Supporting Information

Deep learning-enabled MS/MS spectrum prediction facilitates automated identification of novel psychoactive substances

Fei Wang[1,2], Daniel Pasin[3], Michael A. Skinnider[4,5,6], Jaanus Liigand[7,8], Jan-Niklas Kleis[9], David Brown[10,11], Eponine Oler[7], Tanvir Sajed[7], Vasuk Gautam[7], Stephen Harrison[10], Russell Greiner[1,2], Leonard J. Foster[4,12], Petur Weihe Dalsgaard[3], David S. Wishart[1,7,13,14,15*]

[1] Department of Computing Science, University of Alberta, Edmonton, Alberta, T6G2E8, Canada

[2] Alberta Machine Intelligence Institute, Edmonton, Alberta, T5J 3B1, Canada

[3] Section of Forensic Chemistry, Department of Forensic Medicine, University of Copenhagen, Copenhagen, 2100, Denmark

[4] Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, V6T 1Z4 , Canada

[5] Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, 08544, USA

[6] Ludwig Institute for Cancer Research, Princeton University, Princeton, NJ, 08544, USA

[7] Department of Biological Sciences, University of Alberta, Edmonton, Alberta, T6G 2E9, Canada

[8] Institute of Chemistry, University of Tartu, Tartu, 50411, Estonia

[9] Institute of Forensic Medicine, Forensic Toxicology, Johannes Gutenberg University Mainz, Mainz, 55131,Germany

[10] Forensic Science Laboratory, ChemCentre, Bentley, Western Australia, 6102, Australia

[11] School of Molecular and Life Sciences, Curtin University, Bentley, Western Australia, 6009, Australia

[12] Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, British Columbia, V6T 2A1, Canada

[13] Department of Laboratory Medicine and Pathology, University of Alberta, Edmonton, Alberta, T6G 1C9, Canada

[14] Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, Alberta, T6G 2C8, Canada

[15] Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, 99354, USA

*Corresponding author: Dr. David S. Wishart, Dept. of Biological Sciences, CW-405, Biological Sciences Building, University of Alberta, Edmonton, AB, Canada, T6G 2E8

Telephone: 1-780-492 8574

Email: dwishart@ualberta.ca

Table of Contents

Figure for training and testing dataset

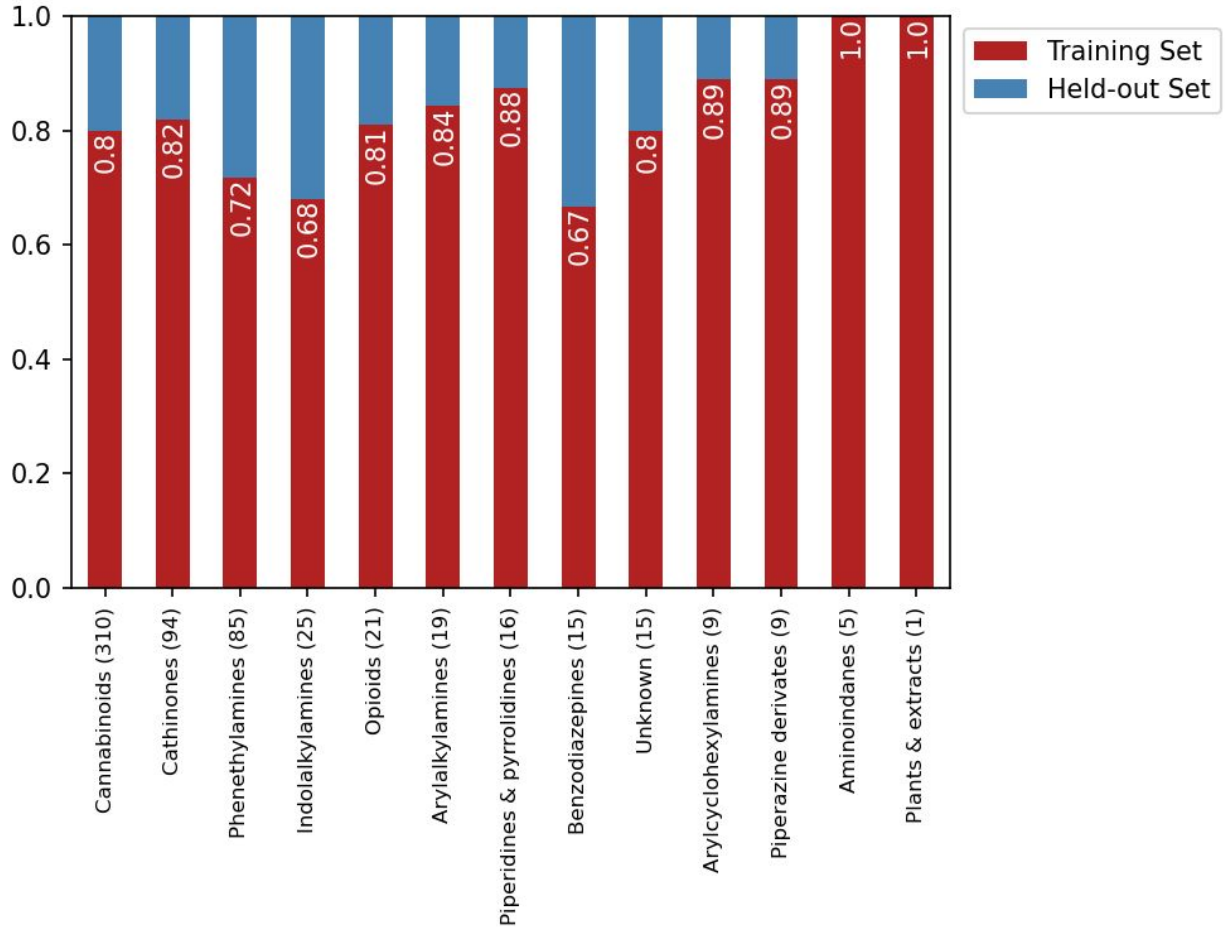

Figure S1. Proportion of compounds from each EMCDDA drug class in the training and test datasets.

Inspection of representative MS/MS spectra predicted by NPS-MS

In general, dot products tend to be higher than the Dice coefficients. Notably, the dot product takes into account relative intensities in addition to the m/z values of the peaks themselves, and consequently, higher-intensity peaks have far more influence over dot product than lower-intensity peaks. For example, the predicted spectrum of 5F-ADBICA at 40 eV has a Dice score of 0.44, since only two of the four predicted peaks correspond to an experimentally observed peak (Figure S2c). However, the two matching peaks have the greatest intensities, whereas the unmatched peaks in both the predicted and experimental spectra are of lower intensity, resulting in a dot product of 0.90.

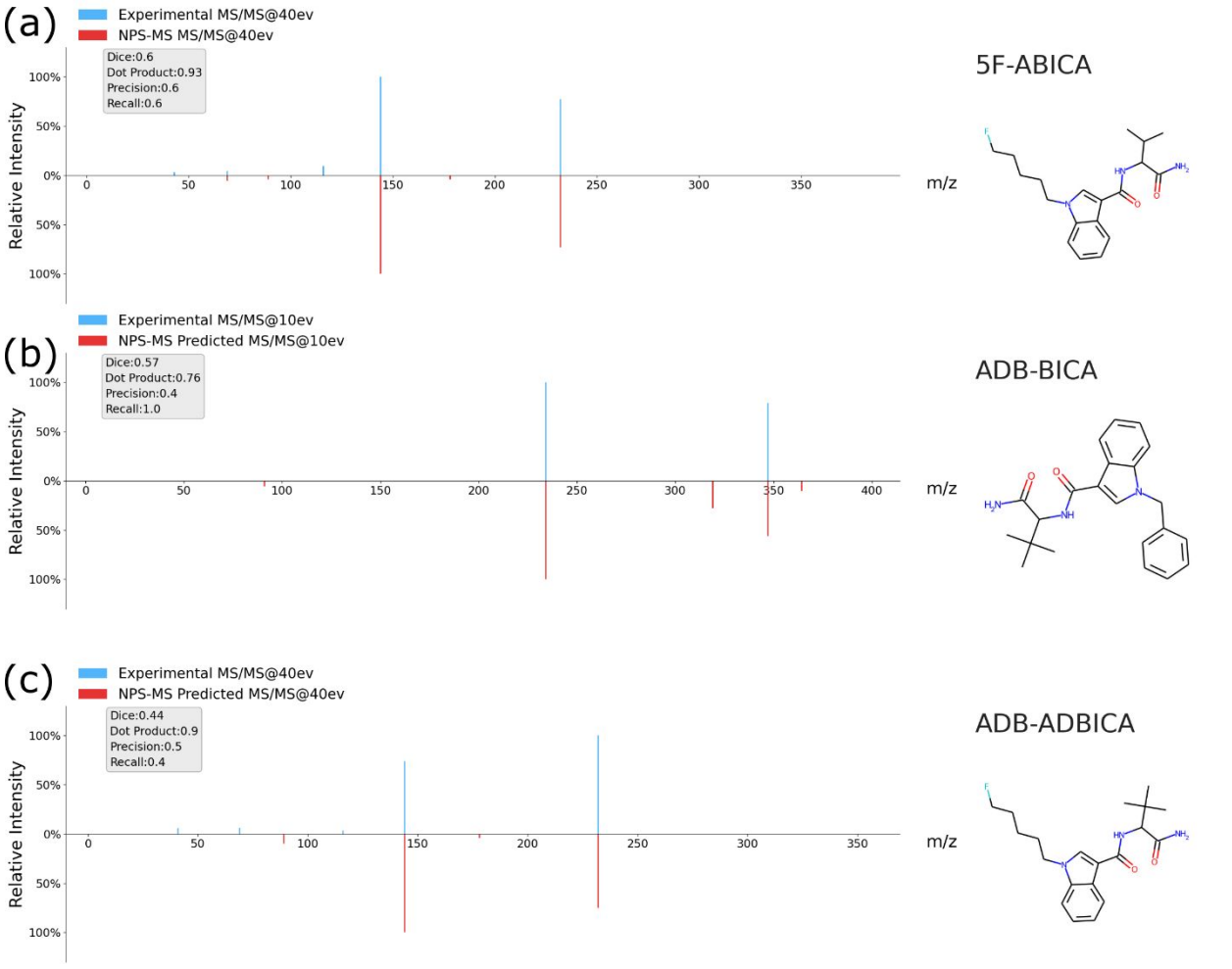

Figure S2. Inspection of representative MS/MS spectra predicted by NPS-MS. Mirror plots of experimental (top) and predicted (bottom) spectra for (a) 5F-ABICA at 40 eV and (b) ADB-BICA at 10 eV, illustrating spectra with Dice coefficients close to the mean value across the test dataset. (c) Mirror plot for 5F-ADBICA at 40 eV, demonstrating that low Dice coefficients may be due to low-abundance peaks, however, the major peaks have been predicted thus yielding a high dot product.

Details for cost score

The performance of each C2MS model was evaluated using a cost score. Cost scores are commonly used to evaluate the performance of models that produce a confidence score or rank. In this setting, a cost score was assigned to each compound identification based on the rank relative to the correct answer (or ground truth), considering the possibility of equally ranked candidates. This cost score reflects the amount of expected MS/MS experiments required to find ground truth compound given a list of identification result for a single task. Equation 1 defines the cost score. Here $Rank_i$ is the rank of the ground truth compound and $Count(Rank_i)$ is the number of compounds in this rank. $Count(Rank_j)$ denotes the number of compounds in the ranks higher than $Rank_i$ (i.e. $j < i$).

$$Cost\ Score = \frac{Count(Rank_i) + 1}{2} + \sum_{j<i}^{n} Count(Rank_j) \qquad \text{(Equation S1)}$$

For example, if the correct structure achieved the highest score but was ranked equally along four other candidates, its cost score would equal 3. However, if the correct structure was the highest scoring spectral match with no other candidates achieving the same score, then its cost score would equal 1. This scoring system circumvents the issues in using the top-k accuracy to evaluate performance when models may assign the same rank to many candidates.

Figure for histograms displaying the number of candidate compounds in each MS2C task

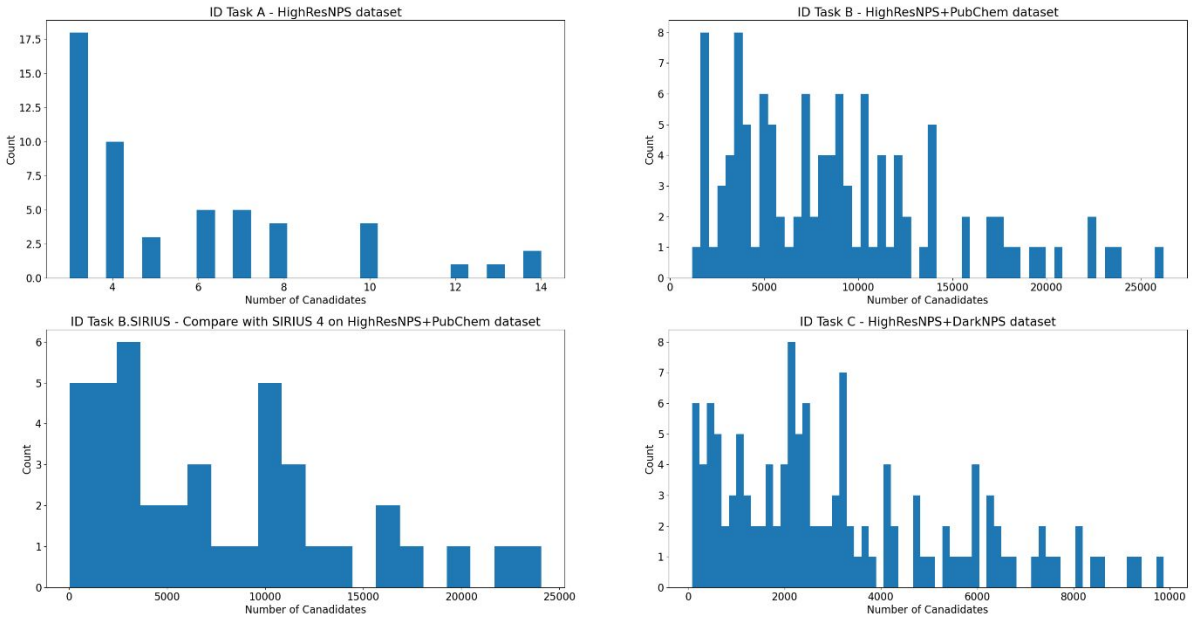

Figure S3. Histograms displaying the number of candidate compounds in each MS2C task. Candidate compounds are chosen based on its precursor ion mass (±10 ppm) from each task's specific candidate dataset.

Figure for MS/MS spectra prediction of the novel PCP derivative, 3-Cl-PCP

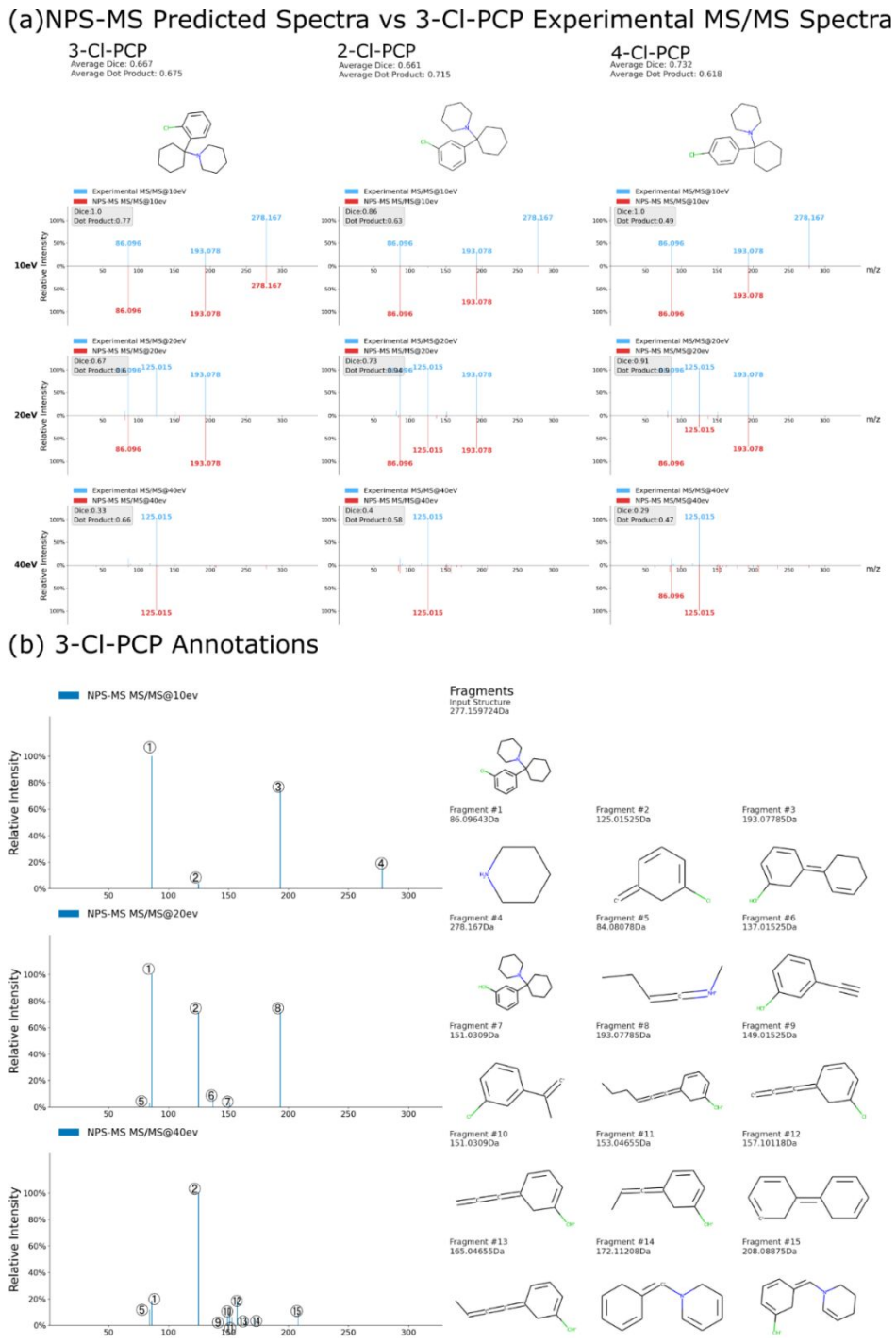

Figure S4. MS/MS spectra prediction of the novel PCP derivative, 3-Cl-PCP. a) Comparison of MS/MS spectra predicted by NPS-MS and experimentally acquired spectra for 3-Cl-PCP at 10, 20, and 40 eV. Experimental MS/MS spectra are shown in blue, while predicted MS/MS spectra are shown in red. b) Structures of hypothetical fragment ions proposed by NPS-MS for each peak in the predicted MS/MS spectra at 10, 20, and 40 eV.