

# MS2Mol: A transformer model for illuminating dark chemical space from mass spectra

Thomas Butler<sup>1,a</sup>, Abraham Frandsen<sup>1,a</sup>, Rose Lightheart<sup>1</sup>, Brian Bargh<sup>1</sup>, James Taylor<sup>1</sup>, TJ Bollerman<sup>1</sup>, Thomas Kerby<sup>1</sup>, Kiana West<sup>1</sup>, Gennady Voronov<sup>1</sup>, Kevin Moon<sup>1</sup>, Tobias Kind<sup>1</sup>, Pieter Dorrestein<sup>2</sup>, August Allen<sup>1</sup>, Viswa Colluru<sup>1</sup>, and David Healey<sup>\*1</sup>

<sup>1</sup>*Enveda Biosciences, Boulder, CO, USA*

<sup>2</sup>*Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA*

<sup>a</sup>*Contributed equally*

## Abstract

The ability to identify small molecules in complex samples from their mass spectra is among the grand challenges of analytical chemistry. Improvements to this ability could significantly advance fields as diverse as drug discovery, diagnostics, environmental science, and synthetic biology. A primary bottleneck is that standard structure elucidation technologies are limited to identifying only those molecules that are contained in databases of known spectra or molecular structures and are therefore not well suited to identifying the vast majority of potentially billions of natural metabolites, whose structures are not yet catalogued. To improve the identification of molecular structures within this vast dark chemical space, we present MS2Mol, a *de novo* structure prediction model based on a generative sequence to sequence transformer. We also release EnvedaDark, a first-of-its-kind for benchmarking identification performance on unknown metabolites. EnvedaDark contains experimental mass spectra from 226 natural products not currently found in major databases. We demonstrate on this challenging dataset that MS2Mol is able to predict 21% of molecular structures to within a close-match accuracy threshold and 62% to within meaningful similarity, both of which are significant improvements over the closest match retrieved using standard database methods. We further present a confidence scorer that enables practical usage for novel molecule discovery and enriches the accuracy on meaningfully-similar and close-match thresholds to 98% and 63%, respectively, for the top 10% most confident predictions.

---

\*Corresponding author

# 1 Introduction

Understanding the small-molecule chemistry of complex samples is a central goal of many scientific endeavors, including natural products discovery, human metabolomics, synthetic biology, food science, and studies of environmental contamination. Tandem mass spectrometry (MS/MS) coupled with liquid or gas chromatography is the principal tool for these studies because it can detect large numbers of small molecules in experimental samples with high sensitivity. The resulting MS/MS fragmentation spectra are the primary experimental signature of the chemical contents of complex chemical mixtures. Yet molecular structures cannot be reliably identified for the majority of spectra from complex samples [1]. One reason for low annotation rates is that standard annotation methods work by predicting matches from databases of reference spectra or known molecules. However, a substantial fraction of the spectra in complex samples correspond to molecules not contained in existing databases. Such compounds, with unknown structure and spectra, can be said to exist in a dark chemical space. Further, in many discovery applications it is often precisely these "dark" compounds that are of greatest interest. [2].

Dark chemical space poses a particular challenge for standard compound identification methods since these methods are limited to predicting only molecular structures contained in their databases. In contrast, our *de novo* structure prediction method uses generative machine learning to predict chemical structures directly from spectra without relying on databases of known compounds. Because complex biological samples often contain a mixture of known and novel compounds, database retrieval methods and *de novo* prediction can be powerfully complementary technologies for untargeted profiling experiments.

The unique utility of *de novo* structure prediction is reflected in the relative sizes of dark and known chemical space. While there is substantial uncertainty about the number of distinct metabolite structures that exist in nature, estimates range in the billions [3, 4]. In contrast, COCONUT [5], a global repository of known natural products, contains around 400,000 compounds, suggesting that we have discovered only a tiny fraction of existing biological metabolites, of which an even smaller number, in the tens of thousands, have experimentally-determined reference spectra available in repositories like NIST2020, MoNa, MassBank, and GNPS. Synthetic chemical space is similarly dark, with large structure databases like ZINC containing in the billions of molecules [6], still a small fraction of the potentially  $10^{60}$  possible molecular structures that could exist. Such non-naturally occurring, but unknown, structures can emerge in critical settings like environmental contaminants. In short, all estimates point to dark chemical space being astronomically larger than known chemical space, supporting a critical role for *de novo* structure prediction.

Compound identification from mass spectrometry can be naturally categorized into three different approaches: spectral reference search, compound library search, and *de novo* generation (see Figure 1).

*Spectral reference search* based systems implement a similarity function for pairs of spectra,  $\text{sim} : S \times S \rightarrow [0, 1]$  where  $S$  is the space of MS/MS spectra. These systems allow users to match experimental unknowns to spectra in reference libraries. Predicted structures then consist of the highest scoring matches from the database. One common similarity function

is a modified cosine similarity function [7], but more sophisticated similarity functions use unsupervised learning to create dense embeddings of spectra, which have been shown to achieve higher accuracy than cosine similarity in some contexts. Examples of these include word2vec-style embeddings [8] and Siamese networks [9, 10]. Spectral library search is limited by the relatively small number of compounds found in publicly available reference libraries.

*Compound library search* based systems attempt to overcome the limited coverage of spectral reference libraries by creating a similarity function between mass spectra and candidate molecular structures,  $\text{sim} : S \times M \rightarrow [0, 1]$ , where  $S$  is the space of mass spectra, and  $M$  is the space of molecular structures. This allows a mass spectrum to be used as a query into molecular databases such as PubChem [11] and COCONUT[5] that lack mass spectral data but contain many more structures.

There are two main approaches to compound library search. The first uses one of many *in silico* fragmentation algorithms to convert molecular libraries into (synthetic) spectral libraries and then performs a spectral library search on the synthetic spectra [12, 13, 14, 15, 16]. The second predicts a molecular fingerprint from the mass spectra, sometimes with prior prediction of molecular formulas and/or fragmentation trees. The molecular fingerprint indicates which of a predetermined set of substructures or features are present in a structure. This method then compares the fingerprint to a library of fingerprints from a structural database of known compounds. Examples of these methods are CSI:FingerID [17], implemented in the SIRIUS software suite, and the tool used in the highest-accuracy submission to the recent Critical Assessment of Small Molecule Identification (CASMI) contest [18]. This is also used by MIST[19], which predicts the same CSI:FingerID fingerprints using a neural network.

While designed to identify known molecules, compound library search may be adapted for identification of novel structures in cases where the candidate set of structures is limited enough that a library containing all possible candidate structures can be enumerated. An example is the discovery of novel bile acid conjugates using CSI:FingerID and COSMIC [20], which used 16,000 hypothesized bile acid conjugates as the molecular library to search. For complex mixtures containing heterogenous molecular structures, however, the space of possible structures is likely to be too large to enumerate.

*De novo structure prediction* systems use machine learning to generate potentially novel molecular structures directly from mass spectra, with three different methods published in recent years. MSNovelist [21] relies on CSI:FingerID [17] to predict a fingerprint for an unknown molecule, then uses a decoder trained on molecular fingerprints from structure libraries to translate the fingerprint into a molecular structure represented by a Simplified molecular-input line entry system (SMILES) string [22]. MassGenie [23] trains a model with a standard transformer encoder-decoder architecture on large amounts of *in silico* fragmentation data generated from chemical structure libraries, finetuned with experimental spectra. Spec2Mol [24], like MSNovelist, uses a separate pretrained decoder, but instead of predicting fingerprints, a convolutional neural network (CNN) spectrum encoder is trained to predict into the learned embedding space of a pretrained GRU SMILES autoencoder, the decoder half of which is then used to generate the output SMILES.

In this work we present MS2Mol, a transformer-based model for the *de novo* generation of

molecular structures from MS/MS spectra. This model uses an encoder-decoder architecture adapted from BART [25] and is trained end-to-end on spectrum-structure pairs. MS2Mol's design differs from previous *de novo* models in several ways. First, we introduce byte pair encoding for SMILES predictions, which provides us with a learnable substructure vocabulary fit on the training dataset, allowing the model to add common substructures and SMILES elements with single tokens. Secondly, we include precursor mass as an input, allowing the model to take advantage of knowing the mass of the unfragmented compound, but avoid overfitting to that information by randomly masking it during training. Third, we encode fragmentation masses using separate integer and fractional mass tokens, which, relative to naive mass binning, maintains the high-resolution mass information needed to disambiguate chemical formulas while representing with a smaller vocabulary. Fourth, we train a small reranking model and use it to rank the MS2Mol predictions generated for a single input spectrum, which improves the accuracy of the top-ranked predictions. Finally, we use a confidence model to prospectively predict the quality of the predictions, enabling users of the model to act on a subset of highly-confident predictions in applications where errors could be costly.

Evaluating the performance of *de novo* structure elucidation methods against database methods is challenging. Evaluation sets must have known structures, so they commonly consist of holdout sets from spectral reference libraries or consist of former challenge spectra from the Critical Assessment of Small Molecule Identification (CASMI) challenges. Since their structures are typically known, database retrieval methods are advantaged on these sets, since these methods drastically reduce the search space to that of known structures. Furthermore, it is common for structure evaluation to only consider exact structure match as a metric of success; however, database search methods by definition cannot retrieve an exact match structure for an unknown compound. These methods can still be useful as baseline methods, though, by identifying analogs from known compounds. Previously published *de novo* methodologies have dealt with these challenges in different ways; MassGenie does not benchmark against a molecular-search baseline. MSNovelist benchmarks against CSI:FingerID on molecules contained in its databases and performs worse at retrieving the exact structure, noting that this is to be expected. Spec2Mol evaluates against CSI:FingerID using similarity rather than exact match as a metric, but to proxy dark chemical space only evaluates on the set of spectra for which SIRIUS fails to predict the correct molecular formula. This disadvantages CSI:FingerID since SIRIUS often correctly predicts molecular formulas even of unknown molecules, and CSI:FingerID relies heavily on accurate molecular formula predictions.

In this work we take a more relevant and stringent evaluation approach, for the first time evaluating *de novo* structure prediction in dark chemical space. We evaluate MS2Mol on a novel set of spectra from 226 molecules that are naturally occurring but whose structures are not contained in structural databases. This mimics directly the intended use case for *de novo* methods. We term this dataset EnvedaDark (for “dark chemical space”). Further, we report as our primary performance metrics accuracy with respect to two Tanimoto similarity thresholds representing meaningful similarity and close match similarity, which were defined to match blinded expert chemist annotations of prediction usefulness for the purposes of drug development prioritization. We include both spectral library search (modified cosine



similarity) and molecular library search (CSI:FingerID) as baselines. We are not able to benchmark against previous *de novo* models since, as of writing, no other *de novo* method has a working public implementation.

On EnvedaDark, MS2Mol predicts a structure to within meaningful similarity of the actual molecule for 62% of spectra, and 21% to within close match, exceeding the highest database retrieval baselines by 44% and 95% relative accuracy, respectively at those thresholds. We evaluate MS2Mol’s confidence model, which enriches the accuracy on meaningfully-similar and close-match thresholds to 98% and 63%, respectively, for the top 10% most confident predictions. We further evaluate MS2Mol on two sets of known molecules: CASMI-2022 and a novel set of spectra experimentally gathered from 454 known natural products, which we call EnvedaLight. We show that despite database methods excelling at retrieving exact matches, MS2Mol performs comparably to database matches at predicting meaningfully similar or close matched analogs even on known molecules. This introduces the possibility of using a single structure elucidation model to predict both known and unknown molecules in complex mixtures.

## 2 Results

### 2.1 MS2Mol generates predicted chemical structures using a transformer encoder/decoder model

MS2Mol casts *de novo* structure prediction from MS/MS data as a neural machine translation problem to map from the "language" of MS/MS fragments into the language of small molecules, as defined by SMILES strings. Natural language frameworks have been observed to be valuable for interpretation of mass spectra [26, 8], since the meaning of mass fragments are contextually dependent on the presence and absence of other fragments, and the fragments are composed into an overall spectrum in a similar way to words in a sentence. The development of transformer encoder-decoder architectures in particular [27] significantly improved machine translation models by enabling more efficient computation while allowing the learning of complex and long ranged dependencies. MS2Mol uses the BART implementation of a transformer based encoder-decoder architecture [25, 28].

Input mass spectra consist of triples of the form {precursor  $m/z$ , fragment  $m/z$  array, intensity array}, where the precursor  $m/z$  is a positive real number corresponding to the mass-to-charge ratio of the unfragmented ion, the fragment  $m/z$  array is an array of positive real numbers that correspond to the mass-to-charge ratios of the MS/MS fragments, and the intensity array is an array of positive real numbers corresponding to the intensity of the fragments. In analogy with natural language processing, MS/MS fragments are represented as a language with a vocabulary of mass values. Input spectra are encoded as pairs of integer and fractional part tokens corresponding to masses and sorted in descending order by intensity with the precursor mass in first position, as shown in Figure 1. This allows a smaller overall vocabulary for a given mass resolution than previous work that has tokenized spectra into mass bins (e.g. [9]).

The output structures are encoded as SMILES string representations of chemical structures

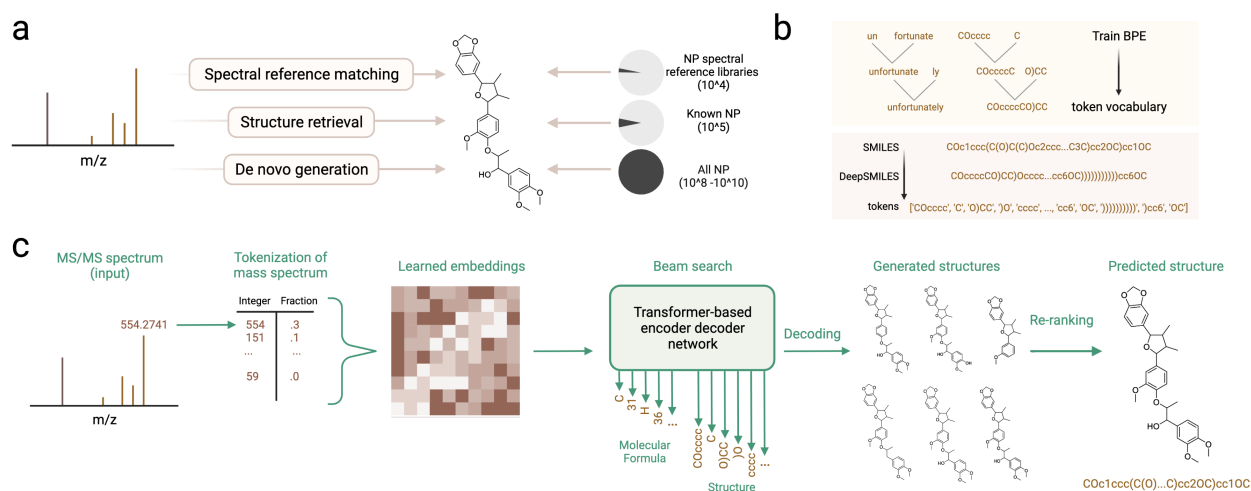


Figure 1: MS2Mol generates predicted chemical structures of unknown molecules from dark chemical space using their fragmentation spectra. **a)** Overview of approaches to structure elucidation. Common approaches to identification of small molecules from their MS/MS spectra include matching to spectral reference libraries (top) and information retrieval from candidate databases of known structures (middle), for which the relevant metabolite databases contain on the order of  $10^4$  and  $10^5$  structures, respectively. In contrast, accurate elucidation of the vast majority of natural metabolites ( $10^8$ - $10^{10}$ ) requires de novo structure prediction. **b)** An illustration of byte-pair encoded (BPE) substructure learning for molecule generation. Similar to BPE in natural language, tokens (masses) that are commonly adjacent are combined into single new tokens. Those tokens may then be further paired with . The process proceeds iteratively, and the resulting vocabulary contains common structural elements of the molecules in the training set encoded as single tokens. **c)** Conceptual overview of MS2Mol encoder/decoder. Spectra are sorted by intensity then tokenized separately into integer and decimal parts, which are then converted to learned fragment embeddings. The spectrum, which includes a probabilistically-masked precursor mass, is then passed into a transformer encoder/decoder. Decoding is accomplished using beam search to produce a ranked set of predicted molecular formulas and chemical structures represented as byte-pair encoded SMILES.

[22] using byte pair encoding (BPE), a method for finding a high coverage vocabulary of substrings in natural language processing [29] through iterative combining of high-frequency pairs of tokens into single tokens. In the context of molecule generation, this results in a token vocabulary that contains common molecular substructures and other common SMILES elements like ring and branch closures encoded as single tokens, learned from the data rather than predefined. This shortens the overall output length and reduces the chances for error during autoregressive generation of structures.

The mass of the unfragmented ion, called the precursor mass, contains valuable information on the identity of the molecule. For example, molecular formula can be disambiguated to a large degree using an accurate precursor mass. However, training on MS/MS data is prone to overfitting if the precursor mass from the MS1 spectrum is naively included in the spectrum. This is because over the training set, the precursor mass is unique or nearly unique for many compounds, allowing the model to memorize the output as a function of precursor mass. To learn from the precursor mass without overfitting, we randomly mask the precursor mass with 50% probability, and leave it unmasked at inference time.

A further complication of available MS/MS libraries is that certain compounds have far more spectra associated with them than others. To avoid overfitting to such compounds, we weight the training objective to account for the multiplicity of spectra per compound by using a modified cross entropy loss, described in Section 4.

MS2Mol is designed to predict chemical structures without ground truth molecular formulas, as those are typically not known in discovery scenarios. However, the output of MS2Mol consists of molecular formula followed by chemical structure. This allows for the case in which the ground truth formula is known or can be accurately predicted. Since the output is generated autoregressively, prompting the output with the ground truth molecular formula allows the chemical structure to be partially conditioned on the correct formula, which increases accuracy. See Appendix D.

To train MS2Mol, we constructed a dataset of 1,020,431 MS/MS spectra paired with 42,995 structures obtained by merging standard commercially available and public datasets spanning instrument types, ion modes, collision energies, and other parameters [7, 30, 31, 32, 33, 34], supplemented with a dataset of 33,634 internally-generated reference spectra from 6,124 purified natural products, for a total of 961,865 spectra from 49,119 structures. We further supplemented the experimental data with weak supervision using publicly available simulated mass spectra generated by CFM-ID 4.0 from natural products contained in the LOTUS database [35, 36, 12]. See Section 4 for further training details.

Ablation studies of each of these choices are presented in the Appendix.

## 2.2 MS2Mol predicts unknown structures with higher accuracy than spectral library search and molecular library search

In order to assess performance on dark chemical space, or compounds that are not presently contained in common molecular databases or annotated spectral libraries, we constructed a dataset of 1010 spectra from 226 purified “novel” natural products across three collision

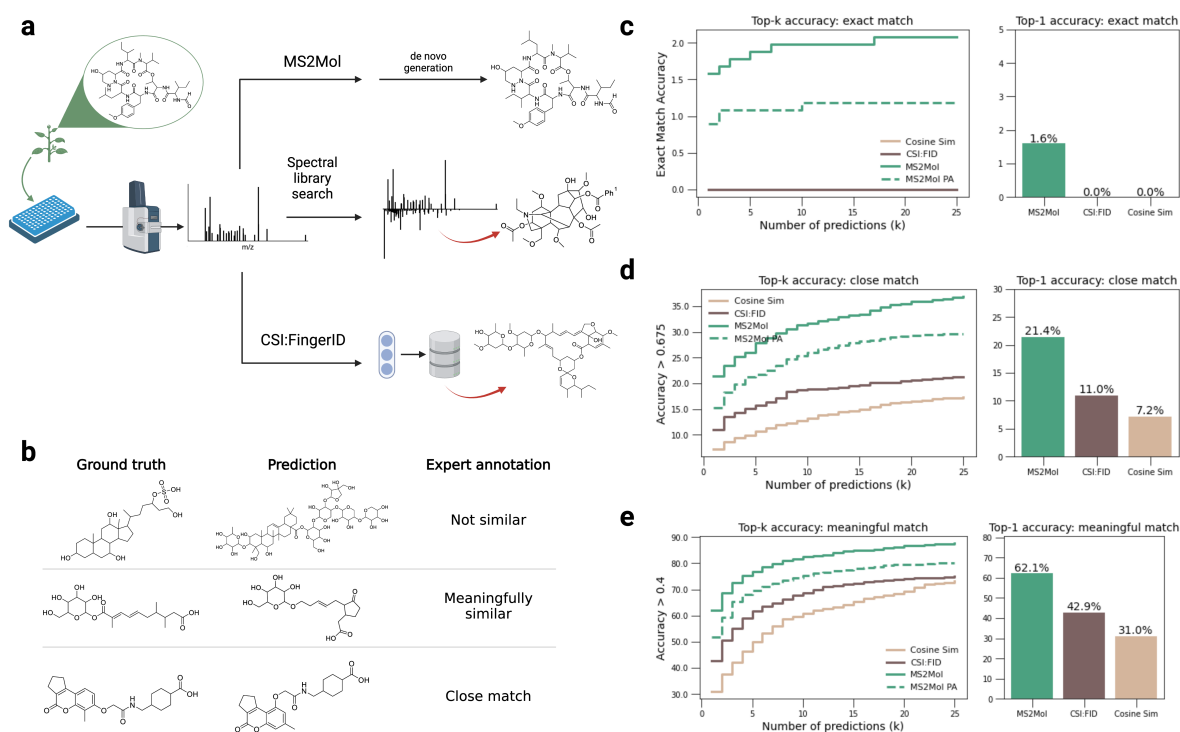


Figure 2: MS2Mol outperforms both spectral reference search and molecule retrieval on predicting structures of unknown natural products at multiple similarity thresholds. **a**) Example structural elucidation using three methods. A test spectrum is gathered experimentally from a sample molecule whose structure is not in a database, then top-1 structure prediction is performed using *de novo* prediction (MS2Mol, top), spectral reference search (modified cosine similarity, middle), or molecular database retrieval (CSI:FingerID, bottom). **b**) Examples demonstrating the expert annotations of pairs of actual and predicted structures that were to define meaningful similarity and close match Tanimoto similarity thresholds (see also Appendix). **c**) - **e**) Performance statistics for Cosine based retrieval, CSI:FingerID, and MS2Mol. MS2Mol indicates the full model, while MS2Mol PA (publicly available) indicates MS2Mol trained only with open and commercially available data as described in the text. **c**) Top-*k* and top-1 similarity between the ground truth and the predicted structure for exact InChIKey-14 match (two dimensional structure). The curves for the two baseline methods are exactly overlaid. **d**) Top-*k* and top-1 similarity between the ground truth and the predicted structure for the close match criterion (see main text) of >0.675 Tanimoto similarity. **e**) Top-*k* and top-1 accuracy for the meaningfully similar criterion (see main text) of >0.4 Tanimoto similarity between ground truth and the predicted structure.

energies and seven possible ion adducts, where we define novelty as compounds whose structures are not contained in common structural repositories Pubchem [11] or COCONUT [5], and therefore not available for database search methods or as training data at the time of writing. We term this dataset EnvedaDark.

To evaluate performance accuracy, we adopt as our primary metrics “meaningfully similar” and “close-match” accuracy. We note that common compound identification metrics often consider exact match success metrics [17, 21, 13, 18], and we report these as well. However, while exact match metrics may be suited to evaluating the ability of database search methods to produce the exact two dimensional structure from a predefined set of candidate structures, when evaluating prediction accuracy on test spectra from dark chemical space, exact-match metrics bias against database methods, which cannot, by definition, produce an exact match prediction corresponding to a structure outside of its databases. Furthermore, in many applications including discovery of novel natural products, a relevant metric of success is whether the prediction is a close enough approximation of the actual molecule to be useful for a given application. Given these observations, we adopt as our primary evaluation framework measuring the fraction of molecules for which the prediction meets or surpasses a suitable structure similarity threshold.

To obtain automated metrics that assess whether a prediction is sufficiently similar to the ground truth molecule to be useful for guiding decision making, we fit two thresholds of Tanimoto similarity (between RDKit topological fingerprints [37]) to expert chemist annotations of pairs of predicted versus actual structures, where predicted structures were generated either from previous generations of MS2Mol, CSI:FingerID, or were chosen randomly from the training set as a control. 1288 such pairs were sequentially shown to six expert chemists. The chemists were aware of the composition of the set, but blinded as to the source of the predictions, and scored pairs of annotations as 1) not similar, 2) meaningfully similar but with errors, and 3) close match. (See the Appendix for a screenshot of the annotation tool, full annotation guidelines, and examples of expert annotations). We chose a Tanimoto similarity threshold of 0.675 to represent “close match accuracy” because it predicts the overall rate of close match predictions correctly. As shown in the Appendix, the threshold that correctly predicts the class balance for close match also balances precision and recall of the similarity scores with respect to the chemist annotations. Likewise, we chose Tanimoto similarity of 0.4 to represent “meaningful similarity”, using the upper 95% confidence bound of chemist annotations as a conservative threshold to minimize erroneously calling "not similar" predictions as being meaningfully similar, which is a more common error in the lower Tanimoto similarity range. See Figure 2 and the Appendix for samples of chemist annotations and bootstrap fitting of Tanimoto similarity thresholds.

We compare MS2Mol’s *de novo* generation approach with standard alternatives in spectral and molecule database search – see Figure 2 and Table 1. While it would be ideal to benchmark against other *de novo* structure prediction methods, as of writing, MSNovelist [21], MassGenie [23], and Spec2Mol [24] lack functioning public implementations. We note that while neither reference lookup nor molecule database search were intended to predict novel structures, novel molecules are nonetheless highly prevalent in nature, and to predict their structures, a reasonable alternative to *de novo* methods is to use database retrieval to find close analogs.

For spectral library search, we use modified cosine using the `matchms` [38] implementation with a spectral reference library consisting of the public and commercially available data described in Section 2.1. We also compare to CSI:FingerID, a state-of-the-art molecular search tool implemented in the SIRIUS software suite [17], and the tool which resulted in the best annotation performance in the most recent Critical Assessment of Small Molecule Identification (CASMI) contest [18]. CSI:FingerID searches molecular databases by using mass spectra to predict a suite of molecular substructures, then searches molecular databases for molecules with similar substructure profiles.

We evaluate both the top-scoring predictions of each method and the top- $k$  predictions (up to  $k = 25$ ) for each of the three methods at the “meaningfully similar” and “close match” thresholds defined above. We also evaluate “exact match” prediction (minus stereochemistry), defined as an exact string match of the first 14 digits of the respective InChiKeys for the predicted and actual structures.

MS2Mol performs strictly better than the baseline methods at all three similarity thresholds on the EnvedaDark dataset. Using the above threshold criteria, MS2Mol’s top predictions are at least meaningfully similar for 62.1% of spectra, compared to 42.9% and 31.0%, respectively, for the top predictions from CSI:FingerID and modified cosine reference search, while being at least close-matched for 21.4% of spectra, compared to 11.0% and 7.2% for CSI:FingerID and modified cosine reference search, respectively. For only 1.6% (16 spectra from 7 molecules) does MS2Mol predict an exact structure match, compared to 0% for the other two, where 0% is expected since the EnvedaDark molecules are not found in those databases.

Furthermore, more predictions per spectra increase the likelihood of predicting a better match. The top-25 predictions from MS2Mol contain a meaningfully-similar prediction for 87.6% of spectra (up from 62.1% top-1), and a close match prediction for 36.8% (up from 21.4% top-1). These improvements between top-25 and top-1 suggest further predictive value to be gained by accurate re-ranking of relatively small numbers of *de novo* predictions. (Figure 2 and Table 1).

Since MS2Mol’s full training set contains spectra generated under the same experimental setting from which EnvedaDark spectra were derived, we also evaluate a version of MS2Mol that was not trained on internal spectra. We call this model MS2Mol PA (“MS2Mol publicly available”); performance of this model is represented with a dotted line in Figure 2). We observe that training on internal spectra provides a lift in performance, which demonstrates the benefits of training on in-domain labeled spectra for real-world applications. However, we show that MS2Mol still comfortably outperforms the baselines even without internal spectra in the training set.

## 2.3 Performance analysis of MS2Mol

We further investigate MS2Mol by examining trends in the performance of predictions on EnvedaDark, noting potential strengths and weaknesses of MS2Mol in predicting natural products in dark chemical space. Firstly, we observe that MS2Mol is able to accurately predict structures even if molecules have fairly complicated extended structures with repeating subunits, as is relatively common for natural products; see Figure 3a. When examining



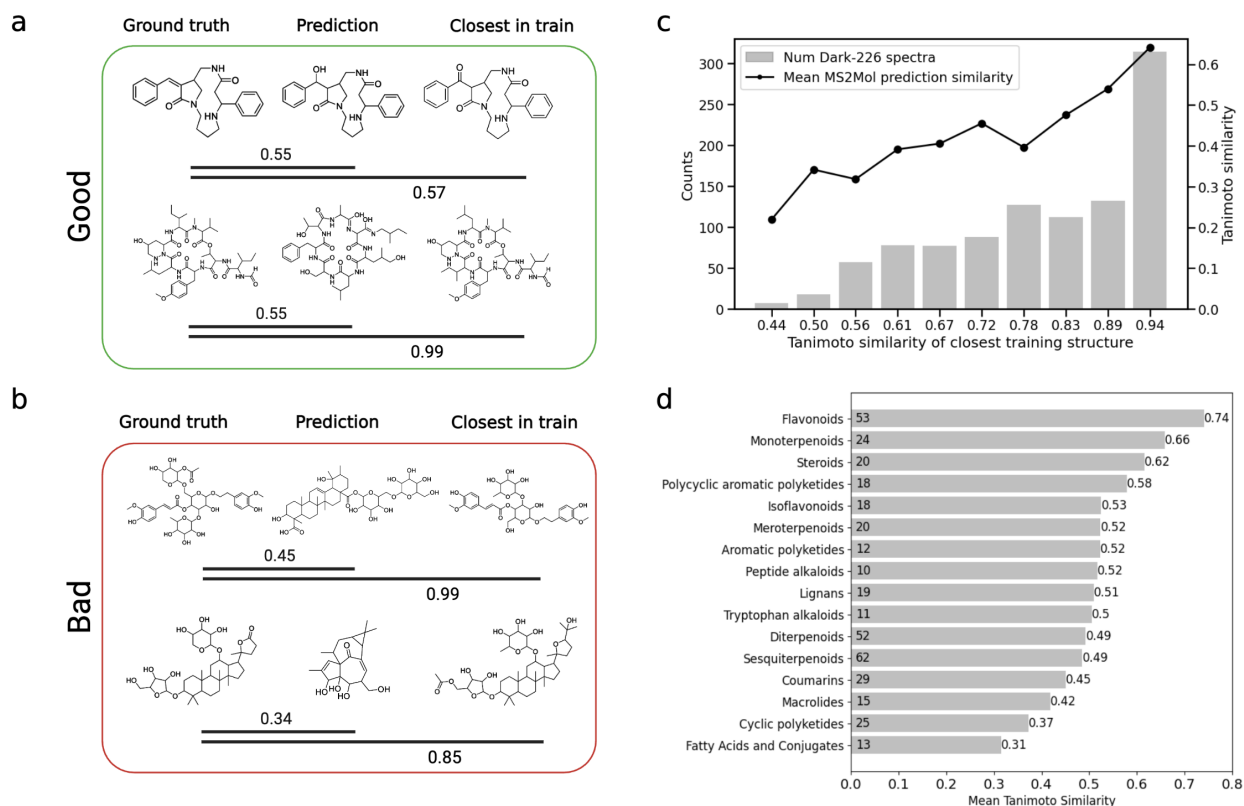


Figure 3: Performance analysis of MS2Mol. **a)** Examples of good MS2Mol predictions. Ground truth structure, predicted structure, and the closest molecule in the training set are shown for two examples from the test set, representative of cases where MS2Mol predicted an exactly or approximately correct structure. The upper example shows a case where MS2Mol predicted the structure exactly correct despite no exact match present in the training data. The lower example demonstrates that MS2Mol can predict well even for relatively complex structures. **b)** Examples of bad MS2Mol predictions. Ground truth structure, predicted structure, and the closest molecule in the training set shown for examples from EnvedaDark, representative of cases where MS2Mol predicted a structure with low Tanimoto similarity to the actual structure. The upper case demonstrates the difficulty of predicting structures far away from anything in the training set. The lower case demonstrates a potential failure mode of MS2Mol in discriminating between linear and cyclic structures. **c)** Mean Tanimoto similarity between predicted and ground truth molecules from EnvedaDark (line) over bins of similarity between the ground-truth molecule to the training set of annotated spectra, overlaid with a histogram of spectrum counts across the same similarity bins. **d)** Mean Tanimoto similarity on a test set of natural products across 16 common natural product classes. Left numbers indicate numbers of distinct compounds represented in each class.

inaccurate predictions, we find spectra from “darker” chemical space, i.e. spectra whose structures are dissimilar to everything in the training set.

To further investigate how the “darkness” of chemical space affects the performance of MS2Mol, we examine the performance of MS2Mol binned by the Tanimoto similarity between the ground truth structure and the closest molecule in the training set (excluding the simulated spectra). As with many machine learning systems, we observe that prediction accuracy is highly correlated with proximity of the unknowns to elements in the training set. The average similarity of predictions to ground truth is nearly 3x higher for spectra from the most similar ( $\geq 0.94$  Tanimoto similarity) structures to train than for the least similar (between 0.44 and 0.50 Tanimoto Similarity). We further note a difference in performance across natural product classes as predicted by NP Classifier [39], with flavonoids, monoterpenoids, and steroids having the highest Tanimoto similarities, while macrolides, cyclic peptides, and fatty acids had the lowest. The latter result reinforces our observations regarding cyclic structures.

## 2.4 Confidence modeling enables practical use of MS2Mol

For many machine learning applications, particularly those in which poor predictions can be costly, it is desirable to not only have aggregate accuracy metrics but to also estimate the reliability of each prediction. This allows expert decision makers to segment predictions by confidence and potentially act only on those that pass a quality threshold. To enable this, in addition to the structure prediction model we trained a separate confidence model that predicts the Tanimoto similarity between the MS2Mol predicted structure to the ground truth structure.

The confidence model is a gradient boosting machine [40] with input features derived from the MS/MS spectrum (such as the number of peaks and the precursor  $m/z$ ), properties of the predicted structure (such as the number of heteroatoms), structural properties predicted directly from the MS/MS spectra by a property prediction model as in [10], and MS2Mol model outputs (such as the log probability of the generated token sequence). More details of the confidence model training are given in Section 4 and the Appendix.

To evaluate the confidence model, we binned predictions on EnvedaDark into confidence deciles (Figure 4a). We find the top 10% most confident predictions achieve a 63.4% close-match accuracy and a 98% meaningful similarity accuracy, compared to the 21% and 62%, respectively, across all confidence levels. Furthermore, we find the confidence model to be well-calibrated in the sense that, within the binned deciles, the actual similarities are close to the predicted similarities (Figure 4b). The confidence model achieves 0.13 mean absolute error between the true and predicted Tanimoto similarity, with an  $R^2$  of 0.4. We next inspected the features of the confidence model most associated with better predictions, using permutation based feature importance [41] (Figure 4c). Feature importance calculations depend heavily on methods and data properties such as collinearity, but the resulting feature importance shows a large dependence on mass difference features, which is chemically reasonable and points to the possibility that generative methods that constrain mass more effectively could be a powerful advance in the future.

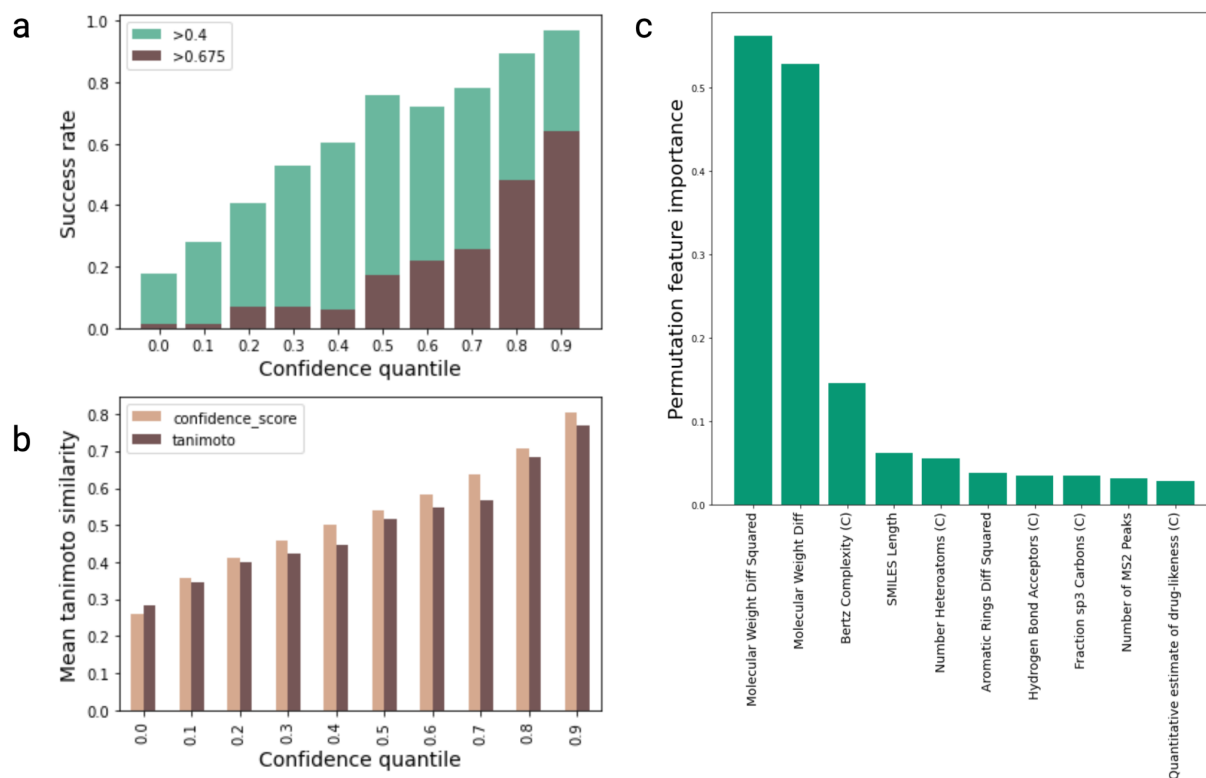


Figure 4: Confidence model performance. **a)** The fraction of predictions with similarity greater than 0.675 and greater than 0.4 as a function of confidence deciles in EnvedaDark. The top 10% highest predicted similarities correspond to predictions that are over 60% accurate at the 0.675 threshold. **b)** The confidence model accurately predicts the Tanimoto similarity at all confidence deciles. Each pair of bars compares the true mean Tanimoto similarities with the predicted similarities grouped by decile bins of the predictions. The slight optimism in the confidence model is due to the difference in domain between the confidence model training set and EnvedaDark. **c)** The relative importance of features used in the confidence model using permutation importance. If the feature name has a (C), that indicates that the feature is calculated from the predicted structure. “Diff” represents a difference between calculated and observed, or in the case where there is no direct observation, the property value predicted using the property prediction model. “Diff squared” is the square difference of the calculated and the observed/predicted properties. The top features are dominated by mass difference features, showing that models that are able to predict structures with higher mass fidelity are a promising avenue for future research. For a full description of the features used in the confidence model, see Section 4 and the Appendix.

## 2.5 Evaluating MS2Mol on known chemical space

In previous sections, we introduced the first evaluation of a *de novo* structure prediction technology on dark chemical space. As a baseline, we compared it to molecular and spectral library search intended to assign known structures to unknown spectra. In this section, we reverse the above scenario to ask how well MS2Mol performs on spectra from known molecules despite the fact that MS2Mol does not benefit from the reduction in search space afforded by limiting the search to structures contained in compound databases. This is a challenge setting, as the model is being evaluated outside of its intended use case.

We evaluate MS2Mol, CSI:FingerID and Modified Cosine spectral reference search on two evaluation sets. The first is the CASMI 2022 data set [18]. For the second evaluation set of known molecules, we created an additional test set that we call EnvedaLight, comprised of 1856 spectra from 454 molecules profiled in our own lab whose structures are available in standard molecular databases, but whose spectra are not contained in spectral reference libraries. The results are summarized in Table 1.

On both test sets, CSI:FingerID outperforms MS2Mol at retrieving the exact match: 13% to 9% and 11% to 2% for CASMI and EnvedaLight, respectively. This is an expected result, as CSI:FingerID is a database retrieval method evaluated on molecules known to be in its database. What is perhaps surprising, however, is that the results are more mixed when considering similarity thresholds rather than exact match. On CASMI, MS2Mol and CSI:FingerID are essentially equal on close-match accuracy for all top- $k$ , while MS2Mol outperforms CSI:FingerID for meaningful-match accuracy. On the EnvedaLight test set, MS2Mol outperforms CSI:FingerID and modified cosine similarity on close-match accuracy for all  $k \geq 2$  both with and without training on any internal data. When training on in-domain data, MS2Mol outperforms by at least 6 points across all  $k$  values.

Thus, when considering heterogeneous complex samples containing both known and unknown molecules, use of database retrieval and *de novo* methods may hinge on the application: known molecules will be better annotated with the exactly-correct structure using database retrieval, with unknown molecules better annotated *de novo*. In this way they can be considered complementary. However, the relative comparability of performance at close-match and meaningfully-similar threshold accuracy raises the intriguing possibility that *de novo* methods may perform well as a single model on heterogeneous samples in applications where close-match accuracy is sufficient.

## 3 Conclusion

Here we have presented MS2Mol, a end-to-end molecular structure prediction model that accurately predicts chemical structures directly from MS/MS spectra. Importantly, it is a true generative, *de novo* structure prediction model that does not rely on compounds existing in any database of known or proposed structures.

For the first time to our knowledge, we systematically benchmark performance on the task of predicting structures of unknown molecules whose structures are not found in common

| EnvedaLight  |             |             |             |             |             |             |             |             |             |
|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Model        | MM@1        | MM@10       | MM@25       | CM@1        | CM@10       | CM@25       | EM@1        | EM@10       | EM@25       |
| MS2Mol       | <b>72.7</b> | <b>88.0</b> | <b>91.5</b> | <b>37.0</b> | <b>50.6</b> | <b>56.6</b> | 2.3         | 8.0         | 8.5         |
| MS2Mol PA    | 64.0        | 83.2        | 86.9        | 27.6        | 42.8        | 47.8        | 1.8         | 4.4         | 4.5         |
| Cosine       | 15.1        | 63.7        | 81.0        | 2.3         | 10.2        | 15.8        | 0.8         | 1.5         | 1.9         |
| CSI:FingerID | 55.5        | 77.2        | 81.7        | 29.7        | 41.4        | 47.2        | <b>10.5</b> | <b>26.1</b> | <b>31.7</b> |

| CASMI 2022   |             |             |             |             |             |             |             |             |             |
|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Model        | MM@1        | MM@10       | MM@25       | CM@1        | CM@10       | CM@25       | EM@1        | EM@10       | EM@25       |
| MS2Mol       | <b>63.9</b> | <b>81.0</b> | <b>85.2</b> | <b>36.1</b> | <b>47.9</b> | <b>51.7</b> | 9.4         | 12.0        | 12.8        |
| MS2Mol PA    | 62.5        | 79.4        | 82.8        | 29.1        | 43.3        | 48.9        | 7.2         | 10.0        | 10.4        |
| Cosine       | 44.9        | 66.1        | 73.1        | 24.2        | 36.9        | 41.3        | 6.6         | 8.0         | 8.2         |
| CSI:FingerID | 57.9        | 71.3        | 74.9        | <b>35.5</b> | <b>48.1</b> | <b>52.7</b> | <b>13.4</b> | <b>28.9</b> | <b>32.3</b> |

| EnvedaDark   |             |             |             |             |             |             |            |            |            |
|--------------|-------------|-------------|-------------|-------------|-------------|-------------|------------|------------|------------|
| Model        | MM@1        | MM@10       | MM@25       | CM@1        | CM@10       | CM@25       | EM@1       | EM@10      | EM@25      |
| MS2Mol       | <b>62.1</b> | <b>82.6</b> | <b>87.6</b> | <b>21.4</b> | <b>31.7</b> | <b>36.8</b> | <b>1.6</b> | <b>2.0</b> | <b>2.1</b> |
| MS2Mol PA    | 51.9        | 75.4        | 80.1        | 15.2        | 26.0        | 29.6        | 0.9        | 1.2        | 1.2        |
| Cosine       | 31.0        | 60.8        | 73.3        | 7.2         | 13.3        | 17.3        | 0.0        | 0.0        | 0.0        |
| CSI:FingerID | 42.9        | 68.8        | 75.1        | 11.0        | 18.8        | 21.2        | 0.0        | 0.0        | 0.0        |

Table 1: Evaluation of MS2Mol and other methods on all test sets. MS2Mol PA denotes MS2Mol trained only on publicly available (ie not including internally-generated spectra), and all other models are as described in the text. The results are for the metrics “Meaningful Match” (MM), “Close Match” (CM) and “Exact Match” (EM) as described in the text. All metrics are assessed according to the best of a list of retrieved candidates for list lengths (1, 10, 25). On all test sets, MS2Mol and MS2Mol PA outperform for meaningful matches (MM), even when compared to molecules with known chemical structures. For close match structures, both variants of MS2Mol perform comparably to methods designed for known chemical space, and outperform substantially on dark chemical space. On known molecules CSI:FingerID outperforms on exact match prediction. On dark chemical space, the intended application domain of MS2Mol, MS2Mol outperforms all competitive methods for top- $k$  accuracy for all metrics, even without in-domain training data.

databases. We demonstrate that MS2Mol outperforms current methods by a wide margin, indicating that *de novo* methods can be powerful additions to search methods. We note that while our results are promising, structure elucidation is far from solved.

There is significant room for improvement, particularly at predicting exact structures. We note that, when considering that dark chemical space in the natural world is orders of magnitude larger than known molecules or spectral reference libraries, even a modest rate of predicting dark space molecules with high confidence can have a very large impact on the total number of molecules that can be identified.

MS2Mol can be improved in many ways. The most promising avenue is to scale up data using both experimentally-generated spectra and synthetic data. Improvements to the accuracy and particularly the scalability of *in silico* spectrum prediction models may well enable future training on orders of magnitude more synthetic spectra [42, 43]. In this work we did not explore pretraining on chemical structures as other methods have, but that remains a promising possibility. The presence of large repositories of unannotated spectra like GNPS raises the possibility of making use of unlabeled spectra as well. Finally, mass errors in the generated molecules are an obvious prediction failure mode, and training or decoding approaches that incorporate mass knowledge more intelligently may improve upon the performance of the system described here.

Because mass spectrometry data is highly variable in its processing and chemical space is diverse, comparisons of methods in structure prediction are intrinsically uncertain. A method that performs comparatively well in one benchmarking setting may perform poorly in another, reversing rankings. Some methods may be more sensitive to noise in the precursor mass measurement, or favor settings with more or fewer peaks, or different regions of molecular space. Because of this variance, the performance of a method on real data is often difficult to identify prospectively, and continues to depend on use case, data processing choices, instrument precision, and other factors.

MS2Mol successfully identifies close analogs and exact matches of compounds that have never before been described or stored in major databases. By applying the confidence model, accurate annotations can often be identified. Considering that even conservative estimates suggest that most of chemical space is still unknown, MS2Mol is an important step toward further illuminating the small molecule chemistry of life.

## 4 Methods

### 4.1 Datasets

The makeup of the spectra from publicly available data sources is described in Section 2.1. We compile the CASMI dataset as outlined in Section G. To prepare the data for training and evaluation, we process spectra by sorting the peaks by intensity and retaining the top 256 peaks. We further drop spectra with fewer than five peaks. We normalize structures by removing stereochemistry, selecting the largest ion where the smiles included salts, and converting to a canonicalized tautomer for each structure using RDKit.



#### 4.1.1 Experimental settings and MS/MS library acquisition for Analyticon library

Purified natural products were purchased from AnalytiCon Discovery (Potsdam, Germany). The compounds were pooled according to different mass ranges and transferred using a Hamilton Star liquid handling system to 384 low dead volume (LDV) plates. Samples were reconstituted in 50:50 Methanol/Water (v/v) and directly injected from the 384 well plates into the LC-MS/MS system. Method blank and quality control samples were inserted across all chromatographic runs.

The Waters ACQUITY UPLC system was run with a methanol/water gradient at 30 °C with a total run time of 7 minutes using a 1.7  $\mu$ m, 2.1 x 50 mm Waters ACQUITY Premier BEH C18 reversed phase column. The timsTOF Pro 2 mass spectrometer (Bruker Daltonics, Bremen, Germany) was equipped with a VIP-HESI ion source (dry gas temperature 220 °C) and data was acquired from 20-1300 m/z in positive and negative ionization mode. Tandem mass spectra were recorded at three different CID voltages (20, 40, 60 eV).

The MS/MS spectra were processed by an in-house developed data processing pipeline. The software combines feature finding with AWS Athena and SQL queries across multiple data sets in parallel. MS/MS spectra were extracted across multiple adduct ion forms ([M+FA-H]<sup>-</sup>, [M+Cl]<sup>-</sup>, [M-H]<sup>-</sup>, [2M-H]<sup>-</sup>, [2M+FA-H]<sup>-</sup>, [M-H<sub>2</sub>O-H]<sup>-</sup>, [M+Br]<sup>-</sup>, [M+NH<sub>4</sub>]<sup>+</sup>, [2M+H]<sup>+</sup>, [M+H]<sup>+</sup>, [M+K]<sup>+</sup>, [M+Na]<sup>+</sup>, [2M+Na]<sup>+</sup>, [M-H<sub>2</sub>O+H]<sup>+</sup>) and individual microscans per MS/MS features were merged and an intensity threshold applied to remove noise peaks. MS/MS spectra were then exported as MSP, CSV and in JSON format with associated structures in SMILES and InChiKey, adduct information, ion mode and additional meta data such as precursor mass error.

For the purposes of computing metrics in this paper we restricted to [M-H]<sup>-</sup>, [M+H]<sup>+</sup>, [M+NH<sub>4</sub>]<sup>+</sup>, [M+Na]<sup>+</sup>, [M+Cl]<sup>-</sup>, [M+K]<sup>+</sup>, and [M-H<sub>2</sub>O+H]<sup>+</sup> adducts, which are the standard adducts for the SIRIUS software. Spectra were gathered in both positive and negative ion mode, at collision energies 20, 40, and 60 eV.

A subset of 226 of these compounds were not found in the main structural libraries Pubchem and COCONUT ([11, 5]). These compounds were used for the evaluation of dark chemical space.

## 4.2 CSI:FingerID

For all CSI:FingerID results, we used SIRIUS version 5.6.3 (Bright Giant, GmbH, Jena, Germany). Query data was prepared as an .mgf file. Each spectrum contained a separate entry for MS1 (containing isotope masses and intensities gathered internally through Enveda's feature calling software) spectra and associated MS/MS spectra. In keeping with all other results, MS/MS spectra were limited to the most intense 256 fragments. Each molecule was represented by at most one spectrum from each available (collision energy, adduct) pair. Both positive and negative ion mode spectra were included. Spectra were evaluated separately, not merged across ion modes, collision energies, or adducts. No timeout was used.

SIRIUS was set to use DB formulas from the Bio Database only, with all possible default

ionizations. Instrument was Q-TOF with a MS2 mass accuracy of 10ppm. All other settings were default for batch compute. CSI:FingerID predictions were calculated with Fallback adducts including [M-H]-, [M+H]+, [M+NH4]+, [M+Na]+, [M+Cl]-, [M+K]+, and [M-H2O+H]+ and search DBs was set to the Bio Database. Spectra for which CSI:FingerID returned no predicted structures were counted as not passing the requisite Tanimoto similarity threshold for the purposes of calculating accuracy.

### 4.3 MS2Mol training

We use the Hugging Face implementation of the facebook/bart-base model [25, 44]. The model is trained using the Adam optimizer [45] and teacher forcing [46]. We train the model for 20 epochs with a batch size of 512 and a constant learning rate of 5e-5.

In order to avoid biasing the model to predict the structures that are over-represented in our dataset (i.e. have many spectra associated with them), we weight each training example inversely proportional to the number of times each structure occurs in the training set.

While the precursor m/z provides important signal pertaining to the molecular formula and structure of the underlying compound, early experiments with prepending the precursor m/z to the input found that the model can become too reliant on it while ignoring information contained in the fragmentation spectra. We obtained our best results by randomly including the precursor m/z as part of the input 50% of the time and requiring the model to predict the structure without it otherwise. At inference time we always include the precursor m/z as part of the input.

To tune the hyperparameters of the model, we used random hyperparameter search [47] as implemented in the Ray Tune library [48]. Details of the search space and parameters chosen are in the Appendix.

See the Appendix for ablation studies on model design and dataset composition.

### 4.4 Reranker model training

The reranker model is a gradient boosted machine [40] that was trained as a “learning to rank” model with a normalized discounted cumulative gain (ndcg) loss [49] using the xgboost package [50]. To train the reranker model, we split the overall MS2Mol training dataset into two structure-disjoint subsets, call them A and B. Set B consisted of publicly available spectra corresponding to 11,602 structures and internally-generated spectra corresponding to 667 structures, for a total of 95,949 spectra. We first trained a version of MS2Mol on set A, and then used it to generate structure predictions on set B. Set B was then used to train the reranker model. Reranker hyperparameters were selected via 100 iterations of random search, using 5% of set B as a validation set, and the average Tanimoto similarity of the top-ranked prediction to the true structure as the validation metric. See the Appendix for further details about this training process.

## 4.5 Confidence model training

The confidence model is likewise a gradient boosted machine [40] that was trained as a regression model to predict the Tanimoto similarity between the true structure and the MS2Mol prediction. The training set and hyperparameter search space were the same as those used to train the reranker model. See Section F in the Appendix for more details.

## 4.6 Inference

At inference time we use a beam search decoding strategy with 25 beams, beyond which performance no longer improves. The model outputs a sequence of tokens that represent the predicted molecular formula and then the predicted structure. The beam search gives us a ranked list of 25 possible structures, encoded as BPE tokens of molecular “subwords” [29]. We decode these to deepSMILES [51], decode the deepSMILES to SMILES [22], and then use RDKit [37] to check if the SMILES strings correspond to valid molecules. We then employ the reranker model to order the 25 candidate structures, throwing out any invalid SMILES predictions. The confidence model is then applied to the top-ranked prediction.

MS2Mol supports additional functionality that we do not evaluate in the main text. When the molecular formula is known at inference time, one can optionally require the beam search to output the correct molecular formula. Since each successive output token is conditioned on the previous outputs, this makes the structure prediction conditional on the given molecular formula and can improve predictive accuracy. Although we do not utilize this functionality in our main results, we do show the effect of formula-conditioned structure prediction in Section D of the Appendix.

## 5 Acknowledgements

We gratefully acknowledge Martin Hoffmann, Marcus Ludwig (Bright Giant GmbH, Jena, Germany) and Kai Duhrkop for answering our questions and assisting with benchmarking CSI:FingerID, and Connor Coley and Rafael Gomez-Bombarelli for useful discussions. We also acknowledge the Enveda Chemistry team for annotations.

## References

- [1] Ricardo R da Silva, Pieter C Dorrestein, and Robert A Quinn. Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences*, 112(41):12549–12550, 2015.
- [2] Atanas G Atanasov, Sergey B Zotchev, Verena M Dirsch, and Claudiu T Supuran. Natural products in drug discovery: advances and opportunities. *Nature reviews Drug discovery*, 20(3):200–216, 2021.
- [3] Farit Mochamad Afendi, Taketo Okada, Mami Yamazaki, Aki Hirai-Morita, Yukiko Nakamura, Kensuke Nakamura, Shun Ikeda, Hiroki Takahashi, Md Altaf-Ul-Amin, Latifah K Darusman, et al. Knapsack family databases: integrated metabolite-plant

- species databases for multifaceted plant research. *Plant and Cell Physiology*, 53(2):e1–e1, 2012.
- [4] Shiva Abdollahi Aghdam and Amanda May Vivian Brown. Deep learning approaches for natural product discovery from plant endophytic microbiomes. *Environmental microbiome*, 16(1):1–20, 2021.
  - [5] Maria Sorokina, Peter Merseburger, Kohulan Rajan, Mehmet Aziz Yirik, and Christoph Steinbeck. Coconut online: collection of open natural products database. *Journal of Cheminformatics*, 13(1):1–13, 2021.
  - [6] John J Irwin, Khanh G Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R Wong, Munkhzul Khurelbaatar, Yurii S Moroz, John Mayfield, and Roger A Sayle. Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of chemical information and modeling*, 60(12):6065–6073, 2020.
  - [7] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapon, Tal Luzzatto-Knaan, et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature biotechnology*, 34(8):828–837, 2016.
  - [8] Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H Spaaks, Faruk Diblen, Simon Rogers, and Justin JJ Van Der Hooft. Spec2vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS computational biology*, 17(2):e1008724, 2021.
  - [9] Florian Huber, Sven van der Burg, Justin JJ van der Hooft, and Lars Ridder. Ms2deepscore: a novel deep learning similarity measure to compare tandem mass spectra. *Journal of cheminformatics*, 13(1):84, 2021.
  - [10] Gemady Voronov, Rose Lightheart, Joe Davison, Christoph A Kretzler, David Healey, and Thomas Butler. Multi-scale sinusoidal embeddings enable learning on high resolution mass spectrometry data. *arXiv preprint arXiv:2207.02980*, 2022.
  - [11] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.
  - [12] Fei Wang, Jaanus Liigand, Siyang Tian, David Arndt, Russell Greiner, and David S Wishart. Cfm-id 4.0: more accurate esi-ms/ms spectral prediction and compound identification. *Analytical chemistry*, 93(34):11692–11700, 2021.
  - [13] Liu Cao, Mustafa Guler, Azat Tagirdzhanov, Yi-Yuan Lee, Alexey Gurevich, and Hosein Mohimani. Moldiscovery: learning mass spectrometry fragmentation of small molecules. *Nature Communications*, 12(1):1–13, 2021.
  - [14] Sebastian Wolf, Stephan Schmidt, Matthias Müller-Hannemann, and Steffen Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC bioinformatics*, 11:1–12, 2010.

- [15] Samuel Goldman, John Bradshaw, Jiayi Xin, and Connor W Coley. Prefix-tree decoding for predicting mass spectra from molecules. *arXiv preprint arXiv:2303.06470*, 2023.
- [16] Samuel Goldman, Janet Li, and Connor W Coley. Generating molecular fragmentation graphs with autoregressive neural networks. *arXiv preprint arXiv:2304.13136*, 2023.
- [17] Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu, and Sebastian Böcker. Searching molecular structure databases with tandem mass spectra using csi: Fingerid. *Proceedings of the National Academy of Sciences*, 112(41):12580–12585, 2015.
- [18] Oliver Fiehn. Critical Assessment of Small Molecule Identification 2022. <http://http://www.casmi-contest.org/2022/index.shtml/>, 2022. [Online; accessed 12-December-2022].
- [19] Samuel Goldman, Jeremy Wohlwend, Martin Stražar, Guy Haroush, Ramnik J Xavier, and Connor W Coley. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *bioRxiv*, pages 2022–12, 2022.
- [20] Martin A Hoffmann, Louis-Félix Nothias, Marcus Ludwig, Markus Fleischauer, Emily C Gentry, Michael Witting, Pieter C Dorrestein, Kai Dührkop, and Sebastian Böcker. High-confidence structural annotation of metabolites absent from spectral libraries. *Nature Biotechnology*, 40(3):411–421, 2022.
- [21] Michael A Stravs, Kai Dührkop, Sebastian Böcker, and Nicola Zamboni. Msnovelist: De novo structure generation from mass spectra. *Nature Methods*, pages 1–6, 2022.
- [22] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.
- [23] Aditya Divyakant Shrivastava, Neil Swainston, Soumitra Samanta, Ivayla Roberts, Marina Wright Muelas, and Douglas B Kell. Massgenie: A transformer-based deep learning method for identifying small molecules from their mass spectra. *Biomolecules*, 11(12):1793, 2021.
- [24] Eleni Litsa, Vijil Chenthamarakshan, Payel Das, and Lydia Kavradi. Spec2mol: An end-to-end deep learning framework for translating ms/ms spectra to de-novo molecules. 2021.
- [25] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.
- [26] Justin Johan Joias van Der Hooft, Joe Wandy, Michael P Barrett, Karl EV Burgess, and Simon Rogers. Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences*, 113(48):13738–13743, 2016.
- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

- [28] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.
- [29] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, 2016.
- [30] SS Mehta. Massbank of north america (mona): An open-access, auto-curating mass spectral database for compound identification in metabolomics presentation, 2020.
- [31] Yuji Sawada, Ryo Nakabayashi, Yutaka Yamada, Makoto Suzuki, Muneo Sato, Akane Sakata, Kenji Akiyama, Tetsuya Sakurai, Fumio Matsuda, Toshio Aoki, et al. Riken tandem mass spectral database (respect) for phytochemicals: a plant-specific ms/ms-based data resource and database. *Phytochemistry*, 82:38–45, 2012.
- [32] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, et al. Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry*, 45(7):703–714, 2010.
- [33] Anzor Mikaia, Principal Edward White V EI, Vladimir Zaikin EI, Damo Zhu EI, O David Sparkman EI, Pedatsur Neta, Igor Zenkevich RI, Peter Linstrom, Yuri Mirokhin, Dmitrii Tchekhovskoi, et al. Nist standard reference database 1a. *Standard Reference Data, NIST, Gaithersburg, MD, USA* <https://www.nist.gov/srd/nist-standard-reference-database-1a>, 2014.
- [34] Colin A Smith, Grace O’Maille, Elizabeth J Want, Chuan Qin, Sunia A Trauger, Theodore R Brandon, Darlene E Custodio, Ruben Abagyan, and Gary Siuzdak. Metlin: a metabolite mass spectral database. *Therapeutic drug monitoring*, 27(6):747–751, 2005.
- [35] Adriano Rutz, Maria Sorokina, Jakub Galgonek, Daniel Mietchen, Egon Willighagen, Arnaud Gaudry, James G Graham, Ralf Stephan, Roderic Page, Jiří Vondrášek, et al. The lotus initiative for open knowledge management in natural products research. *Elife*, 11:e70780, 2022.
- [36] Pierre-Marie Allard, Tiphaine Péresse, Jonathan Bisson, Katia Gindro, Laurence Marcourt, Van Cuong Pham, Fanny Roussi, Marc Litaudon, and Jean-Luc Wolfender. Integration of molecular networking and in-silico ms/ms fragmentation for natural products dereplication. *Analytical chemistry*, 88(6):3317–3323, 2016.
- [37] Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8, 2013.
- [38] Florian Huber, Stefan Verhoeven, Christiaan Meijer, Hanno Spreeuw, Efraín Castilla, Cunliang Geng, Justin JJ van der Hooft, Simon Rogers, Adam Belloum, Faruk Diblen, et al. matchms-processing and similarity evaluation of mass spectrometry data. *Journal of Open Source Software*, 5(52), 2020.



- [39] Hyun Woo Kim, Mingxun Wang, Christopher A Leber, Louis-Félix Nothias, Raphael Reher, Kyo Bin Kang, Justin JJ Van Der Hooft, Pieter C Dorrestein, William H Gerwick, and Garrison W Cottrell. Npclassifier: A deep neural network-based structural classification tool for natural products. *Journal of Natural Products*, 84(11):2795–2807, 2021.
- [40] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [41] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- [42] Michael Murphy, Stefanie Jegelka, Ernest Fraenkel, Tobias Kind, David Healey, and Thomas Butler. Efficiently predicting high resolution mass spectra with graph neural networks. *arXiv preprint arXiv:2301.11419*, 2023.
- [43] Christoph A Kretzler and Gerhard G Thallinger. A map of mass spectrometry-based in silico fragmentation prediction and compound identification in metabolomics. *Briefings in Bioinformatics*, 22(6):bbab073, 2021.
- [44] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.
- [45] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015.
- [46] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.
- [47] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.
- [48] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.
- [49] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, oct 2002.
- [50] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.
- [51] Noel O’Boyle and Andrew Dalke. Deepsmiles: an adaptation of smiles for use in machine-learning of chemical structures. 2018.
- [52] Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino. “found in translation”: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science*, 9(28):6091–6098, 2018.
- [53] Gennady Voronov, Abe Frandsen, Brian Bargh, David Healey, Rose Lightheart, Tobias Kind, Pieter Dorrestein, Viswa Colluru, and Thomas Butler. Ms2prop: A machine

learning model that directly predicts chemical properties from mass spectrometry data for novel compounds. *bioRxiv*, pages 2022–10, 2022.

- [54] Yuanyue Li, Tobias Kind, Jacob Folz, Arpana Vaniya, Sajjan Singh Mehta, and Oliver Fiehn. Spectral entropy outperforms ms/ms dot product similarity for small-molecule compound identification. *Nature Methods*, 18(12):1524–1531, 2021.
- [55] G Richard Bickerton, Gaia V Paolini, J  r  my Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.
- [56] Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1(1):1–11, 2009.
- [57] Steven H Bertz. The first general index of molecular complexity. *Journal of the American Chemical Society*, 103(12):3599–3601, 1981.
- [58] Scott A Wildman and Gordon M Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences*, 39(5):868–873, 1999.

## A Datasets and Code Availability

Code for training and deploying MS2Mol, a version of MS2Mol trained on open data, and the EnvedaDark and EnvedaKnown benchmarking datasets will be made publicly available via github repository at the time of publication. We hope the community will find these datasets and the model useful for benchmarking and research purposes.

## B Disclosures

All authors except PCD and KJM are employed by Enveda Biosciences, PCD is an advisor to Cybele, consulted for MSD animal health in 2023, and he is a Co-founder and scientific advisor for Ometa Labs, Arome, and Enveda with prior approval by UC San Diego. This work is part of Enveda’s drug discovery platform.

## C Ablations

We explore the impact of different architecture and dataset designs on MS2Mol. To facilitate experiments, we use the un-reranked MS2Mol model trained on our full training dataset as the baseline, and all ablation models are similarly evaluated without the final reranking step. For each ablation, we evaluate model performance on three test sets and report the fraction of predicted structures whose Tanimoto similarity to the correct structure exceeds 0.4 and 0.675. See Table 2.

| Ablation               | EnvedaDark  | EnvedaLight | CASMI       |
|------------------------|-------------|-------------|-------------|
| MS2Mol                 | 0.62 / 0.21 | 0.73 / 0.37 | 0.64 / 0.36 |
| MS2Mol_un-reranked     | 0.64 / 0.19 | 0.70 / 0.35 | 0.62 / 0.32 |
| no_precursor_masking   | 0.57 / 0.15 | 0.68 / 0.33 | 0.57 / 0.27 |
| full_precursor_masking | 0.58 / 0.20 | 0.69 / 0.35 | 0.59 / 0.30 |
| unweighted_loss        | 0.62 / 0.17 | 0.70 / 0.34 | 0.61 / 0.31 |
| single_m/z_token       | 0.62 / 0.21 | 0.71 / 0.35 | 0.61 / 0.31 |
| no_formula_prepending  | 0.58 / 0.22 | 0.70 / 0.36 | 0.57 / 0.31 |
| no_bpe                 | 0.06 / 0.00 | 0.08 / 0.00 | 0.09 / 0.02 |
| no_internal            | 0.53 / 0.15 | 0.65 / 0.26 | 0.63 / 0.30 |
| no_lotus               | 0.62 / 0.17 | 0.69 / 0.35 | 0.61 / 0.31 |
| no_lotus_no_internal   | 0.50 / 0.12 | 0.63 / 0.26 | 0.63 / 0.29 |

Table 2: MS2Mol ablation evaluations on three test sets. For each ablation and test set, two numbers are reported: the fraction of predictions with Tanimoto similarity to the true structure greater than 0.4 and 0.675.

**ranked vs. un-reranked** The top two lines of Table 2 compare MS2Mol with and without the final reranking step. On five out of the six metrics, we see that the reranker provides improved predictions. As mentioned above, each subsequent ablation model is evaluated without reranking, and so should be compared to MS2Mol\_un-reranked.

**no\_precursor\_masking** For this ablation, we set the masking probability of the precursor m/z tokens to 0, so that the model can always access these tokens during both training and inference. This can potentially lead to the model over-fitting to the precursor m/z or even memorizing which structure in the training set is associated with a given m/z. Such over-fitting is especially problematic when the test set contains novel compounds that have similar masses to compounds in the training set. Indeed, we see that this ablation performs uniformly worse than the baseline on our evaluation sets.

**full\_precursor\_masking** Here we set the masking probability of the precursor m/z tokens to 1 for both training and inference, so that the model never has access to these tokens, and must predict the structure entirely from the MS/MS spectrum peaks. One would expect this to negatively impact model performance, since the precursor m/z is a key attribute of the structure. On the other hand, the precursor m/z can show up as a peak in the MS/MS spectrum, and so is sometimes redundant. We observe that fully masking the precursor m/z negatively impacts model performance on all metrics except for the 0.675 Tanimoto similarity threshold for EnvedaDark and EnvedaLight. This suggests that the baseline model may still be somewhat over-fitting to the precursor m/z, which is especially impactful when trying to identify truly novel compounds that aren’t found in the training set. Note also that fully masking out the precursor m/z leads to uniformly better performance than never masking it during training.

**unweighted\_loss** The baseline MS2Mol model weights the training loss for each MS/MS spectrum in the training set according to the inverse frequency of its molecular structure in the training set, which is a common technique for addressing class imbalance in prediction tasks. For this ablation, we weight the training loss for each spectrum equally, thereby allowing highly-represented compounds in the training set to exert more influence on the model updates during training. We observe that this unweighted loss function leads to a model with slightly worse performance than the baseline, confirming that addressing class imbalance in the training set is beneficial.

**single\_m/z\_token** In this ablation, we represent each m/z (including the precursor m/z and the MS/MS peaks) with a single token corresponding to its value rounded to a single decimal place. By contrast, the baseline MS2Mol model represents each m/z value with two tokens, one for its integer part and one for its fractional part. The single-token representation leads to a larger vocabulary and potentially sparser coverage for each token in the training set: for example, there may be only a handful of MS/MS spectra containing a rounded m/z of 513.3, which may hinder the model's ability to learn a robust embedding for this token. On the other hand, a single token representation does reduce the sequence length for the MS/MS spectra, which can improve computational speed. We didn't observe meaningful computational performance differences between the two tokenization schemes. However, the single-token representation does lead to slightly better performance compared to the baseline for 2 out of the 6 reported metrics. The optimal way to represent MS/MS spectra in transformer models remains an interesting open problem.

**no\_formula\_prepending** For this ablation, we remove the molecular formula tokens from the label sequence during training. This means the training signal is derived only from the (tokenized) SMILES strings rather than from both the molecular formulas and SMILES strings. Removing the molecular formula label also precludes molecular-formula-conditional inference, as discussed in Section D. It's not clear a priori whether including the molecular formula in the label at train time will benefit model performance beyond enabling conditional inference. On the one hand, predicting the molecular formula is an easier task than predicting the full molecular structure, and a common intermediate step when trying to elucidate the structure, so guiding the model to first predict the formula and then further predict the full SMILES string is somewhat natural. On the other hand, it's also sensible to align the training process as closely as possible with the inference task, which in this case is simply structure prediction. We observe that removing the molecular formula tokens from the training labels improves model performance on 2 out of the 6 reported metrics. This suggests that either choice is viable, but as we elaborate in Section D, our baseline MS2Mol design enables additional functionality that can be beneficial in certain scenarios.

**no\_bpe** It has become standard practice in NLP and beyond to employ byte-pair encoding (BPE) and related tokenization schemes in conjunction with transformer models, as we do in our MS2Mol model. For this ablation, we remove the BPE step and represent the SMILES strings as simple single-atom token sequences. In particular, we split each SMILES string using a simple regular expression given in [52]. We do not convert the SMILES strings to

deepSMILES in this case. We find that omitting BPE causes a drastic drop in performance. In many cases, no valid SMILES strings are generated, and in the other cases, they are usually not structurally similar to the true compound. We leave it to future work to investigate what drives this difference.

**no\_internal** In this ablation, we remove the 33,634 internally-generated spectra from the training set. As expected, this significantly degrades model performance on EnvedaDark and EnvedaLight, both of which are composed of similarly internally-generated spectra. For CASMI, the close-match accuracy is higher when training on internal data. These results confirm the intuitive point that the model performs better for spectra that have experimentally-similar examples in the training set. Note that the 33,634 internal spectra we include are only a small fraction of the entire training set, but the model is still able to effectively learn how to generalize on these spectra while not compromising performance on unrelated test sets like CASMI.

**no\_lotus** Similar to the no\_internal ablation, here we remove the in-silico LOTUS spectra from the training set. Since these spectra are not experimental, we don't expect them to be very spectrally similar to the test sets. Rather, the likely benefit of the LOTUS spectra is to increase the chemical representation in the training set and enable the model to learn to generate a larger diversity of compounds. We find that omitting LOTUS indeed degrades model performance slightly across the board.

**no\_lotus\_no\_internal** This ablation simply excludes both the in-silico LOTUS spectra and the internal-generated experimental spectra from the training set. In other words, this model is trained only on publicly and commercially available experimental MS/MS spectra. We see that this model performs worse by a decent margin compared to the baseline on five out of six metrics. When comparing this model to the no\_internal model, we see that it performs somewhat worse on all evaluation sets, particularly EnvedaDark. This shows the benefit of using in-silico spectra to augment the training set, especially when the test set is quite different from the experimental train set structurally and experimentally.

## D Molecular formula conditioning

In many cases, it is possible to accurately predict the molecular formula for a given MS/MS spectrum. MS2Mol can utilize this partial information by generating a SMILES string conditioned on both the MS/MS spectrum and the molecular formula. This functionality is possible due to our inclusion of the molecular formula tokens in the label sequence during training along with the autoregressive generation process of transformer models. Table 3 shows the results of conditioning MS2Mol predictions on the true molecular formula. Such conditional generation leads to significantly more accurate predictions across the board. We emphasize that in this experiment, we are using the true molecular formula. In practice, one must first predict the molecular formula, and conditioning on inaccurate formula predictions can degrade the structure predictions.

| True Formula Conditioning | EnvedaDark  | EnvedaLight | CASMI       |
|---------------------------|-------------|-------------|-------------|
| no                        | 0.64 / 0.19 | 0.70 / 0.35 | 0.62 / 0.32 |
| yes                       | 0.70 / 0.23 | 0.75 / 0.43 | 0.68 / 0.36 |

Table 3: Benefit of conditioning MS2Mol predictions on ground truth molecular formulas. For each test set, two numbers are reported: the fraction of predictions with Tanimoto similarity to the true structure greater than 0.4 and 0.675.

| Parameter           | Selected value        | Candidate range       |
|---------------------|-----------------------|-----------------------|
| max_input_peaks     | 256                   | [64, 128, 256]        |
| structure_bpe_vocab | 256                   | [64, 128, 256]        |
| learning_rate       | $2.82 \times 10^{-4}$ | $[10^{-5} - 10^{-3}]$ |
| weight_decay        | 0.0                   | [0.0, 0.01, 0.1]      |
| batch_size          | 1024                  | [512, 1024]           |
| d_model             | 768                   | [384, 768]            |
| encoder_layers      | 6                     | [3, 6]                |
| decoder_layers      | 3                     | [3, 6]                |
| encoder_ffn_dim     | 3072                  | [512, 1024, 3072]     |
| decoder_ffn_dim     | 3072                  | [512, 1024, 3072]     |
| encoder_attn_heads  | 12                    | [6, 12]               |
| decoder_attn_heads  | 6                     | [6, 12]               |
| dropout             | 0.2                   | [0.1, 0.15, 0.2]      |

Table 4: Hyperparameters and hyperparameter search space for the main model.

## E Hyperparameters

The architecture of our model is taken from the NLP literature. However, the domain of MS/MS spectra and molecular structures is significantly different from that of natural language. As such, it is necessary to adapt the model hyperparameters to our task. While there are many potential hyperparameters to tune in the BART model, we considered the search space shown in Table 4, which also lists the selected parameters. The algorithm used for hyperparameter search is described in Section 4.

## F Reranker and confidence models

The reranker and confidence models utilize a suite of features derived from the MS2Mol prediction and model outputs, the original input MS/MS spectrum, as well as structural property predictions calculated directly from the MS/MS spectrum as in [53].

The features based on MS2Mol model outputs are as follows:

- log probability of the generated sequence



- beam rank of the predicted structure
- length of predicted SMILES string
- molecular weight of predicted structure

The features based on the input MS/MS spectrum are as follows:

- difference and squared difference between the experimental precursor  $m/z$  and the mass of the predicted structure, and whether this difference matches a known adduct
- spectral entropy of the MS/MS spectrum [54]
- number of peaks in the MS/MS spectrum

In addition to the above features, we also make use of a separate model from [53] that accurately and directly predicts molecular properties (but not the full molecular structure) from the MS/MS spectrum. We emphasize that the property prediction model used here was trained solely on publicly and commercially available data, and never had access to internally-generated data. This property prediction model can often produce predictions quite close to the corresponding properties of the ground-truth molecule, and so it provides a valuable indication of the accuracy of the MS2Mol structure prediction. To utilize this fact, we include as features the difference and squared difference between the MS2Mol structure prediction value and the directly-predicted value, along with the raw MS2Mol structure prediction value, for the following properties:

- polar surface area
- quantitative estimate of drug likeness [55]
- synthetic accessibility score [56]
- BertzCT complexity [57]
- atomic logP [58]
- number of hydrogen bond acceptors
- fraction of  $sp^3$  hybridized carbons
- number of hetero atoms
- number of rotatable bonds
- number of aliphatic rings
- number of aromatic rings

To train the confidence and reranker models, we used the hyperparameters and search space defined in Table 5. Hyperparameters for the confidence model were evaluated by the chosen `eval_metric` while the reranking model hyperparameters were evaluated by reranking on a small validation set and measuring the average Tanimoto similarity to ground-truth of the top-ranked predictions. The confidence model was trained with a mean squared error loss

| Parameter        | Confidence value | Reranker value | Candidate range           |
|------------------|------------------|----------------|---------------------------|
| learning_rate    | .1               | .01            | [.01, .05, .1, .2]        |
| colsample_bytree | .8               | .4             | [.1, .2, .4, .6, .8, 1.0] |
| max_depth        | 12               | 8              | [2,4,8,12,16],            |
| subsample        | .95              | .75            | [.65,.75,.85,.95]         |
| eval_metric      | 'rmse'           | 'mae'          | ['mae', 'rmse']           |

Table 5: Hyperparameters and hyperparameter search space for the confidence and reranker models.

while the reranker model was trained with a normalized discounted cumulative gain (ndcg) loss [49].

## G CASMI dataset preparation

The CASMI dataset we use in this work is extracted from CASMI 2022 trial <sup>1</sup>. The original dataset is given as a collection of mzML files containing MS/MS runs collected in both positive and negative ionization mode and under three collision energies. For each of 500 selected compounds, the retention time, precursor m/z, and mzML file are given. To extract the MS/MS spectrum for a given compound, we search in the specified mzML file for all MS/MS spectra whose precursor m/z is within 0.005 Da of the given value and whose retention time is within 0.1 minutes of the given value, and then select the matching spectrum with the highest total intensity. These spectra are then further processed in the same way as our other datasets.

We also extract the MS1 isotope envelopes for the 500 compounds as follows. For each compound, extract from the specified mzML file the MS1 spectrum corresponding to its MS/MS spectrum, then identify the monoisotopic peak by selecting the most intense peak within a window of 0.05 Da around the given precursor m/z, and finally search for up to three additional isotopic peaks at  $1 \times iso\_gap$ ,  $2 \times iso\_gap$ , and  $3 \times iso\_gap$  Da greater than the precursor m/z (within a tolerance of 0.05 Da), where *iso\_gap* is 1.003354835 divided by the charge of the precursor ion. These isotopic envelopes are given as input for the CSI:FingerID benchmark on CASMI.

## H Annotation tool for molecular similarity annotations and sample annotations

The annotation tool for molecular similarity is shown in Figure 5. We also show a sample of 25 chemist annotations of similarity across diverse molecules and annotation values, including a couple of random predictions in Figure 6.

<sup>1</sup><https://fiehnlab.ucdavis.edu/casmi/casmi-2022-results>

### SpecBart Prediction Annotator

Each example includes our model's prediction of the structure (left) as well as the true structure (right). The purpose of this annotation exercise is solely to evaluate the accuracy of our predictions, NOT to evaluate the quality of the compounds.

Many of the predictions are from our models, but a fraction of them are from a competitor model, and a fraction are simply random compounds that are completely unrelated to the true compound. You will not know the source of the predictions.

Please score predictions using the 1-3 scale below according to the estimated usefulness of this prediction were it to be used in your work:

1. Not useful or misleading for making decisions about what to do about the molecule
2. Useful, but some material errors. Knowing only this and making decisions about prioritizing a molecule will still materially help a chemist make the right decision about the molecule, for the right reasons.
3. This structure is a quite reliable guide to making choices about prioritizing the molecule, and is close in the ways that count for our decisions.

☒ Show smiles
 ☐ Show tanimoto (keep hidden if annotating)

I'm not an annotator - just looking (nothing will be recorded)

| Example # | Score (1-3) |
|-----------|-------------|
| 0         | 1           |

☐ I want to make specific comments about this example (optional)
 

Submit

**Example 0 Prediction:**  
Fc1cc(F)c2ccnc2c1Cl

**Example 0 Ground truth:**  
Fc1ccc2nccc(Cl)c2c1

Figure 5: The annotation tool, including annotation guidelines. Annotations were over a broad range of molecules, sampled from the large dataset of publicly available spectral libraries.



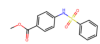
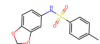




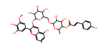
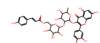




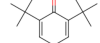
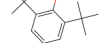




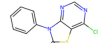
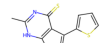




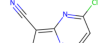
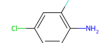




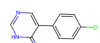
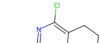




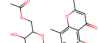
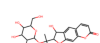




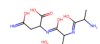
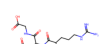




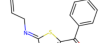
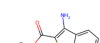




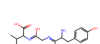
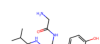


| Annotation: Not similar (1)   |      |   | Annotation: Meaningful match (2)  |      |   | Annotation: Close match (3)   |      |   |
|---|------|---|---|------|---|---|------|---|
|    | 0.19 |    |    | 0.52 |    |    | 0.72 |    |
|    | 0.65 |    |    | 0.86 |    |    | 0.96 |    |
|    | 0.18 |    |    | 0.1  |    |    | 0.41 |    |
|    | 0.09 |    |    | 0.32 |    |    | 1.0  |    |
|    | 0.11 |    |    | 0.09 |    |    | 0.68 |    |
|  | 0.36 |  |  | 0.17 |  |  | 0.87 |  |
|  | 0.13 |  |  | 0.6  |  |  | 1.0  |  |
|  | 0.2  |  |  | 0.68 |  |  | 0.71 |  |
|  | 0.13 |  |  | 0.25 |  |  | 0.89 |  |
|  | 0.52 |  |  | 0.83 |  |  | 0.77 |  |

Figure 6: 30 example annotations with 10 randomly sampled from each possible annotation score. Within the columns, ground truth is on the left, the predicted structure is on the right, and the Tanimoto similarity is in between.

# I Tanimoto threshold derivations from Chemist annotations

To arrive at the definitions of “meaningful” match and “close match” from chemist annotations, we scored predictions of chemical structures as a match if

$$\text{tanimoto\_sim}(\text{fp}_{\text{pred}}, \text{fp}_{\text{groundtruth}}) > t \quad (1)$$

where  $t$  is a real number between 0 and 1. Exact matches require agreement between the first 14 characters of the InChiKeys. Equation 1 can be used to predict the annotations of chemists shown above based on Tanimoto similarity.

To select the values of  $t$ , we chose a threshold where the rate of positive predictions from the similarity threshold model of Equation 1 matches the rate of positive annotations from the chemists. This choice coincides with selecting a threshold where precision and recall are equal. This follows from using the values of precision and recall as estimated probabilities,

$$\text{Prec} = \text{Pr}(C = P | \hat{C} = P) \quad (2)$$

$$\text{Rec} = \text{Pr}(\hat{C} = P | C = P) \quad (3)$$

where  $C$  is the true class (positive or negative, e.g. a close match or not as annotated by chemists),  $P$  indicates positives, and  $\hat{C}$  indicates the estimate of the predictor. In this case the prediction is a positive if the similarity of the molecules exceeds the threshold  $t$ . Therefore,  $\hat{C}$  is a function of the threshold  $t$ . The probabilistic definitions of precision and recall are then connected by Bayes’ theorem,

$$\text{Pr}(C = P | \hat{C} = P) = \frac{\text{Pr}(\hat{C} = P | C = P)P(C = P)}{P(\hat{C} = P)} \quad (4)$$

It then follows immediately from Equation 4 that to match the rate of positive predictions from the model in Equation 1 to the rate of positive annotations from chemist annotations, the threshold  $t$  must be chosen such that precision and recall are equal.

We chose two values of the parameter  $t$  corresponding to two success criteria (Figure 7). The first criterion was whether the prediction was likely to be scored as a “3” (Close match). The second criterion was whether the prediction was likely to be scored as a “2” (meaningfully similar) or greater. We bootstrapped by annotator, and then by annotation to get estimates of error. For the “close match”, or “3” annotation criterion, we selected the precision-recall cross over point, rounded to 0.025. The result is the 0.675 threshold reported in the main text.

For “meaningful match”, or “2+” annotations, we took a more cautious approach. This is because for lower thresholds, errors are much more likely to be molecules that would be annotated as “1”, or actually misleading (see Figure 7). We therefore dropped all annotators who annotated any random examples as >1, and selected the upper confidence band of the 95% bootstrap confidence interval as the resulting threshold, and rounded to the nearest 0.1. This provides a conservative estimate of the rate of meaningful matches, while maximally

| criterion    | threshold | precision | recall | “1” errors |
|--------------|-----------|-----------|--------|------------|
| annotation=3 | 0.675     | 0.79      | 0.81   | 0.004      |
| annotation>1 | 0.4       | 0.92      | 0.71   | 0.08       |

Table 6: Metrics for the selected similarity thresholds for meaningful and close match. Precision, recall, and rate of “1” errors, which are the identification of predictions as successes that the annotators label as meaningless or misleading

suppressing the particularly impactful error of identifying a prediction as meaningful match when it is in fact meaningless or actually misleading. Metrics for both thresholds are in Table 6

We finally note that the Tanimoto similarity thresholds for close and meaningful matches cannot be directly compared to similarity thresholds for analog search in databases of molecules, even with the same underlying fingerprint. This observation emerges from analyzing the similarities and annotation scores of random pairs of unrelated molecules that were inserted as controls in our annotation exercise. As described above, annotators were shown pairs of molecules and asked to rate their similarity as in Figure 5. Approximately 10% of those pairs were random control pairs. We separated the control pairs from the pairs where one of the pair was a prediction of the other based on MS/MS data. We then analyzed how the average annotator rating changed as a function of Tanimoto similarity for both groups. We found that for a given Tanimoto similarity, annotators scored the predictions meaningfully higher than they scored the control pairs (Figure 7). These results suggest that Tanimoto similarity does not capture all aspects of similarity that are meaningful to practicing chemists and are captured in the annotations. It also indicates that the similarity thresholds derived in the present article should not be expected to generalize perfectly to other contexts.



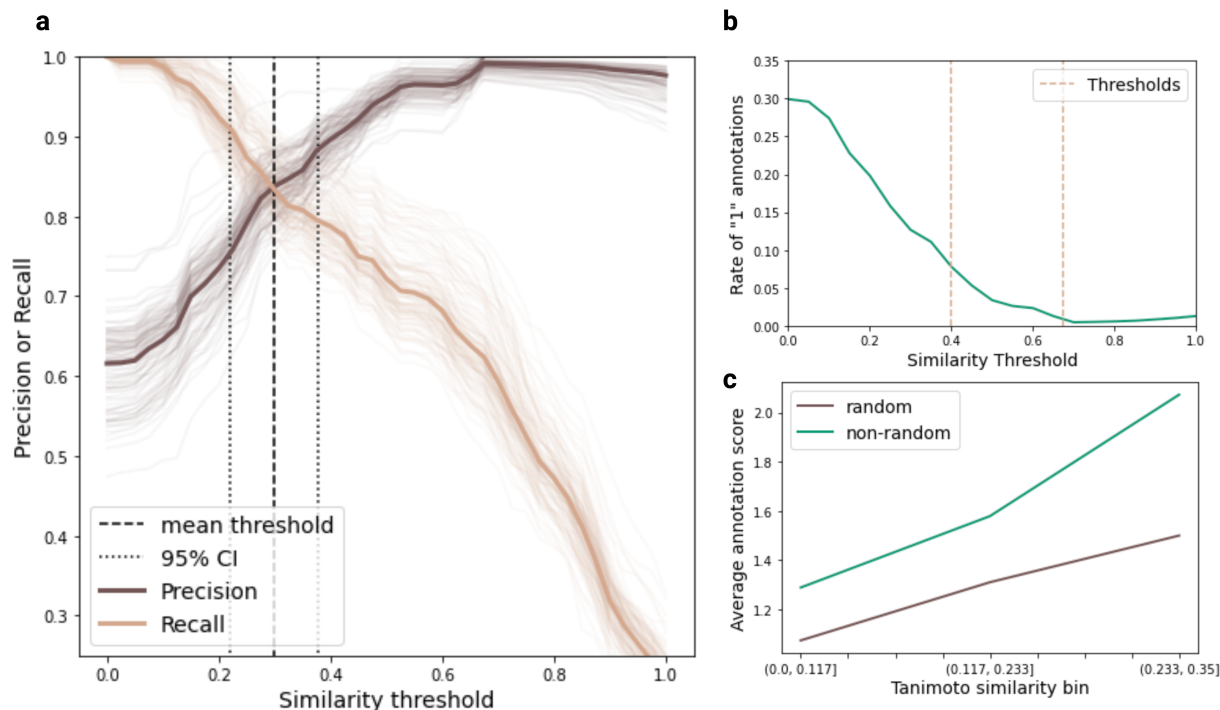


Figure 7: a) Visualization of the precision and recall of Tanimoto similarity based predictions of chemist annotations as a function of the Tanimoto threshold. The light colored lines correspond to the annotator and annotation level bootstrap analysis described in the text, while the heavy lines correspond to the primary sample. The vertical lines correspond to the optimal threshold computations as well as the 95% confidence intervals. We select the rounded upper confidence level for the threshold. b) The rate at which annotations scored as "1", or misleading/incorrect occur as a function of Tanimoto similarity threshold. For each threshold, this is the rate at which they are completely incorrect, and score predictions that are meaningless as meaningful matches or close matches. By the time the threshold of 0.675, for close match, is reached, this happens at only a very low rate. For meaningful match, choosing the upper confidence band for the threshold brings the error rate for meaningless annotations down to below 10%. c) How annotators rate predictions as a function of Tanimoto similarity, split by whether the prediction came from a prediction technology (non-random), or the prediction is simply an unrelated random structure (random), inserted into the annotation as a control. We see that for any given Tanimoto similarity in the range spanned by our data, that the annotators rate the real predictions substantially higher than the random controls. This indicates that the Tanimoto similarity scores of the real predictions to the ground truth molecular structures cannot be simply compared to the similarity scores between random pairs of molecules. The limited range of similarities on the x-axis is because there were not enough high similarity random predictions to extend the analysis to higher similarities.