

## Supplementary Materials

### S1. Feature Selection Models

- The Multi-Layer Perceptron (MLP) model used for  $m/z$  bin selection contains three ReLU-activated layers (256, 128, 64 nodes) and a sigmoid-activated dense layer. The learning rate is set to 0.001.
- The Convolutional Neural Network (CNN) model used for fingerprint selection contains four ReLU-activated 1D convolutional layers (32, 64, 128, and 256 nodes respectively), a dropout layer, a data flattening layer, three LeakyReLU-activated dense layers (256, 512 and 1,024 nodes), and a sigmoid-activated dense layer (4,606 nodes). The model learning rate is set to 0.0005.

### S2. Hyperparameters of Deep Learning Models for Molecular Fingerprint Prediction

We considered the following deep learning models:

- a) Deep Neural Network (DNN): The input vectors first go through a data flattening process, ensuring the data are in the correct format for the fully connected layers. This is followed by three fully connected deep learning layers with 1,024, 2,048, and 4,096 nodes, respectively, designed to extract and learn increasingly complex features. Each deep learning layer is followed by a LeakyReLU activation function to prevent the vanishing gradient problem and improve learning, as well as a dropout layer to reduce overfitting by randomly deactivating neurons during training. The model ends with a sigmoid-activated dense layer consisting of 4,606 nodes, producing class probabilities. The learning rate is set to 0.0005. The model uses the Adam optimizer, which combines adaptive learning rates and momentum to enhance training efficiency and performance.
- b) CNN: The input vectors go through four ReLU-activated 1D convolutional layers with 32, 64, 128, and 256 nodes, respectively. Each convolutional layer has a kernel size of 3 and includes L2 kernel regularization (set to 0.001) to prevent overfitting by penalizing large weights. These layers are followed by max-pooling layers to reduce dimensionality while retaining the most important features. After the convolutional layers, a dropout layer with a rate of 0.1 is applied to further mitigate overfitting. The data then pass through a flattening layer and three LeakyReLU-activated dense layers with 256, 512, and 1,024 nodes, respectively, enabling robust feature extraction. The model ends with a sigmoid-activated dense layer containing 4,606 nodes. The learning rate is set to 0.0005. The Adam optimizer is used to achieve efficient optimization. This architecture is optimized for extracting local patterns from the input data.
- c) Recurrent Neural Network in Long Short-Term Memory Architecture (RNN/LSTM): The input vectors first go through three LSTM layers with 64, 128, and 256 nodes, each with an L2 kernel regularizer set to 0.001 to prevent overfitting. Dropout layers are applied after each LSTM layer to enhance regularization by reducing the risk of co-adaptation of neurons during training. These LSTM layers are designed to capture sequential and contextual relationships in the data. Following the LSTM layers, the data are passed through a fully connected dense layer with 512 nodes, activated by a LeakyReLU function, and are equipped with an L2 kernel regularizer set to 0.001. A dropout layer is applied after this dense layer for additional regularization. The model ends with a sigmoid-activated dense layer consisting of 4,606 nodes, providing the final probability outputs. The learning rate is set to 0.0005. The model uses the Adam optimizer for training. This architecture is particularly suitable for tasks that involve sequential or temporal dependencies in the input data.

**Table S1.** Top-1 ranking accuracy of candidates for the challenges in the CASMI 2016, CASMI 2017 and CASMI 2022 benchmark datasets based on formula prediction by SIRIUS.

	CASMI 2016	CASMI 2017	CASMI 2022
Positive	81%	17%	60%
Negative	60%	29%	43%

**Table S2.** Top-*k* ranking results for a subset of the challenges in the CASMI 2016 dataset whose MS/MS spectra were acquired in the positive ionization mode. The best result in each of the testing conditions is indicated in bold font.

Method	Formula Unknown				Formula Predicted				Formula Known				
	Random	DNN	CNN	RNN	DNN	CNN	RNN	CSI:FingerID	Random	DNN	CNN	RNN	CSI:FingerID
<b>Top 1</b>	1%	4%	22%	<b>26%</b>	28%	23%	25%	<b>39%</b>	15%	43%	<b>46%</b>	<b>46%</b>	43%
<b>Top 3</b>	8%	13%	28%	<b>30%</b>	46%	33%	31%	<b>56%</b>	24%	58%	57%	57%	<b>61%</b>
<b>Top 5</b>	14%	20%	34%	<b>35%</b>	51%	37%	25%	<b>58%</b>	30%	62%	64%	62%	<b>66%</b>
<b>Top 10</b>	29%	37%	<b>46%</b>	<b>46%</b>	59%	52%	50%	<b>63%</b>	35%	70%	<b>74%</b>	<b>74%</b>	72%

**Table S3.** Top-*k* ranking results for a subset of the challenges in the CASMI 2016 dataset whose MS/MS spectra were acquired in the negative ionization mode. The best result in each of the testing conditions is indicated in bold font.

Method	Formula Unknown				Formula Predicted				Formula Known				
	Random	DNN	CNN	RNN	DNN	CNN	RNN	CSI:FingerID	Random	DNN	CNN	RNN	CSI:FingerID
<b>Top 1</b>	3%	19%	25%	<b>28%</b>	26%	28%	30%	<b>60%</b>	13%	40%	46%	48%	<b>73%</b>
<b>Top 3</b>	11%	28%	31%	<b>35%</b>	43%	38%	36%	<b>63%</b>	32%	63%	64%	67%	<b>80%</b>
<b>Top 5</b>	18%	37%	35%	<b>38%</b>	53%	44%	47%	<b>63%</b>	36%	69%	72%	75%	<b>80%</b>
<b>Top 10</b>	25%	44%	44%	<b>49%</b>	<b>69%</b>	56%	57%	63%	41%	78%	81%	<b>85%</b>	80%

**Table S4.** Top-*k* ranking results for a subset of the challenges in the CASMI 2017 dataset whose MS/MS spectra were acquired in the positive ionization mode. The best result in each of the testing conditions is indicated in bold font.

Method	Formula Unknown				Formula Predicted				Formula Known				
	Random	DNN	CNN	RNN	DNN	CNN	RNN	CSI:FingerID	Random	DNN	CNN	RNN	CSI:FingerID
<b>Top 1</b>	4%	6%	18%	<b>33%</b>	6%	<b>46%</b>	30%	9%	7%	18%	<b>49%</b>	36%	44%
<b>Top 3</b>	13%	9%	23%	<b>33%</b>	11%	<b>54%</b>	31%	14%	17%	25%	<b>58%</b>	42%	57%
<b>Top 5</b>	16%	13%	26%	<b>35%</b>	14%	<b>57%</b>	33%	14%	24%	33%	59%	45%	<b>62%</b>
<b>Top 10</b>	26%	22%	30%	<b>40%</b>	24%	<b>61%</b>	39%	15%	38%	40%	62%	50%	<b>65%</b>

**Table S5.** Top-*k* ranking results for a subset of the challenges in the CASMI 2017 dataset whose MS/MS spectra were acquired in the negative ionization mode. The best result in each of the testing conditions is indicated in bold font.

Method	Formula Unknown				Formula Predicted				Formula Known				
	Random	DNN	CNN	RNN	DNN	CNN	RNN	CSI:FingerID	Random	DNN	CNN	RNN	CSI:FingerID
<b>Top 1</b>	2%	10%	14%	<b>19%</b>	13%	15%	19%	<b>23%</b>	5%	22%	22%	<b>28%</b>	27%
<b>Top 3</b>	9%	13%	<b>23%</b>	22%	20%	27%	24%	<b>29%</b>	15%	30%	29%	<b>35%</b>	34%
<b>Top 5</b>	18%	13%	<b>27%</b>	24%	26%	<b>33%</b>	27%	31%	26%	41%	40%	<b>43%</b>	38%
<b>Top 10</b>	30%	20%	31%	<b>32%</b>	31%	<b>37%</b>	34%	33%	34%	50%	49%	<b>55%</b>	43%

**Table S6.** Top-*k* ranking results for a subset of the challenges in the CASMI 2022 dataset whose MS/MS spectra were acquired in the positive ionization mode. The best result in each of the testing conditions is indicated in bold font.

Method	Formula Unknown				Formula Predicted				Formula Known				
	Random	DNN	CNN	RNN	DNN	CNN	RNN	CSI:FingerID	Random	DNN	CNN	RNN	CSI:FingerID
<b>Top 1</b>	3%	6%	30%	<b>35%</b>	11%	27%	<b>36%</b>	13%	9%	27%	47%	<b>56%</b>	18%
<b>Top 3</b>	8%	19%	37%	<b>48%</b>	28%	32%	<b>40%</b>	21%	21%	43%	54%	<b>60%</b>	28%
<b>Top 5</b>	15%	25%	39%	<b>50%</b>	33%	36%	<b>43%</b>	23%	28%	49%	56%	<b>62%</b>	31%
<b>Top 10</b>	23%	32%	51%	<b>55%</b>	44%	50%	<b>51%</b>	27%	43%	54%	67%	<b>70%</b>	34%

**Table S7.** Top-*k* ranking results for a subset of the challenges in the CASMI 2022 dataset whose MS/MS spectra were acquired in the negative ionization mode. The best result in each of the testing conditions is indicated in bold font.

Method	Formula Unknown				Formula Predicted				Formula Known				
	Random	DNN	CNN	RNN	DNN	CNN	RNN	CSI:FingerID	Random	DNN	CNN	RNN	CSI:FingerID
<b>Top 1</b>	6%	15%	13%	<b>23%</b>	13%	10%	<b>21%</b>	1%	15%	41%	36%	<b>46%</b>	7%
<b>Top 3</b>	12%	18%	<b>28%</b>	<b>28%</b>	15%	23%	<b>26%</b>	2%	25%	46%	<b>51%</b>	49%	13%
<b>Top 5</b>	18%	26%	<b>36%</b>	<b>36%</b>	23%	<b>33%</b>	<b>33%</b>	3%	30%	54%	<b>64%</b>	62%	16%
<b>Top 10</b>	34%	31%	<b>51%</b>	41%	31%	<b>46%</b>	38%	4%	41%	59%	<b>64%</b>	62%	20%