Reusability report: Annotating metabolite mass spectra with domain-inspired chemical formula transformers

Janne Heirman[1], **Wout Bittremieux**[1,*]

[1]Department of Computer Science, University of Antwerp, 2020 Antwerpen, Belgium.

Corresponding author: wout.bittremieux@uantwerpen.be

We present an in-depth exploration of the Metabolite Inference with Spectrum Transformers (MIST) tool for annotating small molecule mass spectrometry (MS) data, focusing on its reproducibility and generalizability. MIST innovates by integrating a "chemical formula transformer" to process MS/MS spectra, aiming to bridge the substantial knowledge gap in untargeted MS studies, where only a fraction of spectra are confidently annotated. Here, we critically assess MIST's reproducibility by following the tool's original training and testing protocols, encountering minor challenges but largely succeeding in replicating results. We also evaluate MIST's generalizability by applying it to an external dataset from the CASMI 2022 challenge, revealing insights into the model's performance on previously unseen data. An ablation study further investigates the impact of various model features on database retrieval performance, suggesting that some algorithmic complexities may not significantly enhance performance. Through rigorous evaluation, this work underscores the challenges and considerations in developing robust computational tools for MS data analysis. We advocate for community-wide efforts in benchmarking, transparency, and data sharing to foster advancements in metabolomics and computational biology.

Introduction

Despite substantial advancements in small molecule mass spectrometry (MS) in recent decades, the field faces a persistent challenge. Although an untargeted MS study can acquire thousands to millions of MS/MS spectra, current state-of-the-art methods are only able to confidently annotate around 5 % to 10 % of these spectra on average.[1] This limitation underscores a significant knowledge gap, constraining our capacity to accurately determine the molecular structures present in such studies and, by extension, diminishing the potential impact

of numerous biological investigations.

To address this challenge, Goldman et al. [2] have recently introduced the Metabolite Inference with Spectrum Transformers (MIST) tool to annotate MS/MS spectra. MIST incorporates a novel "chemical formula transformer" aimed at integrating domain-specific knowledge into the architecture of deep neural networks. This approach deviates from traditional methods that represent peaks in MS/MS spectra as discrete, binned values to process them with neural networks. Instead, MIST represents spectra through the chemical formulas associated with their peaks, which are then processed by a transformer neural network.[3] MIST further enriches its model by encoding neutral loss relationships between fragment ions and incorporating substructure prediction as an auxiliary training objective. To address the challenge of limited large-scale training datasets in the domain of small molecule MS, MIST employs a strategy to simulate spectra, thereby expanding the size and diversity of its training corpus. Additionally, rather than directly predicting full-length fingerprints, MIST adopts a progressive inference methodology by first predicting lower-resolution fingerprints, which are subsequently refined into their full-resolution counterparts.

The fingerprints predicted by MIST can be directly utilized for spectrum annotation, employing nearest neighbor searching to match the MS/MS-derived fingerprints against those extracted from molecular structures within chemical databases, such as PubChem.[4] MIST also harnesses a metric learning approach, utilizing a contrastive model to transform both spectral and chemical fingerprints into embeddings, which are then subjected to nearest neighbor searching.

Initial findings by Goldman et al. [2] indicate that the fingerprints predicted by MIST from MS/MS spectra surpass those generated by CSI:FingerID[5] in terms of accuracy. However, it is only with the application of contrastive fine-tuning that MIST demonstrates a superior performance in database retrieval tasks. To understand these results, this reusability report aims to critically assess the MIST reproducibility and evaluate its novel neural network architecture contributions. Specifically, we describe the extent to which the originally reported results can be replicated and apply MIST to an external dataset from the CASMI 2022 challenge in order to assess its generalizability and potential broader impact on the field.

Reproducibility

To evaluate the MIST reproducibility, we followed the training and testing instructions as delineated in the documentation on its GitHub repository (`https://github.com/samgoldman97/mist`). It is crucial to note that the version corresponding to the published paper by Goldman et al. [2] is MIST v1.0.1. The latest version on GitHub, MIST v2.0.0, introduces a few enhancements, including the elimination of the dependency on SIRIUS[6]

for annotating MS/MS spectrum peaks with their corresponding sub-formulas. For our reproducibility analysis, we focused on MIST v1.0.1 to ensure alignment with the version detailed in the publication by Goldman et al. [2].

To install MIST, we successfully followed the repository's instructions for cloning and local installation. Note that this required manual installation from source, as MIST is not packaged as a standalone Python tool. Pretrained model weights are provided on Zenodo and are automatically downloaded as part of the quick start demo, after which they can be easily used for inference. Hence, using the pretrained version of MIST is a relatively user-friendly process for scientists familiar with typical command-line bioinformatics software.

Next, we retrained MIST's various components as outlined in the original study: the fingerprint prediction model, a baseline feed-forward neural network model, the contrastive model for embedding generation, and the forward model for simulating MS/MS data. In addition to public data from GNPS[7] and MoNA, the primary MIST model was also trained on proprietary data from the commercial NIST20 spectral library. As this model is not publicly available, instead we utilized only data coming from public sources to retrain the models. While weights for the fingerprint prediction model and the contrastive model corresponding to the public data subset have also been provided on Zenodo by Goldman et al. [2], the baseline and simulation model were not provided.

Goldman et al. [2] mention that the training process was undertaken using a three-fold structure-disjoint cross-validation methodology, although they did not specify which cross-validation split was employed or whether an integration strategy was used to obtain the reported results. For our analysis, we consistently utilized the first cross-validation split. Training the MIST model and the baseline model was successful, using the provided data excluding simulated spectra. Retraining the contrastive model required initial preparation of a decoy chemical database and the resolution of several code discrepancies.

When including the forward simulation model, we encountered several challenges due to sparse documentation and hard-coded file paths in the corresponding scripts. After these issues were addressed, incorporating simulated spectra during training of the fingerprint prediction model resulted in a reduction of the training time until convergence from approximately four hours to approximately three hours, indicating potential efficiency gains through data augmentation.

Upon completion of the training phase, we conducted an evaluation using the test set of the first cross-validation split, consisting of 819 MS/MS spectra, comparing the performance of our retrained models against the pretrained models that are shared publicly. This revealed a high degree of similarity and correlation in the fingerprints predicted by both the original and retrained models (figure 1), affirming the reproducibility of MIST's training process, notwithstanding some inherent variability associated with neural network training

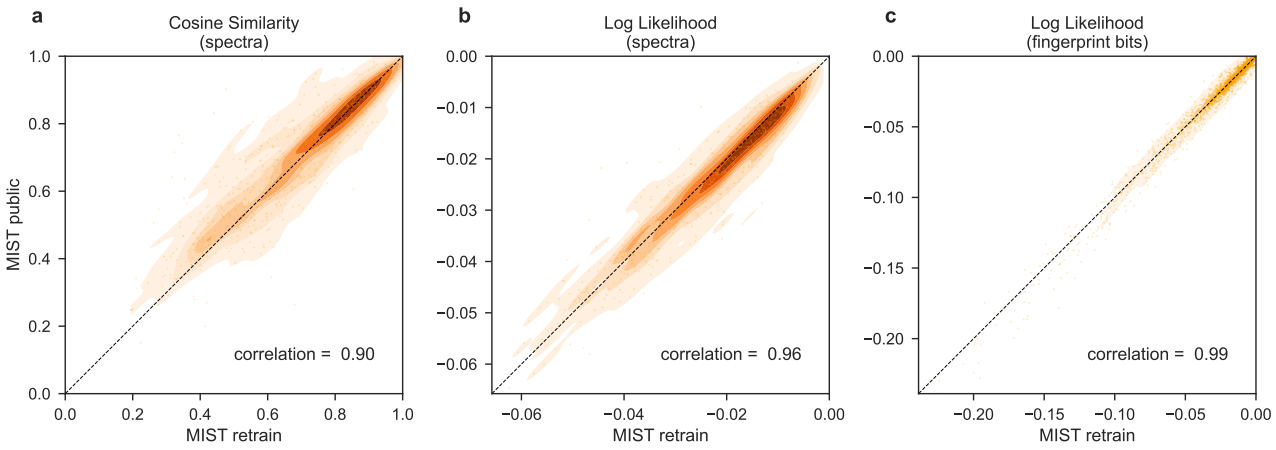

Figure 1: The retrained version of MIST exhibits similar performance as using the public MIST model weights. **(a, b)** Molecular fingerprints were predicted for 819 MS/MS spectra in the test set using the original MIST model and our retrained MIST model. The fingerprint similarity for each spectrum compared to fingerprints for its molecule based on the cosine similarity (a) and the log likelihood (b) are evaluated and plotted. Both models show a very high correlation in their predicted fingerprint similarities. **(c)** Equivalent evaluation showing the log likelihood of predicting each fingerprint bit correctly across all test spectra.

across different systems, hardware configurations, and datasets.

Reusability

To evaluate the generalizability of MIST, we applied it to predict fingerprints and perform database annotation for a set of novel MS/MS spectra from the CASMI 2022 challenge. This dataset, released subsequent to the training of the published version of MIST, offers a stringent benchmark for evaluating MIST's performance on previously unseen and independent data.

Applying MIST to this new dataset presented certain challenges, primarily due to a scarcity of documentation on how to preprocess and integrate new data into MIST's workflow. This included a reliance on SIRIUS[6] for data preprocessing—including the creation of its input files—and a retrieval database, necessitating an examination of the source code to deduce the requisite steps. Although not considered in this analysis, it is worth noting that the most recent MIST v2.0.0 has removed the dependency on SIRIUS, streamlining the process. Additionally, modifications were required to address hard-coded file paths in the codebase.

We employed both the publicly available MIST model and our retrained version to predict fingerprints for 170 protonated, singly charged spectra from CASMI 2022. The results from both models were highly comparable (figure 2a–c), further underscoring MIST's reproducibility, albeit with the caveat that certain assumptions regarding the architecture and hyperparameters of the pretrained model had to be made due to the lack of explicit documentation. Our retrained model incorporated all optional features, including pairwise neutral loss,

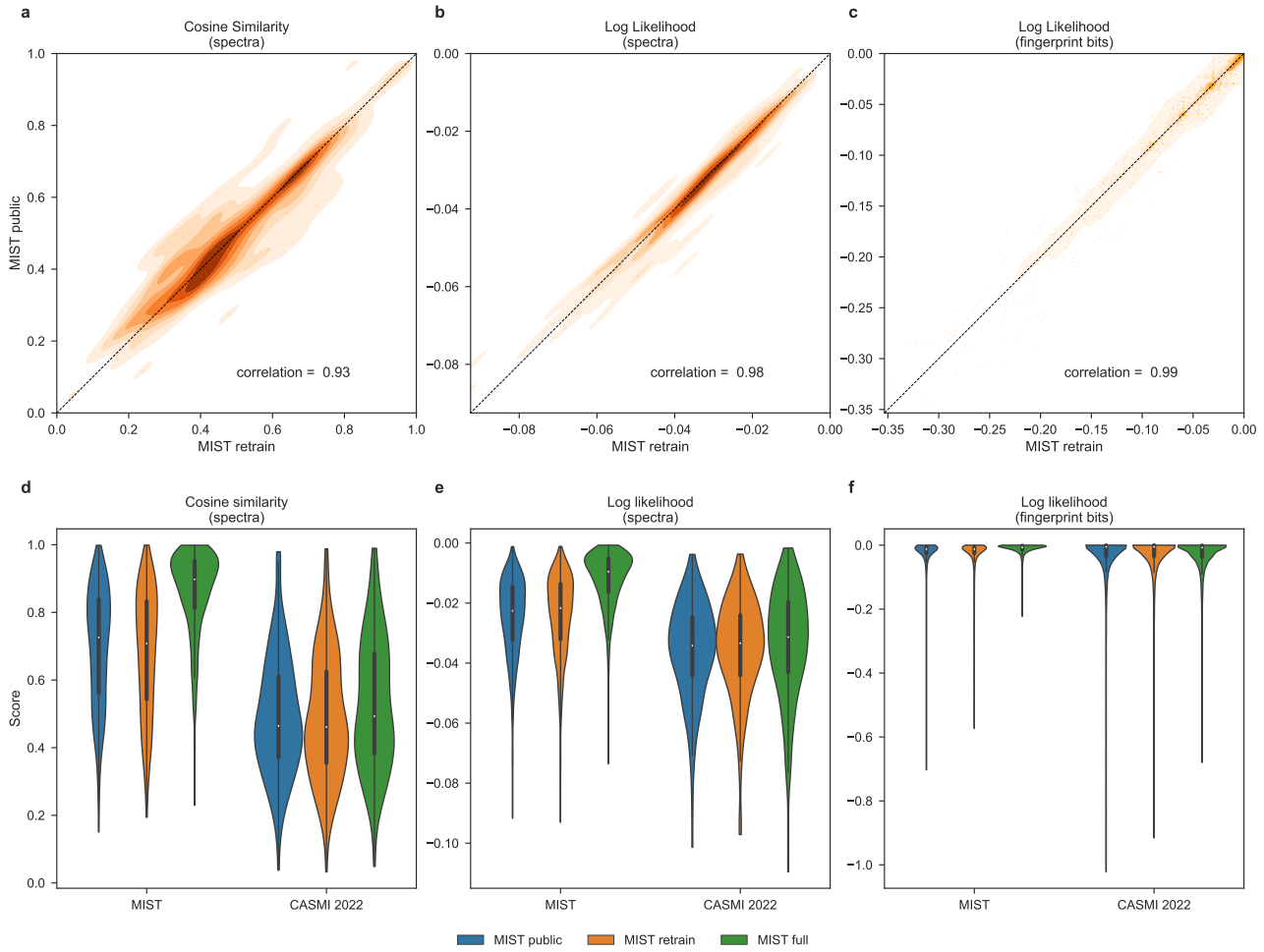

Figure 2: MIST fingerprint prediction performance on the CASMI 2022 dataset. **(a, b, c)** Molecular fingerprints were predicted for 170 MS/MS spectra using the public MIST model and our retrained MIST model. Both models achieve similar performance when evaluating the similarity between the spectral fingerprints and the chemical fingerprints based on the cosine similarity (a), the log likelihood (b), and the log likelihood of predicting each fingerprint bit correctly across all test spectra (c). **(d, e, f)** Evaluation of the fingerprint prediction performance on the original MIST test data and the external CASMI 2022 data. Whereas the full MIST model obtained higher performance on the MIST test data, all three MIST versions exhibit a significant performance decrease in terms of the cosine similarity (d), the log likelihood (e), and the individual fingerprint log likelihood (f) when applied to the CASMI 2022 data.

MAGMa[8] substructure annotation, data augmentation, and unfolding.

An interesting observation is a significant decrease in performance of both the public and retrained MIST models on the CASMI 2022 dataset (figure 2a–c), however, which is in contrast to the performance levels reported in the original MIST publication and on the provided test set (figure 1). To delve deeper into this discrepancy, we also assessed the full MIST model described by Goldman et al. [2], which was trained on both public and commercial data, and which was generously made available by the authors. Although this model demonstrates an enhanced performance on the original MIST test data, its efficacy on the CASMI 2022 dataset mirrored the substantial decline observed with the other models.

Next, we evaluated the MIST database retrieval performance, including a comparison with the benchmark

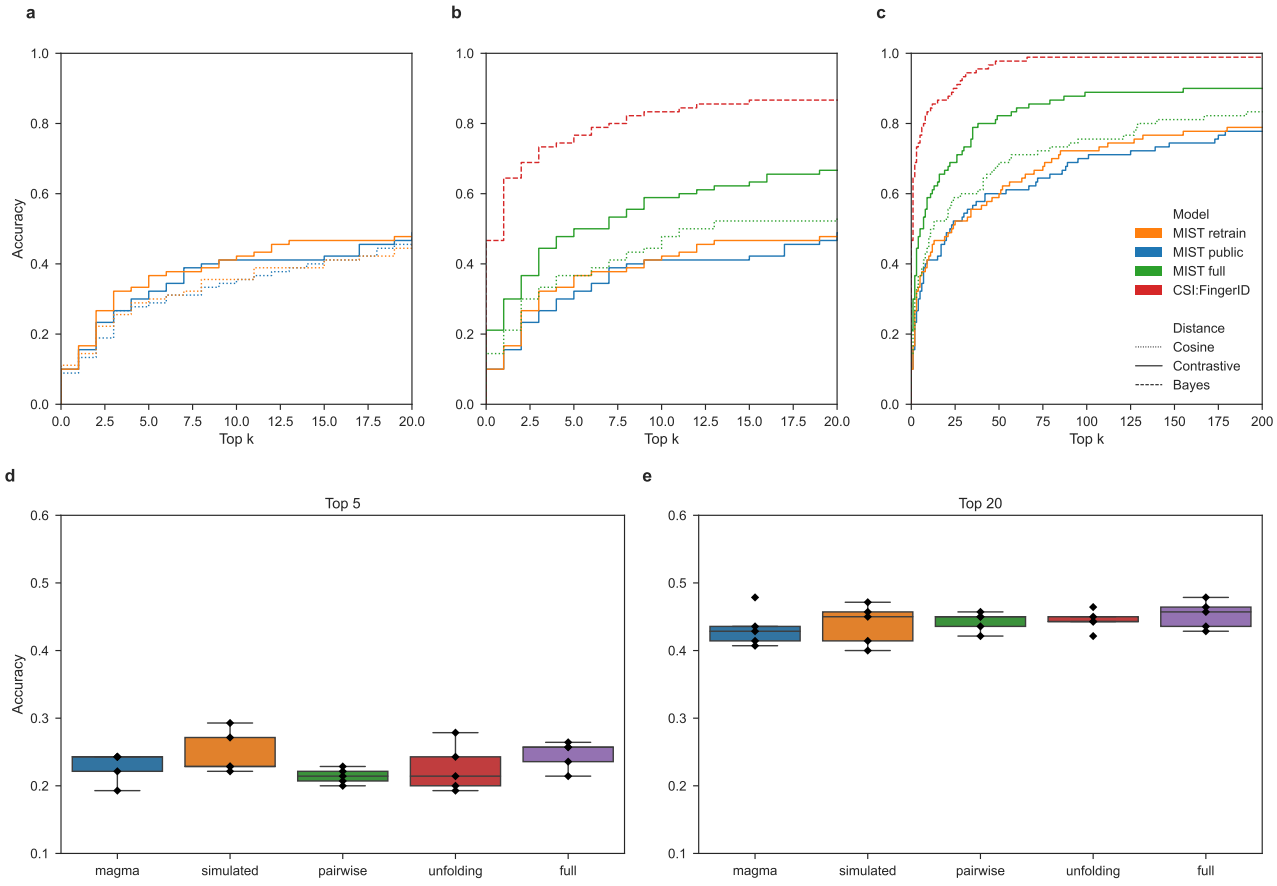

Figure 3: Database retrieval performance for various MIST models and CSI:FingerID on the CASMI 2022 data. Top-k accuracy is computed by ranking candidate compounds for each spectrum and computing the fraction of spectra for which the true compound appears in the top-k rankings. Cosine distances use fingerprint predictions for retrieval, contrastive distances uses contrastive latent spaces, and "Bayes" is a custom CSI:FingerID fingerprint distance function. **(a)** Comparison between the public MIST model and our retrained MIST model, both achieving similar database retrieval performance. **(b, c)** Similar to (a), but now also including the full MIST model and CSI:FingerID. The full MIST model outperforms both MIST models that were only trained on public data. CSI:FingerID achieves the best performance, strongly outperforming MIST. **(d, e)** Ablation results for MIST without MAGMa substructure prediction, without simulated data, without neutral loss featurization (pairwise), and without fingerprint unfolding. Each model was trained five times with different random seeds.

standard, CSI:FingerID.[5] Initial comparisons among the public and retrained MIST versions reveal that their performances are highly similar (figure 3a), indicating consistent spectrum annotation capabilities across versions. Specifically, when utilizing predicted fingerprints for database retrieval using the cosine distance, the pretrained MIST model achieves a top-1 annotation accuracy of 8.9 %, in contrast to 11.0 % for the retrained model. When employing contrastive embeddings for retrieval, both models exhibit identical top-1 annotation accuracy of 10.0 %.

Compared to MIST trained exclusively on public data, the full MIST model, which includes both public and proprietary data in its training, achieves a significantly stronger performance (figure 3b). However, it is noteworthy that across the CASMI 2022 dataset, the MIST database retrieval performance does not match the higher levels previously reported by Goldman et al. [2] on a different test set. This observation is anticipated, considering the lower precision of fingerprint predictions on this novel dataset. Furthermore, CSI:FingerID significantly

outperforms all MIST models on the CASMI 2022 dataset, contrary to the similar performance between MIST and CSI:FingerID reported by Goldman et al. [2].

Building on the initial ablation study to evaluate the impact of various MIST features (pairwise neutral loss, MAGMa[8] substructure annotation, data augmentation, and unfolding) on fingerprint prediction accuracy,[2] we extended this analysis to the database retrieval level. Fingerprint prediction is only an intermediate step during database annotation, and as previously reported,[2] improvements during this step might not necessarily correlate with enhanced performance on the full task.[2] Given that certain features of MIST introduce considerable algorithmic complexity with only marginal improvements in fingerprint prediction accuracy, we sought to understand how this translates to changes in performance during database retrieval.

Our findings from evaluating the top-5 annotation accuracy across all models (figure 3d) suggest a uniform performance. The absence of MAGMa substructure labeling had only a minimal impact on retrieval accuracy, whereas a somewhat more notable reduction in annotation accuracy can be observed when the fingerprint unfolding and neutral loss featurization components are removed. This consistent performance pattern persisted in the evaluation of top-20 annotation accuracy, revealing negligible differences between the models (figure 3e). Statistical analysis using Cochran's Q test to compare the efficacy of multiple classifiers yielded p-values of 0.564 and 0.308 for top-5 and top-20 annotation accuracy levels, respectively, which do not meet the threshold for statistical significance to reject the null hypothesis that all models perform equally well. These results suggest that the added computational complexity introduced by various MIST features, including dependencies on external tools, may not provide a commensurate improvement in database retrieval performance, indicating the need for a critical reassessment of these components within the model's architecture.

Discussion

In this study, we have meticulously assessed the reproducibility and reusability of MIST for the prediction of chemical fingerprints from MS data. Our endeavors to replicate the training and fingerprint prediction of MIST were largely successful, albeit not without encountering several minor challenges. These challenges, including small bugs in the source code and occasional gaps in the documentation, underscore the inherent complexities involved in developing advanced bioinformatics tools. Despite these hurdles, the overall good quality of MIST's documentation and source code facilitated the resolution of these issues, albeit requiring expert effort. This experience demonstrates the significant disparity between the development of a research prototype and the refinement of a tool to a production-ready state, emphasizing the necessity for robustness to ensure successful application across varied scientific environments. The absence of tests within the MIST code repository further highlights this gap, suggesting that the integration of software engineering best practices could greatly enhance

the development process, ensuring the production of high-quality scientific software while saving time in the long term.

Our findings corroborate previous results by Goldman et al. [2] regarding the discrepancy between MIST's fingerprint prediction accuracy and its database retrieval performance. This observation, particularly highlighted through MIST's underperformance on an external test set from the CASMI 2022 challenge, raises questions about the model's generalizability and the nuanced challenges of model evaluation in computational biology. The differential performance on this independent dataset underscores the complexities associated with assessing learned models' applicability to diverse and heterogeneous data sources.

However, it is imperative to view these findings not as a critique of MIST but rather as a catalyst for broader community engagement in the development and benchmarking of computational tools. The initiative by Goldman et al. [2] to release their training datasets, cross-validation splits, and model configurations sets a commendable precedent for transparency and reproducibility in computational biology and machine learning. This open sharing of resources is crucial for fostering a community-wide effort towards robust benchmarking standards.

Data availability presents another critical challenge, as highlighted by the partial reliance of the main MIST model on a combination of public and proprietary datasets, which restricts the full replicability of its reported performance. The increasing reliance of state-of-the-art machine learning models on large-scale datasets underscores the pivotal role of data availability in driving progress within the field of metabolomics and beyond. This scenario suggests that future advancements may hinge not only on algorithmic innovation but also significantly on the accessibility of comprehensive and accurate datasets.

Finally, MIST has introduced innovative algorithmic approaches for integrating domain-specific knowledge into neural network models, exemplified by the chemical formula transformer and other MS-specific features. Through our ablation study aimed at discerning the impact of these various features on database retrieval performance, it became apparent that the additional complexity introduced by certain features may not substantively enhance model performance. Notably, the reliance on external tools, such as MAGMa,[8] and the adjustments made in MIST v2.0.0 to remove its dependency on SIRIUS,[6] highlight an important area for continued development.

In conclusion, while MIST presents a promising framework for fingerprint prediction from MS/MS data, further exploration and validation are required to fully understand its generalizability and efficacy across diverse datasets, thereby reinforcing its potential contribution to the field of computational mass spectrometry.

Data availability

The public MS/MS data used to train MIST can be downloaded following the instructions in MIST's documentation, available at https://github.com/samgoldman97/mist/tree/v1.0.1. The CASMI 2022 dataset, as initially processed by Young et al. [9], is available at https://github.com/Roestlab/massformer, and its version compatible to be used with MIST at https://zenodo.org/records/10794423.

Code availability

The forked MIST code is available as open source on GitHub under the MIT license at https://github.com/Janne98/mist/tree/report_main and has been deposited to Zenodo for permanent archival at https://zenodo.org/records/10801861.

Acknowledgments

We want to thank Samuel Goldman for sharing the model weights for the full MIST model. We want to thank Adamo Young for sharing the preprocessed CASMI 2022 data. J.H. and W.B. acknowledge support by the University of Antwerp Research Fund (BOF DOCPRO4 49298).

Author contributions

Janne Heirman: methodology, software, validation, investigation, writing – original draft, writing – review & editing, visualization. **Wout Bittremieux**: conceptualization, methodology, writing – original draft, writing – review & editing, supervision, project administration, funding acquisition.

Competing interests

The authors declare no competing interests.

References

(1) Bittremieux, W., Wang, M., Dorrestein, P. C. The Critical Role That Spectral Libraries Play in Capturing the Metabolomics Community Knowledge. *Metabolomics* **2022**, *18*, DOI: 10.1007/s11306-022-01947-y.

(2) Goldman, S., Wohlwend, J., Stražar, M., Haroush, G., et al. Annotating Metabolite Mass Spectra with Domain-Inspired Chemical Formula Transformers. *Nature Machine Intelligence* **2023**, *5*, 965–979, DOI: 10.1038/s42256-023-00708-3.

(3) Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al. In *Advances in Neural Information Processing Systems*, Curran Associates, Inc.: 2017; Vol. 30.

(4) Kim, S., Chen, J., Cheng, T., Gindulyte, A., et al. PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Research* **2021**, *49*, D1388–D1395, DOI: 10.1093/nar/gkaa971.

(5) Dührkop, K., Shen, H., Meusel, M., Rousu, J., et al. Searching Molecular Structure Databases with Tandem Mass Spectra Using CSI:FingerID. *Proceedings of the National Academy of Sciences* **2015**, *112*, 12580–12585, DOI: 10.1073/pnas.1509788112.

(6) Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A. A., et al. SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information. *Nature Methods* **2019**, *16*, 299–302, DOI: 10.1038/s41592-019-0344-8.

(7) Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., et al. Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **2016**, *34*, 828–837, DOI: 10.1038/nbt.3597.

(8) Ridder, L., van der Hooft, J. J. J., Verhoeven, S. Automatic Compound Annotation from Mass Spectrometry Data Using MAGMa. *Mass Spectrometry (Tokyo, Japan)* **2014**, *3*, S0033, DOI: 10.5702/massspectrometry.S0033.

(9) Young, A., Wang, B., Röst, H. MassFormer: Tandem Mass Spectrum Prediction for Small Molecules Using Graph Transformers http://arxiv.org/abs/2111.04824 (accessed 03/07/2024), preprint.

(10) McKinney, W. In *Proceedings of the 9th Python in Science Conference*, ed. by van der Walt, S., Millman, J., Austin, Texas, USA, 2010, pp 51–56, DOI: 10.25080/Majora-92bf1922-00a.

(11) Cochran, W. G. The Comparison of Percentages in Matched Samples. *Biometrika* **1950**, *37*, 256–266, DOI: 10.1093/biomet/37.3-4.256.

Supplementary information

MIST installation

The installation process of MIST was executed following the instructions provided in its GitHub repository's README file, by cloning the repository, creating a dedicated Conda environment, installing the dependencies, and ultimately installing MIST itself from the cloned source. While creating the Conda environment, some package version conflicts emerged, notably with the `torchvision` dependency, which could safely be removed. Furthermore, several version incompatibilities with third-party dependencies arose during the installation of additional software packages using pip. While we were able to ascertain that these issues were operating system-specific, their occurrence underlines the dynamic nature of software dependencies and the importance of controlling compatible versions to ensure a seamless installation process. After installing MIST, a shell script can be used to automatically download publicly available model weights from Zenodo.

MIST training

MIST was retrained on the first structure-disjoint cross-validation split of its publicly available dataset, comprising 6810 training, 401 validation, and 819 test spectra. All model training was performed on an AMD Epyc 7452 Zen2 node, using 12 CPU cores and a single NVIDIA Ampere A100 GPU. Using multiple GPUs was not possible due to serialization issues.

First, we retrained MIST excluding its data augmentation component. MIST trains until convergence on the validation data, with a patience of 30 epochs during which the performance does not improve anymore to stop training. The training duration spanned approximately four hours over 342 epochs, culminating in a combined MIST loss of 0.3204 on the test data.

Next, the forward spectrum prediction model was retrained to incorporate simulated spectra for data augmentation. This task presented considerable challenges due to limited documentation, including implicit command-line parameters, having to examine the source code to determine its exact functioning. Additionally, as the used hyperparameters are not specified, we employed the default hyperparameters. Despite these hurdles, we were able to make the necessary script modifications, including fixing hard-coded file paths and skipping the fingerprint comparison to SIRIUS due to missing documentation. This augmented training process demonstrated a reduction in training time to three hours, converging after 250 epochs, and achieving a slight improvement in test loss to 0.3171.

Retraining the contrastive model required the construction of a decoy database of isomers, derived from PubChem. During this step, we had to address several minor coding errors and manage the extensive computational resources required for processing the PubChem database. After fixing these issues, training durations were approximately 12 and 35 minutes, achieving contrastive losses of 1.9430 and 1.7349 on the test set without and with data augmentation, respectively.

Database retrieval for CASMI 2022

The Critical Assessment of Small Molecule Identification (CASMI) 2022 challenge was an international competition aimed at evaluating and improving the methods used for identifying small molecules from MS data (`https://fiehnlab.ucdavis.edu/casmi`). While only the raw data and the ground truth compound information is available from the official CASMI 2022 website, we retrieved a Pandas[10] dataframe with annotated spectra from Young et al. [9], who matched compound annotations to the corresponding MS/MS spectra. Next, as MIST has currently only been trained on protonated, singly charged spectra, we filtered the CASMI 2022 data for spectra with adduct "[M+H]+", resulting in a dataset of 170 spectra.

Applying MIST to the CASMI 2022 dataset necessitated preprocessing with SIRIUS, a step that was somewhat obfuscated by the lack of explicit documentation and required investigating the code to determine the required data types and directory structure. Note that the recently updated MIST v2.0.0 has simplified this process by eliminating the SIRIUS preprocessing dependency. These challenges underscore that MIST, in its current form, functions more as a research prototype than a fully fledged tool designed for end-user application.

MIST offers two database retrieval methods: fingerprint retrieval and contrastive retrieval. Fingerprint retrieval uses the raw predicted fingerprints to rank all isomers within the PubChem database (version April 2022) that correspond to the chemical formula of the precursor compound, as derived by SIRIUS. This ranking is determined by calculating the cosine distance between the predicted spectral fingerprint and the candidate chemical fingerprints, with the compound having the lowest rank being selected.

Contrastive database retrieval involves projecting both spectral and chemical fingerprints into a joint latent space using contrastive learning. Next, candidates are ranked based on the cosine distance within this latent space, akin to the procedure for fingerprint retrieval. Finally, MIST integrates both the fingerprint distance $d_{fp}(S, M)$ and the contrastive distance $d_c(S, M)$, using hyperparameter $\lambda_r = 0.3$:

$$d(S, M) = \lambda_r d_{fp}(S, M) + (1 - \lambda_r) d_c(S, M),$$

where S represents the spectral fingerprint and M denotes the chemical fingerprint.

Cochran's Q test

To critically assess the impact of distinct components on the MIST performance, we conducted an in-depth ablation study. This investigation entailed the strategic disabling of specific model features—namely, MAGMa substructure annotation, simulated training data, neutral loss featurization, and fingerprint unfolding—to discern their individual contributions to the model's efficacy. Subsequently, both the fingerprint prediction model and the contrastive model were retrained across five iterations, each employing a different random seed to ensure the robustness of our findings.

To quantitatively evaluate the statistical significance of performance differences among the variants of the model, we applied Cochran's Q test, a non-parametric statistical test designed to verify whether multiple treatments have identical effects.[11] This test was performed for both top-5 and top-20 annotation accuracy levels, where a spectrum annotation was deemed correct if the correct compound ranked within the top k predictions. Conversely, annotations falling outside this threshold were classified as incorrect.

The computation of the Cochran's Q test yielded a Q-statistic of 2.963 for the top-5 level, corresponding to a p-value of 0.564, and a Q-statistic of 4.809 for the top-20 level, with a p-value of 0.308. These results indicate the absence of statistically significant differences in model performance across the various configurations tested at both annotation accuracy levels. This outcome suggest that while the features under investigation contribute to the algorithmic complexity of MIST, their individual impact on the model's database retrieval performance does not significantly deviate from the collective performance of the full model, underscoring the need for a careful approach towards optimization of model performance against the computational and conceptual complexity introduced.