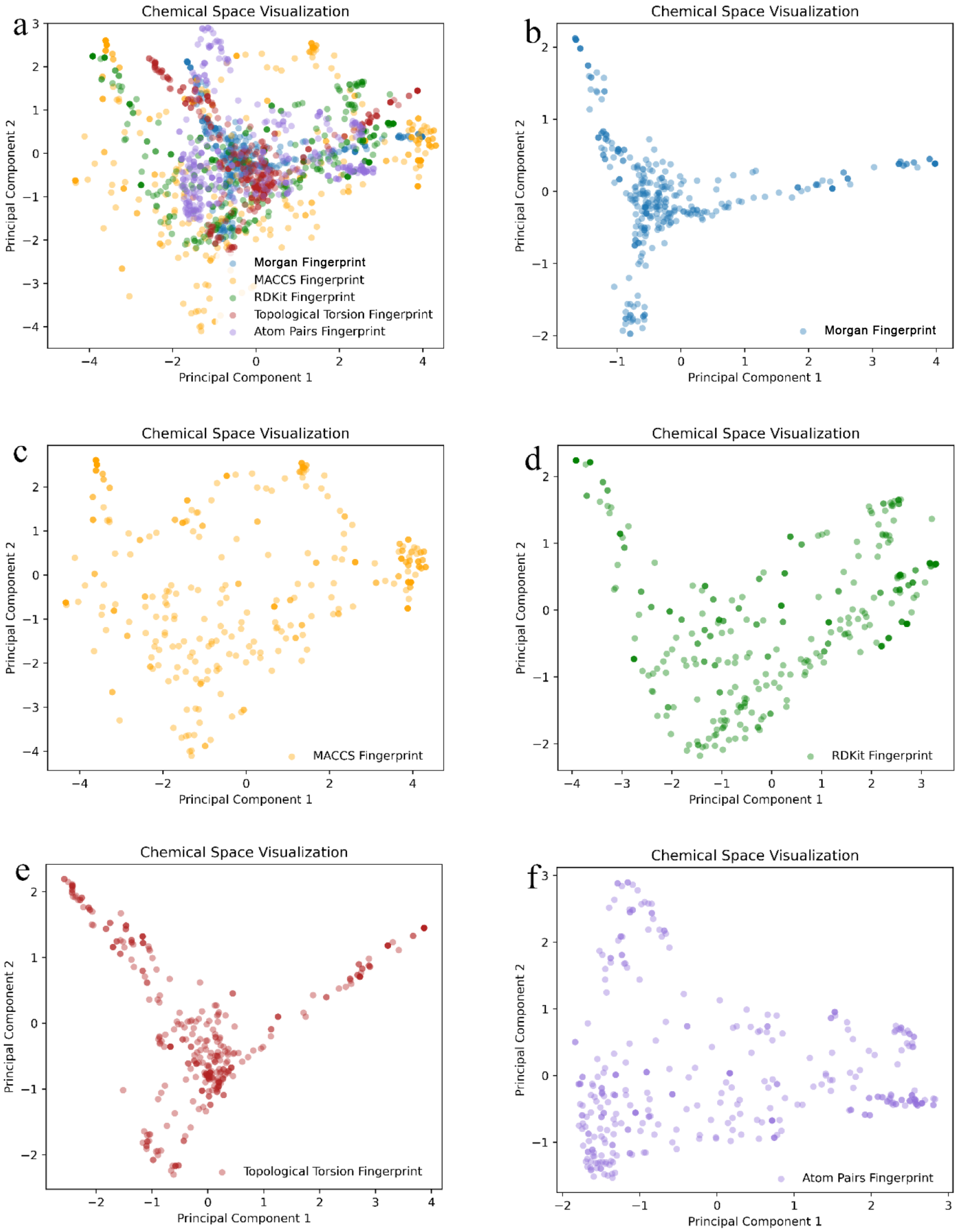


Figure 10. Chemical space of the BNC with different molecular fingerprints. (a) Comparison of chemical spaces constructed using five different molecular fingerprints. (b) The chemical space of the BNC with Morgan fingerprint. (c) The chemical space of the BNC with MACCS fingerprint. (d) The chemical space of the BNC with RDKit fingerprint. (e) Topological Torsion fingerprint. (f) The chemical space of the BNC with Atom Pairs fingerprint.

fingerprints, molecular scaffolds, physicochemical properties, and the relative number of compounds. The global diversity indicated that the FDB dataset has richer structural diversity. It is consistent with the size of the dataset, and the source of compounds in FDB is also richer than in BNC.

3.6. Chemical Space of Betel Nut Components. In order to study the chemical space of betel nut constituent compounds, the all-atom structures were aggregated by hierarchical clustering (Figure 9). The similarity scores (Tanimoto coefficients) of the compounds in the BNC were calculated pairwise.[40] For the similarity cutoff value of 0.7, a total of 214 clusters were generated, and five different sets were primarily generated.

A collection of independent views of the chemical space of compounds in the BNC drawn using different molecular fingerprints is shown in Figure 10, with each point in the view representing a compound. From the visualization results, it is clear that different encoding methods can provide completely different views of the chemical space. Among them, the MACCS keys provided the best dispersion of compounds, and the chemical space generated by the Morgan fingerprint and the Topological Torsion fingerprint had a similar-view profile, showing three distinct clusters of compounds.

4. CONCLUSION

For the first time, compounds from betel nut were integrated and analyzed using molecular fingerprints and chemical space. The PCA map of the BNC dataset was drawn based on the six basic physicochemical properties, with the natural product database FDB as a comparison. The physicochemical property analysis for the compounds in the BNC corroborates the drug and toxicological properties of the betel nut as a medicinal plant and edible hobby product. Most of the compounds in the BNC have drug-like properties, but drug development for the source of betel nut compounds should pay particular attention to the potential toxicity profile. As for the structural diversity of the BNC compounds, we performed molecular scaffolds and structural analysis based on different molecular fingerprints. Although the size of the BNC dataset is much smaller than that of the FDB, the BNC still shows good flexibility in terms of scaffolds, and there exist a few more notable scaffolds of compounds, such as benzene, flavonoids, and pyridines, which all have satisfactory medicinal chemistry properties, especially flavonoids and triterpenoids.[41−43] In particular, the glycosidic forms of flavonoids and triterpenoids are often considered necessary for natural products to exhibit beneficial pharmacokinetic properties.[40,44] Finally, we visualized the chemical space of the BNC using different molecular fingerprints and found that different fingerprints have a significant effect on the view of the chemical space, and compounds in the BNC are more dispersed in the chemical space mapped by the substructure-based MACCS keys. Overall, the application of chemical space in the food field still has some limitations, and the study of the compound dataset of betel nut constituents can provide a basis for better investigation of pharmacological or toxicological effects related to betel nut constituents.

■ ASSOCIATED CONTENT

Data Availability Statement

BNC and FDB datasets are available at Supporting Information. The software we used is open-source and can be found at https://www.rdkit.org/.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c09386.

The SMILES of the compounds from BNC and FDB (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Chaoyang Ma − *School of Food Science and Technology, Jiang Nan University, Wuxi, Jiangsu 214122, China; State Key Laboratory of Food Science and Resources, Jiangnan University, Wuxi, Jiangsu 214122, China;*
Email: machaoyang24@jiangnan.edu.cn

Authors

Yubing Li − *School of Food Science and Technology, Jiang Nan University, Wuxi, Jiangsu 214122, China; State Key Laboratory of Food Science and Resources, Jiangnan University, Wuxi, Jiangsu 214122, China;* orcid.org/0009-0002-3767-1572

Xinyue Wang − *School of Food Science and Technology, Jiang Nan University, Wuxi, Jiangsu 214122, China; State Key Laboratory of Food Science and Resources, Jiangnan University, Wuxi, Jiangsu 214122, China*

Haixuan Sun − *School of Food Science and Technology, Jiang Nan University, Wuxi, Jiangsu 214122, China; State Key Laboratory of Food Science and Resources, Jiangnan University, Wuxi, Jiangsu 214122, China*

Hongxin Wang − *School of Food Science and Technology, Jiang Nan University, Wuxi, Jiangsu 214122, China; State Key Laboratory of Food Science and Resources, Jiangnan University, Wuxi, Jiangsu 214122, China;* orcid.org/0000-0002-7616-9537

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsomega.4c09386

Author Contributions

Y.L.: Methodology, experimental design, formal analysis, and writing—original draft. X.W.: Methodology and formal analysis. H.S.: Formal analysis. H.W.: Resources. C.M.: Resources.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Y.L. acknowledges the support and founding from the School of Food Science and Technology, Jiangnan University, and Hunan Wuzizui Industrial Group Co., Ltd. (grant NO.230256), as well as the useful discussions with X.W. and C.M.

■ ABBREVIATIONS

BNC, The dataset of the compounds in betel nut; FDB, The dataset of the compounds in FooDB; TT, The Topological Torsion fingerprints; AP, The Atom Pair fingerprints; HBD, The number of hydrogen bond donors; HBA, The number of hydrogen bond acceptors; SlogP, The octanol/water partition coefficient; MW, The molecular weight; RB, The number of rotatable bonds; TPSA, Topological polar surface area; Tc, Tanimoto coefficient of similarity; CDP, Consensus Diversity Plot; Rof, Lipinski's Rule; KDS, The known drug chemical space; CSR, Cyclic System Retrieval; CDF, Cumulative distribution function

■ REFERENCES

(1) Singh, A.; Dikshit, R.; Chaturvedi, P. Betel Nut Use: The South Asian Story. *Subst. UseMisuse* **2020**, *55* (9), 1545−1551.

(2) Myers, A. L. Metabolism of the areca alkaloids - toxic and psychoactive constituents of the areca (betel) nut. *Drug Metab. Rev.* **2022**, *54* (4), 343−360.

(3) Liu, P. F.; Chang, Y. F. The Controversial Roles of Areca Nut: Medicine or Toxin? *Int. J. Mol. Sci.* **2023**, *24* (10), 8996.

(4) Zhang, P. Z.; Sari, E. F.; McCullough, M. J.; Cirillo, N. Metabolomic Profile of Indonesian Betel Quids. *Biomolecules* **2022**, *12* (10), 1469.

(5) Song, F.; Tang, M. M.; Wang, H.; Zhang, Y. F.; Zhu, K. X.; Chen, X. A.; Chen, H.; Zhao, X. M. UHPLC-MS/MS identification,

quantification of flavonoid compounds from *Areca catechu* L. extracts and in vitro evaluation of antioxidant and key enzyme inhibition properties involved in hyperglycemia and hypertension. *Ind. Crop. Prod.* **2022**, *189*, 115787.

(6) Machova, M.; Bajer, T.; Silha, D.; Ventura, K.; Bajerova, P. Volatiles Composition and Antimicrobial Activities of Areca Nut Extracts Obtained by Simultaneous Distillation-Extraction and Headspace Solid-Phase Microextraction. *Molecules* **2021**, *26* (24), 7422.

(7) Olmedo, D. A.; González-Medina, M.; Gupta, M. P.; Medina-Franco, J. L. Cheminformatic characterization of natural products from Panama. *Mol. Diversity* **2017**, *21* (4), 779−789.

(8) Avellaneda-Tamayo, J. F.; Chavez-Hernández, A. L.; Prado-Romero, D. L.; Medina-Franco, J. L. Chemical Multiverse and Diversity of Food Chemicals. *J. Chem. Inf. Model.* **2024**, *64* (4), 1229−1244.

(9) *RDKit: Open-Source Cheminformatics Software*. http://www.rdkit.org.

(10) Brown, N.; Jacoby, E. On scaffolds and hopping in medicinal chemistry. *Mini-Rev. Med. Chem.* **2006**, *6* (11), 1217−1229.

(11) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887−2893.

(12) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.

(13) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273−1280.

(14) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsion: A new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27* (2), 82−85.

(15) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25* (2), 64−73.

(16) Mellor, C. L.; Marchese Robinson, R. L.; Benigni, R.; Ebbrell, D.; Enoch, S. J.; Firman, J. W.; Madden, J. C.; Pawar, G.; Yang, C.; Cronin, M. T. D. Molecular fingerprint-derived similarity measures for toxicological read-across: Recommendations for optimal use. *Regul. Toxicol. Pharmacol.* **2019**, *101*, 121−134.

(17) González-Medina, M.; Prieto-Martínez, F. D.; Owen, J. R.; Medina-Franco, J. L. Consensus Diversity Plots: A global diversity analysis of chemical libraries. *J. Cheminf.* **2016**, *8*, 63.

(18) Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Scior, T. Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-Based Measure. *QSAR Comb. Sci.* **2009**, *28* (11−12), 1551−1560.

(19) Naveja, J. J.; Rico-Hidalgo, M. P.; Medina-Franco, J. L. Analysis of a large food chemical database: Chemical space, diversity, and complexity. *F1000Research* **2018**, *7*, 993.

(20) Naveja, J. J.; Medina-Franco, J. L. ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds. *F1000Research* **2017**, *6*, 1134.

(21) Wang, Z.; Guo, Z.; Luo, Y.; Ma, L.; Hu, X.; Chen, F.; Li, D. A review of the traditional uses, pharmacology, and toxicology of areca nut. *Phytomedicine* **2024**, *134*, 156005.

(22) Peng, W.; Liu, Y. J.; Wu, N.; Sun, T.; He, X. Y.; Gao, Y. X.; Wu, C. J. *Areca catechu* L. (Arecaceae): A review of its traditional uses, botany, phytochemistry, pharmacology and toxicology. *J. Ethnopharmacol.* **2015**, *164*, 340−356.

(23) Kiuchi, F.; Miyashita, N.; Tsuda, Y.; Kondo, K.; Yoshimura, H. Studies on crude drugs effective on visceral larva migrans. I. Identification of larvicidal principles in betel nuts. *Chem. Pharm. Bull.* **1987**, *35* (7), 2880−2886.

(24) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'Rule of Three' for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8* (19), 876−877.

(25) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46* (1−3), 3−26.

(26) Bade, R.; Chan, H. F.; Reynisson, J. Characteristics of known drug space. Natural products, their derivatives and synthetic drugs. *Eur. J. Med. Chem.* **2010**, *45* (12), 5646−5652.

(27) Bhandare, A. M.; Kshirsagar, A. D.; Vyawahare, N. S.; Hadambar, A. A.; Thorve, V. S. Potential analgesic, anti-inflammatory and antioxidant activities of hydroalcoholic extract of *Areca catechu* L. nut. *Food Chem. Toxicol.* **2010**, *48* (12), 3412−3417.

(28) Liu, Y. J.; Peng, W.; Hu, M. B.; Xu, M.; Wu, C. J. The pharmacology, toxicology and potential applications of arecoline: A review. *Pharm. Biol.* **2016**, *54* (11), 2753−2760.

(29) Serikuly, N.; Alpyshov, E. T.; Wang, D.; Wang, J.; Yang, L.; Hu, G.; Yan, D.; Demin, K. A.; Kolesnikova, T. O.; Galstyan, D.; Amstislavskaya, T. G.; Babashev, A. M.; Mor, M. S.; Efimova, E. V.; Gainetdinov, R. R.; Strekalova, T.; de Abreu, M. S.; Song, C.; Kalueff, A. V. Effects of acute and chronic arecoline in adult zebrafish: Anxiolytic-like activity, elevated brain monoamines and the potential role of microglia. *Prog. Neuro-Psychopharmacol. Biol. Psychiatry* **2021**, *104*, 109977.

(30) Hughes, J. D.; Blagg, J.; Price, D. A.; Bailey, S.; DeCrescenzo, G. A.; Devraj, R. V.; Ellsworth, E.; Fobian, Y. M.; Gibbs, M. E.; Gilles, R. W.; Greene, N.; Huang, E.; Krieger-Burke, T.; Loesel, J.; Wager, T.; Whiteley, L.; Zhang, Y. Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorg. Med. Chem. Lett.* **2008**, *18* (17), 4872−4875.

(31) Shih, Y. T.; Chen, P. S.; Wu, C. H.; Tseng, Y. T.; Wu, Y. C.; Lo, Y. C. Arecoline, a major alkaloid of the areca nut, causes neurotoxicity through enhancement of oxidative stress and suppression of the antioxidant protective system. *Free Radical Bio. Med.* **2010**, *49* (10), 1471−1479.

(32) Adil, N.; Ali, H.; Siddiqui, A. J.; Ali, A.; Ahmed, A.; El-Seedi, H. R.; Musharraf, S. G. Evaluation of cytotoxicity of areca nut and its commercial products on normal human gingival fibroblast and oral squamous cell carcinoma cell lines. *J. Hazard. Mater.* **2021**, *403*, 123872.

(33) Singh, N.; Guha, R.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A.; Medina-Franco, J. L. Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J. Chem. Inf. Model.* **2009**, *49* (4), 1010−1024.

(34) Wang, R.; Pan, F.; He, R.; Kuang, F.; Wang, L.; Lin, X. Arecanut (*Areca catechu* L.) seed extracts extracted by conventional and eco-friendly solvents: Relation between phytochemical compositions and biological activities by multivariate analysis. *J. Appl. Res. Med. Aromat. Plants* **2021**, *25*, 100336.

(35) Sari, E. F.; Prayogo, G. P.; Loo, Y. T.; Zhang, P.; McCullough, M. J.; Cirillo, N. Distinct phenolic, alkaloid and antioxidant profile in betel quids from four regions of Indonesia. *Sci. Rep.* **2020**, *10* (1), 16254.

(36) Liu, R.; Zheng, M. Y.; Yuan, L.; Liu, Z. L.; Bao, J. Q.; Yang, W. C.; Kong, H. L.; Feng, J. G. Determination of the main alkaloids and their insecticidal activity of extract of nuts against. *Int. J. Trop. Insect Sci.* **2022**, *42* (5), 3563−3570.

(37) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983−996.

(38) Boldini, D.; Ballabio, D.; Consonni, V.; Todeschini, R.; Grisoni, F.; Sieber, S. A. Effectiveness of molecular fingerprints for exploring the chemical space of natural products. *J. Cheminf.* **2024**, *16* (1), 35.

(39) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7* (17), 903−911.

(40) Zhang, R.; Lin, J.; Zou, Y.; Zhang, X. J.; Xiao, W. L. Chemical Space and Biological Target Network of Anti-Inflammatory Natural Products. *J. Chem. Inf. Model.* **2019**, *59* (1), 66−73.

(41) Abdel-Raheem, S. A. A.; Drar, A. M.; Hussein, B. R. M.; Moustafa, A. H. Some oxoimidazolidine and cyanoguanidine compounds: Toxicological efficacy and structure-activity relationships studies. *Curr. Chem. Lett.* **2023**, *12* (4), 695−704.

(42) Drar, A. M.; Abdel-Raheem, S. A. A.; Moustafa, A. H.; Hussein, B. R. M. Studying the toxicity and structure-activity relationships of some synthesized polyfunctionalized pyrimidine compounds as potential insecticides. *Curr. Chem. Lett.* **2023**, *12*, 499−508.

(43) Hussein, B. R. M.; Moustafa, A. H.; Abdou, A.; Drar, A. M.; Abdel-Raheem, S. A. A. Preparation, Agricultural Bioactivity Evaluation, Structure−Activity Relationships Estimation, and Molecular Docking of Some Quinazoline Compounds. *J. Agric. Food Chem.* **2024**, 72 (16), 8973−8982.

(44) Ertl, P.; Schuffenhauer, A.Cheminformatics analysis of natural products: Lessons from nature inspiring the design of new drugs*Natural Compounds as Drugs*Springer200866217−235