

小分子質量分析結果の報告のための HUPO-PSI 標準フォーマットの進化

西田孝三¹、Philippine Louail²、Ozgur Yurekten³、Steffen Neumann⁴、
高橋みき子⁵、早川英介⁶、松沢佑紀⁷、津川裕司⁷、Nils Hoffmann⁸

1. 理研BDR、2. Eurac Research、3. EMBL-EBI、4. IPB Halle、5. 理研CSRS、6. 九州工業大学、7. 東京農工大学、8. Research Center Jülich

背景

メタボローム(細胞/生体内 small molecule の総体)定量の実験技術の中核を担うのは質量分析(MS)
しかしMSデータ解析結果の記録・共有するためのデータフォーマットの標準化は進んでいない。そのため質量分析定量結果の利活用・データ統合は未発達

mzTab-Mとは

- HUPO-PSI (Human Proteome Organization Proteomics Standards Initiative) が策定した
質量分析メタボロミクスにおける定量結果の共有のためのデータ標準規格

- A) Metadata [MTD]
(実験条件、サンプル情報等のテーブル)
- B) Small Molecule [SML]
(「同定された分子」レベルでの特徴量テーブル)
- C) Small Molecule Feature [SMF]
(「MSのPeak Feature」レベルでの特徴量テーブル)
- D) Small Molecule Evidence [SME]
(「同定に用いた」情報のテーブル)

の4種のテーブルを1つのテキストファイル (.mztab)中に記述したデータフォーマット

- SML レベルだけでなく SMF レベルのデータも扱え、統制語彙を用いたメタデータの補完も行えるため再解析者の多様な要望に応じることが可能

- 機械可読性向上のための「JSON変換/逆変換」もリファレンス実装 jmzTab-M により可能
- そのJSONスキーマを元にした OpenAPI も定義
- そのOpenAPIを用いたR/Pythonパッケージも存在し、容易に再解析ワークフローを実装可能

- MetaboLights とのデータ互換性に関しては、「mzTab-M -> ISA-TAB」方向の変換の自動化が実現済み (mztabm2mtbls Pythonパッケージにて)

mzTab-Mのこれまでとこれから

- 2019年: mzTab-M 2.0 と、その論文の発表

- 2024年:** MetaboLights、Metabolomics Workbench等の国際リポジトリを統合的に活用し、データ標準やAPIを通じてメタボロームデータを利活用するためのハブ的プロジェクト／プラットフォーム Metabolomics Hub が発足（Chan Zuckerberg Initiative [CZI] による資金提供が開始）

mzTab-M に基づくデータ標準化、利活用推進の再起動のきっかけに

- 2025年: mzTab-M 2.1 に向けたワーキンググループの発足、開発の開始 (mzTab 1.0はプロテオミクス用、mzTab 2.0でメタボロミクス拡張[==mzTab-M])

1	MTD	mzTab-version	2.2.0-M
2	MTD	mzTab-ID	JetBike Test
3	MTD	software[1]	[MS.MS:1002879,Progenesis QI.2.6.505.48857]
4	MTD	ms_run[1]-location	file:///D:/Data%20sets/Metabolomics/MTBLS263/3injections_inj1_POS.ms2ML
5	MTD	ms_run[1]-scan_polarity[1]	[MS.MS:1000130,positive_scan]
6	MTD	ms_run[2]-location	file:///D:/Data%20sets/Metabolomics/MTBLS263/3injections_inj2_POS.ms2ML
7	MTD	ms_run[2]-scan_polarity[1]	[MS.MS:1000130,positive_scan]
8	MTD	ms_run[3]-location	file:///D:/Data%20sets/Metabolomics/MTBLS263/3injections_inj3_POS.ms2ML
9	MTD	ms_run[3]-scan_polarity[1]	[MS.MS:1000130,positive_scan]
10	MTD	ms_run[4]-location	file:///D:/Data%20sets/Metabolomics/MTBLS263/3samples_sample1_POS.ms2ML
11	MTD	ms_run[4]-scan_polarity[1]	[MS.MS:1000130,positive_scan]
12	MTD	ms_run[5]-location	file:///D:/Data%20sets/Metabolomics/MTBLS263/3samples_sample2_POS.ms2ML
13	MTD	ms_run[5]-scan_polarity[1]	[MS.MS:1000130,positive_scan]
14	MTD	ms_run[6]-location	file:///D:/Data%20sets/Metabolomics/MTBLS263/3samples_sample3_POS.ms2ML
15	MTD	ms_run[6]-scan_polarity[1]	[MS.MS:1000130,positive_scan]

Various columns are available for describing what is known about the identity of the molecule. If ambiguity remains, then several values may be provided (separated by bars) in these columns (not shown).

Experimental data from Progenesis Q1

A) MTD table

3 X 3 replicate, LC-MS/MS in positive mode; global profiling of metabolites

B) SML table

Quantitative data about individual runs (Assays) and averaged over replicates (Study Variables) reported as the “final” values per molecule reported.

[illegible]

Adduct grouping

Measurement data including retention time, experimental mass to charge, assumed adduct status.

Quantitative data about molecules quantified by mass spectrometry and software, prior to adduct grouping.

C) SMF table

id	SMF	SME_ID	REFS	SME_ID_REF_ambiguity_co	adduct_ion	isotopomer	exp_mass_to_charge	charge	retention_time_in_secs	retention_time_start	retention_time_end	abundance_1	abundance_2	abundance_3	abundance_4	abundance_5	abundance_6
83	SMF	6	1		[M+H] ⁺	null	114.0654	1	413.81	393.55	447.87	59579141	63545698	61393681	59131129	63693769	66308066
84	SMF	7	1		[M+H] ⁺	null	116.0699	1	607.75	587.39	648.11	59134552	57461952	60069536	56567781	69988852	60837200
85	SMF	13	4	null	null	null	132.1014	1	483.77	455.72	541.07	65543883	65425790	67694648	59237651	56077561	54207405
86	SMF	17	13	null	null	null	147.0758	1	735.86	719.74	768.94	27880437	23839759	27913260	39816664	4011277	43411662
87	SMF	18	6	null	null	null	118.0857	1	984.32	558.3	621.66	27175406	23522323	24850183	18844152	22892660	21607739
88	SMF	49	9	null	null	null	120.0649	1	704.42	691.29	719.74	8652081	8068845	8480355	7567116	92999261	8805714
89	SMF	50	18	11	[M+H] ⁺	null	124.0132	1	1282.28	1240.17	1327.7	18367576	18367576	18367576	18367576	18367576	18367576
90	SMF	102	11	null	[M+H] ⁺	null	103.0674	1	117.05	70.2	727.4	8552284	7430405	8442646	10530095	10328389	12041933
91	SMF	117	5	null	null	null	150.0578	1	513.76	496.87	555.22	8419806	795485	8318227	8007518	7319791	7364748
92	SMF	153	8	null	null	null	169.0349	1	653.98	630.69	667.64	6594547	6679574	6147242	6238970	6287508	7032912
93	SMF	155	12	null	[M+Na] ⁺	null	173.0781	1	768.94	757.1	121074	1417953	121074	1424344	0	0	0
94	SMF	523	19	null	[M+Na] ⁺	null	169.0943	1	1279.2	1261.83	1314.56	1141208	1127351	111897	8128685	1002102	1019391
95	SMF	566	17	null	null	null	444.0304	1	919.02	910.32	930.76	776953.5	826544	80955239	555138.9	726945.9	682373.5
96	SMF	774	10	null	null	null	126.0217	1	714.99	704.42	727.4	5928287	5417155	5820193	6318313	7554883	6698600
97	SMF	780	12	null	null	null	133.0602	1	733.64	722.81	751.1	2465240	2478660	2478660	1261328	1504851	1670066
98	SMF	782	3	null	[M+Na] ⁺	null	137.0452	1	440.1	429.32	455.72	1164487	1376478	1513499	1398304	1635382	2129397
99	SMF	937	2	null	[M+Na] ⁺	null	136.0475	1	413.81	402.83	419.78	236014	227593	236957.5	0	180978.7	12643.06

Evidence for identification

Data sourced from external
databases about the molecular
identity.

Summary info about the data that has been matched.

Identification scores or statistics from the software or approach.

		SME		database_id	chemical_formula	smiles	inchi	name	url	derivat	adduct	exp_mass	to_charge	charge	theoretical	mass	spectra	identificati	id_conf	id_conf	id_conf	opt_global_r	opt_global_r	
		ID	evidence_input_id	id											ref			ms_level	id_conf	id_conf	id_conf	opt_global_r	opt_global_r	
																			measure	measure	measure	ententi	ententi	
																			[1]	[2]	[3]	nds	nds	
104	SEH	2		database_16737	C4H7NO3			Creatinine	null	[M+H]	114.0624	1	114.0662	ms_run	[Progenesi	[MS.MS-10:0			56.442	0	99.606	1	413.81	414
106	SME	2		database_16737	C4H7NO3			Creatinine	null	[M+Na]	136.0475	1	136.0481	ms_run	[Progenesi	[MS.MS-10:0			56.442	0	99.606	1	413.81	414
107	SME	3	440.10_137.0452m/z	database_17368	C5H4NO4			Hypoxanthine	null	[M+H]	137.0452	1	137.0458	ms_run	[Progenesi	[MS.MS-10:0			44.025	0	93.137	1	440.1	444
108	SME	4	483.77_132.1014m/z	database_25017	C6H13NO2			Leucine	null	[M+H]	132.1014	1	132.1019	ms_run	[Progenesi	[MS.MS-10:0			48.726	0	98.838	1	483.77	486
109	SME	5	513.76_150.0578m/z	database_16811	C5H11NO2S			Methionine	null	[M+H]	150.0578	1	150.0583	ms_run	[Progenesi	[MS.MS-10:0			47.605	0	95.753	1	513.76	516
110	SME	6	584.32_118.0857m/z	database_27266	C5H9NO2			Valine	null	[M+H]	118.0857	1	118.0863	ms_run	[Progenesi	[MS.MS-10:0			48.711	0	98.949	1	584.32	582
111	SME	7	607.75_116.0699m/z	database_26271	C5H11NO2			Proline	null	[M+H]	116.0699	1	116.0706	ms_run	[Progenesi	[MS.MS-10:0			45.887	0	97.758	1	607.75	612
112	SME	8	653.98_169.0349m/z	database_17775	C5H4NO4			Urate	null	[M+H]	169.0349	1	169.0356	ms_run	[Progenesi	[MS.MS-10:0			44.524	0	97.087	1	653.98	648
113	SME	9	704.42_120.0649m/z	database_26986	C4H9NO3			Threonine	null	[M+H]	120.0649	1	120.0655	ms_run	[Progenesi	[MS.MS-10:0			49.309	0	98.25	1	704.42	702
114	SME	10	714.99_126.0217m/z	database_15891	C2H7NO3S			Taurine	null	[M+H]	126.0217	1	126.0219	ms_run	[Progenesi	[MS.MS-10:0			52.852	0	91.869	1	714.99	714
115	SME	11	717.05_303.0674m/z	database_39033	C8H13NO2S2			PIPES (Asp-Hg)	null	[M+H]	303.0674	1	303.0679	ms_run	[Progenesi	[MS.MS-10:0			47.621	0	61.445	1	717.05	714
116	SME	12	733.64_133.0602m/z	database_27553	C4H8NO2			Asparagine	null	[M+H]	133.0602	1	133.0608	ms_run	[Progenesi	[MS.MS-10:0			50.797	0	94.345	1	733.64	732
117	SME	13	735.86_147.0758m/z	database_28300	C5H12NO3			Glutamine	null	[M+H]	147.0758	1	147.0764	ms_run	[Progenesi	[MS.MS-10:0			47.44	0	98.279	1	735.86	732
118	SME	14	754.37_173.0206m/z	database_17270	C3H9O6P2			Glycerol-2-pho	null	[M+H]	173.0206	1	173.0211	ms_run	[Progenesi	[MS.MS-10:0			46.796	0	95.369	1	754.37	750
119	SME	15	768.94_325.0424m/z	database_16695	C9H12NO3OP			UMP	null	[M+H]	325.0424	1	325.0431	ms_run	[Progenesi	[MS.MS-10:0			53.396	0	92.748	1	768.94	768
120	SME	16	865.23_203.0520m/z	database_17268	C6H21NO6			myo-Inositol	null	[M+Na]	203.052	1	203.0505	MS	MDT	id_confidence_measure[1]	[MS.MS-1002889,Progenesi,MetaScope score,]							
																59	MDT	id_confidence_measure[2]	[MS.MS-1002890,fragmentation score,]					
																60	MDT	id_confidence_measure[3]	[MS.MS-1002891,Isotopic fit score,]					

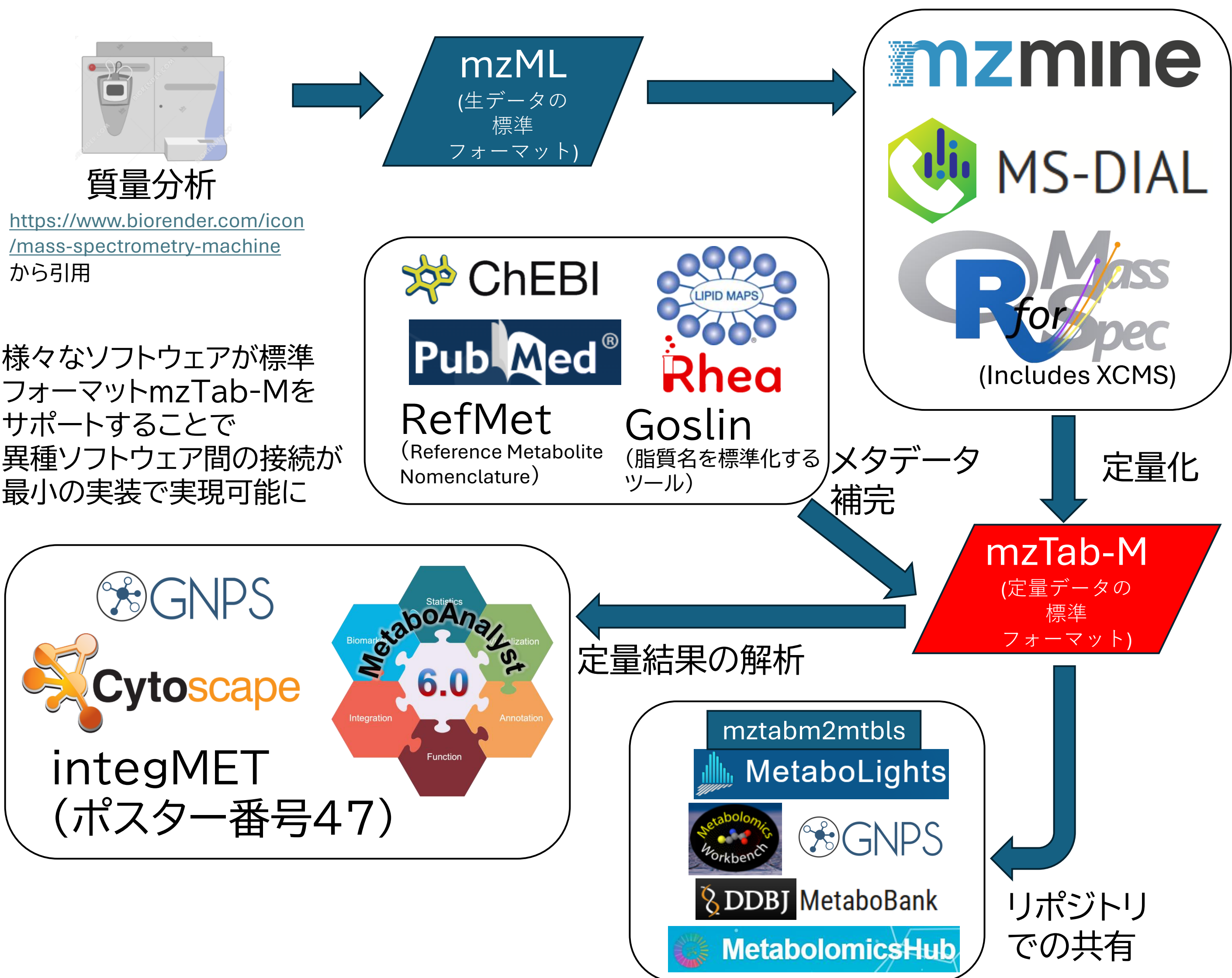
https://huipo-psi.github.io/mzTab/2_0-metabolomics-

Score type or units

https://hupo-psi.github.io/mzTab/2_0-metabolomics-release/mzTab_format_specification_2_0-M_release.html から引用

Score type or units
encoded in the meta
data

mzTab-M を取り巻くメタボロミクス研究環境



mzTab-M 2.1 のロードマップ

- **2.1 での戦略的強化項目**
 1. Universal Spectrum Identifier (USI) スペクトル列のSMEへの追加
(同定の証拠を外部のスペクトルライブラリで検証可能に)
 2. ISA-TAB との相互運用性・変換レイヤーの協調の強化
(mzTab-MとISA-TABの間の情報欠落を防ぐための
両標準化グループ間でのメタデータ粒度の調整等)
 3. リピドミクス定量データの詳細記述標準化の推進
(脂質特有の複雑な同定表現の標準化等)

- 2025年11月 BioHackathon Europe
 - 現在利用可能な mzTab-M プロデューサーとコンシューマーの相互運用性の網羅的テスト
 - 現存する mzTab-Mのバリデーションエラーや不足情報のアップストリームへの報告と修正

- 2025年12月 4th BioHackathon Germany
 1. mzTab-M へのメタデータ補完のためのNLPモデルのトレーニング
(出版物から構造化データセットを抽出するNLP/LLMパイプラインの実装。
モデルトレーニング用構造化データセットの作成)