# Automatic solid form classification in pharmaceutical drug development

**Julius Lange**[1] **Leonid Komissarov**[1] **Rene Lang**[1]

**Dennis Dimo Enkelmann**[2] **Andrea Anelli**[1]

[1] Roche Pharmaceutical Research and Early Development
[2] Pharma Technical Development
F. Hoffmann-La Roche Ltd.
Basel, Switzerland
andrea.anelli@roche.com

## Abstract

In materials and pharmaceutical development, rapidly and accurately determining the similarity between X-ray powder diffraction (XRPD) measurements is crucial for efficient solid form screening and analysis. We present SMolNet, a classifier based on a Siamese network architecture, designed to automate the comparison of XRPD patterns. Our results show that training SMolNet on loss functions from the self-supervised learning domain yields a substantial boost in performance with respect to class separability and precision, specifically when classifying phases of previously unseen compounds. The application of SMolNet demonstrates significant improvements in screening efficiency across multiple active pharmaceutical ingredients, providing a powerful tool for scientists to discover and categorize measurements with reliable accuracy.

## 1 Introduction

Advancements in artificial intelligence have revolutionized various fields by automating complex, experience-driven processes. In pharmaceutical solid-state development, similar to materials science, there is a growing need to automate the analysis of X-ray powder diffraction (XRPD) patterns [1], which are critical for identifying and characterizing the solid form landscape of active pharmaceutical ingredients (APIs) [2–7]. Here, multiple patterns from crystallization experiments of the same API are compared and assigned to a set of reference diffractograms, each representing a distinct solid form of the compound of interest, as depicted in Figure 1. Traditional methods for XRPD pattern comparison often require expert knowledge and manual inspection, making the process time-consuming and susceptible to human error [8, 9]. Moreover, manual interpretation becomes impractical when dealing with hundreds or thousands of samples generated from high throughput crystallization experiments aimed at discovering new solid forms [10, 11].

Previous work to automate the classification of XRPD patterns includes measures operating directly on the signals [12, 13], CNN classifiers [14, 15] and Siamese networks [16]. A key limitation in many of these works consists in framing this exercise into a classification task where the reference forms are known *a priori* [17, 18]. In practice however, a new unseen form can be observed at every measurement during experimental screening. Further, most of the research in this field focuses on highly crystalline inorganic materials that typically exhibit simpler and sharper diffraction

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

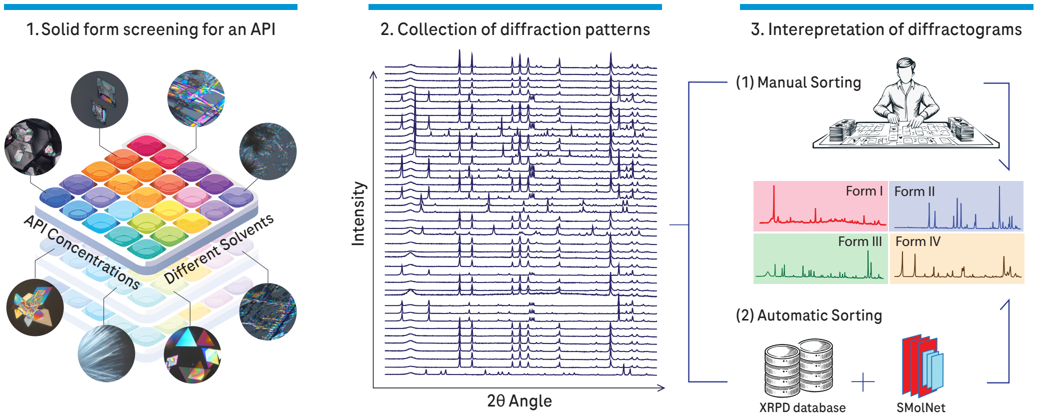

Figure 1: Example of a solid form screening project: (1) A candidate molecule is mixed in different solutions to explore its solid state landscape. Measurable crystalline residues that are formed are investigated using XRPD - creating a large number of measurements of potentially different materials (2). To identify the different forms obtained, typically a manual sorting is employed (3-1). Alternatively, we propose to leverage previous projects' data to train SMolNet, which can be used to perform the same sorting task at higher throughput (3-2).

peaks due to high symmetry and rigid crystal structures. Here the generation of synthetic XRPD patterns is a common technique for enrichment of training data [14, 19–21]. Small molecule APIs however, typically produce more complex XRPD patterns with broader peaks and additional artifacts, stemming from their intricate molecular structures, lower symmetry, and variations such as amorphous backgrounds, height preparation shifts, preferential orientation, and particle size variance. These complexities are especially evident in real-life data obtained from solid form screening routines of organic APIs. Such variations pose significant challenges for automated pattern recognition and classification, as methods developed for inorganic materials may not transfer directly to the organic domain. [22, 23].

To address these challenges, we introduce SMolNet (**S**olid-form **Mol**ecules **Net**work), a novel deep learning framework specifically designed for the pairwise classification and identification of XRPD patterns of organic crystalline materials. Our approach leverages Siamese networks to capture subtle similarities and differences between complex diffraction patterns, even with limited training data. By training directly on experimental XRPD data from organic crystals, our model bypasses the need of synthetic data augmentation and demonstrates reliable generalization to new, unseen patterns and chemical spaces. Our contributions are as follows:

- We propose a Siamese network architecture, trained on a modified SigLIP loss [24] for the first time to handle organic XRPD patterns in a zero shot learning setting.
- Robust training and testing on experimental data using a leave-two-compounds-out routine (L2CO)
- Improved classification performance: We demonstrate significant improvements over non data-driven methods and previous architectures in discriminating across different forms.

By bridging the gap between machine learning and materials science, our work advances the application of AI in pharmaceutical solid-state development. This not only accelerates the solid-form screening process but also paves the way for more efficient and accurate materials discovery.

## 2 Methods

### 2.1 Data

The dataset constructed for this study is proprietary and contains $3750$ experimental XRPD patterns measured across $16$ organic compounds that constitute drug candidates in a pharmaceutical research and development setting, for a total of $24$ different solid forms. Each pattern constitutes a $1$-d signal of $1950$ points, covering a scattering angle of $2\theta \in [3, 42)$. Additional experimental details are provided in Appendix D. For each compound, patterns that belong to the same phase were manually labelled by experimentalists. All pattern intensities were normalized to be within the $[0, 1]$ range. We consider pairwise combinations of individual patterns, amounting to roughly $7 \times 10^6$ pairs and a respective label specifying whether the two patterns belong to the same phase (positive class) or not (negative class). The ratio of positive to negative classes is $1 : 9.5$, creating a substantially imbalanced dataset. We remark how this large curated database constitutes, to the best of our knowledge, a first example on the feasibility of training a foundational model for organic solid diffraction patterns recognition.

### 2.2 Model architecture

We aim to classify whether two diffraction patterns $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ belong to the same phase or not. To that end we consider $\boldsymbol{z} = f(\boldsymbol{x}, \boldsymbol{\theta})$ – a function that translates individually normalized input intensities of a pattern $\boldsymbol{x} \in \mathbb{R}^{1950}$ to a latent embedding $\boldsymbol{z} \in \mathbb{R}^n$, where $f$ can additionally be parameterized by $\boldsymbol{\theta}$. For two patterns representing the phase we would like the distance of their latent embeddings $d(\boldsymbol{z}_1, \boldsymbol{z}_2)$ to be small and vice versa. We find the best architecture for this task through hyperparameter optimization. Our model constitutes four 1-dimensional convolutional layers with a kernel size of 8 to 128, each followed by a batch normalization, Mish activation [25] and dropout with a probability of $0.2$. All convolutional outputs are concatenated and passed on to a multilayer perceptron, consisting of two dense layers with 2048 hidden neurons. The final output embedding is $\boldsymbol{z} \in \mathbb{R}^{128}$. When trained on the Contrastive loss (*cf.* following section), the last layer embeddings are additionally passed through a Sigmoid activation function. A detailed diagram of the architecture is provided in Apppendix A.

### 2.3 Training

To simulate a prospective application setting as closely as possible while guaranteeing statistical significance of the results, a leave-two-compounds-out (L2CO) cross-validation technique is employed. At each fold a unique combination of two out of the total of 16 compounds are selected and all measurements belonging to one of the two compounds are used for testing, resulting in a total of 120 folds (*cf.* Sec. 2.1). This ensures presence of at least two forms in the test set to have both positive and negative pairs. Of the remaining 14 compounds, 11 are used for training and three, randomly chosen, for validation. The model weights are optimized with the Adam [26] optimizer at an initial learning rate of $5 \times 10^{-3}$. Best model parameters are saved every time a new minimal loss is observed on the validation set. The learning rate is reduced by a factor of $0.1$ if no improvement in the validation loss is observed after 15 evaluations. The model is trained for up to 20 epochs.

We train and compare the SMolNet performance after training on one of three different loss functions, each previously applied to contrastive learning tasks. The functions tested are the Contrastive [27, 28], NT-XEnt/InfoNCE [29–31] and SigLIP [24] losses. Considering our application we propose two modifications to the SigLIP loss. First, rather than using a cosine similarity for training the embeddings, the Euclidean distance is used to generate the logit values. This is motivated by the improved classification performance, as reported in Appendix C. Second, since we are not dealing with a self-supervised task, we relax the criterion that all matching pairs are only found along the diagonal of the pairwise matrix of labels. We present a pseudo-implementation of the modified SigLIP algorithm in Appendix B. When training with the Contrastive loss, weighted sampling is employed during the training to account for class imbalance. Training times with the Contrastive loss are approximately 15-25 mins for one fold on a NVIDIA A100 / LS40S GPU. Significant speedup by a factor of 5-10 is achieved when training with one of the other loss functions. This is due to the explicit generation of input pairs when training with the Contrastive loss, effectively resulting in $\sim N^2$ data entries.

To provide a lower bound for predictivity, we compare the performance of SMolNet to a naive baseline model where the score is directly computed from the euclidean distance such that: $d(\boldsymbol{z}_1, \boldsymbol{z}_2) = ||\boldsymbol{x}_1 - \boldsymbol{x}_2||$. We follow the same cross-validation procedure as above, computing all pairwise euclidean distances on the raw 1-d input signals in the training data, interpreting the results as classifier probabilities. We then choose a threshold that maximizes the F1 score and apply it to the test set patterns in order to derive predicted binary labels. Additionally, we compare the performance of SMolNet to a previously published CNN architecture [15, 16] and a similarity measure by de Gelder *et al.* [12]. Both methods were developed specifically for the classification of XRPD patterns.

## 3 Results

Average results of the L2CO cross-validation are presented in Table 1, showing the area under the ROC curve (AUROC), area under the precision-recall curve (AUPRC), accuracy and F1 score. Note that the AUPRC value associated with a random model is variable, equaling to the ratio of positive labels in the test set, $\text{AUPRC}_{\text{random}} = {}^{N_{\text{positive}}}\!/_{N_{\text{total}}}$. We thus report this metric as the delta above the baseline value: $\text{AUPRC} = \text{AUC}(R, P) - \text{AUPRC}_{\text{random}}$.

Our findings reveal that the naive and de Gelder models already perform well in separating positive from negative classes with an above-average precision. Hard classifier metrics, *i.e.* Accuracy and F1 score are best for these models, hinting that threshold finding might be easiest when using simple approaches that operate directly on the input patterns. Nevertheless, machine learning approaches clearly improve on both AUROC and AUPRC metrics. This is particularly important when evaluating cases as ours, where a high class-imbalance (*cf.* Section 2.1) can produce overly-optimistic accuracy and F1 scores. We therefore argue that evaluating models on the soft classifier metrics is more meaningful if bias to a particular class should be avoided. We observe marginal improvements in all metrics when comparing SMolNet to a previously published architecture [15, 16] when using a Contrastive loss. Notably, training on the NT-XEnt and SigLIP loss functions appears to be highly advantageous when high average precision and class separability are desired. Our best-performing model is SMolNet trained on the SigLIP loss, resulting in average AUROC and AUPRC values of 0.98 and 0.65, respectively. Conversely, applying the SigLIP loss to a previously published architecture [15, 16] leads to further degradation of performance, as reported in Appendix C. Although the de Gelder method produces the best F1 score, we argue that the increased AU{ROC,PRC} scores of the machine learning models is particularly significant in typical human-in-the-loop scenarios where partial labeling is performed preliminarily on a subset of data points. This allows for further threshold optimization on test data, making SMolNet especially valuable in real-world applications where efficiency and accuracy are paramount. For additional experiments and relevant loss funciton parameters please refer to Appendix C.

Table 1: Average L2CO performance and their 95% confidence interval for various tested approaches. Showing the area under the ROC curve (AUROC), area under the precision-recall curve (AUPRC), Accuracy and F1 score. Best and second-best results in bold and underlined, respectively.

| Method | Loss function | AUROC | AUPRC | Acc | F1 |
|---|---|---|---|---|---|
| Naive | | $0.88_{\pm.14}$ | $0.34_{\pm.26}$ | $\underline{0.89}_{\pm.14}$ | $\underline{0.90}_{\pm.13}$ |
| De Gelder *et al.*[12] | | $0.92_{\pm.09}$ | $0.46_{\pm.21}$ | $\mathbf{0.92}_{\pm.10}$ | $\mathbf{0.93}_{\pm.11}$ |
| Schützke *et al.*[15,16] | Contrastive | $0.95_{\pm.14}$ | $0.41_{\pm.25}$ | $0.86_{\pm.20}$ | $0.87_{\pm.18}$ |
| SMolNet | Contrastive | $0.97_{\pm.09}$ | $0.43_{\pm.24}$ | $0.87_{\pm.20}$ | $0.89_{\pm.17}$ |
| SMolNet | NT-XEnt | $\underline{0.97}_{\pm.08}$ | $\underline{0.63}_{\pm.33}$ | $0.89_{\pm.20}$ | $0.79_{\pm.39}$ |
| SMolNet | SigLIP | $\mathbf{0.98}_{\pm.09}$ | $\mathbf{0.65}_{\pm.31}$ | $0.89_{\pm.21}$ | $0.80_{\pm.38}$ |

## 4 Conclusion & Outlook

We have provided a model architecture and training setup with extensive proof of its performances across different pharmaceutical materials. The ability to reliably distinguish between similar and dissimilar patterns across different organic compounds and compositions constitutes a fundamental first step in the direction of tackling this problem with a foundational model approach. Our findings further report that the incorporation of contrastive loss functions into the model training procedure is highly beneficial. While our findings show that SMolNet provides superior class separation based on soft metrics, we recognize that optimal threshold determination for classification remains a challenge. Threshold selection is strongly influenced by training procedures and is highly dependent on the test data, particularly in zero-shot learning scenarios where test compounds are not present in the training data. Addressing this issue requires further investigation into adaptive thresholding methods and calibration techniques to improve generalization to unseen compounds. Having established a reliable architecture in complex pure phase projects, we plan in future to explore the integration of measurements from public databases, as well as synthetic data coming from crystal structure prediction simulations. Finally, by combining the multiple single phase data we aim to tackle the challenging task of mixture determination.

## Acknowledgments and Disclosure of Funding

L. K. acknowledges funding by the Roche Postdoctoral Fellowship (RPF) programme. The authors thank Leoni Grossman, Andre Egli and Joost van den Ende for inspiring discussions.

## References

[1] J. Leeman *et al.*, “Challenges in high-throughput inorganic materials prediction and autonomous synthesis,” *PRX Energy*, vol. 3, no. 1, p. 011 002, 2024.

[2] S. Bhattacharya, H. Brittain, and R. Suryanarayanan, “Thermoanalytical and crystallographic methods,” English (US), in *Polymorphism in Pharmaceutical Solids*. CRC Press, Apr. 2016, pp. 318–346.

[3] S. Byrn, R. Pfeiffer, and J. Stowell, *Solid-state Chemistry of Drugs*. SSCI, Incorporated, 1999.

[4] R. Hilfiker, S. M. De Paul, and T. Rager, “Analytical tools to characterize solid forms,” in *Polymorphism in the Pharmaceutical Industry*. John Wiley & Sons, Ltd, 2018, ch. 14, pp. 415–446.

[5] J. Bernstein, *Polymorphism in Molecular Crystals* (IUCr monographs on crystallography). Clarendon Press, 2002.

[6] D. E. Braun, T. Gelbrich, V. Kahlenberg, R. Tessadri, J. Wieser, and U. J. Griesser, “Conformational polymorphism in aripiprazole: Preparation, stability and structure of five modifications,” *Journal of Pharmaceutical Sciences*, vol. 98, no. 6, pp. 2010–2026, 2009.

[7] D. E. Braun *et al.*, “Unraveling complexity in the solid form screening of a pharmaceutical salt: Why so many forms? why so few?” *Crystal Growth & Design*, vol. 17, no. 10, pp. 5349–5365, 2017.

[8] H. P. Klug and L. E. Alexander, *X-ray diffraction procedures: for polycrystalline and amorphous materials*. John Wiley & Sons, 1954.

[9] A. Guinier, *X-ray diffraction in crystals, imperfect crystals, and amorphous bodies*. Dover Publications, 1994.

[10] A. J. Florence, “Approaches to high-throughput physical form screening and discovery,” English (US), in *Polymorphism in Pharmaceutical Solids*. CRC Press, 2009, pp. 139–184.

[11] S. M. Reutzel-Edens and R. M. Bhardwaj, “Crystal forms in pharmaceutical applications: olanzapine, a gift to crystal chemistry that keeps on giving,” *IUCrJ*, vol. 7, no. 6, pp. 955–964, 2020.

[12] R. de Gelder, R. Wehrens, and J. A. Hageman, “A generalized expression for the similarity of spectra: Application to powder diffraction pattern classification,” *Journal of Computational Chemistry*, vol. 22, no. 3, pp. 273–289, 2001.

[13] H. Karfunkel, B. Rohde, F. Leusen, R. J. Gdanitz, and G. Rihs, “Continuous similarity measure between nonoverlapping x-ray powder diagrams of different crystal modifications,” *Journal of computational chemistry*, vol. 14, no. 10, pp. 1125–1135, 1993.

[14] 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*, vol. 11, no. 1, p. 86, 2020.

[15] H. Wang *et al.*, “Rapid identification of x-ray diffraction patterns based on very limited data by interpretable convolutional neural networks,” *Journal of Chemical Information and Modeling*, vol. 60, no. 4, pp. 2004–2011, Mar. 2020, ISSN: 1549-960X. DOI: `10.1021/acs.jcim.0c00020`.

[16] J. Schuetzke, A. Benedix, R. Mikut, and M. Reischl, “Siamese networks for 1d signal identification,” in *Proceedings-30. Workshop Computational Intelligence: Berlin*, vol. 26, 2020, p. 27.

[17] J. Schuetzke *et al.*, “Accelerating materials discovery: Automated identification of prospects from x-ray diffraction data in fast screening experiments,” *Advanced Intelligent Systems*, vol. 6, no. 3, p. 2 300 501, 2024. DOI: `https://doi.org/10.1002/aisy.202300501`.

[18] N. J. Szymanski, S. Fu, E. Persson, and G. Ceder, “Integrated analysis of x-ray diffraction patterns and pair distribution functions for machine-learned phase identification,” *npj Computational Materials*, vol. 10, no. 1, p. 45, 2024.

[19] J.-W. Lee, W. B. Park, M. Kim, S. P. Singh, M. Pyo, and K.-S. Sohn, “A data-driven xrd analysis protocol for phase identification and phase-fraction prediction of multiphase inorganic compounds,” *Inorganic Chemistry Frontiers*, vol. 8, no. 10, pp. 2492–2504, 2021.

[20] P. M. Maffettone *et al.*, “Crystallography companion agent for high-throughput materials discovery,” *Nature Computational Science*, vol. 1, no. 4, pp. 290–297, Apr. 2021, ISSN: 2662-8457. DOI: `10.1038/s43588-021-00059-2`. [Online]. Available: `http://dx.doi.org/10.1038/s43588-021-00059-2`.

[21] J. Schuetzke, N. J. Szymanski, and M. Reischl, “A universal synthetic dataset for machine learning on spectroscopic data,” *arXiv preprint arXiv:2206.06031*, 2022.

[22] B. Warren, “Diffraction by imperfect crystals,” in *X-ray Diffraction*. Dover Publications, 1990, 251 ff.

[23] M. Ermrich and D. Opper, “Xrd data collection,” in *XRD for the Analyst: Getting Acquainted with the Principles*. PANalytical, 2013, pp. 44–52.

[24] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, *Sigmoid loss for language image pre-training*, 2023. arXiv: `2303.15343 [cs.CV]`. [Online]. Available: `https://arxiv.org/abs/2303.15343`.

[25] D. Misra, “Mish: A self regularized non-monotonic neural activation function,” *CoRR*, vol. abs/1908.08681, 2019. arXiv: `1908.08681`. [Online]. Available: `http://arxiv.org/abs/1908.08681`.

[26] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. arXiv: `1412.6980 [cs.LG]`. [Online]. Available: `https://arxiv.org/abs/1412.6980`.

[27] S. Chopra, R. Hadsell, and Y. LeCun, “Learning a similarity metric discriminatively, with application to face verification,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, 2005, 539–546 vol. 1. DOI: `10.1109/CVPR.2005.202`.

[28] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR’06)*, IEEE, vol. 2, 2006, pp. 1735–1742.

[29] K. Sohn, “Improved deep metric learning with multi-class n-pair loss objective,” in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016. [Online]. Available: `https://proceedings.neurips.cc/paper_files/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf`.

[30] A. van den Oord, Y. Li, and O. Vinyals, *Representation learning with contrastive predictive coding*, 2019. arXiv: `1807.03748 [cs.LG]`. [Online]. Available: `https://arxiv.org/abs/1807.03748`.

[31] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, *A simple framework for contrastive learning of visual representations*, 2020. arXiv: `2002.05709 [cs.LG]`. [Online]. Available: `https://arxiv.org/abs/2002.05709`.

# A SMolNet Architecture

Architecture of SMolNet described in Sec. 2.

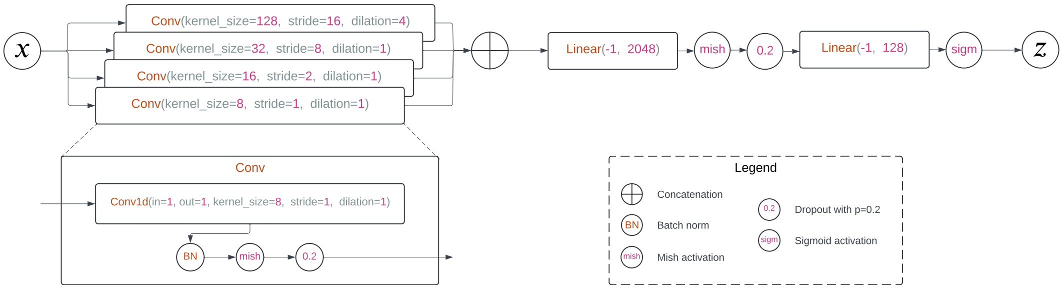
$x$
Conv(kernel_size=128, stride=16, dilation=4)
Conv(kernel_size=32, stride=8, dilation=1)
Conv(kernel_size=16, stride=2, dilation=1)
Conv(kernel_size=8, stride=1, dilation=1)
Linear(-1, 2048)
mish
0.2
Linear(-1, 128)
sigm
$z$
Conv
Conv1d(in=1, out=1, kernel_size=8, stride=1, dilation=1)
BN
mish
0.2
Legend
Concatenation
Batch norm
BN
Mish activation
mish
0.2
Dropout with p=0.2
sigm
Sigmoid activation

Figure 2: SMolNet architecture diagram with hyperparameters used in this study.

# B Modified SigLIP Pseudo-Implementation

```
# n : batch size
# emb1 : model embeddings to be compared with emb2 [n, dim]
# emb2 : model embeddings to be compared with emb1 [n, dim]
# labels1 : labels of emb1 [n,1]
# labels2 : labels of emb2 [n,1]
# t_prime, b : learnable temperature and bias

t = exp(t_prime)
z1 = l2_normalize(emb1)
z2 = l2_normalize(emb2)
logits = l2_distance(z1, z2) * t + b # [n,n]

labels = int(labels1 == labels2.T) # [n,n], positive pairs: 1
labels[labels==0] = -1 # negative pairs: -1

loss = -sum(log_sigmoid(labels * logits)) / n
```

# C Additional Experimental Results and Parameters

Table 2: Average L2CO performance and their 95% confidence interval for various tested approaches, and parameters used. Showing the area under the ROC curve (AUROC), area under the precision-recall curve (AUPRC), Accuracy and F1 score.

| Method | Loss function | Parameters | AUROC | AUPRC | Acc | F1 |
|---|---|---|---|---|---|---|
| Naive | | | $0.88_{\pm.14}$ | $0.34_{\pm.26}$ | $0.89_{\pm.14}$ | $0.90_{\pm.13}$ |
| De Gelder *et al.*[12] | | $l = 0.005$ | $0.92_{\pm.09}$ | $0.46_{\pm.21}$ | $0.92_{\pm.10}$ | $0.93_{\pm.11}$ |
| Schützke *et al.*[15,16] | Contrastive | $m = 8.0$ | $0.95_{\pm.14}$ | $0.41_{\pm.25}$ | $0.86_{\pm.20}$ | $0.87_{\pm.18}$ |
| Schützke *et al.* | SigLIP | L2 distance | $0.69_{\pm.29}$ | $0.19_{\pm.40}$ | $0.39_{\pm.48}$ | $0.41_{\pm.44}$ |
| SMolNet | Contrastive | $m = 8.0$ | $0.97_{\pm.09}$ | $0.43_{\pm.24}$ | $0.87_{\pm.20}$ | $0.89_{\pm.17}$ |
| SMolNet | NT-XEnt | L2 distance, $T = 5.0$ | $0.97_{\pm.08}$ | $0.63_{\pm.33}$ | $0.89_{\pm.20}$ | $0.79_{\pm.39}$ |
| SMolNet | SigLIP | L2 distance | $0.98_{\pm.09}$ | $0.65_{\pm.31}$ | $0.89_{\pm.21}$ | $0.80_{\pm.38}$ |
| SMolNet | SigLIP | Cosine similarity | $0.94_{\pm.20}$ | $0.56_{\pm.43}$ | $0.83_{\pm.36}$ | $0.74_{\pm.45}$ |

# D Experimental Setup for XRPD measurements

The dataset used in this study consists of X-ray diffraction (XRD) measurements collected from 2006 to the present, across various research and development projects, using STOE STADI P diffractometers. These systems were equipped with curved germanium (Ge$(111)$) monochromators and Cu K$\alpha 1$ radiation sources ($\lambda$ = $1.54060$Å). The instruments operated at a voltage of $40$ kV, with currents ranging from $40$ mA to $50$ mA, depending on the specific experimental setup. Samples weighing $1$ to $5$ mg were placed in sample cells with aperture diameters between 3 mm and 5 mm and a sample depth of $0.45$ mm. For wet samples, Kapton film clips were used, while cellulose acetate film clips were employed for dry samples. Data were recorded in transmission mode, spanning an angular range of $3\deg$ to $42\deg$ in $2\theta$, using a moving Position Sensitive Detector (PSD) with a fixed omega angle. Over the years, two detector systems were employed for data collection. Initially, the STOE Linear Position Sensitive Detector (PSD) was used, covering up to $6\deg$ in $2\theta$ per scan, allowing for rapid data acquisition. A step size of $0.5\deg$ in $2\theta$ was applied, with dwell times ranging from 5 to $40$ seconds per step, depending on the measurement mode (standard or rapid). Subsequent measurements were performed with Dectris MYTHEN K1 and MYTHEN K2 strip detectors, which offered a broader angular coverage of $12.5\deg$ in $2\theta$ and a high resolution of $0.01\deg$ in $2\theta$. For both MYTHEN detectors, the step size remained at $0.5\deg$ in $2\theta$, with dwell times ranging from $5$ to $20$ seconds per step, depending on whether standard or rapid measurements were taken.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Compared to the baseline we report significantly improved AUROC and AUPRC scores in Table 1 across measurements of chemically distinct compounds.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations in threshold finding and inability to detect mixtures is addressed in the Conclusion & Outlook Section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Complete description of model architecture and hyper-parameters as well as training procedure is described.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The authors' contractual obligations prevent the release of data and code at this point in time, however the best efforts were undertaken to describe the methodology in full detail.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: *cf.* Methods Section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: 95% CIs reported in Tables.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the model training section we report the average compute resources and architecture used in the training routines.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The data and results presented in this manuscript have no personal information and are not directly related to the NeurIPS code of conduct.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The article does not have direct positive or negative societal impacts. The impact of the presented model is an algorithm which detects similarity between diffraction patterns with higher accuracy than state of the art. It has no potential for dual use.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The work does not pose any of the above risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The authors' employer is the owner of the data and the code. We adhere to our company's standard for publishing and have safeguarded the rights to the data and code.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing, nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.