Optimizing Associative Experimental Design for Protein Crystallization Screening

İmren Dinç*, Marc L. Pusey†, Ramazan S. Aygün*
*DataMedia Research Lab, Computer Science Department,
University of Alabama in Huntsville,
Huntsville, Alabama 35899, United States
†iXpressGenes, Inc., 601 Genome Way, Huntsville, Alabama 35806, United States
*{id0002, aygunr}@uah.edu †marc.pusey@ixpressgenes.com

***Abstract*—The goal of protein crystallization screening is the determination of the main factors of importance to crystallizing the protein under investigation. One of the major issues about determining these factors is that screening is often expanded to many hundreds or thousands of conditions to maximize combinatorial chemical space coverage for maximizing the chances of a successful (crystalline) outcome. In this paper, we propose an experimental design method called "Associative Experimental Design (AED)" and an optimization method includes eliminating prohibited combinations and prioritizing reagents based on AED analysis of results from protein crystallization experiments. AED generates candidate cocktails based on these initial screening results. These results are analyzed to determine those screening factors in chemical space that are most likely to lead to higher scoring outcomes, crystals. We have tested AED on three proteins derived from the hyperthermophile Thermococcus thioreducens, and we applied an optimization method to these proteins. Our AED method generated *novel* cocktails (count provided in parentheses) leading to crystals for three proteins as follows: Nucleoside diphosphate kinase (4), HAD superfamily hydrolase (2), Nucleoside kinase (1). After getting promising results, we have tested our optimization method on four different proteins. The AED method with optimization yielded 4, 3, and 20 crystalline conditions for holo Human Transferrin, archaeal exosome protein, and Nucleoside diphosphate kinase, respectively.**

***Index Terms*—Associative Experimental Design, Protein Crystallization, Screening, Screen Optimization, Experimental Design.**

I. Introduction

Crystallization is usually the bottleneck process in the determination of the 3 dimensional structure of a protein. One of the major difficulties in macromolecular crystallization is setting up the cocktails that yield a single large crystal for X-ray data collection [1], [2]. The problem is in the number of parameters that need to be tested. Physical, chemical and biochemical factors such as type of precipitants, type of salts, concentrations, pH value of the buffer, temperature of the environment, genetic modifications of the protein, etc. influence the crystallization process. Since each protein has a unique primary structure, it is quite challenging to determine the parameters of the experiment that can yield a crystal for a protein [3], [4]. The cost in time and materials renders exhaustive trial of all possible combinations of conditions practically impossible. As a result determination of the crystallization conditions is conducted using screening experiments.

The parameters for protein crystallization experiments are usually set by two main techniques [5], [6]: incomplete factorial experiments (IFE) [7], [8] or sparse matrix sampling (SMS) [2], [9]. Systematic grid screening (GS) of crystallization conditions is a complete factorial screen over a defined range. The incomplete factorial approach aims to determine the important factors of the experiments and to significantly reduce the number of experiments compared to full factorial design experiments [7]. The IFE is a beneficial tool especially when there are not enough resources, such available protein, to carry out those many experiments or it is practically discouraging to set up many experiments [10]. The IFE method generates balanced experiments with respect to the important factors of the experiments. The sparse matrix sampling (SMS) method [2] utilizes a wider range of major reagents conditions (i.e., pH values, type of precipitants, type of salts, etc.) in experiments. In SMS, type of salts, pH, and type of precipitants and their values are selected based on past experience to have resulted in protein crystallization. The reagents appear based on their frequency in the sparse matrix [9]. The sparse matrix approach was first put forth by Jancarik and Kim (1991), and their original screen, plus a wide range of variations, has been commercialized [11]. Grid screening of crystallization conditions is an early method that methodically varies a set of solution components over a range of conditions. This typically requires some insight into those parameters likely to produce crystals, and is more often carried out as part of the end game process following the successful determination of lead conditions by sparse matrix methods.

Once the results of these methods are obtained, a set of optimization methods can be applied [10]. The details of those optimization techniques can be found in the literature. These studies in macromolecular crystallization try to generate new cocktails or optimize available cocktails, which are supposed to yield crystals. The optimization steps in the literature generally involve changing the pH, concentration, concentrations of precipitants and salts. Basically, the traditional optimization techniques do not consider combinations of new reagents. In this paper, we propose an experimental design method called "Associative Experimental Design (AED)" [12] with the opti-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2016.2536030, IEEE Transactions on NanoBioscience

mization for protein crystallization experiments. By analyzing the outcome of preliminary experiments, the AED generates candidate cocktails identifying screening factors that are most likely to lead to higher scoring outcomes, crystals. Thus, AED is not just an optimization method for crystallization conditions, since it could generate novel conditions leading to crystals.

The AED method is not for initial screening operations, but for the determination of new conditions based upon analysis of the initial screens. Basically, the AED analyzes other possible interactions between reagents to determine new crystallization conditions. The output of AED is optimized by eliminating prohibited combinations and prioritizing reagents based upon their performance in the input screens. After identifying initial AED screens, the combinations known to produce a precipitate are eliminated. These combinations are identified either from the literature (for example [13], [14], [15]) or by empirical observation based on lab experiments. The rest of the AED screens are prioritized based on the association of participating reagents with better scoring outcomes. Thus, AED is different from GS, IFE, or SMS, which are currently in use. For example, in GS, the experts generally focus on a small chemical space and generate finer samples for a small set of reagents, making this impractical for covering extensive chemical space. On the other hand, in IFE, balanced crystallization screening experiments are generated using selected reagents, which allows analysis of a broad chemical space. One of the drawbacks of IFE, occurrence of each reagent for a factor is equal in the experiments; however, in the real world, some reagents might be more favorable for the crystallization trials compare to others [16]. SMS tries to overcome the limitations of IFE by increasing the occurrence of the reagents that are more favorable for the experiments based on existing experiment results. The frequency of each chemical used in SMS is generally calculated based on accumulated experimental results. Since the AED analyzes possible interactions between reagents to determine new crystallization conditions based on existing SMS results, it is different than the methods currently in use.

In this study, we tested the AED method in a wet lab, and ranking analysis is performed based on their distance to the conditions used as input for AED. After observing results in the wet lab, we confirmed that the AED generates novel crystalline conditions that did not appear in any of the commercial screens. In our preliminary study the AED method generated *novel* cocktails (count provided in parentheses) leading to crystals for three proteins as follows: Nucleoside diphosphate kinase (4), HAD superfamily hydrolase (2), Nucleoside kinase (1). After getting promising results, we have tested our optimization method on four different proteins. The AED method with optimization yielded 4, 3, and 20 crystalline conditions for holo Human Transferrin, archaeal exosome protein, and Nucleoside diphosphate kinase, respectively.

In this paper, for the wet lab experiments, trace fluorescent labeling was employed to assist the crystal finding and results interpretation process [17], [18]. The previous TFL experiments on a range of proteins showed no effect on the protein crystallization process or the X-ray diffraction results [18], [19]. As the method is used to identify crystals and likely crystallization conditions, if there is any concern then once these have been found subsequent crystals for diffraction analysis can be grown using protein which has not been labeled. As only the protein was fluorescently labeled, this enables rapid discrimination between protein and non-protein crystals in the drops. The central paradigm in trace fluorescent labeling is that the local fluorescence intensity is proportional to the concentration of the fluorescing species. As crystals are the most densely packed form for the protein, then they will fluoresce with the greatest intensity.

II. Background

A brief explanation of the phase diagram would help the reader to understand problem domain and the AED method better. In addition, we explain the scoring for protein crystallization experiments.

A. Phase Diagram

Normally, a protein is going to dissolve in a liquid up to a solubility limit which is a function of the protein and the solution conditions. The solubility is an equilibrium concentration defined in the presence of the solid (crystalline) phase. If the concentration of a solution is below the solubility limit, then that solution is said to be **under-saturated**; if it is exactly on the solubility limit, then it is called **saturated**. When the solution reaches the solubility limit, it is possible to increase its solubility by changing some physical factors such as pH, temperature, etc. If the concentration of the solution is above the solubility limit, then the solution will be **supersaturated**. This is the only region where a protein crystal can be grown. However, a supersaturated solution is not enough for crystallization to proceed by itself. A specific amount of activation energy and an ordered sequence of intermolecular interactions are required for initiating protein crystal nucleation, that eventually yields a protein crystal [20], [21].

A phase diagram illustrates the behavior of the protein with respect to the solution components and conditions. It is very important to locate solubility curve based on these parameters as the proteins can grow only in supersaturated solutions [20], [22]. Thus, the phase diagram is a very useful representation to set experiment parameters properly for X-ray diffraction studies [23]. A visual representation of a phase diagram is shown in Figure 1.

The two main zones of the phase diagram are the under-saturated and supersaturated regions. The supersaturated region has three subdivisions shown in Figure 1. In the labile zone, crystal nuclei can form and grow if the proper conditions are provided. Once the nucleation starts, protein crystals grow using the nutrients of the solution leading to a reduction in the protein concentration of the solution. The solution goes to the metastable region as the protein concentration decreases. In the metastable region, if there are nuclei that have already formed, the crystals may continue to grow until the concentration equals the solubility limit. This also means that new nuclei cannot form in that region [20]. If the supersaturation is too high, amorphous precipitates can also appear in precipitation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2016.2536030, IEEE Transactions on NanoBioscience

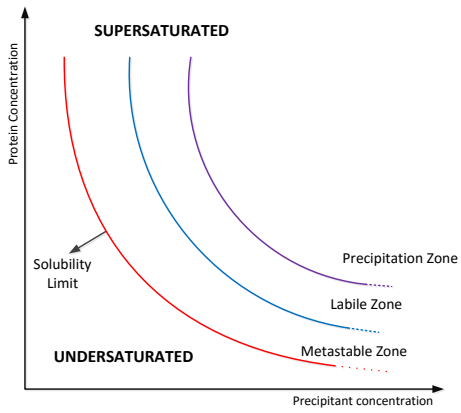

Fig. 1. Phase Diagram.

zone rather than formation of crystals, which is not a desirable result for crystallization process [20].

B. Scoring of Protein Crystallization Outcomes

In the literature, there are a variety of different scoring schemes for protein crystallization experiments. A results scoring scheme for use in analysis should show a linear scaling of the score with the desirability of the outcome. Sigdel et. al summarize some of the different scoring (or classification categories) methods in [24]. Since Hampton scoring [25] provides additional crystal subcategories with respect to other scoring scales, a revised version of Hampton scoring is used as shown in Table I for optimization and AED analysis. This revised scoring scale is put forth as better representing a progression in outcome desirability with higher numerical values [6]. Clear solutions (score = 2) can as well be solutions in the metastable region of the phase diagram where crystals simply have not nucleated, as well as being undersaturated protein. Distinguishing between a heavy and light precipitate is a judgment call, where consistency is more important than that it meets an (as yet undefined) absolute criteria. The bright spots outcome (score = 4) represents high intensity spots observed in the plates under fluorescent imaging which have no corresponding structure when viewed under white light imaging. As intensity is proportional to structure [18], [26] these can be considered as cryptic, or non-obvious, leads. Non faceted structures include items commonly referred to as dendrites, spheroids, and urchins. We provided some sample fluorescently labeled microscopic images for each score described above in Figure 2.

III. Associative Experimental Design(AED)

A. Motivation

Three different proteins were used with a single 96 condition screen to initially test our approach. For the preliminary testing we ignored crystal cocktails that have more than one type of salt or precipitant. This data consists of 9 different salt concentration values, 23 different type of salts, 9 different buffers, 26 different precipitant concentration values, 38 different precipitants, and 3 different protein concentration values.

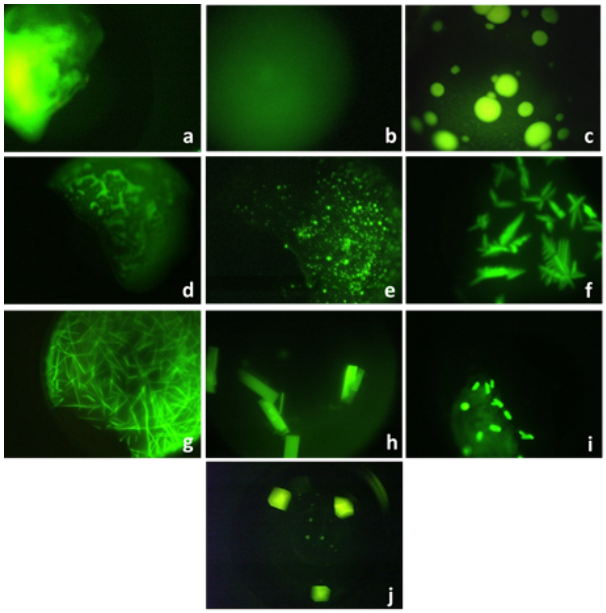

Fig. 2. Sample microscopic images of the protein crystallization outcomes. a- heavy amorphous precipitate, b- clear solution, c- phase change, d- light precipitate, e- bright spots or granular precipitate, f- spheroids, dendrites, urchins, g- needles, h- 2D plates, i- small 3D crystals, and j- large 3D crystals.

The concentrations and pH values are continuous data and the other features are categorical data. Since the type of buffer is generally correlated with pH value, it is not considered. Full factorial design for this single protein would require setting approximately 5,521,932 different experiments based on this dataset without considering the continuity of some of the variables, which is not feasible.

B. Method

Associative experimental design generates a new set of experiment conditions by analyzing the scores of screening experiments already carried out in the lab. Plate results are scored over the range 0 to 9, as listed in Table I. Since we are using trace fluorescent labeling (TFL) [19], a score of 4 is assigned to outcomes giving "bright spot" lead conditions. For AED let

$$D = \{(C_i, H_i) \mid (C_1, H_1), (C_2, H_2), ..., (C_n, H_n)\} \quad (1)$$

be our dataset containing the pairs that include features of the conditions C_i and their scores H_i for the i^{th} solution in the dataset. For simplicity we did not include conditions that have more than one type of salt or precipitant. AED uses the three main components of the remaining conditions: type of precipitant, type of salt and pH value of the solution, while separating their concentrations. Let

$$C_i = \{S_i [sc_i], pH_i, P_i [pc_i]\} \quad (2)$$

be the set of reagents of i^{th} crystal cocktail where i is $1 \leq i \leq n$, n is the number of samples in our dataset, $S_i [sc_i]$ represents type of salt with the concentration of sc_i, pH_i value represents the pH of i^{th} solution, and $P_i [pc_i]$ represents type

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2016.2536030, IEEE Transactions on NanoBioscience

TABLE I
LIST OF HAMPTON AND REVISED SCORES.

Hampton Scoring	Revised	Outcome	Hampton Scoring	Revised	Outcome
	0	heavy amorphous precipitate	5	5	spheroids, dendrites, urchins
1	2	clear solution	6	6	1D needles
2	3	phase change (oiling out)	7	7	2D plates
3	1	precipitate (light)	8	8	3D crystals small, $< 200\mu$m
4	4	bright spots or granular precipitate	9	9	3D crystals large, $> 200\mu$m

of precipitant with the concentration of pc_i. Let R be a subset of D that contains the crystal cocktail pairs with a score greater than or equal to low_H and less than or equal to $high_H$:

$$R = \{(C_i, H_i) \mid (C_i, H_i) \in D, low_H \leq H_i \leq high_H, 1 \leq i \leq n\} \quad (3)$$

In our preliminary experiments, we set the low score to 4 ($low_H = 4$) and the high score to 7 ($high_H = 7$). Therefore, the samples that have a score of 8 or 9 are excluded to generate unbiased conditions for the proteins. However, there is no harm to include these scores, as well. Similarly, for simplicity the samples with scores from 1 to 3 have not been included to the result set.

The AED analysis process consists of two major phases. In the first phase, we process the data to reduce its size as we stated before. Let

$$R_c = \{C_i \mid (C_i, H_i) \in R\} \quad (4)$$

denote the set of conditions of R, $SC_i = \{sc_1, sc_2, ..., sc_k\}$ represents the all unique concentration values of the i^{th} salt, and $PC_i = \{pc_1, pc_2, ..., pc_k\}$ represents the all unique concentration values of i^{th} precipitant. Then, we compare all C_i and C_j condition pairs in R_C where $i \neq j$. If C_i and C_j has a common component, then we generate the candidate conditions set Z based on these two sets. For example, assume that $C_i = \{S_i\,[SC_i]\,, pH_i, P_i\,[PC_i]\}$ and $C_j = \{S_j\,[SC_j]\,, pH_j, P_j\,[PC_j]\}$ where $S_i = S_j$ (i.e., the type of salt is common in C_i and C_j). We generate two new conditions Z by swapping the other components among each other. Therefore,

$$Z = \{\{S_i\,[SC_i]\,, \mathbf{pH_j}, \mathbf{P_i}\,[\mathbf{PC_i}]\}\,, \{S_i\,[SC_i]\,, \mathbf{pH_i}, \mathbf{P_j}\,[\mathbf{PC_j}]\}\} \quad (5)$$

is the set of candidate crystal cocktails for the pair C_i and C_j. In a similar way, candidate cocktails can be generated where pH value or precipitant is common between the pairs as well. After generating candidate combinations using these components, we remove conditions that are replicated or are already in the screening data (i.e., have known outcomes). In the second phase of our method, we assign unique values of concentrations, generating SC_i and PC_i, and unique type of buffers that were used in the preliminary data to generate finalized crystal cocktails. At the end, we merge generated results from two phases of the method. Then, if the number of candidate conditions are more than the desired number of cocktails or there are some bad combinations which are proved empirically, we apply an optimization method to generate a set of conditions. Our optimization method is described in detail in the following section. Examples of bad combinations are those known to result in a phase separation or where the two reagents react to form salt crystals. The basic steps of AED are shown below:

AED:

1) Data pre-processing,
2) Generate a list of cocktails score between 4 and 7
3) Generate triplets of salt, type of precipitant and pH value,
4) Find common reagents between each triplet pairs,
5) Generate two new cocktails by swapping different reagents,
6) Generate unique concentration values for each specific reagent,
7) Assign concentration values,

Optimization:

1) Eliminate prohibited combinations,
2) Prioritize remaining combinations,
3) Optimize the concentration values,
4) Rank prioritized cocktails

In order to increase robustness, after we get the preliminary results from AED, we have generated the family of the conditions from the cocktails having score 7, 8 or 9 for some of the proteins. Basically, the cocktails in a family consist of the same type of buffer, precipitant and salt with different concentrations. According to our results, we could get multiple crystals for a single family. In other words, the number of crystal in a family shows the robustness, the stability, and the reproducibility of that family. In Section V-C, we will provide brief information about these family of conditions.

Running time analysis: Since we are comparing each condition with the remaining conditions to find the common agent, the complexity of our algorithm is $O(n^2)$ where $n = |R|$. Considering today's plate sizes (up to 1536-well plate), we do not expect n is a very large number. Therefore, this implies $O(n^2)$ is a reasonable time for this problem.

1) Sample Scenario: Figure 3 shows the scores from four experiments using a commercial screen. The figure shows a partial graph of scores for common pH value of 6.5. These conditions generated four scores: 1, 1, 4, and 4. As it can be seen, none of the conditions lead to a good crystallization outcome for these conditions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2016.2536030, IEEE Transactions on NanoBioscience

Our AED method determines the common reagent between solutions that could lead crystallization conditions. In this example, there are only two promising conditions (with score 4): $[Zn(O_2CCH_3)_2, PEG\ 8K, pH = 6.5]$ and $[(NH_4)_2SO_4, PEG\ MME\ 5K, pH = 6.5]$. The AED draws a rectangle where these conditions (with score 4) are the two corners of this rectangle (Figure 3) and the other corners represent the candidate conditions. This scenario has two possible candidate conditions. One of them ($[(NH_4)_2SO_4, PEG\ 8K, pH = 6.5]$) already appeared in the commercial screen and yielded a low score. After conducting the experiment for the other condition ($[Zn(O_2CCH_3)_2, PEG\ MME\ 5K, pH = 6.5]$), we were able to get a score of 7 after optimizations. The experiments have not been conducted for others in the figure since they were not on the corners of conditions with promising scores.

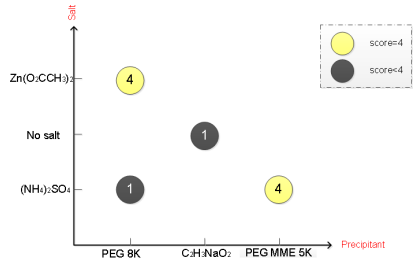

Fig. 3. Visual example for AED.

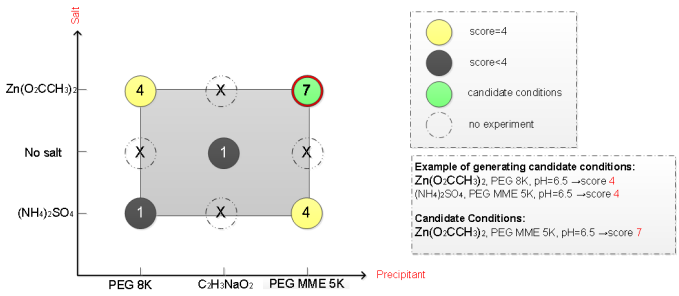

Fig. 4. Visual example for AED.

IV. Optimization of AED Cocktails

In AED analysis, the number of candidate cocktails depends on the number of cocktails that have scores from 4 to 7 in the input data. When AED generates more cocktails than the desired number (e.g., the number of wells in a plate) of cocktails, the experts may want to try the most promising candidate cocktails that need to be set. For example, if AED generates 150 candidate cocktails, the expert may want to know 96 cocktails to be tried for a 96-well plate. To resolve this problem, we employed an optimization process to eliminate cocktails having poor combinations of reagents and to prioritize the remaining conditions based on a metric. In this section, we will briefly explain the optimization method on AED cocktails.

A. Elimination of Prohibited Combinations

The output from the AED analysis usually results in more solution combinations than were present in the initial screen(s). The AED analysis indicates all of the possible unique combinations, and these are reduced to the final solutions by two processes. First is to remove "prohibited" combinations of reagents, such as mixtures known, either from the literature (for example [14], [14], [15]) or by empirical observation, to produce a precipitate, to produce a phase separation (e.g., high concentrations of PEG and a salt), those known to produce a precipitate, such as mixtures of divalent cations with particular anions such as phosphate or sulfate, or those that would tend to remove one or more of the components as unique entities in the solution, such as mixing divalent cations with diacid chelators such as EDTA or citrate. Additional unfavorable pairings are added to this list as they are empirically determined. Additionally, the output does not (yet) take into account the feasibility of attaining the final solutions on the basis of the available stock solution used for formulation. Thus, for example, stock trisodium citrate is 1.6M. A solution calling for 0.1 M buffer, 1.6M citrate, and possibly a third component cannot be made using the available stocks. Redundant outputs are also removed, such as 0.1M citrate buffer with citrate as precipitants 1 and 2.

B. Prioritization of Reagents

The second step of the optimization is a simple prioritization of the reagents for their association with better scoring outcomes. In this stage, the list of the reagents and scores is sorted with respect to the class of reagent being analyzed (buffer, precipitant, salt, etc.). For a candidate cocktail C that consists of precipitant p, buffer b, and salt s as reagents, the ratio of the average of the scores for the component of interest vs. all other scores is determined for the ranking. Let δ_p, δ_s, and δ_b represent the scores of the cocktails having precipitant p, salt s, and buffer b for a given screen file, respectively. Let Δ represent all scores of the input file. Then, the significance ratio, $\rho(\delta_r)$ for each class of reagent: precipitant, salt, and buffer, is computed as $\frac{\mu(\delta_p)}{\mu(\Delta-\delta_p)}, \frac{\mu(\delta_s)}{\mu(\Delta-\delta_s)}, and \frac{\mu(\delta_b)}{\mu(\Delta-\delta_b)}$, respectively. Those with significance ratio greater than 1 ($\rho(\delta_r) > 1$) perform better than the average while those with significance ratio less than 1 ($\rho(\delta_r) < 1$) perform worse. After identifying the components with highest significance ratios for each category, those components appearing with high significance ratios are tried in the wet lab.

Once the composition of the 96 conditions for the AED optimization screen has been determined, a pipetting table is generated to produce a block of 96 solutions of 1 mL volume, using the desired final concentrations for each reagent and the stock solution concentrations. In some cases, the stock reagent concentrations are not sufficiently high to produce the desired final solutions, typically indicated by a negative value for the calculated distilled water addition to bring the solution to the final volume. In such cases, either the concentration of one of the precipitants is reduced or an alternative set of solutions are used.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2016.2536030, IEEE Transactions on NanoBioscience

C. Ranking of Prioritized Conditions

In screen designing, it is important to know that whether a result cocktail is close to another cocktail in the input screen data to make a judgment about its outcome or priority. Chemical distance is a useful tool to evaluate the relationship between cocktails [27]. In this study, we applied a ranking method to the prioritized cocktails generated by AED based on how close they are to the crystal cocktails in the preliminary data. For example, in Figure 5, assume that the green points indicate the crystal cocktails with scores 4, 5, 6, 7, or 8, and red points indicate the AED results. The candidates close to the green points may have a higher chance to yield a good crystal compared to the other candidates.

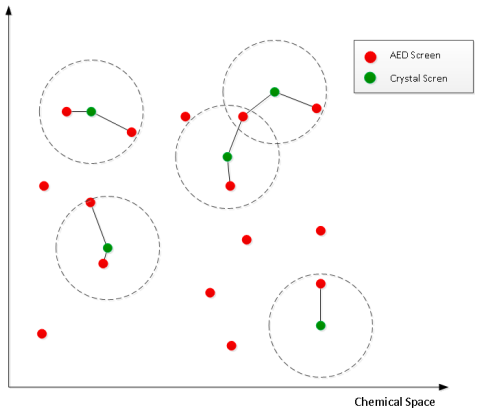

Fig. 5. Selecting the candidate cocktails.

For analyzing crystallization likelihood, we first calculate the distance from AED cocktails (red points) to all crystal cocktails (green points). At this point, we did not exclude any specific score from the input list, because even if the AED generates candidate cocktails using crystal conditions having score 4 to 7, it is still able to generate some cocktails that are close to 3D crystals in the chemical space. To calculate the distance between two cocktails, we used cocktail distance coefficient (CD_{coeff}) [28] given in Eq. 6:

$$CD_{coeff} = \frac{1}{sum(\omega)}((\frac{|E(pH_i)-E(pH_j)|}{14})\omega_1 + BC(F_i,F_j)\omega_2) \quad (6)$$

where $\omega = \omega_1, \omega_2$, $\omega_i \geq 0$ and $sum(\omega) > 0$. $E(pH_i)$ is the estimation of the pH in the cocktail, and $BC(F_i, F_j)$ is the Bray-Curtis dissimilarity measure[29] of fingerprints of the chemicals as in shown Eq. 7 and Eq. 8:

$$BC(F_i,F_j) = \sum_k |F_{ik} - F_{jk}| \sum_k |F_{ik} + F_{jk}| \quad (7)$$

and,

$$F_k = \sum_{i=1}^{n} f_{ik}[c_i] \quad (8)$$

where f_{ik} is the frequency count of descriptor k from the extended-connectivity fingerprints of component i, and c_i is the molar concentrations of the i^{th} component of the chemical. The detailed information about the calculation of the CD_{coeff} is provided in [28] and at the website[2].

Once the distances are calculated, for each AED cocktail, the minimum distance to each crystal class and also to all crystal classes in the preliminary data are taken. In this way, we obtain a matrix of distances to each crystal in the preliminary data. By using the minimum distance to any crystal, the lists are sorted in ascending order. This analysis is performed on the prioritized candidate cocktails.

One difficulty that we came across during the computation of distances is lack of formal standards of naming chemicals for commercial cocktails [30]. Since CD_{coeff} needs the molecular weight, molecular structure information of chemicals, etc., it uses a small database to calculate distances. The major problem of the computation is that the naming convention of the commercial cocktails does not match with the naming convention of the database. For that reason, we had to go through the screen files, and manually edited the chemicals based on the chemical naming provided with the program. Newman et al. has pointed out the problems of the lack of standards in this area [30].

D. Optimizing Concentration Values

The (current) goal of the optimization screen is to test the leading combinations over several concentrations. Thus, for precipitant X in buffer Y, with additive Z, the concentration of buffer Y and additive Z are kept constant (typically at 0.1 and 0.2 M respectively) while the concentration of precipitant X is varied. Concentrations of X are varied over three solutions, starting at the highest concentration indicated from either the AED analysis or by reference to the original screen compositions, and reducing by typically 20-25% for each of the next two solutions. Thus, a 96 condition screen results in 32 unique combinations of X, Y, and Z at 3 different concentrations of X.

A rapid reduction in the AED analysis listing can be carried out using the methods given. Output conditions are listed in order of their calculated priority scores, highest to lowest. Those with the highest priority scores are the mixtures containing the components judged most likely to result in crystals, while those with the lowest are the least likely. The final screen conditions are arrived at by going through the AED analysis and working down the priority listing. The AED analysis on its own gives new and unique combinations not present in the original screens, while the prioritization process gives the reagents associated with the highest scores. Optimization screens based solely on prioritization lead to a "cookie cutter" approach to optimization screen generation, where the same mixtures of precipitants are used with different buffers. Thus use of both approaches together is necessary for the most comprehensive optimization screen. Regardless, the initial screen conditions are constantly referred to when generating the AED optimization screen, primarily as a guide to reagent concentrations.

Three commercial screens were chosen to have a diverse array of precipitants with some overlap as defined by the

[2]https://github.com/ubccr/cockatoo/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2016.2536030, IEEE Transactions on NanoBioscience

C6 webtool [27]. The measured diversities are: $HRHT$ to $JCSG+ = 0.527$, $HRHT$ to $MCSG-3 = 0.489$, $JCSG+$ to $MCSG-3 = 0.367$. Some repetition of conditions is present, and these are used as internal controls for scoring and reproducibility. The fourth, Screen4a, was devised by examination of the components of the 3 commercial screens. A number of components are only present once or twice, and Screen4a was devised to increase the overall occurrence of these low frequency components, so that conclusions about their efficacy are not based upon a single result.

V. Experiments

A. Tools, Materials and Methods

We have developed a C# application that uses .Net Framework 4.0 in Visual Studio 2012. Basically, the users input the Microsoft Excel sheet that contains the screen information and the scores for each trial. Then, the program generates an output Excel sheet that contains ranked candidate cocktails for the future experiments.

Proteins were originally subjected to crystallization screening using a single 96 condition screen as previously reported [18]. Subsequent efforts have used four 96 conditions screens; Hampton Research High Throughput ($HRHT$, cat. $\#HR2-130$ [31]), Molecular Dynamics $JCSG+$ screen (cat. $\#MD1-40$ [32]), Microlytics $MCSG-3$ Screen (cat. $\#MCSG-3$ [33]), and a 96 condition screen under development in-house identified as Screen4a. All proteins were trace fluorescently labeled with the dye 5-(and 6)-carboxyrhodamine 6G (Molecular Probes cat.$\#C-6157$) prior to screening [19], [18]. Crystallization screening plates were set up using 96 well plates having 3 drop positions per well (Corning CrystalEX, cat. $\#3553$), with the protein: precipitant ratios (v/v) for the drops being 1:1, 2:2, and 4:1. Plates were imaged using the in-house developed Crystal X2 imager [24] (iXpressGenes/Molecular Dimensions), with the first set of images immediately after set up, on days 1, 2, 4, and thence on a weekly basis for the next 6 weeks. Plates were scored by visual observation, with the scores then adjusted by reference to the fluorescent images [18]. Thus the primary function of the fluorescent images was to remove non-protein objects from the data, the discovery of crystals that were missed by visual examination, and the assignment of scores of 4.

TABLE II
Parameters of the proteins

Protein	pI	MW	% -Helix	% -Sheet	% Coil
Tt82	4.85	27,900	34	5.8	24.5
Tt106	5.71	22,500	31.9	7.7	18.8
Tt189	5.8	19,600	24.1	6.5	25.9

B. Proteins for Preliminary Experiments

The proteins were chosen to have a range of scoring outcomes based upon a single crystallization screen. The three proteins employed in collecting preliminary data are: $Tt189$, annotated as a nucleoside diphosphate kinase; $Tt82$, annotated as a HAD superfamily hydrolase, and $Tt106$, annotated as a nucleoside kinase. These proteins were chosen as being facile, moderately difficult, and difficult crystallizers, respectively. Secondary structure predictions were made using NetSurfP [34]. Protein molecular weights and pI's were calculated using the ExPASY server [35]. A cutoff prediction of 0.8 was used to estimate the percent of secondary structural features for each protein. The protein parameters are given in Table II. In the case of $Tt106$, no crystals were obtained in the initial screening experiments, which involved 6 replicate plates [18].

C. Results for Preliminary Data

Optimization screens were devised based upon the AED analysis of the scored screening results, the 96 condition AED screens were then prepared and set up. For these preliminary data sets, the AED optimization screen conditions covered a broader range, with both precipitants 1 and 2 being varied over a range of conditions. Each grouping represents a family of screen conditions around a common theme, consisting of the same buffer and precipitants 1 and 2. Results analysis, as shown in Table IV, count the "families" where crystals were found, not the individual conditions. The results for Tt189 are shown in Figure 7, with each family of conditions outlined in red. For all three proteins the AED derived conditions were judged to be novel relative to the starting screen. When compared to all commercially available screens 7 of the 8 conditions were found to be novel, i.e., not occurring elsewhere. For the protein Tt106, the AED optimization screen only resulted in crystals after a second optimization round using additives with the AED-derived conditions.

Success and Novelty of AED Screens. The crystallization screen components that were determined to have the greatest positive effect were determined by the AED software, and a 96 condition optimization screen generated using those components for each protein. Optimization was in 96 well sitting drop plates, with the protein being TFL'd to facilitate results analysis. The successful conditions were identified and scored. Those conditions giving 2D and 3D crystals were then used to search the C6 database [27] for similar conditions across all commercially available screens as a determination of their uniqueness. Some sample images are provided in Figure 6. As the optimization screens had different concentration ratios for the same precipitant pairs, each ratio where a hit was obtained was searched and the lowest C6 score was used.

Table III shows the score distribution of preliminary data versus AED results. According to the table, AED generated more crystals than the preliminary data. Although AED results generated more crystals, not all cocktails are novel compared to all commercial cocktails. Table IV shows the number of novel conditions generated by AED. The numerical values in the first two columns after the protein name refer to the number of conditions with that score in the original screening experiment (numerator) vs. those with that score in the optimization screen (denominator). The third column lists the number of optimization conditions that are novel compared to the original screen, while the last column lists those that are novel compared to all available screens. All

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2016.2536030, IEEE Transactions on NanoBioscience

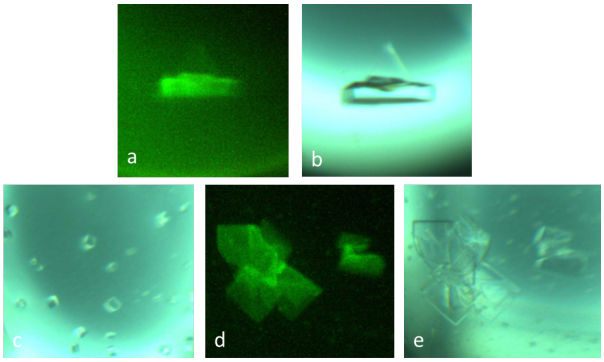

Fig. 6. Sample protein images (a-b) Tt82, (c-e) Tt189.

TABLE III
DATA DISTRIBUTION

	Tt189		Tt82		Tt106	
Score	**AED**	**HSHT**	**AED**	**HSHT**	**AED**	**HSHT**
0	0.00%	0.00%	10.42%	31.25%	0.00%	18.75%
1	3.13%	68.75%	65.63%	47.92%	30.21%	44.79%
2	40.63%	0.00%	13.54%	6.25%	32.29%	21.88%
3	6.25%	8.33%	5.21%	0.00%	4.17%	0.00%
4	3.13%	12.50%	0.00%	4.17%	0.00%	10.42%
5	23.96%	5.21%	2.08%	6.25%	12.50%	3.13%
6	1.04%	0.00%	2.08%	0.00%	0.00%	1.04%
7	12.50%	0.00%	0.00%	4.17%	10.42%	0.00%
8	9.38%	5.21%	1.04%	0.00%	10.42%	0.00%

found conditions were judged to be novel compared to the original screen on the basis of our cutoff score criteria. For $Tt189$, one optimization condition was identical to an existing commercial screen condition, but had no identity with any of the original input screen conditions.

The preliminary data indicated that scored results from commercially available screens can be analyzed, and that components that may contribute to the crystallization of the macromolecule can be derived. Not surprisingly, a number of novel conditions were found for the facile crystallizer ($Tt189$). However, conditions were also found for both the moderate and difficult crystallizers, one of which had not shown any results of needles or better in the original screens ($Tt106$). For all three proteins, crystallization conditions were obtained that were novel combinations of the identified factors.

D. Expanded screen analysis

The proteins employed are a protein from the archaeal exosome complex RrP42 plus the three described above from the hyperthermophilic archaeon *Thermococcus thioreducens* [36], an inorganic pyrophosphatase from *Staphylococcus aureus*, and human holo transferrin (hTFN, Sigma, cat.# T-4132).

The proteins were subjected to the expanded screen tests and the results obtained are given in Table V. In this case only outcomes giving faceted 3D crystals are used for an endpoint. For these proteins the AED optimization screen conditions were in groups of 3, and each condition giving a crystal was counted.

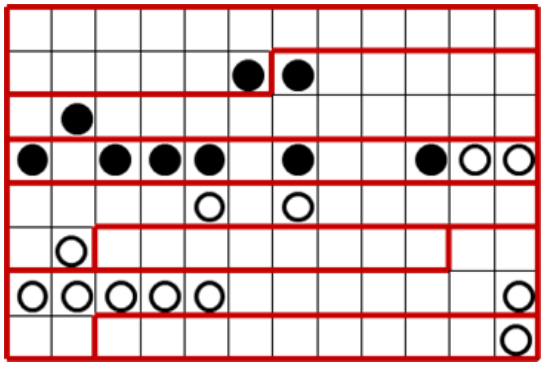

Fig. 7. Results for the preliminary data AED screen of protein Tt189. The filled black circles represent conditions where 3D crystals were obtained, while the open circles are those where 2D plate crystals were obtained. Each family of conditions is outlined in red.

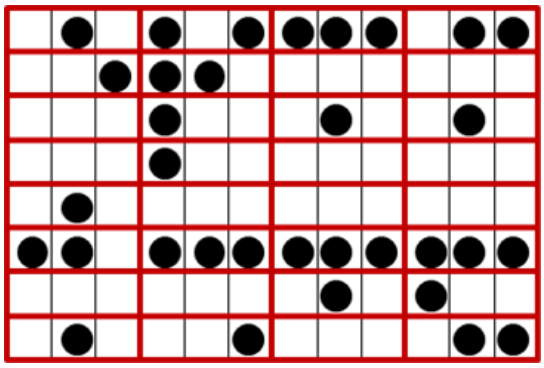

Fig. 8. AED optimization screen for protein Tt189, the screen in this case is generated using the combined results from 4 different 96 condition crystallization screens. The individual families of conditions are outlined in red. Only those conditions resulting in 3D crystals are shown.

Protein Tt189, from the preliminary results, was repeated. The results (Figure 8) indicate that of the 32 families of conditions optimized 20 of them resulted in 3D crystals (63%) compared to the 3 out of 7 (43%) from the preliminary data. The results shown in Figure 8 also indicate where the more "robust" crystallization conditions are to be found, those where all three concentrations of precipitant 1 resulted in crystals.

For Staphylococcus aureus IPPase ($SaIPP$), 2 crystals were obtained in the 4 screens or the AED optimized screen. However, the AED screen did result in a number of conditions that had a score of 5, non-faceted crystals. The analysis had indicated that low MW polyethylene glycols, divalent cations, and basic pH's were the lead factors for obtaining crystals. The AED derived screen results confirmed the high pH and low MW polyethylene glycols, and further indicated that Ca++, but not Mn++ or Mg++, was the best divalent cation. Every well containing Ca++ resulted in spheroids or rough non-faceted crystals, while none of those containing Mn++ or Mg++ had any. While these are not suitable for diffraction analysis, they can be used as a source of seed crystals [37]. The optimization conditions were subsequently tested using crystallization by capillary counter diffusion [38], which resulted in the two hits obtained.

For three of the four proteins more crystallization conditions

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2016.2536030, IEEE Transactions on NanoBioscience

TABLE IV
SUMMARY OF EXPERIMENTS

Protein Annotated Function	$HSHT$ Screen [2] Score = 7	Optimize Screen Score = 8, 9	Novel Cond. vs HSHT Screen*	Novel Cond. vs All Screens*
$Tt189$ (Nucleoside diphosphate kinase)	0 / 2	5 / 3	5	4
$Tt82$ (HAD superfamily hydrolase)	1 / 1	0 / 1	2	2
$Tt106$ (Nucleoside kinase)	0 / 0	0 / 1	1	1
			* Using C6 tool for scores of 7, 8, & 9	threshold value of 0.3

TABLE V
OPTIMIZATION RESULTS.

Protein	# of Crystals 4X screens / AED conditions (family's)
holo Human Transferrin	1 /5 (4)
RrP42 (archaeal exosome protein)	4 /15 (7)
Tt189 (nucleoside diphosphate kinase)	10 / 33 (20)
Tt106 (nucleoside kinase)	1 /9 (6)
Tt82 (HAD superfamily hydrolase)	8/ 3 (2)
Stapylococcus aureus inorganic pyrophosphatase	0 / 2 (2)

were determined by the AED screen than were found using the 4 "set" screens. In two of these cases, more families of conditions were determined.

E. Evaluation of Ranked Results

1) $Bin-Recall$ Metric: In order to evaluate the reliability of our ranking results, we compared our results with real outcome of the experiment. For this purpose, we developed a new metric called "$Bin-Recall$" to evaluate the performance of the ranking method. The traditional ranking methods are sensitive to irrelevant samples appearing before relevant samples. For protein crystallization, if all relevant cocktails are included in a well-plate, it is not critical to have screens not leading to crystals. We partition the list of cocktails into bins and analyze the number of relevant (crystalline) screens in each bin. Ideally, the good candidate cocktails should appear in bins that correspond to the top of the ranked list. Since we use bins and check how many of the relevant cocktails are included in a bin, we name our metric as $Bin-Recall$. Hence, $Bin-Recall$ measures how "close" the cocktails that yield crystals to the top of the ranked list. It generates a normalized value, which is close to 1 (or 100%) when the ranking results are similar to the best case, and it is 0 when the results are far from the best case.

$Bin-Recall$ is computed based on the formulation given in Eq. 9:

$$R_{bin} = \frac{\sum_{j=1}^{|B|} \delta_j (\sum_{i=S_{cmin}}^{S_{cmax}} \omega_i n_{i,j}) - \sum_{i=1}^{n} S_i \delta_{\lfloor \frac{n-i}{binSize} \rfloor} \omega(S_i)}{\sum_{i=1}^{n} S_i \delta_{\lfloor \frac{i}{binSize} \rfloor} \omega(S_i) - \sum_{i=1}^{n} S_i \delta_{\lfloor \frac{n-i}{binSize} \rfloor} \omega(S_i)} \tag{9}$$

where $|B|$ is the number of bins, δ_j is the weight of the bin j, ω_i is the weight of the score i, and $n_{i,j}$ is the number of score i in bin j. S is the list of ordered scores, where $cmin$ is the minimum crystal score and $cmax$ is the maximum crystal score. The denominator of the expression is used to normalize the measure dividing by the best scenario (i.e., all crystalline conditions appear in the top bin) minus the worst scenario (i.e., all crystalline conditions appear in the lower bins). The numerator computes the value based on the distribution of the scores to the bins and subtracts the worst case. $Bin-Recall$ measure allows to give high weights (ω_i) to cocktails or samples having high scores. Similarly, bins can also be assigned weights (δ_j) based on where all crystalline conditions should appear. In our case, the top bin having low distances to crystals is given the highest weight.

To illustrate the evaluation process, an example is provided in Figure 9. In the figure, the actual distribution of the scores is provided on left, when we split data 50% into Bin #1 and 50% into Bin #2, the best possible partitioning will be as in the second column of the figure. We evaluate this set of cocktails with the scores with respect to 2 sample cases given in the figure using $Bin-Recall$. As we mentioned before, $Bin-Recall$ measures the ranking results considering the appearance of the crystals in the highest group (Bin #1). When we have two different ranking such as Case 1 and Case 2, $Bin-Recall$ helps us compare those rankings. We may assign some weights for the bins and scores as shown in the figure, which can be determined by the experts based on their priorities. In our cases, we give the highest weight to the "Score 8" among scores and to Bin #1 among bins, which means the appearance of score 8 in the first 50% of the data has the highest priority. Based on these weights, the computation of $Bin-Recall$ for Case 1 is:

$$R_{bin} = \frac{2*\begin{bmatrix}8\\7\\6\\5\end{bmatrix}^T * \begin{bmatrix}8\\4\\5\\3\end{bmatrix} + 1*\begin{bmatrix}8\\7\\6\\5\end{bmatrix}^T * \begin{bmatrix}2\\2\\4\\2\end{bmatrix} - 1*\begin{bmatrix}8\\7\\6\\5\end{bmatrix}^T * \begin{bmatrix}10\\6\\9\\5\end{bmatrix}}{2*\begin{bmatrix}8\\7\\6\\5\end{bmatrix}^T * \begin{bmatrix}10\\6\\9\\5\end{bmatrix} - 1*\begin{bmatrix}8\\7\\6\\5\end{bmatrix}^T * \begin{bmatrix}10\\6\\9\\5\end{bmatrix}} = 68.16\% \tag{10}$$

and $Bin-Recall$ for Case 2 is:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2016.2536030, IEEE Transactions on NanoBioscience

$$R_{bin} = \frac{2*\begin{bmatrix}8\\7\\6\\5\end{bmatrix}^T * \begin{bmatrix}3\\2\\1\\1\end{bmatrix} + 1*\begin{bmatrix}8\\7\\6\\5\end{bmatrix}^T * \begin{bmatrix}7\\4\\8\\4\end{bmatrix} - 1*\begin{bmatrix}8\\7\\6\\5\end{bmatrix}^T * \begin{bmatrix}10\\6\\9\\5\end{bmatrix}}{2*\begin{bmatrix}8\\7\\6\\5\end{bmatrix}^T * \begin{bmatrix}10\\6\\9\\5\end{bmatrix} - 1*\begin{bmatrix}8\\7\\6\\5\end{bmatrix}^T * \begin{bmatrix}10\\6\\9\\5\end{bmatrix}} = 24.38\% \quad (11)$$

According to the $Bin - Recall$ results for Case 1 and Case 2, we can conclude that Case 1 provides a better ranking than Case 2 using given weights. The results of the real experiments are provided next.

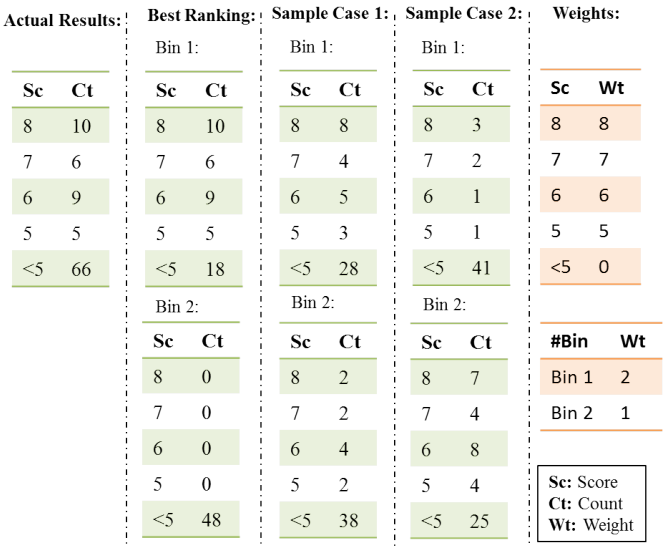

Fig. 9. Sample case for $Bin - Recall$.

2) Evaluation of AED Screens using Bin-Recall: We applied $Bin - Recall$ to our preliminary results. After we rank the cocktails, for each ranking scheme, we divided the results into 3 equal groups ($|B| = 3$ in Eq. 9) starting from the best cocktail and partitioned the data into bins using the crystal scores to evaluate the performance of ranking. Table VI shows the partitioning into bins for the different methods of ranking of Tt189, Tt82 and Tt106.

According to the table, Bin #1 shows the bin of the best 33.3% of the data, Bin #2 shows the second best 33.3%, and so on. In order to compare the different ranking schemes, we calculated the $Bin - Recall$ metric of each ranking. While we are calculating the metric, we used the actual scores as the weights of the scores (ω_i in Eq. 9). For example, the weight of score 6 is 6. We used 3, 2, and 1 as weights (δ_j in Eq. 9) of the Bin #1, Bin #2, and Bin #3, respectively, to give more importance to the bin appearing at the top. That means Bin #1 should have the most promising cocktails.

We calculated $Bin - Recall$ using 3-Bin (33.3%, 33.3%, and 33.3%) and 2-Bin partitioning (66.6%, and 33.3%). Table VII shows the $Bin - Recall$ results for each protein. In the table, the first and second columns show $Bin - Recall$ metric for 3-Bin and 2-Bin partitioning, respectively.

TABLE VI
3-BIN PARTITION OF THE PROTEINS BASED ON DIFFERENT RANKING SCHEMES.

		Score	Bin 1	Bin 2	Bin 3
Protein Tt189	**Min distance to score 4**	5	9	6	8
		6	1	0	0
		7	4	5	3
		8	2	2	5
	Min distance to score 5	5	6	10	7
		6	0	1	0
		7	2	4	6
		8	6	2	1
	Min distance to score 8	5	7	10	6
		6	1	0	0
		7	0	4	8
		8	2	5	2
	Min distance to all crystals	5	10	7	6
		6	1	0	0
		7	2	4	6
		8	4	3	2
Protein Tt82	**Min distance to score 4**	5	0	2	0
		6	1	1	0
		7	0	0	0
		8	1	0	0
	Min distance to score 5	5	1	1	0
		6	0	2	0
		7	0	0	0
		8	0	0	1
	Min distance to score 7	5	0	2	0
		6	0	1	1
		7	0	0	0
		8	0	0	1
	Min distance to all crystals	5	0	2	0
		6	1	1	0
		7	0	0	0
		8	1	0	0
Protein Tt106	**Min distance to score 4**	5	4	6	2
		6	0	0	0
		7	3	6	1
		8	6	2	2
	Min distance to score 5	5	5	4	3
		6	0	0	0
		7	4	4	2
		8	5	4	1
	Min distance to all crystals	5	4	6	2
		6	0	0	0
		7	3	6	1
		8	6	2	2

According to the Table VII, the ranking based on minimum distance to class 4 or 5 may also give good results as much as the ranking based on the minimum distance to all crystal classes. However, the ranking based on the minimum distance to all crystal classes is more consistent than the other rankings. When we consider the ranking scheme with 3 Bins, we can reach 74.15%, 73.33%, and 63.33% $Bin - Recall$, for Tt189, Tt82 and Tt106, respectively. When we consider 2-Bin, we get 84.12%, 100.00%, and 84.29% $Bin - Recall$ for each protein. Please note that the first 66.6% of the cocktails for Tt82 contains the all conditions that yield crystals.

Ideally, the goal is to obtain $Bin - Recall$ value of 100%. It depends on the expert to determine the number of bins for analysis. The results show that 66.6% of the prioritized screens can cover all crystalline outcomes of prioritized screens.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2016.2536030, IEEE Transactions on NanoBioscience

TABLE VII
$Bin - Recall$ RESULTS OF 3-BIN AND 2-BIN PARTITION

Protein:	Method	3 Bins (33.3%, 33.3%, 33.3%)	2 Bins (66.6%, 33.3%)
Protein Tt189	Min dis to score 4	48.58%	52.36%
	Min dis to score 5	73.24%	84.66%
	Min dis to score 8	66.45%	81.59%
	Min dis to all	74.15%	84.12%
Protein Tt82	Min dis to score 4	73.33%	100.00%
	Min dis to score 5	45.00%	73.33%
	Min dis to score 7	26.67%	53.33%
	Min dis to all	73.33%	100.00%
Protein Tt106	Min dis to score 4	63.33%	84.29%
	Min dis to score 5	41.19%	82.38%
	Min dis to all	63.33%	84.29%

VI. SUMMARY

There are reasons beyond simply obtaining a crystal for using a method such as the AED analysis:

- **Finding more robust conditions.** Crystal nucleation is a stochastic process, and it is not uncommon to set up the same condition multiple times with varying outcomes [27], [18]. The AED analysis approach not only helps to find new crystallization conditions, but also, as implemented herein, finds more "robust" crystallization conditions, i.e., those that are less sensitive to the concentration of one or more of the components present. This is shown in Figure 8, where for each family, there are three different concentrations of precipitant #1. Those conditions that are more sensitive are identified by only one outcome having 3D crystals in a family, and those that are less sensitive have crystals in all three concentrations of precipitant #1.
- **Improving existing conditions.** The existing found crystallization conditions may not be readily repeatable, or may not give crystals diffracting to a sufficient resolution. AED analysis can reveal an expanded range of conditions, some or many of which may resolve these problems.
- **Possibly new space groups (to facilitate binding analysis).** Binding studies where potential ligands are soaked into a crystal to determine their location upon diffraction analysis, require that the binding sites be available, not occluded by crystallographic contents. Space groups obtained in initial screening experiments may not be suitable for these studies, prompting a search for new packing arrangements.
- **Improved diffraction resolution.** Having good looking crystals does not automatically translate to good diffraction resolution. However, having crystals where previously one had none, such as with the protein Tt106, does markedly improve one's chances of obtaining a structure. Thus, a primary reason for the AED analysis is to find crystallization conditions where there previously were none. Additionally, crystal nucleation is a stochastic process. From Figure 7 and Figure 8, we see that there are families having many crystallization conditions, and families only having 1 or none. It is intuitively apparent that those with many conditions are more robust, less sensitive to component concentrations and more likely to result in crystals, than those with few conditions. This is important when carrying out additional screening trials and optimizations for improved diffraction resolution and for studies such as for substrate binding or drug development.
- **Improved crystal size (for neutron diffraction).** Although not shown in the data presented, the AED optimization results yielded a range of crystal sizes. Neutron diffraction requires crystals $\leq 1mm^3$ in size. Conditions that favor larger crystals can be determined from these results and are likely a more favorable starting point for growth of large volume crystals.

As shown by comparing Figures 7 and 8, using more screens in the initial search gives a larger search space for the AED analysis. Commercially available screens have a finite number of precipitants present. Increasing the number of screens results in exposure to an expanded range of conditions, although some are only present in 1 or 2 of the conditions. For this reason we formulated Screen4a, to increase the occurrence of these occasional precipitants to complement the other 3 screens.

Not all proteins yielded crystals upon AED optimization screening. In the case of Tt106, the crystals were obtained from the AED-identified conditions after additional optimization using crystallization additives. In the case of SaIPP, the AED analysis indicates those conditions, which should be most likely to result in crystals, and as such is the starting point for subsequent screening experiments. AED analysis results in screen conditions, thus screens, that are formulations of the components most likely to yield crystals for that protein.

Ranking of prioritized cocktails is also an important feature for the experts, and distance metric for cocktails is an efficient tool to rank the cocktails. Once the cocktails are ranked, it is important to evaluate accuracy of the ranked results. $Bin - Recall$ is an effective metric to compare different ranking approaches based on our results.

VII. CONCLUSION & FUTURE WORK

Although the results obtained to date are promising, the AED analysis is currently under development. The results and practical considerations indicate a several promising avenues for future development. For example, a near term goal is the separation of anion and cation effectiveness in salts. If analysis indicates that iodide is a more effective anion than chloride, and potassium is a more effective cation than sodium, then potassium iodide may be inferred to be a more effective solution component even if it does not occur in any of the screens employed. To carry this logic further, one could prepare the buffers with potassium or iodide as the counter ions.

Another direction to be explored is to feed back the AED screen results through a second round of analysis. Using this approach might be further improved by using more permutations, increasing the first round number of families of conditions to 48, or even just 96, to keep the first round of

AED optimization screening broad. We are currently using just 32 conditions for the optimization screen. Use of 3 different levels of precipitant #1 provides a limited amount of a more systematic grid screening data during the first round AED optimization trial. Reduction in the number of variations set up for a given AED analysis output condition would result in an expansion of the lead components that are explored at the expense of this limited grid screen data. Would the first round optimization screen be better as 2 variations for a total of 48 sets, possibly 4 conditions with both precipitants #1 and #2 being varied, or a straight use of just the AED output parameter for 96 new conditions.

VIII. Acknowledgment

This research was supported by National Institutes of Health (GM090453) grant and (GM116283) grant.

References

[1] J. S.-I. Kwon, M. Nayhouse, P. D. Christofides, and G. Orkoulas, "Modeling and control of protein crystal shape and size in batch crystallization," *AIChE Journal*, vol. 59, no. 7, pp. 2317–2327, 2013.

[2] J. Jancarik and S.-H. Kim, "Sparse matrix sampling: a screening method for crystallization of proteins," *Journal of applied crystallography*, vol. 24, no. 4, pp. 409–411, 1991.

[3] R. Giegé, "A historical perspective on protein crystallization from 1840 to the present day," *FEBS Journal*, vol. 280, no. 24, pp. 6456–6497, 2013.

[4] A. McPherson and B. Cudney, "Optimization of crystallization conditions for biological macromolecules," *Structural Biology and Crystallization Communications*, vol. 70, no. 11, pp. 1445–1467, 2014.

[5] R. C. Stevens, "High-throughput protein crystallization," *Current opinion in structural biology*, vol. 10, no. 5, pp. 558–563, 2000.

[6] D. E. Brodersen, G. R. Andersen, and C. B. F. Andersen, "*Mimer*: an automated spreadsheet-based crystallization screening system," *Acta Crystallographica Section F*, vol. 69, no. 7, pp. 815–820, Jul 2013. [Online]. Available: http://dx.doi.org/10.1107/S1744309113014425

[7] C. W. Carter Jr and C. W. Carter, "Protein crystallization using incomplete factorial experiments," *J. Biol. Chem*, vol. 254, no. 23, pp. 12 219–12 223, 1979.

[8] C. Abergel, M. Moulard, H. Moreau, E. Loret, C. Cambillau, and J. C. Fontecilla-Camps, "Systematic use of the incomplete factorial approach in the design of protein crystallization experiments." *Journal of Biological Chemistry*, vol. 266, no. 30, pp. 20 131–20 138, 1991.

[9] J. A. Doudna, C. Grosshans, A. Gooding, and C. E. Kundrot, "Crystallization of ribozymes and small rna motifs by a sparse matrix approach," *Proceedings of the National Academy of Sciences*, vol. 90, no. 16, pp. 7829–7833, 1993.

[10] J. R. Luft, J. Newman, and E. H. Snell, "Crystallization screening: the influence of history on current practice," *Structural Biology and Crystallization Communications*, vol. 70, no. 7, pp. 835–853, 2014.

[11] E. H. Snell, R. M. Nagel, A. Wojtaszcyk, H. O'Neill, J. L. Wolfley, and J. R. Luft, "The application and use of chemical space mapping to interpret crystallization screening results," *Acta Crystallographica Section D: Biological Crystallography*, vol. 64, no. 12, pp. 1240–1249, 2008.

[12] İ. Dinç, M. L. Pusey, and R. S. Aygün, "Protein crystallization screening using associative experimental design," in *Bioinformatics Research and Applications*. Springer, 2015, pp. 84–95.

[13] S. Raja, V. R. Murty, V. Thivaharan, V. Rajasekar, and V. Ramesh, "Aqueous two phase systems for the recovery of biomolecules–a review," *Science and Technology*, vol. 1, no. 1, pp. 7–16, 2011.

[14] J. A. Asenjo and B. A. Andrews, "Aqueous two-phase systems for protein separation: a perspective," *Journal of Chromatography A*, vol. 1218, no. 49, pp. 8826–8835, 2011.

[15] J. Asenjo and B. A. Andrews, "Aqueous two-phase systems for protein separation: phase separation and applications," *Journal of Chromatography A*, vol. 1238, pp. 1–10, 2012.

[16] A. McPherson and J. A. Gavira, "Introduction to protein crystallization," *Acta Crystallographica Section F: Structural Biology Communications*, vol. 70, no. 1, pp. 2–20, 2014.

[17] M. L. Pusey, M. S. Paley, M. B. Turner, and R. D. Rogers, "Protein crystallization using room temperature ionic liquids," *Crystal growth & design*, vol. 7, no. 4, pp. 787–793, 2007.

[18] M. Pusey, J. Barcena, M. Morris, A. Singhal, Q. Yuan, and J. Ng, "Trace fluorescent labeling for protein crystallization," *Structural Biology and Crystallization Communications*, vol. 71, no. 7, 2015.

[19] E. Forsythe, A. Achari, and M. L. Pusey, "Trace fluorescent labeling for high-throughput crystallography," *Acta Crystallographica Section D: Biological Crystallography*, vol. 62, no. 3, pp. 339–346, 2006.

[20] N. Asherie, "Protein crystallization and phase diagrams," *Methods*, vol. 34, no. 3, pp. 266–272, 2004.

[21] B. Rupp, "Origin and use of crystallization phase diagrams," *Acta Crystallographica Section F: Structural Biology Communications*, vol. 71, no. 3, pp. 247–260, 2015.

[22] H. Yang and Å. C. Rasmuson, "Phase equilibrium and mechanisms of crystallization in liquid–liquid phase separating system," *Fluid Phase Equilibria*, vol. 385, pp. 120–128, 2015.

[23] K. Baumgartner, L. Galm, J. Nötzold, H. Sigloch, J. Morgenstern, K. Schleining, S. Suhm, S. A. Oelmeier, and J. Hubbuch, "Determination of protein phase diagrams by microbatch experiments: Exploring the influence of precipitants and ph," *International journal of pharmaceutics*, vol. 479, no. 1, pp. 28–40, 2015.

[24] M. Sigdel, M. L. Pusey, and R. S. Aygun, "Real-time protein crystallization image acquisition and classification system," *Crystal growth & design*, vol. 13, no. 7, pp. 2728–2736, 2013.

[25] "Hampton Research Scoring," https://hamptonresearch.com/ documents/product/hr005490_2-130_user_guide.pdf, accessed: 2015-11-01.

[26] A. Meyer, C. Betzel, and M. Pusey, "Latest methods of fluorescence-based protein crystal identification," *Acta Crystallographica Section F: Structural Biology Communications*, vol. 71, no. 2, pp. 121–131, 2015.

[27] J. Newman, V. J. Fazio, B. Lawson, and T. S. Peat, "The c6 web tool: a resource for the rational selection of crystallization conditions," *Crystal Growth & Design*, vol. 10, no. 6, pp. 2785–2792, 2010.

[28] A. E. Bruno, A. M. Ruby, J. R. Luft, T. D. Grant, J. Seetharaman, G. T. Montelione, J. F. Hunt, and E. H. Snell, "Comparing chemistry to outcome: the development of a chemical distance metric, coupled with clustering and hierarchal visualization applied to macromolecular crystallography," 2014.

[29] J. R. Bray and J. T. Curtis, "An ordination of the upland forest communities of southern wisconsin," *Ecological monographs*, vol. 27, no. 4, pp. 325–349, 1957.

[30] J. Newman, T. S. Peat, and G. Savage, "What's in a Name? Moving Towards a Limited Vocabulary for Macromolecular Crystallisation," *Australian Journal of Chemistry*, vol. 67, no. 12, p. 1813, Jul. 2014.

[31] "Hampton Research Screen HT," https://hamptonresearch.com/documents /product/hr000783_crystal_screen_2.xls, accessed: 2015-11-01.

[32] "Molecular Dynamics JCGS+ Screen," http://www.moleculardimensions.com/applications/upload/Md1-40%20JCSG%20Plus%20HT-96.pdf, accessed: 2015-11-01.

[33] "Microlytics MCSG-3 Screen," http://www.microlytic.com/sites/default /files/MCSG3_Formulations_0_0_0.pdf, accessed: 2015-11-01.

[34] B. Petersen, T. N. Petersen, P. Andersen, M. Nielsen, and C. Lundegaard, "A generic method for assignment of reliability scores applied to solvent accessibility predictions," *BMC structural biology*, vol. 9, no. 1, p. 1, 2009.

[35] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, and A. Bairoch, *Protein identification and analysis tools on the ExPASy server*. Springer, 2005.

[36] E. V. Pikuta, D. Marsic, T. Itoh, A. K. Bej, J. Tang, W. B. Whitman, J. D. Ng, O. K. Garriott, and R. B. Hoover, "Thermococcus thioreducens sp. nov., a novel hyperthermophilic, obligately sulfur-reducing archaeon from a deep-sea hydrothermal vent," *International journal of systematic and evolutionary microbiology*, vol. 57, no. 7, pp. 1612–1618, 2007.

[37] A. D'Arcy, T. Bergfors, S. W. Cowan-Jacob, and M. Marsh, "Microseed matrix screening for optimization in protein crystallization: what have we learned?" *Acta Crystallographica Section F: Structural Biology Communications*, vol. 70, no. 9, pp. 1117–1126, 2014.

[38] J. D. Ng, J. A. Gavira, and J. M. Garcıa-Ruız, "Protein crystallization by capillary counterdiffusion for applied crystallographic structure determination," *Journal of structural biology*, vol. 142, no. 1, pp. 218–231, 2003.