LSM1-MS2: A Foundation Model for MS/MS, Encompassing Chemical Property Predictions, Search and *de novo* Generation

*Gabriel Asher[1], Mimoun Cadosch Delmar[1], Jennifer M. Campbell[1], Jack Geremia[1] and Timothy Kassis[1]**

[1]Matterworks Inc., Somerville, MA

{gabriel, mimoun, jenn, jack, timothy}@matterworks.ai

Abstract

We present LSM1-MS2, a pre-trained self-supervised foundation model designed for tandem mass spectrometry (MS/MS) utilizing a transformer architecture with custom tokenization for masked MS2 peak reconstruction. Our model is fine-tuned on smaller, labeled datasets for tasks such as compound property prediction, spectral matching, and *de novo* molecular generation. LSM1-MS2 demonstrates superior performance compared to traditional supervised models, achieving high accuracy with minimal labeled data. It outperforms conventional methods in database lookups and molecular query retrievals and shows promising results in the opening field of *de novo* molecular generation. The model's efficiency in spectral lookup tasks, with significantly reduced evaluation times, underscores its potential for large-scale applications. Our findings highlight the transformative capability of self-supervised pre-training in enhancing the predictive power of models for mass spectrometry, particularly in data-limited scenarios. The success of LSM1-MS2 in property prediction, database spectral lookup, and molecular generation paves the way for its application in metabolomics and drug discovery, facilitating robust and scalable analysis with reduced data requirements.

1) Introduction

Advances in Deep Learning and Artificial Intelligence, such as transformer-based architectures[1,2], pioneered in the Natural Language Processing (NLP)[1,3,4] and Computer Vision (CV)[2,5] fields are increasingly being applied to chemical and biological data[6–9]. This is for good reason. Deep learning strategies promise better predictive models for life science problems by exploiting the highly unstructured information content of biological and chemical data. Due to their complexity, and despite routine collection, we hypothesize that most of the unstructured information content present in chemical and biological data, e.g., mass spectrometry data, is left unexploited by most predictive modeling and insight generation efforts.

One faces immediate challenges adapting deep learning methodologies from NLP and CV to the life sciences. The structure and abundance of genes, transcripts, proteins, and biomolecules (c.f., metabolites,

lipids, drugs, drug metabolites, toxins, bioactives, etc.) are often continuous valued, sparsely non-zero, and span orders of magnitude in dynamic range. These challenges make deep learning on life science data particularly challenging[10]. That said, there has recently been rapid progress in the application of deep learning methods to life sciences data, including for mass spectrometry data[7,11–15].

A more fundamental challenge facing deep learning efforts in the life sciences, however, is that most methods pioneered for NLP and CV problems were borne of environments with abundant labeled data (e.g., text and image data obtained from the web, mining corporate documents, driving data etc.). Many of the AI advances generating current excitement in text and image processing have been trained on millions, tens of millions, or hundreds of millions of labeled data to achieve state-of-the-art performance[3,4,16]. In the life sciences, it is rare to come by high-quality labeled data sets of comparable size. In this context, labeled data means that the data are annotated with all relevant metadata - be it chemical, biological, pharmacological or clinical. Pre-clinical datasets in drug discovery might include a few thousand labeled points, whereas clinical datasets may only house a few hundred. Even where large data archives have been curated, such as gene and metabolite databases, annotation is often sparse and inconsistent. For example, less than 2% of compounds detected in typical high-resolution liquid chromatographic mass spectrometric (LC/MS) and tandem mass spectrometric (LC/MS/MS) metabolomics experiments are readily annotated using available databases[17]. Tandem mass spectrometry data has a specific problem due to lack of precursor specificity in the collected data. Precursor selection in MS/MS is typically 1 Da, while precursor and fragment detection ranges from 1-10 ppm. The unintentional leakage of fragments from poorly selected precursors creates a large challenge in both the curation of MS/MS spectra and the use of databases for identification. This large low-quality data, while not ideal for supervised machine learning techniques, empowers the use of self-supervised models that can be trained at scale then fine-tuned on high quality curated data for a variety of downstream predictive tasks.

Thus, for the life sciences to capitalize on advanced deep learning, there is a fundamental need to reconcile the healthy data appetites of transformer-based architectures with the practical size limitations of real-world datasets. Again, we look to advances made in the NLP and CV communities: self-supervised pre-training of large semantic foundation models using unlabeled data[3,4,18]. While labeled biological datasets are typically small, the aggregation of these datasets across diverse applications and experiments is quite large. For instance, through a combination of internally generated and externally sourced[19] data acquisitions, we have accumulated more than 100 million unlabeled MS and MS/MS spectra, cumulative over a wide variety of underlying applications. These data are abundant, but unlabeled.

It thus remained to be shown that pre-training on these (or subsets thereof) unlabeled datasets yields an advantage for predictive modeling when focused on a specific task using a relatively small volume of task-specific labeled data. Here, we do just that. We provide evidence that self-supervised pre-training of a large semantic model (a.k.a., a foundation model) followed by fine-tuning a task-specific model using only a relatively small, labeled dataset can work. Additionally, we show the potential for this method to yield superior predictive power versus a standalone fully supervised deep learning model trained on a very large, labeled dataset.

As a concrete demonstration, we report the self-supervised training of an LSM1-MS2, an early version of our Large Spectral Model (LSM) built for MS2, representative of the overall foundation models our team is building for primary omics data. We fine-tune the LSM1-MS2 on the specific task of chemical property prediction, and we compare its performance versus recent highly-successful fully-supervised models for chemical property prediction[7,11]. Like the standalone deep learning model results, our fine-tuned predictive model outperforms property prediction obtained by spectral similarity searching of large reference databases, i.e., spectral look-up methods[7,11–15]. Separately, we show that LSM1-MS2 outperforms both a supervised transformer architecture and a heuristic-based method (cosine similarity) in the more conventional spectral lookup tasks with substantially less training data than is required to train a fully supervised model. Finally, we also show potential in using the LSM1-MS2 for *de novo* molecular generation directly from individual MS2 spectra. These tasks are demonstrated on a variety of test datasets to mimic common metabolomics workflows. The first dataset is an '*Unknown*' dataset, which is both spectral disjoint and molecular disjoint. Many molecules in our dataset have more than one representative experimental spectrum. This multiplicity is caused by each spectrum representing a unique collision energy, ionization mode, liquid chromatography method or other unique experimental variable. Also, separate instances of the same molecule may appear in multiple databanks. Evaluation on this dataset is analogous to performing analysis on a completely unseen set of data (i.e., new molecular matter). Our next dataset is a '*Known*' dataset which is spectral disjoint, meaning the experimental instance of the spectra themselves are not in the training dataset, however, it is not molecular disjoint. This Known dataset mimics an experiment where the set of possible molecules analyzed are known but data distributions may be different to the test set. Our third dataset was the '*CASMI 2022'* dataset - used in the field of metabolomics as a challenge dataset[20]. For model development and validation, we also had the *'CASMI 2017'*[21]. We also ensured that none of the CASMI 2022 or 2017 samples were present in the training dataset to prevent leakage (see Methods for full dataset details).

2) Results

2.1) Leveraging Reconstruction-Based Pre-Training for Improved Contextual Learning in Mass Spectrometry

The strength and value of our method lies in our pre-training. Pre-training primes models with contextual information for downstream fine-tuning tasks, empirically improving performance compared to supervised training only methods. We approach our pre-training in a reconstruction-based masked-signal-modeling approach[3,22] described in Figure 1a. For fine-tuning on the property prediction and spectral-lookup tasks we employ the architecture modifications in Figure 1b.

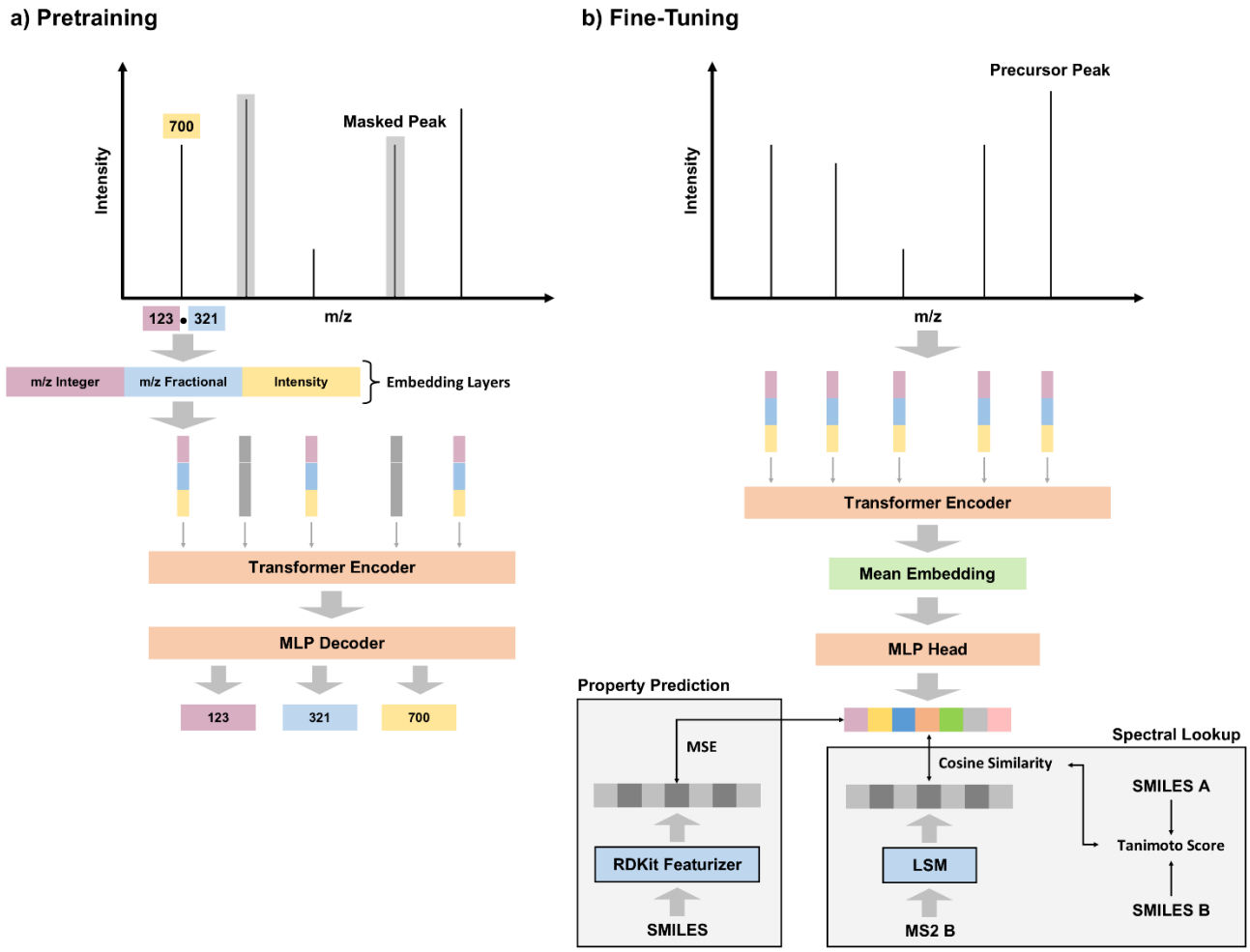

Figure 1: Model Architecture, Pre-Training and Fine-Tuning. a) Our spectrum embedding method. We embed the m/z values using two separate embedding layers, one for each of the integer part (nominal mass) and the fractional (mass defect) part of the value. We use a single embedding layer for binned intensity. All three layers have a vocabulary size of 1,000. We then project the concatenated embeddings through a linear layer before passing them through a transformer encoder. Random peaks are masked by replacing the token with a learnable token shared across all masked peaks. Three separate neural network classification heads are then used to reconstruct the peak's m/z and intensity values. Loss is calculated on all peaks, both masked and unmasked. The precursor peak is never masked. b) We use the mean of all encoded token embeddings for downstream tasks by adding a fully connected MLP head. For property prediction, a set of 209 properties are calculated through a featurizer (RDKit), the LSM1-MS2 is fine-tuned to predict these properties. For spectral lookup, paired spectra are each fed through the LSM, then a projection head to generate a smaller molecular embedding. The model is trained to make the cosine similarity match the Tanimoto similarity of their respective SMILES.

While performance on downstream tasks is difficult to predict strictly based on the pre-training metrics, we show the reconstruction error on both masked and unmasked peaks for the pretrained model on three different test datasets '*Known*', '*Unknown*' and '*CASMI 2022*'. Given that unmasked peaks are inputs to the model, it is expected that reconstruction performance will be good (Figure 2a) compared to that of the masked peaks (Figure 2b). We also observe that the masked peak reconstruction for m/z values has

a highest error in CASMI 2022 compared to both the Known and Unknown test sets. This might partially explain why CASMI 2022 specific tasks did not match the performance of the other two datasets in downstream tasks.

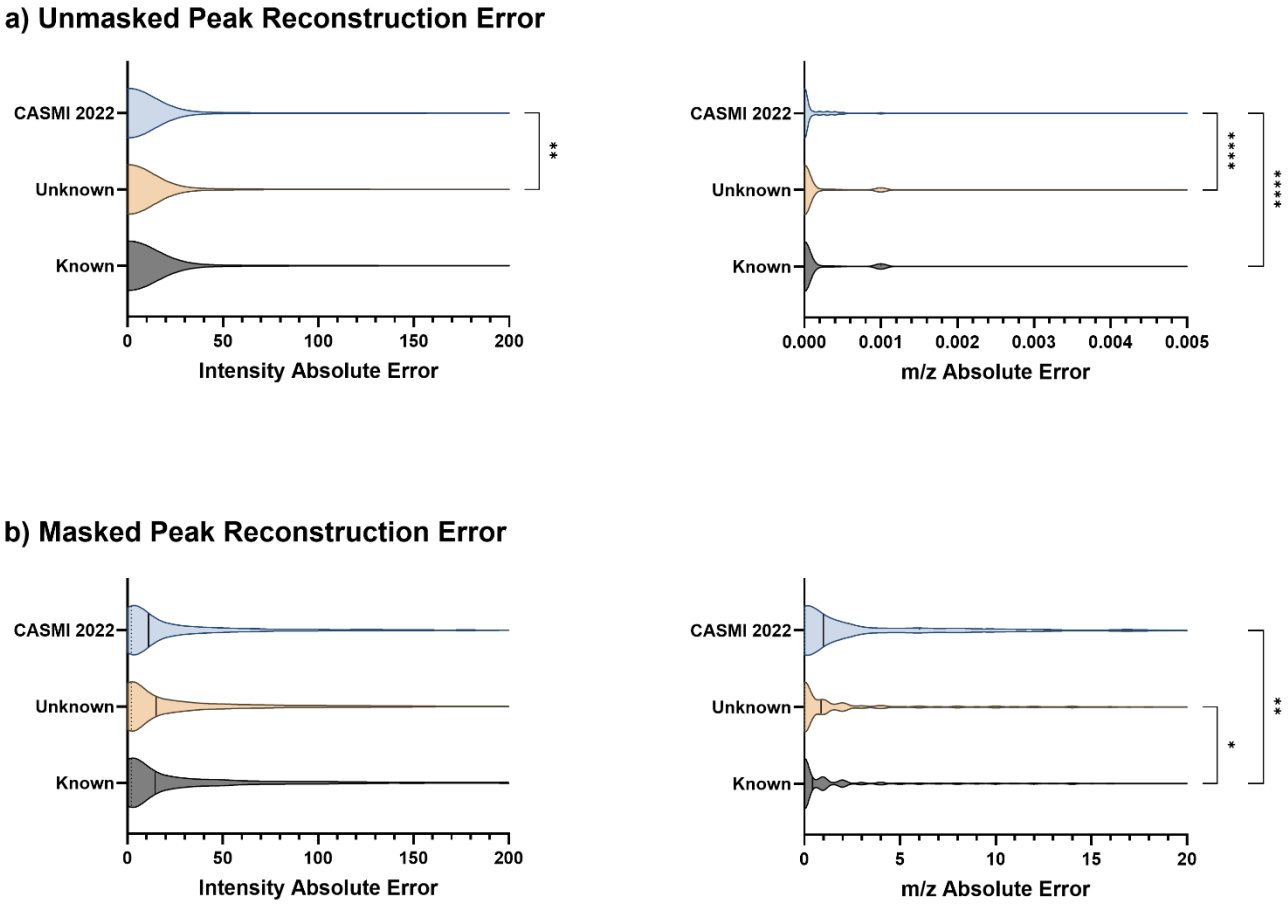

Figure 2: Reconstruction Performance during Pre-Training. a) Absolute reconstruction error for both intensity and m/z values for unmasked peaks. Error differences between the three datasets are not meaningful even though some comparisons show statistical significance. b) Same error calculations across the masked peaks, as expected the errors for masked peaks are higher, with m/z errors being substantially higher than those of unmasked peaks. Additionally, there is a clear difference between the three datasets in terms of m/z reconstruction error where the CASMI dataset is the most challenging. N=5,000 peaks.

2.2) Superior Performance in Property Prediction

For property prediction, we use three baseline methods, a re-implementation of a recently published highly performant supervised transformer-based model MS2Prop[7], a supervised-only version of our LSM1-MS2 model, and modified cosine similarity. In our re-implementation of MS2Prop, we train a model using the hyperparameters and token embedding strategy described in the original paper. For our supervised-only baseline model, we use the same hyperparameters as in our fine-tuned LSM. Finally, to evaluate cosine similarity baseline, we retrieve the most similar spectrum in the training dataset for every query spectrum

in our test dataset. Then, we impute the query's molecular properties from the properties of the retrieved molecule. In our results, we only report best-in-class performance. Thus, given that our MS2Prop and supervised-only LSMs are similar supervised-only transformer architectures, we only report the top performing supervised model representation (MS2Prop).

We broke down the RDKit[23] properties/descriptors into continuous values (if the values for the test dataset ground truth had more than 2 possible values) and binary otherwise (Figure 3). The fine-tuned LSM performs significantly better, on average, across continuous properties than both the fully supervised transformer model and cosine similarity (Figure 3a) with the Known dataset performing significantly better than both the Unknown and CASMI 2022. For the binary properties/descriptors we did not see any major differences across the three approaches, (Figure 3b) most likely since these tend to be structural descriptors and are much easier to predict than continuous values with a large dynamic range. Additionally, there exists a large number of features in the RDKit descriptors that simply are unrelated to anything that an MS/MS signal can be informative of and hence might be impossible to predict. We do not isolate these however and treat all outputs as being equally likely to be predicted through a single MS/MS spectrum in a single mode and single collision energy. Figure 3c shows a single example property (QED: Quantitative Estimation of Drug-Likeness) compared across the three datasets and three approaches (for all 209 outputs see Supplementary Information).

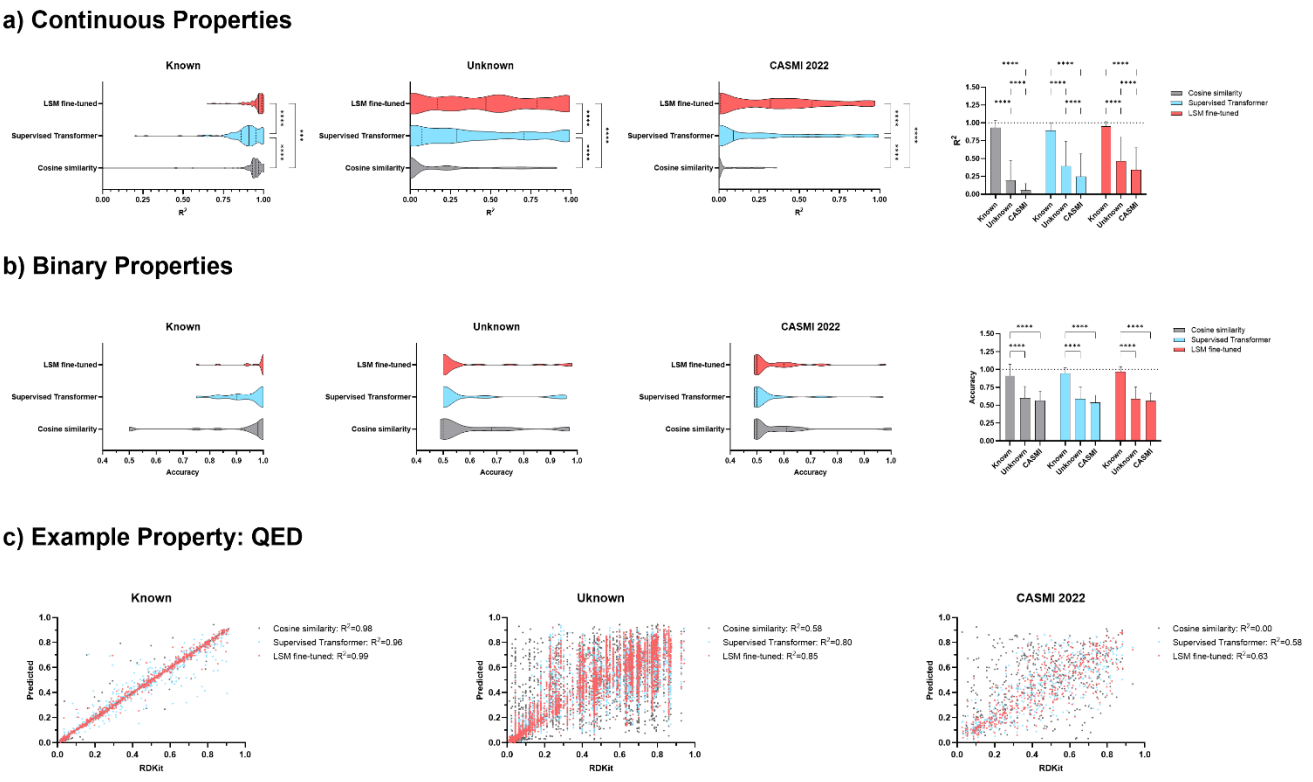

Figure 3: Property Prediction Task. a) Prediction performance for continuous RDKit properties. In all cases, the LSM fine-tuned model outperforms both cosine similarity and the fully supervised transformer model. For all three approaches average prediction across all properties is substantially better for the Known dataset and drops of for the Unknown and CASMI 2022 datasets. b) For binary properties we do not see any difference between the three approaches. c) An example property shown

(QED: Quantitative Estimation of Drug-Likeness). The same performance trends apply across all properties (see Supplementary Information) where R^2 values are higher for the LSM vs other approaches even on a challenging dataset such as CASMI. N=~160 for continuous properties and ~49 for binary properties (but both vary depending on the values in the test set). Error bars represent standard deviations.

2.3) Efficient and Accurate Database Spectral Lookup Outperforming Conventional Methods in Speed and Scalability

We evaluate database spectral lookup task through the creation of a spectral database using LSM1-MS2 embeddings generated with our training set. For each spectrum in our evaluation datasets, we retrieve the most similar spectrum in our training dataset as follows: for each query spectrum's LSM1-MS2 molecular embedding, we find the cosine similarity of this embedding to all the molecular embeddings in the training set. We then narrow down our training set search space to a threshold of the query's precursor m/z, since the precursor mass is always known for data dependent acquisition of MS/MS spectra. Finally, we return the molecule in the precursor-filtered training dataset with the highest cosine similarity to the query.

For the Known and Unknown test datasets the LSM fine-tuned model outperforms conventional cosine similarity. However, for the challenging CASMI 2022 dataset, cosine similarity outperforms the LSM (Figure 4). Finally, it is worth mentioning that although modified cosine had comparable results for many of our metrics on this task, we believe that LSM1-MS2 still is the superior method due to its speed. For the spectral lookup task, our modified cosine similarity took roughly 27.4 seconds per sample, with a spectral database of size 742,049, compared to 2 milliseconds per sample using LSM1-MS2 (~13,700x speedup). Thus, this significant increase in speed makes LSM1-MS2 much more scalable, agile, and useful for large quantities of data - and demonstrates its potential for use in on-the-fly decision making during spectral acquisition. Finally, given the rise of sophisticated vector search algorithms and databases being used with Large Language Models (LLMs) for Retrieval Augmented Generation (RAG)[24–26], the LSM1-MS2 embeddings can easily be utilized in any of these frameworks for very advanced queries and substantially faster lookups at immense scale (billions of spectra if needed).

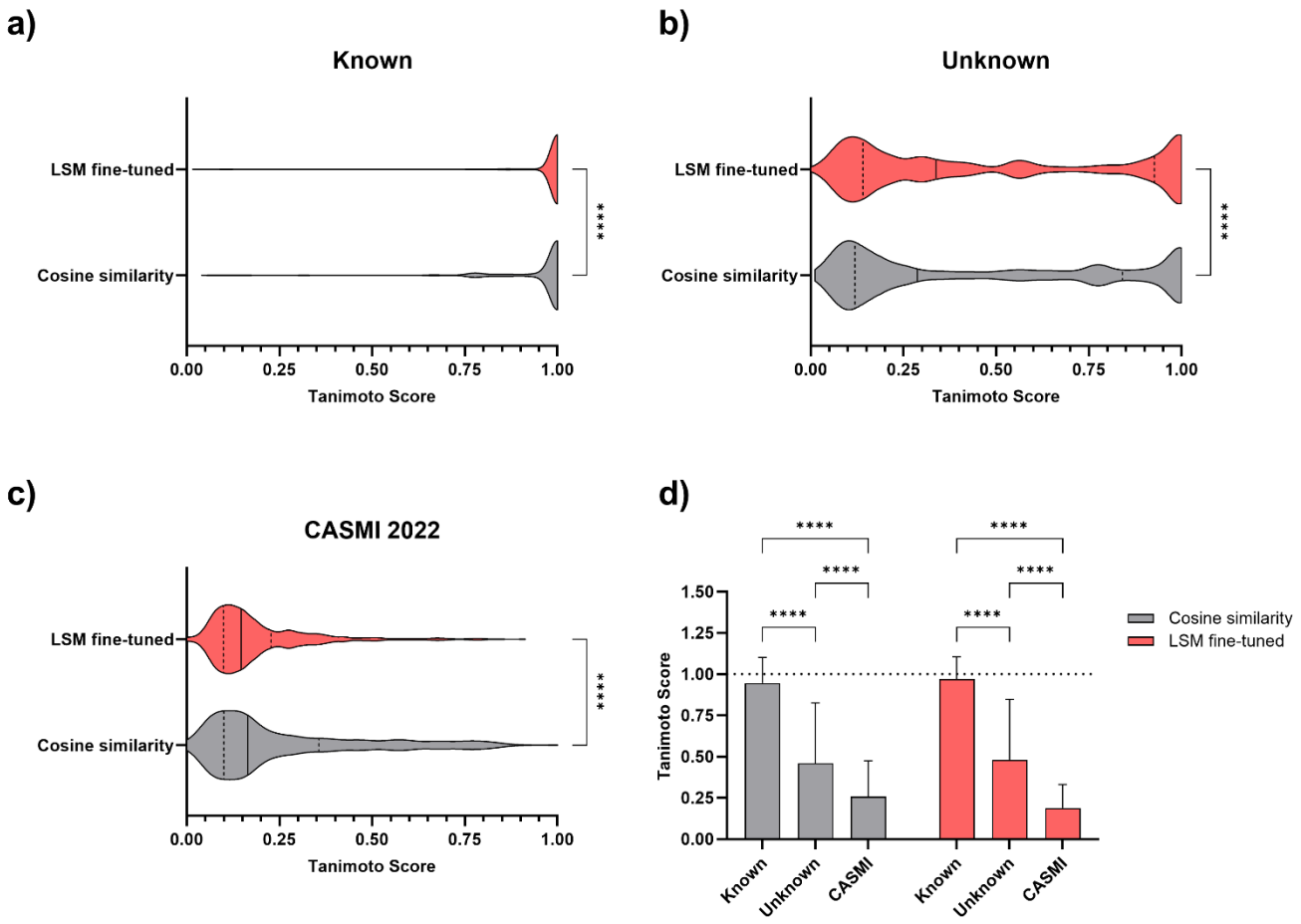

Figure 4: Database Spectral Lookup. LSM generally performs better than cosine similarity for a) Known and b) Unknown datasets but slightly worse for c) CASMI 2022. d) Tanimoto score predictions drop as the dataset becomes more out of distribution (Known -> Unknown -> CASMI). N=1,000 for Known, N=12,274 for Unknown and N=464 for CASMI. Error bars represent standard deviations.

2.4) Experimentally Useful *De Novo* Molecular Generation

The holy grail of Mass Spectrometry based identification is the ability to directly infer molecular identities from the "dark" metabolic space. To this end, we've adapted the LSM1-MS2 to perform *de novo* molecular generation. The process begins by converting molecules from SMILES to SELFIES[27], using SELFIES tokens as the vocabulary for subsequent models. We then train a BERT-style encoder-decoder model[3] to predict masked SELFIES tokens (Figure 5a). With a fixed-weight BERT encoder, we proceed to train a conditional GPT-2 decoder. This decoder generates initial predictions based on context embeddings from the BERT encoder and learns in an autoregressive manner (Figure 5b). Next, we align the LSM1-MS2 embeddings with the BERT encoder embeddings and feed the LSM1-MS2 embeddings into the pre-trained GPT-2 decoder as context embeddings (Figure 5c). This allows the model to learn autoregressively.

Ultimately, this training enables the LSM1-MS2 and GPT-2 decoder to predict *de novo* molecular identities from LSM1-MS2 spectral embeddings during inference.

While ideally, molecular generative capabilities would perform optimally with single-shot generations, they still hold tremendous value in experimental settings by limiting the space of possibilities and informing meaningful downstream validation experiments. To evaluate generative performance, we use beam search to generate 100 SMILES representations and then re-rank them based on simple precursor mass (see methods). We consider the top 1, top 10, and top 100 candidates. We assess the quality of the generations by calculating the maximum Tanimoto score achieved in each setting with respect to the ground truth query molecule (Figure 5d). For in-distribution data exemplified by the Unknown dataset, we achieve a mean maximum Tanimoto score of 0.63 for the top 100 predictions and 0.48 for the top 1 prediction. For the more challenging CASMI 2017 dataset, we achieve a score of 0.46, and for CASMI 2022, a score of 0.38 for the top 100. While the score requirements vary by application, a Tanimoto score of greater than 0.5 is generally considered useful. In experiments focused on automatic synthesis, there is a singular question of synthetic success. In these experiments, both the presence and absence of substructures are of high value, as they can significantly influence the synthetic feasibility and overall success rate of the generated molecules.

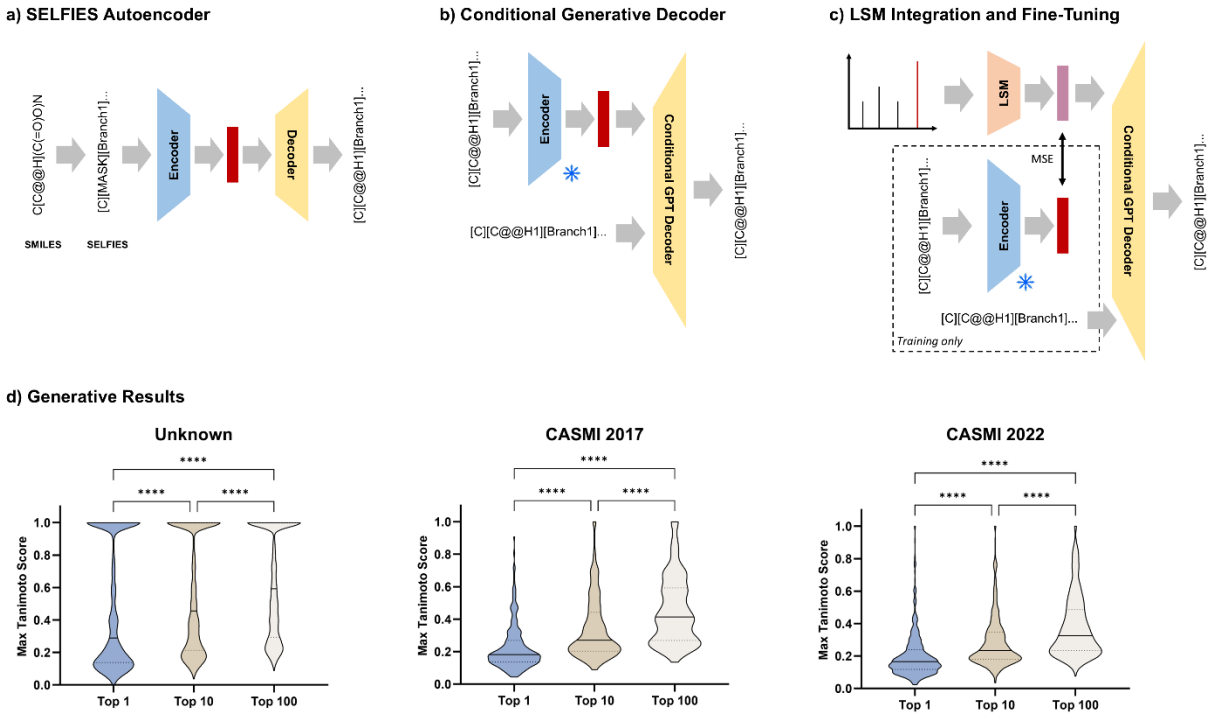

Figure 5: De Novo Generation. a) An autoencoder is trained on SELFIES representations using masked language modeling. b) A GPT-like decoder model is trained to produce a SELFIES representation based on context embeddings from the autoencoder. c) During training we finetune the GPT-like decoder model using MS2-LSM embeddings, which we simultaneously align with the corresponding masked language model embeddings, as our context embeddings. During inference, we generate de novo SELFIES representations using the MS2-LSM embeddings as context embeddings. d) Maximum Tanimoto score achieved for a given dataset based on 1, 10 and 100 de novo generations. Higher scores can be achieved with more generations. N=12,274 for Unknown, N=243 for CASMI 2017 and N=464 for CASMI 2022.

The development of a robust method for predicting molecular structures from mass spectra acknowledges that the correct structure may not be obtained from a single point prediction, but rather from an ensemble approach where the user triangulates the correct structure through commonalities across the top N predicted structures. For this purpose, we have employed a consensus scoring method to evaluate the reliability of structural predictions. The consensus score for a given substructure is defined as the fraction of generated predictions that include the substructure, providing a normalized measure from 0 (no agreement) to 1 (unanimous agreement). A high consensus score must indicate a high probability that the given substructure is present, and a low consensus score must indicate the reverse. For all three datasets, the 100 top generated predictions have a high consensus score when the substructure is present and a low score when not present indicating the ability to predict substructures reliably (Figure 6a). A breakdown of the 18 curated substructures shows that certain substructures are much easier to predict than others (Figure 6b) in terms of both true positive predictions (high consensus when present) and false positive predictions (high consensus when not present). For example, c1ccccc1 tends to be reliably identified when present, but [CX3](=O)[NX3H2] is challenging. Additionally, another example is something like [OX2H] which tends to be falsely identified as present when not.

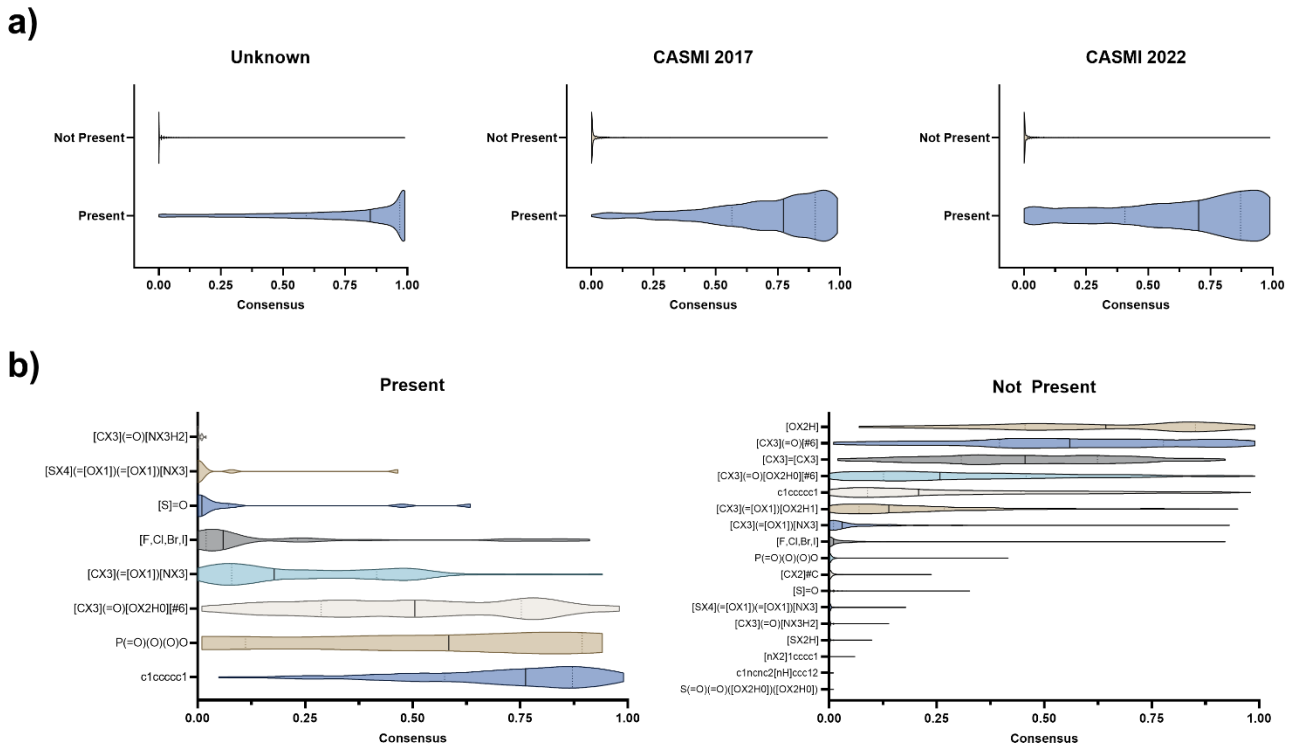

Figure 6: Substructure Consensus. a) The consensus score for each of the datasets representing what percent of the 100 predicted structures contained the pre-determined sub-structure in the ground truth molecule. If a substructure is present there is a high consensus score agreeing of its presence in all dataset, conversely, if a substructure is not present there is a low consensus score (of it being not present) in all datasets. Results include all 18 pre-determined substructures. b) A breakdown of the 18 substructures shown. Only substructures that have at least 5 occurrences in the test data are shown. N=12,274 for

Unknown, N=243 for CASMI 2017 and N=464 for CASMI 2022. The number of occurrences of a certain substructure vary between 5 and 216.

2.5) Pre-Training Allows Smaller Datasets for Fine-Tuning

One of the key benefits of our self-supervised approach is that it requires significantly less fine-tuning data for MS2-based downstream tasks. We show this benefit where we evaluate dataset size against downstream performance on our three datasets for a property prediction task on continuous properties (Figure 7). We fine-tune our pre-trained LSM1-MS2 using 1, 5, 10, 25, and 50% of our full annotated training data (7,420, 37,102, 74,204, 185,512 and 371,024 spectra respectively). For the Known dataset we match the performance of all other approaches by fine-tuning on as little as 50% of their training data. For the Unknown dataset we surpass cosine performance with as little as 1% and the supervised models with as little as 10%. For the challenging CASMI 2022 dataset we match the supervised transformer model with only 5% of the training data and our LSM supervised-only version with slightly less than 50% of the training data. Additionally, we train a single linear layer on top of the fixed embeddings to better understand the representation being captured, while end-to-end fine-tuning in all cases offers substantially better performance, if a limited dataset is being used (<7,420) then fixed embeddings give equal performance and would be preferred given the reduced compute requirements and reduced overfitting potential.

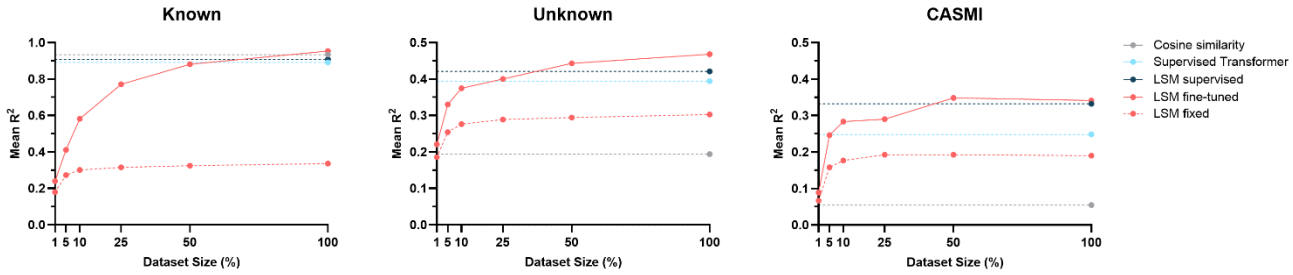

Figure 7: Performance with Dataset Size. LSM1-MS2 achieved similar or superior performance to fully supervised state of the art models with substantially less data. For both Known and Unknown datasets, 50% of the fine-tuning dataset gives equal performance to all other approaches. For the challenging CASMI 2022 dataset, our approach matches the previously published supervised transformer model with as little as 5% of the fine-tuning data. LSM supervised, LSM fine-tuned and LSM fixed are all based on the same architecture. The supervised variant is not pre-trained, and fixed variant is not fine-tuned.

3) Discussion

The results of this study underscore the transformative potential of the LSM1-MS2 model in tandem mass spectrometry. By leveraging a self-supervised pre-training approach, LSM1-MS2 significantly enhances performance in downstream applications such as spectral lookup, enables broader use of property prediction, and paves the way for *de novo* molecular generation.

One of the primary strengths of LSM1-MS2 lies in its robust pre-training methodology. The reconstruction-based masked-signal-modeling technique used for pre-training enabled the model to generalize well across different datasets, including the Known, Unknown, and CASMI datasets. This foundational training provided a substantial boost to the model's performance in fine-tuning tasks, demonstrating the effectiveness of this approach in handling the complexity and diversity of mass spectrometry data.

The results indicate significant improvements in predicting continuous RDKit properties compared to baseline methods such as cosine similarity and state of the art deep learning models such as MS2Prop. The fine-tuned LSM1-MS2 model outperformed these methods across all datasets, particularly excelling in the Known dataset while maintaining robust performance in the Unknown and CASMI datasets. This underscores the model's ability to capture meaningful chemical information that generalizes well across different data distributions, a critical feature for practical applications in metabolomics and drug discovery.

A key advantage of LSM1-MS2 is its efficiency in spectral lookup tasks. The fine-tuned model not only outperformed traditional cosine similarity methods in terms of accuracy but also achieved substantial speed improvements. Reducing the evaluation time by several orders of magnitude makes LSM1-MS2 highly scalable and suitable for real-time applications. This efficiency is crucial for large-scale spectral databases and high-throughput screening, where quick and accurate identification of compounds is essential.

The capability of LSM1-MS2 for *de novo* molecular generation further highlights its potential. By leveraging a BERT-style encoder-decoder architecture and fine-tuning on MS/MS spectra, the model can generate novel molecular structures with relatively high accuracy in both the case of predicting the presence of a certain substructure and the absence of. This ability is particularly valuable for exploring the "dark" chemical space, where traditional methods often fail to identify unknown compounds. Our evaluation shows that LSM1-MS2 performs well in both in-distribution and somewhat in out-of-distribution settings, making it a powerful tool for discovering new molecules.

Despite its strengths, LSM1-MS2 has several limitations that warrant further research. The current pre-training dataset, although large, is still a fraction of the size used in large language models (LLMs), suggesting that performance could improve with more extensive pre-training. Additionally, incorporating more detailed experimental parameters such as ionization modes, adduct charge states, and collision energies could enhance the model's accuracy. Exploring alternative spectrum embedding strategies, including multiple input spectra, and increasing sequence length are other promising avenues for future work. The use of ALiBi[28] positional encoding should facilitate these improvements, allowing for longer sequences without a significant computational overhead. With regards to using such a model, given that it is capable of three tasks (property prediction, spectral look-up and *de novo* molecular generation) one can envision using all three jointly to improve identification-based applications. For example, one can condition the *de novo* generation on the predicted properties and/or closest spectral match. This we believe would significantly improve the generative performance, but we leave that for future work.

The broader implications of adopting pre-trained foundation models in mass spectrometry are profound. The LSM1-MS2 model sets the stage for broader adoption of pre-trained models, enabling high-performance machine learning applications even with limited labeled data. Future research should focus on scaling the pre-training datasets, optimizing hyperparameters, and integrating additional experimental data to further enhance the model's capabilities. Potential applications extend beyond those explored in this study, such as on-the-fly decision-making during spectral acquisition and integration with sophisticated vector search algorithms for advanced queries.

In conclusion, the LSM1-MS2 model represents a significant advancement in the application of machine learning to mass spectrometry. By demonstrating strong results in spectral lookup, property prediction, and *de novo* molecular generation, LSM1-MS2 exemplifies the potential of pre-trained foundation models in this field. Future work should continue to build on these foundations, expanding the scope and performance of these models to fully realize their potential in diverse applications within mass spectrometry and beyond.

4) Methods

4.1) Dataset Preparation

For fine-tuning, we created a labeled dataset consisting of 1,282,758 spectra from MassBank of North America[29] and CompMS MS-DIAL[30,31]. Filtering to exclude any spectra missing sufficient identity information and to exclude any spectra whose precursor molecular mass was above 1,000 m/z resulted in a total of 790,713 labeled spectra. From the molecular identity of these labeled spectra, we further annotated the 790,713 spectra with 209 chemical property descriptors obtained via the RDKit package[23]. After this initial dataset preparation, we curated our datasets similarly to the procedures described MS2Prop[7] and MS2DeepScore[12]. Additionally, a large, diverse, unlabeled, dataset was used for pre-training.

For evaluation, we created three separate datasets to mimic common metabolomics workflows. The first dataset is an '*Unknown*' dataset, which is both spectral disjoint and molecular disjoint - meaning all molecules (and hence any relevant spectral data) are not present in any form in the training dataset. This Unknown dataset includes 12,274 spectra of 2,026 molecules. Many molecules in our dataset have more than one representative experimental spectrum. This multiplicity is caused by each spectrum representing a unique collision energy, ionization mode, liquid chromatography method or other unique experimental variable. Also, separate instances of the same molecule may appear in multiple databanks. We ensure molecule disjointedness for our Unknown dataset by splitting on the first 14 keys of the InChiKey[32]. Evaluation on this dataset is analogous to performing analysis on a completely unseen set of data (i.e., new molecular matter). Our next dataset is a '*Known*' dataset, which consists of 1,000 spectra randomly sampled from our dataset. This dataset is spectral disjoint, meaning the experimental instance of the spectra themselves are not in the training dataset, however, it is not molecular disjoint - meaning other experimental instances of the same molecules are. This Known dataset mimics an experiment where the set of possible molecules analyzed are known but data distributions may be different to the test set. Our

third dataset was the '*CASMI 2022* dataset' - used in the field of metabolomics as a challenge dataset[20]. We were able to extract 464 of the 500 CASMI 2022 molecules. For generative tasks we also used the CASMI 2017[21] dataset created in the same manner. While pre-processing, we also ensured that none of the CASMI 2022 or 2017 datasets were present in the training dataset to prevent leakage.

4.2) Spectrum Tokenization

A key component of our model is the tokenization strategy for MS data. Given the continuous and large range of m/z and intensity of peaks in mass spectrometry, common tokenization strategies in NLP and CV struggle to properly encode MS spectra. Prior papers approach the tokenization of MS spectra through binning[12], sinusoidal position embedding[11], or a codebook for integer (nominal mass) and decimal (mass defect) parts of m/z and intensity values[8] . We choose to adopt the last of these tokenization strategies (with some slight modifications). We first sort spectra by m/z and prune fragment peaks over 1,000 m/z. Then for each peak in a spectrum, we embed the integer and decimal parts of each m/z peak separately using a learnable codebook. Furthermore, we use a codebook to embed the intensity values of the peaks, which are scaled to a maximum of 1,000. Finally, for each peak we concatenate the integer m/z embedding, decimal m/z embedding, and intensity embedding, then pass this concatenated embedding vector through a single linear layer to create a peak token embedding. Each spectrum is represented by a sequence of these tokens corresponding to its peaks. Furthermore, we prepend a special precursor token to the beginning of every spectrum token sequence. We generate the precursor token similarly to other peak tokens. However, all precursors have a precursor-unique preset intensity value of 2,000.

4.3) Pre-Training

We tokenize our input spectra and if any of the spectra have fewer than 64 peaks, we also pad the sequence with 0 values to reach 64 (the padded values are not attended to during training). Then, we randomly mask 25% of peaks with a learnable mask token. We pass this partially masked token sequence through transformer layers. Finally, we pass each token in the transformer output through three heads, which predict an integer m/z, decimal m/z, and intensity value. The cross-entropy loss between the predicted and ground-truth for all non-pad peak integer m/z, decimal m/z, and intensity values are aggregated into our final pre-training loss function:

$$L_{final} = \lambda \cdot L_{mzI} + \beta \cdot L_{mzD} + \gamma \cdot L_{Int}$$

L_{mzI}, L_{mzD}, and L_{Int} are the integer m/z, decimal m/z, and intensity losses respectively, furthermore λ, β and γ are weights applied to the parts of our loss function. We set $\lambda = 100$, $\beta = 1$ and $\gamma = 1$ respectively. Finally, for all fine-tuning tasks, we use the pre-training model checkpoint with the lowest total loss, L_{final}, on our unseen molecule validation dataset. Model hyperparameters were as follows:

Parameter	Selected Value
max_input_peaks	64
learning_rate	$1x10^{-6}$
batch_size	448
d_model	1,024
encoder_layers	16

encoder_attn_heads	16
mask_pct	0.25
alpha	100
beta	1.0
omega	1.0

4.4) Fine-Tuning

4.4.1) Chemical Properties Prediction

To fine-tune our model for property prediction given a SMILES-labeled spectrum, we pass this spectrum through our pre-trained LSM1-MS2 to generate an output spectrum embedding. The mean of all these token values (including the precursor) is computed, and the resultant embedding is passed through a single linear-layer classification head, which outputs a vector size of 209, which represents the number of properties predicted. These predicted properties are then compared to the ground-truth properties vector computed by RDKit on the SMILES identifier of the MS/MS spectrum. Each property is normalized between 0 and 1 to ensure that our model does not overfit certain properties. The loss for this fine-tuning task is calculated via the Mean Square Error (MSE) of these two resultant vectors.

For property prediction, we use three baseline methods, a re-implementation of MS2Prop, a supervised-only version of our LSM1-MS2 model, and modified cosine similarity. In our re-implementation of MS2Prop, we train a transformer model using the hyperparameters and token embedding strategy described in their paper[7]. Namely, our transformer backbone uses 32 heads, 6 layers, a hidden dimension of 512, and no positional encoding. Furthermore, we embed MS/MS peaks into tokens as follows. We first round peak m/z values to the nearest 0.1, then we feed these values through a learnable lookup table, then we concatenate the intensity value (normalized to 1.0 for non-precursor tokens and 2.0 for precursor tokens) to the m/z embedding, then we finally pass this concatenated token vector through a linear layer of depth 1 to project it to the hidden dimension. We also use the first token in the sequence as a classification token.[7] It is worth noting that this re-implementation of MS2Prop has 3 key differences from the original implementation: the training dataset, training hyperparameters (batch size, learning rate, number of epochs), and the number of outputs (we predict on 209 properties instead of 10). Given that the MS2Prop paper does not indicate what learning rate or number of epochs is used, we use a learning rate of 0.00025 for 50 epochs. This learning rate was selected after performing a learning rate grid-search to minimize Mean Absolute Error (MAE) in our unknown validation dataset. For our supervised-only baseline model, we use the same hyperparameters as in our fine-tuned LSM. Finally, to evaluate cosine similarity baseline, we retrieve the most similar spectrum in the training dataset for every query spectrum in our test dataset. Then, we impute the query's molecular properties with the properties of the retrieved molecule. In our results, we only report best-in-class performance. Thus, given that our MS2Prop and supervised-only LSMs are similar supervised-only transformer architectures, we only report the top performing representation (MS2Prop).

4.4.2) Database Spectral Lookup

For the spectral lookup task our inputs are two SMILES-labeled spectra. We pass each of these spectra through the LSM, generating two spectrum embeddings. Then we take the mean of all the tokens and pass the resulting vector through a single linear-layer head, which creates a dimensionally smaller molecular embedding. Finally, we calculate the cosine similarity of the two molecular embeddings. We calculate the loss of this fine-tuning task by taking the mean-square error of the predicted cosine similarity versus the ground-truth Tanimoto similarity. Ground-truth similarity is calculated using RDKit with 2048-bit Morgan fingerprints and a maximum atomic radius of 2.

DL-based library spectral lookup techniques involve the comparison of two annotated spectra from the dataset. These spectra are embedded into a latent space, and are trained to minimize the Tanimoto similarity of the respective compounds they represent[12–14]. From here, one can perform database retrieval on the most similar embedding to a given query molecule. Most current approaches to MS/MS analysis rely on spectral lookup within annotated MS/MS spectra databases. Query spectra are matched against annotated spectra in databases using heuristic algorithms to find close molecular matches, which then serve to identify the query molecules. One such heuristic algorithm is known as modified cosine similarity. Modified cosine similarity aligns the peaks of a query spectrum with the peaks of reference spectra based on a Dalton value threshold. Subsequently, the cosine similarity of the aligned peaks' intensities is calculated, with the most similar reference spectrum, and thus the identified molecule, being returned as the output[33,34] . We use the Matchms[35,36] modified cosine to evaluate modified cosine similarity.

4.4.3) *De Novo* Molecular Generation

Firstly, we convert our molecules from SMILES into SELFIES representations[27,37], and use SELFIES tokens as our vocabulary for subsequent models. Then, we train a BERT-style encoder-decoder model which takes in these SELFIES tokens and learns to predict masked tokens. Using a fixed-weight version of this encoder, we then train a conditional GPT-2 decoder[38]. This decoder learns auto-regressively using the BERT embedding as context. The BERT embedding is analogous to the context created from a prompt in an NLP setting. This decoder is adapted from an open-source implementation[39]. We then learn to align the LSM1-MS2 embeddings with the BERT encoder embeddings and feed the LSM1-MS2 embeddings into our pre-trained decoder as context embeddings to learn auto-regressively from. Thus, we train the LSM1-MS2 and GPT decoder[4,38] to predict *de novo* molecular identities from just an LSM1-MS2 embedding of a spectrum. Finally, given that our decoder can generate multiple different molecules for a single sample, we leverage a heuristic re-ranker for evaluating performance at different number of generations. The reranking of generated molecules is achieved by minimizing the absolute difference between the mass of the generated molecule M_g and the precursor's mass with adduct adjustments ΔM_{adduct}. For a given adduct, the equation for the mass difference:

$$\Delta M = min_{adducts}\left|M_g - (M_p + \Delta M_{adduct})\right|$$

where M_p is the precursor mass from the mass spectrometry data and ΔM_{adduct} is the mass adjustment for a specific adduct, which may involve addition or subtraction of the adduct's mass. The possible adducts

are: $[M]^+$, $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, $[M+NH4]^+$, $[M+2H]^{2+}$, $[M+H+Na]^{2+}$, $[M+2Na]^{2+}$, $[M-H]^-$, $[M+Cl]^-$, $[M+FA]^-$,$[M+Br]^-$.

The generated molecule with the smallest ΔM is considered the best match to the precursor ion. We use supplemental data to train the BERT and conditional GPT models. We use subset of 100M molecules[40] to train these models. We convert this subset from SMILES to SELFIES representations, then train the model as described above. Furthermore, we evaluate this model on three datasets of unknown molecules, both in and out-of-distribution settings. In the in-distribution evaluation, our *de novo* molecular generation model is tasked with identifying molecules from our Unknown dataset. The challenge presented by the Unknown dataset is comparable to inferring the structures of novel molecules based on spectral signatures obtained under experimental parameters previously encountered during the LSM's fine-tuning phase. This setup simulates a practical scenario in which a model is applied to identify new molecules using an LSM1-MS2 version which has been fine-tuned on a spectral library created under familiar experimental conditions. Furthermore, we also use two out-of-distribution evaluation datasets, the CASMI 2017 and 2022 challenge sets. The distribution and molecules in these datasets do not exist in our training data. Thus, this setup simulates a practical scenario where *de novo* molecular generation is inferred from completely novel experimental settings. For validation during training, we use the unknown validation dataset as described in prior parts of this paper. Furthermore, we report performance results for n=1, 10, and 100 generations per spectrum.

Substructure consensus metrics were calculated based on a compiled list of 18 substructures. These substructures were selected by an MS expert user for their plausible observability via standard fragmentation mechanisms in mass spectrometry. The substructures are specified as SMARTS[41], a language designed to describe molecular substructures and run substructure queries. Not meant to be exhaustive, this list is to be adapted to a user's needs based on substructures they expect to find that will help them prioritize between structural predictions. For example, a user can prioritize structural predictions containing the [OH]c1ccccc1 substructure if they expect the structure to contain a phenol group. Next, we generated 100 predictions using the LSM-MS2 structure generation task for every structure in the CASMI 2017, CASMI 2022, and the Unknown datasets. Using the RDKit HasSubstructureMatch function, which queries the presence of a particular substructure in a given molecule, we determined whether each of the substructures is present in the set of 100 predictions generated for every structure.

4.5) Statistical Analysis

Figure 2: N=5,000 randomly sampled peaks from each dataset are presented in both masked and unmasked scenarios. A non-parametric one-way ANOVA (Kruskal-Wallis test) with Dunn's multiple comparisons correction is carried out to compare each dataset against every other dataset.

Figure 3: A set of 209 properties are considered, if for a given test dataset there are <=2 unique possible values it is treated as a binary task and balanced accuracy is used as the metric. If a property has >2 unique possible values then it is treated as a continuous output and R^2 (coefficient of determination) is used as a metric. N~=160 for continuous properties and N=~49 for binary properties (both vary by dataset). A non-parametric one-way ANOVA (Friedman test) with Dunn's multiple comparisons

correction is carried out to compare each approach (LSM fine-tuned, Supervised transformer and Cosine similarity) to every approach. For the summary figures, comparing performance across datasets (Known, Unknown and CASMI 2022), a two-way ANOVA with Tukey's multiple comparisons correction is carried out.

Figure 4: The sample size (number of unique MS2 spectra) varied by dataset (Known=1,000, Unknown=12,274 and CASMI 2022=464). For all three datasets a paired non-parametric t-test (Wilcoxon test) was used to compare LSM vs Cosine for each dataset. For the comparison figure, comparing performance across datasets (Known, Unknown and CASMI), a two-way ANOVA with Tukey's multiple comparisons correction is carried out.

Figure 5: The sample size (number of unique MS2 spectra) varied by dataset (Unknown=12,274, CASMI 2017=243 and CASMI 2022=464). For all three datasets a non-paired non-parametric one-way ANOVA (Kruskal-Wallis test) with Dunn's multiple comparisons correction is carried out to compare each of the number of top rankings to each other.

Figure 6: For the Unknown dataset N=44,075 for 'present' and 164,583 for 'not present'. For CASMI 2017 N=806 and 3,325 respectively, and for CASMI 2022 N=1,662 and 6,226 respectively. The number of substructures being assessed varied where there were as little none present in the dataset for a particular substructure to as much as 216 for c1ccccc1.

In all cases if $p>0.05$ (NS) no asterisk or relation line is shown, otherwise: p=0.01-0.05 (*), p=0.001-0.01 (**), p=0.0001-0.001 (***) and $p<0.0001$ (****). The solid black bar in the middle of the violin plot represents the median.

5) Acknowledgements

The authors would like to thank Aubrey Brueckner for thoroughly editing the manuscript.

References

1. Vaswani, A. *et al.* Attention Is All You Need. *arXiv [cs.CL]* (2017).

2. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv [cs.CV]* (2020).

3. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [cs.CL]* (2018).

4. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *arXiv [cs.CL]* (2020).

5. He, K. *et al.* Masked Autoencoders Are Scalable Vision Learners. *arXiv [cs.CV]* (2021).

6. Cui, H. *et al.* scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI. *bioRxiv* 2023.04.30.538439 (2023) doi:10.1101/2023.04.30.538439.

7. Voronov, G. *et al.* MS2Prop: A machine learning model that directly predicts chemical properties from mass spectrometry data for novel compounds. *bioRxiv* 2022.10.09.511482 (2022) doi:10.1101/2022.10.09.511482.

8. Butler, T. *et al.* MS2Mol: A transformer model for illuminating dark chemical space from mass spectra. *ChemRxiv* (2023) doi:10.26434/chemrxiv-2023-vsmpx-v2.

9. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).

10. Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform.* 19, 1236–1246 (2018).

11. Voronov, G. *et al.* Multi-scale Sinusoidal Embeddings Enable Learning on High Resolution Mass Spectrometry Data. *arXiv [cs.LG]* (2022).

12. Huber, F., van der Burg, S., van der Hooft, J. J. J. & Ridder, L. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *J. Cheminform.* 13, 84 (2021).

13. Guo, H., Xue, K., Sun, H., Jiang, W. & Pu, S. Contrastive Learning-Based Embedder for the Representation of Tandem Mass Spectra. *Anal. Chem.* 95, 7888–7896 (2023).

14. Goldman, S. *et al.* Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence* 5, 965–979 (2023).

15. Bushuiev, R., Bushuiev, A., Samusevich, R., Šivic, J. & Pluskal, T. Emergence of molecular structures from self-supervised learning on mass spectra. *ChemRxiv* (2023) doi:10.26434/chemrxiv-2023-kss3r.

16. Zhai, X., Kolesnikov, A., Houlsby, N. & Beyer, L. Scaling Vision Transformers. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 1204–1213 (IEEE, 2022).

17. da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences of the United States of America* vol. 112 12549–12550 (2015).

18. Touvron, H. *et al.* Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv [cs.CL]* (2023).

19. Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* 44, D463-70 (2016).

20. Critical Assessment of Small Molecule Identification (CASMI) - 2022. *http://www.casmi-contest.org/2022/index.shtml* http://www.casmi-contest.org/2022/index.shtml.

21. Critical Assessment of Small Molecule Identification (CASMI) - 2017. *http://www.casmi-contest.org/2017/index.shtml* http://www.casmi-contest.org/2017/index.shtml.

22. Chen, X. & He, K. Exploring simple Siamese representation learning. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 15745–15753 (2020).

23. *Rdkit: The Official Sources for the RDKit Library*. (Github).

24. Jing, Z. *et al.* When Large Language Models Meet Vector Databases: A Survey. *arXiv [cs.DB]* (2024).

25. Kukreja, S. *et al.* Vector Databases and Vector Embeddings-Review. in *2023 International Workshop on Artificial Intelligence and Image Processing (IWAIIP)* 231–236 (IEEE, 2023).

26. Xie, X., Liu, H., Hou, W. & Huang, H. A Brief Survey of Vector Databases. in *2023 9th International Conference on Big Data and Information Analytics (BigDIA)* 364–371 (IEEE, 2023).

27. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation. *arXiv [cs.LG]* (2019).

28. Ofir Press, Smith, N. A. & Lewis, M. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. *arXiv [cs.CL]* (2021).

29. MoNA - MassBank of North America. https://mona.fiehnlab.ucdavis.edu/.

30. Tsugawa, H. *et al.* A lipidome atlas in MS-DIAL 4. *Nat. Biotechnol.* 38, 1159–1163 (2020).

31. Tsugawa, H. *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* 12, 523–526 (2015).

32. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* 7, 23 (2015).

33. Watrous, J. *et al.* Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. U. S. A.* 109, E1743-52 (2012).

34. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* 34, 828–837 (2016).

35. Huber, F. *et al.* Matchms - processing and similarity evaluation of mass spectrometry data. *J. Open Source Softw.* 5, 2411 (2020).

36. de Jonge, N. F., Hecht, H., van der Hooft, J. J. J. & Huber, F. Reproducible MS/MS library cleaning pipeline in matchms. *ChemRxiv* (2023) doi:10.26434/chemrxiv-2023-l44cm.

37. Krenn, M. *et al.* SELFIES and the future of molecular string representations. *arXiv [physics.chem-ph]* (2022).

38. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners. (2019).

39. Roberta Zinc Decoder. *Hugging Face* https://huggingface.co/entropy/roberta_zinc_decoder.

40. Irwin, J. J. *et al.* ZINC20-A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* 60, 6065–6073 (2020).

41. David, L., Thakkar, A., Mercado, R. & Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminform.* 12, 56 (2020).