



# Deep learning prediction of electrospray ionization tandem mass spectra of chemically derived molecules

Received: 12 March 2024

Accepted: 16 September 2024

Published online: 27 September 2024

Check for updates

Bin Chen<sup>1,4</sup>, Hailiang Li<sup>1,4</sup>, Rongfu Huang<sup>2</sup>, Yanan Tang<sup>3</sup>✉ & Feng Li<sup>1</sup>✉

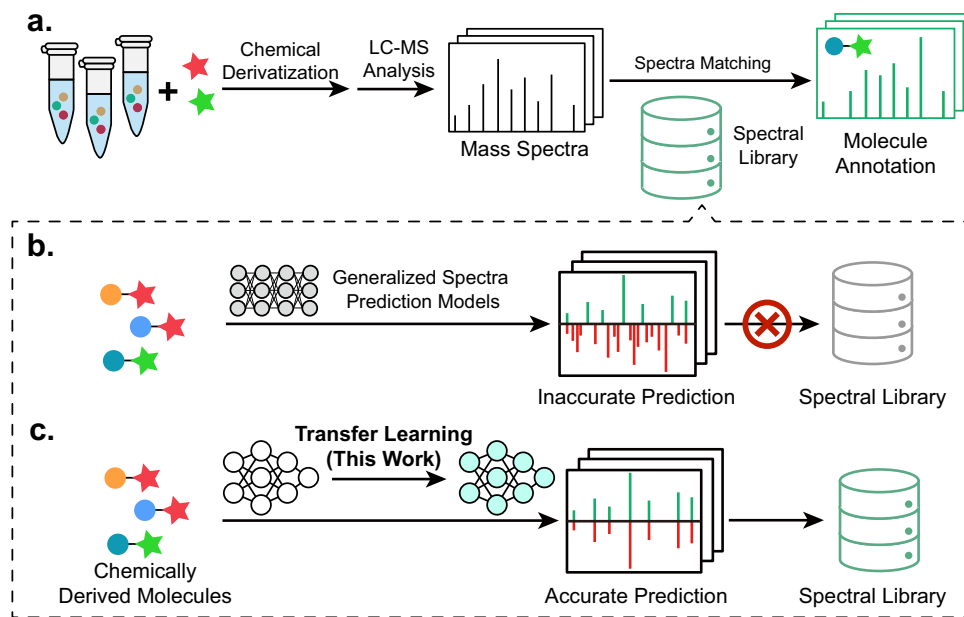
Chemical derivatization is a powerful strategy to enhance sensitivity and selectivity of liquid chromatography-mass spectrometry for non-targeted analysis of chemicals in complex mixtures. However, it remains impossible to obtain large sets of reference spectra for chemically derived molecules (CDMs), representing a major barrier in real-world applications. Herein, we describe a deep learning approach that enables accurate prediction of electrospray ionization tandem mass spectra for CDMs (DeepCDM). DeepCDM is established by transfer learning from a generic spectrum predicting model using a small set of experimentally acquired tandem mass spectra of CDMs, which converts a generic model with low predictability for CDMs into a specialized model with high predictability. We demonstrate DeepCDM by predicting electrospray ionization tandem mass spectra of dansylated molecules. The success in establishing Dns-MS further enables the development of DnsBank, a dansylation-specialized *in silico* spectral library. DnsBank achieves significant increases of accurate annotation rates of dansylated molecules, facilitating discovery of new hazardous pollutants from an environmental study of leather industrial wastewater. DeepCDM is also highly versatile for other classes of CDMs. Therefore, we envision that DeepCDM will pave a way for high-throughput identification of CDMs in non-targeted analysis to dig unknowns with potential health impacts from emerging anthropogenic chemicals.

Chemical derivatization is a common practice in analytical chemistry to enable the analysis of compounds that otherwise cannot be easily separated or detected using chromatographic or spectroscopic technologies<sup>1–3</sup>. When deployed into liquid chromatography-mass spectrometry (LC-MS) analysis that has been widely used in biomarker discovery<sup>4–6</sup>, untargeted metabolomics<sup>7–9</sup> and environmental research<sup>10–12</sup>, high-throughput identification of chemically derived molecules (CDMs) presents an unmet challenge in data mining and molecule identification. This is because typical pipelines to interpret

MS and tandem MS (MS/MS) data rely on spectrum matching in libraries (Fig. 1a), and only molecules for which a reference sample has been measured and its spectrum has been recorded can be confidently annotated<sup>13,14</sup>. However, there are very limited number of CDM reference spectra and there is no specialized spectral library for CDMs so far. Moreover, CDMs often generate significantly different MS/MS spectra from the original molecules. For instance, low similarity with weighted cosine similarity (WCS) value of only  $0.02 \pm 0.01$  (median, 95% CI) was found between the reference spectra of 231 underived

<sup>1</sup>Key Laboratory of Green Chemistry & Technology of Ministry of Education, College of Chemistry, Sichuan University, Chengdu, Sichuan 610064, China.

<sup>2</sup>Sichuan Provincial Key Laboratory of Universities on Environmental Science and Engineering, MOE Key Laboratory of Deep Earth Science and Engineering, College of Architecture and Environment, Sichuan University, Chengdu, Sichuan 610064, China. <sup>3</sup>Analytical & Testing Center, Sichuan University, Chengdu, Sichuan 610064, China. <sup>4</sup>These authors contributed equally: Bin Chen, Hailiang Li. ✉ e-mail: [yanantang@scu.edu.cn](mailto:yanantang@scu.edu.cn); [wintalker\\_1205@scu.edu.cn](mailto:wintalker_1205@scu.edu.cn)



**Fig. 1 | Barrier lies in the liquid chromatography-mass spectrometry (LC-MS) based non-targeted analysis of chemically derived molecules (CDMs).** **a** A schematic workflow of molecule annotation in LC-MS based non-targeted analysis of CDMs. Spectral library for CDMs is the key to fast and high confident molecule annotation from massive data. **b** Current predicting tools are not designed for predicting CDMs spectrum, and the predicted spectra with bad quality cannot be

used for building specialized *in silico* spectral library. **c** The development of specialized *in silico* spectral library for CDMs by DeepCDM. Transfer learning can cost-effectively reuse a model for new purpose<sup>43,55,56</sup>, thus can establish new predicting model specialized for CDMs with improved accuracy, allowing the rapid construction of *in silico* spectral library. The colored dots represent original molecules and the colored stars are chemical derivatizing reagents.

molecules recorded in the MassBank of North America (MoNA), MassBank and Human Metabolome Database (HMDB) and the experimentally obtained spectra of their dansylated forms (Fig. S1a). Consequently, non-targeted analysis of CDMs often results in low confidence levels (level 4–5)<sup>15,16</sup>. Therefore, there is an urgent need to build specialized spectral libraries for CDMs but experimentally measuring MS/MS spectra of CDMs at a repository scale remains unrealistic.

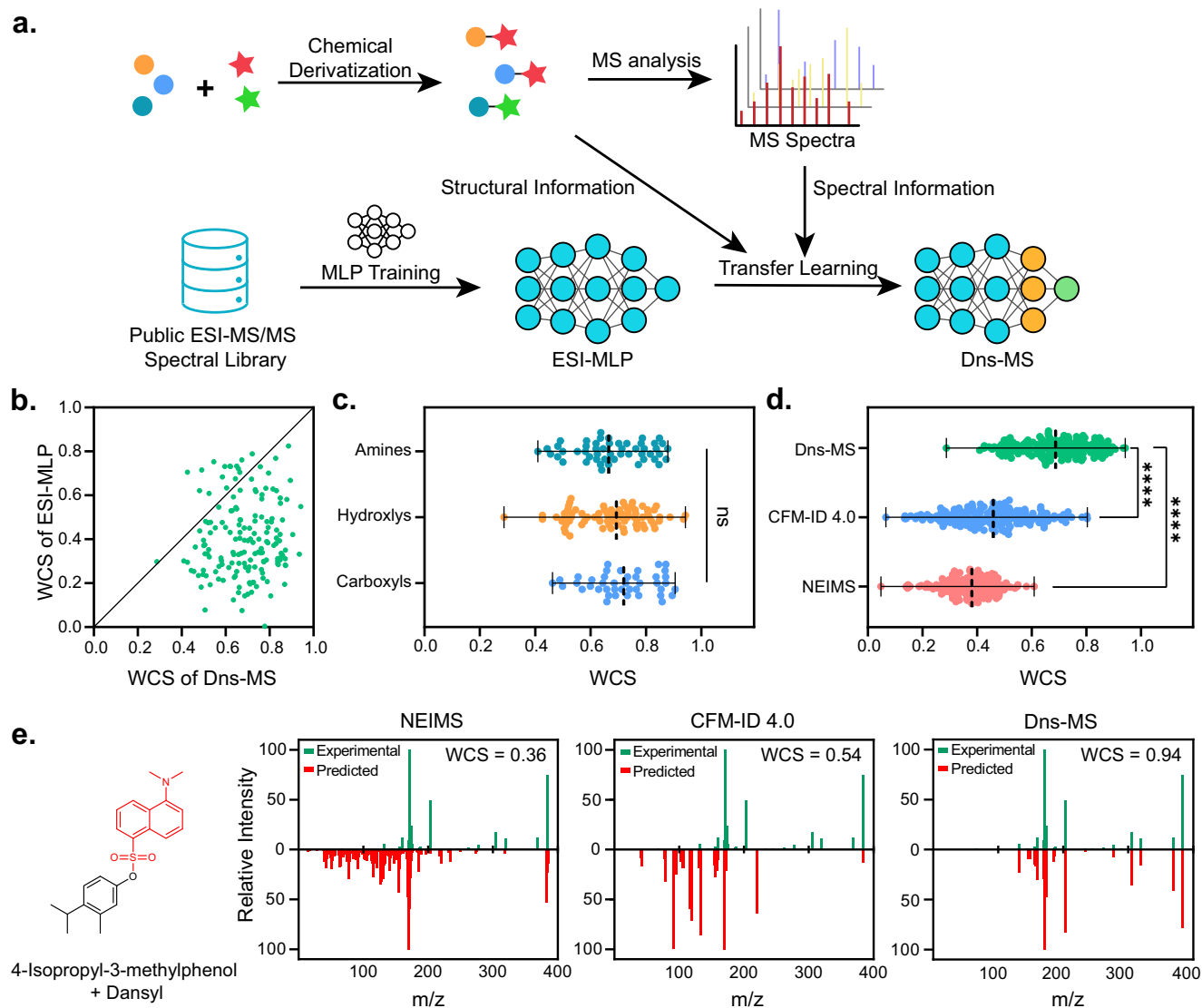
Recently, *in silico* methods were developed, allowing rapid prediction of MS/MS spectra in a scalable manner (Fig. 1b). Current predicting tools are developed based on quantum chemistry simulation or machine learning. The quantum chemistry approaches such as quantum chemistry electron ionization mass spectra (QCEIMS) are cost intensive, thus hamper efficiency and application<sup>17–19</sup>. Machine learning approaches are more efficient<sup>17</sup>. However, existing machine learning tools including neural electron-ionization mass spectrometry (NEIMS)<sup>20</sup> and competitive fragmentation modeling for metabolite identification (CFM-ID)<sup>21–23</sup> are general for underived molecules, offering low predictability for CDMs (Fig. S1b). Herein, we describe a deep learning approach that enables accurate prediction of electrospray ionization tandem mass spectrum (ESI-MS/MS) for CDMs, named as DeepCDM. Rather than developing a new predicting tool from scratch, DeepCDM focuses on re-training existing algorithms using a small set of experimentally obtained CDM spectra via transfer learning. We demonstrate that DeepCDM is accurate and scalable, enabling the establishment of a dansylation-specialized model, Dns-MS, which improves upon the generic tools, such as NEIMS and CFM-ID, in the compound-to-spectrum (C2MS) benchmark of dansylated molecules, and allowing the rapid construction of a specialized *in silico* spectral library containing 294,647 MS/MS spectra for dansylated molecules (Fig. 1c). Moreover, we also demonstrate that DeepCDM is highly versatile, which can be generalized for predicting MS/MS spectrum of other classes of CDMs by switching the training set of CDMs. We further demonstrate that DeepCDM is fully adaptable with non-targeted

analysis, which enables the rapid discovery of new hazardous chemicals in authentic wastewaters from leather industry.

## Results

### DeepCDM workflow for dansylated molecules

As a proof-of-concept, we aim to develop and validate DeepCDM for predicting MS/MS spectra of dansylated molecules. Dansylation is one of the most widely used chemical derivatization approaches in LC-MS based non-targeted analysis<sup>9,12,24</sup>, but there are only 315 MS/MS spectra of dansylated molecules available in MYCompoundID<sup>25</sup> and 6 in MassBank<sup>26</sup>. To accurately predict MS/MS spectra of dansylated molecules, DeepCDM retrained the multilayer perceptron (MLP) of NEIMS into a dansylation-specialized MS/MS predicting model (Dns-MS). The goal of Dns-MS is to predict collision-energy-merged ESI-MS/MS spectra of dansylated molecules because such spectra are most often acquired in non-targeted analysis<sup>27–32</sup>. As NEIMS is an electron ionization (EI) spectrum predicting model<sup>20</sup>, we first trained the MLP of NEIMS using 6790 ESI-positive MS/MS spectra collected from MoNA, so that this pre-trained model (ESI-MLP) was suitable for ESI-MS/MS spectrum prediction (Table S1, S2). Additionally, multiple MoNA spectra of one molecule at different collision energies were merged to generate collision-energy-merged spectra for training. ESI-MLP was further fine-tuned into Dns-MS, where a training set containing 507 experimentally obtained spectra of dansylated molecules synthesized by labeling chemicals with amine, hydroxyl, and carboxyl groups using dansyl chloride (DnsCl) and/or dansyl hydrazine (DnsHz) (Fig. 2a) was employed. The original molecular weights of the training set were ranged from 59.0735 to 665.6982 Da, containing 62 classes of molecules such as benzene and substituted derivatives, carboxylic acids and derivatives, and phenols. The molecules had an average pairwise Tanimoto similarity of 0.11 on their Extended-Connectivity Fingerprints (ECFP4) (Fig. S2). For each dansylated molecule, spectra were collected at multiple collision energies ranging from 10 eV to 50 eV and merged automatically by HESI Q-Exactive Plus Orbitrap (Thermo Scientific, USA).



**Fig. 2 | DeepCDM workflow for dansylated molecules and model evaluation.**

**a** The illustration of DeepCDM to establish a dansylation-specialized spectrum predicting model (Dns-MS) for predicting collision-energy-merged electrospray ionization tandem mass spectra (ESI-MS/MS) of dansylated molecules. The multi-layer perceptron (MLP) of neural electron-ionization mass spectrometry (NEIMS) was first trained using ESI-positive MS/MS spectra collected from the MassBank of North America (MoNA) into a pre-trained model, ESI-MLP, which was further fine-tuned into Dns-MS with a small training set of experimentally obtained spectra of dansylated molecules. The colored dots represent original molecules and the colored stars are dansylating reagents. **b** Weighted cosine similarity (WCS) comparisons between predicted spectra and experimental spectra of 167 dansylated molecules in a hold-out test set computed by Dns-MS and ESI-MLP. **c** WCSs of Dns-MS predicted spectra for dansylated molecules with amine, hydroxyl and carboxyl in the test set. Amines:  $0.66 \pm 0.03$ ,  $n = 51$ ; hydroxyls:  $0.69 \pm 0.03$ ,  $n = 74$  and

carboxyls:  $0.72 \pm 0.04$ ,  $n = 42$  (median, 95% CI). No significant difference was observed between the 3 groups ( $p = 0.1467$ ) through one-way analysis of variance (ANOVA). **d** Compound-to-spectrum (C2MS) capability of Dns-MS was benchmarked on hold-out dansylated molecules in the test set with generic tools. WCSs of predicted spectra generated by NEIMS, CFM-ID 4.0 and Dns-MS for dansylated molecules in the test set matching with the 167 experimental spectra. NEIMS:  $0.38 \pm 0.01$ , CFM-ID 4.0:  $0.46 \pm 0.02$  and Dns-MS:  $0.69 \pm 0.02$  (median, 95% CI). Dns-MS had significantly enhanced WCSs in comparison with CFM-ID 4.0 ( $p = 1.20 \times 10^{-38}$ ) and NEIMS ( $p = 8.19 \times 10^{-78}$ ). Unpaired two-tailed *t*-tests were performed to evaluate statistical differences between 2 groups ( $n = 167$  in every group). **e** The predicted spectra of dansylated 4-isopropyl-3-methylphenol generated by NEIMS, CFM-ID 4.0 and Dns-MS, matching with its experimental spectrum. ns:  $p > 0.05$ , \*\*\*\* $p \leq 0.0001$ . Source data are provided as a Source Data file.

Fine-tuning at the last layer of ESI-MLP showed slightly better performance over fine-tuning at all layers and was thus employed as an optimal fine-tuning strategy for Dns-MS (Fig. S3). Besides Dns-MS, we also developed Dns-MS- $\beta$  by directly training the MLP of NEIMS using a combined set of 6790 MoNA spectra and 507 experimental spectra of dansylated molecules (Fig. S4). Nevertheless, Dns-MS demonstrated much better predictability over Dns-MS- $\beta$  (Figs. S4, S5), suggesting that fine-tuning using even a small set of spectra of CDMs is critical to ensure the success of DeepCDM.

### Evaluation of DeepCDM for dansylated molecules

The compound-to-spectrum (C2MS) capability of Dns-MS was examined using a hold-out test set containing 167 experimental ESI-MS/MS spectra of dansylated molecules. Weighted cosine similarity (WCS) between Dns-MS predicted spectra and experimental spectra in the test set was determined to evaluate the predictability of Dns-MS (Table S3). Although fine-tuned only with a small-sized training set of dansylated molecules, much improved predictability of Dns-MS (median WCS =  $0.69 \pm 0.02$ , 95% CI,  $n = 167$ ) was achieved over ESI-MLP (median WCS =  $0.37 \pm 0.03$ , 95% CI,  $n = 167$ ) (Figs. 2b, S4–S5).

Moreover, Dns-MS showed equally high spectrum predictability for dansylated molecules synthesized through labeling chemicals with amine, hydroxyl and carboxyl groups (Fig. 2c), suggesting that Dns-MS can be broadly used with no structural bias.

In a 4-fold cross-validation, Dns-MS demonstrated robust performances for 4 randomly split dansylated dataset (Fig. S6a). The Dns-MS predicted spectrum and experimental spectrum of every dansylated molecule in the test set was also matched with the experimental spectrum of a structurally most similar molecule in the training set. The predicted spectra of query molecules had low similarity with the experimental spectra of their structurally similar molecules with a median WCS of  $0.27 \pm 0.02$  (95% CI,  $n = 167$ ). Similar observation was made between the experimental spectra of query molecules in the test set with those of structurally similar molecules in the training set (median WCS =  $0.22 \pm 0.03$ , 95% CI,  $n = 167$ , Fig. S6b, c). That confirmed the good predictability of Dns-MS for dansylated molecules that haven't been exposed to the model.

Then, we compared the performance Dns-MS with two widely used spectrum predicting tools, NEIMS and CFM-ID 4.0 (web server), using the test set of 167 experimental spectra. Figure 2d and Fig. S7 revealed better predictability of Dns-MS with a median WCS value of 0.69 over two other machine learning-based tools with median WCS values of 0.38 for NEIMS and 0.46 for CFM-ID 4.0, respectively. Figure 2e shows the predicted spectra of a representative molecule, dansylated 4-isopropyl-3-methylphenol using Dns-MS, NEIMS, and CFM-ID 4.0, as well as matching with the experimental spectrum of this dansylated standard. NEIMS had low predictability of dansylated 4-isopropyl-3-methylphenol with over-prediction of small fragments (WCS = 0.36). CFM-ID 4.0 produced a clean spectrum but with several mismatches (WCS = 0.54). By contrast, the predicted spectrum by Dns-MS showed an excellent match with the standard, which returned a WCS value of 0.94. The spectra matching results of all 167 dansylated molecules were summarized in Supplementary Data-01, and 73.65% of predicted spectra by Dns-MS were of high quality with WCS values over 0.6.

To evaluate the capability of DeepCDM to selectively enhance the MS/MS spectrum predictability of CDMs, we harnessed Dns-MS to predict spectrum for underived molecules in the MoNA dataset ( $n = 6790$ ). As shown in Fig. S8, the predictability of Dns-MS was drastically decreased for underived molecules with a median WCS of 0.38, confirming that our DeepCDM approach is specialized for desired CDMs.

### Versatility of DeepCDM

To demonstrate that DeepCDM is a generalizable approach for different classes of CDMs, we also established a spectrum predicting model for benzylation, another widely used chemical derivatization method for amine- and hydroxyl-containing molecules<sup>33–35</sup>. This model, termed Bz-MS, was first trained using the MoNA dataset containing 6790 merged ESI-MS/MS positive spectra of underived molecules, and then fine-tuned using a training set containing 503 experimental spectra of benzyolated molecules (Fig. S9). Evaluated with a hold-out test set containing 168 benzyolated molecules, Bz-MS achieved a median WCS value of  $0.74 \pm 0.03$  (95% CI,  $n = 168$ ) in predicting spectra for benzyolated molecules, which was significantly improved from the pre-trained model, ESI-MLP (median WCS =  $0.53 \pm 0.02$ , 95% CI,  $n = 168$ ) (Fig. S10a). It also outperformed NEIMS and CFM-ID 4.0 with over 1.5-fold increase of median WCS in a C2MS benchmark (Figs. S10, S11). Moreover, there was no bias in spectrum prediction for benzyolated amine- and hydroxyl-containing molecules (Fig. S10b). The robustness of Bz-MS was confirmed in a 4-fold cross-validation with 4 randomly split benzyolated datasets (Fig. S12). Bz-MS was also a highly specialized tool for benzyolated molecules with predictability significantly better than underived molecules and dansylated molecules (Fig. S13). Collectively, our success in establishing Bz-MS demonstrates that DeepCDM is a

highly versatile approach generalizable from one class of CDMs to another.

### DeepCDM-based spectral library for dansylated molecules

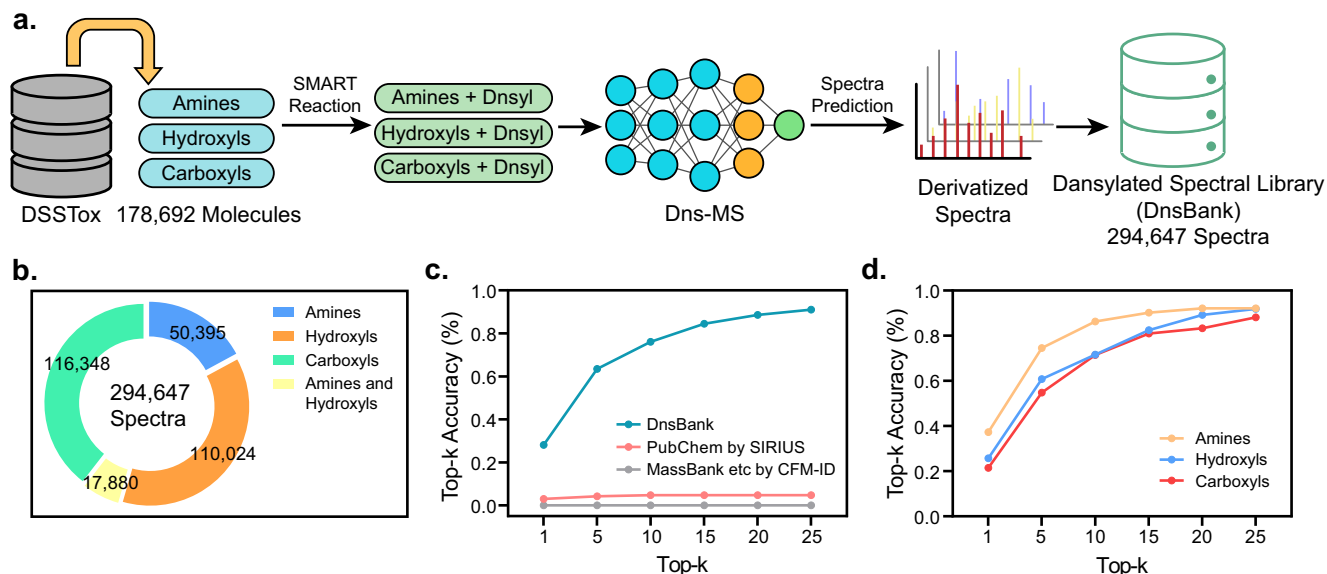
Haven demonstrated that DeepCDM is a highly robust and versatile approach to enable MS spectrum prediction models for desired class of CDMs, we next aim to establish a database that specialized for a given class of CDM. Here, a dansylation-specialized spectral library (DnsBank) was established, which contained 294,647 ESI-MS/MS spectra predicted by Dns-MS. To construct DnsBank, we first screened 178,629 molecules from a public chemical database, Distributed Structure-Searchable Toxicity (DSSTox)<sup>36</sup>, with three criteria, including (1) molecules that were composed by 11 elements: H, C, O, N, P, S, Cl, F, Br, I, Si; (2) molecules that were not ionized or containing ionized adducts; (3) molecules that contained amine, hydroxyl or carboxyl groups. The 178,692 molecules then underwent virtual dansylation reactions, where compounds containing amine or hydroxyl groups were derived using DnsCl and those containing carboxyl groups were derived using DnsHz. The virtual reactions were achieved by converting the Simplified Molecular-Input Line-Entry System (SMILES) of each molecule through "SMART reaction" by RDKit (<http://www.rdkit.org/>) (Fig. 3a). A total of 294,647 dansylated molecules were generated through virtual reactions, the spectra of which were then predicted using Dns-MS and stored in DnsBank. As such, DnsBank contained 50,395 spectra for dansylated amines (17.10%), 110,024 spectra for dansylated hydroxyls (37.34%), 116,348 predicted spectra for dansylated carboxyls (39.49%), and 17,880 spectra for didansylated molecules (6.07%) (Fig. 3b and Table S4).

The spectrum-to-compound (MS2C) capability of DnsBank was evaluated against the 167 experimental spectra of dansylated molecules using WCS as the measure. DnsBank returned correct annotation rates of 63.47% (106/167) in top 5 output and 91.02% in top 25, respectively (Fig. 3c). The highest correct annotation rate was found for dansylated amines (74.51% in top 5), followed by dansylated hydroxyls (60.81% in top 5) and carboxyls (54.76% in top 5) (Fig. 3d). By contrast, public databases (HMDB, MassBankJP/MassBankEU and MoNA) paired with CFM-ID 4.0 as a compound identification module failed to correctly annotate any dansylated molecules in the test set. Failure was also made when *in silico* spectral libraries (ChEBI, DSSTox and STOFF-IDENT) generated by CFM-ID 4.0 were used, though 3 dansylated molecules were found in these libraries upon manual checking (Supplementary Data-02). We also compared the searching result of DnsBank with that obtained using SIRIUS 4, which computes possible molecular fingerprints based on the MS/MS spectrum through a fragmentation tree and then searches the predicted fingerprints in chemical databases, such as PubChem<sup>37</sup>. Despite SIRIUS 4 drastically widens the coverage of chemicals, it only correctly annotated 7 dansylated molecules in top 5 output and 8 in top 25. Among the 159 incorrect annotations, 7 dansylated molecules were found in PubChem but not annotated by SIRIUS 4 and 13 were unable to be computed by SIRIUS (Supplementary Data-02). Collectively, our results demonstrated the superior MS2C performance of DnsBank over other tools for dansylated molecules, demonstrating the power of our DeepCDM for annotating CDMs.

### Wastewater analysis using DeepCDM-assisted non-targeted analysis

DeepCDM allows us to expand the annotation of CDMs beyond the space of known molecules, making it possible to explore novel chemical processes. To demonstrate this, we used DeepCDM to facilitate the discovery of new hazardous chemicals in authentic environmental samples.

Anthropogenic organic chemicals have improved the quality of human life but many also pose inherent dangers to human and



**Fig. 3 | Dnsylation-specialized in silico spectral library (DnsBank) constructed by DeepCDM.** **a** Structural information of molecules was selected and extracted from a public chemical database, Distributed Structure-Searchable Toxicity (DSSTox), which were then virtually derived. The spectra of these virtually dansylated molecules were predicted by the dansylation-specialized spectrum predicting model (Dns-MS) for constructing DnsBank. **b** The spectral composition of DnsBank. 17.10% spectra belonged to dansylated amine-containing molecules, 37.34% for dansylated hydroxyl-containing molecules, 39.39% for dansylated carboxyl molecules and 6.07% for didansylated molecules. **c** Spectrum-to-compound (MS2C) capability of DnsBank was benchmarked with public databases and compound identification tools using the hold-out test set. The top-k accuracy of annotations from 167 query spectra in the test set through searching in DnsBank, PubChem by SIRIUS 4, and 6 libraries by CFM-ID 4.0, including Human Metabolome Database (HMDB), MassBankJP/MassBankEU, the MassBank of North America

(MoNA), and the predicted libraries generated by CFM-ID 4.0 from ChEBI, DSSTox and STOFF-IDENT. 28.14% of the test set were correctly annotated at top 1 output by DnsBank, 63.47% were in top 5, 76.05% were in top 10 and 91.02% were ranked in top 25. CFM-ID 4.0 failed to annotate any dansylated molecules in the test set, although 3 molecules were found in its libraries by manual checking. SIRIUS 4 correctly annotated 7 dansylated molecules (4.19%) in top 5 output and 8 (4.79%) in top 25. In total, 15 dansylated molecules in the test set were found in PubChem by manual checking, 8 of them were annotated by SIRIUS 4. **d** The top-k accuracy of the 167 query spectra in test set for molecules containing amine, hydroxyl and carboxyl matched in DnsBank. Amine: 37.25% at top 1, 74.51% in top 5, 86.27% in top 10 and 92.16% in top 25. Hydroxyl: 25.68% at top 1, 60.81% in top 5, 71.62% in top 10 and 91.89% in top 25. Carboxyl: 21.43% at top 1, 54.76% in top 5, 71.43% in top 10 and 88.10% in top 25. Source data are provided as a Source Data file.

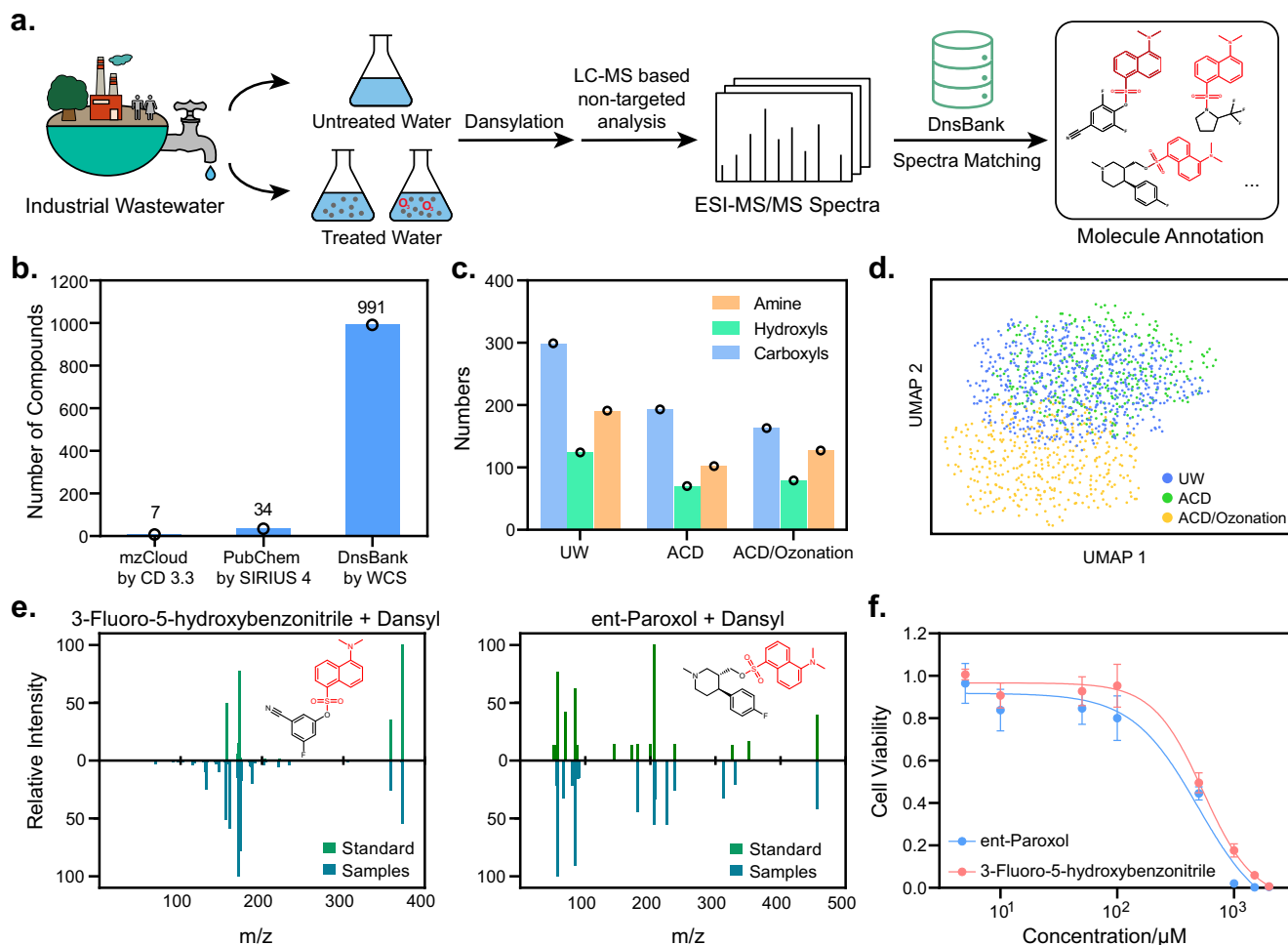
environment. Leather industry is an important contributor to pollutants emission due to the booming global market of leather goods<sup>38</sup>. Although industrial wastewater must be treated before emitting into environment, a significant amount of chemicals persist and can be further transformed during the treatment<sup>39</sup>. Therefore, high-throughput identification of massive organics in wastewater before and after treatments without prior knowledge is the first and foremost step for discovering chemicals that may eventually expose to human and biota.

We explored chemicals in wastewaters of leather industry before and after treatments by collecting and analyzing three types of samples, including untreated wastewater (UW), wastewater treated by active carbon adsorption (ACD), and that treated by ACD and ozonation (ACD/ozonation) (Fig. 4a). All water samples were first dansylated to enhance the retention and ionization of acids, phenolics, and aromatic amines that were heavily involved in the production of leather goods<sup>40,41</sup>. LC-MS based non-targeted analysis was then performed, where DnsBank was employed for spectrum matching and molecule annotation. Using  $WCS \geq 0.6$  as a threshold<sup>42</sup> since it provided reasonable number of annotations for further analysis (Fig. S14), we tentatively annotated 614 dansylated molecules in UW, 365 in ACD, and 369 in ACD/ozonation with confidence level 3 (Table 1). The average annotation efficiency was determined to be 19.42% (Table S5). A total of 991 unique dansylated molecules were annotated throughout all samples (Supplementary Data-03). As comparisons, SIRIUS 4<sup>37</sup> returned 34 tentative annotations (average annotation efficiency: 0.56%) and mzCloud (<https://www.mzcloud.org/>) paired with Compound Discoverer™ 3.3 (CD 3.3, Thermo Fisher Scientific) annotated

only 7 dansylated molecules (average annotation efficiency: 0.22%) for the same set of samples (Fig. 4b and Table S4).

The lower number of annotations in treated wastewater samples by either ACD or ACD/ozonation was consistent with the decreasing chemical oxygen demand (COD) (Fig. 4c and Table 1), suggesting the effective removal of chemicals from wastewater. However, there were 87 tentative annotations detected in all 3 samples, indicating the exist of persistent chemicals stable against ACD and ozonation treatments (Table S6). DeepCDM-assisted non-targeted analysis also revealed significant differences in treatment techniques, where ACD was largely overlapped with UW in uniform manifold approximation and projection (UMAP) analysis but ACD/ozonation showed a significant different profile from the other two samples (Fig. 4d). This result indicated that ozonation made a significant change to chemical compositions in wastewater with 142 unique annotations which were likely to be new transformation products (Table S7).

Of all 991 tentative annotations, there were many known pollutants, such as per- and polyfluoroalkyl substances (PFASs) and halogenated compounds (Table S8 and Fig. S15). We also found 865 chemicals with no toxicity record. A further investigation of this group of chemicals may lead to the discovery of new hazardous pollutants. We selected 4 annotations to confirm their identities using dansylated chemical standards and evaluate their cytotoxicity using Hep G2 cells through cell counting kit-8 (CCK-8) (Table S9). Two compounds, 3-fluoro-5-hydroxybenzotrile and ent-paroxol, were found to be of high cellular toxicity with the 50% inhibitory concentration (IC50) determined to be 545.9  $\mu$ M and 494.4  $\mu$ M, respectively (Figs. 4e, f and S16), thus could be new hazardous pollutants.



**Fig. 4 | DeepCDM-assisted environmental non-targeted analysis.** **a** Dansylation-specialized *in silico* spectral library (DnsBank) enabled confidence level 3 annotation of chemicals in liquid chromatography-mass spectrometry (LC-MS) based non-targeted analysis of leather industrial wastewater after dansylation. **b** Total annotation numbers of 3 wastewater samples obtained from DnsBank spectrum matching, SIRIUS 4 and Compound Discoverer™ 3.3 (CD 3.3) database searching. **c** Composition of tentative annotations in the untreated wastewater (UW), activate carbon adsorption (ACD) wastewater and ACD/ozonation wastewater obtained

from DnsBank. **d** The uniform manifold approximation and projection (UMAP) of chemicals detected in the 3 wastewater samples. **e** Identity confirmation of 3-fluoro-5-hydroxybenzonnitrile and ent-paroxol by matching tandem mass spectra with the dansylated chemical standards. **f** Concentration-response curves of Hep G2 cells incubating with 3-fluoro-5-hydroxybenzonnitrile and ent-paroxol. Each concentration had 6 biological replicates. Data are presented as mean values  $\pm$  standard deviation. Source data are provided as a Source Data file.

## Discussion

Current predicting tools are not designed to predict MS/MS spectra of CDMs. Here we introduced the DeepCDM workflow to establish a specialized model for CDMs *via* transfer learning a generic spectrum predicting model. Using dansylation and benzylation as model systems, we demonstrated that high-quality MS/MS spectrum prediction could be achieved using a small-sized training set of variable CDMs. Unlike existing MS/MS spectrum predicting tools, such as NEIMS and CFM-ID 4.0, that are generic for all molecules, DeepCDM aims to

enhance the spectrum predictability of a more focused group of molecules undergone given chemical derivatizations. Both Dns-MS and Bz-MS are highly specialized to dansylated molecules and benzyolated molecules, respectively, which improved upon generic tools in predicting ESI-MS/MS spectra for CDMs with over 150% enhanced WCS values. The superior performance of DeepCDM is achieved upon transfer learning that converts a generic tool into a highly specialized one for CDMs. For example, although both Dns-MS and Bz-MS were built upon the architecture of NEIMS, each of the two models demonstrated high predictability of its corresponding CDMs rather than the other group. Therefore, DeepCDM is a powerful addition to the current toolbox for MS/MS spectrum prediction. It is also worth noting that DeepCDM is not bound to NEIMS but can be adapted to any other exiting prediction algorithm for addressing specific challenges. Moreover, unlike many current predicting tools that work at a fixed collision energy<sup>21,23,43</sup>, ESI-MS/MS spectra predicted using DeepCDM are collision-energy-merged, which is more compatible with experimental conditions acquired by LC-MS based non-targeted analysis<sup>27–32</sup>.

Because of the remarkable MS2C capability, DeepCDM facilitates the annotation of CDMs. Current compound identification tools, such as SIRIUS 4, CFM-ID 4.0 compound identification module and CD 3.3

**Table 1 | Numbers of tentative annotations in 3 wastewater samples after dansylation**

Sample	Annotation No. in DnsCl labeled sample	Annotation No. in DnsHz labeled sample	Total annotations	COD (mg·L <sup>-1</sup> )
UW	315	299	614	442
ACD	172	193	365	310
ACD/ozonation	206	163	369	329

UW untreated water, ACD activated carbon adsorption, COD chemical oxygen demand.

access databases containing limited information of CDMs. Due to the enormous differences between MS/MS spectra of CDMs and spectra of underived molecules, annotating CDMs using these compound identification tools is difficult and low in efficiency. But the highly scalable DeepCDM enabled the construction of a dansylation-specialized in silico spectra library containing 294,647 ESI-MS/MS spectra. We also demonstrated that DeepCDM could be readily employed into non-targeted analysis workflow for high-throughput annotation of CDMs in complex environmental or biological samples.

The power of DeepCDM was further represented by analyzing authentic environmental samples. A total of 991 chemicals were annotated by DeepCDM upon non-targeted analysis with an average annotation efficiency 30 times better than that of SIRIUS 4 and 80 times better than that of CD 3.3. Besides the high annotation efficiency, DeepCDM also demonstrated an unbiased annotation among dansylated amines, hydroxyls, and carboxyls. By contrast, the two commercial tools failed to annotate any dansylated carboxylic acids<sup>41</sup>. CFM-ID 4.0 was not employed here because its server did not offer batch searching and its docker could not access public databases.

Once integrated to non-targeted environmental analysis, the highly efficient and unbiased nature of our DeepCDM ensures comprehensive monitoring of chemical processes and high-throughput discovery of new hazardous pollutants. Here DeepCDM reveals that ACD and ozonation are effectively strategies to remove organic pollutants. For instance, 2-[perfluoro(propoxy)ethoxy]acetic acid, a fluorinated alternative for polyfluorinated carboxylic acid (PFCA)s<sup>44,45</sup>, was removed by ACD and ozonation from leather industrial wastewater. As no removal techniques can eliminate all chemicals, many compounds, such as the carcinogenic 2,3,3,4,4,5,5-heptafluoropentanoic acid<sup>45</sup>, remain persistent during ACD and ozonation. In further investigation of new pollutants, we annotated and confirmed the identity of ent-paroxol as a persistent chemical, and its toxicity to Hep G2 cells suggested it could be a new hazardous pollutant. In addition to persistent pollutants, DeepCDM also reveals significant numbers of transformation products upon ozonation, which could direct an important source of potential pollutants. To name a few, 3-fluoro-5-hydroxybenzoxonitrile, detected in the ACD/ozonation wastewater and shown significant cellular toxicity, could contribute to the total toxicity of industrial effluent to human and environment as a toxic transformation product.

## Methods

### DeepCDM

First, spectra were downloaded from public MassBank of North American (MoNA, [ucdavis.edu](http://ucdavis.edu)) and filtered with criteria: (1) positive mode, (2) ESI ionization, (3) MoNA rank >4. Then the downloaded spectra was processed to merge all fragments produced at different collision energy for one molecule<sup>37,46</sup>. That produced a MoNA dataset containing 6790 merged electrospray ionization tandem mass spectra (ESI-MS/MS) of 6790 underived molecules, which was used to train the multilayer perception (MLP) of NEIMS<sup>20</sup>. In this step, MoNA dataset was chosen because it is an open database accessible to researchers across diverse research fields. Second, 585 chemical standards were labeled by dansyl chloride (DnsCl) and/or dansyl hydrazine (DnsHz) to produce 674 dansylated molecules, whose experimental merged spectra were acquired using HESI Q-Exactive Plus Orbitrap (Thermo Scientific, USA) at collision energies ranged from 10 to 50 eV. 507 experimental spectra of dansylated molecules were used for fine-tuning the trained MLP as the training set with 10% for validation and 167 spectra were holdout from the training set for model evaluation (test set). Dataset was randomly split without considering structural similarity of dansylated molecules. The MLP was fine-tuned by freezing all but the last layer in Dns-MS.

To modify DeepCDM for benzoylated molecules, the trained MLP by MoNA dataset was fine-tuned using 503 experimental spectra of benzoylated molecules with 10% for validation. A hold-out test set

containing 168 experimental spectra of benzoylated molecules was used for model evaluation. The benzoylated dataset was collected on HESI Q-Exactive Plus Orbitrap at collision energies ranged from 10 to 50 eV, using 671 benzoylated molecules labeled by benzoyl chloride (BzCl). Dataset was randomly split without considering structural similarity of benzoylated molecules. The MLP was also fine-tuned by freezing all but the last layer in Bz-MS.

### Model evaluation

3 metrics were tested for model evaluation: weighted cosine similarity (WCS), cosine similarity and spectra entropy. WCS was calculated using Eq. (1)<sup>47</sup>:

$$\text{Weighted Cosine Similarity}(I_q, I_l) = \frac{\sum_{k=1}^{M_{\max}} m_k^{0.5} I_{qk}^{0.5} \cdot m_k^{0.5} I_{lk}^{0.5}}{\sqrt{\sum_{k=1}^{M_q} (m_k^{0.5} I_{qk}^{0.5})^2} \sqrt{\sum_{k=1}^{M_l} (m_k^{0.5} I_{lk}^{0.5})^2}} \quad (1)$$

Cosine similarity was calculated from Eq. (2)<sup>47</sup>:

$$\text{Cosine Similarity}(I_q, I_l) = \frac{\sum_{k=1}^{M_{\max}} m_k I_{qk}^{0.5} \cdot m_k I_{lk}^{0.5}}{\sqrt{\sum_{k=1}^{M_q} (m_k I_{qk}^{0.5})^2} \sqrt{\sum_{k=1}^{M_l} (m_k I_{lk}^{0.5})^2}} \quad (2)$$

Where  $m_k$  was the mass-to-charge ratio (m/z) of fragment;  $I_{qk}$  and  $I_{lk}$  was the intensity of fragment (m/z = k) in the query spectrum and in the reference spectrum;  $M_l$  and  $M_q$  were the largest nonzero indices of  $I_q$  and  $I_l$ ;  $M_{\max}$  was the largest index of  $M_l$  and  $M_q$ .

Spectra entropy  $S$  was calculated from all ion intensities  $I_p$  by Eq. (3)<sup>48</sup>:

$$S = - \sum_p I_p \ln I_p \quad (3)$$

Consistent results were found among the 3 metrics in the comparison of 3 models with slightly better WCSs, using the hold-out test set of 167 dansylated molecules. Therefore, WCS was selected in this study to evaluate the predictability of models.

### Virtual reaction and library building

A structural dataset was downloaded from Distributed Structure-Searchable Toxicity (DSSTox, [https://epa.figshare.com/articles/dataset/Chemistry\\_Dashboard\\_Data\\_DSSTox\\_SDF/5491516](https://epa.figshare.com/articles/dataset/Chemistry_Dashboard_Data_DSSTox_SDF/5491516), accessed on March 31, 2022). The DSSTox dataset were screen for chemicals with amine, hydroxyl and carboxyl groups. Then, the structures of dansylated molecules were computed by RDKit (<http://www.rdkit.org/>) using a derivatization Python script "SMART reaction" and output in Simplified Molecular Input Line Entry System (SMILES). The "SMART reaction" simulated dansylation to reacting groups using defined templates: DnsCl for amine and hydroxyl, DnsHz for carboxyl. The dansylated SMILES structures were exported in SDF files, then input to Dns-MS to predict collision-energy-merged ESI-MS/MS spectra to 1 Da resolution. The predicted spectra were accumulated for DnsBank (Table S4).

### Benchmarks

C2MS benchmarks between Dns-MS, Bz-MS, NEIMS and CFM-ID 4.0 (web server) were performed on hold-out test sets. NEIMS was acquired from ref. 20. and CFM-ID 4.0 (web server, <http://cfmid.wishartlab.com/predict>) was used for spectrum prediction.

MS2C benchmark was performed between DnsBank and other public spectral libraries accessed by CFM-ID 4.0 compound identification module (<http://cfmid.wishartlab.com/identify>) and PubChem accessed by SIRIUS 4 on the hold-out test set containing 167 dansylated spectra. In silico spectra libraries of ChEBI, DSSTox and

STOFF-IDENT predicted by CFM-ID 4.0 and public spectral libraries including HMDB, MassBank and MoNA were used by CFM-ID 4.0 compound identification module for annotating dansylated molecules through spectrum matching. ESI was selected as the spectrum type, ion mode was positive and adduct type was set to  $[M + H]^+$  in CFM-ID 4.0, with 10 ppm mass tolerance and DotProduct as the scoring function. For molecule annotation by SIRIUS 4, parameters were set to be instrument: orbitrap, database: PubChem, possible ionization:  $[M + H]^+$ . CSI:FingerID was used for fingerprint prediction and structure database search.

### Dansylation of chemical standards

For DnsCl reaction, amine- or hydroxyl-containing chemicals were dissolved in 50% acetonitrile to a concentration of 10 mM individually. 30  $\mu$ L standard solution was mixed with 15  $\mu$ L acetonitrile, 15  $\mu$ L sodium carbonate/sodium bicarbonate buffer (0.5 M, pH 9.4) and 30  $\mu$ L 20 mg·mL<sup>-1</sup> DnsCl. The reaction solution was shaken at 60 °C for 60 min before quenching by 10  $\mu$ L 250 mM sodium hydroxide. The excess sodium hydroxide was neutralized by 50  $\mu$ L 425 mM formic acid.

For DnsHz reaction, carboxyl chemicals were dissolved in acetonitrile to 10 mM individually. 25  $\mu$ L standard solution was mixed with 25  $\mu$ L 50 mM 1-ethyl-3(3-(dimethylamino)propyl)carbodiimide, 25  $\mu$ L 10 mM 1-hydroxy-7-azabenzotriazole in 2-(N-morpholino)ethanesulfonic acid buffer and 25  $\mu$ L 10 mg mL<sup>-1</sup> DnsHz. The reaction solution was shaken at 25 °C for 90 min before quenching at -80 °C for 10 min.

### Benzoylation of chemical standards

For benzoyl chloride (BzCl) reaction, amine- or hydroxyl-containing chemicals were dissolved in 50% acetonitrile to a concentration of 10 mM individually. 200  $\mu$ L standard solution was mixed with 100  $\mu$ L sodium carbonate buffer (0.1 M, pH = 9.5), 100  $\mu$ L 2% BzCl and 1  $\mu$ L formic acid at room temperature for 5 min.

### MS analysis

MS and tandem MS (MS/MS) spectra of dansylated molecules were acquired by HESI Q-Exactive Plus Orbitrap (Thermo Scientific, USA) in positive mode. Precursor ions were selected for MS/MS scan with  $\pm 0.25$  Da mass tolerance. 10 MS/MS scans were accumulated to generate a MS/MS spectrum with collision energies ranged from 10 to 50 eV.

### Wastewater treatments

The leather industrial wastewater was first treated with activated carbon adsorption (ACD) and then ozonation. 100 mL leather industrial wastewater was diluted 10 times with high-performance liquid chromatography (HPLC) graded water and then mixed with 20 g activated carbon granules followed by incubating at 300 r min<sup>-1</sup> for 24 h at room temperature. After filtration, 500 mL ACD treated wastewater was reacted with 500 mL 64.29  $\pm$  1.36 mg L<sup>-1</sup> ozone solution at room temperature for 24 h. The chemical oxygen demand (COD) of wastewater were measured before and after treatments.

### Sample handling

The wastewater samples were extracted with Supelclean™ ENVI™-18 solid phase extraction cartridge (500 mg, 3 mL, Sigma-Aldrich, USA). 40  $\mu$ L eluent was divided into two aliquots: 20  $\mu$ L was reacted with DnsCl and 20  $\mu$ L with DnsHz. After dansylation, the samples were desalted using HPLC (Agilent 1260 Infinity II, USA) with C18 column (Supelco, 2.1 mm  $\times$  50 mm, 5  $\mu$ m) at 25 °C, flow rate: 1 mL/min. The mobile phase A was 0.1% (v/v) formic acid in water and mobile phase B was 0.1% (v/v) formic acid in 95% acetonitrile. Liquid chromatography gradient was:  $t = 0$  min, 5% B;  $t = 3$  min, 5% B;  $t = 3.01$  min, 95% B;  $t = 5$  min, 95% B;  $t = 5.01$  min, 5% B;  $t = 8$  min, 5% B.

### Non-targeted LC-MS analysis

Non-targeted analysis was performed using Dionex Ultimate 3000 UHPLC coupled to HESI Q-Exactive Plus Orbitrap mass spectrometer (Thermo Scientific, USA). A C18 column (Thermo Scientific, 2.1 mm  $\times$  100 mm, 1.8  $\mu$ m) was used at 25 °C, with a flow rate of 300  $\mu$ L min<sup>-1</sup>. Mobile phase A was 0.1% (v/v) formic acid in water, mobile phase B was 0.1% (v/v) formic acid in acetonitrile. Liquid chromatography gradient was:  $t = 0$  min, 5% B;  $t = 3$  min, 5% B;  $t = 4$  min, 15% B;  $t = 14$  min, 65% B;  $t = 16$  min, 95% B;  $t = 18$  min, 95% B;  $t = 18.01$  min, 5% B;  $t = 20$  min, 5% B. MS/MS spectra were accumulated by 10 MS/MS scans with collision energies ranged from 10 to 50 eV.

### Data processing

Molecule classification based on structural features was performed using ClassyFire<sup>49</sup>. Peak lists of non-targeted data were extracted using MZmine 3<sup>50</sup>. Molecule annotation at confidence level 3 was achieved by matching exact mass and MS/MS spectra in DnsBank. The annotation confidence level was complied with small molecule identifications defined by Schymanski, et al., in 2014<sup>14</sup>. With the spectra matching threshold  $WCS \geq 0.6$ <sup>42</sup>, annotations were exported in csv file. Annotation efficiency was calculated to evaluate the percentage of query spectra confidently assigned to molecules by spectra matching or database searching<sup>22</sup>.

$$\text{Annotation Efficiency} = \frac{\text{number of molecule annotations}}{\text{number of query spectra}} \quad (4)$$

### Statistics

The unpaired *t*-test was performed in Prism (GraphPad Software, USA) to compare spectral WCS between two groups of data. For comparisons between three or more groups, ordinary one-way ANOVA was performed. The uniform manifold approximation and projection (UMAP) analysis was performed to analyze the distances between organic profiles of 3 wastewater samples. For all analyses, a two-tailed *p* value < 0.05 was considered to be statistically significant.

### In vitro cytotoxicity study

The cell counting kit-8 (CCK-8) assay was used to monitor the cell viability of human hepatocellular cell line (Hep G2) in 96-wells incubating with various concentrations of chemicals for 72 h. The endpoint absorbance of each well was measured 6 times at 450 nm by Epoch 2 microplate spectrophotometer (Agilent BioTek, USA).

### Implementation details

All DeepCDM models were constructed and trained in Tensorflow<sup>51</sup>. The hyperparameters were optimized on the hidden units and layers of MLP. Training hyperparameters were listed in Table S2. The optimized MLP architecture using ReLU activation had 3 layers of 500 nodes. Extended-Connectivity Fingerprints (ECFP4) with a length of 4096 and radius of 2 were set for the input of molecules. Further details on hyperparameter configurations were included in the code repository and can be obtained through Github (<http://github.com/ADNLab-SCU/DeepCDMs>).

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

The processed MoNA data obtained from public source, the demo datasets containing experimental acquired spectra of dansylated molecules and the raw data of wastewater samples have been deposited in Zenodo at <https://doi.org/10.5281/zenodo.13382796> (Ref. 52). The processed data generated in this study are provided in the

Supplementary Information/Source Data file. Source data are provided with this paper.

## Code availability

All code created in this work is available on GitHub (<http://github.com/ADNLab-SCU/DeepCDMs>)<sup>53</sup> with a DOI of <https://doi.org/10.5281/zenodo.13381990> (Ref. 54).

## References

- Zhang, T.-Y. et al. Derivatization for liquid chromatography-electrospray ionization-mass spectrometry analysis of small-molecular weight compounds. *TrAC Trends Anal. Chem.* **119**, 115608 (2019).
- Shanaiah, N. et al. Class selection of amino acid metabolites in body fluids using chemical derivatization and their enhanced <sup>13</sup>C NMR. *Proc. Natl. Acad. Sci.* **104**, 11540–11544 (2007).
- Lin, L., Sheng, J. & Huang, Z. Nucleic acid X-ray crystallography via direct selenium derivatization. *Chem. Soc. Rev.* **40**, 4591–4602 (2011).
- Keshishian, H. et al. Quantitative, multiplexed workflow for deep analysis of human blood plasma and biomarker discovery by mass spectrometry. *Nat. Protoc.* **12**, 1683–1701 (2017).
- Zhang, W. et al. Online photochemical derivatization enables comprehensive mass spectrometric analysis of unsaturated phospholipid isomers. *Nat. Commun.* **10**, 79 (2019).
- Slieker, R. C. et al. Identification of biomarkers for glycaemic deterioration in type 2 diabetes. *Nat. Commun.* **14**, 2533 (2023).
- Arifuzzaman, M. et al. Inulin fibre promotes microbiota-derived bile acids and type 2 inflammation. *Nature* **611**, 578–584 (2022).
- Pruss, K. M. et al. Host-microbe co-metabolism via MCAD generates circulating metabolites including hippuric acid. *Nat. Commun.* **14**, 512 (2023).
- Zhao, S. & Li, L. Chemical derivatization in LC-MS-based metabolomics study. *TrAC Trends Anal. Chem.* **131**, 115988 (2020).
- Hollender, J. et al. Nontarget screening with high resolution mass spectrometry in the environment: ready to go? *Environ. Sci. Technol.* **51**, 11505–11512 (2017).
- Wang, B. et al. Derivatization for nontargeted screening of acids in oilfield refinery wastewater: identification and behaviors of recalcitrant chlorinated naphthenic acids. *Environ. Sci. Technol.* **53**, 1022–1030 (2019).
- Wang, B. et al. Evaluating a tap water contamination incident attributed to oil contamination by nontargeted screening strategies. *Environ. Sci. Technol.* **50**, 2956–2963 (2016).
- Sumner, L. W. et al. Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG). *Metabolomics Standards Initiative (MSI)*. *Metabolomics* **3**, 211–221 (2007).
- Schymanski, E. L. et al. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ. Sci. Technol.* **48**, 2097–2098 (2014).
- Liu, Z. et al. Stable isotopic labeling and nontarget identification of nanogram/liter amino contaminants in water. *Anal. Chem.* **91**, 13213–13221 (2019).
- Zhao, T. et al. HDPairFinder: a data processing platform for hydrogen/deuterium isotopic labeling-based nontargeted analysis of trace-level amino-containing chemicals in environmental water. *J. Environ. Sci.* **136**, 583–593 (2024).
- Zou, Z. et al. A deep learning model for predicting selected organic molecular spectra. *Nat. Comput. Sci.* **3**, 957–964 (2023).
- Koopman, J. & Grimme, S. From QCEIMS to QCxMS: a tool to routinely calculate CID mass spectra using molecular dynamics. *J. Am. Soc. Mass Spectrom.* **32**, 1735–1751 (2021).
- Wang, S. et al. Predicting in silico electron ionization mass spectra using quantum chemistry. *J. Cheminformatics* **12**, 63 (2020).
- Wei, J. N. et al. Rapid prediction of electron-ionization mass spectrometry using neural networks. *ACS Cent. Sci.* **5**, 700–708 (2019).
- Allen, F., Greiner, R. & Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **11**, 98–110 (2015).
- Djombou-Feunang, Y. et al. CFM-ID 3.0: significantly improved ESI-MS/MS prediction and compound identification. *Metabolites* **9**, 72 (2019).
- Wang, F. et al. CFM-ID 4.0 – a web server for accurate MS-based metabolite identification. *Nucleic Acids Res.* **50**, W165–W174 (2022).
- Guo, K. & Li, L. Differential <sup>12</sup>C-/<sup>13</sup>C-isotope dansylation labeling and fast liquid chromatography/mass spectrometry for absolute and relative quantification of the metabolome. *Anal. Chem.* **81**, 3919–3932 (2009).
- Huan, T. et al. DnsID in MyCompoundID for rapid identification of dansylated Amine- and Phenol-containing metabolites in LC-MS-based metabolomics. *Anal. Chem.* **87**, 9838–9845 (2015).
- Horai, H. et al. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
- Ledvina, A. R. et al. Increased throughput of proteomics analysis by multiplexing high-resolution tandem mass spectra. *Anal. Chem.* **83**, 7651–7656 (2011).
- Blaženović, I. et al. Structure annotation of all mass spectra in untargeted metabolomics. *Anal. Chem.* **91**, 2155–2162 (2019).
- Mattoli, L. et al. Suspect screening analysis to improve untargeted and targeted UHPLC-qToF approaches: the biodegradability of a proton pump inhibitor medicine and a natural medical device. *Sci. Rep.* **14**, 51 (2024).
- Yang, Y. et al. Data acquisition methods for non-targeted screening in environmental analysis. *TrAC Trends Anal. Chem.* **160**, 116966 (2023).
- Wright Muelas, M. et al. An untargeted metabolomics strategy to measure differences in metabolite uptake and excretion by mammalian cell lines. *Metabolomics* **16**, 107 (2020).
- Klávus, A. et al. “Notame”: workflow for non-targeted LC-MS metabolic profiling. *Metabolites* **10**, 135 (2020).
- Özdestan, Ö. & Üren, A. A method for benzoyl chloride derivatization of biogenic amines for high performance liquid chromatography. *Talanta* **78**, 1321–1326 (2009).
- Song, P. et al. In vivo neurochemical monitoring using benzoyl chloride derivatization and liquid chromatography-mass spectrometry. *Anal. Chem.* **84**, 412–419 (2012).
- Widner, B. et al. Quantification of Amine- and alcohol-containing metabolites in saline samples using pre-extraction benzoyl chloride derivatization and ultrahigh performance liquid chromatography tandem mass spectrometry (UHPLC MS/MS). *Anal. Chem.* **93**, 4809–4817 (2021).
- Grulke, C. M. et al. EPA’s DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research. *Comput. Toxicol.* **12**, 100096 (2019).
- Dührkop, K. et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
- Leather Goods Market Size, Share & Trends Analysis Report By Type (Genuine Leather, Synthetic Leather, Vegan Leather), By Product, By Region, And Segment Forecasts, 2023–2030*. Report ID: GVR-3-68038-061-3, <https://www.grandviewresearch.com/industry-analysis/leather-goods-market>.
- Lu, L. et al. Wastewater treatment for carbon capture and utilization. *Nat. Sustain.* **1**, 750–758 (2018).

40. China, C. R. et al. Alternative tanning technologies and their suitability in curbing environmental pollution from the leather industry: a comprehensive review. *Chemosphere* **254**, 126804 (2020).
41. Rydin, S. Chemicals in leather: international trends on risk-based control and management. in *Global risk-based management of chemical additives II: risk-based assessment and management strategies* (eds. Bilitewski, B., Darbra, R. M. & Barceló, D.) 245–262 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013). [https://doi.org/10.1007/698\\_2012\\_201](https://doi.org/10.1007/698_2012_201).
42. Treen, D. G. C. et al. SIMILE enables alignment of tandem mass spectra with statistical significance. *Nat. Commun.* **13**, 2510 (2022).
43. Wang, F. et al. Deep learning-enabled MS/MS spectrum prediction facilitates automated identification of novel psychoactive substances. *Anal. Chem.* **95**, 18326–18334 (2023).
44. Wang, Z. et al. A never-ending story of per- and polyfluoroalkyl substances (PFASs)? *Environ. Sci. Technol.* **51**, 2508–2518 (2017).
45. Evich, M. G. et al. Per- and polyfluoroalkyl substances in the environment. *Science* **375**, eabg9065 (2022).
46. Dührkop, K. et al. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci.* **112**, 12580–12585 (2015).
47. Stein, S. E. & Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **5**, 859–866 (1994).
48. Li, Y. et al. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nat. Methods* **18**, 1524–1531 (2021).
49. Djoumbou Feunang, Y. et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminformatics* **8**, 61 (2016).
50. Schmid, R. et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nat. Biotechnol.* **41**, 447–449 (2023).
51. Abadi, M. et al. TensorFlow: a system for large-scale machine learning. in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation* 265–283 (USENIX Association, USA, 2016).
52. Chen, B. et al. Deep learning prediction of electrospray ionization tandem mass spectra of chemically derived molecules. ADNLab-SCU. Public Data files for DeepCDM. *Zenodo* <https://doi.org/10.5281/zenodo.13382796> (2024).
53. Chen, B. et al. Deep learning prediction of electrospray ionization tandem mass spectra of chemically derived molecules. ADNLab-SCU. DeepCDM. *GitHub* <http://github.com/ADNLab-SCU/DeepCDMs> (2024).
54. Chen, B. et al. Deep learning prediction of electrospray ionization tandem mass spectra of chemically derived molecules. ADNLab-SCU. ADNLab-SCU/DeepCDMs: DeepCDM. *Zenodo* <https://doi.org/10.5281/zenodo.13381990> (2024).
55. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
56. Jablonka, K. M. et al. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **6**, 161–169 (2024).

## Acknowledgements

The study was supported by the National Natural Science Foundation of China (22006104, Y.T. and 22074099, F.L.), the Fundamental Research

Funds for the Central Universities (No. YJ201975, F.L.), and the Institutional Research Fund from Sichuan University (2021SCUNL105, F.L.). The authors would like to thank Prof. Bin Yan (College of Biomass Science and Engineering, Sichuan University) for kindly providing us with leather industrial wastewater samples and the Analytical & Testing Center of Sichuan University for mass spectrometry analysis.

## Author contributions

Y.T. and F.L. conceived the idea, designed all experiments, and supervised the overall project. Y.T. designed DeepCDM approach. B.C. and H.L. derivatized chemical standards and performed MS analysis. B.C. and H.L. trained, validated and evaluated Dns-MS and Bz-MS for danylated molecules and benzoylated molecules respectively. B.C. constructed DnsBank. R.H. treated the leather industrial wastewater. Y.T. and B.C. performed LC-MS analysis of wastewater samples and analyzed data. H.L. performed in vitro cytotoxicity experiments.

## Competing interests

Yanan Tang, Bin Chen, and Feng Li are inventors on a pending patent for DeepCDM (Patent Number: CN202410365114.X). The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-52805-5>.

**Correspondence** and requests for materials should be addressed to Yanan Tang or Feng Li.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024