# Mixture-of-Experts Ensemble with Hierarchical Deep Metric Learning for Spectroscopic Identification

**Masaki Adachi**
University of Oxford
Toyota Motor Corporation
masaki@robots.ox.ac.uk

## Abstract

A mixture-of-experts ensemble of hierarchical deep metric learning models is introduced in order to identify materials from X-ray diffraction spectra. In previous studies, the identification accuracy of the 1D convolutional neural networks model deteriorates significantly as the number of classes increases. To overcome this problem, a hierarchical deep metric learning model was developed that can identify approximately 10,000 classes with an average top-1 accuracy of 87%. Furthermore, this new model was employed to create expert models for 73 general chemical elements, which in turn were used to construct a mixture-of-experts ensemble. This ensemble model successfully identified materials from 136,899 classes with a top-1 accuracy of 98%.

## 1 Introduction

Spectroscopic identification plays a vital role in various scientific disciplines, such as forensic science [1], materials science [2], astrophysics [3], and earth science [4]. For example, a tiny piece left at a crime scene can break an alibi in a murder case, leading to the identification of the culprit. A material synthesized by a new method can only be proven to be genuinely unique by identifying its crystal structure via spectral analysis. Nonetheless, in many cases, a myriad of spectra from planetary probes, deep-sea explorers, or synchrotron radiation facilities have left or just been used for visualisation without fully exploiting the data, which could possibly lead to more scientific discovery [4].

The reason is that this spectroscopic identification is still not possible without human intervention. The accuracy is determined mainly by the following two factors: the search algorithm of the spectral database, and how noiseless the spectra are. XRD (X-Ray Diffraction) [5] is a standard spectral analysis method employed to identify inorganic materials by matching their spectra with the database. Each material has a unique spectral pattern, the so-called "fingerprint spectra" [6], which enables human experts to identify the material. However, complex noise patterns, such as pattern shifts or peak losses [7], are often contained in samples with insufficient quantities or fast-scanned cases, deteriorating the identification accuracy by a significant amount.

In recent years, 1D-CNN (One Dimensional Convolutional Neural Networks) has been developing spectroscopic identification models, which have attracted significant attention [7–10]. 1D-CNN is capable of robust inference under such noisy spectra quickly, usually in the order of milliseconds. Still, the number of identifiable materials by the previous models has been limited to around 100 [7, 8]. In actual usage, identifying all 136,899 materials registered in the ICSD (Inorganic Crystal Structure Database) [11] is essential.

For cases with a vast number of classes, deep metric learning models have been applied, as is well known for the case of face recognition tasks [12–14]. Among the models, AdaCos [14] is considered

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

one of the state-of-the-art models in this field. In this way, a combination of 1D-CNN and AdaCos was expected to address the problem of large-scale databases.

Nevertheless, the study presented here was able to determine that the combination of vanilla 1D-CNN and AdaCos alone could not identify more than a three-digit number of classes. Instead, the following additional techniques were crucial in order to make identifiable classes 100 times larger: 1D-RegNet [15, 16] and hierarchical deep metric learning. In this case, using AdaCos as well as static angular penalty softmax loss [12] within the two-stage hierarchical structure, made it possible to identify around 10,000 classes with a mean top-1 accuracy of 87%. Yet, even for this hierarchical model, identifying a six-digit number of classes was still challenging. Therefore, by constructing an ensemble of the above models trained to identify specific compounds as an MoE (Mixture-of-Experts) [17], it has been enabled to identify 136,899 materials with 98% top-1 accuracy.

The main contributions in this report are as follows:

- A hierarchical deep metric learning model for up to a five-digit number of classes.
- An MoE ensemble for over six-digit number of classes.

The code is available on GitHub: `https://github.com/ma921/XRDidentifier`

## 2 Dataset

CIFs (Crystallography Information File) [18] of all 136,899 materials in the ICSD were used for the dataset, which was converted to XRD by pymatgen [19]. DWFs (Debye-Waller Factor) [20] of each element were randomly selected. In this way, a total of five spectra with different DWFs were generated per CIF. The spectra were generated in the conditions where the two theta ranged from 0 to 120 degrees with 0.02-degree increments, resulting in 6,000 length $\times$ 1 channel tensors. Only the chemical composition is identified, not the full atomic structure. Different polymorphs of a compound would be identified as a single class.

Different datasets were prepared for training the expert models and the MoE model. The 73 expert models were created to be tailored to 73 general chemical elements, excluding radioactive elements and noble gases. Hydrogen and oxygen, which have more than 20,000 compounds, are excluded from the expert models because they become inaccurate due to the extensive number of identification classes. (Nevertheless, this is negligible impact on identification accuracy because these hydrides and oxides include the other elements in almost all cases, except for H2O or their simple gas H2, O2.) The lithium compounds database was used for investigating expert models. The number of lithium compounds is 6,800 and the number of XRD spectra is 34,000 (6,800 $\times$ 5). The dataset was split in the ratio of train/validation/test = 70/15/15, holding all class information in every dataset but divided by the DWF variations. The train dataset alone was subjected to random physics-informed data augmentation [7]; thus, the model was trained using spectra with different transformations for each epoch. The models were compared by the metric of top-1 accuracy [21] of the test dataset. The training of the MoE model was performed on 684,495 XRD spectra from all 136,899 ICSD materials, with 5 randomized DWFs. The dataset was split in the ratio of train/validation/test = 70/15/15, with no physics-informed data augmentation on all datasets. The combination of samples in batches was shuffled each time for the train dataset but not for the validation and test datasets.

## 3 Expert Models using Hierarchical Deep Metric Learning

The proposed expert models comprise of 1D-RegNet [16] as encoder, AdaCos [14] as the last layer, and Angular Penalty Softmax Loss (CosFace) [12] as a loss function, as shown in Figure 1(a). In addition, Adam [22] was adopted as an optimizer, iterating for 100 epochs and applying early-stopping when it reached a plateau. Training was done with 2 GPUs of NVIDIA A100 for NVLink 40GiB HBM2, taking approximately seven days to converge. Figure 1(b) shows the top-1 accuracies of the 73 expert models. The identification accuracy is strongly influenced by the number of compounds, approaching zero when it exceeds 20,000. By observing the accuracy of each model in Figure 1(c), it is clear that elements having more than 10,000 compounds are extremely challenging to identify, particularly for mundane elements, such as carbon, iron, and nitrogen. In contrast, the accuracy of the chlorine model is only 0.687, even though the number of compounds is merely 8,767.

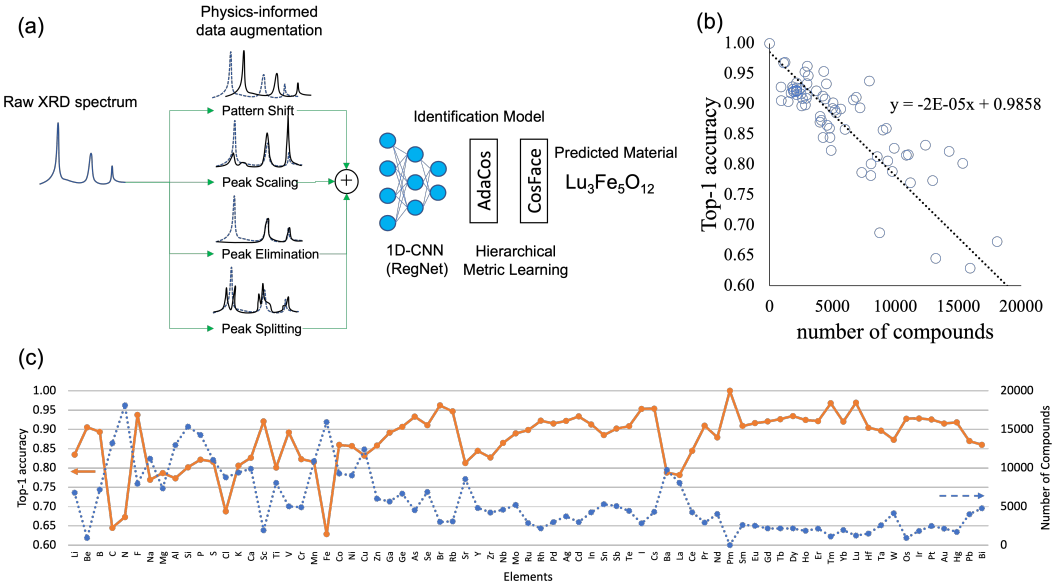


Figure 1: (a) Schematic illustration of hierarchical deep metric learning model, (b) the top-1 accuracies over the number of compounds for 73 expert models, (c) the top-1 accuracies for each element expert.

Table 1: Ablation Study

| # | Method | Top-1 accuracy |
|---|---|---|
| 1 | Ours | **0.834** |
| 2 | Ours replaced with vanilla 1D-CNN from 1D-RegNet | 0.075 |
| 3 | Ours replaced with single AdaCos from hierarchical metric learning | 0.398 |
| 4 | Ours replaced with single CosFace from hierarchical metric learning | 0.329 |
| 5 | Ours replaced with Cross Entropy Loss from hierarchical metric learning | 0.682 |
| 6 | 1D-CNN | 0.021 |

The reason for this could be that the spectra of chlorine compounds have similar shapes, making classification difficult. Thus, it is suggested that the number of compounds and the similarity of the spectra significantly influence the identification accuracy: the average accuracy of the 73 expert models is 0.872.

The ablation study shown in Table 1 compares the network architectures of the expert model on the lithium compounds dataset. The accuracy of the models was evaluated by the top-1 accuracy of the test dataset. The accuracy of a previously reported method with vanilla 1D-CNN (#6) [7] that was employed with limited datasets has been significantly reduced to 0.0206 in this study, indicating that the method is not applicable for large-scale datasets (164 classes in the previous study [7] and 6,800 classes in this study). When 1D-RegNet instead is applied as an encoder architecture (#5), the accuracy is improved to 0.6822, which shows that the CNN network architecture is influential. Interestingly, for the variants of the conventional single deep metric learning (AdaCos #3 and CosFace #4), the accuracies are lower than when neither is employed. This is likely due to the fact that some of the 6,800 compounds show indistinguishably similar spectra, while others show distinct patterns. The different levels of distinguishability are mixed in the same dataset, making it challenging to identify them with a single metric to differentiate. Therefore, by applying two layers of deep metric learning (i.e. both AdaCos and CosFace #1), the model becomes capable of both discriminating between the dissimilar spectra in the first layer, and then also distinguishing the differences between similar spectra in the second layer. Despite this, employing this two-stage hierarchical deep metric learning without 1D-RegNet results in a very low accuracy (#2). In this way, it can be clearly observed that the best model is the variant with both 1D-RegNet and a two-stage hierarchical deep metric learning, which can achieve a top-1 accuracy of 0.8343 (#1).

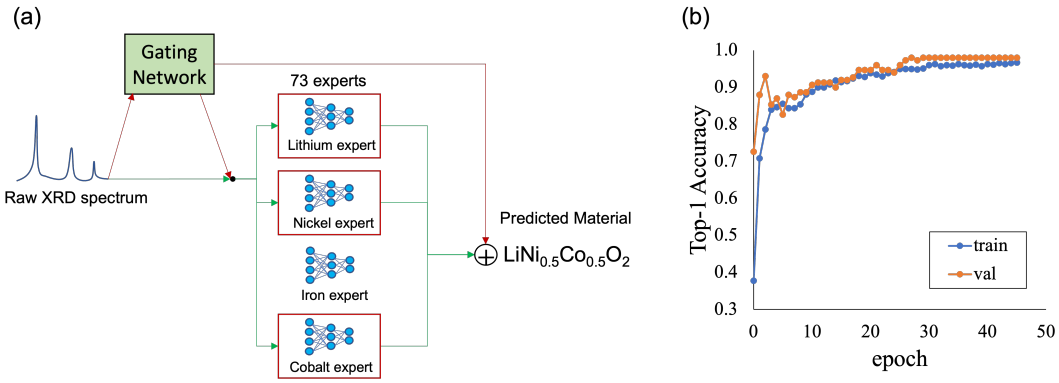


Figure 2: (a) Schematic illustration of the MoE ensembles, (b) the learning curve.

Table 2: Benchmark results

| # | Method | Top-1 accuracy |
|---|---|---|
| 1 | MoE | **0.980** |
| 2 | 1D-RegNet + hierarchical deep metric learning | 0.002 |
| 3 | Inference from the expert models with highest accuracy and confidence | 0.892 |
| 4 | Correlation search and cosine similarity | 0.686 |

## 4 Full Material Identification Model with the MoE Ensemble

In the MoE model, 73 expert models were combined in the sparsely-gated network [17]. The training took about seven days to converge. Therefore, the whole model requires two weeks to re-train in accordance with database updates. Nonetheless, the update frequency of ICSD is in the order of a year or more, meaning that not a big problem for operation in service. The top-1 accuracy of the MoE ensemble model reached 98% at the 50th epoch. Moreover, inference was made quickly, within 1.586 ms per spectrum. Thus, this full material identification model achieved high speed and high accuracy simultaneously.

Table 2 illustrates that the benchmark results with the existing method evaluated with all 136,899 classes. As is clearly shown in #2, the above hierarchical deep metric learning model itself failed to identify a six-digit number of classes. By contrast, the #3 model marked the top-1 accuracy of 0.892. In this model, the class was identified by the following equation:

$$\text{class\_index} \;=\; \text{argmax}\left(\frac{\theta_{acc} \times v}{\text{std}\,(v)}\right) \tag{1}$$

where $v$ is the output from the model before activation, $\theta_{acc}$ is the top-1 accuracy in each of expert datasets. This calculation leads to the natural selection of the inference from the single model with the highest accuracy and confidence. However, this model cannot adaptively combine the multiple inference results from the expert models, thus the accuracy is lower than an MoE ensemble. Furthermore, the model with a classical method [23] in #4 was also examined, which was much less accurate than an MoE. This result is considered to be derived from the classical model [23] that cannot deal with noisy data, especially pattern shift. To make matters worse, it took over 30 minutes per spectrum to infer, showing the explicit advantage of adopting an MoE and deep learning model.

In this paper, the present method considers only pure samples with possibly imperfect spectra. Samples in the wild will always be mixtures of many different compounds, so that the extension of this method to the multi-class multi-label task should be discussed in the future.

## 5 Broader Impact

In previous studies, noisy XRD spectral analysis could not be performed without human intervention because the identification was only possible for around 100 compounds. However, the present method achieves 98% top-1 accuracy and fast identification in the ICSD database of 136,899 materials. The method is expected to be applied to a wide range of fields such as forensic science, materials science, planetary exploration, and deep-sea exploration, which have not been sufficiently explored due to the lack of fast and accurate spectroscopic identification methods. Furthermore, this method can be applied not only to XRD but also to various other spectroscopic analyses such as FT-IR, Raman, XPS, NMR, mass spectrometry, and other spectroscopic techniques, which should attract the interest of many scientists.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 3, 4
   (b) Did you describe the limitations of your work? [Yes] See Section 3, 4
   (c) Did you discuss any potential negative societal impacts of your work? [No] Might need to state this could cause losing spectroscopy technicians' jobs
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See GitHub link
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 2
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Because it takes weeks to run once
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Reference
   (b) Did you mention the license of the assets? [Yes] See GitHub link
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See GitHub link
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] I've followed the instructions on how to use database (ICSD) and GitHub codes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] There is no personal information: all data is scientific experimental data.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

## References


[1] C. K. Muro, K. C. Doty, J. Bueno, L. Halámková, and I. K. Lednev. Vibrational spectroscopy: Recent developments to revolutionize forensic science. *Analytical Chemistry*, 87(306), 2015. URL `https://pubs.acs.org/doi/full/10.1021/ac504068a`.

[2] G. U. Bublitz and S. G. Boxer. Stark spectroscopy: Applications in chemistry, biology, and materials science. *Annual Review of Physical Chemistry*, 48(213), 1997. URL `https://www.annualreviews.org/doi/abs/10.1146/annurev.physchem.48.1.213`.

[3] P. J. Sarre. The diffuse interstellar bands: A major problem in astronomical spectroscopy. *Journal of Molecular Spectroscopy*, 238(1), 2006. URL `https://www.sciencedirect.com/science/article/abs/pii/S0022285206000828`.

[4] M. E. Schaepman, S. L. Ustin, A. J. Plaza, T. H. Painter, J. Verrelst, and S. Liang. Earth system science related imaging spectroscopy—an assessment. *Remote Sensing of Environment*, 113(S123), 2009. URL https://doi.org/10.1016/j.rse.2009.03.001.

[5] M. S. Smyth and J. H. Martin. x ray crystallography. *Molecular pathology*, 53(8), 2000. URL https://doi.org/10.1136/mp.53.1.8.

[6] A. Aarva, V. L. Deringer, S. Sainio, T. Laurila, and M. A. Caro. Understanding x-ray spectroscopy of carbonaceous materials by combining experiments. *Chemistry of Materials*, 31(9243), 2019. URL https://doi.org/10.1021/acs.chemmater.9b02049.

[7] F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, S. Ramasamy, B. L. DeCost, S. I. P. Tian, G. Romano, A. G. Kusne, and T. Buonassisi. Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks. *npj Computational Materials*, 5(60), 2019. URL https://doi.org/10.1038/s41524-019-0196-x.

[8] J. W. Lee, W. B. Park, J. H. Lee, S. P. Singh, and K. S. Sohn. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic xrd powder patterns. *Nature Communications*, 11(86), 2020. URL https://doi.org/10.1038/s41467-019-13749-3.

[9] M. H. Mozaffari and L. L. Tay. A review of 1d convolutional neural networks toward unknown substance identification in portable raman spectrometer. In *arXiv*, 2006. URL https://arxiv.org/abs/2006.10575.

[10] Z. Shen and R. A. V. Rossel. Automated spectroscopic modelling with optimised convolutional neural networks. *Scientific Reports*, 11(208), 2021. URL https://doi.org/10.1038/s41598-020-80486-9.

[11] National Institute of Standards and Technology. Nist inorganic crystal structure database. URL https://doi.org/10.18434/M32147.

[12] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, 2018. URL https://arxiv.org/abs/1801.09414.

[13] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. URL https://arxiv.org/abs/1801.07698.

[14] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *CVPR*, 2019. URL https://arxiv.org/abs/1905.00292.

[15] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár. Designing network design spaces. In *CVPR*, 2020. URL https://arxiv.org/abs/1905.00292.

[16] S. Hong, Y. Xu, A. Khare, S. Priambada, K. Maher, A. Aljiffry, J. Sun, and A. Tumanov. Holmes: Health online model ensemble serving for deep learning models in intensive care units. In *KDD*, 2020. URL https://arxiv.org/abs/2008.04063.

[17] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017. URL https://arxiv.org/abs/1701.06538.

[18] I. D. Brown and B. McMahon. The crystallographic information file (cif). *Data Science Journal*, 5(174), 2006. URL http://doi.org/10.2481/dsj.5.174.

[19] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. Chevrier, K. A. Persson, and G. Ceder. Python materials genomics (pymatgen) : A robust, open-source python library for materials analysis. *Computational Materials Science*, 68(314), 2013. URL https://doi.org/10.1016/j.commatsci.2012.10.028.

[20] V. F. Sears and S. A. Shelley. Debye-waller factor for elemental crystals. *Acta Crystallographica*, A47 (441), 1991. URL https://scripts.iucr.org/cgi-bin/paper?S0108767391002970.

[21] M. Lapin, M. Hein, and B. Schiele. Top-k multiclass svm. In *NeurIPS*, 2015. URL https://arxiv.org/abs/1511.06683.

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. URL https://arxiv.org/abs/1412.6980.

[23] C. Carey, T. Boucher, S. Mahadevan, P. Bartholomew, and M. D. Dyar. Machine learning tools formineral recognition and classification from raman spectroscopy. *Journal of Raman Spectroscopy*, 46 (894), 2015. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/jrs.4757.