
Article:
Wilson, Julie Carol orcid.org/0000-0002-5171-8480, Bruno, A, Charbonneau, P et al. (6 more authors) (2018) Classification of crystallization outcomes using deep convolutional neural networks. PLOS one. e0198883. ISSN 1932-6203

https://doi.org/10.1371/journal.pone.0198883

Classification of crystallization outcomes using deep convolutional neural networks

Andrew E. Bruno[1], Patrick Charbonneau[2], Janet Newman[3], Edward H. Snell[4], David R. So[5], Vincent Vanhoucke[5,*], Christopher J. Watkins[6], Shawn Williams[7], Julie Wilson[8]

1 Center for Computational Research, University at Buffalo, Buffalo, New York, United States of America
2 Department of Chemistry and Department of Physics, Duke University, Durham, North Carolina, United States of America
3 Collaborative Crystallisation Centre, CSIRO, Parkville, Victoria, Australia
4 Hauptman-Woodward Medical Research Institute and SUNY Buffalo, Department of Materials, Design, and Innovation, Buffalo, New York, United States of America
5 Google Brain, Google Inc., Mountain View, California, United States of America
6 IM&T Scientific Computing, CSIRO, Clayton South, Victoria, Australia
7 Platform Technology and Sciences, GlaxoSmithKline Inc., Collegeville, Pennsylvania, United States of America
8 Department of Mathematics, University of York, York, United Kingdom

* Corresponding author
E-mail: vanhoucke@google.com (VV)

Abstract

The Machine Recognition of Crystallization Outcomes (MARCO) initiative has assembled roughly half a million annotated images of macromolecular crystallization experiments from various sources and setups. Here, state-of-the-art machine learning algorithms are trained and tested on different parts of this data set. We find that more than 94% of the test images can be correctly labeled, irrespective of their experimental origin. Because crystal recognition is key to high-density screening and the systematic analysis of crystallization experiments, this approach opens the door to both industrial and fundamental research applications.

Author summary

Protein crystal growth experiments are routinely imaged, but the mass of accumulated data is difficult to manage and analyze. Using state-of-the-art machine learning algorithms on a large and diverse set of reference images, we manage to recapitulate the labels of a remarkably large fraction of the set. This automation should enable a number of industrial and fundamental applications.

1 Introduction

X-ray crystallography provides the atomic structure of molecules and molecular complexes. These structures in turn provide insight into the molecular driving forces for small molecule binding, protein-protein interactions, supramolecular assembly and other biomolecular processes. The technique is thus foundational to molecular modeling and

design. Beyond the obvious importance of structure information for understanding and altering the role of biomolecules, it also has important industrial applications. The pharmaceutical industry, for instance, uses structures to guide chemistry as part of a "predict first" strategy [1], employing expert systems to reduce optimization cycle times and more effectively bring medicine to patients. Yet, despite decades of methodological advances, crystallizing molecular targets of interest remains the bottleneck of the entire crystallography program in structural biology.

Even when crystallization is facile, it is microscopically rare; for macromolecules it is also uncommon [2–5]. Experimental trials typically involve: (i) mixing a purified sample with chemical cocktails designed to promote molecular association, (ii) generating a supersaturated solution of the desired molecule via evaporation or equilibration, and (iii) visually monitoring the outcomes, before (iv) optimizing those conditions and analyzing the resultant crystal with an X-ray beam. One hopes for the formation of a crystal instead of non-specific (amorphous) precipitates or of nothing at all. In order to help run these trials, commercial crystallization screens have been developed; each screen generally contains 96 formulations designed to promote crystal growth. Whether these screens are equally effective or not [5, 6] remains debated, but their overall yield is in any case paltry. Typically fewer than 5% of crystallization attempts produce useful results (with a success rate as low as 0.2% in some contexts [7]).

The practical solution to this hurdle has been to increase the convenience and number of crystallization trials. To offset the expense of reagents and scientist time, labs routinely employ industrial robotic liquid handlers, nanoliter-size drops, and record trial outcomes using automated imaging systems [5, 8–11]. Hoping to compensate for the rarity of crystallization, commercially available systems readily probe a large area of chemical space with minimal sample volume with a throughput of ~ 1000 individual experiments per hour.

While liquid handling is readily automated, crystal recognition is not. Imaging systems may have made viewing results more comfortable than bending over a microscope, but crystallographers still manually inspect images and/or drops, looking for crystals or, more commonly, conditions that are likely to produce good crystals when optimized. This human cost makes crystal recognition a key experimental bottleneck within the larger challenge of crystallizing biomolecules [7]. A typical experiment for a given sample includes four 96-well screens at two temperatures, i.e., 768 conditions (and can have up to twice that [12]). Assuming that it takes 2 seconds to manually scan a droplet (and noting that the scans have to be repeated, as crystallization is time dependent), simply looking at a single set of 96 trials over the lifetime of an experiment can take the better part of an hour [1]. For the sake of illustration, the U.S. Structural Science group at GlaxoSmithKline performs ~ 1200 96-well experiments per year. If the targeted observation schedule were rigorously followed, the group would spend a quarter of the year staring at drops, of which the vast majority contains no crystal. Recording outcomes and analyzing the results of the 96 trials would further increase the time burden. Current operations are already straining existing resources, and the approach simply does not scale for proposed higher-density screening [10].

Crystal growth is also sufficiently uncommon that the tolerance for false negatives is almost nil. Yet most crystallographers are misguided in thinking that they themselves would never miss identifying a crystal given an image containing an crystal, or indeed miss a crystal in a droplet viewed directly under a microscope [13]. In fact, not only do crystallographers miss crystals due to lack of attention through boredom, they often disagree on the class an image should be assigned to. An overall agreement rate of

[1] This estimate is based on personnal communication with five experienced crystallographers at GlaxoSmithKline: 2 seconds/observation × 8 observations × 96 wells. Note that current technology can automatically store and image plates at about 3 min/plate.

$\sim 70\%$ was found when the classes assigned to 1200 images by 16 crystallographers were compared [13]. (When considering only crystalline outcomes, agreement rose to $\sim 93\%$.) Consistency in visual scoring was also considered by Snell et al. when compiling a $\sim 150,000$ image dataset [14]. They found that viewers give different scores to the same image on different occasions during the study, with the average agreement rate for scores on a control set at the beginning and middle of the study being 77%, rising to 84% for the agreement in scores between the middle and end of the study. Crystallographers also tend to be optimistically biased when scoring their own experiments [15]. A better use of expert time and attention would be to focus on scientific inquiry.

An algorithm that could analyze images of drops, distinguish crystals from trivial outcomes, and reduce the effort spent cataloging failure, would present clear value both to the discipline and to industry. Ideally, such an algorithm would act like an experienced crystallographer in:

- recognizing macromolecular crystals appropriate for diffraction experiments;
- recognizing outcomes that, while requiring optimization, would lead to crystals for diffraction experiments;
- recognizing non-macromolecular crystals;
- ignoring technical failures;
- identifying non-crystalline outcomes that require follow up;
- being agnostic as to the imaging platform used;
- being indefatigable and unbiased;
- occurring in a time frame that does not impede the process;
- learning from experience.

Such an algorithm would further reduce the variance in the assessments, irrespective of its accuracy. A high-variance, manual process is not conducive to automating the quality control of the system end-to-end, including the imaging equipment. Enhanced reproducibility enables traceability of the outcomes, and paves the way for putting in place measurable, continuous improvement processes across the entire imaging chain.

Automated crystallization image classifications that attempt to meet the above criteria have been previously attempted. The research laboratories that first automated crystallization inspection quickly realized that image analysis would be a huge problem, and concomitantly developed algorithms to interpret them [16–19]. None of these programs was ever widely adopted. This may have been due in part to their dependence on a particular imaging system, and to the relatively limited use of imaging systems at the time. Many of the early image analysis programs further required very time consuming collation of features and significant preprocessing, e.g., drop segmentation to locate the experimental droplet within the image. To the best of our knowledge, there was also no widespread effort to make a widely available image analysis package in the same way that that the diffraction oriented programs have been organized, e.g., the CCP4 package [20].

Can a better algorithm be constructed and trained? In order to help answer this question, the Machine Recognition of Crystallization Outcomes (MARCO) initiative was set up [21]. MARCO assembled a set of roughly half a million classified images of crystallization trials through an international collaboration with five separate institutions. Here, we present a machine-learning based approach to categorize these images. Remarkably, the algorithm we employ manages to obtain an accuracy exceeding

Table 1. Breakdown of data sources and imaging technology per institution contributing to MARCO.

Institution	Technical Setup	# of Images
Bristol-Myers Squibb	Formulatrix Rock Imager (FRI)	8719
CSIRO	Sitting drop, FRI, Rigaku Minstrel [22, 23]	15933
HWMRI	Under oil, Home system [14]	79632
GlaxoSmithKline	Sitting drop, FRI	83126
Merck	Sitting drop, FRI	305804

94%, which is even above what was once thought possible for human categorization. This suggests that a deployment of this technology in a variety of laboratory settings is now conceivable. The rest of this paper is as follows. Section 2 describes the dataset and the scoring scheme, Sec. 3 describes the machine-learning model and training procedure, Secs. 4 and 5 describe and discuss the results, respectively, and Sec. 6 briefly concludes.

2 Material and Methods

Image Data

The MARCO data set used in this study contains 493,214 scored images from five institutions (See Table 1 [21]). The images were collected from imagers made from two different manufacturers (Rigaku Automation and Formulatrix), which have different optical systems, as well as by the in-house imaging equipment built at the Hauptman-Woodward Medical Research Institute (HWMRI) High-Throughput Crystallization Center (HTCC). Different versions of the setups were also used – some Rigaku images are collected with a true color camera, some are collected as greyscale images. The zoom extent varies, with some imagers set up to collect a field-of-view (FOV) of only the experimental droplet, and some set for the FOV to encompass a larger area of the experimental setup. The Rigaku and Formulatrix automation imaged vapor diffusion based experiments while the HTCC system imaged microbatch-under-oil experiments. A random selection of 50,284 test images was held out for validation. Images in the test set were not represented in the training set. The precise data split is available from the MARCO website [21].

Labeling

Images were scored by one or more crystallographers. As the dataset is composed of archival data, no common scoring system was imposed, nor were exemplar images distributed to the various contributors. Instead, existing scores were collapsed into four comprehensive and fairly robust categories: clear, precipitate, crystal, and other. This last category was originally used as a catchall for images not obviously falling into the three major classes, and came to assume a functional significance as the classification process was further investigated. Examination of the least classifiable five percent of images indeed revealed many instances of process failure, such as dispensing errors or illumination problems. These uninterpretable images were then labelled as “other” during the rescoring, which added an element of quality control to the overall process [24].

Relabeling

After a first baseline system was trained (see Sec. 3), the 5% of the images that were most in disagreement with the classifier (independently of whether the image was in the training or the test set), were relabeled by one expert, in order to obtain a systematic eye on the most problematic images.

Because no rules were established and no exemplars were circulated prior to the initial scoring, individual viewpoints varied on classifying certain outcomes. For example, the bottom 5% contained many instances of phase separation, where the protein forms oil droplets or an oily film that coats the bottom of the crystallization well. Images were found to be inconsistently scored as "clear", "precipitate", or "other" depending on the amount and visibility of the oil film. This example highlights the difficulty of scoring experimental outcomes beyond crystal identification. A more serious source of ambiguity arises from process failure. Many of the problematic images did not capture experimental results at all. They were out of focus, dark, overexposed, dropless, etc. Whatever labeling convention was initially followed, for the relabeling the "other" category was deemed to also diagnose problems with the imaging process.

A total of 42.6% of annotations for the images that were revisited disagreed with the original label, suggesting somewhat high (1 to 2%) label noise in this difficult fraction of the dataset. For a fraction of this data, multiple raters were asked to label the images independently and had an inter-rater disagreement rate of approximately 22%. The inherent difficulty of assigning a label to a small fraction of the images is therefore consistent with the results of Ref. [13]. Table 2 shows the final image counts after relabeling.

Table 2. Data distribution. Final number of images in the dataset for each category after collapsing the labels and relabeling.

	Number of images	
Label	Training	Validation
Crystals	56,672	6632
Precipitate	212,541	23,892
Clear	148,861	16760
Other	24,856	3,000

3 Machine Learning Model

The goal of the classifier here is to take an image as an input, and output the probability of it belonging to each of four classes (crystals, precipitate, clear, other) (see Fig. 1). The classifier used is a deep Convolutional Neural Network (CNN). CNNs, originally proposed in Ref. [25], and their modern 'deep' variants (see, e.g., Refs. [26, 27] for recent reviews), have proven to consistently provide reliable results on a broad variety of visual recognition tasks, and are particularly amenable to addressing data-rich problems. They have been, for instance, state of the art on the very competitive ILSVRC image recognition challenge [28] since 2012.

This approach to visual perception has been making unprecedented inroads in areas such as medical imaging [29] and computational biology [30], and have also shown to be human-competitive on a variety of specialized visual identification [31, 32]. The chosen classifier is thus well suited for the current analysis.

Fig 1. Conceptual Representation of a Convolutional Neural Network. A CNN is a stack of nonlinear filters (three filter levels are depicted here) that progressively reduce the spatial extent of the image, while increasing the number of filter outputs that describe the image at every location. On top of this stack sits a multinomial logistic regression classifier, which maps the representation to one probability value per output class (Crystals vs. Precipitate vs. Clear vs. Others). The entire network is jointly optimized through backpropagation [33], in general by means of a variant of stochastic gradient descent [34].

Model Architecture

The model is a variation on the widely-used Inception-v3 architecture [35], which was state of the art on the ILSVRC challenge around 2015. Several more recent alternatives were tried, including Inception-ResNet-v2 [36], and automatically generated variants of NASNet [37], but none yielded any significant improvements. An extensive hyperparameter search was also conducted using Vizier [38], also without providing significant improvement over the baseline.

The Inception-v3 architecture is a complex deep CNN architecture described in detail in Ref. [35] as well as the reference implementation [39]. We only describe here the modifications made to tailor the model to the task at hand.

Standard Inception-v3 operates on a 299x299 square image. Because the current problem involves very detailed, thin structures, it is plausible to assume that a larger input image may yield better outcomes. We use instead 599x599 images, and compress them down to 299x299 using an additional convolutional layer at the very bottom of the network, before the layer labeled `Conv2d_1a_3x3` in the reference implementation. The additional convolutional layer has a depth (number of filters) of 16, a 3×3 receptive field (it operates on a 3×3 square patch convolved over the image) and a stride of 2 (it skips over every other location in the image to reduce the dimensionality of the feature map). This modification improved classification absolute accuracy by approximately 0.3%. A few other convolutional layers were shrunk compared to the standard Inception-v3 by capping their depth as described in Table 3, using the conventions from the reference implementation.

Table 3. Limits applied to layer depths to reduce the model complexity. In each named layer of the deep network – here named after the conventions of the reference implementation – every convolutional subblock had its number of filters reduced to contain no more than these many outputs.

Layer	Max depth
`Conv2d_4a_3x3`	144
`Mixed_6b`	128
`Mixed_6c`	144
`Mixed_6d`	144
`Mixed_6e`	96
`Mixed_7a`	96
`Mixed_7b`	192
`Mixed_7c`	192

While these parameters are exhaustively reported here to ensure reproducibility of the results, their fine tuning is not essential to maximizing the success rate, and was mainly motivated by improving the speed of training. In the end, it was possible to train the model at larger batch size (64 instead of 32) and still fit within the memory of a NVidia K80 GPU (see more details in the training section below). Given the large

Fig 2. Classifier Accuracy. Accuracy on the training and validation sets as a function of the number of steps of training. Training halts when the performance on the evaluation set no longer increases ('early stopping'). As is typical for this type of stochastic training, performance increases rapidly at first as large training steps are taken, and slows down as the learning rate is annealed and the model fine-tunes its weights.

number of examples available, all dropout [40] regularizers were removed from the model definition at no cost in performance.

Data Preprocessing and Augmentation

The source data is partitioned randomly into 415990 training images and 47062 test images.

The training data is generated dynamically by taking random 599x599 patches of the input images, and subjecting them to a wide array of photometric distortions, identical to the reference implementation:

- randomized brightness ($\pm$ 32 out of 255),
- randomized saturation (from 50% to 150%),
- randomized hue ($\pm$ 0.2 out of 0.5),
- randomized contrast (from 50% to 150%).

In addition, images are randomly flipped left to right with 50% probability, and, in contrast to the usual practice for natural scenes which don't have a vertical symmetry, they are also flipped upside down with 50% probability. Because images in this dataset have full rotational invariance, one could also consider rotations beyond the mere 90°, 180°, 270° that these flips provide, but we didn't attempt it here, as we surmise the incremental benefits would likely be minimal for the additional computational cost. This form of aggressive data augmentation greatly improves the robustness of image classifiers, and partly alleviates the need for large quantities of human labels.

For evaluation, no distortion is applied. The test images are center cropped and resized to 599x599.

Training

The model is implemented in TensorFlow [41], and trained using an asynchronous distributed training setup [42] across 50 NVidia K80 GPUs. The optimizer is RmsProp [43], with a batch size of 64, a learning rate of 0.045, a momentum of 0.9, a decay of 0.9 and an epsilon of 0.1. The learning rate is decayed every two epochs by a factor of 0.94. Training completed after 1.7M steps (Fig. 2) in approximately 19 hours, having processed 100M images, which is the equivalent of 260 epochs. The model thus sees every training sample 260 times on average, with a different crop and set of distortions applied each time. The model used at test time is a running average of the training model over a short window to help stabilize the predictions.

4 Results

Classification

The original labeling gave rise to a model with 94.2% accuracy on the test set. Relabeling improved reported classification accuracy by approximately 0.3% absolute,

Table 4. Confusion Matrix. Fraction of the test data that is assigned to each class based on the posterior probability assigned by the classifier. For instance, 0.8% of images labeled as Precipitate in the test set were classified as Crystals.

True Label	Predictions			
	Crystals	Precipitate	Clear	Other
Crystals	**91.0%**	5.8%	1.7%	1.5%
Precipitate	0.8%	**96.1%**	2.3%	0.7%
Clear	0.2%	1.8%	**97.9%**	0.2%
Other	4.8%	19.7%	5.9%	**69.6%**

Table 5. Standard Deviation of the predictions across data sources. Note in particular the large variability in the consistency of the label 'Other' across datasets, which leads to comparatively poor selectivity of that less well-defined class.

True Label	Predictions			
	Crystals	Precipitate	Clear	Other
Crystals	**5%**	4%	1%	1%
Precipitate	2%	**4%**	1%	2%
Clear	1%	3%	**5%**	1%
Other	7%	15%	6%	**21%**

with the caveat that the figures are not precisely comparable since some of the test labels changed in between. The revised model thus achieves 94.5% accuracy on the test set for the four-way classification task. It overfits modestly to the training set, reaching just above 97% at the early-stopping mark of 1.7M steps. Table 4 summarizes the confusions between classes. Although the classifier does not perform equally well on images from the various datasets, the standard deviation in performance from one set to another is fairly small, about 5% (see Table 5), compared to the overall performance of the classifier.

The classifier outputs a posterior probability for each class. By varying the acceptance threshold for a proposed classification, one can trade precision of the classification against recall. The receiver operating characteristic (ROC) curves can be seen in Fig. 3.

Validation

At CSIRO C3 a workflow [44] has been set up which uses a variation of the analysis tool from DeepCrystal [45] to analyze newly collected crystallisation images and to assign either no score, 'crystal' score or 'clear' score. A total of 37,851 images were collected in Q1 2018 and assigned a human score by a C3 user were used as an independent dataset to test the MARCO tool. Within this dataset, 9746 images had been identified as containing crystals. The current, DeepCrystal tool (which assigns only 'crystal' or 'clear' scores) was found to have an overall accuracy rate of 74%, while the MARCO tool has 90%. Although this retrospective analysis doesn't allow for a direct comparison of the ROC, the precision, recall and accuracy of the two tools all favor the MARCO tool, as shown in table 6. The precision achieved by MARCO on this dataset is also very similar to that seen for the CSIRO images in the training data.

Pixel Attribution

We visually inspect to what parts of the image the classifier learns to attend by aggregating noisy gradients of the image with respect to its label on a per-pixel basis.

Table 6. Validation at C3 Precision, recall and accuracy from an independent set of images collected after the MARCO tool was developed. The 38K images of sitting drop trials were collected between January 1 and March 30, 2018 on two Formulatrix Rock Imager (FRI) instruments.

DL tool	Precision	Recall	Accuracy
DeepCrystal	0.4928	0.4520	0.7391
MARCO	0.7777	0.8663	0.9018

Fig 3. Receiver Operating Characteristic Curves. (Q) Percentage of the correctly accepted detection of crystals as a function of the percentage of incorrect detections (AUC: 98.8). 98.7% of the crystal images can be recalled at the cost of less than 19% false positives. Alternatively, 94% of the crystals can be retrieved with less than 1.6% false positives. (B) Percentage of the correctly accepted detection of precipitate as a function of the percentage of incorrect detections (AUC: 98.9). 99.6% of the precipitate images can be recalled at the cost of less than 25% false positives. Alternatively, 94% of the precipitates can be retrieved with less than 3.4% false positives.

The SmoothGrad [46] approach is used to visualize the focus of the classifier. The images in Fig. 4 are constructed by overlaying a heat map of the classifier's attention over a grayscale version of the input image.

Note that saliency methods are imperfect and do not in general weigh faithfully all the evidence present in an image according to their contributions to the decision, especially when the evidence is highly correlated. Although these visualizations paint a simplified and very partial picture of the classifier's decision mechanisms, they help confirm that it is likely not picking up and overfitting to cues that are irrelevant to the task.

Inference and Availability

The model is open-sourced and available online at [47]. It can be run locally using TensorFlow or TensorFlow Lite, or as a Google Cloud Machine Learning [48] endpoint. At time of writing, inference on a standard Cloud instance takes approximately 260ms end-to-end per standalone query. However, due to the very efficient parallelism properties of convolutional networks, latency per image can be dramatically cut down for batch requests.

Fig 4. Sample heatmaps for various types of images. (A) Crystal: the classifier focuses on some of the angular geometric features of individual crystals (arrows). (B) Precipitate: the classifier lands on the precipitate (arrows). (C) Clear: The classifier broadly samples the image, likely because this label is characterized by the absence of structures rather than their presence. Note the slightly more pronounced focus on some darker areas (circle and arrows) that could be confused for crystals or precipitate. Because the 'Others' class is defined negatively by the the image being not identifiable as belonging to the other three classes, heatmaps for images of that class are not particularly informative.

5 Discussion

Previous attempts at automating the analysis of crystallisation images have employed various pattern recognition and machine learning techniques, including linear discriminant analysis [49, 50], decision trees and random forests [51–53], and support vector machines [19, 54]. Neural networks, including self-organizing maps, have also been used classify these images [16, 55], with the most recent involving deep learning [56]. However, all previous approaches have required a consistent set of images with the same field of view and resolution, in order to identify the crystallization droplet in the well [22], and thereby restrict the analysis. Various statistical, geometric or textural features were then extracted, either directly from the image or from some transformation of the region of interest, to be used as variables in the classification algorithms.

The results from various studies can be difficult to compare head-to-head because different groups present confusion matrices with the number of classes ranging from 2 to 11, only sometimes aggregating results for crystals/crytalline materials. There is also a tradeoff between the number of false negatives and the number of false positives. Yet most report classification rates for crystals around 80-85% even in more recent work [8, 53, 57], in which missed crystals are reported with much lower rates. This advance comes at the expense of more false positives. For example, Pan et al. report just under 3% false negatives, but almost 38% false positives [54].

As the trained algorithms are specific to a set of images, they are also restricted to a particular type of crystallisation experiment. Prior to the curation of the current dataset, the largest set of images (by far) came from the Hauptman-Woodward Medical Research Institute HTCC [14]. This dataset, which contains 147,456 images from 96 different proteins but is limited to experiments with the microbatch-under-oil technique, has been used in a number of studies [56, 58]. Most notably, Yann et al. used a deep convolutional neural network that automatically extracted features, and reported a correct classification rates as high as 97% for crystals and 96% for non-crystals. Although impressive, these results were however obtained from a curated subset of 85,188 *clean* images, i.e., images with class labels on which several human experts agreed [56]. In order to validate our approach, we retrained our model to perform the same 10-way classification on that subset of the data alone without any tuning of the model's hyperparameters and achieved 94.7% accuracy, compared to the reported 90.8%.

In this context, the current results are especially remarkable. A crystallographer can classify images of experiments independently of the systems used to create those images. They can view an experiment with a microscope or look at a computer image and reach similar conclusions. They can look at a vapor diffusion experiment or a microbatch-under-oil setup and, again, asses either with confidence. Here, we show that this can be accomplished equally well, if not better, using deep CNNs. A benchtop researcher can classify many images, especially if they relate to a project that has been years in the making. For high-throughput approaches, however, that task becomes challenging. The strength of computational approaches is that each image is treated like the previous one, with no fatigue. Classification of 10,000 images is as consistent as classification of one. This advance opens the door for complete classification of all results in a high-throughput setting and for data mining of repositories of past image data.

Another remarkable aspect of our results is that they leverage a very generic computer vision architecture originally designed for a different classification problem – categorization of natural images – with very distinct characteristics. For instance, one can presume that the global geometric relationships between object parts would play a greater role in identifying a car or a dog in an image, compared to the very local, texture-like features involved in recognizing crystal-like structures. Yet no particular specialization of the model was required to adapt it to the widely differing visual

appearances of the samples originating from different imagers. This convergence of approaches toward a unified perception architecture across a wide range of computer vision problems has been a common theme in recent years, further suggesting that the technology is now ready for wide adoption for any human-mediated visual recognition task.

6 Conclusion

In this work, we have collated biomolecular crystallization images for nearly half a million of experiments across a large range of conditions, and trained a CNN on the labels of these images. Remarkably, the resulting machine-learning scheme was able to recapitulate the labels of more than 94% of a test set. Such accuracy has rarely been obtained, and has no equal for an uncurated dataset. The analysis also identified a small subset of problematic images, which upon reconsideration revealed a high level of label discrepancy. This variability inherent to using human labeling highlights one of the main benefits of automatic scoring. Such accuracy also make conceivable high-density screening.

Enhancing the imaging capabilities by including UV or SONICC results, for instance, could certainly enrich the model. But several research avenues could also be pursued without additional laboratory equipment. In particular, it should be possible to leverage side information that is currently not being used.

- The four-way classification scheme used is a distillation of 38 categories which are present in the source data. While these categories are presumed to be somewhat inconsistent across datasets, they could potentially provide an additional supervision signal.

- Because one goal of this classifier is to be able to generalize *across* datasets, it would be worthwhile to investigate the contribution of techniques that have been designed to specifically reduce the effect of domain shift across data sources on the classification outcomes [59, 60].

- Each crystallization experiment records a series of images taken over times. Using the timecourse information could enhance the success rate of the classifier [61].

Note in closing that the current study focused on crystallization as an outcome, which is but a small fraction of the protein solubility diagram. Patterns of precipitation, phase separation, and clear drops, also provide information as to whether and where crystallization might occur. The success in identifying crystals, precipitate and clear can be thus also be used to accurately chart the crystallization regimes and to identify pathways for optimization [58, 62, 63]. The application of this approach to large libraries of historical data may therefore reveal patterns that guide future crystallization strategies, including novel chemical screens and mutagenesis programs.

Acknowledgments

We acknowledge discussions at various stages of this project with I. Altan, S. Bowman, R. Dorich, D. Fusco, E. Gualtieri, R. Judge, A. Narayanaswamy, J. Noah-Vanhoucke, P. Orth, M. Pokross, X. Qiu, P. F. Riley, V. Shanmugasundaram, B. Sherborne and F. von Delft. PC acknowledges support from National Science Foundation Grant no. NSF DMR-1749374.

References

1. Harrison S, Lahue B, Peng Z, Donofrio A, Chang C, Glick M. Extending 'predict first' to the design make-test cycle in small-molecule drug discovery. Future Med Chem. 2017;9:533–536.
2. McPherson A. Crystallization of Biological Macromolecules. Cold Spring Harbor: CSHL Press; 1999.
3. Chayen NE. Turning protein crystallisation from an art into a science. Curr Opin Struct Biol. 2004;14(5):577–583.
4. Fusco D, Charbonneau P. Soft Matter Perspective on Protein Crystal Assembly. Colloids Surf B: Biointerfaces. 2016;137:22–31.
5. Ng JT, Dekker C, Reardon P, von Delft F. Lessons from ten years of crystallization experiments at the SGC. Acta Cryst D. 2016;72:224–235.
6. Fazio VJ, Peat TS, Newman J. Lessons for the future. Methods Mol Biol. 2015;1261:141–156.
7. Newman J, Bolton EE, Muller-Dieckmann J, Fazio VJ, Gallagher DT, Lovell D, et al. On the need for an international effort to capture, share and use crystallization screening data. Acta Cryst F. 2012;68(3):253–258.
8. Kotseruba Y, Cumbaa CA, Jurisica I. High-throughput protein crystallization on the World Community Grid and the GPU. J Phys Conf Ser. 2012;341(1):012027.
9. Newman J. One plate, two plates, a thousand plates. How crystallisation changes with large numbers of samples. Methods. 2011;55(1):73 – 80.
10. Zhang S, Gerard CJJ, Ikni A, Ferry G, Vuillard LM, Boutin JA, et al. Microfluidic platform for optimization of crystallization conditions. J Cryst Growth. 2017;472:18 – 28.
11. Thielmann Y, Koepke J, Michel H. The ESFRI Instruct Core Centre Frankfurt: Automated high-throughput crystallization suited for membrane proteins and more. J Struct Funct Genomics. 2012;13(2):63–69.
12. Snell EH, Lauricella AM, Potter SA, Luft JR, Gulde SM, Collins RJ, et al. Establishing a training set through the visual analysis of crystallization trials. Part II: crystal examples. Acta Cryst D. 2008;64(11):1131–1137.
13. Wilson J. Automated Classification of Images from Crystallisation Experiments. In: Perner P, editor. Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining. Springer Berlin Heidelberg;. p. 459–473.
14. Snell EH, Luft JR, Potter SA, Lauricella AM, Gulde SM, Malkowski MG, et al. Establishing a training set through the visual analysis of crystallization trials. Part I: ˜150 000 images. Acta Cryst D. 2008;64(11):1123–1130.
15. Hargreaves D. Private communication;.
16. Spraggon G, Lesley SA, Kreusch A, Priestle JP. Computational analysis of crystallization trials. Acta Cryst D. 2002;58(11):1915–1923.

17. Cumbaa C, Jurisica I. Automatic Classification and Pattern Discovery in High-throughput Protein Crystallization Trials. J Struct Funct Genomics. 2005;6(2):195–202.

18. Kawabata K, Saitoh K, Takahashi M, Asama H, Mishima T, Sugahara M, et al. Evaluation of protein crystallization state by sequential image classification. Sensor Rev. 2008;28(3):242–247.

19. Buchala S, Wilson JC. Improved classification of crystallization images using data fusion and multiple classifiers. Acta Cryst D. 2008;64(8):823–833.

20. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, et al. Overview of the CCP4 suite and current developments. Acta Cryst D. 2011;67(4):235–242.

21. MAchine Recognition of Crystallization Outcomes (MARCO); 2017. Available from: `https://marco.ccr.buffalo.edu`.

22. Vallotton P, Sun C, Lovell D, Fazio VJ, Newman J. DroplIT, an improved image analysis method for droplet identification in high-throughput crystallization trials. J Appl Crystallogr. 2010;43(6):1548–1552.

23. Rosa N, Ristic M, Marshall B, Newman J. Keeping Crystallographers App-y. Acta Cryst F;submitted.

24. Mele K, Li R, Fazio VJ, Newman J. Quantifying the quality of the experiments used to grow protein crystals: the iQC suite. Journal of Appl Cryst. 2014;47(3):1097–1106.

25. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural computation. 1989;1(4):541–551.

26. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436.

27. Rawat W, Wang Z. Deep convolutional neural networks for image classification: A comprehensive review. Neural Comput. 2017;29(9):2352–2449.

28. Berg A, Deng J, Fei-Fei L. Large scale visual recognition challenge (ILSVRC); 2010. Available from: `http://www.image-net.org/challenges/LSVRC`.

29. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.

30. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol. 2016;12(7):878.

31. Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. arXiv:171001711 [csCV]. 2017;(preprint).

32. Liu Y, Gadepalli K, Norouzi M, Dahl GE, Kohlberger T, Boyko A, et al. Detecting cancer metastases on gigapixel pathology images. arXiv:170302442 [csCV]. 2017;(preprint).

33. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533.

34. Bottou L. Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010. Springer; 2010. p. 177–186.

35. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 2818–2826.

36. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv:160207261 [csCV]. 2017;(preprint).

37. Zoph B, Vasudevan V, Shlens J, Le QV. Learning transferable architectures for scalable image recognition. arXiv:170707012 [csCV]. 2017;(preprint).

38. Golovin D, Solnik B, Moitra S, Kochanski G, Karro J, Sculley D. Google vizier: A service for black-box optimization. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2017. p. 1487–1495.

39. Silberman N, Guadarrama S. TensorFlow-Slim image classification model library; 2017. Available from: `https://github.com/tensorflow/models/tree/master/research/slim`.

40. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–1958.

41. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al.. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems; 2015. Available from: `https://www.tensorflow.org`.

42. Dean J, Corrado G, Monga R, Chen K, Devin M, Mao M, et al. Large scale distributed deep networks. In: Advances in neural information processing systems; 2012. p. 1223–1231.

43. Tieleman T, Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning. 2012;4(2):26–31.

44. Watkins C. C4, C3 Classifier Pipeline. v1. CSIRO. Software Collection.; 2018. Available from: `https://doi.org/10.4225/08/5a97375e6c0aa`.

45. DeepCrystal; 2017. Available from: `http://www.deepcrystal.com`.

46. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: Removing Noise by Adding Noise. arXiv:170603825 [csLG]. 2017;(preprint).

47. Vanhoucke V. Marco repository in TensorFlow Models; 2018. Available from: `http://github.com/tensorflow/models/tree/master/research/marco`.

48. Google Cloud Machine Learning Engine; 2018. Available from: `https://cloud.google.com/ml-engine`.

49. Cumbaa CA, Lauricella A, Fehrman N, Veatch C, Collins R, Luft J, et al. Automatic classification of sub-microlitre protein-crystallization trials in 1536-well plates. Acta Cryst D. 2003;59(9):1619–1627.

50. Saitoh K, Kawabata K, Asama H, Mishima T, Sugahara M, Miyano M. Evaluation of protein crystallization states based on texture information derived from greyscale images. Acta Cryst D. 2005;61(7):873–880.

51. Bern M, Goldberg D, Stevens RC, Kuhn P. Automatic classification of protein crystallization images using a curve-tracking algorithm. J Appl Cryst. 2004;37(2):279–287.

52. Liu R, Freund Y, Spraggon G. Image-based crystal detection: a machine-learning approach. Acta Cryst D. 2008;64(12):1187–95.

53. Cumbaa CA, Jurisica I. Protein crystallization analysis on the World Community Grid. J Struct Funct Genomics. 2010;11(1):61–69.

54. Pan S, Shavit G, Penas-Centeno M, Xu DH, Shapiro L, Ladner R, et al. Automated classification of protein crystallization images using support vector machines with scale-invariant texture and Gabor features. Acta Cryst D. 2006;62(3):271–279.

55. Po MJ, Laine AF. Leveraging genetic algorithm and neural network in automated protein crystal recognition. In: Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS'08 - "Personalized Healthcare through Technology"; 2008. p. 1926–1929.

56. Yann MLJ, Tang Y. Learning Deep Convolutional Neural Networks for X-Ray Protein Crystallization Image Analysis. In: Thirtieth AAAI Conference on Artificial Intelligence; 2016.

57. Hung J, Collins J, Weldetsion M, Newland O, Chiang E, Guerrero S, et al. Protein crystallization image classification with elastic net. In: SPIE Medical Imaging. vol. 9034. SPIE;. p. 14.

58. Fusco D, Barnum TJ, Bruno AE, Luft JR, Snell EH, Mukherjee S, et al. Statistical Analysis of Crystallization Database Links Protein Physico-Chemical Features with Crystallization Mechanisms. PLoS ONE. 2014;9(7):e101123.

59. Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning; 2015. p. 1180–1189.

60. Bousmalis K, Trigeorgis G, Silberman N, Krishnan D, Erhan D. Domain separation networks. In: Advances in Neural Information Processing Systems; 2016. p. 343–351.

61. Mele K, Lekamge BMT, Fazio VJ, Newman J. Using Time Courses To Enrich the Information Obtained from Images of Crystallization Trials. Cryst Growth Des. 2014;14(1):261–269.

62. Snell EH, Nagel RM, Wojtaszcyk A, O'Neill H, Wolfley JL, Luft JR. The application and use of chemical space mapping to interpret crystallization screening results. Acta Cryst D. 2008;64(12):1240–1249.

63. Altan I, Charbonneau P, Snell EH. Computational crystallization. Arch Biochem Biophys. 2016;602:12–20.