# De Novo Crystallization Condition Prediction with Deep Learning

**Hyunmin Lee[1,2], Zhen Hao Wu[1,2], Carles Corbi-Verge[2], Mac Mok[2], Sidney Kang[3], Shun Liao[1,2] Zhaolei Zhang[1,2,4], Michael Garton[3]**

Department of Computer Science[1]
Donnelly Center for Cellular & Molecular and Biomolecular Research[2]
Institute of Biomaterials and Biomedical Engineering[3]
Department of Molecular Genetics[4]

University of Toronto

## Abstract

X-ray crystallography is the most commonly used technique to determine three-dimensional structures of macromolecules. Despite its prevalence, protein crystallization is still a trial and error process that requires random screening using an extensive number of conditions. This motivates us to model the relationship between protein sequence and its crystallization conditions. Using deep learning on crystallization conditions extracted from the Protein Data Bank, we demonstrate that it is possible to predict crystallization conditions for a given sequence. Accurate condition prediction would facilitate more focused condition screening, which would reduce the time and cost required to generate diffraction quality crystals.

## 1 Introduction

Elucidating three-dimensional macro-molecular structures is of central importance to understanding their biological function, but it remains a major challenge. One of the most powerful methods to solve three-dimensional structures is X-ray crystallography, with more than 90% of the structures found in Protein Data Bank (PDB) solved with this method to this date [3]. However, producing a diffraction-quality protein crystal remains a bottleneck for X-ray crystallography. A large number of parameters affect crystallization outcome and current methods rely on trial and error sampling of the chemical space within crystallization conditions [16]. Crystallization protocols involve randomly screening hundreds, if not thousands, of buffer conditions using kits and robots [18]. ]. Even so, less than 2-10% of cloned proteins yielded diffraction quality crystals [19]. The time, cost and success rate associated with crystallization hampers the viability of X-ray crystallography and thus our efforts to understand the fundamental chemistry of life.

We aim to reduce the crystallization condition search space by learning the mapping between protein sequence and the conditions required to form its crystal. To achieve our goal, we extract sequence and crystallization conditions from the PDB and train a neural network to learn the mapping between the two. This approach can drastically reduce the number of conditions screened by identifying a subset of conditions that maximize the probability of crystal yield.

We formulate the crystallization condition prediction problem as a multi-label classification task. Given an input protein sequence, X, the goal is to predict binary vector, Y, where each position corresponds to a different crystallization term. We define crystallization term as the elements found within crystallization conditions, which includes technique, buffer, salts and other precipitating agents such as PEG, that are used to crystallize the protein.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

In this work, we evaluate several deep learning architectures for predicting crystallization conditions, including feed-forward neural networks (FFN) and convolutional neural networks (CNN). Through various metrics, we demonstrate that crystallization conditions can be predicted using a machine learning-based approach. Our findings suggest that computationally driven condition screening is feasible.

## 2 Related Work

We review computational predictions applied to various aspects of protein crystallization.

**Crystallization Propensity.** Some protein sequences tend to have higher crystallization propensity than others. Kurgan and Mizianty [15] found that only 4.6% of crystallization attempts successfully yielded diffraction-quality crystals. Janhandideh and Mahdavi [10] trained random forest models to predict crystallization propensity using various features, including amino acid compositions, isoelectric point, sequence length, and molecular weight. Wang $et\ al$ [20] applied support vector regression (SVR) models to predict the propensity, as well as the success rates of other experimental procedures including sequencing cloning, protein material production, protein purification, crystal production, and structure determination. More recently, DeepCrystal applied a Convolutional Neural Network (CNN) to predict crystallization propensity using only the protein sequence [9].

**Crystallization Outcome Classification.** Crystallization robots can screen through thousands of different conditions per day. To further increase the throughput, automated image classification algorithms can be used instead of human experts to identify the formation of crystals. Cumbaa $et\ al$ applied various models, including decision trees and support vector machines, to classify the presence of crystals from an image [7]. More recently, Bruno $et\ al$ found that CNN can achieve 94% test accuracy on the Machine Recognition of Crystallization Outcomes (MARCO) data set [4].

**Crystallization Conditions.** To reduce the number of trial and errors, different approaches attempted to identify ideal crystallization conditions for an arbitrary protein. Kimber $et\ al$ suggested building screens using conditions that had high success rates [12]. Cumbaa $et\ al$ applied association mining to group together proteins based on their crystallization conditions [8]. Other work explored correlating buffer pH with the protein isoelectric point to predict the crystallization conditions [11, 5, 21, 14]. However, recent studies suggests that the previous findings were not significant since, on average, crystallization occurs at neutral pH [13]. More recently, Abrahams and Newman [2] found no correlation between BLAST sequence similarity score and crystallization conditions for non-identical sequences.

To the best of our knowledge, direct prediction of crystallization conditions from raw sequence has not been explored.

## 3 Processing Protein Data Bank Files

We generated our dataset by extracting protein sequences and crystallization conditions from headers SEQRES and REMARK 280, respectively. Since crystallization conditions are written in free text format, we used a two step-approach to transform the text into structured data: 1). We generate the output domain by filtering for words (chemicals and techniques) with at least 50 occurrences, as words below this cutoff were mostly entered incorrectly (i.e spelling mistakes). 2). We implemented annotation mapping to remove the redundancy resulting from the use of either chemical formulas or names, as well as the same ions appearing in more than a single compound. This process resulted in 192 crystallization terms that consist of chemicals and crystallization techniques. After filtering out PDB files without any annotation and files with sequence lengths greater than 2000, we were able to curate data from 71659 PDB files.

## 4 Experiments

### 4.1 Model Architectures and Input Processing

The following hyperparameters were used for all models: Relu and sigmoid activation functions were used for our hidden layer and output layer, respectively. All models were trained for 500 epochs. We

also used ADAM [13], with a learning rate of 0.0005 and mini-batch size of 64. We train the models by minimizing the binary cross-entropy loss. The models were implemented using Keras [6] with a Tensorflow backend [1].

Our FFN architecture consists of three layers with 2048, 1024, and 512 nodes. Our CNN architectures consist of N number of 1-D convolutions with 16 filters and a window size of 50, followed by two feed-forward layers with 1024 and 512 nodes.

We process the protein sequences and their respective crystallization terms as one-hot encoded vectors. The protein sequences were padded with zeros to the max length (2000). We randomly split the data set into 90% and 10% for training and testing sets, respectively. 10% of the training set was randomly chosen to be used as the validation set.

### 4.2 Evaluation

Table 1: Weighted Precision, Recall and F1-Scores for crystallization condition prediction different architectures on predicting crystallization conditions. Top values are in bold for each metric.

| **Architecture** | **Validation Set** | | | **Test Set** | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1-Score | Precision | Recall | F1-Score |
| FFN | **0.628** | 0.325 | 0.356 | **0.631** | 0.337 | 0.364 |
| 1-CNN | 0.551 | **0.414** | **0.454** | 0.550 | **0.423** | **0.461** |
| 3-CNN | 0.518 | 0.421 | 0.447 | 0.524 | 0.431 | 0.456 |
| 5-CNN | 0.503 | 0.417 | 0.440 | 0.551 | 0.423 | 0.447 |

Table 1 shows the performance comparison between FFN and CNN with a varying number of convolutional layers. To address the label imbalance, we calculated the weighted precision, recall, and f1-score for each model: for each label, the metrics are calculated and weighted by their frequency. FNN model achieved F1-score of 0.364 while the 1,3 and, 5-layer CNN achieves 0.461, 0.456 and 0.447, respectively. We observe that models with convolutional layers achieve higher F1-score, which suggests the importance of sequence locality for predicting crystallization conditions.

### 4.3 Filter Visualization

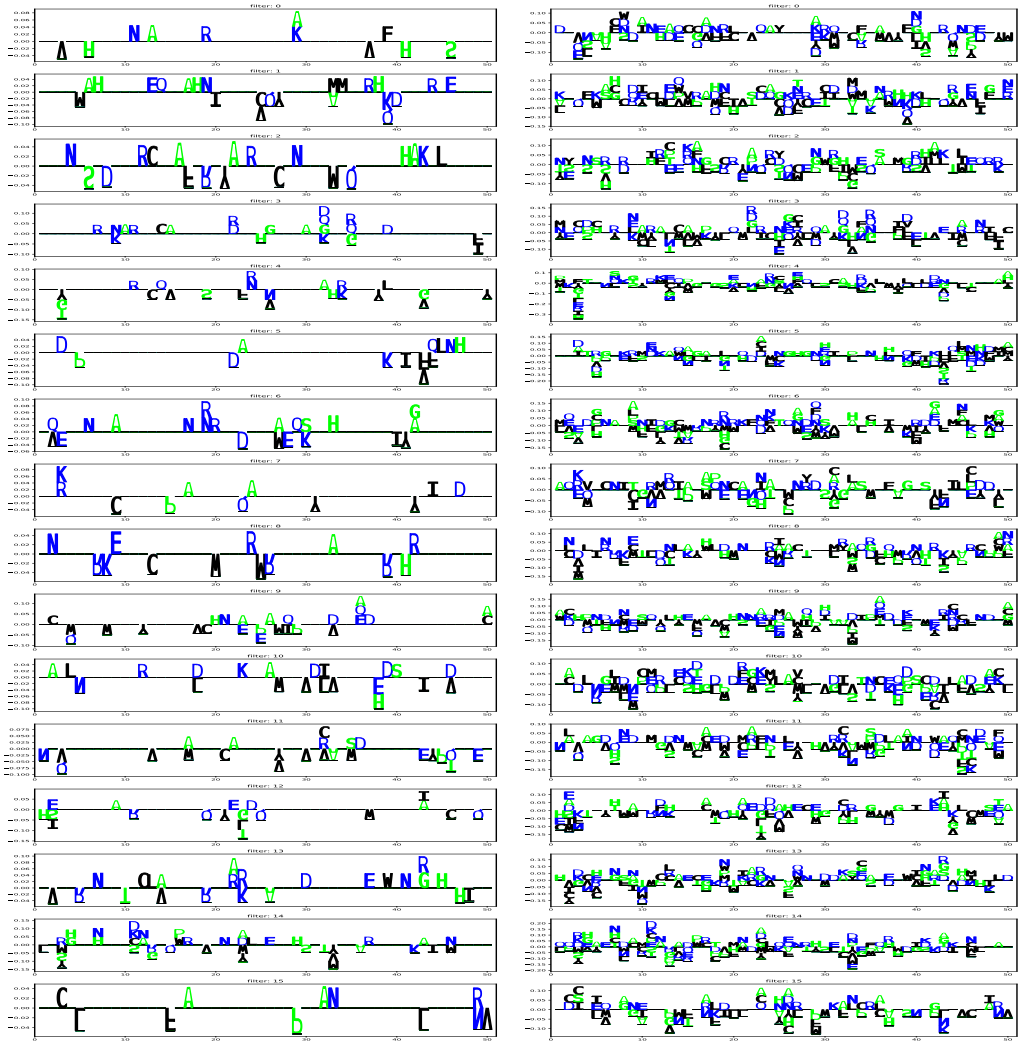
Figure 1 shows the filter position weight matrices (PWMs) for the first convolutional layer, using top 1% and top 5% of the filter weights. Interestingly, amino acids with the highest weights correspond to hydrophilic and neutral residues, whereas the majority of negative weights correspond to hydrophobic residues. Intuitively, we expect hydrophilic residues to have a strong influence on the buffer conditions since these residues affect the isoelectric point of a protein.

## 5 Conclusion and Future Work

Here we introduce a new problem of predicting crystallization conditions and demonstrate that deep learning methods enable accurate prediction of those conditions from sequence alone. We describe how data was generated for training and evaluation by parsing through PDB files using a two-step approach. Furthermore, we found that convolutional neural networks achieved the highest performance.

Price et al [17] suggest that the dominant factor for crystallization propensity is the presence of well-ordered surface epitopes that mediate inter-protein interaction. The features extracted from the filters may reflect local secondary structures and influence the conditions required. Further investigation is required to understand the influence of secondary structures on crystallization conditions.

Though here we focused exclusively on the crystallization terms, in future we may extend our model to include other variables such as temperature and pH. Also, rather than treating buffers as a class, we can train our models to directly infer the concentration of each buffer. We believe that our work complements other methods that focus on crystallization propensity and crystal image processing. We propose that a combination of these tools can be used to accelerate structure determination by X-ray crystallography.



(a) Top/Bottom 1% (b) Top/Bottom 5%

Figure 1: Filter position weight matrices from the first convolutional layer using top/bottom 1% and 5% of the filter weights. Blue, green and black represent hydrophilic (RKDENQ) , neutral (SGHTAP) and hydrophobic (YVMCLFIW) residues, respectively.


### Acknowledgments

We thank NVIDIA Corporation for providing GPUs used for this research.

## References


[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[2] G. J. Abrahams and J. Newman. Blasting away preconceptions in crystallization trials. *Acta Crystallographica Section F: Structural Biology Communications*, 75(3), 2019.

[3] H. Berman, K. Henrick, H. Nakamura, and J. L. Markley. The worldwide protein data bank (wwpdb): ensuring a single, uniform archive of pdb data. *Nucleic acids research*, 35(suppl_1):D301–D303, 2006.

[4] A. E. Bruno, P. Charbonneau, J. Newman, E. H. Snell, D. R. So, V. Vanhoucke, C. J. Watkins, S. Williams, and J. Wilson. Classification of crystallization outcomes using deep convolutional neural networks. *PLOS one*, 13(6):e0198883, 2018.

[5] M. Charles, S. Veesler, and F. Bonneté. Mpcd: a new interactive on-line crystallization data bank for screening strategies. *Acta Crystallographica Section D: Biological Crystallography*, 62(11):1311–1318, 2006.

[6] F. Chollet et al. Keras. `https://keras.io`, 2015.

[7] C. Cumbaa and I. Jurisica. Automatic classification and pattern discovery in high-throughput protein crystallization trials. *Journal of structural and functional genomics*, 6(2-3):195–202, 2005.

[8] C. A. Cumbaa and I. Jurisica. Protein crystallization analysis on the world community grid. *Journal of structural and functional genomics*, 11(1):61–69, 2010.

[9] A. Elbasir, B. Moovarkumudalvan, K. Kunji, P. R. Kolatkar, R. Mall, and H. Bensmail. Deepcrystal: a deep learning framework for sequence-based protein crystallization prediction. *Bioinformatics*, 35(13):2216–2225, 2018.

[10] S. Jahandideh and A. Mahdavi. Rfcrys: Sequence-based protein crystallization propensity prediction by means of random forest. *Journal of theoretical biology*, 306:115–119, 2012.

[11] K. A. Kantardjieff and B. Rupp. Protein isoelectric point as a predictor for increased crystallization screening efficiency. *Bioinformatics*, 20(14):2162–2168, 2004.

[12] M. S. Kimber, F. Vallee, S. Houston, A. Nečakov, T. Skarina, E. Evdokimova, S. Beasley, D. Christendat, A. Savchenko, C. H. Arrowsmith, et al. Data mining crystallization databases: knowledge-based approaches to optimize protein crystal screens. *Proteins: Structure, Function, and Bioinformatics*, 51(4):562–568, 2003.

[13] J. Kirkwood, D. Hargreaves, S. O'Keefe, and J. Wilson. Analysis of crystallization data in the protein data bank. *Acta Crystallographica Section F: Structural Biology Communications*, 71(10):1228–1234, 2015.

[14] J. Kirkwood, D. Hargreaves, S. O'Keefe, and J. Wilson. Using isoelectric point to determine the ph for initial protein crystallization trials. *Bioinformatics*, 31(9):1444–1451, 2015.

[15] L. Kurgan and M. J. Mizianty. Sequence-based protein crystallization propensity prediction for structural genomics: review and comparative analysis. *Nat Sci*, 1(2):93–106, 2009.

[16] J. R. Luft, E. H. Snell, and G. T. DeTitta. Lessons from high-throughput protein crystallization screening: 10 years of practical experience. *Expert opinion on drug discovery*, 6(5):465–480, 2011.

[17] W. N. Price Ii, Y. Chen, S. K. Handelman, H. Neely, P. Manor, R. Karlin, R. Nair, J. Liu, M. Baran, J. Everett, et al. Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data. *Nature biotechnology*, 27(1):51, 2009.

[18] R. C. Stevens. High-throughput protein crystallization. *Current opinion in structural biology*, 10(5):558–563, 2000.

[19] T. C. Terwilliger, D. Stuart, and S. Yokoyama. Lessons from structural genomics. *Annual review of biophysics*, 38:371–383, 2009.

[20] H. Wang, L. Feng, Z. Zhang, G. I. Webb, D. Lin, and J. Song. Crysalis: an integrated server for computational analysis and design of protein crystallization. *Scientific reports*, 6:21383, 2016.

[21] C.-Y. Zhang, Z.-Q. Wu, D.-C. Yin, B.-R. Zhou, Y.-Z. Guo, H.-M. Lu, R.-B. Zhou, and P. Shang. A strategy for selecting the ph of protein solutions to enhance crystallization. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 69(7):821–826, 2013.