UNIVERSTIY OF HAMBURG

MASTER THESIS

Predicting Protein Crystallization Conditions using Machine Learning

Author:
Michael HÜPPE

Supervisors:
Prof. Dr. Fabian Michael Kern
Prof. Dr. Aymelt Itzen

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Machine Learning in Bio Informatics
Department of Computer Science

Thursday 27th November, 2025

Declaration of Authorship

I, Michael HÜPPE, declare that this thesis titled, "Predicting Protein Crystallization Conditions using Machine Learning" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"This one is for the boys with the booming system"

Nicki Minaj

UNIVERSTIY OF HAMBURG

Abstract

Faculty Name
Department of Computer Science

Master of Science

Predicting Protein Crystallization Conditions using Machine Learning

by Michael HÜPPE

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Acknowledgements

The acknowledgments and the people to thank go here, don't forget to include your project advisor...

Contents

List of Figures

List of Tables

List of Abbreviations

Physical Constants

Speed of Light $c_0 = 2.997\,924\,58 \times 10^8\,\mathrm{m\,s^{-1}}$ (exact)

List of Symbols

a	distance	m
P	power	W ($\mathrm{J\,s^{-1}}$)
ω	angular frequency	rad

For/Dedicated to/To my…

Chapter 1

Decoding the Crystal Recipe: Predicting Protein Crystallization Conditions via Machine Learning

1.1 Introduction

Protein crystallization is the process of arranging purified protein molecules into a highly ordered, repeating lattice that forms a crystal. This crystalline state is essential for X-ray crystallography, the most widely used method for determining high-resolution protein structures. When X-rays are diffracted by a protein crystal, the resulting patterns allow reconstruction of the electron density and ultimately the atomic structure of the protein.
Crystallization is crucial because structural information provides fundamental insights into protein function and interactions which build the basis of structure-based drug design. Here binding sites derived from the structure enables rational development of small molecules that inhibit or modulate the protein's biological function.
Although AlphaFold predictions often align remarkably well with experimentally determined structures, they are not a substitute for them. Terwilliger *et al*. (2024) argue that AlphaFold accelerates, but cannot replace, experimental structure determination due to its varying local accuracies and occasional failures at the global structural level. This reinforces that, despite the breakthroughs brought by AlphaFold, further methodological innovation in structure determination and crystallography remains essential. Beyond crystallography, protein crystals are also used in neutron diffraction, cryo-electron microscopy benchmarking and biophysical studies of stability and folding.
Because most proteins do not readily form diffraction-quality crystals, the main bottleneck in structural biology is identifying the crystallization conditions under which a given protein will form suitable crystals (Mall *et al*., 2025). Despite advances in structure prediction, identifying suitable crystallization conditions, such as buffer type, pH, salts, and precipitants remains largely empirical and often requires screening hundreds of combinations. This process is both time-consuming and costly, with a high failure rate.

At the same time, extensive data from the Protein Data Bank (PDB) and accurate structural predictions from models like AlphaFold2 have become widely available (Jumper *et al*., 2021). These resources offer a unique opportunity to explore whether machine learning models can predict suitable crystallization conditions directly from protein sequence and/or structure.

1.2 Related Works

1.2.1 Challenge of X-ray Crystallography

Crystallizing a protein is often the bottleneck in X-ray crystallography. Only a small fraction (roughly 2–10%) of proteins produce diffraction-quality crystals, meaning over 90% of crystallization trials fail (Mall *et al.*, 2025). This trial-and-error process is costly, where >70% of the total expense in structure determination is spent on attempts not producing crystals of diffraction quality (Mall *et al.*, 2025). Achieving crystals requires finding the right crystallization conditions (e.g. precipitant chemicals, salts, pH, temperature), but currently these must be determined empirically by screening hundreds or thousands of conditions, which demands a lot of protein (McPherson & Gavira, 2013). This reality motivates computational approaches to predict either whether a protein is likely to crystallize (crystallization propensity) or even which specific conditions might lead to crystals. The former is a widely researched topic with models such as CrystalP2 (Kurgan *et al.*, 2009) or PPCPred (Mizianty & Kurgan, 2011), the latter however has received little to no attention (Jin *et al.*, 2022). While crystallization robots have eased the burden of manual search, discovering optimal conditions still means testing thousands of solutions, wasting protein in the process (Wilson & DeLucas, 2014). Any such predictive model could significantly reduce experimental screening, saving time and cost.

1.2.2 Predicting Protein Crystallization

As mentioned predicting crystallization propensity is a well established research area. Early **classical ML** relied on hand-crafted features such as amino-acid composition, the proteins isoelectric point, hydrophobicity, disorder, predicted structure and classifiers (SVM, RF, LR, GB). Matinyan *et al.* (2024) summarizes multiple tools such as:

- *XtalPred*/XtalPred-RF: feature-distribution scoring.
- *TargetCrys*: two-layer SVM ensemble.
- *Crysalis* (2016): integrated predictions + mutation suggestions.
- *BCrystal* (2020): XGBoost + SHAP-based feature selection.
- *DCFCrystal* (2021): cascaded RF stages (expression → purification → crystallization), with a membrane-protein branch.

These achieved moderate accuracy (60–75%, MCC 0.4–0.6) but demanded expert feature engineering. Deep learning automatically extracts sequence patterns as shown in superior performances presented by:

- *DeepCrystal* (2019): multi-scale CNN on one-hot sequences; ∼ 83% accuracy, MCC 0.66.
- *CLPred* (2020): CNN+BLSTM; accuracy 85%, MCC 0.70.
- *ATTCrys* (2021): adds multi-head attention; MCC 0.72.
- Structure-infused: *SADeepCry* (2022) uses autoencoder+self-attention on sequence+predicted structure; *GCmapCrys* (2023) employs GNNs on AlphaFold2 contact maps.

1.2.3 Predicting Crystallization Conditions from Sequence (and Structure)

While predicting "will it crystallize?" is useful, a more ambitious goal is to predict the actual crystallization conditions that would make a given protein form crystals. This is a multi-output prediction (the combination of reagents, concentrations, pH, etc. that will work) and more challenging. Thousands of successful crystallization recipes are known (recorded in databases like the Protein Data Bank), but each protein is unique and may crystallize in different conditions, often unpredictably with no known patterns to predict crystallization conditions (Zhang *et al.*, 2022). However, this does not stem from a lack of attention, Kirkwood *et al.* (2015) for example studied correlations between a protein's isoelectric point and the pH of its crystallization buffer. However, only weak or inconsistent trends were found. Ultimately, Kirkwood *et al.* (2015) conclude that these trends are not sufficiently robust to guide initial crystallization-pH selection. Interestingly, even proteins with high sequence similarity did **not** necessarily crystallize under similar conditions, highlighting that small sequence/structure differences can lead to different optimal crystallization cocktails indicating a multidimensional problem.

Liao and Sun (2025) emphasize the importance of crystal packing and intermolecular packing interfaces in determining crystallization conditions. In their work, they present Molecular Assembly Simulation in Crystal Lattice (MASCL), a framework for simulating crystal packing using AlphaFold combined with symmetrical docking. Crystallization conditions are predicted using a patch-based method that quantifies molecular interface similarity between proteins. For a given target protein, proteins with the most similar physicochemical interface descriptors are used as reference points from which test crystallization conditions are chosen. Meaning no de-novo crystallization conditions are constructed.

This pipeline of constructing a "crystal fingerprint" for a given protein and comparing it to previously crystallized proteins was evaluated on lysozyme, a common model protein in crystallization studies. In this test case, the proposed AAI-PatchBag approach successfully identified conditions yielding crystals with the desired packing characteristics. However,the use of lyzosome as a model system has been repeatedly criticized (Chayen & Saridakis, 2001). Particularly in the context of assessing prediction accuracy lysozome benchmarks should be assed with caution, because it is well known and used for its unusually high crystallizability (Ghosh, 2023). It crystallizes across a broad range of pH values without loss of crystal quality (Iwai *et al.*, 2008), and also tolerates wide variations in temperature and salt concentration (Ataka & Asai, 1988).

Nevertheless, for lysozyme specifically, Liao and Sun (2025) show that similarity in crystal packing information has a stronger influence on predicting successful crystallization conditions than sequence or structural homology alone.

In contrast Lee *et al.* (2019) introduced a proof-of-concept deep learning model to map protein sequence to de-novo crystallization conditions. They parsed crystallization records from PDB entries and framed the task as a multi-label classification: for a given protein sequence, predict which "crystallization terms" (e.g. specific buffers, salts, precipitants like PEG, etc.) appeared in successful recipes for that protein. Essentially, the model learns associations between sequence features and the types of reagents or techniques that tend to be used. Remarkably, this sequence-to-condition model did show predictive power. A simple 1-layer CNN could achieve a weighted F1-score around 0.46 on held-out proteins, substantially better than random guessing in this high-dimensional space. The CNN outperformed a fully-connected network, suggesting that local sequence motifs (captured by convolutional filters) were

informative for certain crystallization agents. For example, the authors noted that hydrophilic or charged residues in the sequence strongly influenced buffer and salt predictions (likely because they affect the protein's isoelectric point and solubility). This indicates real biochemical signal: proteins rich in acidic/basic residues might require certain pH buffers or salt conditions, etc., whereas hydrophobic patches might correlate with needing precipitating agents like PEGs or additives.

It's important to emphasize that this research is still early-stage. An F1 of 0.45 means the model is far from perfectly pinpointing the exact crystallization recipe, but it is better than trial-and-error alone and demonstrates that sequence patterns can inform what conditions are likely to work. As a future direction, the authors suggested incorporating more variables (e.g. predicting optimal pH and temperature as continuous values, not just class labels) and using the approach to focus screening on a smaller set of candidate conditions. Nonetheless, this is a promising frontier: even a modest predictor that suggests, say, the top 10 most likely crystallization cocktails for a new protein (instead of blindly testing 1000) would be hugely valuable.

Chapter 2

Theory

2.1 Proteins

2.1.1 Protein Crystallization

2.2 Trees

Chapter 3

Data

3.1 The Protein database

The Protein Data Bank (PDB) is the central repository for experimentally determined three-dimensional structures of proteins, nucleic acids, and complexes. As of 2025, it contains over 220,000 entries, the majority of which are proteins solved via X-ray crystallography. Each entry includes not only the atomic coordinates of the protein structure but also extensive metadata about the experimental conditions used during crystallization. The following outlines the data acquisition, format, and structure to enable data preprocessing and analysis.

3.1.1 Acquisition

Using the RCSB Search API, all entries solved by X-ray diffraction (about 80%) were queried. Entries that did not contain any information about the crystallization at all or the entry was otherwise incomplete were not downloaded. A Python script then converted these identifiers into download links for the mmCIF files (.cif.gz), wrote them to text files, and split the list into several parts to enable parallel downloads via wget. The complete set of structures was downloaded in compressed mmCIF format and stored locally (after decompression to .cif where needed). Due to the large number of entries, the download process took approximately two days. However, the total download time naturally depends on the available internet bandwidth.

3.1.2 Format

The PDB provides structural data in two formats: PDB and CIF. The newer CIF format was chosen because, unlike the fixed-width PDB format, it does not impose size limitations on large structures. In addition, it can represent complex features such as branched carbohydrates and offers greater detail and flexibility than its predecessor. The data items are in the format of '_' + category name + '.' + attribute name. Data categories can be either saved in either key-value or in tabular format. These can be easily parsed into a column vector as is in CSV files. For example the crystallization conditions are saved in key-value format and look like the following for 3P4V:

```
_exptl_crystal_grow.crystal_id      1
_exptl_crystal_grow.method          'VAPOR DIFFUSION, SITTING DROP'
_exptl_crystal_grow.apparatus       None
_exptl_crystal_grow.atmosphere      None
_exptl_crystal_grow.pH              9.5
_exptl_crystal_grow.temp            298.0
_exptl_crystal_grow.pdbx_details    '3.2M (NH4)2SO4, 0.1M Glycine, ...'
_exptl_crystal_grow.time            None
```

A category is stored in tabular format when a token defines multiple values. In this case, loop_ is followed by rows of data-item names, with data values separated by whitespace. Notably, the protein's structural information i.e., the atom_sites category is represented in this format, as shown in the following example:

```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
ATOM   1    N  N   . VAL A 1 1   ? 6.204   16.869  4.854   1.00 49.05 ...
ATOM   2    C  CA  . VAL A 1 1   ? 6.913   17.759  4.607   1.00 43.14 ...
ATOM   3    C  C   . VAL A 1 1   ? 8.504   17.378  4.797   1.00 24.80 ...
ATOM   4    O  O   . VAL A 1 1   ? 8.805   17.011  5.943   1.00 37.68 ...
```

3.1.3 Structure

In structural biology, one rarely needs to work with all proteins deposited in the PDB. Consequently, the PDB is designed as an entry-oriented resource: users typically retrieve and parse the complete dataset for a single protein of interest. This design is well aligned with the standard workflow in structural biology, where researchers focus on a small number of proteins but require all available structural, experimental, and metadata associated with those entries.

In this project, however, the goal is fundamentally different: to identify patterns across many proteins, using all deposited entries in the database. Parsing a single CIF file takes ~100 ms using the fastest available library, Gemmi which amounts to approximately 9 hours for the entire PDB. Repeating such operations would be not feasible for data analysis where data typically has to be read multiple times across sessions.

To address this, the first step after downloading the database was to restructure it into a feature-based data model, in which information is grouped by category rather than by entry. Each category is stored as a separate Parquet file (also convertible to CSV), where each attribute becomes a column spanning all proteins. This enables highly targeted access: if an analysis requires only a single attribute, the corresponding column can be read directly without loading irrelevant data.

Although the initial restructuring (reading, grouping, and writing) takes ~11 hours, it drastically improves downstream performance. For example, extracting all crystallization-condition information from the entire database now takes ~10s, compared to the 9 hours required to parse every CIF individually.

Storage efficiency is also greatly improved. The full PDB occupies ~46 GB and allocates ~110 GB, making it impossible to load into memory. In contrast, the derived

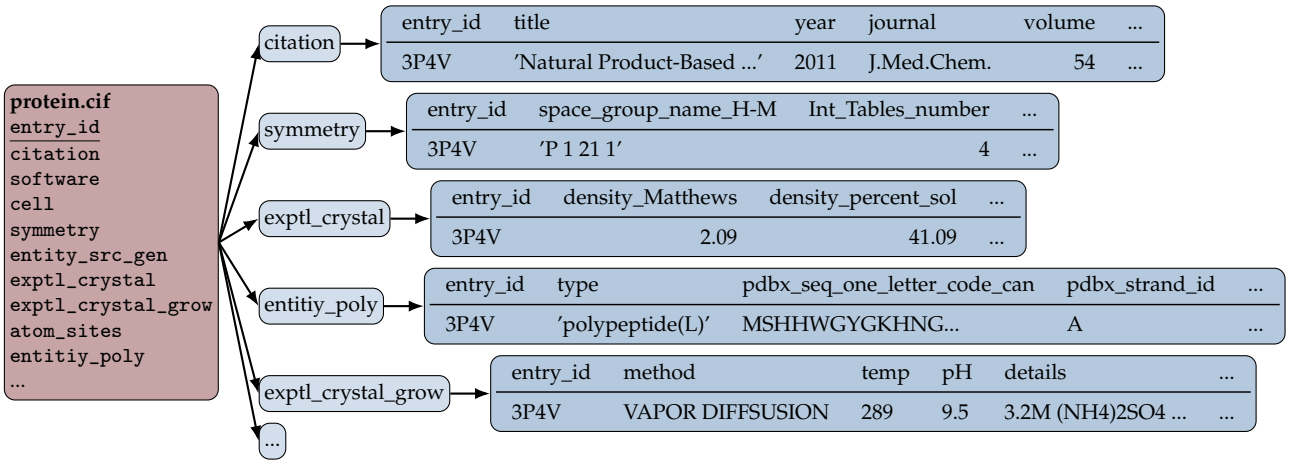


FIGURE 3.1: Visualization of the restructured database format derived from the `protein.cif` file. The `protein.cif` serves as the point of origin and contains all available information for a given protein. In the restructured schema, each CIF category is stored as a separate table whose columns correspond to the category attributes, while each row represents a protein entry. To enable relational operations, all category tables share a common `entry_id` column.

dataset containing only crystallization conditions is about 20 MB, small enough to be memory-resident and repeatedly queried without overhead.

This dramatic reduction in both time and storage requirements arises because most information in CIF files is irrelevant for crystallization-condition prediction. Metadata such as software provenance or publication information is not needed for this project and can be safely excluded from the feature-based representation.

However, to enable analyses that combine multiple categories, the restructured database must be joinable in a way that allows the original information to be reconstructed. This was accomplished using a star schema, in which each category table is indexed by the unique `entry_id` of the protein. This design ensures that information from different categories (such as crystallization conditions and sequence data) can be linked by performing joins on the shared entry identifier. The restructured database format is depicted in Figure 3.1.

3.1.4 Description

After filtering for majorly incomplete entries the total number of deposited proteins was 196743 as of 01.09.2025. Each of these entries has around 700 attributes across 68 categories. However, as mentioned before the minority of these attributes/categories is of importance when predicting crystallization conditions. More specifically, the important categories and their attributes are the following:

Input Categories

The `entity_poly` category of the mmCIF format describes each polymeric entity in a macromolecular structure. A polymer represents a sequence of linked monomers (e.g., amino acids or nucleotides), and this category provides essential metadata about its type, sequence, and representation in the structural model. `entity_id` Unique identifier linking the polymer to the corresponding entry in the `entity` category.

`type` Specifies the polymer type, such as polypeptide(L), DNA, RNA, or polysaccharide.
`nstd_linkage` Indicates whether the polymer contains non-standard chemical linkages (e.g., cross-links or modified connectivity).
`nstd_monomer` Flags the presence of non-standard or modified monomers within the sequence.
`pdbx_seq_one_letter_code` The polymer sequence in one-letter code, including symbols for modified residues.
`pdbx_seq_one_letter_code_can` Canonicalized one-letter sequence where modified residues are mapped to their closest standard equivalents.
`pdbx_strand_id` Lists the chain identifiers (e.g., A, B) representing this polymer in the structure.
`pdbx_target_identifier` Optional external identifier used mainly in structural genomics pipelines.

The `atom_site` category contains the atomic coordinates and related information that define the three-dimensional structure of the macromolecule. Each row corresponds to a single atom and records properties such as atom name, element, residue identifier, chain, Cartesian coordinates, occupancy, and atomic displacement parameters.

Label Categories

The category `exptl_crystal_grow` defines the crystallization conditions. It contains 17 attributes that describe the experimental setup and methodology used to grow the protein crystal. However, as seen in Table 3.1 for 13 of the attributes the majority of entries do not contain any information.

Attribute	Missing percentage
pdbx_details	0.04
method	10.78
temp	10.79
pH	22.30
pdbx_pH_range	75.24
temp_details	98.56
details, seeding	99.83
apparatus, atmosphere, method_ref, pressure, pressure_esd, seeding_ref, temp_esd, time	99.84

TABLE 3.1: The percentage of missing values for each attribute in the crystallization condition category. Only 5 attributes have sufficient data.

Thus, the attributes of relevance are the following:
`method`: Describes the crystallization technique employed, such as vapor diffusion, batch crystallization, or microbatch methods (in free text).
`pH`: Specifies the pH of the crystallization solution, which strongly influences protein stability and crystal formation (numerical).
`temp`: Records the temperature at which the crystallization experiment was performed, typically given in Kelvin or Celsius (in free text).

`pdbx_details`: Provides free-text experimental details, such as buffer components, precipitants, additives, or other conditions important for reproducing the crystallization setup (in free text).

Moreover, it might be of importance how well the protein has crystallized. Crystal quality measures are primarily provided in the `exptl_crystal` category, which describes the physical properties of the crystal, and in the `diffrn` and `reflns` categories, which contain diffraction statistics such as resolution, completeness, and R-factors that reflect the overall quality of the crystal.

3.2 Details Parsing

entry_id	pdbx_details
3P4H	Crystallization was carried out in sitting-drop vapor-diffusion setups with 1:1 mixtures of protein solution containing 0.7 mM Cko and 1.8 mM MnCl2 and reservoir solution containing 20% PEG 3350 and 0.2 M Na2HPO4 , pH 9.5, VAPOR DIFFUSION, SITTING DROP, temperature 298K
3P4V	3.2M (NH4)2SO4, 0.1M Glycine, pH 9.5, VAPOR DIFFUSION, HANGING DROP, temperature 289K
3P62	100mM sodium cacodylate, 100 mM sodium acetate, 16-18% isopropanol, pH 6.2, VAPOR DIFFUSION, SITTING DROP, temperature 293K
3PCA	pH 8.4
6LJR	PEG
3PF1	15-17% PEG 4K 0.2M KCl, protein dialysed in 10mM NaOAc 50mM NaCl, 10% glycerol 0.4%C8E4, pH 5.5, VAPOR DIFFUSION, HANGING DROP, temperature 295K

TABLE 3.2: Examples of free text descriptions of crystallization conditions given in the PDB. The examples show the high variety in possible expressions of crystallization conditions.

The examples in Table 3.2 illustrate the complexity involved in parsing protein crystallization conditions from free text into a uniform, machine-readable chemical cocktail representation. The first entry (3P4H) demonstrates that crystallization conditions are often embedded in long, narrative-style descriptions that mix experimental setup, protein concentration, co-factors, and reservoir composition in a single sentence. Extracting a structured cocktail from such text requires an algorithm that can reliably identify and segment chemically relevant entities (salts, buffers, precipitants, additives) and distinguish them from procedural information.

The second entry (3P4V) highlights the lack of uniformity in how units and concentrations are reported. Here, components are given in varying molar units, whereas other entries in the table use percentages. A robust parsing pipeline must therefore handle multiple concentration formats, convert them into a common representation. The third entry (3P62) further emphasizes semantic heterogeneity in naming: chemicals may be referred to by systematic names (e.g. “sodium cacodylate”, “sodium acetate”) rather than by their chemical formulas, in contrast to entries that use formula-based names (e.g. $MnCl_2$, Na_2HPO_4). Peat *et al.* (2005) illustrated

the difference in naming conventions by showing the 30 different ways ammonium sulfate is spelled in the dataset. Using string matching the updated dataset presented 141 different spelling attributed to ammonium sulfate. This necessitates normalization against a chemical dictionary or ontology to map synonymic names and formulas to a shared canonical identifier.

Entry 6LJR and 3PCA illustrates that some `pdbx_details` fields contain only minimal useful information. Such cases require the parser to handle incomplete cocktails and to distinguish between genuinely missing data and conditions that were simply not reported. The variability in reporting detail is substantial: while the average description length is 93 characters (SD = 72), the lengths range from as little as a single character to as much as 1758 characters. This large spread underscores the heterogeneity and inconsistency in the level of detail provided across entries.

Finally, the entry 3PF1 illustrates more subtle challenges, including the presence of concentration ranges, ambiguous abbreviations (PEG 4K instead of PEG 4000), and incomplete specifications (e.g. "%" without an explicit indication of whether it is w/v or v/v). Additionally, the same text string can interleave multiple components without clear delimiters, making tokenization and assignment of units to the correct solute non-trivial. These examples underscore that converting free-text crystallization descriptions into uniform chemical cocktails is not a simple extraction task, but a complex natural language processing and normalization problem that must account for heterogeneous syntax, inconsistent units, synonymic naming, incomplete information, and domain-specific ambiguities.

3.2.1 Pipeline

Lynch *et al*. (2020) constructed a multi-stage extraction pipeline designed to transform the unstructured crystallization metadata contained in the PDB into a consistent format. The goal of their proposed workflow is to impose a controlled vocabulary on the free text description and thereby enable large-scale analyses of crystallization conditions. The complete procedure consists of the four major steps 1. data acquisition, 2. details parsing and text normalization, 3. curating a compound dictionary to create a controlled vocabulary, and finally 4. the construction of the crystallization details dataset. In the following, the focus is on the actual parsing as their dataset was not used and the controlled vocabulary was heavily expanded upon using string matching algorithms instead of manual specification.

The address the variation in wording, punctuation, spelling and chemical names Lynch *et al*. (2020) designed a custom parsing function that performs multiple passes over the raw text. This function extracts: chemical component names, their reported concentrations (if present) and incubation temperatures as well as pH level. The parser explicitly handles inconsistent spacing and irregular phrasing that commonly appear in the deposited records. However, it does not handle punctuation or typographical errors. These were added to the pipeline to improve the parsing. To identify misspellings and variant forms of chemical names in the dataset, a fuzzy string-matching approach was employed. First, all chemical names occurring more than 200 times were treated as reliable entries and were collected into a reference set of *correct chemicals*. All remaining chemical names, which appeared less frequently, were considered potential misspellings or variants.

For each infrequent chemical name c_{var}, the algorithm computed its similarity to all names in the reference set using a normalized edit-distance metric. Specifically,

the similarity score was based on the Levenshtein ratio, which measures the minimum number of character insertions, deletions, or substitutions required to transform one string into another, normalized by the maximum string length. This score ranges from 0 (no similarity) to 100 (identical strings).

For every variant c_{var}, the algorithm selected the most similar reference chemical c_{ref} and evaluated whether their similarity exceeded a threshold of 90 %. If this condition was met, the two names were considered variants of the same underlying chemical, and a mapping $c_{\text{var}} \mapsto c_{\text{ref}}$ was created. This procedure ensured that common spelling errors, hyphenation differences, and minor typographical variations were systematically corrected and unified across the dataset.The string similarity between chemical names can be misleadingly high, which required manual inspection of the resulting mappings. For example, ammonia and ammonium share a similarity of approximately 80 %, yet they are chemically distinct and should not be merged. Likewise, even with a threshold of 90 %, names such as disodium malonate and sodium malonate or hexanediol and heptanediol were incorrectly matched, despite referring to different compounds. To avoid such false positives, mappings in which the difference between source and target consisted solely of common chemical modifiers (e.g., di-, tri-, mono-, hydrogen, etc.) or while chemical names ('sodium ammonium tartrate' -> 'sodium potassium tartrate') were removed, as these often indicate genuinely different forms rather than typographical variants. This resulted in around 25 % of the false chemicals being successfully mapped to a more common name.

However, some examples were too unique to be added to the parsing pipeline. An example is 8C9L for which the entered description is "'0.1MBis Tris Propane pH 6.50.02 MSodium potassium phosphate pH 7.520 % w/vPEG 335010% v/vEthylene glycol'". Here the problem is that after each numerical value a space is missing not being clear if a pH of 7.5 is meant. Implementing to many anomalies interfered with the parsing of the more typical descriptions.

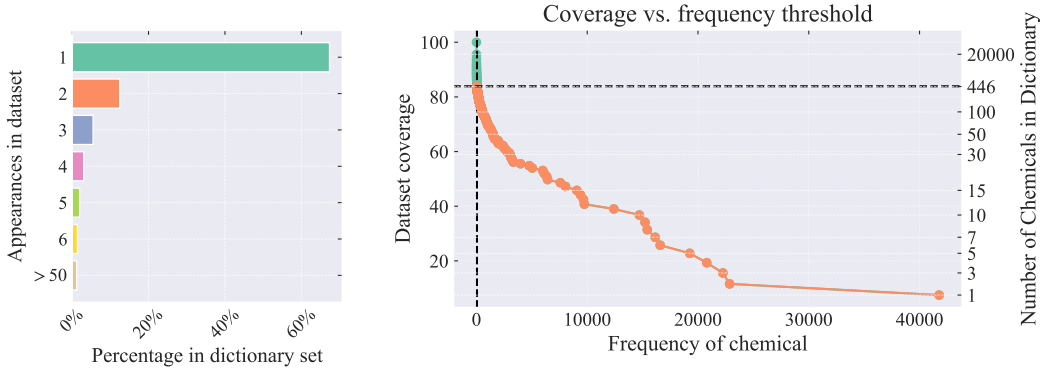
3.2.2 Parsing Quality

However, the parsing was not perfect in that it identified words as chemicals even though they were not. Originally, the parsing resulting in the majority of chemicals being parsed being present only once. This resulted in a total unique chemical count of around 30 000. This can be seen in Figure 3.2a.

Similar to the string matching approach described in subsection 3.2.1 the first step to reduce the number of unique chemicals was to check for common string patterns. This included for example the token "crystal tracking id <crystal id>" with part of the crystal id typically being picked up as the concentration. This removed around 350 of the faulty chemical names. Similarly, sometimes verbs were also identified as proteins. Using Spacy all chemicals that only occured once were screened. Verbs were removed after confirming that the model can successfully distinguish between them and chemical names. This got rid of around 2400 false chemicals.

After applying the string-matching approach, all chemicals occurring fewer than 50 times were removed from the parsed set. This threshold was chosen based on the coverage analysis shown in Figure 3.2b.The plot shows that the top 446 most occurring chemicals cover around 84 % of the dataset. When taking out every name that only occurred once and is thus very likely due to a parsing error the data coverage is above 98%. This increased the confidence that the chemical cocktails described in the free text were accurately captured by the parsed dictionaries.

PH and temperature values were sometimes provided in the free-text field but not in the corresponding numerical fields. After extracting and parsing pH and temperature from the free text, the proportion of missing values decreased to 9.2% for pH and 8.3% for temperature.



(A) Occurrence counts of unique chemical names extracted from free text. Although 25,000 unique names were parsed, 67% occur only once—likely due to typographical or parsing errors.

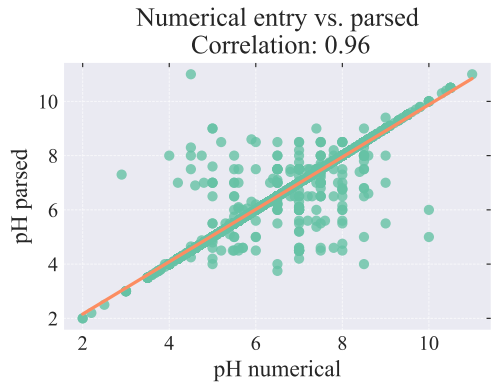
(B) Coverage of the dataset as a function of chemical name frequency. Applying a frequency threshold of 50—i.e., requiring a chemical to be mentioned at least 50 times before inclusion in the chemical dictionary—removes 99% of unique names, leaving 446 entries while still covering 84% of the dataset. From the dictionary set only 1.2% of unique chemical names occur more than 50 times in the dataset.

FIGURE 3.2: As shown in Figure 3.2a, most chemicals appear only a few times, and all parsed names represent only a small fraction of chemically meaningful compositions. This is also seen by the density of points and the almost logarithmic scaling of right y axis in Figure 3.2b.

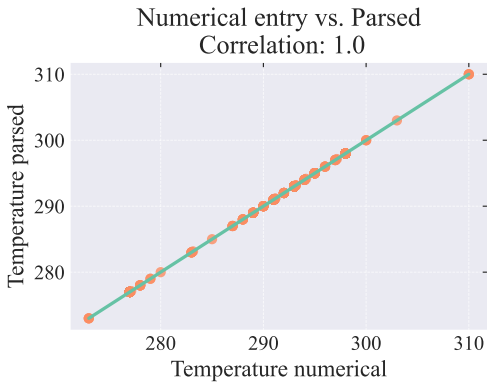
The parsing process aimed not only to extract chemical compounds from the free text but also to recover pH and temperature values. A straightforward way to assess this is to compare the parsed numerical values with those entered in the designated PDB fields, as shown in Figure 3.3. The parsed and entered values correlate strongly approximately 96% for pH and nearly perfectly for temperature as seen in Figure 3.3b. Inspection of the remaining pH (Figure 3.3a) mismatches shows that they primarily come from crystallographers reporting multiple pH values for different solutions, rather than from parsing errors. The few temperature mismatches likewise stem from genuine discrepancies between the entered and reported values. In such cases, the value provided in the numerical field was preferred.

Additionally, concentration values were extracted for all compounds whenever available. However, as shown in Figure 3.4a, the majority of entries associated with PEG did not specify a concentration unit. To integrate these data points into the model, the missing units needed to be imputed. This required analysing the relationship between molecular weight and the convention used for reporting concentrations. Figure 3.4b illustrates this relationship.

In biochemical practice, solids are typically reported as weight-per-volume (w/v) concentrations, whereas liquids are reported as volume-per-volume (v/v). Since the physical state of PEG at room temperature depends strongly on its molecular weight, this convention provides a principled basis for inferring missing units. As shown in Figure 3.9a, crystallization experiments are predominantly performed at room temperature. Under these conditions, PEGs with molecular weights below approximately 800 Da are generally liquid, while those above this threshold are typically



(A) Comparison of pH values parsed from free text with the single pH value recorded in the numerical field.

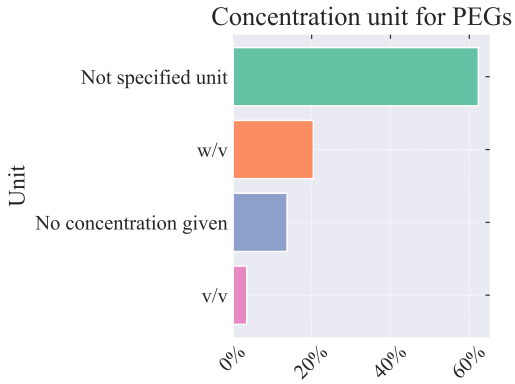


(B) Comparison temperature values parsed from free text with the numerical field entered temperature value.

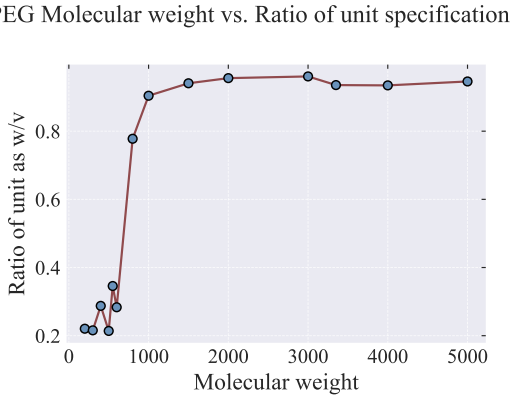
FIGURE 3.3: Multiple pH values reported in free text for different solutions leads to mismatches. In contrast the temperature field for crystallization produced almost no mismatches since only one value was given in free text.

solid. The distribution in Figure 3.4b clearly reflects this transition.

Based on this domain knowledge, missing concentration units were imputed using the following rule: whenever the unit was not specified and the PEG molecular weight exceeded 800 Da, the concentration was assumed to be given in weight-per-volume. This perturbation step ensured consistent and interpretable concentration features for downstream modelling.



(A)



(B)

FIGURE 3.4: Figure 3.4a shows the distribution of named units. For the majority of the parsed PEG concentration no unit was specified. Concentration can be given both in weight per volume and volume per volume. This naming is related to the molecular weight as seen in Figure 3.4b, which shows the relation between molecular weight and how often the unit was specified to be w/v in relation to all named unit appearances.

3.3 Data Analysis

3.3.1 Univariate Analysis

To gain an initial understanding of the dataset and identify characteristic patterns, a univariate analysis was conducted for both the input and the label. Section 3.3.1.1 describes the analysis of protein features which focuses on three major feature groups: sequence features, structural model information, and derived surface properties. Section 3.3.1.2 presents the analysis of the crystallization condition in particular pH and temperature distributions as well as patterns in the chemical cocktails derived from the parsed text. Examining the marginal distributions of these features provides insight into the typical ranges, dominant trends, and potential irregularities present in the data which help the multivariate analysis in Section 3.3.2.

3.3.1.1 Protein features

The input to the downstream model is intended to represent all information available for a protein prior to crystallization. In typical biochemical workflows, only the amino-acid sequence is known at this stage. However, AlphaFold allows highly accurate estimates of the three-dimensional conformation to be computed directly from the sequence. Consequently, the model input consists of (i) a sequence-based representation of the protein and (ii) a structural representation approximating the fold that crystallization aims to resolve.

Since protein crystals form through the periodic assembly of protein molecules into a lattice, intermolecular contacts are primarily mediated by exposed surface regions. Therefore, structural surface features can be explicitly encoded and incorporated as additional inputs. The following section focuses on sequence-, structure-, and surface-derived descriptors.

Sequence construction. Constructing a single sequence representation for a protein complex is not trivial, as crystallized proteins often consist of multiple chains. In homomeric assemblies all chains are identical and the protein is therefore associated with a single unique sequence. In contrast, heteromeric or oligomeric complexes contain multiple distinct chains, each potentially occurring multiple times. To generate a consistent sequence representation for the model, the following procedure was applied.

First, all polymer entities in the corresponding PDB entry were filtered to retain only protein chains, i.e. entities annotated as `polypeptide(L)` or `polypeptide(D)`. Nucleotide and ribonucleotide chains were discarded. Second, for each remaining entity, the amino-acid sequence was repeated according to its stoichiometric multiplicity, inferred from the number of strand identifiers associated with that entity. Finally, all repeated chain sequences were concatenated to form the complete protein sequence used as model input.

Formally, let a protein complex contain K protein chain entities indexed by $i \in \{1, \ldots, K\}$. For each entity i, let

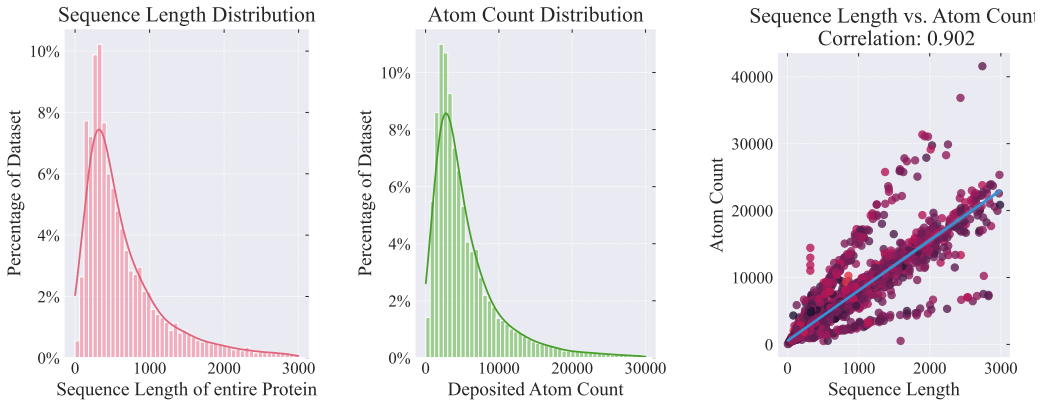
- s_i denote its canonical one-letter amino-acid sequence,
- A_i denote the set of strand identifiers associated with the entity,
- $c_i = |A_i|$ denote the copy number (stoichiometric multiplicity),
- $s_i^{c_i}$ denote the sequence s_i repeated c_i times.

The final sequence representation S for the protein complex is constructed as

$$S = s_1^{c_1} \,\|\, s_2^{c_2} \,\|\, \cdots \,\|\, s_K^{c_K}, \tag{3.1}$$

where "||" denotes concatenation.

Justification of the representation. Concatenation provides a simple yet expressive way of encoding the stoichiometry and composition of a protein complex in a format that is compatible with sequence-based neural models. Alternative representations, such as multisets of chain types or separate input channels per chain would require additional architectural modifications and would complicate downstream processing without providing clear advantages for this task. In contrast, concatenation preserves both chain identity and stoichiometric context while maintaining compatibility with transformer-based and recurrent sequence models.



(A) Sequence Length Distribution (B) Atom count Distribution (C) Atom count vs. Sequence Length

FIGURE 3.5: Visualization of the distribution of sequence lengths (Figure 3.5a) and deposited atom counts (Figure 3.5b). Most proteins have sequences between 50 and 400 residues, with models containing roughly 1,000–8,000 atoms. The relationship is largely linear as seen in Figure 3.5c, though missing or extra atoms in some structures cause deviations in slope.

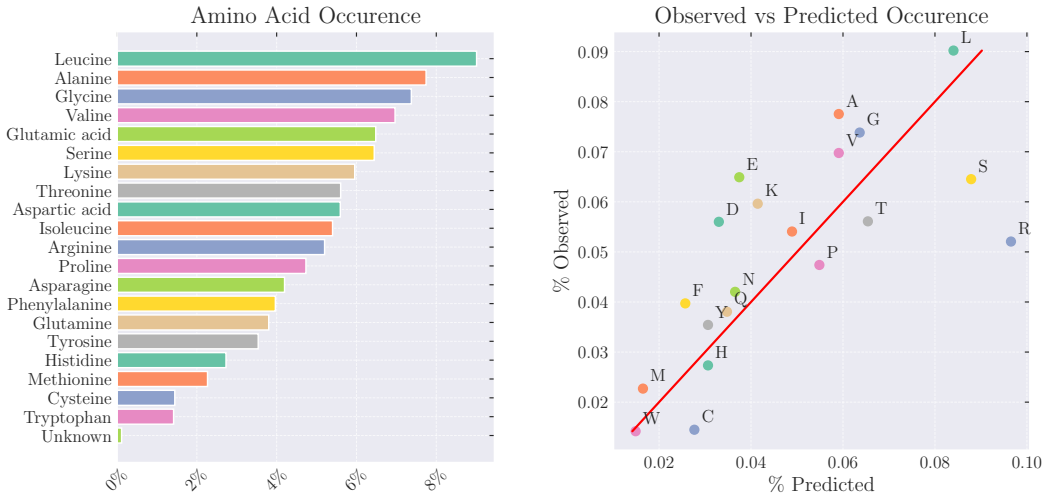
The distribution of protein sequence lengths is shown in Figure 3.5a. Most proteins in the dataset fall within a biologically plausible range of approximately 50 to 700 amino-acid residues. Although the mean sequence length is 784 residues, this value is heavily skewed by a small number of extreme outliers: the maximum observed sequence length is 87,120 residues, and the standard deviation is correspondingly large at 1,198 residues. More robust statistics provide a clearer picture of the dataset: the median sequence length is 474 residues and the mode is 306 residues, which more accurately reflect the typical protein size encountered in structural biology.

These considerations informed the choice of an upper cutoff during preprocessing. Proteins with a total sequence length greater than 4,000 residues were removed. This threshold excludes only exceptionally large and structurally atypical proteins

(e.g., very large multi-domain complexes or concatemeric constructs) while preserving ~99% of the dataset. The cutoff therefore improves model tractability without compromising representativeness.

A related distribution is shown in Figure 3.5b, which reports the number of atoms defined in the structural model for each protein. Atom counts are broadly distributed with a median of 3,987 atoms, a standard deviation of 16,257, and a maximum of 978,720 atoms. The same 99th-percentile–based filtering strategy was applied here. The 99th percentile corresponds to approximately 45,000 atoms, and this value was adopted as the upper inclusion threshold. Observations beyond this point correspond almost exclusively to very large complexes or structures containing auxiliary ligands, cofactors, or model artefacts that would introduce unnecessary computational overhead.

Each atom in the structural model is associated with a residue in the protein sequence. Consequently, one would expect a roughly linear relationship between sequence length and atom count, with the slope determined by the average number of atoms per residue. Deviations seen in Figure 3.5c can arise for several reasons, including incomplete structural models, missing density for flexible regions, artificially truncated or extended constructs, or variations in side-chain composition. Such deviations are visible in the data and provide an additional justification for discarding extreme outliers, as they likely represent structural artefacts rather than biologically meaningful proteins.



(A) Distribution of amino acids in the sequences. Residue frequencies vary substantially—Leucine is over four times more common than Tryptophan.

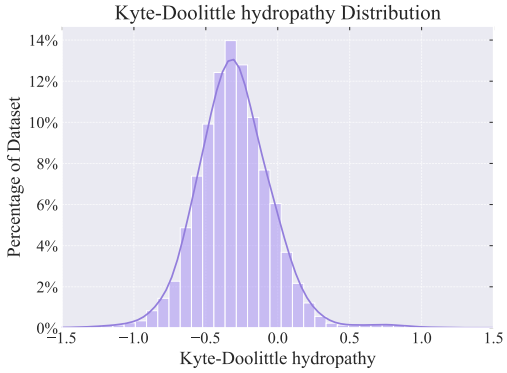
(B) Observed vs. predicted residue frequencies, illustrating the relationship between amino-acid occurrence and the codon patterns that encode them.

FIGURE 3.6: Figure 3.6a shows the distribution of amino acids across all sequences, while Figure 3.6b compares observed residue frequencies with those predicted from codon usage.

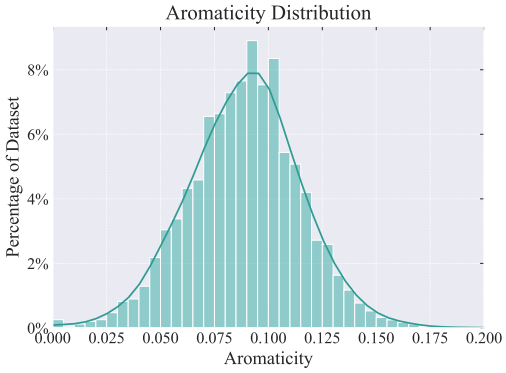
The amino-acid statistics presented in Figure 3.6 demonstrate that the dataset captures realistic and biologically meaningful properties of natural proteins. As shown in Figure 3.6a, the distribution of residues closely matches the characteristic composition observed across diverse proteomes: common hydrophobic residues such as leucine and alanine occur with high frequency, whereas aromatic or functionally specialized residues such as tryptophan and cysteine appear substantially

less often. These proportions are not arbitrary but reflect well-established biochemical constraints and evolutionary pressures, including metabolic cost, structural stability, and functional versatility. The fact that such canonical patterns emerge strongly from the dataset indicates that the underlying sample size is sufficiently large to average out oddities of individual proteins and instead reveal global, biologically grounded trends.

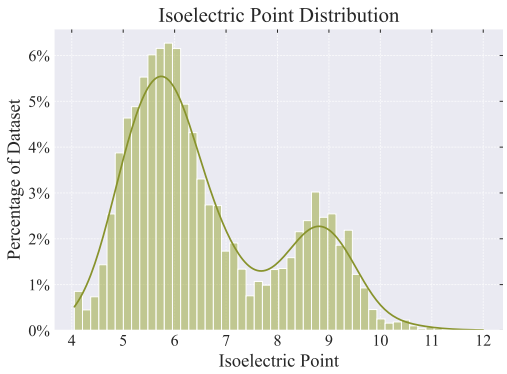
This representativeness is further supported by Figure 3.6b, which compares the observed residue frequencies with those predicted from codon usage statistics. The close correspondence between these distributions confirms that the dataset mirrors the translational and genomic biases embedded in natural organisms. Together, these analyses show that the dataset is both extensive and diverse enough to reflect the biochemical and evolutionary regularities of real protein sequences. This provides a strong foundation for downstream modelling, ensuring that any learned patterns are likely to be biologically plausible rather than artefacts of dataset construction.



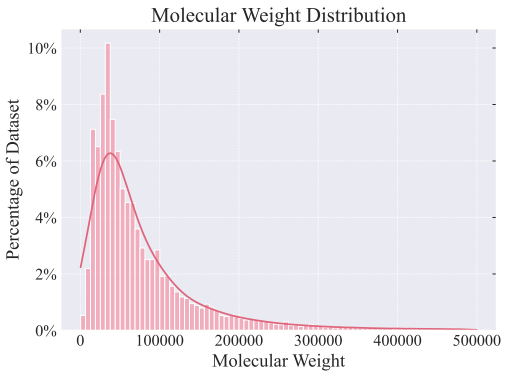
(A) Kyte-Doolittle hydropathy distribution



(B) Aromaticity distribution



(C) Isoelectric point distribution



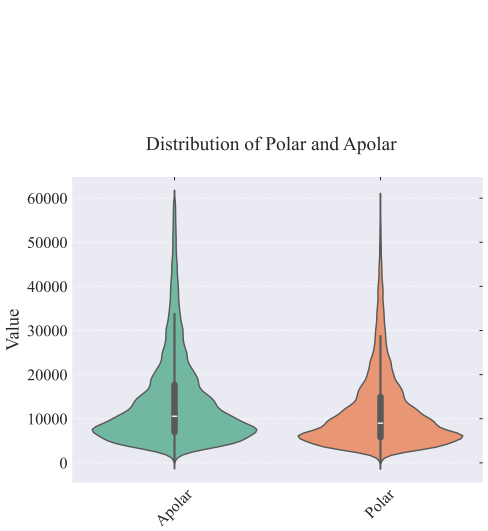
(D) Molecular weight distribution

FIGURE 3.7: Distributions of sequence-derived features across the dataset. Kyte–Doolittle hydropathy scores are roughly normally distributed with a slightly negative mean, while aromaticity values form a mildly positive normal distribution. The isoelectric point shows a bimodal pattern with peaks near pH 6 and pH 9. Molecular weights decrease sharply beyond 50,000 Da, mirroring trends in sequence length and atom count seen in Figure 3.5.

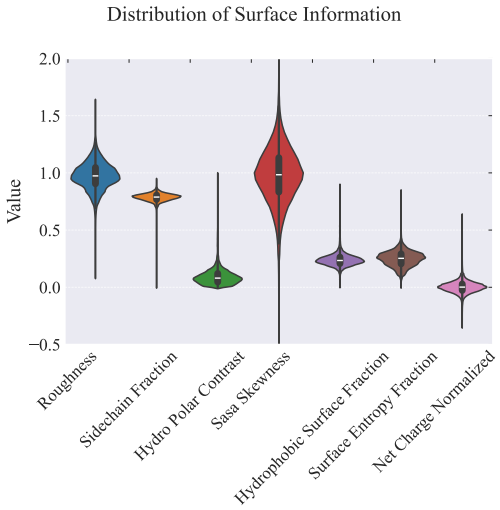
The sequence-derived features shown in Figure 3.7 provide a comprehensive overview of the biochemical properties represented within the dataset. These features—hydropathy, aromaticity, isoelectric point, and molecular weight—capture

fundamental aspects of protein composition that influence folding, stability, solubility, and ultimately the crystallization process. The Kyte–Doolittle hydropathy distribution (Figure 3.7a) is centered slightly below zero, indicating that most proteins exhibit a balanced mixture of hydrophobic and hydrophilic residues. Similarly, the aromaticity distribution in Figure 3.7b reflects a slight presence of aromatic residues involved in stacking interactions and structural stabilization.

The isoelectric point distribution (Figure 3.7c) displays a characteristic bimodal pattern, with peaks near pH 6 and pH 9. Finally, the molecular weight distribution (Figure 3.7d) shows a steep decline beyond 50,000 Da, consistent with the typical size range of monomeric proteins and with the sequence-length and atom-count distributions discussed earlier.



(A) Distribution of apolar vs. polar surface atoms.



(B) Comparison of different surface features.

FIGURE 3.8: Distributions of derived surface features. Apolar and polar surface atoms show approximately normal distributions with comparable overall exposure. Surface roughness, sidechain fraction, and SASA skewness are centered around 1, while the remaining surface descriptors cluster near zero or show slight positive shifts.

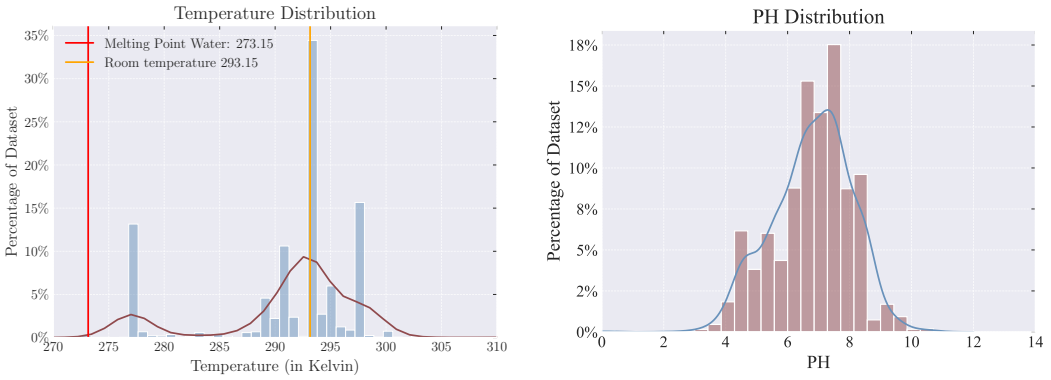
The surface-derived features presented in Figure 3.8 provide a detailed characterization of chemical features that influence intermolecular interactions during crystallization. As shown in Figure 3.8a, the distributions of apolar and polar surface atoms are approximately normal and of comparable magnitude, reflecting the typical balance between hydrophobic and hydrophilic exposure observed in soluble protein structures. This balance is essential for protein–protein contacts, as both hydrophobic patches and polar interaction sites contribute to lattice formation.

The broader set of surface descriptors illustrated in Figure 3.8b further demonstrates that the dataset captures realistic structural diversity. Features such as surface roughness, side-chain fraction, and SASA skewness cluster around values close to one. Other features, including geometric and electrostatic descriptors, are centered near zero or exhibit slight positive shifts, indicating mild asymmetries or directional preferences.

Together, these distributions confirm that the derived surface features span a biologically plausible range and reflect the heterogeneity of real protein surfaces. This diversity is crucial for downstream modelling, as crystallization propensity and conditions are likely o be influenced by surface geometry and chemistry.

3.3.1.2 Condition

After parsing it can be seen that the descriptions that do not contain any information about the chemical cocktail is ~3.5 %. They are very similar to entry 3PCA depicted in Table 3.2 in that they only define the pH or temperature. Additionally, concentrations were attributed to chemicals in ~93% of the times for the chemicals chosen to be in the dictionary set.



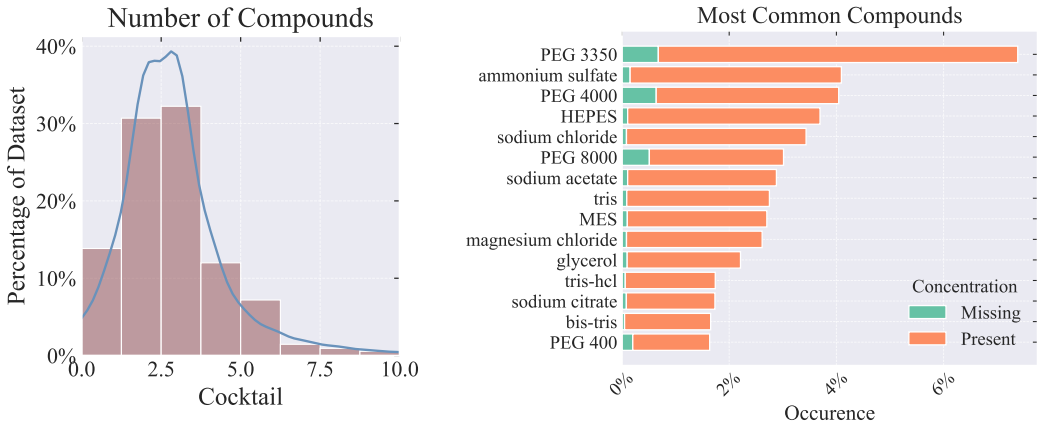
(A) (B)

FIGURE 3.9: Distribution of numerical fields in the condition category. Only a few distinct temperatures (3.9a) are reported, with 35% of samples specifying room temperature and most remaining values clustered around it; a secondary peak appears at 4 °C. In contrast, pH values follow an approximately normal distribution centered near pH~7 (3.9b).

The distributions shown in Figure 3.9 highlight the difference in how temperature and pH are reported across crystallization experiments. As illustrated in Figure 3.9a, the variance in temperature values is extremely low. More than one third of all entries explicitly specify room temperature, and the vast majority of the remaining values lie in a narrow window around it. A small secondary peak at 4,°C reflects experiments performed under cooling conditions, but otherwise temperature is essentially discretized into only a few commonly used settings. This has direct implications for modelling: because the data are dominated by room-temperature crystallizations, a model that predicts "room temperature" for all samples would achieve high accuracy despite providing almost no actionable information. The same holds for crystallization method employed. Figure A.1 in Appendix A depicts the lack of variation when it comes to crystallization methods. More than 95% of proteins in the PDB were crystallized using either sitting-drop or hanging-drop vapor diffusion. Both temperature and crystallization method predictions must therefore be interpreted with caution, as the task is intrinsically imbalanced and carries limited discriminative power.

In contrast, the pH distribution in Figure 3.9b spans a wide and continuous range and is approximately normal with a central tendency near pH,7. This variability reflects genuine experimental diversity, as pH strongly influences solubility, surface charge, and intermolecular interactions, and is intentionally explored across a broad range during crystallization trials. As a result, predicting pH constitutes a substantially more informative and challenging task. Meaningful pH predictions can provide valuable guidance to users by narrowing down the most promising acidity or alkalinity ranges for successful crystallization. Thus, while temperature prediction

is inherently limited by the structure of the data, the pH values offer rich signal that the model can exploit to generate practically useful recommendations.



(A) Number of compounds in a cocktail is on average 3 but sometimes specifies as much as 10 chemicals.

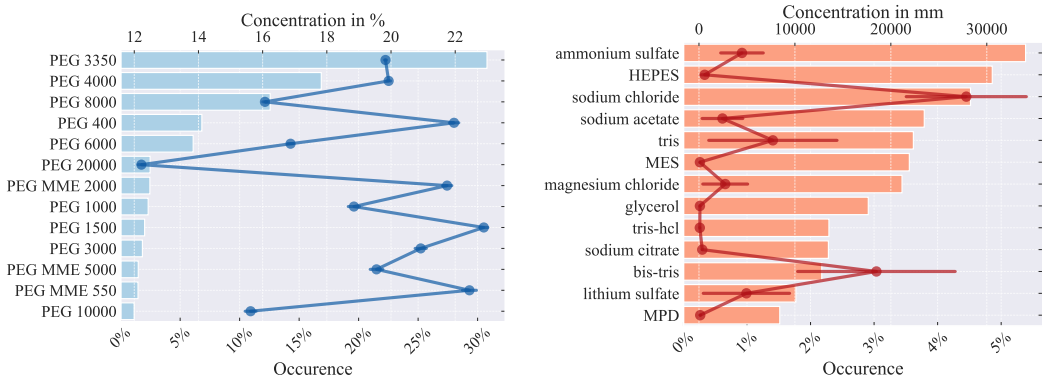
(B) The most common compounds across all cocktails feature multiple PEGs, buffers and salts. For the majority of the compounds a concentration value is given.

FIGURE 3.10: General Cocktail composition

The collection of plots in Figure 3.10–Figure 3.13 provides a detailed overview of the chemicals underlying the crystallization cocktails in the dataset. As shown in Figure 3.10a, a typical cocktail contains around three distinct chemical components, although more complex formulations with up to ten compounds do occur. This level of combinatorial variation is representative of common crystallization screening strategies, which balance chemical diversity with practical limitations on experimental complexity.

The distribution of individual reagents, shown in Figure 3.10b, reveals the dominance of well-established crystallization agents. Multiple PEGs, a range of buffering compounds, and various inorganic salts constitute the majority of entries. Importantly, for most commonly used chemicals, concentration values are available and sufficiently well populated to support reliable modeling. Notably, concentration values are more likely to be missing for PEGs in comparison to other additives. This distribution of most common compounds is important when analyzing the performance of the model. The goal is for the model to build an understanding of common chemicals while still being able to find patterns to determine when less common chemicals are preferred.

A more fine-grained view of compound usage is provided in Figure 3.11. The distribution of molecular weights for PEGs, shown in Figure 3.11a, exhibits a clear bias: the five most frequently used PEG variants account for approximately 60% of all PEG occurrences in the dataset. Despite this strong preference for specific molecular weights, the concentrations at which these PEGs are used vary substantially. Lighter PEGs such as PEG 400, 550, 1000, 1500, and 2000 are typically present at higher concentrations (above 21%), whereas heavier PEGs—including PEG 6000, 8000, 10 000, and 20 000—tend to be used at noticeably lower concentrations (below 17%). This inverse trend between molecular weight and concentration is well known in crystallization practice and reflects differences in their physicochemical effects on protein solubility and osmotic pressure. The dataset therefore captures a meaningful and biologically plausible relationship that the model must learn to reproduce: not only



(A) Most common PEGs and the average concentration value at which they occur

(B) Most common chemicals and the average concentration value at which they occur

FIGURE 3.11: General Cocktail composition

which PEG molecular weight is appropriate, but also the concentration at which it should be applied.

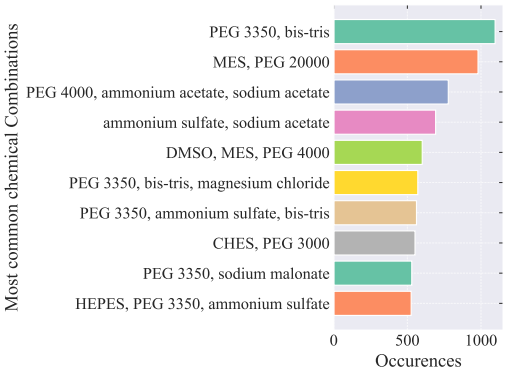
A similar level of variation is observed for non-PEG chemicals. However, two important differences make this category substantially more challenging from a modelling perspective. First, the distribution of individual reagents outside the PEG family is far more uniform. While the most commonly used PEG (PEG 3350) accounts for roughly 30% of all PEG entries, the most frequent non-PEG reagent—ammonium sulfate—appears in only about 6% of its respective category. Thus, predicting which buffer, salt, or additive should be included in a cocktail is intrinsically more difficult than predicting the appropriate PEG molecular weight, since no single reagent dominates the distribution.

Second, the concentration ranges for these non-PEG compounds are considerably broader. Bis–tris, for example, appears across a range from approximately 10 000 mM to 27 000 mM, reflecting the wide experimental latitude with which buffers, salts, and additives are explored in crystallization screens. This substantial variability introduces additional uncertainty: the model must not only identify the correct chemical component but also infer the appropriate concentration within a much wider and less predictable interval.

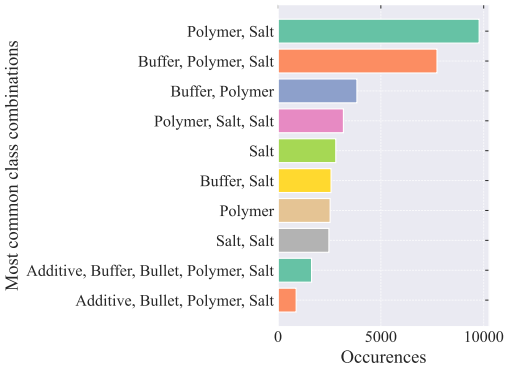
Together, these observations indicate that the prediction of PEG-related parameters, both in molecular weight and concentration, is comparatively biased and thus more learnable, whereas predicting the optimal combination of buffers, salts, and additives is a more complex task with greater combinatorial and quantitative uncertainty.

Beyond individual reagents, the dataset also captures the combinatorial structure of crystallization cocktails. Figure 3.12a highlights the most frequently occurring multi-compound combinations, illustrating characteristic reagent pairings and triplets—for example, PEGs combined with salts or buffers as co-precipitants. The two most common chemical combinations are PEG 3350, bis-tris and MES and PEG 20 000. This is also reflected when classifying the compounds into chemical classes as seen in Figure 3.12b. Here the most common combinations are one polymer i.e. some PEG and a salt as with the previously two mentioned cocktails. Generally polymers, buffers, and salts form the core set of recurring building blocks often used in combination.

Finally, Figure 3.13 provides a higher-order perspective by examining frequent 5-compound subsets centered around the most abundant reagents. For widely used



(A) Most frequent specific chemical combinations found in the crystallization conditions, showing the dominant pairs and triplets of reagents across the dataset.



(B) Most frequent combinations of chemical classes occurring together in crystallization setups, highlighting the typical pairing patterns between polymers, buffers, and salts.

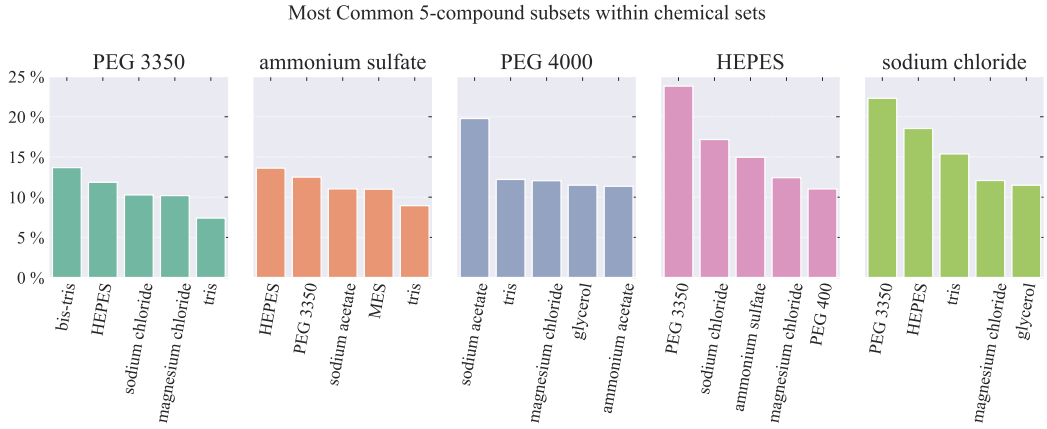


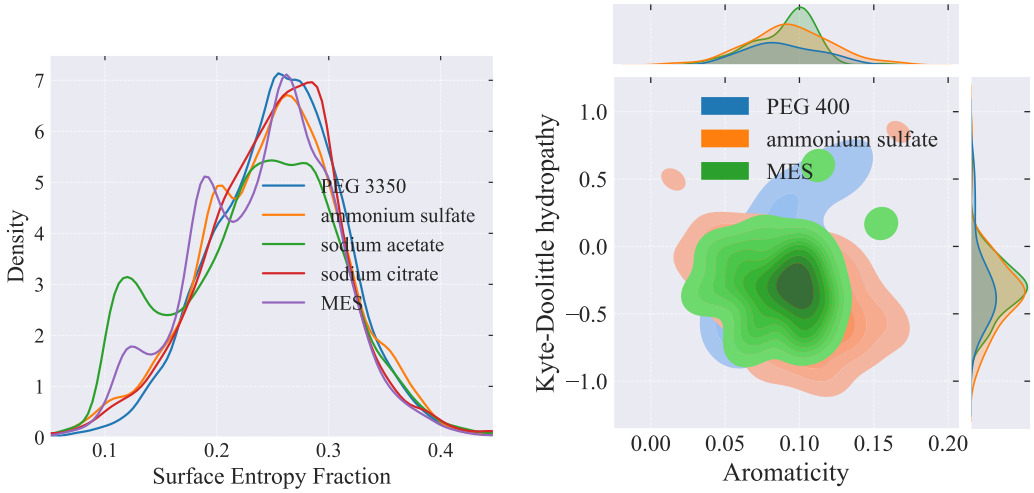
FIGURE 3.13: Most common 5-compound subsets within chemical sets. For several highly frequent chemicals—including PEG 3350, ammonium sulfate, PEG 4000, HEPES, and sodium chloride—the plot shows the compounds most often co-occurring with them in crystallization conditions, highlighting typical pairing patterns among buffers, salts, polymers, and additives.

components such as PEG 3350, ammonium sulfate, HEPES, sodium chloride, and PEG 4000, the plots reveal characteristic "neighborhoods" of co-occurring chemicals, reflecting well-established formulation strategies used to steer crystallization outcomes. The model be able to understand these co-occurrences and model them accurately. As seen in both Figure 3.12b and Figure 3.13 a chemical cocktail typically only has one PEG but might have multiple salts. This can be seen with sodium chloride which often is combined with magnesium chloride or ammonium sulfate and sodium acetate. It is also clear that these chemicals have a preference for certain PEGs.

Together, these analyses confirm that the chemical compositions within the dataset are not only diverse but also representative of real-world crystallization practice. The dataset preserves established chemical patterns while offering sufficient variation for meaningful learning, ensuring a robust foundation for predicting crystallization conditions and understanding reagent synergies. Moreover, some components of the condition set are less diverse than others and thus might be easier to predict than others.

3.3.2 Multivariate Analysis

The goal of this analysis was to examine whether particular chemicals tend to appear preferentially in specific ranges of protein sequence- and surface-derived features. For each chemical, I computed the empirical distributions of these features and compared them across compounds. The underlying hypothesis is that if a chemical consistently interacts with proteins exhibiting certain structural or physicochemical characteristics, these preferences should manifest as systematic shifts in the feature distributions.

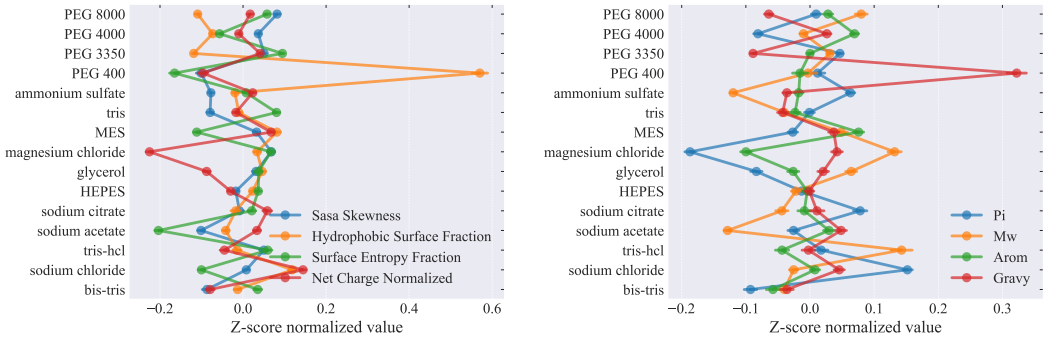


(A) (B)

FIGURE 3.14: The figure demonstrates that surface (Figure 3.14a) and sequence feature distributions encode predictive information regarding the presence of particular chemicals. Although the observed differences are subtle, distinct shifts in these distributions are associated with variations in the probability of specific chemicals occurring. Similar to how Figure 3.14b shows the 2D distributions for different chemicals, the model would combine multiple features to differentiate between the chemicals.

Across most surface and sequence features, the observed differences between chemicals were present but relatively subtle. For many features, the distributions of different chemicals largely overlapped and diverged only in localized regions (Figure 3.14a). This indicates that no single feature alone is strongly predictive. However, when examining two features jointly, the separability increases.
The bivariate distributions (Figure 3.14b) already show clearer clustering for certain compound classes, suggesting that the underlying relationships between chemicals and protein properties are inherently multivariate. In other words, although each single feature contains only weak signal, the combination of multiple features provides stronger discriminatory power.

From these univariate distributions, I next computed the median and its standard error for each feature–chemical pair and normalized them to z-scores. These aggregated statistics visualize how the typical feature value shifts depending on the chemical (Figure 3.15). Some features exhibit noticeably higher between-chemical variability than others, implying that they carry more predictive information. For example, the sequence-derived molecular weight differs substantially across chemicals, whereas other features remain close to the overall mean. Similarly, certain



(A) (B)

FIGURE 3.15: The point plot displays the z-score normalized median values of each feature, with standard errors, across different chemicals. They are derivative of distributions as seen in Figure 3.14a. Features with larger or more distinct deviations reflect stronger predictive influence on the occurrence of specific chemicals. Deviations for certain chemicals occur both in surface features (3.15a) as well as (3.15b.)

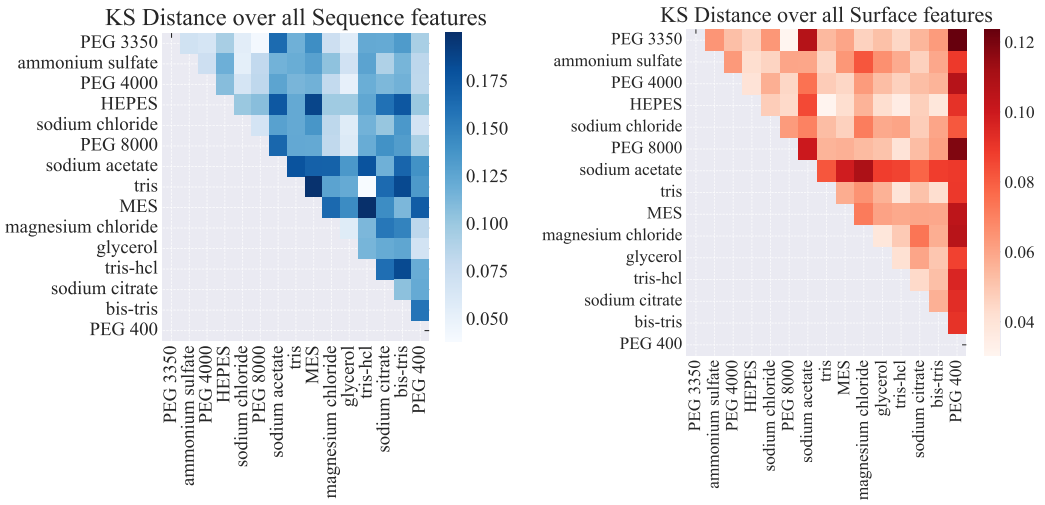
chemicals deviate strongly from the global average regardless of the feature category. PEG 400, in particular, shows large positive shifts for both the hydrophobic surface fraction and the Kyte–Doolittle hydropathy index.

This agreement is expected because the Kyte–Doolittle hydropathy is derived from the intrinsic hydrophobicity of amino acids in the primary sequence, whereas the hydrophobic surface fraction quantifies how much of the protein's solvent-exposed area is contributed by hydrophobic residues. Proteins that contain hydrophobic, high-GRAVY segments tend to bury some hydrophobic residues but still expose proportionally more non-polar surface patches than proteins with overall polar sequences. Thus, these two features capture related biochemical tendencies even though they were derived from different structural levels.

Finally, to quantify how different the full distributions are across chemicals, I computed the pairwise Kolmogorov–Smirnov (KS) distances for all surface and sequence features and averaged them (Figure 3.16). This analysis highlights which chemicals are globally most distinguishable from the others. PEG 400 consistently exhibits the highest KS distances across nearly all surface features, confirming that its usage correlates with a characteristic and unusual pattern of protein properties. Sodium acetate also shows elevated distances, suggesting it may likewise be more predictable than most other compounds.

The KS analysis further reveals chemical pairs that are either highly similar or strongly divergent. For instance, Tris and Tris-HCl are extremely similar, which is expected because they are chemically almost identical buffering agents; the only difference is protonation state, and both are typically used in similar pH ranges. In contrast, Tris differs sharply from MES or HEPES, both of which are sulfonic-acid–based buffering molecules with distinct pKa values and chemical behavior. These compounds stabilize proteins under different physicochemical conditions, which likely explains the characteristic shifts observed in their associated protein feature distributions.

Together, these results indicate that the problem of predicting chemicals from protein properties is not dominated by any single strongly discriminative feature but



(A) Differences of distributions for all sequence features between chemicals.

(B) Differences of distributions for all surface features between chemicals.

FIGURE 3.16: The KS distance between the distributions of one chemical to every other chemical for the 15 most common compounds. The KS distance was calculated for every feature independently. All distances were then averaged.

is instead inherently multivariate. Predictive information arises from subtle but consistent patterns across several sequence- and structure-derived features, supporting the use of machine-learning methods capable of jointly modeling these interactions.

3.4 CRIMS

In addition to the Protein Data Bank–derived dataset described in section 3.1, I incorporated a second, complementary data source: the Crystallization Information Management System (CRIMS) used at the European Molecular Biology Laboratory (EMBL). CRIMS is a laboratory information system designed to record, track, and analyze crystallization experiments performed on-site, particularly those carried out at synchrotron beamlines. Unlike large public repositories, CRIMS captures local experimental outcomes for a defined set of proteins studied within a specific laboratory environment. However, the data that is recorded at EMBL and other crystallization beamline providers using CRIMS is not public. Consequently, the number of proteins represented is smaller as it only covers proteins crystallized from the Itzen work group. However, the data quality and experimental granularity are substantially higher.

A key advantage of CRIMS is that it provides both positive and negative crystallization outcomes, which the PDB inherently lacks. While the PDB contains only successful crystallization conditions—i.e., those that yielded a structure—CRIMS logs the full experimental matrix. Crystallization trials are typically conducted in four 96-condition plates, and CRIMS records for each condition whether crystals were observed or not. Additionally, the conditions are not free text but rather in a uniform categorical/numerical format. This transforms the dataset into a true gold standard for supervised learning, as it provides clean explicit negative samples for model evaluation.

The crystallization drops are systematically monitored over time. Automated imaging systems photograph each drop at multiple intervals—after one hour, after one to three days, after one to four weeks, and up to approximately four months. These time-resolved images enable the use of machine-learning models to classify crystallization outcomes based on visual evidence of crystal formation. Because each condition is linked to the sequence and structural properties of the protein, the system allows us to determine how well a model can discriminate between crystallization conditions that lead to crystal growth and those that do not.

For the purposes of this thesis, CRIMS therefore serves as a critical validation dataset. It allows us to test whether a model trained on sequence- and surface-derived features can generalize beyond predicting which chemicals tend to appear in successful PDB crystallizations and instead evaluate the actual effectiveness of crystallization conditions for specific proteins. As a result, CRIMS provides a robust, experimentally grounded benchmark for assessing the model's ability to distinguish productive from non-productive crystallization conditions.

3.4.1 Acquisition

Unfortunately, CRIMS does not provide a bulk download option for crystallization experiment data. For this reason, a custom web scraper and parser were implemented. Because CRIMS is not a static website its content dynamically changes based on user interactions. The scraper had to simulate a real user session and systematically "check out" each crystallization plate. For every condition on each plate, both the chemical formulation and all available time-resolved images were extracted.

These images were required to obtain an objective quality assessment of each crystallization condition. To this end, a convolutional neural network developed by King *et al.* (2024) was applied to score the probability that a crystal is present in a given drop. Figure 3.17 shows representative examples of these crystallization drops alongside their predicted scores. The model exhibits high sensitivity, meaning false negatives are extremely rare. This behavior is desirable in crystallization screening, where failing to detect a true crystal is far more problematic than incorrectly flagging a non-crystal image as positive.

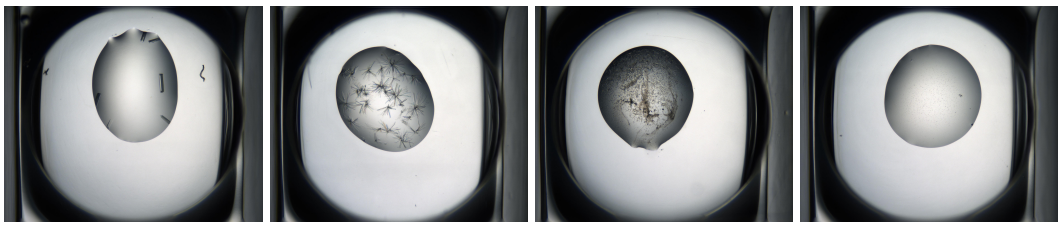


FIGURE 3.17: Scores: 0.99, 0.97, 0.90, 0.3

Chapter 4

Methods

4.1 Feature Engineering

To enable the model to capture the complex relationships between proteins and crystallization conditions, multiple types of feature representations were constructed. These include sequence-based embeddings generated with established protein language models, a geometric surface representation derived using a fast sampling framework, and a contrastive language embedding of the crystallization labels themselves. Together, these complementary feature types provide information from protein sequence, structure, and condition descriptors.

In the following sections, I introduce each of these feature representations and describe how they were generated and integrated into the overall modeling pipeline.

4.1.1 Sequence Embedding

Rives *et al*. (2021) present the Evolutionary Scale Model (ESM) a large-scale protein language modles that demonstrates how biological structure and function can be directly learned from the raw amino-acid sequences. The idea is that protein sequences contain sufficient statistical and evolutionary signal for a model, trained purely through self-supervised learning, to infer aspects of structure, function and mutational constraints.

The model is transformer based and follows the same architectural principles as Bidirectional Encoder Representations from Transformers (BERT) presented by "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (2019). Similarly, it has multiple layers of self-attention where each layer attends over all amino acids in the sequence, allowing the model to capture long-range dependencies that are crucial in protein folding. The input tokens are individual amino acids, and the model outputs a contextualized embedding for each residue as well as a sequence-level embedding obtained by averaging or pooling across positions. The model was trained using a masked language modeling objective on a dataset of ~250 million protein sequences sampled from UniRef. During training a fraction of amino acids is randomly masked. The goal of the model is to predict the correct amino acid at each masked position. By minimizing the prediction error the model aims to understand inherent structures in the protein space. Rives *et al*. (2021) shows that model implicitly leans evolutionary conservation, co-variation between residues, family-specific sequence constraints and patterns associated with secondary and tertiary structure. The model is available in several parameter scales, with the largest variant containing 650 million parameters. For this thesis, embeddings were generated for all proteins in the dataset using ESMs of different sizes.

Specifically, representations were computed using models with 35 million, 150 million, and 650 million parameters to assess the effect of model size on downstream performance.

4.1.2 Surface Embedding

4.1.3 Label Embedding

Chapter 5

Results

Chapter 6

Discussion

Appendix A

Data Appendix

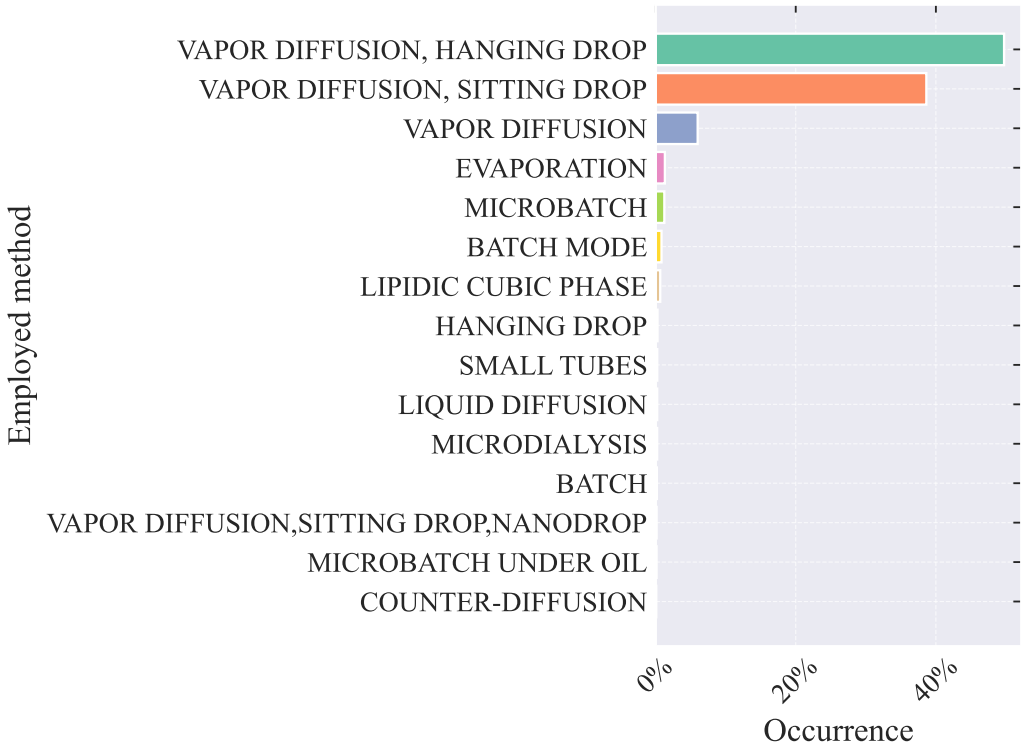


FIGURE A.1: A visualization of the most common methods employed to crystallize a protein. There is virtually no variation in method with more than 95% of entries being crystallized using a variation of vapor diffusion.

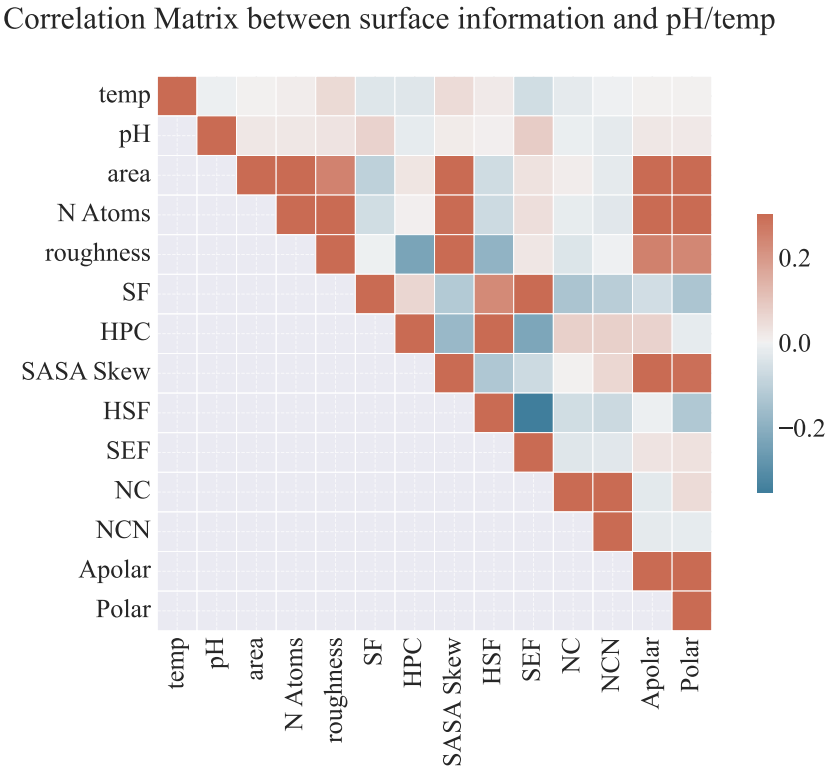
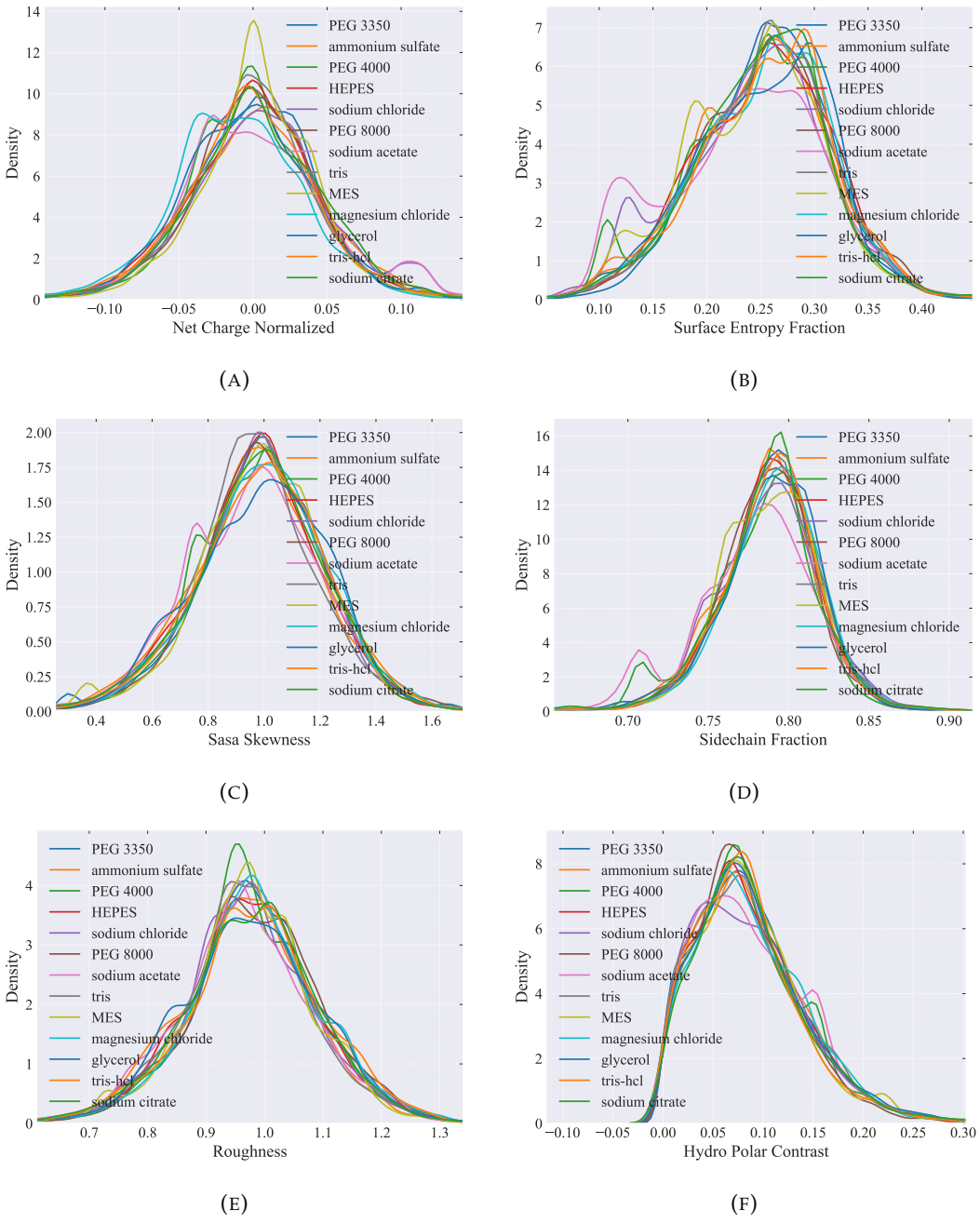


FIGURE A.2: Correlation matrix displaying correlation factors for the temperature and pH. Label values do not exhibit strong correlations directly with any of the features. However, there are some correlation between surface features as well.

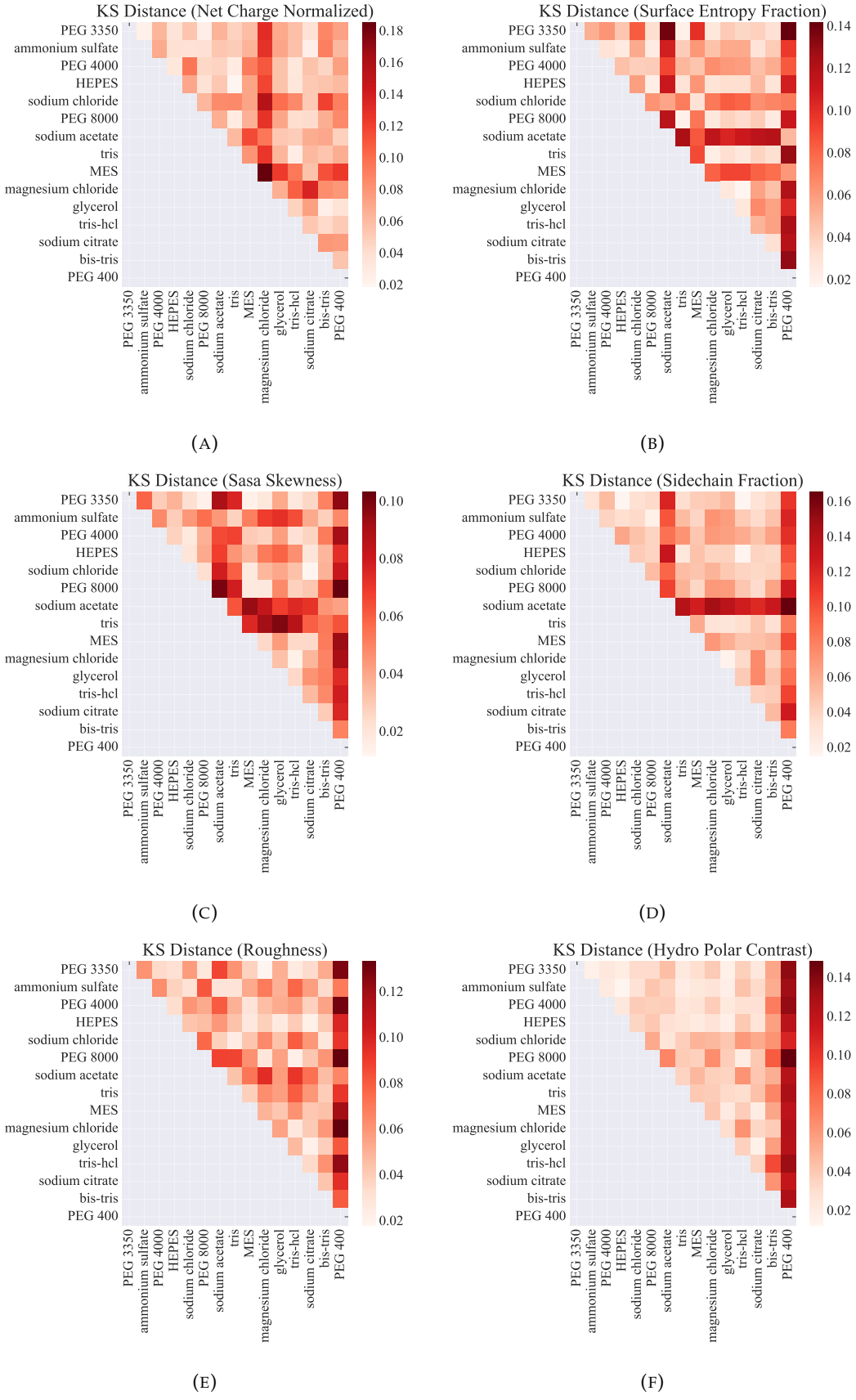


(A) (B)

(C) (D)

(E) (F)

FIGURE A.3: Distributions of different chemicals for different surface features



(A) (B) (C) (D) (E) (F)

FIGURE A.4: Distributions of different chemicals for different surface features

Bibliography

Ataka, M., & Asai, M. (1988). Systematic studies on the crystallization of lysozyme: Determination and use of phase diagrams. *Journal of Crystal Growth*, *90*(1), 86–93. https://doi.org/10.1016/0022-0248(88)90302-8

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [arXiv:1810.04805 [cs]]. (2019, May). https://doi.org/10.48550/arXiv.1810.04805

Chayen, N. E., & Saridakis, E. (2001). Is lysozyme really the ideal model protein? *Journal of Crystal Growth*, *232*(1), 262–264. https://doi.org/10.1016/S0022-0248(01)01203-9

Ghosh, R. (2023). Membrane-Based Micro-Volume Dialysis Method for Rapid and High-Throughput Protein Crystallization. *Processes*, *11*(7), 2148. https://doi.org/10.3390/pr11072148

Iwai, W., Yagi, D., Ishikawa, T., Ohnishi, Y., Tanaka, I., & Niimura, N. (2008). Crystallization and evaluation of hen egg-white lysozyme crystals for protein pH titration in the crystalline state. *Journal of Synchrotron Radiation*, *15*(Pt 3), 312–315. https://doi.org/10.1107/S0909049507059559

Jin, C., Shi, Z., Kang, C., Lin, K., & Zhang, H. (2022). TLCrys: Transfer Learning Based Method for Protein Crystallization Prediction. *International Journal of Molecular Sciences*, *23*(2), 972. https://doi.org/10.3390/ijms23020972

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

King, O. N. F., Levik, K. E., Sandy, J., & Basham, M. (2024). CHiMP: Deep-learning tools trained on protein crystallization micrographs to enable automation of experiments. *Acta Crystallographica Section D: Structural Biology*, *80*(Pt 10), 744–764. https://doi.org/10.1107/S2059798324009276

Kirkwood, J., Hargreaves, D., O'Keefe, S., & Wilson, J. (2015). Using isoelectric point to determine the pH for initial protein crystallization trials. *Bioinformatics*, *31*(9), 1444–1451. https://doi.org/10.1093/bioinformatics/btv011

Kurgan, L., Razib, A. A., Aghakhani, S., Dick, S., Mizianty, M., & Jahandideh, S. (2009). CRYSTALP2: Sequence-based protein crystallization propensity prediction. *BMC Structural Biology*, *9*(1), 50. https://doi.org/10.1186/1472-6807-9-50

Lee, H., Wu, Z. H., Corbi-Verge, C., Mok, M., Kang, S., Liao, S., Zhang, Z., & Garton, M. (2019). De novo crystallization condition prediction with deep learning [Accessed: 2025-07-31]. *Advances in Neural Information Processing Systems*, *32*. https://mlcb.github.io/mlcb2019_proceedings/papers/paper_3.pdf

Liao, K.-J., & Sun, Y.-J. (2025). Using AlphaFold and Symmetrical Docking to Predict Protein–Protein Interactions for Exploring Potential Crystallization Conditions. *Proteins*, *93*(10), 1747–1766. https://doi.org/10.1002/prot.26844

Lynch, M. L., Dudek, M. F., & Bowman, S. E. J. (2020). A Searchable Database of Crystallization Cocktails in the PDB: Analyzing the Chemical Condition Space. *Patterns*, *1*(4), 100024. https://doi.org/10.1016/j.patter.2020.100024

Mall, R., Kaushik, R., Martinez, Z. A., Thomson, M. W., & Castiglione, F. (2025). Benchmarking protein language models for protein crystallization. *Scientific Reports*, *15*(1), 2381. https://doi.org/10.1038/s41598-025-86519-5

Matinyan, S., Filipcik, P., & Abrahams, J. P. (2024). Deep learning applications in protein crystallography. *Acta Crystallographica. Section A, Foundations and Advances*, *80*(Pt 1), 1–17. https://doi.org/10.1107/S2053273323009300

McPherson, A., & Gavira, J. A. (2013). Introduction to protein crystallization. *Acta Crystallographica. Section F, Structural Biology Communications*, *70*(Pt 1), 2–20. https://doi.org/10.1107/S2053230X13033141

Mizianty, M. J., & Kurgan, L. (2011). Sequence-based prediction of protein crystallization, purification and production propensity. *Bioinformatics*, *27*(13), i24–i33.

Peat, T. S., Christopher, J. A., & Newman, J. (2005). Tapping the Protein Data Bank for crystallization information. *Acta Crystallographica. Section D, Biological Crystallography*, *61*(Pt 12), 1662–1669. https://doi.org/10.1107/S0907444905033202

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, *118*(15), e2016239118. https://doi.org/10.1073/pnas.2016239118

Terwilliger, T. C., Liebschner, D., Croll, T. I., Williams, C. J., McCoy, A. J., Poon, B. K., Afonine, P. V., Oeffner, R. D., Richardson, J. S., Read, R. J., & Adams, P. D. (2024). AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nature Methods*, *21*(1), 110–116. https://doi.org/10.1038/s41592-023-02087-4

Wilson, W. W., & DeLucas, L. J. (2014). Applications of the second virial coefficient: Protein crystallization and solubility. *Acta Crystallographica Section F: Structural Biology Communications*, *70*(5), 543–554. https://doi.org/10.1107/S2053230X1400867X

Zhang, X., Xu, Z., Zhou, J., Xing, X., & Li, L. (2022). Enhancement of Protein Crystallization Using Nano-Sized Metal–Organic Framework. *Crystals*, *12*(5), 578. https://doi.org/10.3390/cryst12050578