bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MS2Prop: A machine learning model that directly predicts chemical properties from mass spectrometry data for novel compounds

Gennady Voronov [1], Abe Frandsen [1], Brian Bargh [1], David Healey [1], Rose Lightheart [1], Tobias Kind [1], Pieter Dorrestein [1,2], Viswa Colluru [1], and Thomas Butler *[1]

[1]Enveda Biosciences, 1880 S Flatiron Ct Ste K, Boulder, 80301, CO, USA
[2]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, 9500 Gilman Drive, La Jolla, 92093, CA, USA

October 10, 2022

Abstract

Mass spectrometry is a key analytical tool for the study of complex small molecule mixtures. The contents of these mixtures are the subject of untargeted metabolomics applications ranging from understanding metabolism, disease, biomarkers, and environmental contaminants to natural products based drug discovery. Yet identifying from mixtures the compounds or their properties from mass spectrometry data remains very challenging for most small molecules. For most compounds there will not be an annotation, and nearly all annotation techniques rely on partial matches to spectral and structural databases with limited coverage. However, property prediction of unknowns in untargeted metabolomics relies heavily on those annotations. Here we introduce MS2Prop, a complement to compound identification, that directly predicts chemically relevant properties of compounds for drug discovery and other applications from mass spectrometry data for any mass spectrometry feature, regardless of whether the corresponding compound is in an existing database. On compounds excluded from the training set MS2Prop has an average $R^2 = 0.73$ across ten properties, including synthetic accessibility and quantitative drug likeness properties, and $R^2 = 0.96$ for compounds in the training set, but with disjoint spectra. For compounds excluded from the training set, MS2Prop outperforms predictions based on compound identification by over a factor of three, setting the stage for future use of computational prioritization of compounds for diagnostic and drug discovery applications.

1 Introduction

In liquid chromatography tandem mass spectrometry (MS/MS) based untargeted metabolomics experiments, a majority of fragmentation spectra cannot be assigned an accurate structure with current computational tools [7]. Unidentified compounds may correlate with relevant biological and chemical activity and be of high interest. Natural products drug discovery is similarly limited by the challenge of compound identification, where it is common for many unidentified MS features to show bioactivity and it is often precisely the novel compounds that are of greatest interest [4]. This forces decisions about which compounds to investigate further to be made without structural information. Follow-up experiments often involve expensive and time-consuming compound isolation and nuclear magnetic resonance (NMR) structural elucidation experiments. Because of this, computational methods for characterizing compound properties, such as drug likeness, directly from mass spectrometry data, before embarking on isolation and or synthesis, could assist in prioritization of natural products for drug discovery and other metabolomics applications.

Most existing untargeted compound identification methods either search for spectral matches in tandem spectrum/structure databases [41, 17], or use machine learning to search compound databases using

*Corresponding Author: tom.butler@envedabio.com

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mass spectra as a query [11, 6]. With incomplete annotation of known compounds and the inability to identify any unknown compounds, only a small fraction (2-20% depending in sample type) of actual small metabolite chemistry can presently be annotated by analyzing MS/MS data [7]. Molecular networking extends the range of spectral libraries by visualizing clusters of compounds, allowing some unknown compounds to be linked to related identifiable compounds via clustering. However, it does not quantitatively identify the properties of the unknown compounds [32]. De novo structure prediction technologies that can predict the structures of novel compounds are beginning to be actively investigated, but so far performance is limited [34, 37]. Other methods predict the compound classes of unknown metabolites, but leave the structure and chemical properties undetermined [9].

While compound identification, especially for novel compounds, remains very challenging, metabolomics applications like prioritizing hits in natural products drug discovery may be supported by information that is much lower dimensional than full chemical structures, especially if it is available for compounds that are not in chemical databases. For example, medicinal chemists' decisions about hits to select for further study can be supported by predictions of the properties of the compounds corresponding to the features of interest, including logP, synthetic accessibility [12], Quantitative Estimate of Drug-likeness (QED) [5], fraction of sp3 carbon atoms, and others. While natural product based drugs are rich in examples that violate drug-likeness criteria and other property ranges common in synthetic drug discovery[4], evaluation of FDA approved natural product based drugs show that even for natural products drug discovery, QED and other properties do correlate with FDA drug approval (Supplementary Information). When choosing between bioactive features for additional screening, isolation, structural elucidation or other follow up, medicinal chemists can usefully balance properties like QED with bioactivity and other factors to make higher quality decisions even without full structure information.

Here, we train a machine learning model, MS2Prop, end-to-end for predicting 10 key properties of compounds from MS/MS data directly. MS2Prop performs with high accuracy even for molecules outside of its training set (novel with respect to the model), while referring to no external database. For novel compounds, MS2Prop performance averages $R^2 = 70\%$ across 10 properties compared with only $R^2 = 22\%$ for a multi-stage model approach that predicts structure first, and then extracts properties from the structure. Thus, MS2Prop can be used to predict properties for novel chemical compounds. We also show that for structures contained in spectral libraries, the resulting model is able to reproduce the properties of the compounds nearly perfectly with disjoint spectra ($R^2 = 95\%$). Across both novel and known compounds MS2Prop performance is much higher than multi-stage approaches likely because it avoids predicting the chemical structure or fingerprint as an intermediate product, which is a challenging high-dimensional structured prediction problem [8]. Furthermore, it avoids complex error propagation problems intrinsic to pipeline models in general [2].

MS2Prop is related to models developed in molecular machine learning that predict the properties of molecules from their structures (e.g. [42]) in that it attempts to predict properties of compounds, but differs in that it attempts to predict them from MS/MS data rather than from molecular structure. Additionally, in this paper we focus on properties whose estimates can be calculated directly from the chemical structure, rather than targets such as toxicity that are experimentally derived [42]. To our knowledge MS2Prop is the first model to predict core chemical properties directly from MS/MS data.

Furthermore, the architecture is extensible to any properties where sufficient training data can be obtained. Unlike many machine learning methods on MS data, which may require minutes or more for a single inference call and scale poorly with molecular weight, MS2Prop can carry out inference even for very complex spectra with many peaks in a few milliseconds. Exploiting this computational efficiency, we predict properties for 500 million unlabeled spectra from a range of repositories of metabolomics experiments [29, 3, 14, 17, 22, 24, 25, 27, 31, 33, 35, 39, 41]. A first look at natural products space using these predictions suggests there exist portions of natural products space that are significantly drug-like, relatively synthetically accessible, and largely un-mined by current FDA approved drugs.

2 Results

2.1 MS2Prop model

MS2Prop predicts 10 numerical chemical properties (atomic logP, number of hydrogen bond acceptors, number of hydrogen bond donors, polar surface area, number of rotatable bonds, number of aromatic rings, number of aliphatic rings, fraction of sp3 carbons, quantitative estimate of drug likeness [5], and synthetic accessibility [12]). These were selected for relevance to drug discovery and medicinal chemistry and are easily computed within RDKit [23] from chemical structures. Conceptually, MS2Prop maps

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

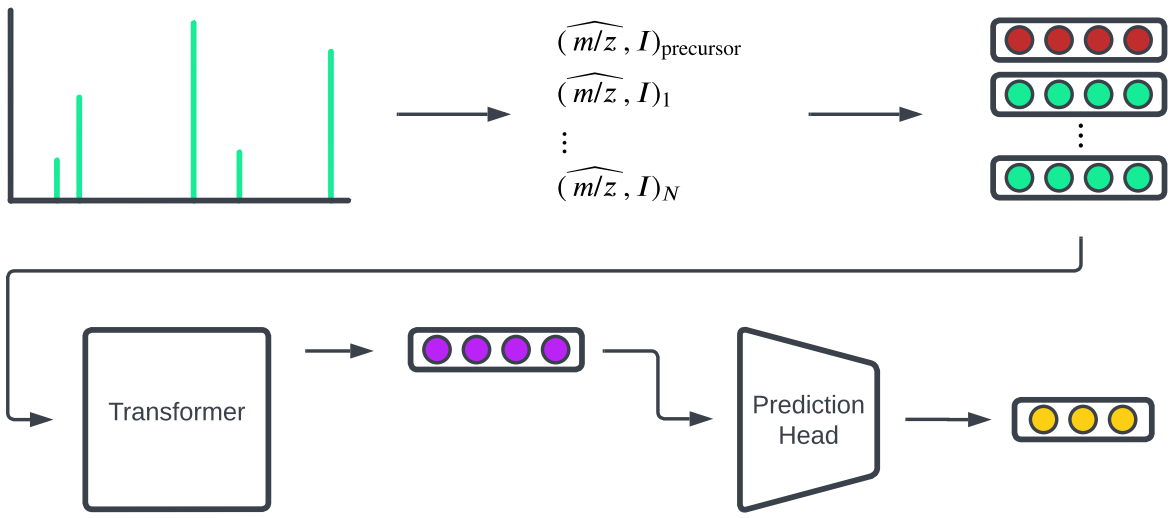


Figure 1: Overview of MS2Prop for property prediction. Masses from MS/MS data are rounded to tokens at a resolution of 0.1 Daltons and paired with normalized intensities. These are then passed through a learned embedding layer that transforms the mass and intensity data into an array of dense vectors. The embeddings are input to a transformer block, which learns patterns corresponding different properties and emits a dense vector that is used by a prediction head to simultaneously predict all target properties.

MS/MS spectra to chemical properties by aggregating information across the fragment peaks to create a latent vector representation of the input spectrum and then using this vector to predict the target properties. The model is built around a deep neural network architecture consisting of three main stages.

First, the input spectrum – represented as a sequence of m/z-intensity pairs along with the precursor m/z – is mapped to a sequence of embeddings, i.e. dense real-valued vectors. In this stage, each m/z value is tokenized by rounding to a fixed precision, following [18], with the resulting m/z token denoted by $\widehat{m/z}$.

Next, the sequence of embeddings is passed through a transformer [40]. The key insight of the transformer architecture is that each layer of the network has an output for each element of the input sequence and can weight those outputs dynamically in the most relevant way for learning. While transformers have largely been associated with natural language processing tasks, they are also a natural choice for adaptive aggregation of information across the peaks of MS/MS spectra. The final layer of the transformer outputs a sequence of vectors, but we retain only the first output vector, which can be considered a latent embedding for the entire MS/MS spectrum.

Finally, the spectrum embedding is passed to a prediction head, which is a feed forward neural network that outputs the predicted chemical properties. Because this final prediction module computes all properties, only a single inference call is needed. The entire MS2Prop model can then be trained by minimizing the mean squared error between the predicted and true properties over a labeled training dataset. See Figure 1 for an outline of the model and Section 4.2 for further details.

2.2 MS2Prop validation

To validate the predictive performance of MS2Prop, we test it on three annotated MS/MS datasets. Traditionally, machine learning systems are validated using a single test set that is simply a random sample from the same pool of data used in training. However, work on molecular property prediction from chemical structures has shown that such test sets over-estimate generalization performance because the data is too similar to the training data compared to realistic use cases [42]. This phenomenon is common in machine learning in complex domains, and similar observations have been made in natural language processing [13].

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To address this phenomenon and accurately estimate the generalization performance of MS2Prop, we use three different test sets that provide increasingly stringent tests of generalization performance. We first test with a random split that is disjoint in spectra, but not in structures ("known"). This split represents the easiest generalization setting, where a lab may be interested in properties across experiments on compounds contained in reference databases created under experimental conditions that partially overlap their own. We also test with a split that is both structure and spectrum disjoint ("novel"). This represents the performance on novel chemical structures, but where the experimental setting may have some overlap with the set of settings used to generate the training data. Finally, we test with a set of experiments from a small molecule identification challenge (CASMI 2022[1]) which were generated in an independent experimental setting, and whose molecules were chosen because they were not in spectral reference libraries. The CASMI data are disjoint in time, structure, experiment, and spectra, and represent a particularly challenging test of generalization performance.

An alternative to a direct spectra-to-property model would be to obtain predicted molecular structures and compute the properties of the predicted structures [30, 19]. Therefore, in order to benchmark MS2Prop, we also apply two standard approaches for structure identification – cosine similarity spectral lookup [41] and CSI:FingerID [11] – and evaluate how well these perform at the property prediction task. We briefly note that we used our own implementation of cosine spectral similarity that is substantially faster than standard software but gives comparable lookup performance. Additionally, since we do not train the CSI:FingerID model ourselves, we cannot guarantee that structures in the novel test set are absent from the data used to train CSI:FingerID. A key motivation for using the CASMI 2022 dataset for evaluation is that we can guarantee that it is disjoint both in structure and experiment from any training data used for any method considered here. In all cases, predicted structures are contained in structure databases available to CSI:FingerID at inference time, but which are not accessed by MS2Prop. Thus, we expect the comparison would be more favorable to MS2Prop on truly novel compounds that are unavailable in compound databases.

As shown in the Figure 2, MS2Prop has overall R^2 values (averaged across all predicted properties) of 0.928, 0.735, and 0.719 on the known, novel, and CASMI 22 test datasets, respectively. This model outperforms the two baselines on every dataset we consider. A direct comparison is possible between MS2Prop and cosine similarity spectral lookup on our known and novel test sets, where MS2Prop shows a strong advantage. A fair comparison is possible between all methods on the CASMI 2022 dataset, and MS2Prop outperforms the rest by 25 percentage points.

To give a more comprehensive overview of property prediction performance, we show the R^2 broken down by individual predicted properties on each test dataset in Figure 3. Observe that certain properties appear easier to predict, such as the number of hydrogen bond acceptors; all methods have strong performance on these, but MS2Prop generally performs best. Atomic $\log P$, a particularly important property for drug discovery, is one of the more challenging properties to predict. MS2Prop outperforms the baselines for this property as well, but there is clear room for improvement. CSI:FingerID outperforms MS2Prop by a small margin on the CASMI 2022 dataset for three properties: the number of hydrogen bond acceptors, polar surface area, and the number of rotatable bonds. We briefly note that R^2s must be compared with caution between different test sets, as the actual property variance can and does differ between the various test sets. Figures 2-3 enable the comparison of distinct model performance differences on each test set separately and comparisons between different test sets are seen more readily from the Mean Absolute Errors (MAE) which we report, by individual properties, in the Supplementary Information provided. Note that MAE carries the dimensions of the underlying properties, and therefore cannot be aggregated across properties.

Finally, we also highlight MS2Prop performance on two properties particularly relevant to drug discovery efforts, namely QED and synthetic accessibility. In Figures 4a and 4b we show scatter plots of predicted vs. actual values for these two properties. The "novel" test set has many MS/MS spectra per structure and this accounts for the vertical lines present in Figure 4.

2.3 Mapping natural product space

A virtue of MS2Prop is that it enables inference of chemical properties on large datasets. On a single NVIDIA Tesla V100 GPU, MS2Prop requires approximately $\sim$2 milliseconds per MS/MS spectrum. Moreover, the computation cost is fixed with respect to the precursor mass and the number of fragments. In contrast, property inference via CSI:FingerID requires, on average, $\sim$27 seconds (with a standard deviation of $\sim$50 seconds) computation time per MS/MS spectra and may require up to hours for larger

[1] `https://fiehnlab.ucdavis.edu/casmi`

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

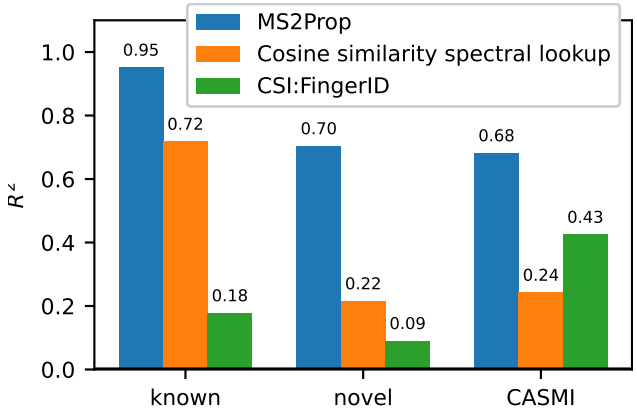


Figure 2: Model performance as measured by R^2 aggregated across all properties (higher is better). Performance is reported across three test datasets: known, novel, and CASMI 22 and three models: MS2Prop, cosine similarity spectral lookup, and CSI:FingerID.

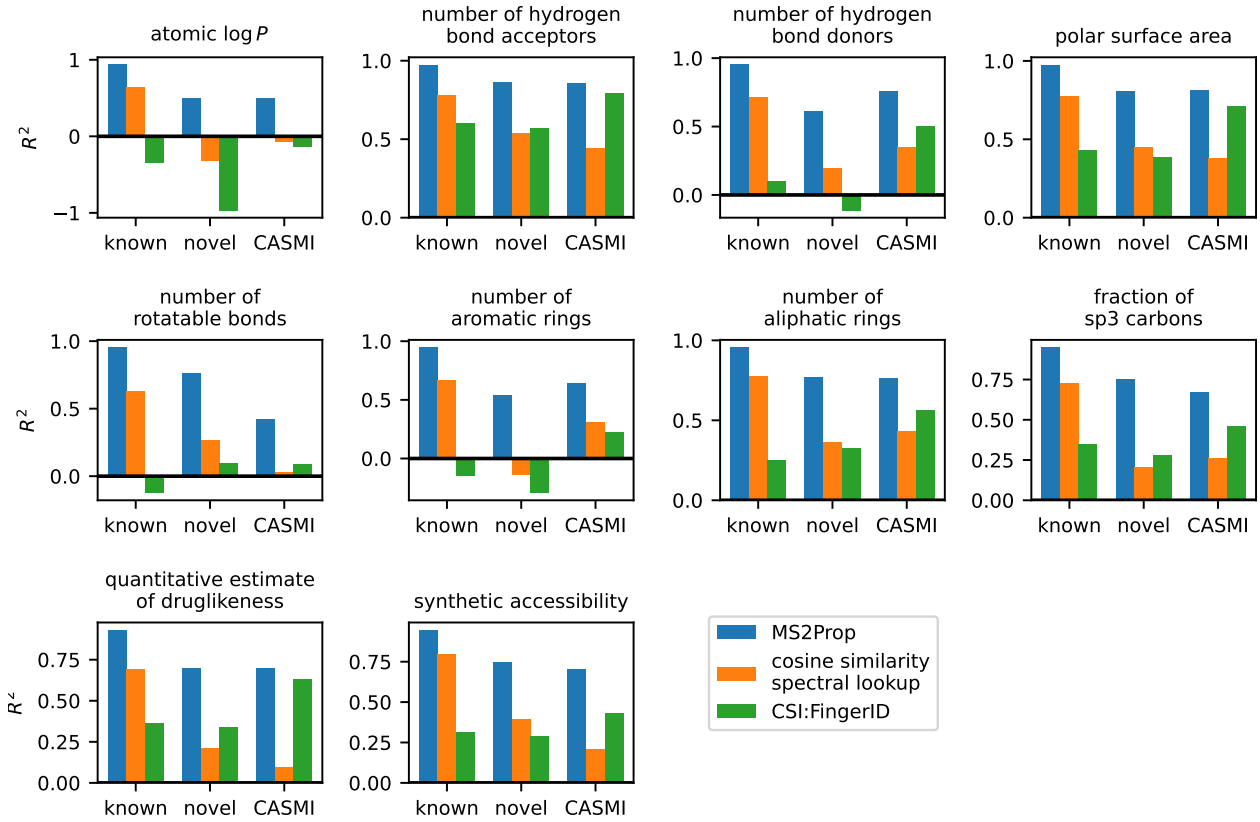
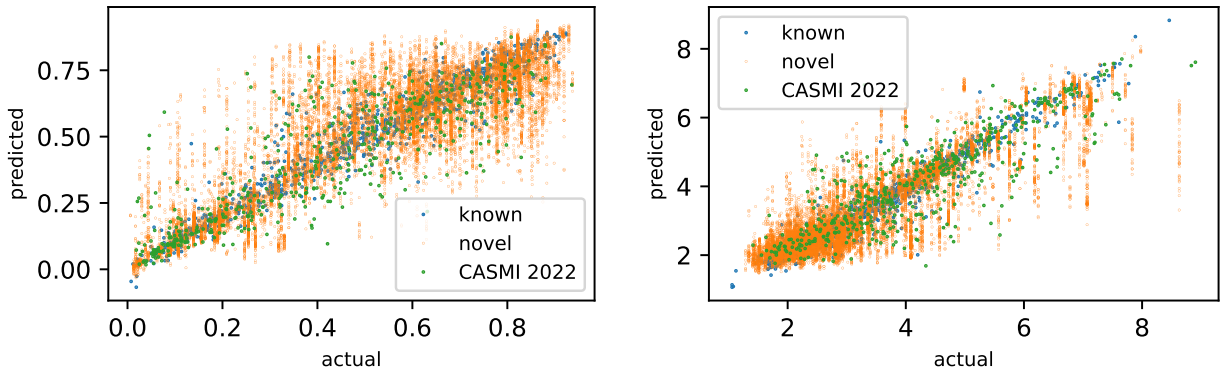


Figure 3: Model performance as measured by R^2 shown dis-aggregated for all properties (higher is better). Performance is reported across three test datasets: known, novel, and CASMI 22 and three models: MS2Prop, cosine similarity spectral lookup, and CSI:FingerID.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



(a) (b)

Figure 4: Predicted vs. actual scatter plots for QED (left) and synthetic accessibility (right). The vertical lines are a result of multiple spectra of varying quality and experimental conditions corresponding to the same structure.

precursor masses. This is nearly 12,000 times slower than MS2Prop on average. Moreover, CSI:FingerID fails to return any output for some inputs. Optimized cosine similarity search is faster, but still more than five hundred times slower than MS2Prop. See Figure 5. Our timing results are consistent with published latency benchmarks for several compound identification methods, including CSI:FingerID, all of which are reported to require O(10) seconds or more for all but the smallest molecules [6]. The favorable computational properties of MS2Prop make it a powerful tool for exploring the molecular properties of natural product chemical space using large public datasets that have been relatively unexplored until now.

To demonstrate the capabilities of MS2Prop, we use it to compute predicted properties on a dataset of 500 million un-annotated MS/MS spectra corresponding to natural products. These unlabeled spectra were collected from GNPS [41], MetaboLights [15], and Metabolomics Workbench [38]. We show the empirical distribution of the QED property in Figure 6a. Bickerton *et al.* [5] suggest a threshold QED $\geq$ 0.8 to discriminate for druglike compounds; we find that 0.637% of spectra have a predicted QED greater than 0.8. This raises the tantalizing possibility that public datasets contain millions of unlabeled MS/MS spectra that correspond to unexplored druglike molecules, and that chemical space generally contains an abundance of appealing molecules that have yet to be explored. A few caveats are in order. First, it is likely that the number of distinct compounds corresponding to these "druglike" spectra is much smaller. However, this consideration applies equally throughout the spectral data. In future work, we aim to dereplicate these spectra in order to better estimate the rate of druglikeness in natural product chemical space. Second, the threshold QED $\geq$ 0.8 is overly restrictive. Indeed, as can be seen in Figure 6b, while increased QED is obviously desirable, a substantial number of natural products with QED $\leq$ 0.8 have been approved by the FDA.

To further characterize the unlabeled MS/MS natural products dataset, we randomly sample 200K spectra and then sample an additional 10K spectra with the requirement that their predicted QED exceeds 0.8. In Figure 7, we plot these 210K spectra using the Uniform Manifold Approximation and Projection (UMAP) [28] dimensionality reduction technique. The sampling is used to make the plotting tractable, but the qualitative results are not sensitive to the sample size. The distances between spectra are specified by their spectral cosine similarity [41]. We produce two versions of this UMAP figure. In the first (Figure 7a), we color individual spectra by their predicted QED. In the second (Figure 7b), we color individual spectra by their predicted synthetic accessibility score (lower means easier to synthesize).

Furthermore, in Figures 7a-7b, we show MS/MS spectra which correspond to FDA approved drugs. In both UMAP figures, these are shown by orange-colored points if they also correspond to a natural product and by red-colored points otherwise. These MS/MS spectra are obtained from our labeled MS/MS dataset, restricted to those compounds that exist in the FDA Orange Book[1] of approved drugs. Whether or not a given compound is a natural product is determined by its presence in the COCONUT[36] Natural Product database, which may also include some NP derived compounds. Explicitly, we join

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

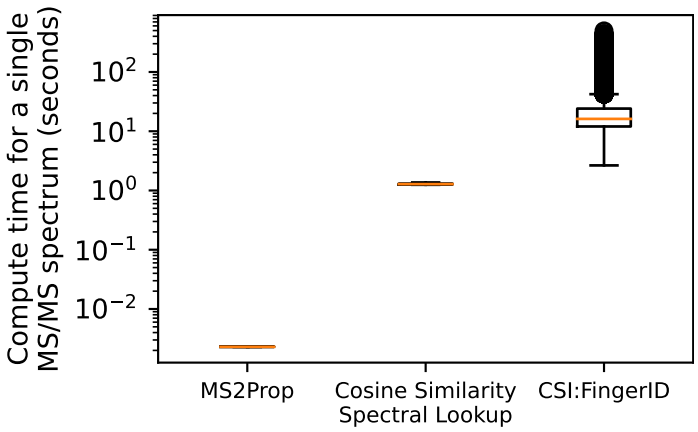


Figure 5: Log-scale comparison of compute time required for inference over a single MS/MS spectrum between the MS2Prop model, cosine similarity spectral lookup, and CSI:FingerID. This is evaluated on our "known" and "novel" test sets. These have an average compute time, per MS/MS spectrum, of 2.3 milliseconds, 1.29 seconds, and 27.3 seconds respectfully. Notably, MS2Prop is nearly $\sim$12K faster than CSI:FingerID and over 550 times faster than similarity lookup.

the relevant datasets by using the first block of 14 characters of the InChIKey[16] corresponding to the relevant compounds. Using this approximate identification scheme, our annotated dataset contains spectra for 708 FDA approved drugs that are also contained in COCONUT, and for 85 FDA approved drugs not contained in COCONUT.

In Figure 7, we highlight a number of regions in natural product chemical space as represented by the UMAP. First we note that regions A and B both contain FDA approved compounds, natural products and otherwise, and that these are mapped close to unlabeled MS/MS spectra which have predicted QED and synthetic accessibility that would lead one to expect promising drug candidates. Focusing in on Region A, where we see many approved drugs clustered tightly, the regions immediately surrounding these approved compounds are predicted to be relatively unsuitable (less drug-like and harder to synthesise).

Furthermore, in Figure 7, we highlight an additional sample set of 3 regions (regions C-E). Each of these regions suggest that there are drug-like natural products which are distinct from FDA approved drugs and have desirable synthetic accessibility scores. Note these regions highlighted in 7 are not exhaustive, they are merely obvious examples visible in a two dimensional representation of a very high dimensional space. It's worth noting that these regions do contain some FDA approved compounds nearby and further work is required to ascertain the degree of relation between these compounds. Figure 7 suggests that there are un-mined regions of natural product chemical space which are both drug-like and relatively easy to synthesize.

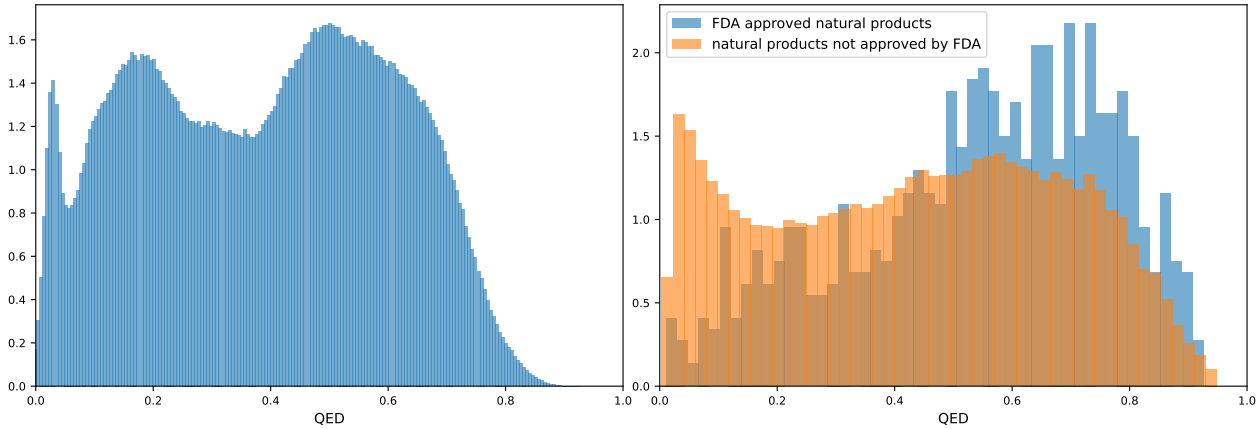
3 Discussion

MS2Prop shows that many of the most relevant chemical properties for novel compounds that cannot be identified with current compound identification technology can be predicted directly from tandem mass spectrometry data. While compound identification remains a crucial task, accurate characterization of compounds in a sample or collection of samples by their properties can provide a complementary view of the chemistry that is more accurate and can scale at inference time to billions of spectra (*i.e.* all known) mass spectrometry datasets.

In a first of kind analysis of hundreds of millions of unlabeled spectra, we showed suggestive evidence that there exist regions of natural product space with appealing synthesis and drug-like properties that are substantially un-mined. Further exploration will be required to validate that hypothesis, but the ability of the analysis to raise the hypothesis in a data-driven manner shows the promise of repository scale machine learning analysis of metabolomics data.

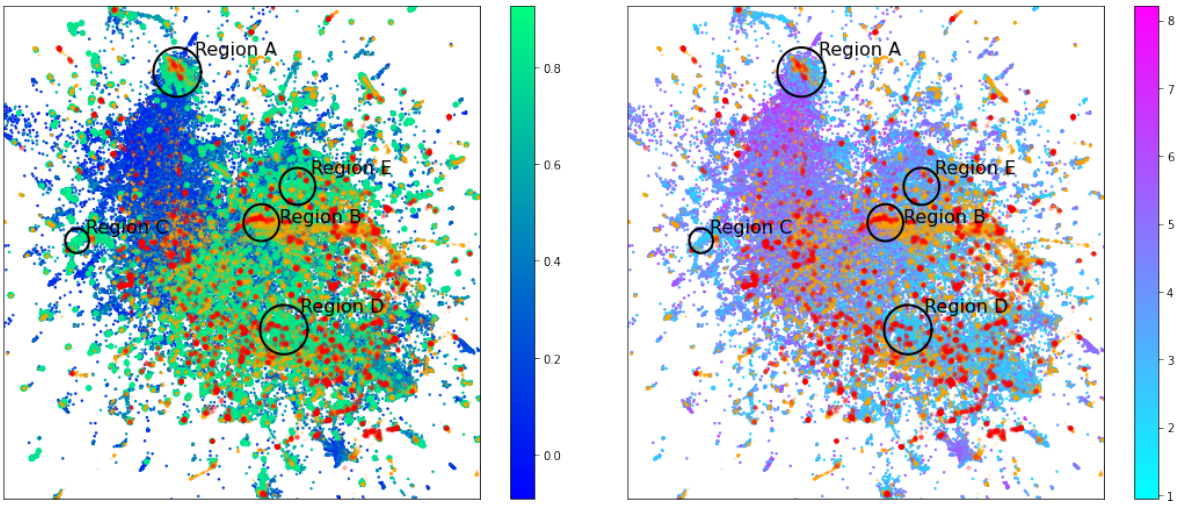
Our approach of directly predicting properties instead of predicting the compounds or fingerprints as intermediate objects has two principal advantages. The first is that it avoids complex, sometimes non-monotonic, propagation of errors that arise in multi-stage machine learning systems [2]. The second

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



(a) (b)

Figure 6: Empirical distribution of predicted QED in large unlabeled MS/MS spectral dataset (left) and empirical distribution of actual QED for natural products with and without FDA drug approval (right).



(a) (b)

Figure 7: UMAP of a sub-sample of unlabeled MS/MS spectra from GNPS. Orange (red) points correspond to MS/MS spectra which are derived from FDA approved compounds that are (not) natural products. Color shading in Figure 7a indicates predicted quantitative estimate of drug likeness (QED). Color shading in Figure 7b indicates predicted synthetic accessibility. The circled regions a sample of regions where compounds are predicted to be highly drug like and have favorable synthetic accessibility scores. For synthetic accessibility, lower scores indicate predictions that compounds are easier to synthesize. Regions A and B contain many FDA approved compounds, while regions C-E are relatively free of FDA approved compounds.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

principal advantage is that fingerprints or compound structures are far more complex and high dimensional objects than the scalar properties predicted by this model. Avoiding the need to predict these complex intermediate objects increases the tractability of the learning problem, and likely plays a large role in the high performance of our method. These factors in conjunction with the use of modern training and transformer architecture pioneered in large language models [20, 40] likely drive MS2Prop's strong performance and generalization.

Despite its strong performance, MS2Prop has several shortcomings. The current implementation incorporates molecular formula information only indirectly through the use of the precursor mass as a feature. Modern tandem MS analysis tools can identify molecular formulas with 90+% accuracy in many contexts [10, 26, 21]. Additionally, MS2Prop ignores all ion mode information and truncates the resolution of the input spectra aggressively. Furthermore, the transformer architecture used here is particularly well-suited to self-supervised learning approaches pioneered in large language models that allow the model to learn from abundant unlabeled mass spectrometry data [20]. We thus see the strong performance of the current model as a baseline for future work, and expect that the accuracy of property prediction from mass spectrometry will improve with further development.

In conclusion, this work shows that chemically relevant properties can be predicted with high accuracy and low latency directly from tandem MS data even for novel chemistry. Unlike many popular methods, it can be applied at repository scale. It functions as a complement to methods that attempt to directly identify structures or compound classes. Applying it to existing metabolomics workflows will support the quantitative characterization of complex samples with many metabolites with real-time latency, expanding the information that can be extracted from complex metabolic samples.

References

[1] Orange book: approved drug products with therapeutic equivalence evaluations. Accessed: 2022-09-01.

[2] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 291–300. IEEE, 2019.

[3] Ryohei Aoyagi, Kazutaka Ikeda, Yosuke Isobe, and Makoto Arita. Comprehensive analyses of oxidized phospholipids using a measured ms/ms spectra library. *Journal of lipid research*, 58(11):2229–2237, 2017.

[4] 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.

[5] G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.

[6] 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.

[7] 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.

[8] Hal Daumé, John Langford, and Daniel Marcu. Search-based structured prediction. *Machine learning*, 75(3):297–325, 2009.

[9] Kai Dührkop, Louis-Félix Nothias, Markus Fleischauer, Raphael Reher, Marcus Ludwig, Martin A Hoffmann, Daniel Petras, William H Gerwick, Juho Rousu, Pieter C Dorrestein, et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology*, 39(4):462–471, 2021.

[10] Kai Dührkop, Kerstin Scheubert, and Sebastian Böcker. Molecular formula identification with sirius. *Metabolites*, 3(2):506–516, 2013.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[11] 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.

[12] 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.

[13] Kyle Gorman and Steven Bedrick. We need to talk about standard splits. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2786–2791, 2019.

[14] Shuo Han, Will Van Treuren, Curt R Fischer, Bryan D Merrill, Brian C DeFelice, Juan M Sanchez, Steven K Higginbottom, Leah Guthrie, Lalla A Fall, Dylan Dodd, et al. A metabolomics pipeline for the mechanistic interrogation of the gut microbiome. *Nature*, 595(7867):415–420, 2021.

[15] Kenneth Haug, Reza M Salek, Pablo Conesa, Janna Hastings, Paula De Matos, Mark Rijnbeek, Tejasvi Mahendraker, Mark Williams, Steffen Neumann, Philippe Rocca-Serra, et al. Metabolights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic acids research*, 41(D1):D781–D786, 2013.

[16] Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. Inchi, the iupac international chemical identifier. *Journal of cheminformatics*, 7(1):1–34, 2015.

[17] 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.

[18] 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.

[19] Vilma Jägerroos et al. Predicting drug bioactivities from tandem mass spectra. 2019.

[20] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[21] Tobias Kind and Oliver Fiehn. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics*, 8(1):1–20, 2007.

[22] Jeremy P Koelmel, Matthew K Paige, Juan J Aristizabal-Henao, Nicole M Robey, Sara L Nason, Paul J Stelben, Yang Li, Nicholas M Kroeger, Michael P Napolitano, Tina Savvaides, et al. Toward comprehensive per-and polyfluoroalkyl substances annotation using fluoromatch software and intelligent high-resolution tandem mass spectrometry acquisition. *Analytical Chemistry*, 92(16):11186–11194, 2020.

[23] Greg Landrum. Rdkit: Open-source cheminformatics software. 2016.

[24] Sangwon Lee, Sungbo Hwang, Myungwon Seo, Ki Beom Shin, Kwang Hoe Kim, Gun Wook Park, Jin Young Kim, Jong Shin Yoo, and Kyoung Tai No. Bmdms-np: A comprehensive esi-ms/ms spectral library of natural compounds. *Phytochemistry*, 177:112427, 2020.

[25] Zhentian Lei, Li Jing, Feng Qiu, Hua Zhang, David Huhman, Zhiqin Zhou, and Lloyd W Sumner. Construction of an ultrahigh pressure liquid chromatography-tandem mass spectral library of plant natural products and comparative spectral analyses. *Analytical chemistry*, 87(14):7373–7381, 2015.

[26] Marcus Ludwig, Louis-Félix Nothias, Kai Dührkop, Irina Koester, Markus Fleischauer, Martin A Hoffmann, Daniel Petras, Fernando Vargas, Mustafa Morsy, Lihini Aluwihare, et al. Database-independent molecular formula annotation using gibbs sampling through zodiac. *Nature Machine Intelligence*, 2(10):629–641, 2020.

[27] Marie Mardal, Mette Findal Andreasen, Christian Brinch Mollerup, Peter Stockham, Rasmus Telving, Nikolaos S Thomaidis, Konstantina S Diamanti, Kristian Linnet, and Petur Weihe Dalsgaard. Highresnps. com: an online crowd-sourced hr-ms database for suspect and non-targeted screening of new psychoactive substances. *Journal of Analytical Toxicology*, 2019.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[28] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. cite arxiv:1802.03426Comment: Reference implementation available at http://github.com/lmcinnes/umap.

[29] 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.

[30] Yuji Nozaki and Takamichi Nakamoto. Odor impression prediction from mass spectra. *PLoS One*, 11(6):e0157030, 2016.

[31] Prasad Phapale, Andrew Palmer, Rose Muthoni Gathungu, Dipali Kale, Britta Brugger, and Theodore Alexandrov. Public lc-orbitrap tandem mass spectral library for metabolite identification. *Journal of Proteome Research*, 20(4):2089–2097, 2021.

[32] Robert A Quinn, Louis-Felix Nothias, Oliver Vining, Michael Meehan, Eduardo Esquenazi, and Pieter C Dorrestein. Molecular networking as a drug discovery, drug metabolism, and precision medicine strategy. *Trends in pharmacological sciences*, 38(2):143–154, 2017.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] Michael A Stravs, Kai Dührkop, Sebastian Böcker, and Nicola Zamboni. Msnovelist: De novo structure generation from mass spectra. *Nature Methods*, pages 1–6, 2022.

[38] Manish Sud, Eoin Fahy, Dawn Cotter, Kenan Azam, Ilango Vadivelu, Charles Burant, Arthur Edison, Oliver Fiehn, Richard Higashi, K Sreekumaran Nair, et al. Metabolomics workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic acids research*, 44(D1):D463–D470, 2016.

[39] Ipputa Tada, Hiroshi Tsugawa, Isabel Meister, Pei Zhang, Rie Shu, Riho Katsumi, Craig E Wheelock, Masanori Arita, and Romanas Chaleckis. Creating a reliable mass spectral–retention time library for all ion fragmentation-based metabolomics. *Metabolites*, 9(11):251, 2019.

[40] 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.

[41] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, 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.

[42] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4 Methods

4.1 MS/MS Datasets and Data Preparation

We construct a set of labeled MS/MS spectra by combining a number of publicly available datasets [29, 3, 14, 17, 22, 24, 25, 27, 31, 33, 35, 39, 41] with our internal proprietary dataset. We allow spectra collected in both positive and negative ion mode, and across a variety of instrument types, collision energies, and other important instrument parameters. For evaluation purposes, we also use data from the CASMI 2022 contest.

In metabolomics experiments, biological samples contain both compounds that have previously been profiled in MS/MS libraries and compounds that have not been profiled. Therefore, we separately characterize model performance on "known" molecules that have MS/MS spectra seen by the model during train time, and "novel" molecules with no spectra seen by the model during train time. To accomplish this, we partition our set of labeled spectra into a training set ("Train") and a test set ("Test"). While the test and training sets are disjoint at the spectrum level, they are *not* disjoint at the molecule level, and so we further partition the test set into "KnownTest" – consisting of spectra corresponding to molecules that are represented in the training set – and "NovelTest" – consisting of spectra corresponding to molecules that are not represented in the training set. We also evaluate model performance on the CASMI 2022 dataset ("CASMI"), which is *experiment* disjoint from the training set.

We also construct a larger set of unlabeled MS/MS spectra by combining publicly available data from GNPS [41], MetaboLights [15], and Metabolomics Workbench [38].

Data Preparation We apply simple filtering steps to ensure uniform quality in our datasets. First, we exclude spectra that have fewer than 3 decimal places of m/z resolution. Next, we exclude spectra with precursor m/z greater than 1000 Daltons, as we are interested in the small molecule domain. Similarly, from each spectrum we exclude peaks with m/z greater than 1000 Daltons. Next, for each spectrum we sort the peaks by intensity and retain only the top 512 peaks. Finally, we exclude spectra that have fewer than 5 peaks remaining after the previous filtering steps.

For each spectrum, we normalize intensities to have a maximum of 1. We discretize all m/z values by rounding them to the nearest 0.1 Dalton, as in Spec2Vec [18]. In this way, each peak is represented by a discrete token $\widehat{m/z}$ and a normalized intensity value.

We strip all stereochemistry from our molecular structure labels, which is a common step taken in MS/MS modeling and allows for better molecule-disjoint splitting. Chemical properties and fingerprints are then computed from the cleaned molecules using RDKit [23]. In total, our resulting labeled data has approximately 1,250,000 spectra corresponding to approximately 45,000 distinct molecules.

4.2 MS2Prop Architecture

Model Input Our MS2Prop model treats an input MS/MS spectrum S as a set

$$S = \left\{ (m/z, I)_{\text{precursor}}, (m/z, I)_{\text{fragment}_1}, \ldots, (m/z, I)_{\text{fragment}_N} \right\}, \tag{1}$$

comprising a precursor m/z and N fragment peaks at various m/z's and intensities (I). We discretize each m/z by rounding to 0.1 Da (denoted by $\widehat{m/z}$), and we normalize the fragment intensity values in each spectrum to have a maximum of 1.0. Since the precursor m/z indicates the molecular mass and doesn't have an experimental intensity value, we always assign it an intensity of 2.0. Hereafter, we refer to each tuple $(m/z, I)$ within a spectrum as a *peak*.

Peak Embedding Given an input MS/MS spectrum S, MS2Prop first embeds each peak into a continuous vector space as follows:

$$\text{PE}\,(m/z, I) = \text{FF}\left(\text{TE}\left(\widehat{m/z}\right) \parallel I\right) \tag{2}$$

Here, TE is an embedding function that maps each $\widehat{m/z}$ token to a vector of dimension $d = 512$ – analogous to word embeddings in natural language processing models – the $\parallel$ operator denotes concatenation, and FF is a standard feed-forward neural network with 1 hidden layer of dimension d, ReLU nonlinearity, and output dimension d. A diagram of PE is shown in Figure 8.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

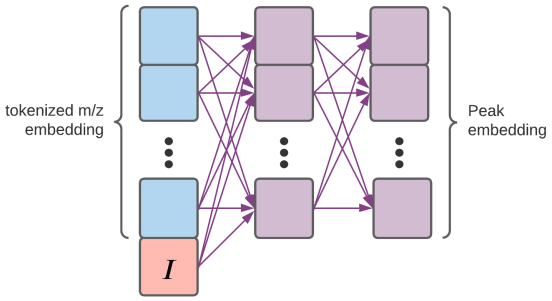


Figure 8: Peak Embedding architecture used to embed MS/MS peaks (composed of a m/z and an intensity I). We use blue to highlight the network components involved in producing an m/z embedding and purple to highlight the full peak embedding after intensity is incorporated.

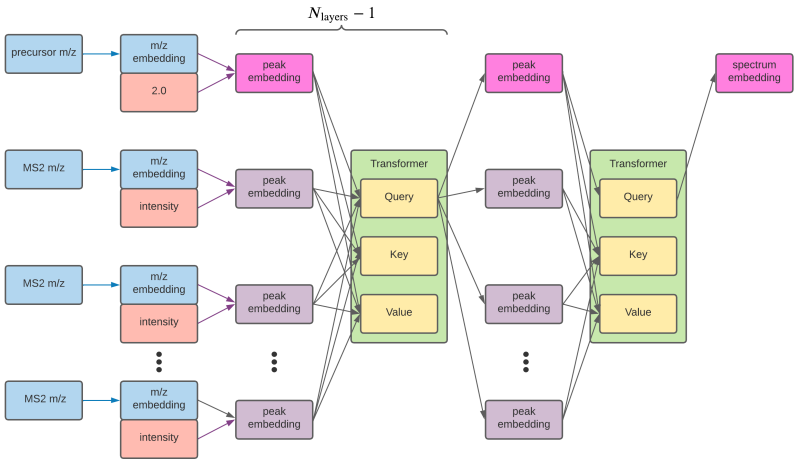


Figure 9: Full model architecture that highlights how we featurize and embed an MS/MS spectrum in a dense vector space.

Spectrum Transformer Encoder Transformer encoders [40] without the positional encoding are fully-symmetric functions and hence are ideally suited to model a set of MS/MS fragmentation peaks [34]. Therefore, MS2Prop passes a sequence of peak embeddings to a transformer encoder as follows:

$$\text{SpectrumEncoder}\left(S\right) = \text{TransformerEncoder}\left(\text{PE}\left(m/z, I\right)_{\text{precursor}}, \ldots, \text{PE}\left(m/z, I\right)_{\text{fragment}_N}\right). \quad (3)$$

Here, TransformerEncoder has embedding dimension d and six layers, each with 32 attention heads and an inner hidden dimension of d. As mass spectra have no intrinsic ordering, we opt to not include a positional encoding.

In order to get a single embedding vector as an output, the final transformer layer query only attends to the first embedding, corresponding to the position of the precursor m/z. A diagram of our full model architecture is shown in Figure 9.

Property Prediction Head After encoding the input spectrum to a single vector representation, MS2Prop finally passes this to a simple feed-forward neural network with 1 hidden layer of dimension d and output dimension 10 to predict the desired properties.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

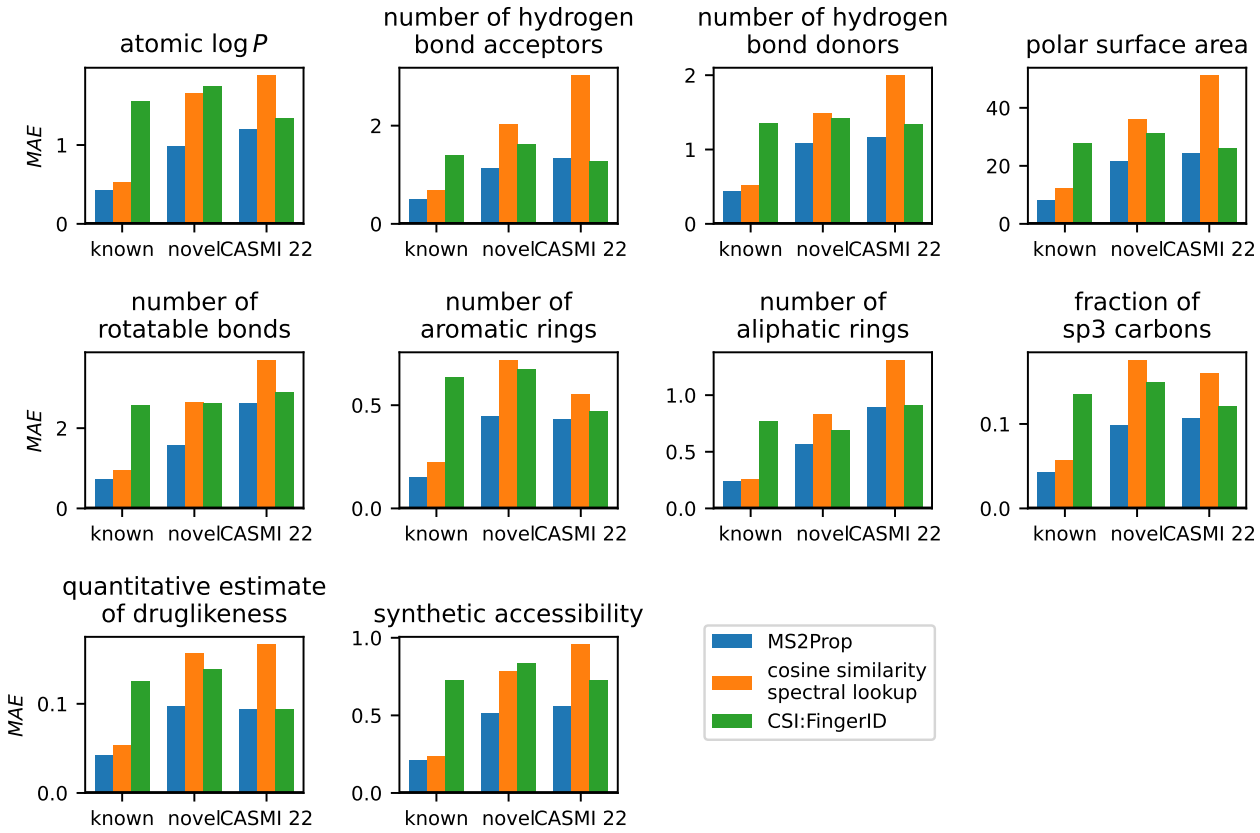


Figure 10: Model performance as measured by MAE shown dis-aggregated for all properties (lower is better). Performance is reported across three test datasets: known, novel, and CASMI 22 and three models: MS2Prop, cosine similarity spectral lookup, and CSI:FingerID.

4.3 Sirius and CSI:FingerID

We benchmarked MS2Prop against the generation of properties from chemical structure predictions provided by SIRIUS and CSI:FingerID (version 4.8.2). To predict structures, We run SIRIUS molecular formula predictions run with a 10 ppm mass tolerance, no database constraints, and default settings with respect to atom distributions and possible adducts. CSI:FingerID was run against all but in silico databases. Notably, we do not provide Sirius with MS1 isotopic distributions (as we are benchmarking only predictions from MS2 spectra); we anticipate both MS2Prop and CSI:FingerID would perform better given accurate molecular formulas as input.

5 Supplementary Information

5.1 Additional property prediction performance metrics

To provide further context on prediction performance, we also report the Mean Absolute Error (MAE) (Figure 10) and the Root Mean Squared Error (RMSE) (Figure 11) for all properties on all test datasets. MAE and RMSE figures establish expectations of accuracy on specific properties for downstream users of MS2Prop. For example, on the QED property, the MAE across our various datasets is MAE ≤ 0.1. This is sufficiently good performance to meaningfully impact drug discovery efforts.

5.2 Further analysis of QED "drug-likeness"

Typically a threshold of QED ≥ 0.8 [5] is used to discriminate for druglike compounds. Applying this threshold to our PropertyMS QED predictions leads to an accuracy of 0.93, 0.88, and 0.96 on our known, novel, CASMI 2022 test sets respectively. However, our train and test datasets are unbalanced with

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

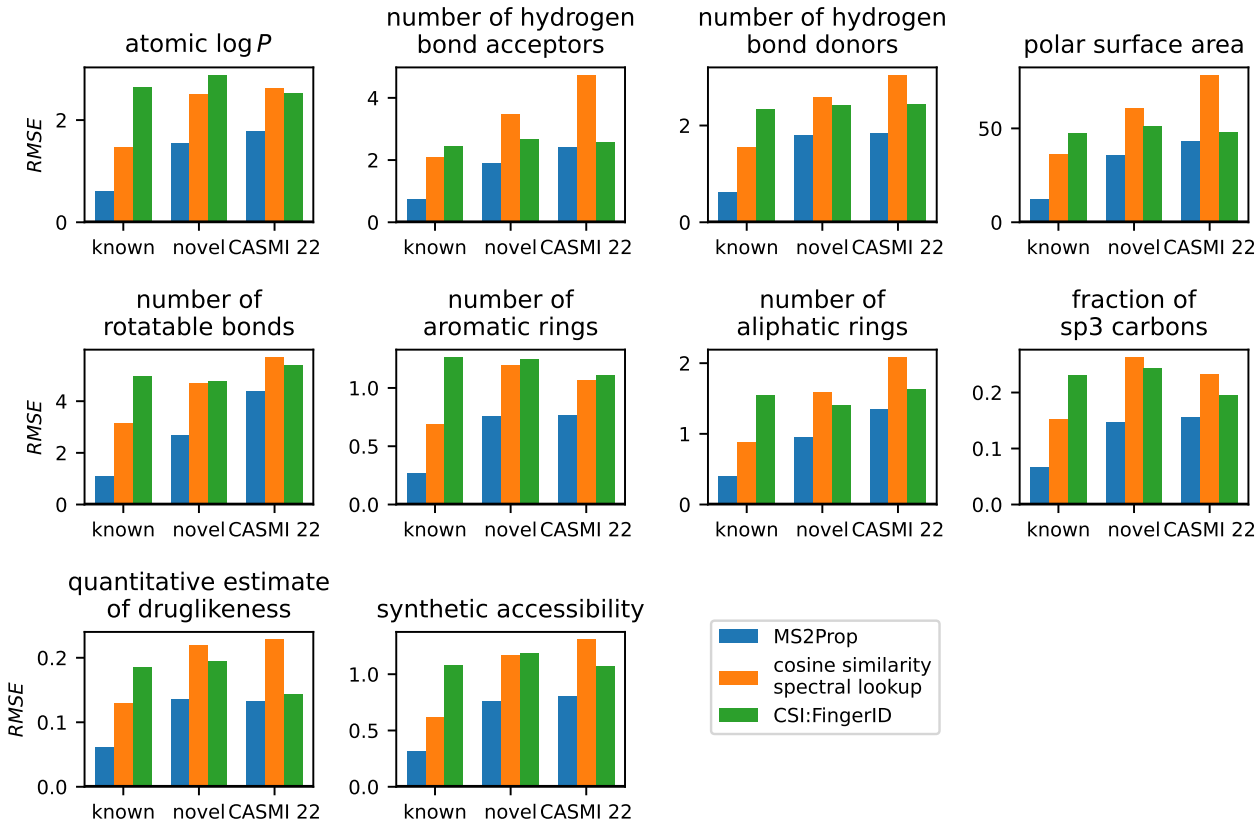
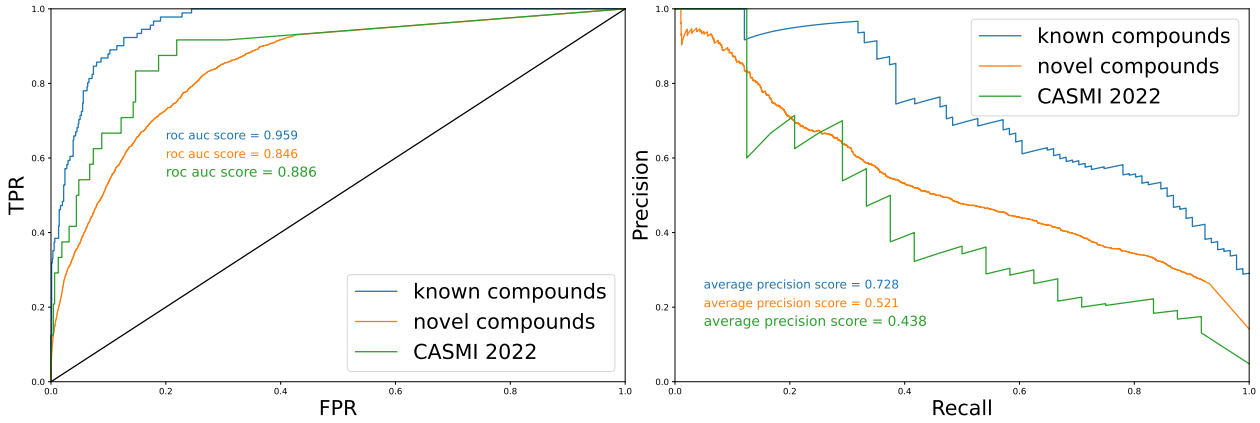


Figure 11: Model performance as measured by RMSE shown dis-aggregated for all properties (lower is better). Performance is reported across three test datasets: known, novel, and CASMI 22 and three models: MS2Prop, cosine similarity spectral lookup, and CSI:FingerID.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



(a) (b)

Figure 12: Evaluation of predicted QED as a class label determined by QED ≥ 0.8. Performance is reported across three test datasets: known, novel, and CASMI 22. (left) ROC curves are shown for each test set. We report ROC-AUC scores of 0.959, 0.846, and 0.886 for the known, novel, and CASMI 2022 sets respectfully. (right) Precision-recall curves are shown for each test set. We report average precision scores of 0.728, 0.521, and 0.438 for the known, novel, and CASMI 2022 sets respectfully

respect to QED druglikeness. Specifically 9.6%, 11.8%, 16.5%, and 4.8% compounds are QED druglike in our train, known test, novel test, and CASMI 2022 datasets respectfully. Due to this we also report a balanced accuracy of 0.74, 0.64, 0.56 on our known, novel, CASMI 2022 test sets respectively. Moreover, we also compute an ROC (Figure 12a) and a precision-recall (Figure 12b) curves. The ROC area-under-the-curve scores range from 0.826 to 0.903 over the various datasets, with the weakest performance being on novel test. The average precision scores are are 0.653, 0.491, and 0.404 for known test, novel test, and CASMI 2022. Reading off from Figure 12b, even on the harder datasets one can still recall around 20% of QED druglike compounds with over 50% precision. These two curves allow one to estimate how reliably will the model identify QED druglike compounds on large unlabeled datasets.

5.3 Correlation between properties and FDA approved drugs

One of the key applications of property prediction from MS/MS is to select from many bioactive MS peaks (features) the ones that are most likely to be appealing for further investigation in the absence of reliable structural elucidation. A medicinal chemist could use PropertyMS to compare the predicted chemical properties across several candidates, and evaluate the balance of appealing properties with bioactivity to make data driven decisions about which features to select of many possibilities for further screening, isolation and experimental structural elucidation.

To gain information on the utility of the properties for this purpose, we directly analyzed a collection of FDA approved natural products drug structures and compared them to a large corpus of natural products structures. The structures we considered were obtained from using all FDA approved drugs from the Orange Book [1] that are also contained in the COCONUT natural products database [36].

We then independently separated the structures into quartiles for each of the properties studied in the paper. For each quartile, we computed the mean number of FDA natural products within the quartile (SI figure 13). From inspection of the figures, it's clear that while drugs exist in all property value ranges, several of the properties are substantially enriched in certain property ranges. Quantitative estimate of drug-likeness in particular is a strong feature.

bioRxiv preprint doi: https://doi.org/10.1101/2022.10.09.511482; this version posted October 11, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

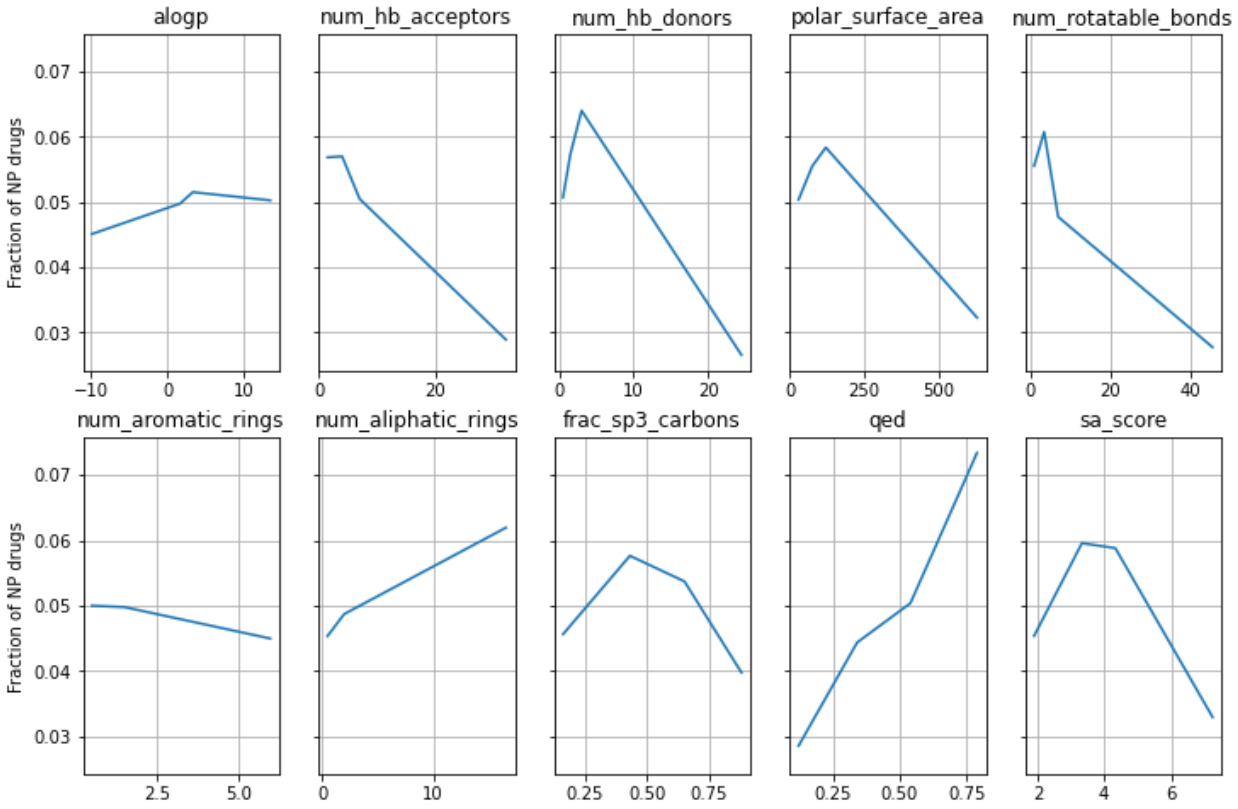


Figure 13: Fraction of Natural Product drugs in each property quartile. X axis labels are midpoints of the quartile over entire collection of natural products.