Journal of Pharmaceutical Analysis 14 (2024) 100897

Contents lists available at ScienceDirect

Journal of Pharmaceutical Analysis

journal homepage: www.elsevier.com/locate/jpa

Short communication

Glycomol: A pervasive tool for structure predication of natural saponin products basing on MS data

Daotong Zhao [a, b, 1], Chunguo Wang [c, d, 1], Hanyun Qu [b], Qinling Rao [c], Bingqing Shen [d], Yinan Jiang [c], Jiayu Gong [c], Yumiao Wang [c], Di Geng [c], Rui Hong [f], Tao Lu [b, ***], Qing Ni [e, **], Xinqi Deng [a, *]

[a] Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, 100700, China
[b] School of Life Sciences, Beijing University of Chinese Medicine, Beijing, 100029, China
[c] School of Chinese Materia Medica, Beijing University of Chinese Medicine, Beijing, 100029, China
[d] Beijing Research Institute of Chinese Medicine, Beijing University of Chinese Medicine, Beijing, 100029, China
[e] Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing, 100053, China
[f] Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215123, China

ARTICLE INFO

Article history:
Received 6 July 2023
Received in revised form
23 October 2023
Accepted 15 November 2023
Available online 28 November 2023

Due to the multitudinous structural types of glycosylated components, accurate identification of glycosylation modifications and secondary metabolite structures in herbs remains a challenge for natural drug analysis and new drug discovery [1]. To solve this problem, we developed an auxiliary spectral analysis strategy based on Glyco-GNN machine learning techniques (Fig. 1). It is named as “Glycomol”, an auxiliary tool for the identification of glycoside molecules. In Glycomol, the structure of glycosylated components was analyzed in terms of parent nucleus, glycan modification sites, and glycan types. Saponins were taken as representative glycosylated components due to their high structural complexity. To estimate the likelihood of results from this modularization strategy, first we extracted the liquid phase and mass spectrum information of 96 structure-known saponins from multi-stage mass spectrometry data. Then, 24 standard saponins were applied for analysis and a 95.8% hit rate was achieved. Compared with traditional spectrogram library matching, neural network-based identification of specific structure of modification sites and quantities of glycosylated compound components can not only overcome the homogeneity of mass spectrogram fragmentation ions of isomers, but also avoid the low coverage of constructing comparative spectrogram libraries.

By applying Glycomol for liquid chromatography-mass spectrometry (HPLC-MS) data analysis, a total of 1,126 unreported saponin signals were obtained. A total of 533 saponin signals were found credible after data scrubbing. Among them, 54 (10%) saponin signals were further processed, and 52 accurate structural formulas (false discovery rates (FDR) < 0.05) were finally obtained. To sum up, we developed a powerful modularization strategy for novel glycosylated component identification in herbs that could be beneficial for in-depth structural study of natural pharmaceutical chemistry studies. And it can be used as a reference for rapid structural characterization of analogs. Though ultrahigh-performance liquid chromatography/high-resolution mass spectrometry (UPLC/HRMS) signals that can be detected come from ion fragment information, it is important for characterizing an unknown Glyco-molecule. The Glycomol displays a novel predictive flow for glycoside compounds based on massive ion fragment information (ions pool). Based on the completed candidate compound library we constructed, Glycomol achieves satisfactory discrimination on optimal compound structure prediction, which is difficult for current analytical instruments and methods (Fig. 1).

First, diagnostic ions and neutral loss of previous reported 110 saponins, including 4 glycosylation modification groups, 3 aglycone parent, and 55 C17 side chain, were collected for candidate compound structure character library construction (Fig. 1B, Step 1). A total of 1,222 experimental signals are taken as saponins signals by structure character information mapping (Fig. 1B, Step 2). Depending on the information of the segmented saponins information, the structural composition of the target compound, including the core structure of saponins and the identification of modification groups, could be determined theoretically (Fig. 1B,

[*] Corresponding author.
[**] Corresponding author.
[***] Corresponding author.
E-mail addresses: taolu@bucm.edu.cn (T. Lu), niqing@ganyy.cn (Q. Ni), xqdeng@icmm.ac.cn (X. Deng).
[1] Both authors contributed equally to this work.

https://doi.org/10.1016/j.jpha.2023.11.004

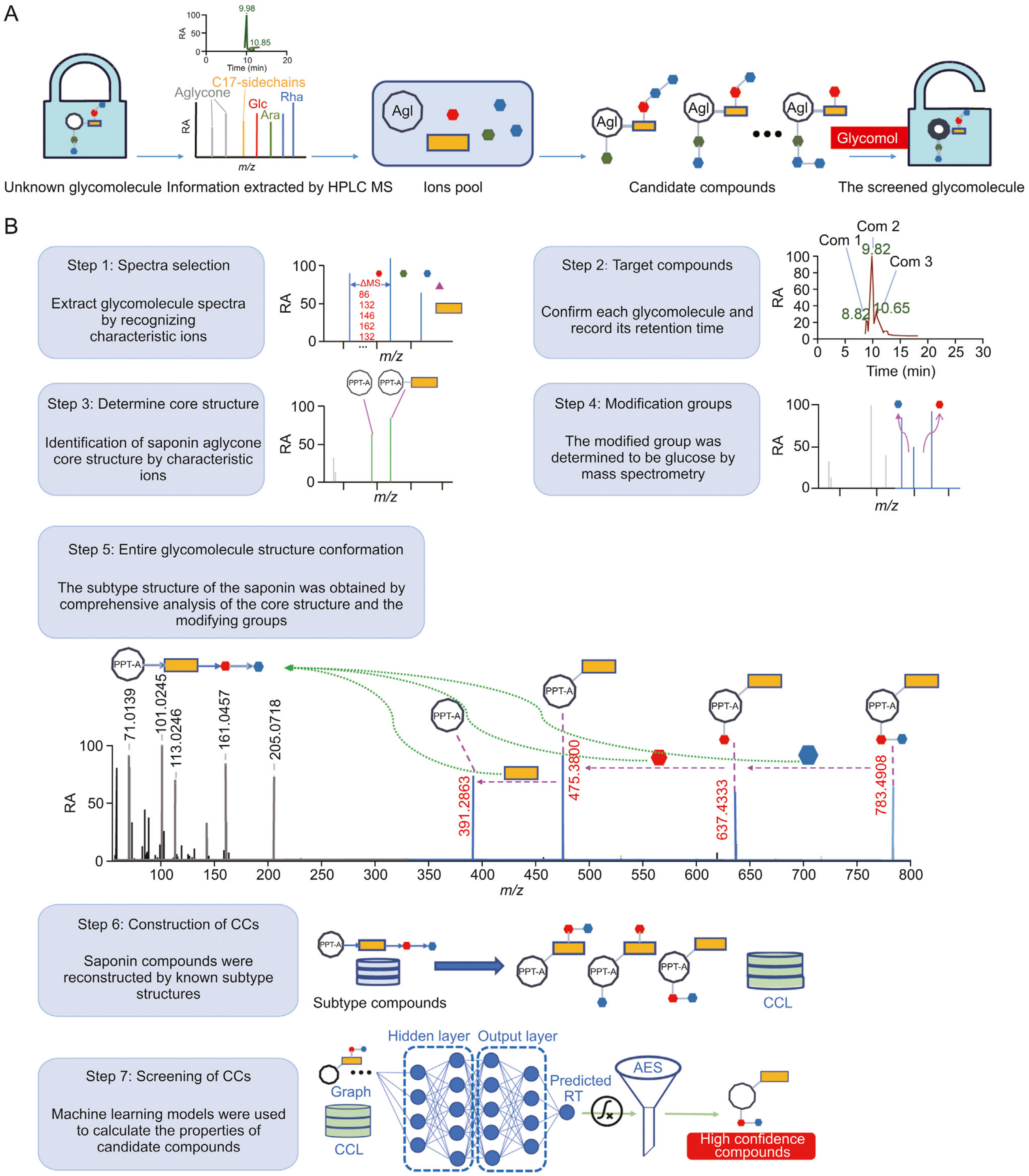
A
9.98
10.85
RA
Time (min)
Aglycone
C17-sidechains
Glc
Ara
Rha
m/z
Agl
Glycomol
Unknown glycomolecule
Information extracted by HPLC MS
Ions pool
Candidate compounds
The screened glycomolecule
B
Step 1: Spectra selection
Extract glycomolecule spectra by recognizing characteristic ions
ΔMS
86
132
146
162
132
...
Step 2: Target compounds
Confirm each glycomolecule and record its retention time
Com 1
Com 2
Com 3
9.82
8.82
10.65
Time (min)
Step 3: Determine core structure
Identification of saponin aglycone core structure by characteristic ions
PPT-A
Step 4: Modification groups
The modified group was determined to be glucose by mass spectrometry
Step 5: Entire glycomolecule structure conformation
The subtype structure of the saponin was obtained by comprehensive analysis of the core structure and the modifying groups
71.0139
101.0245
113.0246
161.0457
205.0718
391.2863
475.3800
637.4333
783.4908
Step 6: Construction of CCs
Saponin compounds were reconstructed by known subtype structures
Subtype compounds
CCL
Step 7: Screening of CCs
Machine learning models were used to calculate the properties of candidate compounds
Graph
Hidden layer
Output layer
Predicted RT
AES
High confidence compounds

Fig. 1. Principle of the new method for structural interpretation of site-specific glycoside molecule. (A) Regional separation of aglycone saponin core and modified group ions under HCD energy fragmentation of intact glycoside molecule and how different ions were used for modular glycan structure interpretation. (B) An example to demonstrate step-by-step interpretation of glycan structures on intact glycoside molecules from MS/MS spectra of low and high MS ranges using the new strategy. RA: relative abundance; CCs: candidate compounds; CCL: candidate compound library; AES: absolute error score; and RT: retention time.

Steps 3, and 4). In this study, 96 of the reported 110 saponins matched with the experimental data. The remaining undefined 1,126 experimental saponins signals were taken as signals of unreported saponins structure (Fig. 1B, Step 5). Next, a total of 533 model-identifiable subtype structure data were obtained following data normalization. Based on the summarized sugar chain fragmentation law, 10% of them were random sampled and applied for candidate structure construction (Fig. 1B, Step 6) and model machine learning (Fig. 1B, Step 7).

Machine learning of Glyco-GNN model was carried out based on encoded molecular sequence. Briefly, user should input mzML file and the retention time (Fig. 2A). (For more information on UPLC-

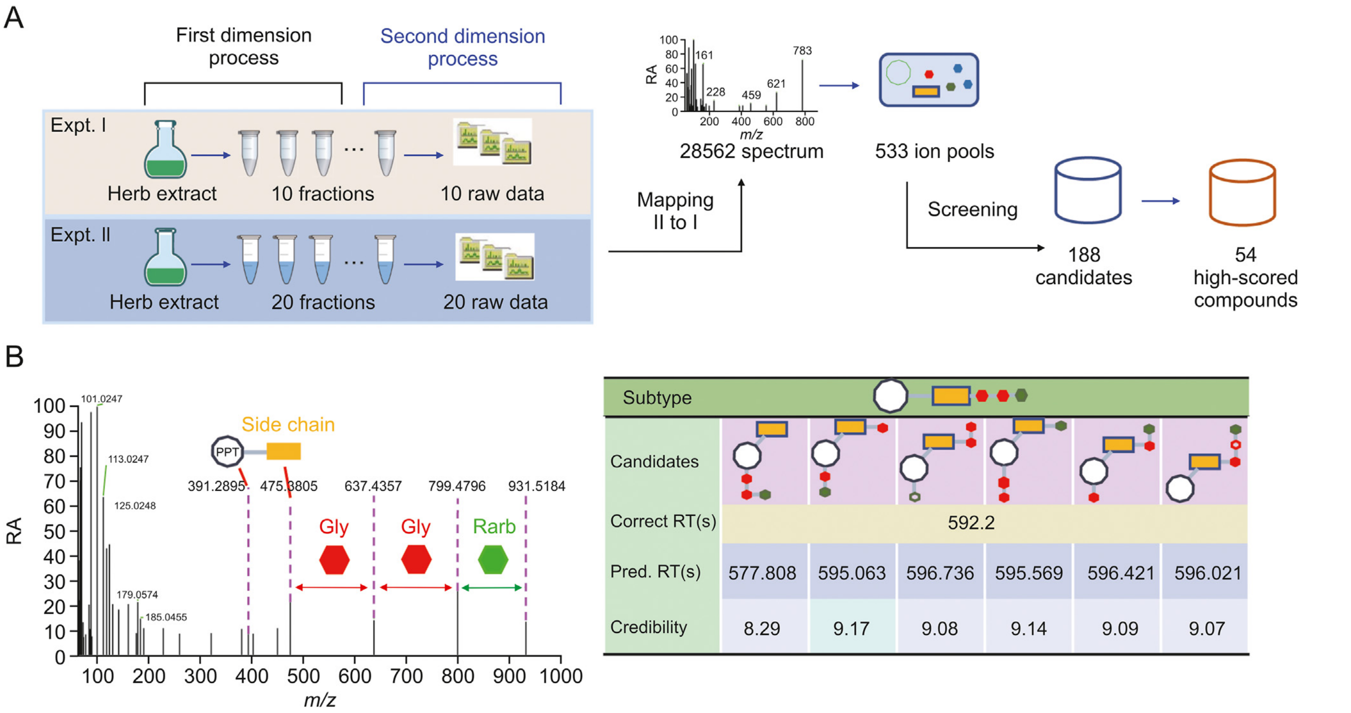

Fig. 2. (A) Acquisition of the normalized data set. This process includes the separation of total saponins of *Panax notoginseng* in 1D liquid phase, the collection of 2D liquid phase mass spectrometry data from experiments I and II, and the interactive retrieval and data normalization processing of experiments I and II. Experiment I captured a total of 1,126 saponin isoforms using neutral loss products and diagnostic ions. Experiment II will perform compound mass spectrometry MS^1 and MS^2 level matching from optimum separation experiment A data. Finally, a total of 533 saponins were screened in Experiment II. (B) Construction and screening of candidate compounds. Using the neutral loss strategy and product ion diagnosis, we identified the obtained fragment component (left), which specifically matched the sugar chain with the ion shedding sequence, and exhaustively listed 6 candidate structures of this isoform. Retention time evaluation was conducted using the Glyco-GNN model (right).

HRMS chromatographic conditions, see the LC-MS Methods and Data Processing section in the Supplementary data.) Chromatograms of target ion, decoy precursor ion, and corresponding fragment ion were extracted to determine potential elution peaks and characteristic ions with top intensity assigned. Being akin to unique peptides of parallel reaction monitoring (PRM) technique, 1–6 characteristic ions were selected from MS^2 fragment and taken as “identification sequence” of each elution peak (Tables S1 and S2, and Fig. S1). One or more molecule sequences that representing each elution peak were included and formed candidate glycoside molecule library for Glyco-GNN model training. In the study, 80,038 data from SMRT dataset of METLIN database and the 96 reported saponins are included into the Glyco-GNN training set (Figs. S2A–C) [2,3].

In prediction analysis, we would distinguish high-confidence target compounds from low-confidence decoys based on the elution peak confidence scores provided by Glyco-GNN (Fig. 2B, detailed construction process for candidate compounds can be found in the Supplementary data “Construction of candidate libraries” section.) Importantly, taken *Panax notoginseng* saponins, a complex glycoside mixture, as the test sample, we found no neural network model showed more significant benefits than Glyco-GNN (Figs. S2D–E, and S3). In addition, the testing data showed that Glyco-GNN displayed satisfactory prediction outcome on segregation of 20% input data for test sets and 80% input data for training set.

P.anax notoginseng is a natural medicinal plant rich in active ingredients. So far, only approximately 110 saponins in *P.anax notoginseng* have been reported, and this is a big deficiency for drug development. In this study, taken *P.anax notoginseng* saponins as an example, the data collection and processing procedure was expounded in detail. Comparing to traditional database searching strategy, Glycomol has obvious advantages in saponin identification (Fig. S4). By testing 24 *P.anax notoginseng* saponin standards, Glycomol's reliability was demonstrated with 95.8% hit rate achieved and it remained at 91.7% after FDR screening (Figs. S5 and 6) [4]. However, there are still limitations. Although we have completed most of the construction of the “small molecule signature signal” (SMSS, including diagnostic ions and neutral loss) library for metabolites of natural products such as *Erigeron breviscapus* (Tables S3–6, and Figs. S7–S9), it is still far from sufficient. For other small molecule families, such as cyanoside, coumarin, lignans, anthraquinones, flavonoids, indoles and many other glycoside molecules, we have not exhausted the “small molecule signature signal”. This requires a large number of researchers to enrich structural information. A complete set of public information collection databases on natural products under community management is therefore needed.

Since novel structure identification requires other tools, including nuclear magnetic resonance (NMR) and infrared spectrum (IR), for further verification, limitations associated with molecule discovery based on MS/MS alone remain to be solved in Glycomol and other MS identification software such as GNPS and Compound Discover. To sum up, Glycomol provides a powerful reference for the discovery of glycoside molecules, as well as a structure prediction strategy for other types of small molecules.

CRediT author statement

Daotong Zhao: Conceptualization, Methodology, and Software; **Chunguo Wang**: Data curation, Resources; **Hanyun Qu**,

Qinling Rao, and **Bingqing Shen**: Validation, and Formal analysis; **Yinan Jiang**, **Jiayu Gong**, **Yumiao Wang**, **Di Geng**, and **Rui Hong**: Investigation, Data curation; **Tao Lu**: Supervision; **Qing Ni**: Resources, Funding acquisition; **Xinqi Deng**: Writing - Original draft preparation, and Reviewing and Editing, Visualization, and Funding acquisition.

Declaration of competing interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant Nos.: 82374076, and 82174354), National Key R&D Program of China (Grant No.: 2023YFC3504100), Natural Science Foundation of Beijing (Grant No.: 7244493), NSFC Cultivating Project of Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences (Grant No.: ZXKT23003), Young Elite Cultivating Project of Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences (Grant No.: ZZ17-YQ-023), and Start-up Fund for New Scientific Researchers of China Academy of Chinese Medical Sciences (Grant No.: ZZ16-XRZ-071).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpha.2023.11.004.

References

[1] J. Cai, A. Jozwiak, L. Holoidovsky, et al., Glycosylation of N-hydroxy-pipecolic acid equilibrates between systemic acquired resistance response and plant growth, Mol. Plant 14 (2021) 440–455.

[2] X. Domingo-Almenara, C. Guijas, E. Billings, et al., The METLIN small molecule dataset for machine learning-based retention time prediction, Nat. Commun. 10 (2019), 5811.

[3] J. Xue, C. Guijas, H.P. Benton, et al., METLIN MS^2 molecular standards database: A broad chemical and biological resource, Nat. Methods 17 (2020) 953–954.

[4] J. Shen, L. Jia, L. Dang, et al., StrucGP: *de novo* structural sequencing of site-specific N-glycan on glycoproteins using a modularization strategy, Nat. Methods 18 (2021) 921–929.