ISSN 2186-7437

NII Shonan Meeting Report

No. 250

Advances and opportunities in artificial intelligence for computational metabolomics

Tomáš Pluskal
Wout Bittremieux
Ichigaku Takigawa

June 29–July 2, 2026

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Advances and opportunities in artificial intelligence for computational metabolomics

Organizers:
Tomáš Pluskal (IOCB Prague, Czech Republic)
Wout Bittremieux (University of Antwerp, Belgium)
Ichigaku Takigawa (The University of Tokyo, Japan)

June 29–July 2, 2026

Long before exchanging GitHub repositories and manuscripts, the participants first left something simpler: a small piece of themselves on the Yosegaki.

Background and Introduction

Metabolomics studies the small molecules that participate in, regulate, and reflect biological processes. These metabolites are central to cellular growth, development, communication, stress responses, and disease mechanisms, and they are also directly relevant to drug discovery because many pharmaceuticals are small molecules. Mass spectrometry is the dominant experimental technology for measuring metabolites at scale, but untargeted metabolomics still faces a severe computational bottleneck: only a small fraction of spectra produced in typical experiments can be confidently annotated.

This gap has created the so-called dark metabolome: a large body of spectral evidence that points to biologically and chemically meaningful molecules, but cannot yet be reliably connected to molecular structures. Classical computational approaches, spectral-library search, and rule-based fragmentation tools have produced important progress, but they do not fully address the diversity, noise, and scale of modern metabolomics data. At the same time, AI and machine learning have transformed adjacent fields such as genomics, proteomics, materials science, and molecular design, suggesting that computational metabolomics may now be ready for a similar shift.

The meeting was designed to catalyze collaboration between computational scientists, mass spectrometrists, chemists, biochemists, and machine-learning researchers. It built on earlier community efforts, including Shonan and Dagstuhl meetings that helped establish resources such as the CompMS community, DreaMS, and MassSpecGym. The central premise was that progress requires both better AI methods and sharper metabolomics-specific problem definitions: robust benchmarks, realistic data splits, uncertainty-aware evaluation, better spectral libraries, and models that respect the physical and chemical structure of mass spectrometry data.

Overview of the Meeting

NII Shonan Meeting No. 250 took place at the Shonan Village Center from June 29 to July 2, 2026. The official meeting description listed potential discussion topics including foundation models for metabolomics, generative AI for molecular structures, uncertainty in molecular identification, data management and standardization, generalization beyond annotated molecular space, integration of quantum chemistry and machine learning, molecular networking, community development, and Kaggle-style competitions. The seminar was organized around collaborative sessions rather than a sequence of fully documented formal talks. Participants proposed topics in a shared document, selected sessions, and recorded notes in separate files. This report summarizes the meeting session by session.

The participants of the meeting are shown in Figure 1. They came from Europe, Asia, and North America, and their backgrounds covered mass spectrometry, chemistry, machine learning, and software development. The full list of names and affiliations is given in the section *List of Participants*.

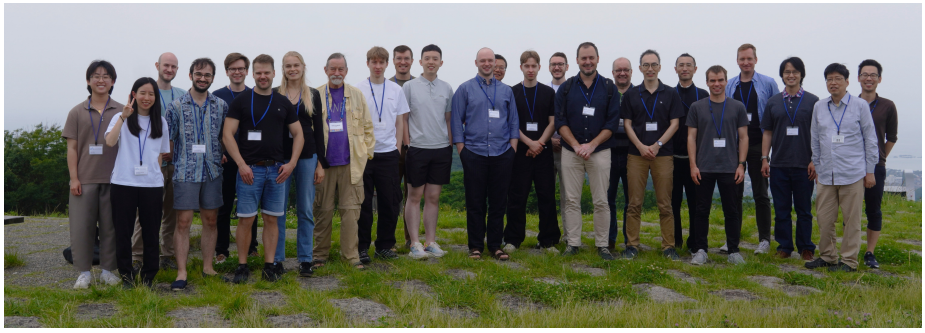

Figure 1: Participants of NII Shonan Meeting No. 250 at the Shonan Village Center.

Contents

1 Benchmarking Structure Elucidation Methods 5

2 MS Libraries and Data Quality 7

3 Generative Models for De Novo Chemical Structure Characterization 8

4 Machine-Learning Interatomic Potentials for Tandem MS 9

5 Self-Supervised Pre-Training for Mass Spectrometry Spectrum Encoders 10

6 Weird Molecules and Non-Canonical Chemical Spaces 11

7 Chemical Space 12

8 Metabolomics Atlas and Modeling After Structure Elucidation 13

9 Kaggle/CASMI Competitions for Metabolomics 14

10 Evaluation Metrics for *De Novo* Structure Prediction 15

11 MassSpecGym v2 16

1 Benchmarking Structure Elucidation Methods

Date: June 29, 2026.

MassSpecGym motivation and community impact

The benchmark discussion began from the role of MassSpecGym as a shared resource for MS/MS-to-structure machine learning. Participants noted that MassSpecGym was created because recent machine-learning papers in small-molecule mass spectrometry had used inconsistent datasets, preprocessing pipelines, splits, metrics, and evaluation protocols. This made method comparison unreliable and made it difficult for machine-learning researchers to enter the field without first reconstructing domain-specific evaluation infrastructure.

The group viewed MassSpecGym as successful in attracting new methods and stimulating community activity. At the same time, the use of MassSpecGym by a growing number of papers has revealed deeper problems in benchmark design and benchmark use. An audit of 26 papers using MassSpecGym found that a large fraction had substantial or minor issues related to leakage, implementation, evaluation, or benchmark interpretation. The conclusion was not that benchmarking had failed, but that the community now needs stricter, clearer, and more realistic evaluation standards.

Spectra-to-molecule abstraction limits

A recurring question was how far mass spectrometry can be abstracted away for machine learning. Several participants pushed back against the idea that ML researchers can safely treat spectra as generic inputs and molecules as generic labels. MS/MS spectra were described as semi-unique, lossy fingerprints rather than deterministic one-to-one encodings of structure. Spectra depend on instrument type, collision energy, tuning, ionization, matrix effects, co-elution, acquisition conditions, and sample complexity.

The group discussed whether spectra from different vendors and instruments should be comparable under a shared benchmark. QTOF and Orbitrap data may be broadly comparable under appropriate tuning, while ion trap data and low-mass cutoffs can create larger differences. No final consensus was recorded, but the notes emphasize that benchmark tasks should not hide domain assumptions that strongly affect generalization.

Data leakage, splitting, and overfitting

The group agreed that leakage is not binary. Exact test-structure leakage is the clearest failure mode, but near-neighbor leakage is also important: a model can benefit from training on molecules that are highly similar to test molecules even if the exact test structure is absent. Similarity thresholds such as Tanimoto 0.7–0.8 were discussed, but the notes do not record a universal cutoff.

A major proposal for MassSpecGym 2.0 is to define explicit allowed pools for training and pretraining, so that every method is evaluated with controlled access to molecules and spectra. Participants also discussed releasing pretrained molecular generator or foundation-model weights trained only on approved data. The group stressed that every pipeline component must respect benchmark

boundaries, including molecular decoders, formula predictors, synthetic-data generators, spectrum simulators, and test-time training procedures.

SpectraVerse, S4, and shortcut learning

The benchmark sessions identified several shortcut-learning artifacts. One involved inconsistent SMILES canonicalization between ground-truth molecules and candidate decoys. Some models could exploit representation artifacts rather than learn the spectrum-structure relationship, and performance dropped sharply when the artifact was corrected. Another shortcut came from PubChem ranking: selecting the top PubChem hit for a formula can recover the correct molecule surprisingly often, reflecting the fact that library molecules tend to be more studied, commercially available, biologically relevant, or database-prominent than random candidates.

Notes from the SpectraVerse/S4 discussion indicate that some retrieval benchmarks can be scored highly even without meaningful use of spectra. In particular, structure-only or library-likeness baselines can perform competitively under some settings. The group therefore emphasized the need to report strong baselines, including modified-cosine nearest-neighbor methods and structure-only controls.

Metrics, candidate sets, and benchmark saturation

Participants debated whether candidate-based retrieval is artificial or realistic. Retrieval can encourage shortcuts, but database search is also part of real metabolomics practice. The likely compromise is to keep both database-ranking tasks and de novo or open-ended tasks. Candidate-set construction was identified as one of the most important design problems for future MassSpecGym releases.

The group also questioned whether top-1 or top-n exact accuracy is sufficient. Proposed alternatives included structural-similarity metrics, maximum common substructure similarity, structural edit distance, scaffold-level accuracy, formula-aware metrics, and metrics that characterize whether top candidates are chemically consistent. A related task between retrieval and full de novo elucidation was also discussed: given a known spectrum-structure pair and a related unknown spectrum, predict the structure of the analog. This analog-discovery task is directly connected to molecular networking and annotation propagation.

2 MS Libraries and Data Quality

Date: June 30, 2026.

The MS libraries session distinguished between two different uses of spectral libraries. Traditionally, libraries were used primarily for annotating spectra. For machine learning, however, the goal is often to collect as much training data as possible, including synthetic compounds and spectra that may not correspond to typical biological metabolites. This creates new questions about data quality, licensing, and benchmark scope.

Low-quality historical spectra is a major concern, including spectra containing only a single peak or essentially only a precursor peak. For MassSpecGym v2 and related benchmarks, the group favored focusing on high-quality signal, including resources such as MSnLib and Enveda data. Participants also discussed whether synthetic compounds should be included in benchmarks. Such compounds may help increase coverage, but models trained or evaluated on them may require explicit adjustment because synthetic chemical space differs from natural-product and metabolomics chemical space.

Licensing and openness were also central. NIST might be able to contribute real challenge examples for CASMI or Kaggle-style competitions, but NIST copyright and licensing constraints may prevent fully open redistribution. The group discussed whether all released data should remain CC0, whether proprietary training data should be forbidden in open benchmarks, and whether competitions should distinguish between unrestricted best-performance tracks and fully open tracks where models and training data must be disclosed.

The session also discussed realistic challenge design. A possible benchmark split could include both molecules absent from common compound libraries and molecules present in libraries, allowing both de novo and retrieval-based approaches to participate. Future challenge editions might initially provide transparent information such as adducts and isotopic patterns, then hide more information in later years. The group suggested keeping the first year simple, for example focusing on protonated positive-mode spectra, while later adding adduct assignment, chimeric spectra, and intentionally messy cases.

3 Generative Models for De Novo Chemical Structure Characterization

Date: June 30, 2026.

The de novo generative-model session focused on the appropriate framing of MS/MS-based structure elucidation. The group agreed that de novo prediction should not be treated as a deterministic mapping from one spectrum to one structure. A single MS/MS spectrum often does not contain enough information to uniquely determine a full molecular structure. Multiple positional isomers, close analogs, or broader equivalence classes may explain the same spectrum.

Participants therefore distinguished several different tasks that are often conflated: database retrieval, near-neighbor discovery, transformation-product or analog discovery, true novel-scaffold discovery, and partial-structure or scaffold-level elucidation. Benchmarks should evaluate these use cases separately. A model that performs well by retrieving close neighbors from pretraining data may be useful for practical annotation, but that does not mean it can discover a genuinely novel natural-product scaffold.

The session compared fingerprint-based intermediate representations, autoregressive SMILES models, diffusion models, graph decoders, and coordinate-based generative approaches. Autoregressive models can work well with BPE tokenization and randomized non-canonical SMILES, but they remain vulnerable to cascading decoding errors. Diffusion or flow models over continuous coordinates may simplify some aspects of molecular generation, but they still require conversion into chemically valid molecular graphs.

A central point was that candidate generation alone is insufficient. Generated molecules need to be re-ranked or refined using forward fragmentation prediction, spectrum simulation, or related scoring methods. Test-time scaling was identified as promising: generate many candidates, score them, refine them, and iterate. However, this strategy depends strongly on the quality of the forward model and on whether the true molecule lies within the model's training distribution.

The group identified confidence estimation as a major unmet need. Models should report when there is a single clear candidate, when there is a small set of indistinguishable candidates, and when the spectrum is insufficient for confident assignment. In some cases, methods should output partial structures, scaffolds, or equivalence classes rather than forcing a complete molecular structure. Additional experimental data, including positive and negative ion modes, multiple spectra per molecule, stochastic fragmentation spectra, and MS^n data, were discussed as ways to make de novo elucidation better posed.

4 Machine-Learning Interatomic Potentials for Tandem MS

Date: June 30, 2026.

This session examined whether machine-learning interatomic potentials (MLIPs) could help predict and interpret tandem mass spectra. Forward spectrum prediction remains difficult, and many current methods are not clearly better than retrieval from close spectral-library matches. Better forward prediction is important for the inverse problem: if an unknown spectrum yields candidate structures, those candidates should be checked by simulating or scoring plausible fragmentation.

The discussion placed MLIPs between graph-only machine-learning models and expensive quantum-chemistry methods. Density functional theory can help interpret fragmentation pathways by evaluating plausible fragments, reaction pathways, and energetics, and automated reaction-path methods such as GRRM/AFIR can explore dissociation channels. However, exhaustive DFT or AFIR-style searches are too expensive for large candidate libraries. MLIPs may offer more chemical physics than graph-only models while remaining much faster than systematic quantum-chemical exploration.

The group identified several possible uses. First, MLIPs could provide 3D information for spectrum prediction, including relaxed conformers, conformer energies, local strain, and activated geometries. Second, they could help rank protonation, deprotonation, adduct, and charge-localization sites, which strongly influence fragmentation. Third, MLIPs could score fragile bonds by approximating local bond weakness, strain release, charge-assisted cleavage, and dissociation energies. Fourth, after candidate fragments have been generated, MLIPs could help score whether each fragment is chemically plausible by estimating relaxation energy, ion stability, neutral-loss stability, or rearrangement costs.

The session therefore framed MLIPs not as a replacement for existing spectrum predictors, but as a physically informed component that could be combined with fragmentation graphs, spectrum simulation, and candidate-ranking systems.

5 Self-Supervised Pre-Training for Mass Spectrometry Spectrum Encoders

Date: June 30, 2026.

This session focused on self-supervised objectives for MS/MS spectrum encoders. The main goal is to learn spectrum representations that improve downstream tasks such as structural elucidation, spectrum retrieval, property prediction, and spectrum similarity search. Participants compared ideas from BERT-style masked modeling, contrastive learning, autoregressive language modeling, diffusion or flow-based generation, and latent-prediction methods such as BYOL or JEPA.

Masked peak modeling was discussed as a natural adaptation of BERT to spectra: treat each spectrum as a sequence or set of peaks, mask a fraction of them, and predict the missing peaks from the rest. This objective is chemically meaningful because it operates directly in peak space, but the notes report only modest gains and identify shortcut risks. If isotope peaks, neutral-loss-related peaks, or other correlated peaks remain visible, the model may reconstruct missing peaks without learning deeper fragmentation chemistry. Isotope-aware and neutral-loss-aware masking were proposed to reduce such shortcuts.

Contrastive learning was discussed as a way to align two views of the same spectrum or molecule. Possible positive pairs include peak subsets, spectra at different collision energies, positive and negative ion spectra of the same molecule, or related MS^n spectra. The major concern is false negatives: in unlabeled data, two spectra treated as negatives may actually be the same compound, an adduct, an isomer, or a close analog. Participants therefore viewed contrastive learning as more reliable when curated labels are available.

Autoregressive peak prediction was identified as a promising alternative. Rather than predicting a canonical next token, a set-based autoregressive loss could reward prediction of any remaining valid peak. Randomized ordering may be preferable because spectra are naturally set-like. The group also discussed using pretraining loss as a diagnostic for spectrum quality, out-of-distribution status, or downstream confidence: spectra that are hard to reconstruct may be noisy, unusual, or intrinsically difficult.

The main takeaway was that no single self-supervised objective is likely to be universal. Pretraining objectives should be chemically grounded, avoid obvious shortcuts, and be evaluated by fast diagnostics such as linear probes, retrieval metrics, embedding-neighborhood structure, and correlations with known molecular relationships.

6 Weird Molecules and Non-Canonical Chemical Spaces

Date: June 30, 2026.

This session addressed molecules that fall outside standard proteomics and metabolomics workflows, including cyclic peptides, polymers, RNA therapeutics, exposomic contaminants, dimers, large natural products requiring MS^n, and glycans. The common theme was that current computational tools are often trained on canonical metabolite or peptide spaces and may fail when confronted with different molecular grammars.

Cyclic peptides were identified as a promising but difficult area for de novo sequencing. They overlap proteomics, metabolomics, and natural-products research, but existing peptide de novo models usually assume the 20 canonical amino acids plus a small number of modifications. Cyclic peptides may contain many non-canonical monomers and modifications. Expanding the vocabulary is possible in principle, but annotated spectra are scarce. Synthetic data, including simulated ring openings and linearizations, may help pretrain models before fine-tuning on limited real spectra.

The session also discussed polymers and pyrolysis coupled with GC-MS, where the sample is not a pure compound but a material producing a series of reproducible but highly similar spectra at specific retention indices. Achieving high-confidence matching will necessitate combining standard spectral matching with novel chromatographic pattern matching metrics.

RNA therapeutics were also described as an emerging biomanufacturing challenge where mass spectrometry can confirm sequence, help characterize modifications, as well as identifying small molecule and peptide contaminants during production. However, intact RNA analysis and modification-aware interpretation is very underdeveloped, both in the experimental and computational domains.

Exposomics and industrial contaminants broaden the chemical space further. Halogenated and fluorinated contaminants, including PFAS-like compounds, are poorly handled by tools trained on ordinary metabolites. Non-covalent dimers and heterodimers were also identified as a recurring source of misleading features: precursor masses can correspond to sums of co-eluting molecules, causing dimers to be mistaken for novel structures. The group proposed quantifying how dimer spectra differ from monomer spectra and deciding whether they belong in libraries, training sets, or only as annotation warnings.

For glycans, participants noted that tree-like branched structures may be better suited to geometric or graph-based models than standard transformer decoders. However, common CID/HCD spectra often provide composition-level information rather than full linkage and branching structure, and some machine-learning claims may rely on labels inferred in silico rather than experimentally resolved structures. A proposed action was to critically review glycan structure-prediction papers to assess whether their labels are experimentally justified.

7 Chemical Space

Date: July 1, 2026.

The chemical-space discussion considered the mismatch between the molecular spaces available to AI models, public compound databases, natural products, and molecules actually observable by mass spectrometry. TMS:4000191he generative space of AI models is larger than resources such as ZINC, natural-product libraries, or PubChem, but this does not mean that all generated molecules are biologically or experimentally relevant.

Participants discussed natural products from the perspectives of biosynthesis and evolution: similar scaffolds may appear across species, and databases such as COCONUT and HMDB provide different windows into biologically plausible chemistry. A practical question was how to determine whether a molecule in ZINC could be formed in natural systems; one suggested approach was to check overlap or similarity with natural-product libraries. The group also raised the question of what molecules cannot be seen by mass spectrometry at all, including limits imposed by ionization, abundance, stability, and instrumentation.

One action item was to create an official platform where people can submit paired MS and NMR spectra for novel molecules. The notes also mention testing models trained on synthetic molecules, such as Enveda compounds, for generalization to natural products.

8 Metabolomics Atlas and Modeling After Structure Elucidation

Date: July 1, 2026.

This session asked whether metabolomics could build an analogue of transcriptomics expression atlases from public datasets. The central idea is to aggregate metabolite or feature measurements across studies, biological conditions, perturbations, sample types, and repositories. Participants agreed that metabolomics is harder to integrate than transcriptomics because of heterogeneous instruments, chromatographic methods, acquisition settings, ionization behavior, feature detection, annotations, and metadata quality.

Metadata was identified as essential. A useful atlas would need information about biological system, perturbation, sample condition, protocol, analytical instrument, acquisition parameters, and experimental variables. Participants debated whether metadata must be formally standardized or merely available and interpretable. Strict ontology mapping may be unrealistic, but lightweight normalization and validation could reduce errors. LLMs were discussed as possible tools for harmonizing heterogeneous metadata terms across studies and languages.

Quantification was a major challenge. Untargeted metabolomics generally does not provide absolute quantification across studies, and abundance estimates are affected by batch effects, instrument sensitivity, chromatography, ionization, acquisition settings, and processing pipelines. The group debated whether an atlas should prioritize quantitative abundance, presence/absence, molecular discovery, or a hybrid model. Presence/absence may be sufficient for some molecule-discovery questions, whereas biomarker discovery likely requires more reliable abundance estimates.

The notes describe an ongoing reanalysis of public metabolomics datasets from repositories such as MassIVE, Metabolomics Workbench, and Asian repositories. Approximately 5,000 datasets were mentioned, with priority given to datasets containing MS/MS data. The pipeline performs automated feature finding, adduct collapsing, alignment, and related preprocessing. The key question is whether the resulting quantitative outputs are scientifically reliable. Follow-up actions assign Ming to provide the reprocessed quantitative data and Kozo and collaborators to evaluate whether the quantification is useful and what validation metrics are needed.

9 Kaggle/CASMI Competitions for Metabolomics

Available notes: proposal-level and action-item notes only.

The proposal document asks how the community can create more Kaggle-style competitions for computational metabolomics. Enveda was expected to organize CASMI2026 on Kaggle, and the group discussed whether research funds could cover prize money, whether application datasets could be connected to actual discovery commitments, and whether competition-winning strategies could be incorporated into real analysis pipelines.

Action items from the seminar indicate that the community intends to organize the next CASMI, contribute pre-release unannotated NIST29 spectra to CASMI, and potentially connect competition design with MassSpecGym 2.0. These discussions align with the broader benchmark theme: public benchmarks support method development, while blind or semi-blind competitions provide stronger tests of real-world generalization.

The discussion also drew parallels with the CASP competition and considered the feasibility of mirroring the CASP setup more closely – for example, by incentivising researchers to contribute novel molecules and their corresponding spectra, aiming for on the order of 100 novel molecules per competition. This would make it possible to measure the practical generalization of methods, while still allowing participants to use all available public data without restriction.

10 Evaluation Metrics for *De Novo* Structure Prediction

Date: June 30, 2026.

The session opened with a discussion of the limitations of the current MassSpecGym 1.0 metrics. First, exact-structure accuracy was considered too strict to differentiate between methods that are still far from correct top-1 predictions. Second, the MCES distance averaged over the validation or test set does not behave well, because the reported values are shaped by the two thresholds involved in its computation: the threshold of 15 that switches from exact to approximate evaluation, and the timeout threshold of 100. Averaging over a set therefore mixes exactly computed distances with clipped and approximated ones, which makes the resulting number difficult to compare across methods. Third, Tanimoto similarity was found hard to interpret.

The remainder of the session focused on how to improve these metrics, and produced two concrete outcomes. The first is that in MassSpecGym 2.0 the averaged MCES metric can be replaced by the fraction of validation or test entries with an MCES distance below a threshold k. The value of k can either be fixed, chosen so that it is informative given current state-of-the-art results, or methods can be compared across a range of k values, for example by reporting the area under the resulting curve. This formulation avoids averaging over clipped values and has a directly interpretable meaning: the share of spectra for which a method produces a prediction within k edge edits of the true structure. This metric can be understood as a "relaxed accuracy", because when $k = 1$ it reduces to the standard accuracy metric.

The second outcome concerns the fingerprint-based similarity metric, which can be improved in two ways. First, the hashing step can be omitted and the bags of fragments compared directly. This makes the metric more interpretable and more accurate, as it eliminates hash-collision artifacts. Second, count-based fingerprints can be used instead of binary ones, as they are generally reported to perform better across studies.

Several further metrics were discussed, including similarity in learned molecular embedding spaces, scaffold accuracy, and accuracy of predicted ClassyFire classes. The agreement, however, was that these metrics are either fuzzy or hard to define for an arbitrary pair of molecules, and are therefore less suitable as primary benchmark metrics.

A related point raised in the session is that 2D InChIKeys are sensitive to tautomers, whereas for evaluation purposes one would like them not to be, so that tautomeric variants of the same structure are not counted as incorrect predictions. Whether a tautomer-insensitive hashing alternative exists that is otherwise as convenient as the InChIKey remains an open question.

11 MassSpecGym v2

Date: July 2, 2026.

The MassSpecGym v2 session consolidated many lessons from the benchmark discussions. The group agreed that the next benchmark should be substantially larger, ideally at least an order of magnitude larger than the current dataset. It should use higher-quality MS signal, include additional sources such as MSnLib, Enveda, MERLIN, and newly annotated data, and avoid low-quality test spectra with too few peaks.

Participants strongly supported adding negative ionization mode, multiple adducts, multiple collision energies, and molecule-level spectral bundles. At the same time, the benchmark must reflect realistic missingness: not every molecule will have every ion mode, adduct, or collision energy. The preferred direction was to provide all available spectra while reporting predefined modality ablations such as positive-only, negative-only, and all-available-spectra performance. A single aggregate leaderboard score could be misleading unless models are tagged by input modalities and external data use.

Split design was a major topic. The current hard split strategy is useful for testing generalization, but it may be unrealistically hard or may distort molecular-size distributions. The group discussed MCES, InChI-based disjointness, clustering, nearest-neighbor similarity, molecule-size-aware thresholds, and calibration using clean baselines. Participants also suggested reporting performance as a function of structural difficulty, showing where methods succeed or fail relative to similarity with training molecules.

The benchmark should also improve standardization: tautomer handling, charge normalization, avoidance of ambiguous 3D stereochemistry, collision-energy metadata recovery, and cleaner handling of adducts. The notes mention adding standard tests for pretraining, releasing large-scale pretrained models trained without leakage, and including educational material on correct model evaluation.

Analog discovery was again discussed as a candidate challenge. Given one known spectrum-structure pair and an unknown related spectrum, the task would be to infer the related analog. This problem is relevant to ModiFinder/ModiStruct-style tasks, matched molecular pairs, and molecular-network propagation. The group did not fully settle whether this belongs in MassSpecGym v2 or a separate specialized benchmark.

The session also highlighted community-facing infrastructure. The MassSpecGym website, documentation, and leaderboard should be improved with clearer task descriptions, updated baselines, visualizations of which compounds are solved or missed, and explicit labels for leakage-avoidance status and external-data use. The notes mention a target connected to the next NeurIPS cycle and possible outreach through tutorials, workshops, or ASMS activities.

List of Participants

The meeting brought together 24 participants from academia, national research institutes, and industry. They are listed below alphabetically by family name, together with their affiliations.

Name	**Affiliation**
Tess Beavers	Novogaia, Inc.
Philipp Benner	Federal Institute for Materials Research and Testing (BAM)
Wout Bittremieux	University of Antwerp
Leopold Brinkmann	Friedrich Schiller University Jena
Anton Bushuiev	Czech Technical University in Prague
Roman Bushuiev	IOCB Prague
Garrison W. Cottrell	University of California San Diego
Gaetan De Waele	Ghent University
Kai Duehrkop	Bright Giant GmbH
Abraham Frandsen	Enveda Biosciences
Filip Jozefov	IOCB Prague
Paul Krzakala	Ecole Polytechnique
Hongxuan Liu	Massachusetts Institute of Technology
Tytus D. Mak	National Institute of Standards and Technology (NIST)
Kozo Nishida	RIKEN
Kouharu Otsuki	Toho University
Tomáš Pluskal	IOCB Prague
Juho Rousu	Aalto University
Michael A. Skinnider	Princeton University
Ash Sze	Tufts University
Ichigaku Takigawa	The University of Tokyo
Haixu Tang	Indiana University Bloomington
Koji Tsuda	The University of Tokyo
Mingxun Wang	University of California Riverside

Meeting Schedule

Check-in Day: June 28, 2026

- Check-in at Shonan Village Center.

Day 1: June 29, 2026

- Benchmarking structure elucidation methods.
- MassSpecGym current state, benchmark audit, SpectraVerse/S4, leakage, data splits, and shortcut learning.

Day 2: June 30, 2026

- Generative models for de novo chemical structure characterization.
- MLIPs for tandem MS prediction and interpretation.
- MS libraries and data quality.
- Self-supervised pre-training for mass spectrometry spectrum encoders.
- Weird molecules and non-canonical chemical spaces.

Day 3: July 1, 2026

- Chemical space.
- Metabolomics atlas and modeling after structure elucidation.

Day 4: July 2, 2026

- MassSpecGym v2.
- Action items and future community organization.

Summary of Discussions

Across the meeting, the strongest recurring theme was that AI for computational metabolomics cannot advance through model architecture alone. The community needs benchmarks that are realistic enough to measure useful progress, but controlled enough to diagnose leakage, shortcut learning, and overfitting. MassSpecGym has already played a central role in making the field accessible to machine-learning researchers, but the audit of recent papers showed that benchmark design, data-access rules, and evaluation tools must become more rigorous.

A second major theme was the difference between clean benchmark tasks and real metabolomics. Real experimental data contain co-elution, chimeric spectra, missing modalities, low-abundance compounds, inconsistent acquisition settings, contaminants, and uncertain annotations. Several sessions proposed intermediate tasks that sit between current clean benchmarks and fully unconstrained real-world analysis: high-quality but larger datasets, molecule-level spectral bundles, analog-discovery tasks, controlled noisy-data challenges, and blind competitions.

The meeting also emphasized that the chemical space relevant to metabolomics is broader than standard spectral libraries and narrower than unconstrained generative chemistry. Natural products, cyclic peptides, glycans, RNA therapeutics, polymers, contaminants, dimers, and synthetic compounds each pose different modeling and evaluation challenges. The group repeatedly returned to the need for models to report uncertainty, candidate sets, partial structures, or equivalence classes rather than always forcing a single exact molecular prediction.

Finally, participants discussed infrastructure and community development: better spectral libraries, shared deposition of MS and NMR data, public-data reanalysis, improved leaderboards, CASMI/Kaggle competitions, workshops, and outreach to machine-learning venues. These action-oriented discussions suggest that one of the main outputs of the meeting will be coordinated community infrastructure rather than a single technical conclusion.

Summary of New Findings

The first major finding is that many current benchmark results in small-molecule mass spectrometry may overstate real progress. Leakage can enter not only through exact test structures, but also through near-neighbor molecules, pretrained decoders, spectrum simulators, formula predictors, synthetic data, canonicalization artifacts, and candidate-set biases. Structure-only and library-likeness baselines can perform surprisingly well, demonstrating that some tasks partly measure database bias rather than spectrum understanding.

The second finding is that de novo elucidation should be reformulated. In many cases, a single MS/MS spectrum does not support a unique full structure. Useful systems should produce ranked candidate sets, confidence estimates, partial structures, scaffold-level assignments, or groups of indistinguishable structures. Near-neighbor discovery, analog discovery, and true novel-scaffold discovery should be evaluated separately.

The third finding is that future benchmarks should operate at the molecule level rather than only the single-spectrum level. Multiple spectra, ion modes, adducts, collision energies, and eventually MS^n data can make tasks more realistic and more informative. However, missing modalities are part of reality and must be represented explicitly rather than hidden by benchmark design.

The fourth finding is that self-supervised pretraining for mass spectrometry needs chemically meaningful objectives. Masked peak modeling, contrastive learning, autoregressive peak prediction, MS^n-based objectives, isotope-aware masking, and latent-prediction methods all have promise, but each has shortcut risks. The group favored objective combinations and fast diagnostics that test whether embeddings preserve chemically meaningful relationships.

The fifth finding is that non-canonical molecules are not an edge case. Cyclic peptides, glycans, polymers, RNA therapeutics, exposomic contaminants, dimers, and large natural products expose limitations in standard metabolomics pipelines. These spaces may require new representations, new libraries, new fragmentation data, and new evaluation standards.

Identified Issues and Future Directions

The meeting produced several concrete future directions. The most developed is MassSpecGym v2: a larger, cleaner, better documented benchmark with controlled training data, high-quality test spectra, negative mode, multiple adducts, multiple collision energies, improved splits, stronger baselines, clearer modality tags, and more informative structural metrics. Related benchmark directions include analog discovery, semi-realistic noisy-data tasks, and independent blind challenges.

Spectral-library development is another major direction. The notes identify the need to publish MSnLib, explore Enveda and NIST contributions, enrich libraries with unique molecules from GNPS submitters, establish minimum acquisition and quality-control requirements, and resolve licensing constraints for AI training. Community members also proposed a shared platform for paired MS–NMR deposition of novel molecules and a letter to the Journal of Natural Products encouraging data sharing.

The metabolomics-atlas discussion suggests a third direction: repository-scale reanalysis of public metabolomics data. This will require defining primary scientific use cases, minimal metadata requirements, feature-collapsing strategies, annotation standards, quality-control filters, and validation metrics. The immediate action item is for Ming to share quantitative outputs and for Kozo and collaborators to evaluate their reliability.

Community organization was also emphasized. Recorded action items include organizing a NeurIPS 2027 workshop, organizing the next CASMI, contributing NIST29 spectra to CASMI, proposing ASMS and MetSoc workshops, exploring another Shonan meeting, and building dashboards or websites for MassSpecGym and target-compound tracking.

Recorded Action Items from Available Notes

- Organize a NeurIPS 2027 workshop for MassSpecGym 2.0.
- Share 180,000 Enveda compounds with the Pluskal lab for MS^n acquisition.
- Enrich MSnLib by contacting authors or submitters of unique molecules deposited to GNPS.
- Create a platform for shared MS–NMR deposition of novel molecules.
- Send a letter to the Journal of Natural Products to encourage data sharing.
- Organize the next CASMI and contribute new NIST29 spectra.
- Evaluate repository-scale quantitative metabolomics data.
- Define minimum MS/MS acquisition and quality-control requirements for benchmark inclusion.
- Improve MassSpecGym documentation, website, leaderboard, and external-data-use labeling.
- Explore future workshops at ASMS, MetSoc, NeurIPS, and possibly another Shonan meeting.