analytical chemistry

# Goslin 2.0 Implements the Recent Lipid Shorthand Nomenclature for MS-Derived Lipid Structures

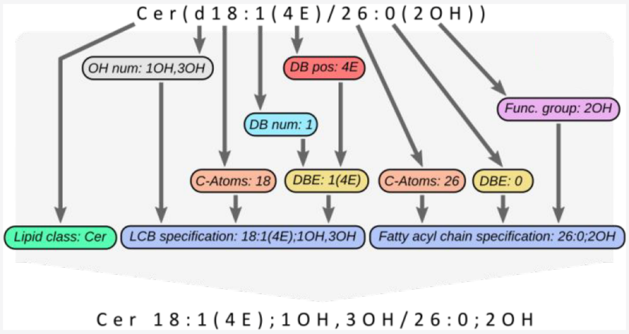
Dominik Kopczynski,[#] Nils Hoffmann,[#] Bing Peng, Gerhard Liebisch, Friedrich Spener, and Robert Ahrends*

Cite This: *Anal. Chem.* 2022, 94, 6097−6101

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Goslin is the first grammar-based computational library for the recognition/parsing and normalization of lipid names following the hierarchical lipid shorthand nomenclature. The new version Goslin 2.0 implements the latest nomenclature and adds an additional grammar to recognize systematic IUPAC-IUB fatty acyl names as stored, e.g., in the LIPID MAPS database and is perfectly suited to update lipid names in LIPID MAPS or HMDB databases to the latest nomenclature. Goslin 2.0 is available as a standalone web application with a REST API as well as C++, C#, Java, Python 3, and R libraries. Importantly, it can be easily included in lipidomics tools and scripts providing direct access to translation functions. All implementations are open source.



## INTRODUCTION

Lipids are essential organic molecules, since they are responsible for the compartmentalization of cells by membrane formation for the storage of energy and for serving as signaling molecules. No other molecule class comprises all these characteristics. Lipidomics is the research field based on large-scale lipid characterization by bioanalytical methods combined with data analysis by bioinformatics and the interpretation of these data in a broader biological context. Thereby, mass spectrometry (MS) drives the field due to its speed, sensitivity, and specificity.[1,2] On the one hand, this field is highly dependent on computational approaches, providing researchers with fast and accurate algorithms for conducting MS-based lipidomics experiments due to the complexity of lipid structures and lipidomes. On the other hand, hierarchical shorthand descriptions of lipids were introduced by Liebisch et al.[3,4] to concisely represent the vast structural variety of lipids and to thereby enable consistent reporting and communication. Recently, this shorthand nomenclature was refined and extended[3] to support more lipid classes and specific lipids with new features. The main updates were the annotation of ring double bond equivalents instead of double bonds and the number of oxygen atoms to permit a hierarchical reporting of oxygenated lipid species. The fatty acyl category was completely updated and now covers fatty acids and conjugates, fatty alcohols, wax monoesters, N-acyl amines, etc. In addition, shorthand notations for functional groups were added to the nomenclature, such as the COOH and OOH groups. Enclosed structures like acyl/alkyl branches or cycles can now be described as well. The hierarchical description levels representing the structural knowledge about a lipid were also updated as shown in Table 1. These changes will require updates of existing lipidomics tools and databases.

Besides this shorthand notation, several other systematic nomenclatures for lipids exist. In a linguistic context, these nomenclatures can be considered languages to describe how lipid names have to be structured. Since these nomenclatures are closely related and produce similar lipid names, we denote them as dialects. It remains a challenging task to correctly identify a lipid by its name among all the different dialects. The lipidomics field currently faces the challenges of integration and reanalysis of the existing results and data sets from multiple tools and data repositories that use different lipid shorthand dialects in order to document and reproduce lipid identification and quantification results. Making lipidomics research data machine-readable and accessible via community-accepted data formats, common shorthand names that encode the structural knowledge of lipids is one of the first steps required to address those challenges in accordance with the mission of the FAIR principles of interoperability and reusability.[5]

In 2020, we introduced Goslin[6] (the "Grammar Of Succinct LIpid Nomenclature"), which is a framework to translate lipid

**Received:** December 15, 2021
**Accepted:** March 21, 2022
**Published:** April 11, 2022

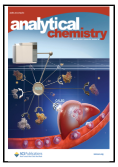


**Table 1. Hierarchical Presentation of a Shorthand Notation for Oxygenated Phosphatidylethanolamine PE 16:1(6*Z*)/16:0;5OH[*R*],8OH[*S*];3oxo[a]**

| level | lipid name |
|---|---|
| category | GP |
| class | PE |
| species level | PE 32:2;O3 |
| molecular species level | PE 16:1_16:1;O3 |
| *sn*-position level | PE 16:1/16:1;O3 |
| structure defined level | PE 16:1(6)/16:0;(OH)2;oxo |
| full structure level | PE 16:1(6Z)/16:0;5OH,8OH;3oxo |
| complete structure level | PE 16:1(6*Z*)/16:0;5OH[*R*],8OH[*S*];3oxo |

[a] From top to bottom, the structural information of the molecule increases. The species level provides information about the head group plus aggregated information on fatty acyl chains. The molecular species level provides aggregated information about constituent fatty acyl chains with unknown *sn*-positions. The *sn*-position level clarifies stereo-specific numbering. Until this level, the double bonds in the functional groups may be aggregated in the double bond equivalent. The structure defined level resolves functional groups in constituent fatty acyl chains. The full structure level adds position information, while the complete structure level adds all stereo-chemical information.

names from different dialects into a standardized name according to the lipid shorthand nomenclature. Its key module is a so-called parser that, based on a dialect-specific formal grammar, disaggregates the input string (i.e., lipid name), checks for correct syntax, and interprets these fragments to generate a standardized lipid name. Each grammar is a set of rules on what valid lipid names must look like according to the nomenclature. RefMet[7] is a database and web application that supports the normalization of names in metabolomics data with support for lipidomics shorthand nomenclature. LipidLynxX[8] provides a web application to normalize lipid names in order to interlink lipidomics data sets with external data sources, e.g., in the context of integration with biological pathways.

Goslin is already well accepted in the community and utilized in different tools such as LipidSuite, Lux Score 2.0, or lipidomics workflows.[9−11] To allow quick adaptation of the computational field and to keep the framework up to date and usable, we released the new version Goslin 2.0 with new features and full support of the latest shorthand nomenclature.

## ■ METHODS

Goslin can be used directly as a web application with an HTML user interface supplemented by a REST API for computational access at https://apps.lifs-tools.org/goslin/. In addition, it is available as a library for the programming languages C++, C#, Java, Python, and R. Programmers can easily include these libraries in their tools and use the provided lipid name translation functions directly. Here, we present the updated version Goslin 2.0 with several new features with an updated set of lipid classes (see Table S1).

Goslin fully supports the new shorthand nomenclature and is, to the best of our knowledge, the only tool capable of handling nested or recursive structures. For example, a typical lipid notation has the following structure: "[*lipid class*] [*chain specification*]/[···]", e.g., "TG 16:0/18:1(9Z)/18:1(9Z)". The *chain specification* itself usually has the structure "[*C-atoms number*]:[*double bond equivalent*];[*functional group 1*];[*functional group 2*];[···]", e.g., "16:0;3OH", where all functional groups are added sequentially at the end and are separated by a semicolon. Usual functional groups are, for instance, hydroxy groups (OH), carboxy groups (COOH), or sometimes whole O-acyl branches such as for TG 16:0;5O(FA 16:0)/18:1(9Z)/18:1(9Z). Here, [*functional group*] corresponds to the pattern [*chain specification*]. The *specification* of the O-acyl branch "O(FA 16:0)" is enclosed as a functional group at the carbon 5 position within the complete *specification* of the first fatty acyl chain "16:0;5O(FA 16:0)". An illustration of this nested

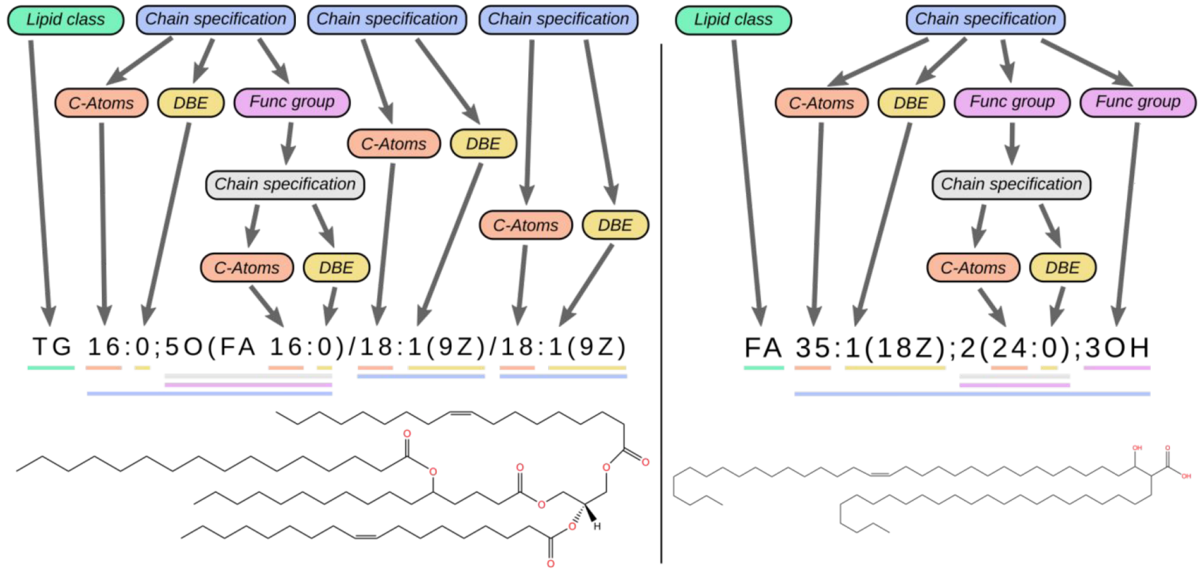


**Figure 1.** Exemplary illustrations of nested patterns: left, triacylglycerol with additional O-acyl linkage;[3] right, branched fatty acyl chain (LIPID MAPS-ID LMFA01160041) can be aligned schematically into the substitution blocks [*lipid class*] and [*chain specification*]. Here, the blocks [*chain specification*] (blue) are substituted into their successors. Some [*functional group*] blocks are again substituted into a [*chain specification*] block (gray) to describe the attached fatty acyl (left) or alkyl (right) branches in their lipids. A chain specification appears within another chain specification (gray within blue).

Table 2. Examples for Lipid Naming by IUPAC-IUB and Standardization by Shorthand Notation

| IUPAC-IUB name | LIPID MAPS | standardized name |
|---|---|---|
| 5-methyl-octadecanoic acid | LMFA01020216 | FA 18:0;5Me |
| 2-docosyl-3-hydroxy-28,29-epoxy-30-methyl-pentacontanoic acid | LMFA01160100 | FA 50:0;2(22:0);30Me;28Ep;3OH |
| 11*R*-hydroxy-9,15-dioxo-2,3,4,5-tetranor-prostan-1,20-dioic acid | LMFA03010032 | FA 15:0;[4-8cy5:0;7OH;5oxo];11oxo;15COOH |
| *N*-((±)-8,9-dihydroxy-5*Z*,11*Z*,14*Z*-eicosatrienoyl)-ethanolamine | LMFA08040030 | NAE 20:3(5Z,11Z,14Z);8OH,9OH |

pattern is provided in Figure 1. We use context-free grammars and parsers, which can recognize these kinds of patterns. According to the Chomsky hierarchy, standard approaches utilizing regular expressions do not have the possibility of recognizing general nested patterns as formally correct[12] and thus are insufficient for our purposes. Incorrect recognition of lipid names (and thus no or incorrect error handling) results in incorrect annotations.

Another new feature is the additional grammar that parses systematic fatty acyl descriptions following the IUPAC-IUB nomenclature,[13] such as that listed in Table 2. This feature enables databases with older entries to convert their fatty acyl IUPAC-IUB names into the newest lipid nomenclature. The third major new feature of Goslin 2.0 is the calculation of the chemical sum formula of lipids and accurate masses (neutral or adduct ions). As in the preceding version, the user can choose at which level the lipid annotation should be generated in accordance with the hierarchy shown in Table 1. All information that can be extracted from a lipid shorthand notation can be obtained from Goslin 2.0 either as a table from the web service at https://apps.lifs-tools.org/goslin/ or as the associated data structures from the programming libraries. Supported lipid dialects are the updated shorthand notations by Liebisch et al.,[3] the original shorthand notation by Liebisch,[4] i.e., LIPID MAPS dialect,[14] Goslin dialect, SwissLipids[15] dialect, HMDB[16] dialect, and the IUPAC-IUB nomenclature dialect for fatty acyl chains.

## RESULTS

We evaluated Goslin 2.0 by taking all fatty acyl chain descriptions from LIPID MAPS along with their chemical sum formula and converting them via Goslin. The LIPID MAPS database contains 10 011 fatty acyls (July 2021). From this set, 8005 lipid names can be specified with the new nomenclature. The remaining entries contain structures such as triple bonds, histidine, aspartate, azaniumyl, etc. that cannot be described by the nomenclature yet. The conversion and computation of the chemical sum formula of all lipid names took less than 3 s (when applying the C++ library) on our standard computing platform (Lenovo Thinkpad X1 Carbon, Intel i7 1.8 GHz octa-core laptop, 16GB main memory). All computed sum formulas did perfectly match with the sum formulas from the database. Additionally, we randomly picked lipid names from the translated list and manually checked their correctness according to the nomenclature specifications. All lipid names were correct. We updated our previous unit tests (>100 000 single tests) to the new nomenclature. All tests passed without problems on each system (C++/C#/Java/Python/R). To check the overall performance of Goslin 2.0, we took the databases from LIPID MAPS, SwissLipids, and HMDB and selected fatty acyls (FAs), glycerolipids (GLs), glycerophospholipids (GPs), sphingolipids (SPs), and sterols (STs) for conversion. For SwissLipids and HMDB, almost the complete selection could be converted into the new nomenclature (Table 3). For LIPID MAPS, about 20% of the lipid names was not recognized, since they are not lipids of confirmed biological origin, contain only trivial names, or simply could not yet be described by the shorthand notation. However, for the SwissLipids and HMDB databases, more than 97% of their lipid names can be converted (Table 3).

Table 3. Number of Parsed Lipids per Database: All Database Snapshots Were Acquired in July 2021

| | LIPID MAPS | SwissLipids | HMDB |
|---|---|---|---|
| total no. of lipids | 45 552 | 777 956 | 90 688 |
| total no. of FA, GL, GP, SP, and ST | 35 556 | 777 956 | 87 775 |
| no. of converted FA, GL, GP, SP, and ST by Goslin 2.0 | 29 098 (81.8%) | 771 287 (99.1%) | 85 179 (97.0%) |

We further compared Goslin 2.0 to Goslin 1.1.2, RefMet, and LipidLynxX 0.9.24 on a selection of data sets sourced from public data repositories and the literature to assess their speed and percentage of converted lipid names in realistic data conversion and normalization scenarios (see Table S2). On average, Goslin 2.0 was slightly faster than Goslin 1.1.2 (avg. of 2.34 s vs 2.45 s), while being on par with RefMet (avg. of 2.18 s). LipidLynxX was the slowest one (avg. of 79.05 s). Concerning the rate of converted lipid names, Goslin 2.0 outperformed the other tools with an average percentage of 84.11. However, for Goslin, this number only contains lipids that are valid following at least one of the supported grammars. Otherwise, Goslin will not attempt to convert them. Higher percentages for RefMet and LipidLynxX are attributable to a number of questionable conversions (see Table S3). We provide an overview of the specific features supported by Goslin 2.0, Refmet, and LipidLynxX in Table S4.

## DISCUSSION AND CONCLUSION

For standardization of lipid annotation, Goslin 2.0 supports the newest shorthand nomenclature and provides new features such as parsing of systematic fatty acyl chain names or computation of chemical sum formulas, molecular, and adduct masses. To the best of our knowledge, Goslin 2.0 is the first available tool that recognizes nested and recursive patterns problem-free within lipid names. In contrast to error-prone systems using regular expressions, it is very robust against incorrect lipid name descriptions, a crucial feature in automated analysis of very large data sets. Lipid names that do not follow the nomenclature are reported to the user instead of being passed with incorrect annotations. Goslin 2.0 is designed to be a real-time module within pipelines for lipidomics analyses. A key feature is its performance as all implementations can convert a regular-sized list of lipid names into the new nomenclature (including the computation of the sum formulas) in less than a second on current standard computers. It is highly suited to streamline computational lipidomics workflows for true high-throughput analytical experiments. The schematic diagram of the Goslin class model can be found in Figure S1. The Goslin implementations are capable of parsing the IUPAC-IUB systematic lipid names of free fatty acyl chains. Lipid databases such as LIPID MAPS, SwissLipids, or HMDB can automatically update their lipid name entries using our libraries.

Tools like LipidCreator[17] already profit from the updated version of Goslin due to its straightforward implementation into other standalone tools. It can serve as a cornerstone of standardization in the field of lipidomics[18] but is still open for updates. All Goslin 2.0 implementations are freely available at https://github.com/lifs-tools/goslin under the terms of liberal open source licenses.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.1c05430.

> List of supported lipid classes; comparison between different converters on lipid datasets; and class diagram of the object model used by all Goslin implementations (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Author

**Robert Ahrends** − *Institute of Analytical Chemistry, University of Vienna, 1090 Vienna, Austria;* orcid.org/0000-0003-0232-3375; Email: robert.ahrends@univie.ac.at

### Authors

**Dominik Kopczynski** − *Institute of Analytical Chemistry, University of Vienna, 1090 Vienna, Austria;* orcid.org/0000-0001-5885-4568

**Nils Hoffmann** − *Center for Biotechnology (CeBiTec), Bielefeld University, 33594 Bielefeld, Germany;* Present Address: Forschungszentrum Jülich, Institute of Bio- and Geosciences, Computational Metagenomics (IBG-5), 33594 Bielefeld, Germany; orcid.org/0000-0002-6540-6875

**Bing Peng** − *Division of Rheumatology, Department of Medicine, Solna, Karolinska Institutet and Karolinska University Hospital, 17176 Stockholm, Sweden;* orcid.org/0000-0001-5006-7041

**Gerhard Liebisch** − *Institute of Clinical Chemistry and Laboratory Medicine, Regensburg University Hospital, 93053 Regensburg, Germany;* orcid.org/0000-0003-4886-0811

**Friedrich Spener** − *Department of Molecular Biosciences, University of Graz, 8010 Graz, Austria; Division of Molecular Biology and Biochemistry, Gottfried Schatz Research Center, Medical University of Graz, 8036 Graz, Austria*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.1c05430

### Author Contributions

#D.K. and N.H. contributed equally to this work. D.K. and N.H. designed and implemented the libraries, grammars, and the web-service. B.P. controlled and supervised the chemical aspect of the libraries. G.L. and F.S. coordinated and verified the updates on the nomenclature. R.A. discussed, supervised, and coordinated the project. All authors wrote and revised the manuscript.

### Funding

Open Access is funded by the Austrian Science Fund (FWF).

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This study was supported by the grants from the Deutsche Forschungsgemeinschaft and FWF der Wissenschaftsfonds (P33298-B). The authors received further support from University of Vienna by seed funding and funding derived from the DosChem doctoral school program faculty of Chemistry.

## ■ REFERENCES

(1) Blanksby, S. J.; Mitchell, T. W. *Annu. Rev. Anal Chem. (Palo Alto Calif)* **2010**, *3*, 433−65.

(2) Wenk, M. R. *Cell* **2010**, *143* (6), 888−95.

(3) Liebisch, G.; Fahy, E.; Aoki, J.; Dennis, E. A.; Durand, T.; Ejsing, C. S.; Fedorova, M.; Feussner, I.; Griffiths, W. J.; Kofeler, H.; Merrill, A. H., Jr.; Murphy, R. C.; O'Donnell, V. B.; Oskolkova, O.; Subramaniam, S.; Wakelam, M. J. O.; Spener, F. *J. Lipid Res.* **2020**, *61* (12), 1539−1555.

(4) Liebisch, G.; Vizcaino, J. A.; Kofeler, H.; Trotzmuller, M.; Griffiths, W. J.; Schmitz, G.; Spener, F.; Wakelam, M. J. O. *J. Lipid Res.* **2013**, *54* (6), 1523−1530.

(5) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J.G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A.C; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. *Sci. Data* **2016**, *3*, 160018.

(6) Kopczynski, D.; Hoffmann, N.; Peng, B.; Ahrends, R. *Anal. Chem.* **2020**, *92* (16), 10957−10960.

(7) Fahy, E.; Subramaniam, S. *Nat. Methods* **2020**, *17* (12), 1173−1174.

(8) Ni, X.; Fedorova, M. LipidLynxX: a data transfer hub to support integration of large scale lipidomics datasets. *Biorxiv* 2020; https://www.biorxiv.org/content/10.1101/2020.04.09.033894v2.

(9) Lam, S. M.; Wang, Z.; Li, B.; Shui, G. *Anal. Chim. Acta* **2021**, *1147*, 199−210.

(10) Marella, C.; Torda, A. E.; Schwudke, D. *PLoS Comput. Biol.* **2015**, *11* (9), e1004511.

(11) Mohamed, A.; Hill, M. M. *Nucleic Acids Res.* **2021**, *49* (W1), W346−W351.

(12) Chomsky, N. *IRE Transactions on Information Theory* **1956**, *2* (3), 113−124.

(13) McNaught, A. D.; Wilkinson, A. *IUPAC: Compendium of Chemical Terminology*; Version 3.0.1; Blackwell Scientific Publications, 2018.

(14) Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E. A.; Glass, C. K.; Merrill, A. H., Jr.; Murphy, R. C.; Raetz, C. R.; Russell, D. W.; Subramaniam, S. *Nucleic Acids Res.* **2007**, *35* (Database issue), D527−32.

(15) Aimo, L.; Liechti, R.; Hyka-Nouspikel, N.; Niknejad, A.; Gleizes, A.; Gotz, L.; Kuznetsov, D.; David, F. P.; van der Goot, F. G.; Riezman, H.; Bougueleret, L.; Xenarios, I.; Bridge, A. *Bioinformatics* **2015**, *31* (17), 2860−6.

(16) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M. A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; Macinnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie,

T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L. *Nucleic Acids Res.* **2007**, *35* (Database issue), D521−D526.

(17) Peng, B.; Kopczynski, D.; Pratt, B. S.; Ejsing, C. S.; Burla, B.; Hermansson, M.; Benke, P. I.; Tan, S. H.; Chan, M. Y.; Torta, F.; Schwudke, D.; Meckelmann, S. W.; Coman, C.; Schmitz, O. J.; MacLean, B.; Manke, M. C.; Borst, O.; Wenk, M. R.; Hoffmann, N.; Ahrends, R. *Nat. Commun.* **2020**, *11* (1), 2057.

(18) Lipidomics Standards Initiative Consortium. *Nat. Metab* **2019**, *1* (8), 745−747.