Adapting Contrastive Learning Methods for Few-Shot Imputation in High-Dimensional Data

Michael Henry Aldorf

MSc Applied Mathematics
TU Delft University Numerical Analysis Group
in collaboration with ASML D&E PWD

March 2025

I would like to express my thanks and eternal gratitude to my supervisor, Alexander Heinlein, for his guidance throughout this thesis. Without his help, this research would largely not have been possible.

I would like to further thank my company supervisor, Sjoerd Wiarda, for his continuous support and entertaining conversations, as well as Giulio Bottegal and Xingang Cao who, despite no formal obligations, kindly provided their time and mentorship—I learned a lot, thanks guys!

Abstract

High-dimensional data imputation is a critical challenge in semiconductor metrology, where secondary measurements are often purposely omitted to optimize throughput. This thesis examines the *Missing By Design* (MBD) framework—an industrially motivated scenario in which data are systematically uncollected to reduce measurement overhead—and investigates a range of imputation solutions tailored to the particular complexities of wafer reflectivity and overlay. After establishing the physical, rank-deficient nature of wafer metrology data through singular-value decompositions and principal component analyses, we explore several classes of methods: linear regressions and matrix-completion techniques for baseline comparisons; deep neural-network regression (MLP) to capture nonlinearities; a contrastive-learning adaptation of CLIP for pairwise matching of primary–secondary measurements; and novel *Bridge Models* that refine coarse CLIP estimates with localized residual translations. Additionally, we integrate overlay-based domain constraints into CLIP via *domain-guided neural network* regularization (DG), ensuring physically coherent tool-to-tool (T2T) predictions.

Comprehensive experiments on proprietary wafer datasets confirm that linear approaches, including regressions and matrix completion methods, despite capturing the low-rank structure of the data, under-perform in downstream overlay and T2T prediction due to subtle nonlinear relationships. Deep neural networks offer strong reconstruction accuracy, yet demand extensive hyperparameter tuning and deeper network structures than contrastive alternatives such as CLIP-like approaches, which yield architecturally efficient, instance-based retrievals, but can lack the precision needed for rigorous overlay alignment. DG regularization, as an extension of the CLIP framework, considerably enhances T2T consistency and reduces raw reconstruction error. Meanwhile, the Bridge Model combines a CLIP-derived coarse imputation with a smaller learnable residual map between encoder domains, bridging global pairwise alignment and localized corrections for improved reconstruction and downstream tasks. Overall, this thesis presents a flexible suite of tools that advance high-dimensional MBD imputation in wafer metrology, offering valuable insights and a robust methodological foundation for future industrial applications.

Contents

1 Introduction

1.1 Motivation

Modern data-driven systems and problems such as semi-conductor machine calibration frequently contend with incomplete datasets, whether arising from sensor failures, budget constraints, or proprietary restrictions. In such scenarios, missing values not only reduce the effective sample size available for training but also risk obscuring essential variability in the data. This absence of information can lead to the learning of ungeneralizable patterns, amplify spurious correlations, and diminish the robustness of predictive models. Many algorithms additionally require complete data matrices for training, further underscoring the importance of suitable imputation methods.

In industrial contexts where data acquisition can be costly or time-consuming, methods that reliably restore missing values can mitigate measurement overhead while preserving model performance. By learning how to extrapolate or estimate unmeasured features, imputation enables more efficient use of the available data, helping to avoid significant performance drops due to incomplete observations. Approaches to data imputation typically fall into either discriminative or generative paradigms. Discriminative methods, such as mean substitution, regression imputation, and k-nearest neighbors [1], often depend on observed similarities among the complete data points. These approaches, however, may struggle to represent the complexity of high-dimensional data, particularly if they rely on simplistic assumptions about the underlying distribution. Generative methods, exemplified by Generative Adversarial Networks (GANs) [2] or Variational Autoencoders (VAEs) [3], can effectively learn and model the data's statistical distributions to generate entirely new data from, but they can be computationally expensive and less easily interpretable.

A range of advanced statistical and machine learning techniques, including principal component analysis (PCA) and Bayesian marginalization, also address missing data. PCA reduces dimensionality by projecting observations onto a small set of principal components, though it may overlook nonlinear relationships and can be ill-suited to very large datasets. Bayesian marginalization introduces prior distributions for missing data and integrates over the resulting uncertainties, although it often requires approximate inference methods due to intractable integrals. Many standard machine learning approaches likewise encounter limitations in few-shot learning scenarios or when the interpretability of model predictions is a priority.

Recent progress in self-supervised representation learning has shown that contrastive approaches can capture intricate data patterns without requiring large amounts of labeled data. Contrastive Language-Image Pre-training (CLIP) [4] exemplifies such advances by aligning textual and visual embeddings within a shared representation space. Building on ConVIRT [5], which sought to remedy inefficiencies in medical image labeling, CLIP has proven flexible in domains such as breast cancer tissue classification [6] and lung nodule analysis [7]. These applications learn a contrastive space linking domain-specific imagery and text, thereby reducing reliance on labeled datasets. Nevertheless, there remains limited exploration of CLIP in data modalities beyond image and text, especially those involving continuous high-dimensional inputs that may be incomplete.

1.2 Objectives

The work presented here is conducted in collaboration with ASML, a major semiconductor equipment manufacturer with a line of state-of-the-art metrology systems, called YieldStar. These systems measure reflected light around the wafer and process the high-dimensional reflectivity data with neural networks to detect anomalies such as **overlay** [1]. These measurements occur as paired data points (denoted the primary and secondary measurement), which together are intended to calibrate machine-to-machine and sensor asymmetries. These pairs should be theoretically identical under certain contextual changes to the Yieldstar, though differ in practicality as a result of the systematic errors. Eliminating these systematic errors from the Yieldstar data is essential for accurate overlay prediction, which is further required for wafer alignment during the manufacturing phase and directly influences the nano-meter scale at which ASML can assure photolithographic printing accuracy. Being able to infer the systematic error by measuring primary point alone and imputing missing secondary measurements therefore promises a reduction in measurement overhead and a more efficient production workflow. This is not a problem unique to ASML's Yieldstar, and is a pervasive challenge in photolithography that yields a significant reward.

There is currently no published investigation or thorough guide of semiconductor metrology imputation that provides the setup necessary to extensively explore and rigorously compare a mixture of relevant methods. This thesis therefore aims to establish a comprehensive study on several *discriminative* imputation methods and their effectiveness in imputing missing data signals for wafer metrology. For each method we explore, the objective is to learn a meaningful mapping from primary to secondary measurements so that, given only the primary inputs, the model can predict and correct for systematic machine errors. Ideally, this should work in few-shot scenarios where only a handful of the total data collected is complete and utilized to impute the incomplete majority. The methods investigated include regularized regression imputation (Lasso, Ridge, Elastic Net), matrix-completion approaches such as Singular Value Thresholding (SVT) and PCA, and machine learning techniques like Multi-Layer Perceptrons (MLPs) as well as adaptations of the CLIP framework for exploring new contrastive learning approaches.

As a further objective through this comparison, the study seeks to extend research on the CLIP contrastive learning framework by assessing its ability to effectively handle the complexities of semiconductor wafer metrology and produce reliable imputations in high-dimensional, continuous numerical data settings beyond text-image pairs.

[1] Overlay is the alignment position of a lithography pattern relative to the underlying printed layers on a semiconductor wafer, and largely determines the minimum structure size that may be incorporated into semiconductor device designs [8]

Collectively, we can summarize the aims of our research with the following overarching research question:

> **RQ:** *How effectively do different discriminative imputation methods, including classical, statistical, and machine learning based methods, reconstruct missing secondary signals in wafer metrology data to reduce systematic machine errors and improve overlay prediction accuracy?*

To address this primary question, the following sub-questions guide the scope and depth of our investigation:

1. **SQ1:** *What is the relative performance of regularized linear models (Lasso, Ridge, Elastic Net) and matrix-completion methods (PCA, SVT) in imputing missing secondary data, and do they sufficiently capture systematic error signals?*

2. **SQ2:** *How robust are these imputation methods with respect to key metrics in semiconductor manufacturing, such as systematic error reconstruction, overlay prediction improvement, and tool-to-tool (T2T) consistency* [2] *?*

3. **SQ3:** *How can the CLIP framework be adapted to better suit continuous value imputation, and to what extent can it account for underlying physical data relationships between primary and secondary data pairs?*

4. **SQ4:** *What lessons can be distilled into practical guidelines for the semiconductor industry regarding the choice, design, and deployment of imputation strategies to minimize measurement overhead while preserving wafer metrology quality?*

[2] Tool-to-tool matching or consistency marks how closely aligned different Yieldstar tools are in inferring overlay for the same point on a wafer under equivalent machine settings. Correcting systematic errors *should* improve this metric.

1.3 Contributions

This paper introduces the first academic, extensive guide on the implementation and rigorous comparison of several discriminative data imputation techniques within the scope of semiconductor wafer metrology. Within this scope, this work proposes an extension of Rubin's [9] missing data framework, identifying Missing By Design (MBD) settings that are prevalent in industrial applications, yet underrepresented in machine learning literature. We further extend research on contrastive learning as an imputation mechanism in wafer metrology, specifically extending the CLIP model to new data modalities. It additionally contributes an adaptation of the CLIP framework that regularizes pairings with overlay inference error for *domain-guided* imputations (DGCLIP), extending CLIP to account for domain constraints characterized by the relationship between metrology data and overlay. Lastly, the study introduces a novel class of *Bridge models* that upgrades the CLIP framework by learning to translate unused residual information in pairings from the primary to the secondary domain for finer-grained [3] imputation. Empirical evaluations are conducted with both graphical analysis and metrics like residual PCA visualizations, systematic error reconstruction loss, overlay inference loss, and overlay tool-to-tool matching loss to gauge performance and reliability, accompanied by practical guidelines that consider data dimensionality, design and degree of missingness, and available computational resources.

[3] At test, CLIP is limited to imputing either an entire secondary point, or a weighted average of several, from the set of complete data pairs (the training set). Limited to this set of possible imputations, this constitutes a *coarse-grained* imputation lying in the interior of the span of the training set. Good imputations are hard to achieve for points that are not close-by or within this interior, and will likely require interpolation of several pairings, the quality of which is highly dependent on the spread of the training set.

2 Background and Literature Review

Herewith, we begin our investigation with a review of relevant work, such as missing data mechanisms, similarly explored imputation settings, and popular discriminative imputation methods, to establish the literature gaps that we aim to address with this thesis.

2.1 Classical Missing Data Mechanisms

Before we consider imputation methods for missing data values directly, we must first understand the mechanisms that lead to them. Rubin (1976) [10] categorized these mechanisms into three distinct types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). In this section, we outline each of these mechanisms with mathematical definitions and discussions of their implications on data imputation.

Consider an arbitrary dataset represented by a matrix $Y \in R^{n\times d}$, where $y_{i,j}$ denotes the entry in the i-th row and the j-th column for $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, d\}$. Let Y_{obs} be the set of all observed entries in Y and let Y_{mis} be the set of all missing entries. In other words, Y can be viewed as the combined dataset in a conceptual sense (encompassing both observed and missing values), whereas Y_{obs} and Y_{mis} denote disjoint portions of that data matrix. To formalize the missingness pattern, define a binary *missingness indicator* matrix $R \in \{0,1\}^{n\times d}$ with row and column indexes i and j, such that:

$$R_{ij} = \begin{cases} 1, & \text{if } Y_{ij} \text{ is observed,} \\ 0, & \text{if } Y_{ij} \text{ is missing.} \end{cases}$$

The distribution of R given the full data Y (both observed and unobserved) characterizes the *missing data mechanism.* In practice, understanding this mechanism can guide the choice of imputation or modeling strategies, and it may influence how reliably different methods can recover the unobserved entries in Y_{mis}. The complete-data likelihood function with missingness mechanism R is given by:

$$L(\theta, \phi \mid Y_{\text{obs}}, R) = P(Y_{\text{obs}}, R \mid \theta, \phi),$$

where θ is the imputation models parameters and ϕ is the set of parameters governing our missingness mechanism. We can further separate the likelihood function into:

$$L(\theta, \phi \mid Y_{\text{obs}}, R) = P(Y_{\text{obs}} \mid \theta) \times P(R \mid Y_{\text{obs}}, \theta, \phi).$$

Rubin defines a missingness mechanism as *ignorable* if it satisfies two conditions:

1. **Distinctness**: The parameters θ and ϕ are distinct, i.e., the parameter space can be partitioned such that inferences about θ do not provide information about ϕ, and vice versa.
2. **Missing at Random**: The missingness mechanism *at most* only depends on Y_{obs} and not on Y_{mis}.

Given the above conditions for a missingness mechanism, the likelihood simplifies and standard estimation procedures can be used without having to model the mechanism explicitly. The simplified observed-data likelihood would then be:

$$L(\theta \mid Y_{\text{obs}}) = P(Y_{\text{obs}} \mid \theta).$$

Consequently, the observed-data likelihood can be factored or integrated over the missing values without explicitly modeling the missingness process, preserving the consistency of standard estimators. Under an ignorable mechanism, it is therefore not necessary to model the details of *why* certain measurements are missing; instead, practitioners can focus solely on the relationship between observed variables to predict the missing samples. In particular, maximum likelihood and Bayesian methods applied solely to the observed data yield unbiased (or more precisely, consistent) parameter estimates, because the act of missingness does not distort the underlying data distribution in a way that depends on unobserved observations. By contrast, in non-ignorable settings, standard discriminative methods risk systematically biased or incomplete imputations if they fail to account for dependence between unobserved data and the missingness pattern.

In the semiconductor context, this distinction could be critical. For example, if one randomly omits half of the secondary metrology readings to reduce runtime, the resulting dataset may satisfy the 'missing at random' condition and allow straightforward imputation with methods like ridge regression or multilayer perceptrons. However, if measurements are missing in a way that correlates with the wafer's underlying properties (e.g., failed measurements on wafers prone to high overlay error), additional modeling of the missingness mechanism becomes essential. Ensuring or confirming the plausibility of an ignorable mechanism thus simplifies the imputation procedure and strengthens confidence in the fidelity of reconstructed wafer measurements.With this in mind, we will now introduce and analyze each of the classical missingness mechanisms.

Missing at Random (MAR) Also the second condition of ignorability, Rubin classifies a missingness mechanism as *Missing at Random* if the probability of an entry missing only depends on the observed data and not on the missing data itself:

$$P(R \mid Y_{\text{obs}}, Y_{\text{mis}}) = P(R \mid Y_{\text{obs}}).$$

Analyses under MAR require methods that model the missingness mechanism as conditioned on the observed data to obtain unbiased estimates. An example might occur if a preliminary inspection is used to decide whether a second, more detailed measurement is required. Here, the decision to skip the additional metrology step might be based on known, observable traits of the wafer (such as a recorded reflectivity range or known processing route). While missing data still reflect a pattern, that pattern is explainable by information in the observed wafer records rather than by unknown or hidden variables.

Given that the missingness mechanism depends directly on the observed data, imputation methods must utilize the observed data to model a predictive relationship. For this classification, techniques like multiple imputation or machine learning approaches would work well where missing values are replaced with plausible estimates that are inferred by a derived or learned distribution conditioned on Y_{obs}.

Missing Completely at Random (MCAR) Rubin further classifies a missingness mechanism as *Missing Completely at Random* if the probability of an entry missing is both independent of the observed data and the missing data itself:

$$P(R \mid Y_{\text{obs}}, Y_{\text{mis}}) = P(R).$$

In other words, under MCAR conditions, the 'missingness' does not depend on any data values observed. The observed data may consequently be considered a simple random sample of the complete data, and any analysis performed on MCAR data is then unbiased . For instance, suppose a measurement station is intermittently offline due to a random hardware glitch or network outage that has no correlation with the wafers being processed. In this case, each missing measurement arises purely at random, leaving no systematic differences between measured and unmeasured data points. Considering the mechanism is completely random, straightforward imputation methods like mean imputation can provide unbiased estimates.

Missing Not at Random (MNAR) Finally, Rubin classifies a missingness mechanism as *Missing Not at Random* if the probability of an entry missing depends on the missing data itself, even after conditioning on the observed data:

$$P(R \mid Y_{\text{obs}}, Y_{\text{mis}}) = P(R \mid Y_{\text{obs}}, Y_{\text{mis}}).$$

Analyzing MNAR data is a challenge because the missingness mechanism must be explicitly modeled and standard techniques may give biased results. As like a previous example, in metrology, one could imagine a tool failing specifically for wafers that have unusually high overlay errors or extreme reflectivity values that are not captured by the primary measurement. Since those wafers' secondary readings are systematically absent in a way that depends on the unobserved (but potentially problematic) measurements, the missingness cannot be accounted for simply by conditioning on observed attributes. This can introduce biases if the model treats missing data as though they were MCAR or MAR.

Addressing MNAR requires modeling the missingness mechanism explicitly, such as incorporating further external data or assumptions about the distribution of Y_{mis} (e.g., Bayesian priors). However, this has the potential to introduce bias. Approximation algorithms are necessary to evaluate the integrals as well, which can be an expensive computational step. However, in MNAR datasets, the missingness mechanism is *non-ignorable* according to the outlined conditions. Therefore, we must consider the full likelihood including Y_{mis}:

$$L(\theta, \phi \mid Y_{\text{obs}}, R) = \int P(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta) \times P(R \mid Y_{\text{obs}}, Y_{\text{mis}}, \phi) \, dY_{\text{mis}}.$$

This integral would likely lack a closed-form solution and require advanced computational methods for approximation, such as the Expectation-Maximization (EM) algorithm, Markov Chain Monte Carlo (MCMC) simulations, or variational inference techniques [11].

2.2 Missing by Design (MBD)

We have introduced the significance of ignorability and the classical mechanisms that shape our imputation approaches in semiconductor metrology. However, each of these classifications only govern missingness as a consequence of an uncontrolled mechanism, and not by design. On Rubins framework, Graham, Hofer, and MacKinnon (1996) later noted that the MCAR condition is rarely satisfied in practice "unless data are missing by design" [12]. The concept of *Missing by Design* (MBD) refers to instances where missingness is intentionally introduced, i.e., purposefully uncollected data, to optimize processes and reduce resource consumption. This classification is also known as *planned missingness* or *designed missingness* and has been extensively explored, with Pokropek (2011) formalizing it within the social sciences [13]. Essentially, MBD is a framework for balancing trade-offs between comprehensive data sampling and practicality. These practical constraints include sampling costs, time limitations, and burdens on participants, which are accounted for in MBD while yet maintaining the integrity of statistical analyses. To achieve this, procedures in the MBD framework use deliberately structured missingness according to a predefined scheme during data collection. This scheme should be designed in a way that appropriate data completion methods can then accurately account for the missingness mechanism. Common designs in social sciences include:

- **Multiform Design**: Where different subsets of variables are collected from different subsets of participants. This is equivalent to taking entirely primary or entirely secondary measurements for each sample wafer, or with each Yieldstar tool.
- **Two-Method Measurement Design**: This design has a subset of participants receive both a comprehensive measurement and a less resource-intensive one, while the rest receive only the latter. While neither primary or secondary measurements are less or more resource intensive than the other, this could be equivalent to taking both primary and secondary measurements for a subset of the measurement points along a wafer, followed by taking only primary measurements for the remainder. This method best aligns with the objective of our research, as we aim to predict secondary measurements for future primary measurements based on data relationships from a complete paired sample subset.
- **Three-Form (Matrix Sampling) Design**: Variables are divided into different forms, and each participant completes only a subset of these forms.

These designs reduce the data collection burden without significantly compromising the ability to estimate statistical models accurately [14] [15]. Now, consider again our arbitrary data matrix $\mathbf{Y}$ that can be separated into two disjoint sets of observed data $\mathbf{Y}_{\text{obs}}$ and missing data $\mathbf{Y}_{\text{mis}}$, where the missing data then corresponds to an MBD design. The missingness indicator matrix $\mathbf{R}$ is defined as:

$$R_{ij} = \begin{cases} 1, & \text{if } Y_{ij} \text{ is observed,} \\ 0, & \text{if } Y_{ij} \text{ is missing by design.} \end{cases}$$

In MBD, the probability of missingness is a deterministic function of the study design and does not depend on the data values themselves:

$$P(R_{ij} = 0 \mid \mathbf{Y}, \mathbf{D}) = f(\mathbf{D}),$$

where $\mathbf{D}$ represents the design parameters, and f is a deterministic function specifying which data points are missing. As the probability of missingness does not depend on the data values $\mathbf{Y}$, this mechanism is considered MCAR. However, unlike traditional MCAR scenarios where missingness is random, MBD involves deliberate missingness introduced through the design $\mathbf{D}$ and poses both challenges and opportunities for imputation. Specifically, because the missingness mechanism is known and controlled, imputation methods can leverage this information to create ignorable missingness and produce unbiased estimates.

Despite its established utility in social sciences and psychology, MBD has not been formally defined in the context of machine learning. Traditional missing data classifications—MCAR, MAR, and MNAR—focus on unintentional missingness arising from random or systematic factors. MBD, by contrast, arises from deliberate design choices aimed at balancing data collection efficiency with operational constraints.

This study builds upon the foundations of planned missing data designs by Rubin and Pokropek through expanding MBD for missing data imputation with machine learning methods in industrial applications. Specifically, in scenarios like advanced metrology systems, secondary measurements can be purposefully omitted during deployment to reduce measurement overhead. These omitted data points are then imputed using predictive models trained on complete data pairs. The intentionality of the missingness distinguishes MBD from other missing data mechanisms and aligns it with the principles of planned missingness. Hence, we define *Missing by Design* (MBD) as a missing data mechanism where data is intentionally omitted based on predefined operational or strategic criteria, independent of the data values:

$$P(R_{ij} = 0 \mid \mathbf{Y}, \mathbf{D}) = f(\mathbf{D}), \quad \forall i, j.$$

Unlike MCAR, MAR, or MNAR, MBD reflects deliberate decisions made to balance resource constraints, measurement costs, and analytical needs. This definition extends the principles of planned missing data designs to data imputation in machine learning contexts, particularly in industrial applications. To expand on this concept further, consider a high-dimensional continuous dataset where each data point consists of two components, $\mathbf{x}_1$ and $\mathbf{x}_2$. Due to operational constraints, we intentionally collect only $\mathbf{x}_1$ during deployment, resulting in missing data $\mathbf{x}_2$ that is *Missing by Design.* Our goal is to impute $\mathbf{x}_2$ using a model trained on complete data pairs $(\mathbf{x}_1, \mathbf{x}_2)$. The experimental design is such that our mechanism is independent of both the observed and unobserved data, and distinct from any parameters of our imputation model, and therefore, our missingness mechanism is ignorable. Assuming a joint distribution $P(\mathbf{x}_1, \mathbf{x}_2)$, we can model the conditional distribution $P(\mathbf{x}_2 \mid \mathbf{x}_1)$. Using this conditional distribution, the expected value of $\mathbf{x}_2$ given $\mathbf{x}_1$ can be computed as:

$$\hat{\mathbf{x}}_2 = E[\mathbf{x}_2 \mid \mathbf{x}_1] = \int \mathbf{x}_2 P(\mathbf{x}_2 \mid \mathbf{x}_1)\, d\mathbf{x}_2,$$

where, in practice, we estimate $P(\mathbf{x}_2 \mid \mathbf{x}_1)$ using methods like regression analysis, Bayesian approaches and machine learning models.

Although the MBD framework can be implemented as an ignorable missingness mechanism, certain design choices may violate Rubin's distinctness assumption if not carefully managed. An immediate example is if both the fraction of missing data (or another design parameter) *and* the imputation model's parameters are optimized simultaneously based on performance, then the missingness mechanism becomes entangled with the data model. In this joint optimization scenario, the way data are omitted depends on the model's outcomes, breaching the independence needed for an ignorable mechanism.

By contrast, if one parameter is held fixed and the other is optimized, the distinctness condition can be preserved. For instance, one could *fix* the fraction of missing data at 50% and optimize only the imputation model architecture; or, given a chosen model architecture, vary the fraction of missingness to assess cost-accuracy trade-offs. In both cases, the fraction remains an external design choice, preventing feedback loops in which the missingness depends on unobserved data or the model's predictive success.

Evidently, although the MBD approach can be crafted to maintain an ignorable mechanism, certain practical implementations may inadvertently violate Rubin's distinctness condition and thus become non-ignorable. Addressing these scenarios falls under our fourth research question, namely: *What lessons can be distilled into practical guidelines for the semiconductor industry regarding the choice, design, and deployment of imputation strategies to minimize measurement overhead while preserving wafer metrology quality?* Below are imputation implementation scenarios we've evaluated, that might be encountered in semiconductor metrology, on how deliberate design choices can tie the missingness pattern to the data or model parameters in ways that break the required independence:

Adaptive Skipping based on Preliminary Results Consider a metrology system that initially collects both primary and secondary measurements for each wafer, yet omits the secondary measurement if a preliminary reflectivity check indicates the wafer is likely within specifications. Although this setup appears MAR at first (since the decision depends on observed partial measurements), it can lean towards non-ignorability if those preliminary signals strongly correlate with the missing secondary values in ways not accounted for by the model. The fraction of missing data thus becomes a function of the wafer's underlying properties, entangling the mechanism with unobserved measurements.

Online Adjustment of Missingness Fraction In a high-throughput setting, the metrology tool might dynamically alter the fraction of omitted secondary measurements based on operational factors such as queue length or overall resource usage. If these operational factors themselves correlate with wafer characteristics (e.g., certain lots are processed during peak loads), the missingness fraction ceases to be purely exogenous. Unobserved wafer traits can indirectly dictate when secondary data are skipped, thereby rendering the mechanism partially non-ignorable.

Performance-Driven Omission A scenario arises when the imputation model is periodically retrained (fine-tuned), and its observed accuracy on validation data informs whether more or fewer secondary measurements should be collected. As the missingness fraction is directly linked to the model's performance, the parameters of the data model (θ) and the missingness design ($\mathbf{D}$) become co-optimized. This feedback loop breaks the distinctness assumption, since the design choice depends on the model's success in imputing missing values.

Cost-Driven Bidding or Pricing In some industrial workflows, wafers may bid for measurement resources according to an internal cost or priority function that depends on partially observed data. Wafers with higher priority gain both measurements, while lower-priority wafers skip the secondary reading. If the cost function substantially depends on unobserved wafer attributes or latent features, the missingness becomes correlated with precisely those missing values.

In each of these examples, the fraction or pattern of missing data is not set independently of the underlying data or model behavior. Instead, a feedback mechanism arises wherein the choice to omit certain measurements depends on observed or potentially unobserved properties of the wafer, or on imputation model performance. Such feedback invalidates the notion of an ignorable design, as the missingness mechanism now influences—and is influenced by—the same parameters the model is attempting to estimate. Consequently, more elaborate selection or pattern-mixture modeling would be required to avoid biased inferences under such non-ignorable MBD regimes.

In summary, Rubin's classification of missingness mechanisms provides a fundamental basis for understanding the complexity of data omission and the modeling required to address it. The MBD framework extends these ideas by deliberately introducing missingness according to a deterministic pattern that, when kept independent from model parameters, preserves an ignorable mechanism. This independence allows practitioners—such as semiconductor metrologists—to maintain unbiased inference, provided that the missingness design and imputation strategies are carefully separated. Having established the theoretical rationale for MBD and clarified the importance of avoiding certain design choices in semiconductor metrology data imputation, we now turn to the core imputation methods.

2.3 Data Imputation Techniques

This chapter presents a suite of data imputation approaches, ranging from classical linear models to more sophisticated methods that incorporate nonlinearities and regularization. We begin by examining established baseline classical techniques, such as regression-based methods and matrix completion, which often perform well under simpler assumptions (e.g., low-rank data or linear relationships) and are based on leveraging statistical properties of the observed data to estimate missing values. They are generally easy to implement and computationally efficient, making them popular choices for initial data preprocessing. However, these methods often rely on strong assumptions and may not capture complex relationships in the data, potentially leading to biased estimates and underestimation of variability. Subsequently, we move to advanced machine learning methods, including neural networks and contrastive-learning-based architectures, which can handle richer data structures and potentially offer greater flexibility in capturing complex wafer characteristics.

2.3.1 Regression Imputation

The most historical of the methods, linear regression analysis dates back to Newton in the 17th century, and has been continually expanded upon through modern times with as recently as Hoerl and Kennard introducing the ubiquitous ridge regression in 1970 [16], and Santosa and Symes' first use of the equally-infamous LASSO (L1) regression in 1986 [17]. Regression imputation is an application of linear regression analysis that explicitly models how missing values depend on other observed variables. Simply put, this method fits a regression model to the available data and then uses the fitted relationship to predict missing outcomes. In the simplest case, one assumes a linear link between a response variable Y (partially observed) and a set of fully observed predictors $X = \{X_1, X_2, \ldots, X_p\}$. The fitted regression function then provides imputations for Y_{mis}, preserving inter-variable correlations in ways that simpler approaches, such as mean substitution, cannot.

Despite the appeal of its straightforward application and deployability, regression imputation has limitations that become apparent when dealing with complex measurement processes like semiconductor wafer metrology. First, classical regression assumptions of linearity may underrepresent the intricate relationships in high-dimensional reflectivity or overlay data. Although the technique can be extended to include interaction terms or non-linear transformations, it still relies on a correctly specified functional form. If the true relationship is more complex or highly nonlinear, the imputed values may fail to capture essential patterns. Second, regression imputation underestimates variability by placing the imputed data directly on a fitted line or surface. This can inflate correlations between the imputed variable and the predictors, potentially skewing analyses that rely on covariance structures. Stochastic variants introduce random noise drawn from residuals [9] to better approximate true data variability, but these methods add further modeling assumptions.

In the broader missing data literature, regression imputation is typically discussed under the *Missing at Random* (MAR) framework, meaning the missingness depends only on observed data [18]. When these conditions are met and the model captures enough of the genuine data relationships, regression imputation often provides more accurate estimates than naive approaches like mean substitution. For example, in wafer metrology, if the missingness in a secondary measurement is driven by observed wafer properties (e.g., certain reflectivity ranges), then a well-specified regression linking these properties to the target measurement could yield meaningful imputations. However, if hidden factors play a major role in driving both the measurement outcome and the probability of missingness, any linear or even nonlinear regression model would risk yielding biased imputations.

Though widely used in social sciences and medical research, regression imputation remains relatively underexplored in the semiconductor manufacturing domain, where high-dimensionality and possible nonlinearity may depart from classical assumptions. As part of this thesis, we will benchmark the performance of regression-based approaches against more advanced or specialized methods to determine whether they can feasibly handle wafer metrology data. In doing so, we aim to assess both the potential strengths of regression imputation (e.g., interpretability, simplicity) and its vulnerabilities, thus identifying where it may fit in the broader suite of available imputation techniques.

2.3.2 k-Nearest Neighbors Imputation

Fix and Hodges introduced the k-nearest neighbors (k-NN) algorithm in 1951 as a new numerical approach for data interpolation and imputation [1]. In contrast to purely parametric approaches, k-Nearest Neighbors (k-NN) imputation directly exploits local data similarity when filling in missing values [19, 20]. Rather than relying on distributional assumptions or linearity, k-NN identifies the k most comparable observations in the dataset through numerical iteration—based on a chosen distance metric—and imputes a missing entry by averaging the known values of those nearest neighbors. This method can better capture complex or nonlinear relationships in data, a characteristic that proves especially relevant in contexts such as semiconductor metrology, where subtle differences in wafer reflectivity or overlay might result from localized process variations not easily approximated by a simple global model.

Despite this flexibility, several practical challenges arise when applying k-NN to industrial data. First, the choice of k and the distance metric can heavily influence the quality of imputations, as too few neighbors may skew results toward outliers, while too many neighbors risk diluting meaningful local structure. Second, high-dimensional scenarios—as in wafer metrology, where measurements can span multiple sensors or process steps—exacerbate the curse of dimensionality, diminishing the effectiveness of distance-based comparisons. Preprocessing steps such as Principal Component Analysis (PCA) can mitigate this issue but introduce further complexity and assumptions about data structure. Moreover, computing distances for each missing entry across large wafer datasets can become computationally expensive, making k-NN less practical for real-time or large-scale applications.

Like other non-mechanism-specific imputation strategies, k-NN requires assumptions akin to Missing Completely at Random (MCAR) or Missing at Random (MAR) for unbiased estimates [18]. Should wafer measurements be systematically missing due to, for example, machine failures triggered by unusual reflectivity ranges, straightforward local averaging cannot correct for that selection bias. Though, when missingness is reasonably ignorable and the data size remains tractable, k-NN might outperform simpler methods by preserving more of the variable interdependencies. This trait is particularly important to maintaining correlations that reflect critical physical relationships among wafer features. However, the literature remains limited on how effectively k-NN handles large-scale, high-dimensional metrology data, suggesting a need for empirical comparisons against advanced neural or matrix completion techniques. Due to time constraints, k-NN was not explored as part of this study. Including this in future research would be valuable as a comparison to the CLIP model, which employs this simpler method as part of its framework.

2.3.3 Principal Component Analysis (PCA) Imputation

Principal Component Analysis (PCA) was first introduced in 1901 by Karl Pearson, and has long since served as a quintessential statistical technique for dimensionality reduction, projecting high-dimensional data onto a smaller set of orthogonal components that capture the majority of variance [21]. In the context of missing data, Troyanskaya et al. (2001) formally extended PCA into an iterative imputation method by viewing the observed matrix as a low-rank structure, then estimating the unobserved entries as the best fit to that low-rank approximation [22]. This iterative process typically involves initializing missing values, computing a partial PCA decomposition based on available entries, and repeatedly refining both the imputed values and the principal components until convergence. By focusing on a reduced number of latent factors, PCA imputations are often more stable than purely local methods, particularly when variables exhibit strong correlations.

In principle, semiconductor metrology data can benefit from PCA-based imputation, since wafer measurements (e.g., reflectivity readings) frequently reveal dominant modes of variation that relate to physical properties or systematic machine effects. If these variations can be captured by a limited set of principal components, PCA offers a straightforward way to reconstruct missing entries without discarding partial observations. This characteristic proves attractive in high-dimensional scenarios, as PCA mitigates the curse of dimensionality by compressing the original feature space. Moreover, by approximating global variance structures, PCA may facilitate the alignment of wafer measurements across multiple sensors or process steps, thus preserving the key relationships that underpin industrial quality control.

However, several factors may limit PCA's suitability for metrology data imputation. First, the underlying assumption that wafer measurements lie near a linear subspace can be restrictive: if nonlinear dependencies are prevalent—perhaps due to complex wafer-layer interactions—an ordinary PCA model may fail to capture essential features. Although kernel or probabilistic extensions of PCA can provide more flexibility [23, 24], these variants raise computational costs and introduce additional modeling choices. Second, standard PCA methods typically assume an ignorable missingness mechanism (MCAR or MAR), yet wafer measurement data can be systematically omitted, for instance, if certain sensors tend to fail on more challenging samples (an MNAR scenario). In such cases, the imputed values might still be biased despite the low-rank assumptions. Lastly, implementing PCA-based algorithms on large-scale industrial data can be computationally intensive, as each iteration requires a partial Singular Value Decomposition (SVD) or EM-like updates.

From a research standpoint, while PCA is widely recognized in statistical literature, its direct application to deliberate or complex missingness in semiconductor metrology data remains understudied. It is not entirely clear whether the potential benefits—such as dimensionality reduction and capturing overall variance trends—offset the cost of possible mis-specifications when wafer measurements deviate from linear assumptions. A sub-goal of this thesis is therefore to ascertain whether PCA can effectively preserve the critical variance structure in wafer data, especially under purposeful or systematically introduced missingness, or whether more specialized methods are necessary to handle the demands of industrial wafer metrology.

2.3.4 Singular Value Thresholding (SVT) Imputation

A relatively recent development, the matrix completion imputation method Singular Value Thresholding (SVT) was formulated by Cai et al. (2010), which operates on a principle very similar to Principal Component Analysis (PCA)—namely, that a data matrix can be approximated by a low-rank representation [25]. However, instead of explicitly decomposing the data into principal components, SVT enforces low-rank structure through the nuclear norm (the sum of singular values) and iterative thresholding. Like PCA-based methods, it benefits from the observation that many real-world datasets, including those from semiconductor metrology, often exhibit a limited set of dominant modes or factors. By iteratively shrinking smaller singular values, SVT aims to recover a coherent, low-rank matrix from incomplete observations, thereby filling in missing entries.

In a wafer metrology setting, SVT can be appealing where systematic errors and prominent measurement variations align with a small number of latent factors. This approach, akin to PCA, offers a clean way to capture broad structural variability with relatively few parameters. Further expanding on their distinction, SVT frames matrix completion as a direct convex optimization under the nuclear norm, rather than extracting principal axes of variation through an eigen-decomposition or SVD at each iteration. While this formulation can be computationally efficient for certain types of sparse data, it still demands careful tuning of the threshold parameter τ and convergence criteria. If the metrology data deviate strongly from a low-rank profile or exhibit significant nonlinearity, SVT may omit subtler process variations, potentially biasing the imputed values.

Moreover, despite strong theoretical guarantees in some settings, the practical efficacy of SVT on large-scale or deliberately incomplete wafer data is not widely reported in existing literature. Evaluating SVT in comparison to both simpler (e.g., linear regression) and more advanced (e.g., neural network) imputation methods thus remains a key objective of this thesis, helping to clarify whether the nuclear norm minimization approach can effectively handle nuanced wafer-level phenomena and systematically designed missing entries.

2.3.5 Bayesian Marginalization

Bayesian marginalization, as the name suggests, owes itself to the historically and mathematically significant Bayes Theorem. First developed by Rubin (1976) in the same seminal paper that introduced missing data mechanisms, Bayesian marginalization distinguishes itself from other imputation strategies by formally treating missing entries as latent variables drawn from a prior distribution [10]. Instead of substituting or inferring point estimates for these missing values, this approach integrates over their full posterior distribution, thereby incorporating uncertainty into parameter estimation. Under a Bayesian view, both the model parameters and the missing data are random variables endowed with prior distributions, and the observed data are used to update these priors via Bayes' rule. Mathematically, one combines the likelihood of the complete dataset (observed as well as missing samples) with the priors to obtain a posterior distribution over parameters and unobserved values. Integrating out the missing data—often through Markov Chain Monte Carlo (MCMC) or variational inference—yields marginal posteriors for the parameters and posterior predictive distributions for the missing observations [26, 27].

For industrial contexts like semiconductor metrology, this method holds conceptual appeal: a Bayesian framework can seamlessly incorporate domain-specific knowledge (via priors) about wafer behavior or reflectivity measurements. Such domain knowledge might reduce reliance on purely data-driven assumptions, potentially mitigating biases when the observed data alone do not capture all relevant variability. Moreover, marginalizing over missing entries explicitly acknowledges that multiple plausible values could exist for the omitted measurements, an important acknowledgment if the missing data mechanism is uncertain or has some unmodeled complexity. The result is an imputation that not only provides point estimates but also quantifies uncertainty, which can be invaluable for downstream decisions where confidence bounds are necessary.

Nevertheless, Bayesian marginalization can become computationally burdensome for large or high-dimensional wafer datasets. MCMC-based sampling, in particular, may converge slowly if the posterior is complex or if many parameters are involved. Variational inference can reduce these computational demands but introduces additional approximation steps, and care must be taken to ensure accurate inference in light of potential constraints such as planned missingness. In addition, practical application in wafer metrology likely requires careful tuning of priors and nuanced modeling choices to avoid misspecification. While there is a substantial body of Bayesian literature on missing data for smaller, lower-dimensional problems, how best to adapt these methods to high-dimensional, systematically missing wafer measurements remains an open question. Although Bayesian Optimization could provide a cornerstone method for metrology imputation, time constraints prevented its implementation in this thesis. By examining Bayesian marginalization alongside more traditional or machine-learning-based techniques, future research may seek to establish whether the theoretical benefits of full posterior integration can be realized in the industrial setting of semiconductor manufacturing.

2.3.6 Multi-Layer Perceptrons (MLPs) for Imputation

A psychologist by trade, Rosenblatt (1958) first theorized the perceptron model in pioneering work for machine learning, laying the theoretical foundation for the framework that has and continues to make waves in modern times, known as deep learning [28]. As a single layer model, the perceptron could not solve linearly separable systems; though, Rumelhart, Hinton and Williams (1986) expanded upon this foundational model by introducing the Multi-Layer Perceptron (MLP) [29]. These models exhibit extraordinary ability to estimate nonlinear relationships in data, and have attracted notable interest for imputation tasks because they can, in principle, learn arbitrarily complex mappings from observed to missing features, provided the training data adequately capture the relevant relationships [30]. In a typical MLP-based imputation framework, the observed features ($\mathbf{X}_{\text{obs}}$) feed into an input layer, which passes information through one or more hidden layers. These hidden layers use nonlinear activation functions (e.g., ReLU, tanh) to detect patterns that may not be apparent through linear transformations alone. The final output layer generates predictions $\hat{\mathbf{X}}_{\text{mis}}$ for the missing entries, and training proceeds by minimizing a loss function—often the Mean Squared Error (MSE)—with respect to observed ground truth. Through iterative gradient-based updates (e.g., stochastic gradient descent or Adam [31]), the network refines its weights to reconstruct missing values more accurately.

Unlike classical models that assume linearity or rely on fixed bases such as principal components, MLPs approximate functions through layered compositions of nonlinear transformations, making them well-suited to heterogeneous or high-dimensional data. Studies comparing MLP-based imputation to classical techniques (e.g., linear regression, k-Nearest Neighbors, or PCA-based methods) often report lower reconstruction errors, especially in domains with complex feature interactions. For instance, [30] evaluated MLP-based imputation on medical records with high missingness rates and found improved predictive power in downstream analyses relative to simpler methods. Similarly, [32] demonstrated that MLPs can capture higher-order interactions across multiple features, effectively reducing imputation bias compared to ordinary least squares regressions.

While MLPs excel in handling tabular data by providing global, end-to-end mappings, certain metrology tasks might benefit from convolution-based architectures (CNNs) if measurement arrays exhibit structured spatial or spectral patterns [33]. In wafer metrology, spatial correlations or repeated measurement grids could align with convolutional kernels, potentially leading to superior local feature extraction. However, adapting CNNs requires that the data be arranged in a way that leverages local neighborhoods or hierarchical patterns—an arrangement not always guaranteed in high-dimensional reflectivity measurements. Given that our specific dataset lacks these qualities, we have chosen not to pursue further research on the subject.

Despite their versatility, neural-network imputation approaches raise several practical questions within semiconductor manufacturing. High model complexity could lead to overfitting if data are sparse or systematically missing, necessitating careful initialization and regularization of the network. Additionally, compared to the classical techniques, the number of model parameters that have to be tuned is substantial and will require significantly more time and computational resources to optimize over. At the scale of metrology datasets, even batch loading the data might not be sufficient to reduce memory usage during training and inference. Unless high-capacity GPUs are available, like the Nvidia A100, the dataset would have to be compressed (e.g., with PCA) to more manageable dimensions, which might bottleneck performance.

Additionally, the black-box nature of MLPs poses interpretability challenges. Process engineers and customers might seek a clear rationale for any deviation in wafer measurements—particularly if large financial or safety consequences hinge on the accuracy of imputed data. Furthermore, from the perspective of inter-variable relationships, while MLPs can approximate a broader range of functional forms than purely linear approaches, if correlation structures are extremely high or if certain wafer measurements in metrology are dominantly governed by physical constraints, purely data-driven approaches might overlook domain knowledge that could further refine imputation outcomes.

Lastly, the *few-shot* nature of some production scenarios can further complicate the use of MLPs. In a Missing by Design (MBD) setting, wafer data may be partially withheld to reduce measurement overhead, so the training set might underrepresent the richer variability encountered during actual device production. If the missingness is extensive and the real wafer space spans more variations than the training subset, e.g., due to subtle differences in layer composition, optical properties, or equipment calibration, the observed portion may not thoroughly span the manifold of wafer measurements. Under such conditions, the model could overfit to the subset of samples that happen to be observed, leading to poor generalization when predicting unseen or rarely encountered wafer configurations. This limitation motivates an examination of newer machine learning frameworks that rely on instance- or pair-level comparisons rather than an all-encompassing regression function. Such methods can potentially avoid the risk of half-baked global mappings by leveraging local similarity structures in the data. In the subsequent sections, this thesis will compare MLP performance with both classical imputation and emerging representation learning strategies, namely *contrastive learning*, evaluating their efficacy in large-scale, MBD wafer metrology datasets.

2.3.7 Contrastive Representation Learning with CLIP

Representation learning focuses on discovering meaningful data embeddings or features directly from raw inputs, reducing the reliance on manual feature engineering [34]. Although deep learning has advanced representation learning by training neural networks on large labeled datasets, many real-world domains—particularly industrial settings—either lack abundant labeled samples or encounter systematically missing measurements. These challenges have shifted attention to *self-supervised* approaches, which learn robust representations from unlabeled data through auxiliary pretext tasks [35]. Among these methods, *contrastive learning* has emerged as a powerful technique that aligns positive (similar) pairs while separating negative (dissimilar) pairs in an embedding space, thereby capturing crucial data structure [36].

In contrastive learning, each data instance is projected into a latent space by an encoder, and a *contrastive loss* function guides the encoder to construct the latent space in a way that brings embeddings of related samples closer while pushing apart unrelated ones [37, 38]. This principle has proved remarkably effective for large-scale image classification [39, 40], text-based tasks [41], and cross-modal modeling [4, 42]. By treating pairs of observations—whether they be two augmented (slightly altered) views of the same image or two different modalities of the same object—as positives, the model learns latent features that capture invariances or underlying relationships. Even without explicit labels, these learned embeddings often transfer well to downstream tasks, improving performance when labeled data are scarce.

Contrastive learning has proven effective for representation learning in diverse domains, often outperforming classical supervised or unsupervised techniques under limited labeling. Early successes in computer vision, such as SimCLR [39] and MoCo [40], illustrated how self-supervised objectives can rival fully supervised training by contrasting augmented image pairs within large unlabeled datasets. This paradigm also extends to natural language processing, where models like SimCSE [41] produce robust sentence embeddings by contrasting different noise-augmented versions of the same text. Researchers have similarly applied contrastive frameworks to multimodal tasks, aligning disparate data modalities—most notably in image-text pairs—with models like CLIP [4] and ALIGN [42], enabling zero-shot classification in new domains.In domains where labeled data are scarce, such as medical imaging, contrastive methods can align data modalities—e.g., radiology images and textual reports—to learn embeddings that enhance tasks like disease classification or anomaly detection [5].

Beyond image and text, contrastive learning techniques have gained traction in audio processing [37, 43], where the model maximizes agreement between temporally adjacent audio segments while discriminating unrelated samples. This approach, often referred to as Contrastive Predictive Coding (CPC), helps uncover latent features that improve performance on downstream tasks like speech recognition. Graph representation learning has likewise benefited from contrastive objectives [44, 45] by encouraging nodes that share structural properties to cluster in the embedding space, leading to more informative node or subgraph embeddings.

Although each of these applications underscores the versatility of contrastive learning, relatively few studies have investigated whether the same principles translate effectively to industrial contexts with designed missingness, such as in semiconductor metrology. The success in other data-rich fields suggests that if contrastive learning can be adapted to handle paired yet incomplete measurements, it might also mitigate the shortcomings of MLPs in few-shot scenarios.

Extending contrastive principles to data imputation requires leveraging a model's ability to align different views or subsets of data into a consistent latent space. For primary and secondary measurements in semiconductor metrology, by training on samples where both measurements are present—and treating them as positive pairs in a contrastive sense—an adapted model learns embeddings that reflect the intrinsic relationships between these two measurement types. When one measurement is missing, the model can then infer its likely embedding position (and thus the missing values) by comparing against samples that occupy similar regions in the latent space.

This approach addresses limitations in classical imputation strategies. Traditional regression-based methods can overfit or fail to capture rich nonlinear relationships among wafer measurements, while matrix completion often assumes a relatively global low-rank structure that may not hold in every localized region of the wafer data. Moreover, deep learning approaches like MLPs might lack guiding principles that incorporate prior information in the construction of their latent spaces. Contrastive learning, by contrast, aligns each observed primary-secondary pair through a specially constructed latent space without imposing strictly linear or low-rank constraints, thus potentially accommodating more fine-grained variation. Moreover, as shown in domains ranging from image-text alignment radford2021learning to audio oord2018representation, contrastive models can remain resilient even if some modalities or views are missing, by focusing on embedding consistency rather than strict functional mapping.

Despite these potential advantages, the literature offers scant direct evidence on whether contrastive methods specifically improve imputation in designed missingness contexts, such as those encountered in wafer metrology. As a result, this thesis aims to bridge the gap by adapting a CLIP-like model to learn aligned embeddings for primary and secondary wafer measurements. If contrastive training proves robust against deliberate data omissions and nonlinear measurement phenomena, it may yield more accurate and less biased imputations than classical regression, matrix-completion frameworks, and MLPs, especially in few-shot, high-dimensional scenarios where all those may struggle. Through experimental evaluations on semiconductor metrology datasets, we seek to clarify whether this novel adaptation of contrastive learning can fulfill the dual goals of preserving local structure in the data and delivering reliable imputation for downstream tasks like overlay prediction.

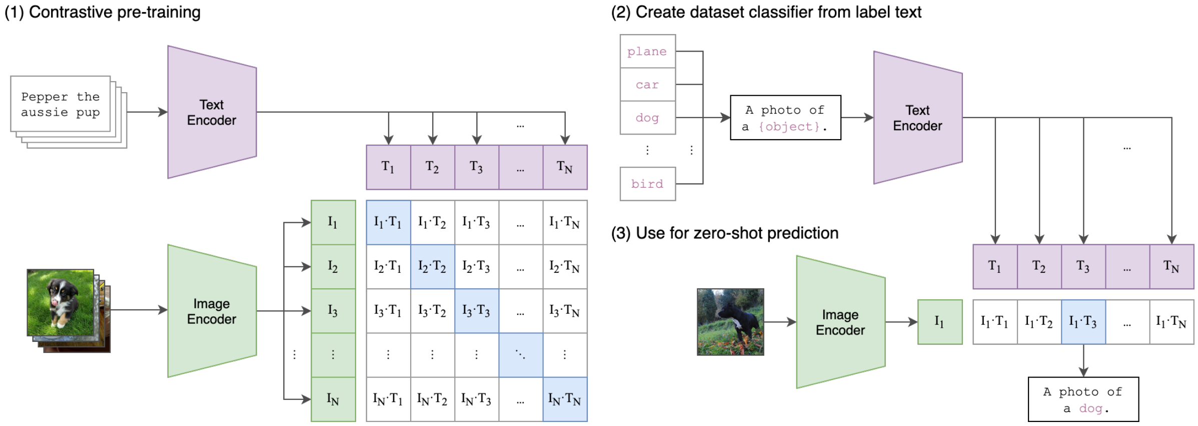


Figure 1: An architectural overview of the CLIP model [4]

Among the contrastive learning frameworks, the Contrastive Language-Image Pre-training (CLIP) model by Radford et al. (2021) [4] represents a notable leap forward in multimodal representation learning, originally designed to associate visual and textual information within a shared embedding space. By training on massive collections of internet-sourced image-caption pairs, CLIP acquires a robust alignment between images and their corresponding linguistic descriptions, facilitating zero-shot inference on downstream tasks. Notably, CLIP's contrastive loss objective encourages correct (image-text) pairs to cluster together in latent space, while mismatched pairs are driven apart. This training regimen yields an embedding space that generalizes to novel tasks without requiring explicit fine-tuning, demonstrating an adaptability that has spurred a flurry of subsequent research.

Although CLIP's impact is most visible in image-text alignment, variations of this framework have shown promise in domains where labeled data is inherently scarce. Researchers have adapted CLIP-like architectures to specialized settings—e.g., medical imaging—by aligning clinical images with textual reports [5, 46]. These adaptations illuminate how contrastive learning can address data limitations common in industry and science, but also underscore the need to accommodate new data modalities and domain constraints. For instance, while ConVIRT [5] aligns radiology images with textual findings for improved diagnostic support, models such as MammoCLIP [46] tailor the CLIP pipeline to mammography images by incorporating domain-specific architectures and augmentations. Despite their success, these medical imaging models rely heavily on semantic textual descriptions—an assumption that may not hold in purely numeric or sensor-based contexts.

In practice, CLIP's architecture comprises two encoders: an image encoder and a text encoder, as seen in the original architectural diagram given by figure 1. The original formulation uses a convolutional or transformer-based network for images, paired with a transformer-based text encoder. A contrastive loss function such as Van Oord's (2018) InfoNCE loss [37] (or a symmetric variant) drives the model to match each image with its corresponding text. Applying this paradigm to wafer metrology would involve substituting images with primary measurements and text with secondary measurements (or vice versa), then training the model to align these two data views whenever a wafer is fully observed. Under missing-by-design conditions, the aim is that the encoder learns associations among distinct wafer measurements even when only partial data are available at inference time.

Extending CLIP to such numeric pairings introduces at least four challenges. First, the image and text encoders must be repurposed to handle continuous tabular data rather than pixels and semantic tokens. Second, the model must preserve the physical or operational relationships among measurements, which may not be adequately captured by general-purpose data augmentations or simple similarity metrics in the objective function. Unlike text-based descriptions that often embody rich semantic content, numeric secondary measurements in metrology may offer more subtle cues for alignment, necessitating carefully tuned similarity metrics or domain-specific augmentations. We must therefore explore whether and how domain knowledge—such as physical constraints, allowable ranges, or known wafer-level correlations—might be integrated into the contrastive loss to enforce physically plausible imputations. Third, despite initial demonstrations of CLIP's versatility in multimodal tasks, there remains limited evidence that it naturally extends to continuous industrial data. Previous explorations have focused primarily on semantic alignment (e.g., image-text captioning) or specialized clinical images with textual reports, leaving open questions about how effectively a contrastive framework can discover latent correspondences in purely numeric data domains.

Lastly, and possibly most importantly, while CLIP embeddings are powerful at suggesting which pairs belong together, they do not inherently provide exact numeric predictions. For many engineering or scientific tasks, approximate proximity in an embedding space may be insufficient if the imputed values must meet rigorous accuracy or regulatory thresholds. This remains a significant gap in current literature. Possible research to address this might be combining a coarse contrastive alignment step with a subsequent fine-tuning or refinement step that narrows the discrepancy between embedded pairings and actual measurements. Such an approach could treat the CLIP encoder as a prior to 'get close' to the target, then deploy a secondary residual to refine the estimate.

Addressing these challenges altogether constitutes our third research subquestion of this paper, namely: *How can the CLIP framework be adapted to better suit continuous value imputation, and to what extent can it account for underlying physical data relationships between primary and secondary data pairs?*

As such, in this thesis, we seek to adapt the core CLIP methodology to handle deliberately incomplete, continuous-valued wafer data. Building on the foundational principle that contrastive learning can align partial observations in a shared space, we examine whether a CLIP-based model can more reliably impute missing measurements than simpler regression or matrix-completion techniques, particularly when the missingness mechanism is purposeful. Our hypothesis is that contrastive representations may better preserve local structure and domain-specific nuances, reducing the risk of overfitting or linearity constraints. We further hypothesize that by regularizing CLIP for domain constraints, like the relationship between measurement pairs and overly, as well as extending CLIP beyond simple pairing for imputation, we might achieve better performance in metrology data imputation. By evaluating performance on high-dimensional metrology datasets with systematically withheld secondary measurements, we aim to clarify whether CLIP and its adaptations indeed offer a superior alternative—or at least a viable complement—to standard approaches in semiconductor manufacturing settings.

This concludes the literature review section of this thesis. Having identified the literature gaps in current research regarding data imputation in wafer metrology, we will now construct the necessary methodology to address them.

3 Methodology

This chapter outlines the methodological framework for addressing the research questions and sub-questions. We begin by describing the structure and statistical properties of our wafer metrology dataset, including the design of our planned missingness. Next, we define the metrics and tools used to gauge overlay accuracy and systematic error reduction, ensuring that each imputation method—ranging from classical regression to advanced neural-network approaches—is evaluated under consistent, industry-relevant criteria. We then detail the data preprocessing steps, the model training procedures, and the specific measures used to quantify reconstruction fidelity, highlighting adaptations necessary to accommodate high-dimensional metrology data. Finally, we discuss the statistical tests applied to determine the significance of any performance differences. Through these steps, we aim for a transparent and replicable methodology that provides a solid foundation for interpreting the subsequent empirical findings. The detailed algorithms for each imputation method can be found in appendix A.

3.1 Data Description, Metrics and Experimental Setup

Central to ASML's lithography machines is the YieldStar metrology tool, which captures optical measurements of semiconductor wafers for quality control. Each measurement site on the wafer $i \in \{1, \ldots, N\}$ yields two p-dimensional measurement vectors: a *primary* measurement $\mathbf{x}_i^{(P)} \in R^p$ and a *secondary* measurement $\mathbf{x}_i^{(S)} \in R^p$. Under ideal, perfectly symmetric conditions, these two vectors would match; in practice, systematic differences arise from equipment asymmetries, environmental fluctuations, or wafer-specific properties.

In our MBD framework, we choose not to measure the secondary point at certain locations. Let $\Omega \subseteq \{1, \ldots, N\}$ be the set of indices for which $\mathbf{x}_i^{(S)}$ is observed. For $i \in \Omega$, the measurement pair $(\mathbf{x}_i^{(P)}, \mathbf{x}_i^{(S)})$ is available; for $i \notin \Omega$, only $\mathbf{x}_i^{(P)}$ is known. This *Missing By Design* (MBD) mechanism stems from practical limitations: acquiring both measurements at every site can be time-consuming and resource-intensive, motivating the selective omission of $\mathbf{x}_i^{(S)}$ at certain locations. Thus, our imputation objective is to learn a mapping

$$\hat{\mathbf{x}}_i^{(S)} = f\big(\mathbf{x}_i^{(P)}\big),$$

where $f(\cdot)$ is estimated using the paired subset $\{(\mathbf{x}_i^{(P)}, \mathbf{x}_i^{(S)}) : i \in \Omega\}$, and subsequently applied to approximate the missing vectors $\{\hat{\mathbf{x}}_i^{(S)} : i \notin \Omega\}$.

In this study, we then partition the data into a *paired* set,

$$\mathbf{X}_{\text{paired}} = \big\{(\mathbf{x}_i^{(P)}, \mathbf{x}_i^{(S)}) : i \in \Omega\big\},$$

and an *unpaired* set of primary-only readings,

$$\mathbf{X}_{\text{unpaired}}^{(P)} = \big\{\mathbf{x}_i^{(P)} : i \notin \Omega\big\}.$$

In practice, we specifically employ a randomized 1:1:4 train-validation-test split of our candidate dataset of complete paired measurements to mimic few-shot conditions. The training dataset is thus composed of complete primary-secondary data pairs, with the validation and test sets only containing primary measurements. Our randomized split mimics random sampling of wafer locations for primary-only measurement on the production line, which is an MAR design mechanism for our MBD framework and is therefore ignorable. We pick the best model parameters by evaluating the validation set at each training epoch, after which we use the trained model to impute secondary measurements for the test set. Each imputation method—whether a classical regression, a matrix-completion algorithm, or a more advanced neural approach—must predict $\hat{\mathbf{x}}_i^{(S)}$ for $i \notin \Omega$ based on $\mathbf{x}_i^{(P)}$. We assess the quality of these imputations not only by direct comparison to ground truth ($\mathbf{x}_i^{(S)}$ is known) but also by evaluating downstream improvements in metrology accuracy, such as overlay inference performance and tool-to-tool overlay differences when training an overlay inference network to the imputed dataset. Since semiconductor processes are complex and typically exhibit nonlinear dependencies, robust performance in both raw reconstruction and practical error correction serves as a strong indicator of an imputation technique's suitability.

Here, the mean squared reconstruction error in imputing the true secondary measurements is defined as:

$$\mathcal{L}_{\mathrm{MSRE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i^{(S)} - \hat{\mathbf{x}}_i^{(S)}\right)^2 . \tag{1}$$

This error metric is used to train each model, except for the CLIP model which uses the infoNCE loss and will be discussed in a subsequent section.

As mentioned, to judge the downstream effect of our reconstruction, we further evaluate the performance of the overlay prediction network trained on the imputed datasets. The loss function of the overlay prediction network is the mean squared overlay error between the true overlay targets y_i and the predicted overlays $\hat{y}_i$:

$$\mathcal{L}_{\mathrm{MSOE}} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 , \tag{2}$$

where y_i is the true overlay target for sample i, and $\hat{y}_i = q(\mathbf{x}_i^{(P)}, \hat{\mathbf{x}}_i^{(S)}; \theta)$ is the predicted overlay, with $q(\cdot;\theta)$ representing the overlay prediction network parameterized by θ. The overlay inference model is a fixed model architecture and trained on each imputation set with $\mathcal{L}_{\mathrm{MSOE}}$ as its training loss. This loss therefore has two uses: as a training loss for the overlay inference model, and as a downstream performance (non-training) metric reflecting the overall impact of the imputed secondary measurements on the accuracy of overlay predictions. Note: it is *not* used as a training loss for the imputation models, except for the domain-guided CLIP model where it reports the overlay regularization loss.

We then define the aforementioned tool-to-tool (T2T) loss that captures the variance in overlay predictions across different tools. Let $\mathcal{T} = \{t_1, t_2, \ldots, t_{N_{\mathrm{tools}}}\}$ denote the set of tools, and let $\mathcal{I}_k$ be the indices of samples associated with tool t_k.

The T2T loss $\mathcal{L}_{\mathrm{T2T}}$ is defined as:

$$\mathcal{L}_{\mathrm{T2T}} = \frac{1}{N_{\mathrm{pairs}}}\sum_{(k,l)}\left(\bar{\hat{y}}_k - \bar{\hat{y}}_l\right)^2 , \tag{3}$$

where $\bar{\hat{y}}_k = \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\hat{y}_i$ is the mean predicted overlay for tool t_k, $N_{\mathrm{pairs}} = \binom{N_{\mathrm{tools}}}{2} = \frac{N_{\mathrm{tools}}(N_{\mathrm{tools}}-1)}{2}$ is the total number of unique tool pairs, and the summation is over all unique pairs (k,l) with $k < l$. This loss measures the average squared difference between the mean overlay predictions of all tool pairs, capturing inter-tool variability. A lower $\mathcal{L}_{\mathrm{T2T}}$ indicates better alignment of overlay predictions across different tools, suggesting that the imputation model effectively mitigates tool-to-tool differences.

Overall, this setup allows us to the research questions with precise, systematic analysis. Specifically, we will compare the MSRE and MSOE between imputation models to determine which are best suited to the data and, in general, how effectively discriminative imputation methods reconstruct missing secondary signals in wafer metrology data to reduce systematic machine errors and improve overlay prediction accuracy.

To further assess how faithfully each imputation approach reproduces the latent structure of our wafer metrology data, we perform Principal Component Analysis (PCA) on both the fully observed (target) dataset and each imputed dataset. Concretely, suppose the target dataset is represented by a matrix

$$\mathbf{X}_{\text{target}} \in R^{N \times d},$$

where N is the number of observations (e.g., wafer sites) and d is the number of features (possibly after dimensionality reduction with PCA). Let $\mathbf{X}_{\text{imp}} \in R^{N \times d}$ denote a corresponding imputed dataset in which missing secondary readings have been replaced with a model's estimates.

Step 1: Principal Components. We apply PCA to each dataset by computing its eigendecomposition or singular value decomposition (SVD). Specifically, for $\mathbf{X}_{\text{target}}$, we decompose the data-centered covariance matrix to obtain principal components $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ and associated eigenvalues $\{\lambda_1, \ldots, \lambda_r\}$, where $r \leq \min(N, d)$ is the effective rank or number of retained components. By convention, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$. A similar procedure is used to derive the principal components $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ and eigenvalues $\{\mu_1, \ldots, \mu_r\}$ for the imputed dataset $\mathbf{X}_{\text{imp}}$.

Step 2: Explained Variance. For each dataset, we define the *explained variance* of the j-th component as

$$\mathrm{EV}_j^{(\text{target})} \;=\; \frac{\lambda_j}{\sum_{k=1}^{r} \lambda_k} \quad \text{and} \quad \mathrm{EV}_j^{(\text{imp})} \;=\; \frac{\mu_j}{\sum_{k=1}^{r} \mu_k}.$$

These ratios quantify how much of the total data variance each principal component captures. The *cumulative explained variance* up to component m is then

$$\sum_{j=1}^{m} \mathrm{EV}_j^{(\text{target})} \quad \text{or} \quad \sum_{j=1}^{m} \mathrm{EV}_j^{(\text{imp})}.$$

Plotting the cumulative difference in explained variance between two datasets as a function of principal components considered allows us to see whether the imputed dataset distributes its total variance across principal components in a manner similar to the target (Figure 14).

Step 3: Cosine Similarity and Similarity Trace. Once we obtain the principal components, we compare the directions (i.e., eigenvectors) of the imputed dataset to those of the target. For the j-th component, let $\mathbf{v}_j \in R^d$ (target) and $\mathbf{u}_j \in R^d$ (imputed). We define the *cosine similarity* as

$$\operatorname{cosim}(\mathbf{v}_j,\, \mathbf{u}_j) \;=\; \frac{\mathbf{v}_j^\top\, \mathbf{u}_j}{\|\mathbf{v}_j\|\, \|\mathbf{u}_j\|}. \tag{4}$$

Higher values reflect closer alignment between the two principal directions. Summing or averaging these individual similarities across components yields a *similarity trace*, which is often plotted as a function of the component index (Figure 13). A steep decline in the similarity trace for higher-order components indicates that the imputed dataset matches the target only in the leading modes of variation.

Step 4: Cosine Similarity Matrix and Heatmaps. To visualize a finer-grained comparison across all pairs of components, we form a *cosine similarity matrix* $\mathbf{S} \in R^{r \times r}$, where

$$\mathbf{S}_{ij} \;=\; \mathrm{cosim}(\mathbf{v}_i,\, \mathbf{u}_j).$$

If components i and j align well, $\mathbf{S}_{ij}$ is close to 1; a value near 0 indicates orthogonality. Figure 15 displays such matrices as heatmaps, revealing how specific components from the imputed dataset do or do not coincide with those from the target. A diagonal band of high similarity suggests close alignment in most principal axes, whereas off-diagonal blocks may indicate that a method shifts certain modes of variation relative to the target.

Collectively, these steps provide a comprehensive evaluation of structural fidelity in the imputed data. By inspecting explained variance, cosine similarities of corresponding components, and the overall similarity matrix, we gain insight into whether each imputation technique preserves both dominant and subtle patterns critical for wafer metrology. This PCA-based methodology thus complements more direct error metrics by highlighting how effectively a model replicates the underlying geometry and variance distribution of the target dataset.Combining all the analyses, we can compare diverse models under consistent conditions, identifying which strategies excel at capturing subtle interactions between primary and secondary measurements to achieve tangible gains in manufacturing precision.

3.2 Dataset Analysis

A crucial step in evaluating how best to impute missing wafer measurements is to characterize the dominant modes of variation between primary and secondary signals, as well as any finer-grained structure not captured by simple low-rank approximations. To this end, we first construct a difference matrix

$$\mathbf{D} \;=\; \mathbf{X}^{(S)} - \mathbf{X}^{(P)},$$

where each row of $\mathbf{X}^{(P)}$ (respectively $\mathbf{X}^{(S)}$) contains the primary (respectively secondary) measurement vector for a given site. Subtracting these vectors isolates the net systematic error incurred by Yieldstar tool asymmetries, calibration drifts, or other machine-specific factors.

Singular Value Decomposition (SVD). To examine the rank-deficient nature of $\mathbf{D}$, we perform an SVD:

$$\mathbf{D} \;=\; \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^{\mathsf{T}},$$

where $\mathbf{U} \in R^{n\times r}$ and $\mathbf{V} \in R^{m\times r}$ have orthonormal columns, and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ contains the nonzero singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. Here, n and m denote the number of wafer sites and total features, respectively, and $r = \mathrm{rank}(\mathbf{D})$. Plotting these singular values in descending order highlights any elbow in the spectrum, suggesting that most variance is concentrated in the first few modes.

Cumulative Energy. To quantify how quickly these modes capture the variance of $\mathbf{D}$, we define the *cumulative energy* function

$$E(k) \;=\; \frac{\sum_{j=1}^{k} \sigma_j^2}{\sum_{j=1}^{r} \sigma_j^2}.$$

A sharp rise in $E(k)$ for small k indicates that a low-rank approximation, constructed by retaining only the leading k singular values and vectors, can explain most of the large-scale differences between primary and secondary measurements. By contrast, slowly growing $E(k)$ would imply more diffuse variance.

Residual Matrix. We further investigate the presence of localized structure by examining the residual

$$\mathbf{R} \;=\; \mathbf{D} \;-\; \mathbf{D}^{(r_{\text{trunc}})},$$

where $\mathbf{D}^{(r_{\text{trunc}})} = \sum_{j=1}^{r_{\text{trunc}}} \sigma_j\,\mathbf{u}_j\,\mathbf{v}_j^{\mathsf{T}}$ is the rank-r_{trunc} truncated SVD reconstruction. Subtracting this approximation from the original difference matrix produces a residual matrix $\mathbf{R}$ that ideally contains only random noise if a low-rank model fully explains the data. When heatmaps of $\mathbf{R}$ reveal faint vertical or horizontal banding, however, we infer that unmodeled systematic effects remain.

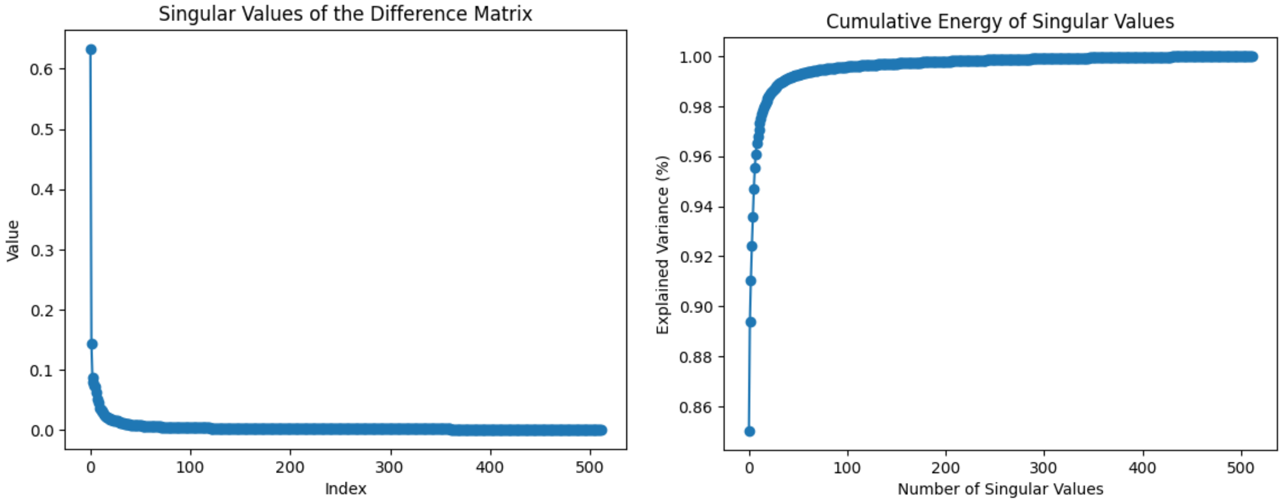
Block Aggregation. To visualize these patterns more effectively, we partition $\mathbf{R}$ into blocks of size 500×50. Concretely, each continuous set of 500 rows and 50 columns is aggregated into a smaller matrix block, which is then displayed as a single tile in the heatmap. By grouping related entries, we can highlight correlated anomalies or banding that might otherwise be drowned out.

Comparing the block-aggregated residual matrix to a synthetic noise matrix—of matching dimensions and variance scale—helps distinguish genuine structured signals from random fluctuations.

Row- and Column-Wise Mean and Variance. To further characterize these residual anomalies, we compute row- and column-wise statistics. Specifically, for each row i in $\mathbf{R}$, we record

$$\text{mean}_i = \frac{1}{m}\sum_{j=1}^{m}\mathbf{R}_{ij} \quad \text{and} \quad \text{var}_i = \frac{1}{m}\sum_{j=1}^{m}\Big(\mathbf{R}_{ij} - \text{mean}_i\Big)^2.$$

Analogously, each column j has its own mean_j and var_j. Spikes or troughs in these statistics indicate non-uniform patterns localized to particular subsets of sites or features, underscoring the possibility that even a robust low-rank approximation may miss nuanced but physically meaningful variations.



(a) (b)

Figure 2: (a) Magnitude of the singular values of the difference matrix $\mathbf{D}$ as a function of their index, and (b) their cumulative energies. The pronounced elbow and the rapid saturation point underscore the matrix's low-rank structure.

By combining the SVD-based low-rank perspective with residual analysis and block aggregation, we gain a comprehensive view of how closely primary and secondary measurements align and where systematic errors persist. The results of the dataset analysis are given by figures 2, 3, 4. As shown in Figure 2(a), the singular values exhibit a pronounced elbow, indicating that the leading principal components account for much of the systematic discrepancy between primary and secondary measurements. In Figure 2(b), the cumulative energy plot confirms that only a few modes can effectively capture the large-scale variance. Although this observation suggests that low-rank approaches, such as PCA or SVT, might accurately reconstruct the majority of primary–secondary deviations, subtracting a truncated SVD from the original difference matrix reveals additional, finer-grained structures in the residual (see Figure 3). Unlike purely stochastic noise, the faint horizontal and vertical banding visible in the residual heatmap implies localized processes—potentially including tool miscalibrations or region-specific wafer anomalies—that remain unmodeled by a simple low-rank fit.

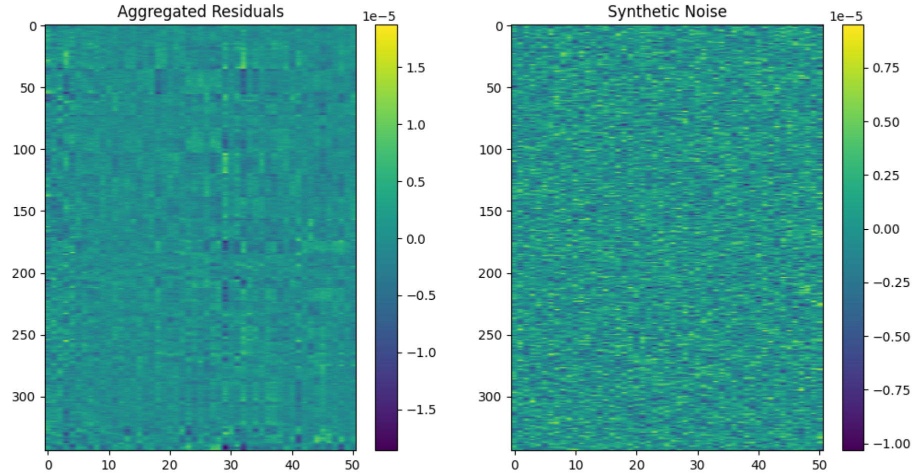


Figure 3: A heatmap of the aggregated blocks of the residual matrix (left) alongside synthetic random noise (right). The residual matrix shows noticeable banding patterns, suggesting non-random structure in the unmodeled data.

Figure 3 gives the heatmap of the aggregated residual matrix blocks of size 500 × 50 a noise heatmap for comparison. The left-hand panel of Figure 3 illustrates clear banding and correlated signals absent from the synthetic noise matrix in the right panel, underscoring the structured nature of these residuals. These appear across the entire block, though more strongly in certain areas, which might suggest particular features or locations around the wafe Finally, row- and column-wise means and variances plotted in Figure 4 verify that certain subsets deviate substantially from uniform behavior, reinforcing the conclusion that a purely low-rank model fails to capture subtle, site-specific phenomena.

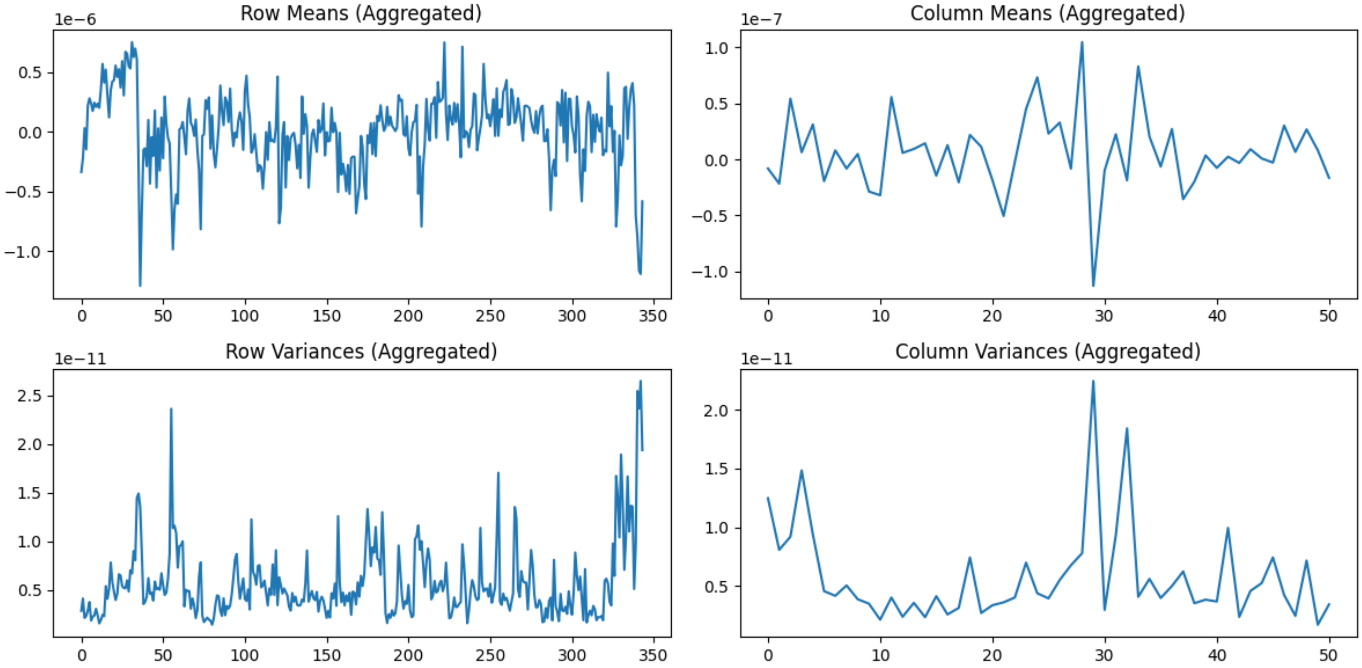


Figure 4: Row and column mean/variance of the aggregated residual matrix blocks. The spikes in mean and variance confirm non-uniform structural patterns rather than random noise.

From the standpoint of imputation methodology, these findings emphasize two key messages. First, the dominance of a few principal directions suggests that matrix-completion methods could potentially address a significant portion of the primary–secondary difference, leveraging the data's rank-deficient structure. Second, persistent localized anomalies indicate that more flexible models—those that extend beyond linear assumptions—may be necessary to handle site- or machine-specific errors that remain visible in the residual. As the results section will show, imputation techniques that reconcile both global rank-deficient signals and smaller-scale systematic patterns have the greatest potential to preserve overlay accuracy across varying industrial conditions.

3.3 PCA Pre-Processing

Building on the findings that our wafer data exhibits both rank-deficient structure and localized anomalies, we next perform a global dimensionality reduction to filter noise and redundant features. Specifically, we apply Principal Component Analysis (PCA) jointly to the primary and secondary measurements, concatenating them horizontally into a matrix

$$\mathbf{X} \in R^{N\times(2p)},$$

where each row corresponds to the combined feature vectors for a single wafer site. By extracting the principal axes from this unified matrix, both primary and secondary signals are projected onto the same lower-dimensional subspace, ensuring consistent representations for subsequent learning tasks.

Beyond simplifying the feature space, PCA helps mitigate risks of overfitting, since components dominated by noise or marginal variance are discarded. To identify an appropriate compression level, we experimentally vary the retained variance threshold from 90% to 99%, then assess the mean squared reconstruction error of a baseline linear regression model under identical hyperparameters. Although a lower threshold (e.g., 90%) reduces dimensionality more aggressively, it may also discard relevant structure essential for capturing subtle primary–secondary relationships. Conversely, pushing beyond 99% can preserve noise. Balancing these considerations, we find that retaining 98% of the total variance offers an optimal trade-off between computational feasibility and accurate modeling of critical wafer features. This choice addresses memory constraints on our available GPUs while preserving the dominant modes of variability established during the preceding rank and residual analyses.

Throughout the forthcoming sections, all models are trained and validated on these PCA-reduced data, thereby ensuring consistency in both dimensionality and variance content. Specific gains from PCA—such as improved training stability, reduced memory usage, and enhanced reconstruction performance—are discussed in the results section, where we quantify the impact of PCA on imputation accuracy and ability to handle systematically missing measurements.

3.4 Hyperparameter Optimization

Stochastic gradient search methods were used to streamline the hyperparameter tuning process and reduce computational burden. Specifically, in this study, we employed Optuna [47], a framework-agnostic, advanced hyperparameter optimization tool, to efficiently and systematically explore our model's hyperparameter space for optimal sets. Optuna supports multiple optimization algorithms and further allows dynamic trial pruning strategies that cease under-performing runs early.

Optuna optimizes hyperparameters by defining an objective function that takes a trial as input, with each trial corresponding to a unique set of hyperparameters sampled from a user-defined search space. The objective function evaluates the performance for a given trial by computing and reporting the best validation loss. based on the feedback from completed trials, Optuna adaptively guides the search toward promising regions of the hyperparameter space using Bayesian search algorithms, for which we used the standard preset 'Tree-structured Parzen Estimators' (TPE). For each model, the objective function evaluated the k-fold cross-validation performance for **30 epochs** (motivated by the fact that each model reached 90% convergence by this point), and ran until the parameter optimization either converged, or reached a fixed cap of **150 trials** (due to time constraints). In combination with Optuna, we further regularized against data split sensitivity using K-fold cross-validation with **k=5 folds**, a standard number for ML studies, for each trial.

We used the Median pruner preset for Optuna's dynamic pruning, which actively evaluates the reported running validation losses for ongoing trials and compares them against the median performance of completed trials at similar evaluation steps. A trial is then pruned if its performance falls below this median threshold, thereby preserving computational resources. The pruning presets **n__startup__trials** and **n__warmup__steps** were set to **5 trials** for a sufficiently representative sample size and **10 epochs** for an adequate number of steps before a pruning decision is made, respectively. Because the Optuna pruning engine cannot distinguish between folds within a trial, only the parameters of the first fold for each trial are reported for pruning consideration.

For SVT and PCA matrix completion methods, pruning was directly handled by the library (scikit-learn) used to implement these and not by Optuna. The following tables summarizes the hyperparameter search space for the models, organized into general and model-specific parameters.

Table 1: General ML Model Hyperparameters

Hyperparameter	**Description**
Number of Hidden Layers	Varied between 1 and 6.
Layer Dimensions	Sampled from [128, 512] neurons per layer (increments of 64).
Layer Normalization	A choice to include layer normalization after a linear transformation.
Activation Functions	ReLU, LeakyReLU, SELU, SiLU, Tanh.
Learning Rate	Logarithmic scale between 10^{-4} and 10^{-6}.
Output Dimension	Defines the dimensionality of the final layer output.

Table 2: Model-Specific Hyperparameters

Model	**Hyperparameter**	**Description**
CLIP Encoder Optimization	Temperature (T)	Cross-entropy loss scaling parameter (0.01 to 1).
	k	Number of CLIP pairings averaged by cosine similarity.
DGCLIP Adaptation (additional to CLIP)	p	Overlay loss scale parameter.
Matrix Completion	Rank τ	PCA matrix completion rank. SVT soft-thresholding parameter.

3.5 Procedure Overview for Imputation and Overlay Prediction

This subsection details the specifics of model training and inference at each stage, from imputation to downstream overlay prediction. Two Nvidia Tesla T4 GPUs with 16 GB memory each were used in parallel for this study. Following our PCA-based dimensionality reduction, we define two key data subsets for the imputation and overlay tasks. First, recall that

$$\Omega \subseteq \{1, \ldots, N\}$$

denotes the set of wafer site indices for which both primary and secondary measurements are available. For ease, we further define the remaining indices $\Omega^c = \{1, \ldots, N\} \setminus \Omega$ comprise those sites missing secondary measurements. Let us further define:

$$\mathcal{D}_{\text{complete}} \;=\; \bigl\{\,(\mathbf{x}_i^{(P)},\,\mathbf{x}_i^{(S)}) \;\big|\; i \in \Omega\bigr\} \quad \text{and} \quad \mathcal{D}_{\text{primary}} \;=\; \bigl\{\,\mathbf{x}_j^{(P)} \;\big|\; j \in \Omega^c\bigr\}.$$

Here, each $\mathbf{x}_i^{(P)}$ and $\mathbf{x}_i^{(S)}$ is assumed to already lie in the PCA-reduced subspace.

Imputation Stage To impute missing secondary readings, we train a model—statistical or machine learning–based—on $\mathcal{D}_{\text{complete}}$. Statistical methods (e.g., matrix completion) directly map $\mathbf{x}_i^{(P)} \mapsto \hat{\mathbf{x}}_i^{(S)}$, while machine learning approaches (e.g., neural networks) learn a parameterized function

$$f : \mathbf{x}_i^{(P)} \;\longmapsto\; \mathbf{x}_i^{(S)}.$$

For specific algorithmic procedures of each method, see appendix ?. Each models best parameters are chosen by taking the parameters at the epoch (training iteration) which gives the best MSRE on the validation set. Once trained, the model is applied to each $\mathbf{x}_j^{(P)}$ in $\mathcal{D}_{\text{primary}}$ to produce imputed complements $\hat{\mathbf{x}}_j^{(S)}$. For test pairs in Ω, we can estimate the MSRE using equation (1). We then form a *tool-corrected* dataset by combining each primary measurement with its newly imputed complement:

$$\tilde{\mathbf{x}}_j \;=\; \mathcal{P}\bigl(\mathbf{x}_j^{(P)},\,\hat{\mathbf{x}}_j^{(S)}\bigr),$$

where $\mathcal{P}$ is the necessary post-processing or alignment function. Collecting these results for all $j \in \Omega^c$ yields $\tilde{\mathcal{D}}_{\text{train}} = \{\tilde{\mathbf{x}}_j : j \in \Omega^c\}$. Following this, we analyze the reconstructed data against the target data using PCA as described in section 3.2.

Overlay Prediction Stage Using $\tilde{\mathcal{D}}_{\text{train}}$, we train an overlay network

$$q\bigl(\tilde{\mathbf{x}}_j;\psi\bigr) \;\approx\; y_j,$$

where y_j is the overlay measurement at site j. A validation set $\tilde{\mathcal{D}}_{\text{val}}$ guides the hyperparameter tuning of ψ, minimizing

$$\mathcal{L}_{\text{overlay}} \;=\; \frac{1}{|\tilde{\mathcal{D}}_{\text{val}}|} \sum_{j\,\in\,\text{val}} \Bigl|\,y_j - \hat{y}_j\Bigr|, \quad \text{where} \;\; \hat{y}_j \;=\; q\bigl(\tilde{\mathbf{x}}_j;\psi\bigr).$$

We further evaluate the consistency of these overlay predictions across different YieldStar tools by measuring an auxiliary loss

$$\mathcal{L}_{\text{aux}} \;=\; \frac{1}{N_{\text{pairs}}} \sum_{(t_k,\,t_l)} \Bigl(\bar{\hat{y}}_k - \bar{\hat{y}}_l\Bigr)^2,$$

where $\bar{\hat{y}}_k$ denotes the mean predicted overlay over all sites measured by tool t_k, and $\binom{N_{\text{tools}}}{2} = N_{\text{pairs}}$ counts distinct pairs of tools. A low $\mathcal{L}_{\text{aux}}$ indicates less inter-tool mismatch—an important objective in wafer metrology.

By combining these two metrics, we capture both the raw accuracy of each imputation method (via $\mathcal{L}_{\text{MSRE}}$) and its practical impact on overlay predictions and tool-to-tool consistency. This dual perspective clarifies whether reconstructing the missing secondary signals truly improves downstream metrology tasks.

Having analyzed the dataset and techniques we employ to characterize each imputation method, we will now outline how each method is applied and expanded in the context of our study.

3.6 Regularized Regression Imputation

In a regularized regression framework, each missing secondary vector $\mathbf{x}_i^{(S)}$ is approximated by a linear mapping from its primary measurement $\mathbf{x}_i^{(P)}$. Specifically, we posit

$$\hat{\mathbf{x}}_i^{(S)} \;=\; \mathbf{x}_i^{(P)}\,\boldsymbol{\beta},$$

where $\boldsymbol{\beta} \in R^p$ represents a matrix of regression coefficients in the p-dimensional PCA-reduced space. To avoid overfitting, we impose a penalty on the size of $\boldsymbol{\beta}$ using a regularization term $\mathcal{R}(\boldsymbol{\beta})$. Common choices include ridge (L2), lasso (L1), or a convex blend of the two (Elastic Net), leading to an objective of the form

$$\mathcal{L}(\boldsymbol{\beta}) \;=\; \frac{1}{N}\sum_{i=1}^{N}\Big\|\,\mathbf{x}_i^{(S)} - \mathbf{x}_i^{(P)}\,\boldsymbol{\beta}\Big\|^2 \;+\; \lambda\,\mathcal{R}\big(\boldsymbol{\beta}\big),$$

where $\lambda > 0$ determines the strength of regularization. In the ridge case, $\mathcal{R}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$ permits a closed-form solution, whereas lasso ($\|\boldsymbol{\beta}\|_1$) and Elastic Net ($\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2$) typically require iterative solvers such as coordinate descent. Feature normalization is helpful to ensure that each dimension is penalized uniformly, and cross-validation on a subset of primary–secondary pairs guides the choice of λ (and α for Elastic Net).

Once $\boldsymbol{\beta}$ is learned, any unpaired primary vector $\mathbf{x}_j^{(P)}$ can be mapped to $\hat{\mathbf{x}}_j^{(S)}$. This technique often serves as a strong linear baseline and should exploit global correlations with relatively low computational overhead. Its effectiveness in high dimensions, however, may hinge on how accurately linear assumptions capture the wafer-specific relationships—an aspect we evaluate by comparing reconstruction errors and subsequent overlay improvements with those achieved by more advanced approaches.

The regularized regression imputation algorithm is given by algorithm ? of appendix ?

3.7 Singular Value Thresholding (SVT) Imputation

In the SVT approach, we treat the combined primary–secondary data matrix as partially observed and seek a low-rank approximation consistent with these observations. Let

$$\mathbf{X}^{(P)} \in R^{N\times p} \quad \text{and} \quad \mathbf{X}^{(S)} \in R^{N\times p}$$

denote the primary and secondary measurements for the N fully observed (paired) sites, and let

$$\mathbf{X}^{(P)}_{\text{unpaired}} \in R^{M\times p}$$

represent the M primary-only rows. Form a matrix

$$\mathbf{X}_{\text{complete}} \in R^{(N+M)\times(2p)}$$

by horizontally concatenating $\big[\mathbf{X}^{(P)}, \mathbf{X}^{(S)}\big]$ for the paired rows and $\big[\mathbf{X}^{(P)}_{\text{unpaired}}, \mathbf{0}\big]$ for the unpaired rows, where zeros fill in the missing secondary portion, filling the last M rows. Formally, we aim to minimize the nuclear norm $\|\mathbf{X}\|_*$ subject to the constraint that $\mathbf{X}$ match all known entries in $\mathbf{X}_{\text{complete}}$. This encourages a low-rank solution:

$$\min_{\mathbf{X}} \ \|\mathbf{X}\|_* \quad \text{such that } P_\Omega(\mathbf{X}) \ = \ P_\Omega\big(\mathbf{X}_{\text{complete}}\big),$$

where $\|\mathbf{X}\|_* = \sum_k \sigma_k$ (the sum of singular values σ_k), Ω indexes observed entries, and P_Ω enforces consistency on those entries.

The SVT algorithm iteratively performs a singular value decomposition of the current estimate $\mathbf{X}_k$, applies a soft-threshold τ to the singular values, and overwrites the known entries with their original values. Convergence occurs when successive iterates differ by less than a predefined tolerance ϵ. Larger τ results in heavier shrinkage of singular values and thus a lower-rank solution. Once converged, we extract the imputed secondary portion from the rows and columns corresponding to unpaired data.

While SVT can effectively recover major low-rank structure with modest computational overhead, it may be less suited to highly nonlinear scenarios where wafer measurements deviate from a strictly low-rank model. Nonetheless, by constraining the solution to align with observed primary–secondary pairs, SVT provides a baseline matrix-completion perspective on imputation in this industrial context.

3.8 Principal Component Analysis (PCA) Imputation

PCA imputation enforces a low-rank representation on the combined primary–secondary data matrix by repeatedly projecting onto a principal-component subspace and overwriting known entries after each reconstruction. Assume the same setup as with SVT. Though very similar in method, at each iteration, we instead approximate $\mathbf{X}_k$ via its principal components up to some rank r, then replace all known entries in the reconstructed matrix with their observed values. We terminate once successive estimates differ by less than a chosen tolerance ϵ, at which point the last reconstructed rows corresponding to the unpaired subset yield the imputed $\hat{\mathbf{X}}^{(S)}$.

Because PCA captures dominant variance directions, this method can reconstruct large-scale correlations between primary and secondary measurements with relative ease. However, as with other low-rank approaches, PCA imputation may omit finer or nonlinear patterns if the data deviate substantially from a strict low-rank assumption. Nonetheless, it provides an effective baseline for matrix-completion-style solutions and clarifies whether purely linear feature compression suffices to recover missing secondary signals in wafer metrology.

3.9 Neural Network Regression (MLP)

In a neural network setting, we treat missing secondary measurements $\mathbf{X}^{(S)}_{\text{unpaired}}$ as the regression target for each unpaired primary vector $\mathbf{X}^{(P)}_{\text{unpaired}}$. Let $\{\mathbf{X}^{(P)}_{\text{paired}}, \mathbf{X}^{(S)}_{\text{paired}}\}$ denote the training set of fully observed primary–secondary pairs. A Multi-Layer Perceptron (MLP) is then trained to learn a nonlinear mapping

$$f_\theta : \mathbf{X}^{(P)}_{\text{paired}} \;\mapsto\; \mathbf{X}^{(S)}_{\text{paired}},$$

where θ encompasses the MLP's weights, biases, and activation parameters (for a detailed equation and representation of the forward and backward algorithms, refer to appendix A). During inference, the trained model imputes each missing $\hat{\mathbf{X}}^{(S)}_{\text{unpaired}} = f_\theta(\mathbf{X}^{(P)}_{\text{unpaired}})$ An MLP consists of an input layer that receives $\mathbf{X}^{(P)}$, multiple hidden layers applying affine transformations plus nonlinear activations (e.g., ReLU or tanh), and an output layer that returns continuous estimates of $\mathbf{X}^{(S)}$. We optimize the model by minimizing mean squared error (MSE) on the training set:

$$\mathcal{L} \;=\; \frac{1}{N}\sum_{i=1}^{N} \Big\| \,\mathbf{X}^{(S)}_{i,\text{paired}} \;-\; f_\theta(\mathbf{X}^{(P)}_{i,\text{paired}})\Big\|_2^2,$$

updating the network parameters θ via backpropagation and gradient descent. Once trained, the MLP imputes $\hat{\mathbf{X}}^{(S)}_{\text{unpaired}}$ for all unpaired primary samples, thereby providing a flexible, nonlinear solution that can surpass linear regression when wafer measurements exhibit complex dependencies. However, achieving good performance requires careful hyperparameter tuning and adequate representation of diverse primary–secondary patterns in the training data.

3.10 CLIP

To extend Contrastive Language-Image Pre-training (CLIP) from its original image–text paradigm to wafer metrology, we replace the standard encoders with dual MLP networks designed for numeric primary and secondary measurements. Let

$$f_\theta : R^p \to R^d \quad \text{and} \quad g_\phi : R^p \to R^d$$

respectively embed primary and secondary vectors into a shared d-dimensional latent space. During training, we form batches of N paired observations $\{(\mathbf{x}_i^{(P)}, \mathbf{x}_i^{(S)})\}_{i=1}^N$, computing embeddings $\mathbf{z}_i^{(P)} = f_\theta(\mathbf{x}_i^{(P)})$ and $\mathbf{z}_i^{(S)} = g_\phi(\mathbf{x}_i^{(S)})$. A contrastive loss (we employ InfoNCE [37] for this study) encourages positive primary–secondary pairs to rank higher in similarity than any mismatched negative pairs:

$$\mathcal{L}_{CE} \;=\; -\frac{1}{N}\sum_{i=1}^{N} \log\Bigl(\frac{\exp\bigl(\text{sim}(\mathbf{z}_i^{(P)},\, \mathbf{z}_i^{(S)})/t\bigr)}{\sum_{j=1}^{N} \exp\bigl(\text{sim}(\mathbf{z}_i^{(P)},\, \mathbf{z}_j^{(S)})/t\bigr)}\Bigr),$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity, and $t > 0$ (temperature) governs the distribution's sharpness. We also apply a symmetrical loss for $\mathbf{z}_i^{(S)} \leftrightarrow \mathbf{z}_j^{(P)}$ matching to ensure both encoders learn reciprocal embeddings. In this InfoNCE framework, each primary vector $\mathbf{x}_i^{(P)}$ and its corresponding secondary vector $\mathbf{x}_i^{(S)}$ form a positive pair, while $\bigl(\mathbf{x}_i^{(P)}, \mathbf{x}_j^{(S)}\bigr)$ for $j \neq i$ are considered negatives. When minimizing this loss function, by exponentiating the cosine similarity in the numerator for the correct pair and summing exponentiated similarities of all pairs in the denominator, the model optimizes by assigning a higher probability to $\bigl(\mathbf{z}_i^{(P)}, \mathbf{z}_i^{(S)}\bigr)$ being matched than to $\bigl(\mathbf{z}_i^{(P)}, \mathbf{z}_j^{(S)}\bigr)$ with $j \neq i$. Maximizing the log probability of the true pair thus compels the encoders oders f_θ and g_ϕ to construct a shared shared d-dimensional latent space that *pulls* positive embeddings closer and *pushes* mismatched embeddings farther apart.

Critical design choices include (1) the embedding dimension d, which too large can impede generalization and too small may underfit; (2) the temperature τ, whose lower values force more confident separation of true and false pairs; and (3) the batch size, since the InfoNCE loss constructs an $N \times N$ similarity matrix for each update and can become memory-intensive.

Once trained, we embed any unpaired primary vector $\mathbf{x}_i^{(P)}$ into $\mathbf{z}_i^{(P)} = f_\theta(\mathbf{x}_i^{(P)})$. To recover $\hat{\mathbf{x}}_i^{(S)}$, we retrieve secondary embeddings $\mathbf{z}_j^{(S)}$ close to $\mathbf{z}_i^{(P)}$ in the latent space. A common strategy is to search among the known secondary embeddings for the nearest neighbor(s), then project back to the original measurement domain:

$$\hat{\mathbf{x}}_i^{(S)} \;=\; \textstyle\sum_{j\in\mathcal{N}_i} p_j\, \mathbf{x}_j^{(S)},$$

where $\mathcal{N}_i$ indexes the top k neighbors by cosine similarity (equation 4), and the weights p_j form a softmax over these similarities. This reconstruction leverages CLIP's contrastive embedding directly, offering an alternative to purely parametric regressions. However, it entails a query-time cost that scales with the size of the training secondary set and may be computationally intensive.

For large-scale production data, smaller embedding sizes are needed to manage memory and time constraints, which we addressed with PCA as discussed in earlier sections. Moreover, because of the computational intensity of the model, we further use mini-batching to reduce GPU workload.

Although CLIP harnesses robust nonlinear alignment, its success here depends on achieving sufficient coverage of the primary–secondary space in training. If certain wafer conditions or machine variations appear rarely in the paired data, the learned embeddings may fail to generalize. Moreover, the InfoNCE objective can struggle if τ or the batch size are not well tuned, leading to suboptimal separation of pairs.

3.11 DGCLIP: Integrating Domain Knowledge for Overlay-Aware CLIP

Adapting CLIP for wafer metrology can be further refined by incorporating overlay-related constraints to ensure that the model's imputed measurements remain consistent with physical conditions. Instead of relying solely on contrastive alignment, the idea is to guide the training process with an additional overlay prediction term that penalizes spurious or physically implausible pairings.

Let f_θ and g_ϕ be the encoders producing embeddings for primary and secondary measurements, respectively, and recall $\mathcal{P}$ as the known function that calibrates these measurements into the corrected measurement. After encoding the primary measurement $\mathbf{x}_i^{(P)}$ to $\mathbf{z}_i^{(P)}$ and the imputed complement $\hat{\mathbf{x}}_i^{(S)}$ to $\mathbf{z}_i^{(S)}$, we form a processed vector $\tilde{\mathbf{x}}_i = \mathcal{P}(\mathbf{x}_i^{(P)}, \hat{\mathbf{x}}_i^{(S)})$. A small projection head q_ψ then predicts an overlay signal $\hat{\mathbf{y}}_i = q_\psi(\tilde{\mathbf{x}}_i)$. By comparing $\hat{\mathbf{y}}_i$ to the observed overlay $\mathbf{y}_i$ where it is available, we regularize the imputation toward physically meaningful solutions. Specifically, we augment the contrastive loss $\mathcal{L}_{\text{contrastive}}$ with an overlay error term $\mathcal{L}_{\text{overlay}}$ and scale factor p:

$$\mathcal{L}_{\text{total}} \;=\; \mathcal{L}_{\text{contrastive}} \;+\; p\,\mathcal{L}_{\text{overlay}}, \quad \text{where}$$

$$\mathcal{L}_{\text{overlay}} = \frac{1}{N}\sum_{i=1}^{N} \Big\| \mathbf{y}_i \;-\; q_\psi\big(\mathcal{P}(\mathbf{x}_i^{(P)},\, \hat{\mathbf{x}}_i^{(S)})\big) \Big\|_2^2.$$

A suitable balance for p ensures that embeddings still obey the contrastive principle while imputations honor the physical relationship with overlay.

To determine whether such domain knowledge is necessary, we examine how naïve (purely contrastive) imputations compare to the true data. For instance, if PCA analyses reveal significant discrepancies—e.g., the low-rank structure of the imputed complements diverges from that of the real measurements—then overlay-based guidance can curb these artifacts. Conversely, if purely contrastive training already yields consistent patterns, adding a domain term might provide only marginal gains or even over-regularize the model.

When domain knowledge proves beneficial, a further strategy is to pretrain q_ψ on a subset of fully observed pairs (with known overlay), then fix or partially fine-tune it while CLIP learns the embedding. This reduces the risk of overshadowing the contrastive objective. However, the computation overhead increases because each forward pass now includes an overlay prediction and an additional term in the loss. Moreover, setting p too high can degrade embedding quality by forcing the model to overemphasize overlay alignment at the expense of learning robust pairwise similarities.

3.12 Bridge Models

Here, we introduce the *Bridge Model*, which addresses two key challenges that arise with high-dimensional data under the MBD framework. First, purely global mappings (such as MLPs) can struggle to capture nuanced, localized relationships when data are scarce or strongly nonlinear. Second, while a CLIP-based framework can align primary and secondary measurements at a coarse level, that alignment alone may fall short of the precision requirements of wafer metrology. The Bridge Model attempts to meet these needs by splitting imputation into two steps: an initial coarse retrieval via CLIP, followed by a localized refinement that translates the discrepancy between the primary inputs into a refined correction for the secondary estimates.

When an unpaired primary measurement $\mathbf{x}_i^{(P)}$ is observed, CLIP retrieves a bridge anchor $\mathbf{x}_j^{(S)}$ from training data, yielding an initial guess $\hat{\mathbf{x}}_i^{(S)}$. This coarse guess likely lacks small-scale accuracy for wafer data. To improve it, the Bridge Model considers how the primary measurement $\mathbf{x}_i^{(P)}$ deviates from $\mathbf{x}_j^{(P)}$. By learning a function

$$h_\phi : \left(\Delta\mathbf{x}_i^{(P)}, \mathbf{u}_i\right) \longmapsto \Delta\mathbf{x}_i^{(S)},$$

the model refines the imputation to $\hat{\mathbf{x}}_{i,bridge}^{(S)} = \hat{\mathbf{x}}_i^{(S)} + \hat{\Delta}\mathbf{x}_i^{(S)}$. Here, $\Delta\mathbf{x}_i^{(P)} = \mathbf{x}_i^{(P)} - \hat{\mathbf{x}}_i^{(P)}$ and $\Delta\mathbf{x}_i^{(S)} = \mathbf{x}_i^{(S)} - \hat{\mathbf{x}}_i^{(S)}$. The term $\mathbf{u}_i$ encodes auxiliary context (e.g., a PCA reduction of $\hat{\mathbf{x}}_i^{(S)}$), preventing ambiguities when two identical primary deltas map to different secondary deltas, such as concatenating location data. Though, we have found for our dataset that this term is redundant, likely because the feature space is rich enough that deltas are highly distinguishable and very unlikely to repeat.

Assume our initial CLIP imputation is not exact, i.e. $\hat{\mathbf{x}}_i^{(S)} \neq \mathbf{x}_i^{(S)}$. Under the assumption that the residual translation is unbiased, we can prove that we expect the new refined imputation $\hat{\mathbf{x}}_{i,bridge}^{(S)}$ to indeed be exact, i.e. $E\left[\hat{\mathbf{x}}_{i,bridge}^{(S)}\right] = \mathbf{x}_i^{(S)}$. Given the initial CLIP imputation $\hat{\mathbf{x}}_i^{(S)}$, we have that

$$E\left[\hat{\mathbf{x}}_{i,bridge}^{(S)}\right] = \hat{\mathbf{x}}_i^{(S)} + E\left[\hat{\Delta}\mathbf{x}_i^{(S)}\right] = \hat{\mathbf{x}}_i^{(S)} + \Delta\mathbf{x}_i^{(S)} = \hat{\mathbf{x}}_i^{(S)} + \left(\mathbf{x}_i^{(S)} - \hat{\mathbf{x}}_i^{(S)}\right) = \mathbf{x}_i^{(S)}$$

where the first equality is given by substituting the refinement equation and removing the given CLIP imputation term from the expectation, and the second equality is given by the unbiasedness assumption on our residual translation.

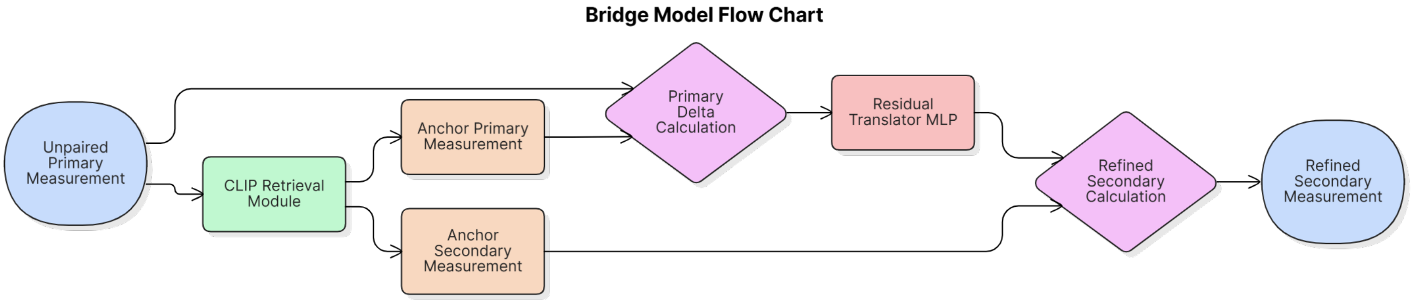

Bridge Model Flow Chart
Unpaired Primary Measurement
CLIP Retrieval Module
Anchor Primary Measurement
Anchor Secondary Measurement
Primary Delta Calculation
Residual Translator MLP
Refined Secondary Calculation
Refined Secondary Measurement

Figure 5: Diagram explaining the bridge model inference pipeline

Figure 5 gives a schematic of the inference pipeline, which is further summarized by the following process description:

Two-Step Framework.

1. *Coarse Imputation via CLIP.* We embed $\mathbf{x}_i^{(P)}$ into the contrastive space and identify a top-k neighbor among the secondary embeddings. The selected or averaged match becomes $\hat{\mathbf{x}}_i^{(S)}$, a quick, instance-level approximation that does not require a direct functional mapping from primary to complement.

2. *Refinement via Residual Translation.* The Bridge Model learns a network h_ϕ that takes $\Delta\mathbf{x}_i^{(P)}$ and outputs $\Delta\mathbf{x}_i^{(S)}$. Minimizing an MSE loss of the form
$$\sum\nolimits_{i=1}^{N} \|\Delta\mathbf{x}_i^{(S)} - h_\phi(\Delta\mathbf{x}_i^{(P)}, \mathbf{u}_i)\|^2$$
gives the model explicit guidance to bridge the gap between coarse anchors and precise secondary values. Finally, we refine the initial CLIP imputation by adding the translated secondary residual:
$$\hat{\mathbf{x}}_{i,bridge}^{(S)} = \hat{\mathbf{x}}_i^{(S)} + \hat{\Delta}\mathbf{x}_i^{(S)}.$$

It should be noted that If $\Delta\mathbf{x}_i^{(P)}$ consistently predicts $\Delta\mathbf{x}_i^{(S)}$, then the expected refined imputation $\hat{\mathbf{x}}_{i,bridge}^{(S)}$ converges to the true $\mathbf{x}_i^{(S)}$. In practice, it is crucial that the model sees nontrivial residuals, ensuring that h_ϕ gains relevant correction signals. Otherwise, if the coarse imputation is perfect on training pairs, residuals become zero, offering no gradient for learning. During training on paired data, we artificially enforce nonzero training residuals by taking the softmax weighted average of the k-best CLIP pairings with $k \geq 2$.

3.13 Conclusion of the Methodology

In this chapter, we introduced a comprehensive methodological framework for imputation in wafer metrology under Missing By Design conditions. We began by analyzing the structure of primary–secondary measurements, highlighting both low-rank patterns and localized anomalies through SVD- and PCA-based investigations. Following these insights, several candidate imputation strategies were presented: regularized regression and matrix-completion methods for handling moderate complexities, neural-network regression for capturing nonlinearities, and adapted contrastive-learning approaches (CLIP) for leveraging instance-level retrieval in high-dimensional data. Finally, we proposed two refinements that enhance these methods further: domain-aware overlay constraints to inject physical knowledge, and Bridge Models that refine coarse CLIP imputations via local residual translation.

By balancing global and local perspectives, linear versus nonlinear modeling, and purely mathematical versus domain-guided constraints, this methodology accommodates diverse scenarios in industrial wafer metrology. In the next section, we evaluate these methods empirically, examining whether each approach faithfully reconstructs missing secondary measurements, improves overlay prediction accuracy, and remains robust to the complexities of real semiconductor manufacturing processes.

4 Results and Discussion

This chapter presents the empirical evaluation of the proposed methods for metrology data imputation in MBD contexts. We begin with the optimization results of the models, followed by comparing their reconstruction performances, which are characterized by the $\mathcal{L}_{\text{MSOE}}$, similarity trace and explained difference graphs, and finally analyzing the downstream $\mathcal{L}_{\text{MSOE}}$ and $\mathcal{L}_{\text{T2T}}$ performances. Note: the optimization results of the matrix completion techniques are omitted for customer confidentiality, as they give exact rank information on the customer dataset.

4.1 Results

4.1.1 Optimization Studies

InfoNCE Loss	0.233
Learning Rate (α)	0.000500
Temperature (t)	0.704
Output Dimension (d	256
Number of Layers	1 (input layer)
Layer dimensions	[416]
Layer norms	[True]
Activations	[Tanh]

Table 3: CLIP Summary of the best trial parameters (shared between encoders)

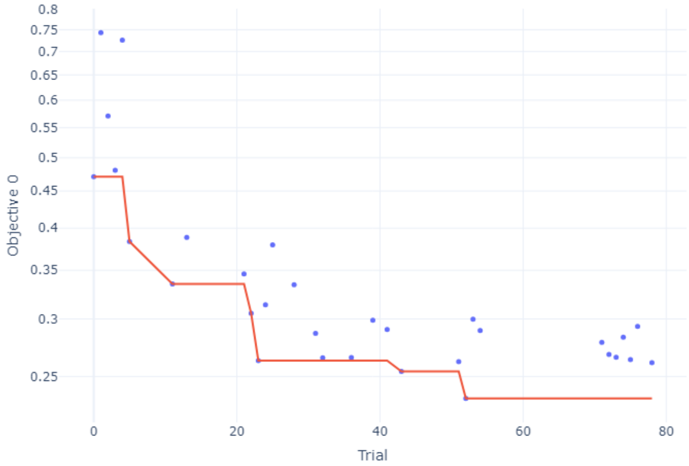


Figure 6: CLIP Objective losses (InfoNCE) of each successive trial. The study was automatically stopped at trial 78 after convergence.

CLIP Table 3 and Figures 6 and 7 summarize the hyperparameter optimization outcomes for the CLIP imputation model. Interestingly, the model favored (1) a shallow architecture (one hidden layer of dimension 416), and (2) a higher temperature (≈ 0.70), in contrast to the typical ≈ 0.07 used in large vision-language datasets [48]. A larger temperature broadens confidence across multiple pairings, while a smaller temperature intensifies confidence in fewer pairings. This elevated temperature may indicate a relatively homogeneous dataset that is easier to learn or, conversely, the presence of challenging negative pairs that benefit from less severe penalization. The results also suggest that the relation between the primary and secondary datasets does not require deep feature extraction, aligning with the preference for fewer layers. Addition-

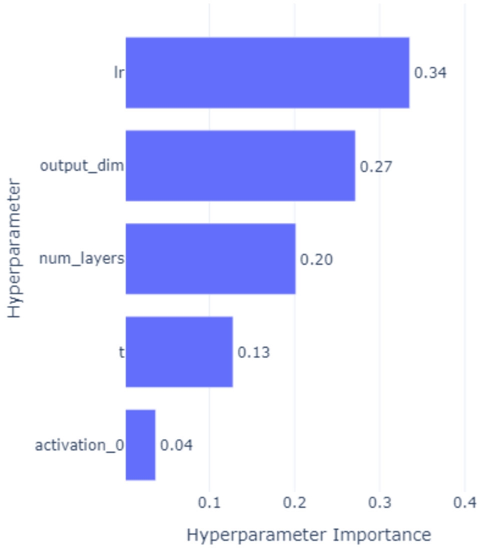


Figure 7: Optuna hyperparameter hierarchy results for CLIP parameter search. Each number represents the percentage of variation in model performance attributed to the hyperparameter.

ally, taking the soft-max weighted average of the best $k = 5$ imputation pairs yielded the best reconstruction performance for CLIP.

Metric	Value
InfoNCE loss	0.207
Learning rate (α)	0.000292
Temperature (t	0.316
DGCLIP scale (p)	225
Output Dimension (d)	256
Number of Layers	1 (input layer)
Layer dimensions	[448]
Layer norms	[True]
Activations	[ReLU]

Table 4: DGCLIP summary of the best trial parameters (shared between encoders)

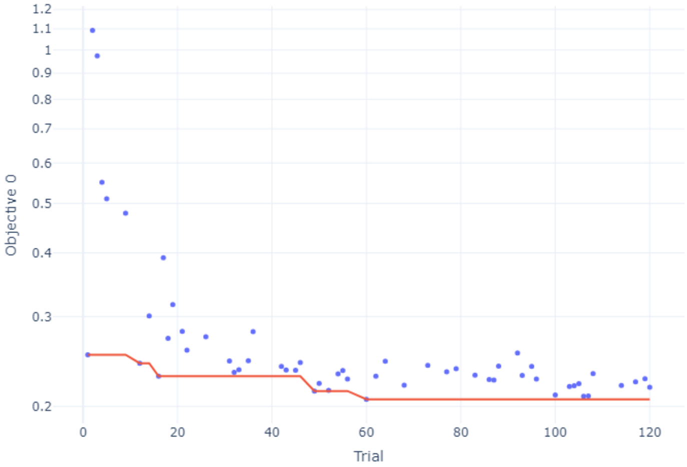


Figure 8: DGCLIP Objective losses (cross entropy) of each trial. The study was automatically stopped at trial 120 after convergence.

DGCLIP Table 4 and Figures 8 and 9 detail results from adding domain-guided regularization to CLIP. The optimal architecture, similar to the primary CLIP, again consisted of a single hidden layer (448 units), but the optimal temperature was notably lower (≈ 0.316).

Figure 10 illustrates the interplay between the DGCLIP scale p and temperature t. As p increases, the best-performing temperature region decreases, until over-regularization forces the model back to higher temperatures. This suggests that stronger physics-based constraints help the model better discriminate between correct and incorrect pairings, particularly those that have high cosine similarity yet large physical overlay discrepancies. Notably, at $p = 1$ (unscaled), the DGCLIP overlay loss added only 0.0017 to the objective, signifying modest yet influential domain regularization.

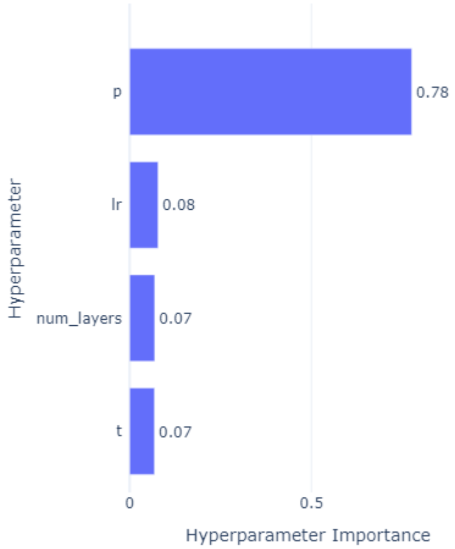


Figure 9: Hyperparameter hierarchy results for DGCLIP parameter search

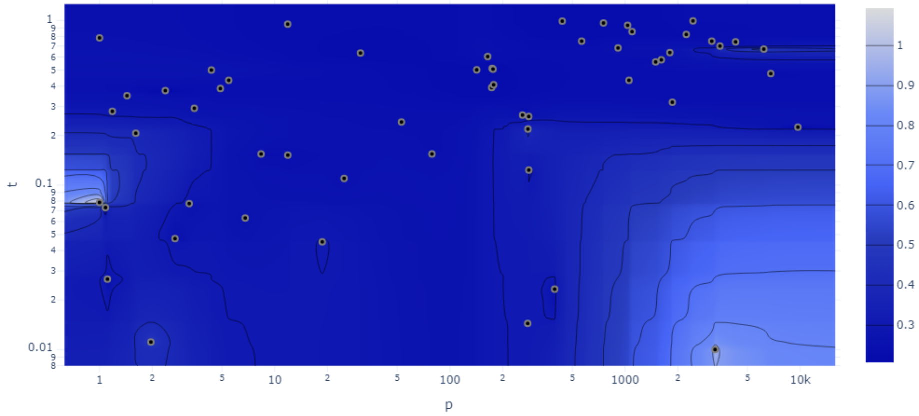


Figure 10: DGCLIP model's Optuna study contour map of objective loss for varying temperature (t) and DGCLIP scale (p)

MLP Table 5 and Figures 11 and 12 show that a moderately deep Multilayer Perceptron (five hidden layers: [320, 512, 448, 64, 256], with mixed activation functions) achieved the best mean-squared error (MSE) scores. The unusual mixture of activation functions is notable as shared activations were also trialed in the optimization runs, and negatively impacts interpretation of the model. Network depth was the most sensitive hyperparameter, reflecting the data's need for multiple layers of feature transformation.

Parameter	Value
Objective Loss (MSE)	0.0000360
Learning rate (α)	0.000350
Number of Layers	5
Layer Dimensions	[320,512,448,64,256]
Layer norms	[False,False,False,False,False]
Activations	[SiLU, ReLU, SiLU, LeakyReLU, SiLU]

Table 5: MLP summary of best trial parameters

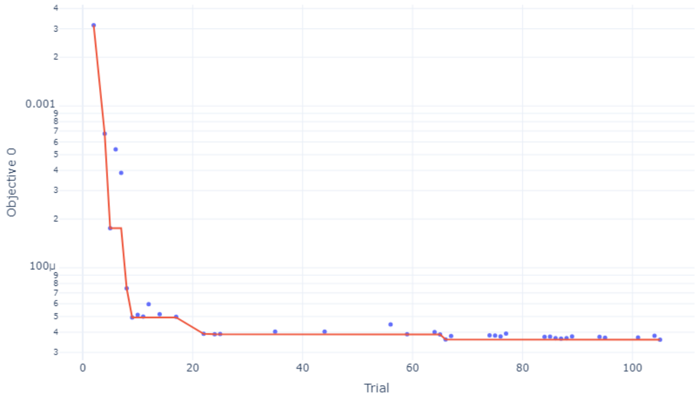


Figure 11: MLP Objective losses (cross entropy) of each trial. The study was automatically stopped at trial 106 after convergence.

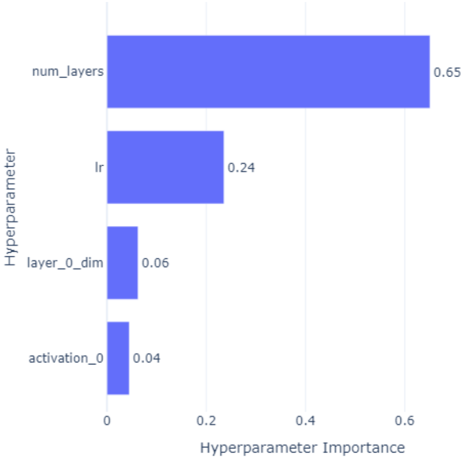


Figure 12: Hyperparameter hierarchy results for MLP parameter search

Bridge For the Bridge imputation model, best-performing parameters for the residual translator MLP (given by table 6) closely mirrored those of the CLIP

Metric	**Value**
Learning Rate	0.0000954
Number of Layers	1 (input layer)
Output dim	=Input dim
Layer norms	[False]
Activations	[Tanh]

Table 6: Bridge model summary of the best trial parameters

encoders, except it omitted layer normalization. Like CLIP, it also benefited from aggregating $k = 5$ pairs. This small network therefore adds minimal computational overhead to the CLIP model. While still technically an MLP, its lack of need for additional layers, in stark contrast to the MLP imputation model, suggests the CLIP model is providing adequately close anchors to 'bridge' easily toward the target imputations. Moreover, this may also imply the residual domain is considerably less complex to map than learning a global function like the MLP does.

4.1.2 Reconstruction Error

Method	Rel. MSE	Train Time (min)	Convergence Epoch
(baseline)	1.00	-	-
Ridge	40.4	<1	-
Lasso	30.9	<1	-
ElasticNet	30.9	<1	-
SVT matrix completion	0.524	11	48
PCA matrix completion	0.519	4	15
MLP	0.433	1	41
CLIP	0.814	3	87
DGCLIP	0.710	3	93
bridge	0.473	1 (for translator)	31

Table 7: Reconstruction errors (MSE) for each method, as a ratio of the error between the primary dataset and target secondary datasets.

Using the optimal hyperparameters for each model, we trained each model and evaluated them on a test set. Table 7 presents the relative MSE (i.e., the ratio of imputed MSE to the baseline MSE between primary and target secondary points, since primary and secondary points should be theoretically identical but are perturbed by systematic errors).

An immediate observation from Table 7 is the stark difference between linear methods and the rest: Ridge, Lasso, and ElasticNet show relative mean-squared errors (MSE) surpassing 30—more than 30 times worse than simply leaving the dataset uncorrected (baseline). This reaffirms that wafer measurement relationships are too nonlinear or complex for global linear mappings to capture. By contrast, even the simplest matrix-completion techniques (PCA and SVT) reduce the baseline error by nearly half, down to around 0.52, indicating that enforcing a low-rank constraint helps recover much of the secondary structure.

Among the more advanced models, the MLP yields the lowest relative MSE, at 0.433, within just one minute of training time and converging at 41 epochs. This suggests that a sufficiently deep, well-tuned neural network can approximate primary-secondary mappings more effectively than either matrix completion or contrastive alignment approaches. The Bridge model also converges quickly (31 epochs) and achieves a 0.473 MSE, a substantial improvement over CLIP alone (0.814). By refining the coarse pair retrieval from CLIP, Bridge provides a middle ground between purely global neural learning (MLP) and one-step contrastive matching (CLIP).

Meanwhile, DGCLIP improves upon CLIP's 0.814 MSE to 0.710, which, although not as low as the MLP or Bridge results, still indicates that adding overlay-based domain constraints yields more faithful reconstructions. Both CLIP and DGCLIP, however, demand more training epochs (87 and 93, respectively), with training times around three minutes each, reflecting the cost of repeated mini-batch contrastive learning.

Interestingly, PCA matrix completion attains 0.519 MSE with only four minutes of total run time, a modest overhead for large-scale wafer data if the primary objective is halving the baseline error. SVT, despite its iterative nuclear-norm thresholding, converges in about 11 minutes to a similar 0.524 MSE—decent performance but slower than simpler MLP-based methods in this setting.

Overall, the table highlights that the MLP achieves the best pure reconstruction error in a short training window, though is only marginally better than the bridge model and matrix completion methods. Bridge offers a significantly better reconstruction by a practical refinement of CLIP's coarse pairing without building a heavy global mapping from scratch. DGCLIP further provides an improvement by domain-physics alignment and raw reconstruction fidelity, while the linear techniques lag far behind for such complex, nonlinear wafer data.

4.1.3 PCA Analysis of Imputations

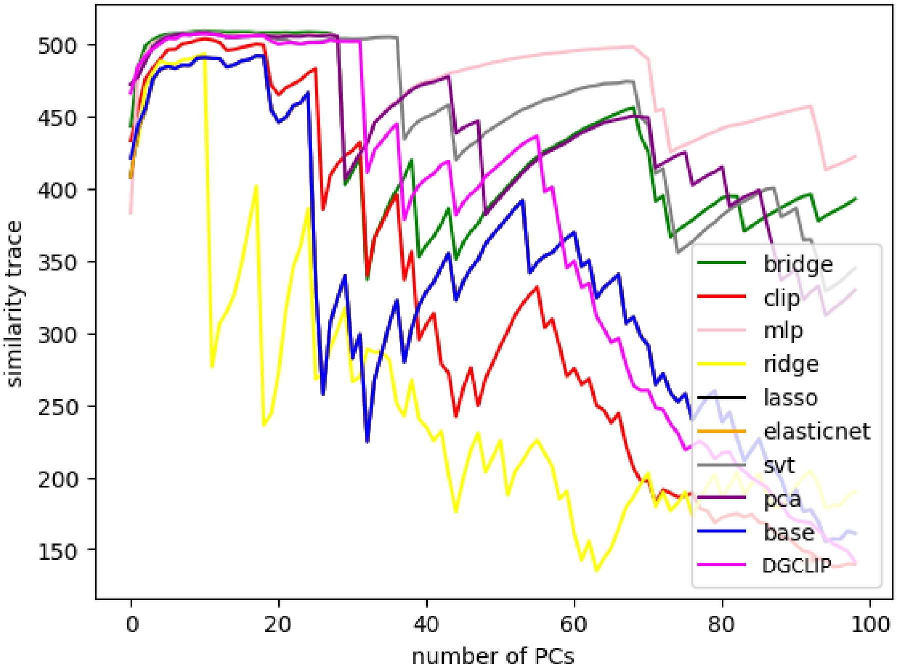


Figure 13: Plot of the cumulative cosine similarity trace between the principal components of the imputed datasets and the target dataset, as a function of the number of components considered.

Figure 13 plots the cumulative similarity trace between the principal components of each imputed dataset and those of the fully observed (target) dataset, as a function of the number of principal components (PCs). Interpreting this curve begins by recognizing that each principal component captures a progressively smaller fraction of the variance in the target data. A high similarity trace at a given number of components thus indicates that the imputed dataset's modes of variation align closely with the target's, not only for the largest variance directions but also for subtler, higher-order modes.

The baseline curve (in blue), which is the the similarity trace between the raw primary and target dataset, often shows decent alignment for the largest modes—base measurements are not wholly dissimilar from secondary signals at a broad scale—but drops significantly for deeper modes, indicating that uncorrected systematic errors degrade the representation of smaller-scale structure.

A broad takeaway from the left region of the graph is that most methods achieve relatively high alignment on the first five to ten PCs. These primary components often represent the dominant rank-deficient structure of wafer measurements, where broad discrepancies between primary and secondary readings are more straightforward to capture. Notably, linear approaches (Ridge, Lasso, ElasticNet) perform similarly to the best nonlinear methods in these early components, reinforcing the idea that large-scale variance can be approximated even by simpler methods.

As the number of components grows beyond approximately 10, more substantial differences emerge among the methods. The linear regression curves (layered on top of each other with ridge in yellow visible) begin to drop off before even the baseline curve, revealing that they actually worsen the error between the primary and target data. This is consistent with the reconstruction error performances we saw in table 7, where all linear regression methods increased the reconstruction error at least 30 times over baseline. In contrast, the matrix completion and nonlinear methods maintain higher similarity traces, suggesting they capture additional layers of local or fine-grained variation that improve the baseline. This disparity is again consistent with the reconstruction error performances, which is likely due to the heightened capacity of these models to handle high-dimensionality.

CLIP (in red) improves upon the baseline curve for the first 40 components, then perturbs the primary dataset's structure such that it dips below the baseline for the remaining components. Therefore, we can infer that CLIP's improvement in raw reconstruction error is likely due to this improvement in the initial components, and is staggered by failing to capture the higher-order components. Comparing Bridge (green) and CLIP curves highlights how a modest residual correction can bring an otherwise purely contrastive approach much closer to the target across many principal components. Beyond 20 PCs, Bridge systematically and significantly outperforms CLIP and the baseline curve, implying that local residual translations consistently reduce the mismatch that coarse CLIP pairings leave uncorrected.

The MLP (pink/magenta) line generally remains the top curve. Its deeper architecture appears well-suited to capturing high-dimensional wafer complexities, though small fluctuations at mid to high PCs might reflect sensitivity to minor modes or numerical instabilities in the SVD computations that underlie the PCA-based similarity measure. These fluctuations seem shared with the matrix completion methods and bridge model at regular intervals, supporting this idea.

Matrix completion methods (SVT in grey, PCA in purple) present remarkably respectable performances in all orders of components, only slightly falling behind the MLP when more principal components are considered. Interestingly, PCA and SVT methods behave considerably different. SVT manages the longest initial run of capturing the first components from any method, though the PCA curve follows closely behind and intersects many times.

Observing DGCLIP (violet) clarifies another dimension of the analysis. Although not always matching the MLP or matrix completion methods in raw similarity values, DGCLIP surpasses the basic CLIP approach for all components. Even more, DGCLIP improved CLIP's similarity trace in the first 30 components to the point it surpassed all methods except SVT. This confirms that domain guidance has, by some degree, quantifiably informed and constrained the network to avoid unphysical pairings.

Taken together, the figure underscores that while nearly all methods capture dominant large-scale trends (i.e., the initial PCs), the real test emerges once the target dataset's subtler variance directions come into play. Linear regression methods suffer a pronounced drop, whereas matrix completion sustain high similarity, though not as well as deep learning approaches like the MLP. This pattern confirms that, while the wafer data's complexity might be low rank enough for matrix completion methods to capture significant structure, subtle nonlinearities may necessitate deep learning approaches like the MLP (the degree depending on how much they affect downstream overlay prediction). Additionally, the bridge model and DGCLIP, in particular, highlight two strategies—residual translation or domain constraints—for addressing CLIP's mid- and high-component shortcomings without necessitating the deep, comprehensive transformation that an MLP requires.

Expanding onthe jaggedness of the curves in more depth, although each curve in Figure 13 begins relatively smooth for the first few principal components—where most data variance resides—several lines grow jagged or oscillatory once they move into higher-order modes. These fluctuations highlight how each model's reconstruction aligns (or fails to align) with increasingly subtle or low-variance directions in the target dataset, often in non-monotonic ways. A few key factors might contribute to this phenomenon.

First, by the time the analysis has advanced beyond roughly 20 or 30 principal components, each component explains only a small fraction of the data's total variance. Consequently, small discrepancies between the imputed datasets and the target can trigger disproportionately large changes in similarity. Numerical noise or micro-scale deviations in wafer data, for instance, can cause abrupt surges or dips in the cosine similarity from one principal component to the next, creating a sawtooth-like shape in the cumulative trace.

Second, certain models are inherently more sensitive to local aspects of the wafer data. Neural networks such as the MLP, Bridge, and DGCLIP may overfit to particular mid-tier components—achieving a momentarily high alignment—only to underrepresent the next few components. This localized success or failure manifests as brief peaks followed by sharp declines in the cosine similarity. In contrast, linear and matrix-completion approaches often yield smoother curves initially, but once they can no longer capture the finer structures, they can display sudden collapses in alignment for entire ranges of components.

Moreover, numerical instabilities in SVD computations become more pronounced when eigenvalues or singular values are small. If several singular values are nearly degenerate, their ordering can vary slightly depending on data splits, floating-point precision, or iterative solver conditions. Such variations can translate to apparent jumps or dips in the cosine similarity measure from one principal component to the next. Even minor arithmetic differences in floating-point operations may reorder these low-energy components, thus fluctuating how cumulative similarity accumulates.

In addition, the wafer data may itself be segmented by partially disjoint modes—e.g., certain wafer layers, distinct tool calibrations, or localized anomalies. If a model captures one of these modes well but only partially addresses a second mode, it might sustain a relatively smooth trace until a principal component suddenly emphasizes features from the underrepresented mode. The similarity trace then dives, only to rebound if subsequent components refocus on a domain the model handles better.

In practical terms, such oscillations do not necessarily invalidate a model's overall success; rather, they signal that a precise, uniform reconstruction of lower-variance features can remain challenging. Modest deviations in those minor modes might be acceptable, so long as the model accurately recovers the major principal components dictating wafer yield and overlay alignment. However, if those finer components correspond to critical but spatially limited wafer features, abrupt dips in similarity may signal a need for more targeted refinement, such as further domain-based constraints. Understanding these reasons behind the jagged pattern helps clarify where further methodological improvements, data expansions, or domain knowledge could shore up the reconstruction of subtle wafer behaviors.

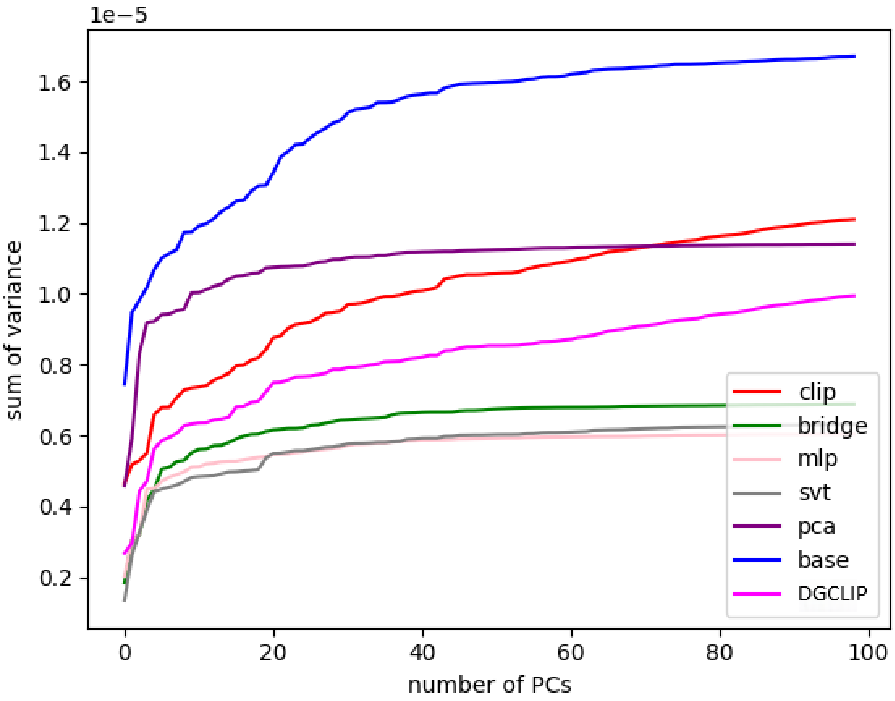


Figure 14: Cumulative difference in explained variance between processed imputed datasets and the target dataset, as a function of the number of components considered.

In graph 14, each curve traces the cumulative *difference* in explained variance between the imputed dataset and the target dataset as we consider more principal components, so lower curves indicate better alignment with the target's overall variance distribution. Viewed in that light, several patterns emerge. The base curve, which indicates the baseline difference in explained variance between the raw primary and target dataset, sits highest across nearly all components. This means the uncorrected wafer data diverge significantly from the target's variance profile, as it retains the largest net discrepancy since none of the systematic offsets are removed. By contrast, methods whose curves remain lower across more principal components are capturing the target's global variance structure more effectively, introducing fewer distortions or imposing less smoothing.

Specifically, SVT consistently places near the bottom of the graph, particularly after about 10–20 components, which is closely followed and then slightly overtaken by the MLP, suggesting they both preserve a broad swath of the target's variance distribution with minimal excess or deficit. Likewise, DGCLIP remains relatively low compared to CLIP, reflecting that its domain constraints help it align with the target's smaller-scale components beyond the initial major modes. Meanwhile, the Bridge model, though not quite as good as MLP or SVT in absolute terms, still outperforms both DGCLIP and CLIP in matching the target variance. This improvement underscores the value of refining coarse pairings through a residual translator rather than relying on CLIP's one-step coarse imputation.

Like with the similarity trace plots, although both SVT and PCA are matrix-completion methods aimed at leveraging the data's low-rank structure, their curves in this variance-difference plot deviate considerably. Notably, SVT's line hovers near the MLP curve over a substantial range of components, indicating that SVT recovers much of the same variance distribution as a deep neural model, despite relying on a more classical iterative shrinkage of singular values. By contrast, PCA's curve stands markedly higher—often more than twice that of SVT—signifying that its strictly linear projection leaves significant unexplained variance relative to the target for most components.

This disparity underscores the difference in how each method imposes low-rank constraints. PCA restricts the entire matrix to a small number of principal directions preselected by global variance, which can discard local or high-order details if they lie outside those major axes. SVT also iteratively shrinks singular values but retains flexibility in how many singular modes survive the thresholding process, enabling it to adapt better to moderate- or mid-range components that still contain meaningful wafer variability. Consequently, SVT manages to align more closely with the target across the whole principal-component spectrum in both the similarity trace and explained variance plots, resulting in a curve that nearly tracks a deeper neural approach like MLP, even if it lacks the same expressive power or local modeling capacity. Meanwhile, PCA imposes a rigid cutoff that captures coarse structure but misses a broader set of subtle modes, causing its variance-difference line to remain comparatively higher throughout.

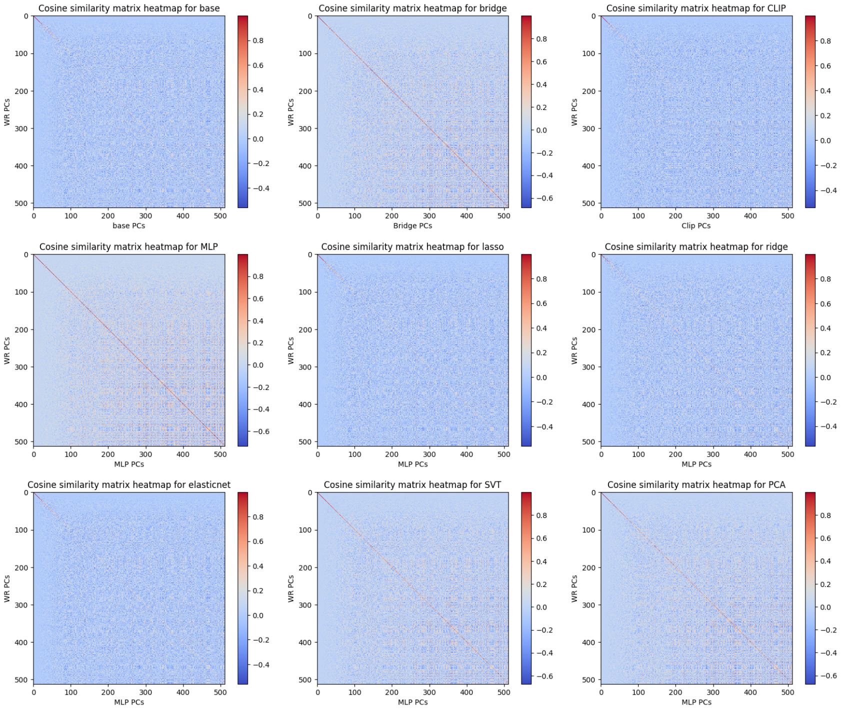


Figure 15: Combined heatmaps of the cosine similarity matrix between each processed imputed set and target set

From these nine heatmaps in 15, each displaying the cosine similarity between the imputed dataset's principal components (horizontal axis) and the target dataset's principal components (vertical axis), one can gauge how well each method's extracted modes correspond (or fail to correspond) to the target's. A dominant, bright diagonal typically indicates a strong one-to-one match of components. By contrast, faint or scattered off-diagonal structures suggest that the method's principal components diverge from the target's ordering or shape.

Looking at the base' dataset (top-left), the similarity matrix mostly exhibits diffuse, low-level color, punctuated by only mild diagonal banding. This confirms that the raw primary measurements have only a modest overlap with the target's component structure: they do not align well beyond the leading modes, and the diagonal is faint overall. Each of the linear regressions (middle-center, middle-right, bottom-left) and CLIP (top-right) show no improvement in aligning the principal components over the raw primary (base) set.

When we shift to Bridge (top-center), we observe a significantly more pronounced diagonal streak than CLIP: its diagonal band is considerably more coherent, reflecting that localized residual refinements reduce mismatches in the mid-range modes. To a slightly greater extent than bridge, the MLP heatmap (middle-left) shows one of the most distinctive diagonal lines, indicating that this deeper neural approach reproduces the target's component directions more consistently across the entire rank spectrum. Some faint cross-structures exist in both bridge and the MLP, but they are overshadowed by the comparatively bright main diagonal, implying the their reconstructions systematically match or align each principal component with the target.

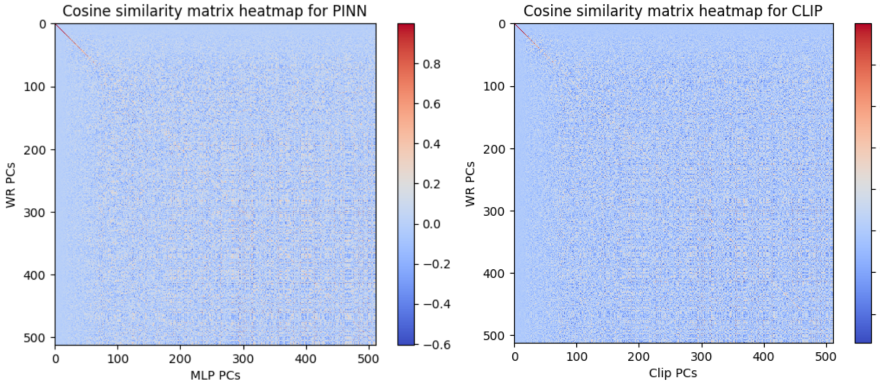


Figure 16: Cosine similarity matrix heatmap for DGCLIP compared to CLIP (right)

Figure 16 gives the heatmap of DGCLIP side by side with CLIP for more fine comparison. As mentioned, we see that CLIP (right) features a faint, discontinuous diagonal mirroring the base heatmap, indicating it struggles to align many of its higher-order principal components with those of the target. By contrast, DGCLIP (left) exhibits a more pronounced diagonal band, suggesting a more robust one-to-one correspondence in both early and moderate-ranked principal components. Though still not nearly as sharp as in some advanced neural architectures, DGCLIP's improvement over CLIP is visible in the diminished expanses of purely blue cells and a more continuous diagonal gradient, reflecting how domain guided overlay constraints steer the model away from purely data-driven mismatches. Effectively, DGCLIP preserves better alignment with the wafer dataset's deeper modes, slightly reducing the random or unphysical alignments that CLIP alone might produce after capturing the main variance directions.

For matrix completion, both give a strong diagonal band indicating good alignment, though SVT (bottom-center) shows a clearer band than PCA (bottom-right). SVT's diagonal remains fairly consistent, indicating an impressive alignment with the target's principal components across much of the rank, echoing the earlier observation that it can adapt to mid-range modes better than PCA. By contrast, PCA's diagonal is present yet notably softer and more disrupted by off-diagonal speckling, suggesting the global directions it extracts diverge more from the target's.

Altogether, these heatmaps clarify how each method's principal components match (or mismatch) the target's. Neural (barring CLIP) or refined matrix-completion methods typically present a stronger, more continuous diagonal, reflecting more faithful reconstructions. Simpler linear regressions exhibit faint or inconsistent diagonals with no improvement upon the baseline, confirming they do not approximate the target's underlying structure beyond the leading few directions.

Summary of Principal-Component Analyses Results The base curve, high in cumulative difference for both principal-component measures, reaffirms how simply ignoring secondary adjustments leaves the wafer data significantly off-target in deeper variance directions. Bringing together the observations from both the cosine similarity traces (Figure 13) and the explained-variance differences (Figure 14), we see that each imputation method exhibits distinct strengths and weaknesses depending on the scope of variance considered. In the early principal components—where the wafer data's largest-scale differences reside—nearly all methods perform decently, indicating that the dominant, rank-deficient structure is not overly challenging. Linear approaches match this leading variance about as well as more complex models, reflecting that low-frequency or coarse-level discrepancies can be approximated with simple global regressions.

However, as we look beyond approximately 10 to 20 principal components, several methods separate themselves by continuing to capture the target dataset's mid- and higher-order modes. The MLP consistently retains alignment with subtle wafer features, as suggested by its high cosine similarity in the tail-end of Figure 13 and correspondingly low difference in explained variance in Figure 14. Bridge and DGCLIP also fare better in these mid-tier modes than baseline CLIP, underscoring that local residual corrections (Bridge) or physically informed overlay constraints (DGCLIP) can significantly reduce the mismatch left unaddressed by contrastive alignment alone. By contrast, the base CLIP approach tends to diverge more strongly in higher components, illustrating that coarse pair retrieval alone can undershoot the finer nuances of wafer data.

Matrix completion methods provide an interesting comparison. SVT surprisingly tracks near the MLP in terms of preserving overall variance for a wide range of components, suggesting iterative singular-value thresholding can adapt well to moderate or mid-level complexities. PCA completion, however, stands out for lagging behind: while it may handle the main low-rank structure, its rigid choice of principal axes can omit numerous localized or higher-order signals, leaving a noticeably larger unexplained variance gap relative to the target.

Notably, many models' traces become more jagged in the higher components, an effect amplified by small singular values and the inherent noise or partial coverage in the wafer data. Minor overfitting to specific local patterns may yield momentary peaks in cosine similarity, only for subsequent principal components to reveal unmodeled variability. Although these dips do not invalidate the models' overall gains, they indicate areas where refined domain knowledge, improved regularization, or more specialized model architectures might be needed.

Taken together, these observations clarify two overarching themes. First, virtually all methods handle the largest principal components of wafer reflectivity data with relative ease; it is the secondary or tertiary modes that truly distinguish advanced methods from simplistic ones. Second, simply aligning primary–secondary pairings (as in unrefined CLIP) can leave considerable mid-range structure unmodeled, while bridging or domain-constrained approaches pick up much of this slack. However, the upshot for practical metrology is that while purely linear regressions entirely risk missing precisely those subtleties that can matter for fine overlay alignment, matrix-completion methods with linear low-rank assumptions, while not quite as good as the MLP, appear to adapt flexibly to the wafer's underlying complexity—and might be more suitable if deep learning approaches introduce too much training overhead or hyperparameter tuning needs.

4.1.4 Overlay and T2T Inference Error

Dataset	MSE Loss	T2T Loss
primary	1.00	1.00
error corrected	0.908	0.734
CLIP	0.919	1.15
DGCLIP	0.990	0.924
Bridge	0.913	1.06
MLP	0.990	0.924
SVT	0.996	1.08
PCA	1.00	1.11
Lasso	1.03	1.09
Ridge	1.04	1.12
ElasticNet	1.02	1.40

Table 8: Comparison of overlay MSE loss and overlay T2T Loss from training and testing the overlay inference network on the primary dataset, error corrected dataset, and each methods imputed dataset, relative to the model performance on the primary input dataset.

Table 8 gives the overlay and T2T prediction performances for the overlay network trained on each methods imputed dataset. The primary dataset, representing uncorrected primary measurements, sets both loss benchmarks at 1.00. By comparison, training on the actual error-corrected dataset (primary data corrected by the observed target secondary data) yields MSE ≈ 0.908 and T2T ≈ 0.734, giving the best-case scenario for training on our imputation sets, i.e. a 9.2% and 26.6% improvement respectively. A quick scan of the table shows that no single model simultaneously minimizes both overlay mean-squared error (MSE) loss and tool-to-tool (T2T) loss significantly.

Among the novel methods, CLIP and bridge reduce the MSE to 0.919 and 0.913 but raise T2T to 1.15 and 1.06 respectively, indicating that while it partially corrects primary–secondary mismatches, it fails to maintain consistent overlay estimates across different tools and might suggest overfitting to over-represented tools in the dataset. By contrast, DGCLIP (0.990 MSE, 0.924 T2T) focuses more on preserving overlay coherence, so although its raw MSE is not as impressive, it yields significantly better T2T performance relative to CLIP. Bridge (0.913 MSE, 1.06 T2T) simultaneously improves MSE over CLIP while tempering T2T error somewhat (though not below the primary dataset's baseline). The MLP, interestingly, hits 0.990 in MSE, the same as DGCLIP, and matches DGCLIP's 0.924 T2T, suggesting that a sufficiently tuned global neural-network regression can also deliver overlay-friendly imputation.

Linear baselines provide mixed outcomes, though generally underperform, with the linear regression methods negatively impacting both overlay and T2T inference. Both matrix completion methods (0.996 MSE, 1.08 T2T for SVT) (1.00 MSE, 1.11 T2T for PCA) similarly prove unhelpful, though to much a lesser degree in T2T.

The final upshot is that though matrix completion methods succeeded in good reconstruction error and principal component reconstruction, subtle nonlinear variations likely carry on to downstream tasks like overlay prediction, necessitating nonlinear models. Moreover, different use cases may favor different trade-offs. Specifically, DGCLIP and the MLP stand out if T2T matching is paramount. CLIP-based or Bridge methods yield a better MSE than DGCLIP but may fall short in T2T consistency—unless additional domain constraints are introduced. Furthermore, CLIP-based models maintain the simplest network architecture and are among the lowest training times of the advanced methods. Compared to the MLP, both DGCLIP and bridge utilize one entire layer less, with uniform use of the tanh activation, which altogether is a significant portion of reduced complexity. The MLP hits a middle ground, retaining physically plausible overlay error while reducing MSE effectively, though at a cost to complexity. Ultimately, each approach strikes its own balance between complexity, pure numerical fidelity, and physically coherent tool-to-tool predictions, reinforcing that overlay-cognizant regularization is essential if wafer metrology requires more than just raw reconstruction accuracy. Overall, however, earlier successes in reconstruction do not seem to carry on as strongly downstream, as these methods appear to largely fall short in capturing a considerable enough portion of the tool asymmetries to reduce T2T errors significantly.

4.2 Limitations and Challenges of the Study

Although this research presents promising approaches for imputing high-dimensional wafer metrology data under Missing By Design conditions, several limitations and challenges remain. These issues reflect constraints in data availability, modeling assumptions, computational resources, and domain-specific considerations, all of which can shape future extensions and refinements of our methods.

A central limitation arises from the restricted scope and characteristics of the dataset itself. While the experiments leveraged a large number of wafer measurements, the overall diversity and representativeness of these samples may not fully reflect the breadth of possible manufacturing conditions. Notably, our data primarily focus on a single process flow with limited variation in wafer types and tool settings. In industrial practice, wafer data can evolve over time or may exhibit seasonal machine drifts, necessitating repeated retraining or more adaptive models. Hence, the techniques proposed might see their performance degrade if encountered with wafer conditions that lie significantly outside the training distribution, especially for neural or contrastive approaches without explicit outlier detection or adaptation modules.

Additionally, the assumption of rank-deficiency and moderately smooth error landscapes—e.g., as exploited by low-rank or matrix-completion methods—may not hold universally across all operational contexts. Although principal component analyses indicate that broad-scale wafer differences lie in a few principal directions, localized or nonlinear anomalies can still challenge simpler matrix-completion techniques. At the same time, purely data-driven neural networks (MLP or CLIP-based) must rely on sufficient paired data to learn these intricate patterns. Under real production constraints, data may be significantly omitted, leading to partial coverage of important operating regimes. Models with high capacity (e.g., deeper MLPs) risk overfitting when unpaired data far outstrips the fully observed pairs, while CLIP-based retrieval, though a much more conservative approach that could fair better than MLPs, might still be misled if only a narrow portion of the secondary space is well-sampled. Consequently, the approach's generalizability hinges on obtaining a sufficiently large, well-distributed training set to capture the full variability of wafer measurements, tool calibrations, and process conditions.

The Bridge Model—while elegantly splitting imputation into coarse retrieval and local refinement—also depends strongly on the quality of initial anchors. If the underlying CLIP retrieval is systematically flawed for certain wafer categories or fails under certain machine conditions, the Bridge approach can inherit these defects.

Moreover, integrating domain knowledge—such as overlay constraints through DGCLIP-style regularization—brought observable benefits in preserving physically coherent tool-to-tool alignment, but this approach adds further complexity to training and tuning. Determining the exact balance between numerical fidelity (e.g., in MSE or cross-entropy losses) and physical plausibility (in T2T overlay consistency) remains nontrivial, subject to trial-and-error in picking scale parameters for overlay or other physical constraints. Overly aggressive domain regularization might overshadow the model's ability to capture subtle data-driven correlations, whereas insufficient regularization fails to enforce domain imperatives effectively.

A further limitation lies in the computational overhead for certain methods. MLP training can be expensive to tune thoroughly, given a wide hyperparameter space that includes layer depth, activation choice, and learning rates. Contrastive methods (CLIP or Bridge) also require large-batch computations of similarity matrices, which can exceed memory limits for extremely high-dimensional data or large wafer-lot sizes. In production contexts, real-time or near-real-time constraint may be crucial, limiting the feasibility of approaches with heavy retrieval or repeated forward passes.

Lastly, interpretability and explainability remain ongoing concerns, especially for deeper neural or contrastive approaches. The bridging step and DGCLIP-based domain constraints do help clarify how certain physics-based or anchor-based corrections arise, but a full industrial deployment might demand a more formal interpretability framework. For instance, engineers might require direct tracing of which wafer features contributed most to an imputation or an overlay correction. Methods like post-hoc saliency or attention analysis can shed light on these decisions, but this thesis primarily focused on predictive performance rather than thorough interpretability workflows.

In summary, the study's constraints reflect the realities of a single dataset, limited exploration of potential out-of-distribution regimes, the inherent complexities in model tuning and domain integration, and the computational and interpretability challenges of scaling high-capacity models. Addressing these gaps—through larger, more diverse data, more adaptive or incremental learning, more robust domain constraints, and refined interpretability tools—constitutes natural directions for future research. By tackling these limitations, subsequent work can bolster the practical viability of the presented imputation models, paving the way for more reliable, efficient, and explanatory solutions in the evolving landscape of semiconductor wafer metrology.

4.3 Applications and Further Research

The imputation strategies explored in this thesis provide a versatile foundation for addressing Missing By Design scenarios in semiconductor metrology while also offering pathways for extension into other high-dimensional, systematically missing data contexts. In metrology specifically, the central application lies in the ongoing drive to minimize measurement overhead while preserving high-accuracy overlay predictions. By substituting costly or time-consuming secondary measurements with imputed values through neural-based methods—fabs can reduce wafer processing times and more flexibly allocate measurement resources. The contrastive approach (such as CLIP or a DGCLIP-augmented CLIP) can also facilitate new procedures in metrology by matching newly captured primary measurements to historical secondary points, creating a kind of multisensor memory that captures long-term variability across different production runs or tool calibrations. Furthermore, overlay correction tasks can be strengthened by embedding domain knowledge (e.g., DGCLIP regularization) that pushes the model to produce physically coherent solutions, thus mitigating the risk of large systematic offsets that degrade yield or tool-to-tool consistency.

Beyond semiconductor fabrication, many industrial domains could adopt these methods when data collection is deliberately sparse or unevenly distributed. For instance, aerospace maintenance might omit selected sensor channels to reduce instrumentation load, while healthcare applications (in the scope of tabular data) often forego certain expensive or invasive tests. Neural contrastive approaches like CLIP can handle a variety of numerical data modalities by first building a shared embedding space that readily identifies matching or similar instances, as we've shown in this study. In such settings, an additional Bridge-style residual correction can further handle local divergences arising from patient-specific or part-specific anomalies that are not reflected in coarse or historical matches. Likewise, if domain physics or safety constraints exist, DGCLIP-like regularizations can force imputed values to obey known laws of motion, fluid flow, or mechanical stress, complementing purely data-driven alignment.

Looking ahead, an important area for further research involves building more adaptive frameworks and missingness designs that dynamically select which measurement sites or features to collect. By integrating active learning or Bayesian optimization techniques, one could direct the metrology system to measure only those wafer sites whose secondary signals are most critical or least predictable from prior data. In doing so, the model's uncertainty estimates become a deciding factor in measurement scheduling, requiring careful attention to the MBD mechanism. Another avenue is refining the overlay-based regularization to account for more complex physical relationships—perhaps modeling wafer warpage or thermal effects through PDE-based constraints, considering wafer stack parameters other than overlay, or by introducing more structured domain embeddings. This would better capture the reality that errors often cluster by wafer region, layer, or processing batch. Along similar lines, a deeper synergy between CLIP embeddings and physically motivated priors could be explored. Rather than simple scalar weighting in the DGCLIP approach, one might incorporate physically interpretable latent variables that encode, for example, known wafer geometry.

In the realm of neural architectures, the methods presented—MLPs, Bridge, and contrastive encoders—do not exhaust the spectrum of possible deep learning designs. Convolutional or attention-based layers could prove more robust when measurement locations follow a spatial arrangement or when certain channels of the primary–secondary data correlate in structured patterns. Graph-based models might also be relevant if wafer sites are connected in some topological or process-driven network. For instance, a graph neural network might encode adjacency relations between different wafer regions, helping the imputation process maintain smooth transitions or detect abrupt layer boundaries. Each of these architectures could also be explored as both direct applications or as the encoders of the CLIP model, instead of MLPs. Meanwhile, the Bridge approach could be extended into multi-bridge systems, where different residual translators operate over various subdomains or sensor types, eventually merging local refinements into a unified final imputation.

Moreover, given its success with CLIP, another next step is to apply domain-guided constraints to the other models, particularly to the bridge model (as it is partially composed by CLIP). Likewise, varying the degree of missingness would allow more thorough analyses of few-shot capabilities for the tested methods. This was not included in the scope of the thesis due to time constraints, as it would necessitate optimizing each model for each tested degree of missingness to preserve ignorability conditions.

Returning to the literature review, the Bayesian marginalization and k-NN frameworks introduced were not applied in the scope of this study due to time constraints. Exploring k-NN would be valuable to evaluate the effectiveness of contrastive learning with CLIP versus classical approaches, considering that CLIP directly employs the k method in its imputation mechanism and can be viewed as a contrastive, deep learning extension of it. Bayesian marginalization was found to be highly complex when scaling to both the size and dimension of the metrology dataset, and proved difficult when managing computational resources and deciding on priors. Though this method was abandoned in lieu of time, it is indeed possible and certainly an area of research that should be undertaken for future study in metrology data imputation.

An additional line of inquiry involves uncertainty quantification. While the present models yield point estimates of missing complements, future solutions might include Bayesian layers or Monte Carlo dropout to provide confidence intervals for each imputed measurement. Such uncertainties could guide process engineers in deciding whether to trust an imputation or instead perform a physical measurement. Tying these uncertainties back into an adaptive measurement system closes the loop, so that the model actively requests real secondary data whenever its imputation variance crosses a reliability threshold. This can be especially valuable in high-stakes metrology, where critical yield decisions hinge upon measurement precision. The question of maintaining interpretability remains salient: domain experts need to diagnose why an imputation might fail for a given wafer region, an explanation that can emerge only if each modeling stage (coarse retrieval, residual translation, domain constraints) is sufficiently transparent and traceable.

Finally, the capacity to transfer or generalize these approaches beyond the single dataset used in this thesis stands as a test of scalability. Given that wafer reflectivity and overlay errors often share certain rank-deficient or repetitive features, the studied methods might carry over smoothly to other layers or wafer designs with minimal re-calibration. Yet transferring them to a domain with drastically different measurement physics may require rethinking the underlying assumptions or adjusting how domain knowledge is embedded in the model. Comprehensive multi-site or multi-tool experiments would verify whether the same bridge concept—coarse instance matching plus localized residual translation—persists as a robust pattern across varied industrial lines. This is an important next step for research in further realizing the potential of contrastive frameworks in industrial data imputation, particularly for CLIP (and by extension bridge), as it is a conservative model explicitly designed for few-shot data settings.

Thus, while the thesis introduces a rich suite of techniques for MBD imputation in semiconductor metrology, it likewise opens numerous avenues for enhancements. These include broadening data coverage, deepening the integration of physical laws, refining the bridging approach for more dynamic or multi-constraint contexts, varying degrees of missingness, adding uncertainty estimates, and improving computational efficiency for real-time requirements.

Beyond the scope of metrology, bridge might also present opportunities in the original language-image data setting where CLIP is derived. A potential extension is to simultaneously train a decoder network for the language encoder, which would project latent bridged imputations into new textual captions outside our bag of known labels. This would constitute translating the pixel differences between a new image and a similar known image (paired by CLIP) into semantic differences that adjust the caption of the similar known image.

5 Conclusion

This thesis investigated how multiple classes of discriminative imputation methods—spanning regularized linear regressions, matrix-completion frameworks, and advanced machine-learning approaches—can reconstruct omitted secondary signals in wafer metrology to reduce systematic machine errors and improve overlay accuracy. By grounding our inquiry in a Missing By Design context, we highlighted the practical and industrially driven rationale for deliberately skipping measurements in high-dimensional data settings. Here, we reflect on the primary research question and its associated sub-questions, synthesizing the empirical results and methodological lessons in light of the goals of wafer metrology data imputation.

Primary Research Question: *How effectively do different discriminative imputation methods, including classical, statistical, and machine learning based methods, reconstruct missing secondary signals in wafer metrology data to reduce systematic machine errors and improve overlay prediction accuracy?*

Our results confirm that, although straightforward linear regressions (Ridge, Lasso, ElasticNet) are computationally lightweight, they struggle to capture the strongly nonlinear relationships between primary and complementary signals. By contrast, matrix completion (PCA, SVT) improved upon these linear baselines by enforcing low-rank constraints, halving reconstruction errors and aligning principal components better than baseline, yet could not resolve local or higher-order wafer features.

Neural approaches outperformed simpler methods in both raw imputation error and principal component alignment—namely the MLP, which excelled in reconstructing subtle wafer variations, and the CLIP-based models augmented by domain constraints or localized residuals. The overlay inference experiments clarified that introducing domain knowledge (e.g., via an overlay-focused regularization as in DGCLIP) allows the model to reduce tool-to-tool (T2T) errors, thus promoting physically consistent corrections. This domain guided approach echoes a broader industrial tension: purely minimizing numeric reconstruction error does not necessarily yield the best overlay alignment across multiple tools, underscoring the importance of embedding operational objectives within the learning pipeline.

SQ1: *What is the relative performance of regularized linear models (Lasso, Ridge, Elastic Net) and matrix-completion methods (PCA, SVT) in imputing missing secondary data, and do they sufficiently capture systematic error signals?*

We found that regularized linear regressions notably lagged behind matrix completion methods, significantly exceeding the baseline errors considered. Despite their simplicity and negligible training time, linear regression proved entirely inadequate to capture wafer reflectivity's rank-deficient structure fully. Although this did not carry to overlay tasks, matrix completion methods successfully addressed global, rank-dominant behaviors in the data, capturing significant variations in the principal components. This could suggest that nonlinear adaptations of matrix completion methods like kernelized SVT would allow better performance downstream.

SQ2: *How robust are these imputation methods with respect to key metrics in semiconductor manufacturing, such as systematic error reconstruction, overlay prediction improvement, and tool-to-tool (T2T) consistency?*

While nearly all methods captured some fraction of the wafer's dominant low-rank structure, their performance varied widely in overlay and T2T tests. Matrix completion methods, even if they cut reconstruction error, produced higher T2T error for the overlay prediction network than leaving the dataset uncorrected, implying poor cross-tool consistency. Some purely data-driven neural methods (bridge and CLIP) lowered reconstruction and overlay prediction errors, but did not always achieve strong T2T performance without additional constraint and sacrificing moderate overlay accuracy. DGCLIP, by contrast, embedded a physics-informed overlay constraint that improved T2T consistency, albeit at a minor cost to absolute MSE compared to CLIP. The MLP performed consistently the best at reconstruction metrics, though was equal to DGCLIP ultimately at downstream overlay tasks. These differences illustrate the trade-off between numerical fidelity and physically coherent transformations: in production settings, an imputation method that preserves overlay alignment across multiple machines can be more valuable than one that merely minimizes raw error. Lastly, the difference in complexity, namely the MLPs comparatively larger architecture to DGCLIP, necessitates further testing for generalization to other datasets and production lines to draw any concrete conclusions.

SQ3: *How can the CLIP framework be adapted to better suit continuous value imputation, and to what extent can it account for underlying physical data relationships between primary and secondary data pairs?*

We addressed CLIP's coarse, uninformed imputations by two main adaptations. First, the Bridge approach introduced a residual translator that refines CLIP's retrieved secondary measurements, capturing local deviations otherwise missed by a single step of contrastive alignment. This improved the alignment to finer wafer modes while keeping training overhead modest. Second, DGCLIP wove domain knowledge directly into CLIP's loss function, ensuring physically valid overlay consistency. Although neither adaptation fully overtook the MLP's raw reconstruction accuracy, each overcame some of CLIP's mid-range and overlay shortcomings. In principle, the results suggest that marrying bridge models with multi-constraint domain-guided modules could further steer the model away from unphysical pairings, indicating ample room for future research.

SQ4: *What lessons can be distilled into practical guidelines for the semiconductor industry regarding the choice, design, and deployment of imputation strategies to minimize measurement overhead while preserving wafer metrology quality?*

Several practical guidelines emerged. First, the design of missingness in the operational flow directly influences the methods available for imputation, and must be carefully chosen to maintain ignorable conditions. Second, linear methods including regressions and matrix completion are inadequate to capture the necessary subtleties of metrology data for accurate overlay prediction and T2T alignment, strongly implying nonlinearities in the data that necessitate more advanced methods. Third, deeper neural regressions can capture these complex nonlinear structures but demand greater attention to hyperparameter tuning and remain yet to be tested for generalization on multi-line, multi-tool datasets. Fourth, contrastive-learning frameworks might not work without follow-up refinement or domain constraint, which provided quantifiable improvements to tool asymmetry correction.

Overall, these findings highlight the importance of adapting to the wafer data's strong rank deficiency, nonlinearity, local variations, and domain constraints when designing MBD imputation solutions in semiconductor metrology.

Disclaimer

The contents of this report were written with the help of LLMs for spelling, grammar, and paraphrasing.

A Imputation Method Algorithms

Algorithm 1 Linear Regression

```
Require: Observed primary measurements $\mathbf{X}^{(P)} \in R^{N \times p}$,
  Observed secondary measurements $\mathbf{X}^{(S)} \in R^{N \times p}$,
  Unpaired primary measurements $\mathbf{X}^{(P)}_{\text{unpaired}} \in R^{M \times p}$,
  Regularization parameter $\lambda > 0$
Ensure: Predicted secondary measurements $\hat{\mathbf{X}}^{(S)}$
  Compute $\mathbf{A} = {\mathbf{X}^{(P)}}^T \mathbf{X}^{(P)} + \lambda \mathbf{I}$
  Compute $\mathbf{b} = {\mathbf{X}^{(P)}}^T \mathbf{X}^{(S)}$
  Solve $\hat{\beta} = \mathbf{A}^{-1}\mathbf{b}$
  Predict secondary measurements: $\hat{\mathbf{X}}^{(S)} = \mathbf{X}^{(P)}_{\text{unpaired}} \hat{\beta}$
  return $\hat{\mathbf{X}}^{(S)}$
```

Algorithm 2 Principal Component Analysis (PCA) Algorithm

```
Require: Paired primary measurements $\mathbf{X}^{(P)} \in R^{N \times p}$,
Paired secondary measurements $\mathbf{X}^{(S)} \in R^{N \times p}$,
Unpaired primary measurements $\mathbf{X}^{(P)}_{\text{unpaired}} \in R^{M \times p}$,
Tolerance $\epsilon > 0$
Ensure: Imputed secondary measurements $\hat{\mathbf{X}}^{(S)}_{\text{unpaired}}$
Concatenate paired primary and secondary measurements:
    $$\mathbf{X}_{\text{paired}} = \begin{bmatrix} \mathbf{X}^{(P)} & \mathbf{X}^{(S)} \end{bmatrix}.$$
Concatenate unpaired primary measurements with zeros for missing secondary measurements:
    $$\mathbf{X}_{\text{unpaired}} = \begin{bmatrix} \mathbf{X}^{(P)}_{\text{unpaired}} & 0 \end{bmatrix} \in R^{M \times 2p}.$$
Initialize combined dataset:
    $$\mathbf{X}_0 = \begin{bmatrix} \mathbf{X}_{\text{paired}} \\ \mathbf{X}_{\text{unpaired}} \end{bmatrix}.$$
Set iteration counter $k = 0$
repeat
    Perform PCA on $\mathbf{X}_k$ to compute:
        $$\mathbf{X}_k \approx \mathbf{U}\Lambda\mathbf{V}^T$$
    Reconstruct the dataset:
        $$\mathbf{X}_{k+1} = \mathbf{U}\Lambda\mathbf{V}^T$$
    Enforce consistency with observed data:
        $$(\mathbf{X}_{k+1})_{ij} = (\mathbf{X}_0)_{ij}, \quad \forall (i,j) \in \Omega,$$
    where $\Omega$ represents the observed entries in $\mathbf{X}_0$, i.e., $\mathbf{X}^{(P)}, \mathbf{X}^{(S)}, \mathbf{X}^{(P)}_{\text{unpaired}}$.
    Update iteration counter: $k \leftarrow k + 1$
until Convergence criterion met:
    $$\|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F \leq \epsilon$$
Extract imputed secondary measurements:
    $$\hat{\mathbf{X}}^{(S)}_{\text{unpaired}} = (\mathbf{X}_{k+1})_{N+1:N+M,p+1:2p}.$$
return $\hat{\mathbf{X}}^{(S)}_{\text{unpaired}}$
```

Algorithm 3 Singular Value Thresholding (SVT) imputation

```
Require: Observed primary measurements $\mathbf{X}^{(P)} \in R^{N \times p}$,
    Observed secondary measurements $\mathbf{X}^{(S)} \in R^{N \times p}$,
    Unpaired primary measurements $\mathbf{X}^{(P)}_{\text{unpaired}} \in R^{M \times p}$,
    Threshold $\tau > 0$, tolerance $\epsilon > 0$
Ensure: Predicted secondary measurements for unpaired primary measurements $\hat{\mathbf{X}}^{(S)}_{\text{unpaired}}$
Concatenate paired primary and secondary measurements:
    $\mathbf{X}_{\text{paired}} = \begin{bmatrix} \mathbf{X}^{(P)} & \mathbf{X}^{(S)} \end{bmatrix} \in R^{N \times 2p}.$
Concatenate unpaired primary measurements with zeros for missing secondary measurements:
    $\mathbf{X}_{\text{unpaired}} = \begin{bmatrix} \mathbf{X}^{(P)}_{\text{unpaired}} & 0 \end{bmatrix} \in R^{M \times 2p}.$
Combine paired and unpaired data:
    $\mathbf{X}_0 = \begin{bmatrix} \mathbf{X}_{\text{paired}} \\ \mathbf{X}_{\text{unpaired}} \end{bmatrix} \in R^{(N+M) \times 2p}.$
Set iteration counter $k = 0$
repeat
    Compute SVD of $\mathbf{X}_k$:
        $\mathbf{X}_k = \mathbf{U} \Sigma \mathbf{V}^T,$
    where $\mathbf{U} \in R^{(N+M) \times r}$, $\Sigma \in R^{r \times r}$, $\mathbf{V} \in R^{2p \times r}$
    Apply soft-thresholding to singular values:
        $\Sigma' = \max(\Sigma - \tau, 0)$
    Reconstruct $\mathbf{X}_{k+1}$:
        $\mathbf{X}_{k+1} = \mathbf{U} \Sigma' \mathbf{V}^T$
    Enforce consistency with observed data:
        $(\mathbf{X}_{k+1})_{ij} = (\mathbf{X}_0)_{ij}, \quad \forall (i, j) \in \Omega,$
    where $\Omega$ represents the observed entries in $\mathbf{X}_0$, i.e., $\mathbf{X}^{(P)}, \mathbf{X}^{(S)}, \mathbf{X}^{(P)}_{\text{unpaired}}$.
    Update iteration counter: $k \leftarrow k + 1$
until Convergence criterion met:
    $\|\mathbf{X}_{k+1} - \mathbf{X}_k\|_F \leq \epsilon$
Extract imputed secondary measurements:
    $\hat{\mathbf{X}}^{(S)}_{\text{unpaired}} = (\mathbf{X}_{k+1})_{N+1:N+M, p+1:2p}.$
return $\hat{\mathbf{X}}^{(S)}_{\text{unpaired}}$
```

Algorithm 4 MLP Training and Imputation

```
Require: Paired primary and secondary measurements $\mathbf{X}^{(P)}_{\text{paired}} \in R^{N \times p}$, $\mathbf{X}^{(S)}_{\text{paired}} \in R^{N \times p}$,
Unpaired primary measurements $\mathbf{X}^{(P)}_{\text{unpaired}} \in R^{M \times p}$,
Learning rate $\eta > 0$, number of epochs $T$
Ensure: Imputed secondary measurements $\hat{\mathbf{X}}^{(S)}_{\text{unpaired}}$
Normalize $\mathbf{X}^{(P)}_{\text{paired}}$ and $\mathbf{X}^{(P)}_{\text{unpaired}}$ for feature scaling
for each epoch $t = 1, \ldots, T$ do
    Forward Pass:
        • Compute activations for each hidden layer:
              $\mathbf{h}_i = \sigma(\mathbf{W}_i \mathbf{h}_{i-1} + \mathbf{b}_i), \quad \text{for } i = 1, \ldots, L,$
          where $\mathbf{h}_0 = \mathbf{X}^{(P)}_{\text{paired}}$.
        • Compute output:
              $\hat{\mathbf{X}}^{(S)}_{\text{paired}} = \mathbf{W}_L \mathbf{h}_L + \mathbf{b}_L.$
    Compute Loss:
        $\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{X}_{\text{secondary},i\text{paired}} - \hat{\mathbf{X}}_{\text{secondary},i\text{paired}} \right\|_2^2$
    Backward Pass:
        • Compute gradients of the loss with respect to each parameter using backpropagation.
        • Update parameters:
              $\mathbf{W}_i \leftarrow \mathbf{W}_i - \eta \nabla \mathbf{W}_i, \quad \mathbf{b}_i \leftarrow \mathbf{b}_i - \eta \nabla \mathbf{b}_i.$
end for
Use the trained MLP to predict secondary measurements for unpaired primary measurements:
    $\hat{\mathbf{X}}^{(S)}_{\text{unpaired}} = f_{\text{MLP}}(\mathbf{X}^{(P)}_{\text{unpaired}}).$
return $\hat{\mathbf{X}}^{(S)}_{\text{unpaired}}$
```

Algorithm 5 Adapted CLIP Model

```
Require: Training data $\mathcal{D}_{\text{train}}$, learning rate $\eta > 0$, temperature $t > 0$, number of epochs $T$
Ensure: Trained encoder networks $f_\theta$ and $g_\phi$
for epoch $t = 1, \ldots, T$ do
    Sample mini-batch $\{(\mathbf{x}_i^{(P)}, \mathbf{x}_i^{(S)})\}_{i=1}^N$ from $\mathcal{D}_{\text{train}}$
    Compute embeddings:
        $$\mathbf{z}_i^{(P)} = f_\theta(\mathbf{x}_i^{(P)}), \quad \mathbf{z}_i^{(S)} = g_\phi(\mathbf{x}_i^{(S)}), \quad \forall i \in \{1, \ldots, N\}$$
    Compute similarity matrix:
        $$\mathbf{S}_{ij} = \frac{\mathbf{z}_i^{(P)} \cdot \mathbf{z}_j^{(S)}}{\|\mathbf{z}_i^{(P)}\| \|\mathbf{z}_j^{(S)}\|}, \quad \forall i, j \in \{1, \ldots, N\}$$
    Compute contrastive loss:
        $$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} - \log \frac{\exp(\mathbf{S}_{ii}/t)}{\sum_{j=1}^{N} \exp(\mathbf{S}_{ij}/t)}$$
    Update parameters:
        $$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}, \quad \phi \leftarrow \phi - \eta \nabla_\phi \mathcal{L}$$
end for
return $f_\theta, g_\phi$
```

Algorithm 6 Bridge Model

```
Require: Training set $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i^{(P)}, \mathbf{x}_i^{(S)})\}$,
Pre-trained CLIP model, bridge model $h_\phi$, PCA transformation
Ensure: Refined secondary measurements $\hat{\mathbf{x}}_i^{(S)}$ for unseen primary measurements
Training Procedure:
for each $(\mathbf{x}_i^{(P)}, \mathbf{x}_i^{(S)}) \in \mathcal{D}_{\text{train}}$ do
    Generate initial imputation $\hat{\mathbf{x}}_i^{(S)}$ using the CLIP model
    Compute $\Delta\mathbf{x}_i^{(P)} = \mathbf{x}_i^{(P)} - \hat{\mathbf{x}}_i^{(P)}$
    Compute $\Delta\mathbf{x}_i^{(S)} = \mathbf{x}_i^{(S)} - \hat{\mathbf{x}}_i^{(S)}$
    Minimize the loss:
        $$\mathcal{L} = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \left\| \Delta\mathbf{x}_i^{(S)} - h_\phi(\Delta\mathbf{x}_i^{(P)}) \right\|_2^2$$
end for
return $\hat{\mathbf{x}}_i^{(S)} = \hat{\mathbf{x}}_i^{(S)} + \hat{\Delta}\mathbf{x}_i^{(S)}$
```

References

[1] E. Fix and J. L. Hodges. Discriminatory analysis: Nonparametric discrimination: Consistency properties. *International Statistical Review*, 57(3):238–247, 1989.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[3] D. P. Kingma and M. Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019.

[4] A. Radford, J. W. Kim, C. Hallacy, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[5] Y. Zhang, Y. Lu, X. Liao, L. Zhang, J. Huang, and E. Xing. Contrastive learning of medical visual representations from paired images and text. https://arxiv.org/abs/2010.00747, 2020. arXiv:2010.00747.

[6] Y. Zhang et al. Mammoclip: Contrastive learning with multi-modality views for mammogram analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 373–382, 2021.

[7] Y. Lei, Z. Li, Y. Shen, J. Zhang, and H. Shan. Clip-lung: Textual knowledge-guided lung nodule malignancy prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 403–412, 2023.

[8] C. A. Bode, B. S. Ko, and T. F. Edgar. Run-to-run control and performance monitoring of overlay in semiconductor manufacturing. *Control Engineering Practice*, 12(7):893–900, 2004.

[9] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* Wiley, 1987.

[10] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[11] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* Wiley, 2nd edition, 2002.

[12] Paul D Allison. Missing data. *The SAGE handbook of quantitative methods in psychology*, 23:72–89, 2009.

[13] Artur Pokropek. Missing by design: Planned missing-data designs in social science. *Ask: Research & Methods*, 20(1):81–105, 2011.

[14] [Author Name?] Graham. Planned missing data designs in psychological research. *International Journal of Behavioral Development*, 38(5):411–422, 2014. Check author names; Sage Publications, London, England.

[15] Paul J. Silvia, Thomas R. Kwapil, Molly A. Walsh, and Inez Myin-Germeys. Planned missing data designs in experience sampling research: Monte carlo simulations of efficient designs for assessing within-person constructs. *Behavior Research Methods*, 46(1):41–54, 2014.

[16] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[17] Fadil Santosa and William W Symes. Linear inversion of band-limited reflection seismograms. *SIAM journal on scientific and statistical computing*, 7(4):1307–1330, 1986.

[18] J. L. Schafer and J. W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002.

[19] Olga Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[20] Pablo J. García-Laencina, Jesús L. Sancho-Gómez, and André R. Figueiras-Vidal. Pattern classification with missing data: A review. *Neural Computing and Applications*, 19(2):263–282, 2010.

[21] I. T. Jolliffe and J. Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[22] Olga Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[23] Bernhard Schölkopf, Alex Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[24] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[25] Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[26] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, 1995.

[27] David M. Blei, Alireza Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[28] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[29] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[30] Brett K Beaulieu-Jones, Jason H Moore, and Pooled Resource Open-Access ALS Clinical Trials Consortium. Missing data imputation in the electronic health record using deeply learned autoencoders. In *Pacific symposium on biocomputing 2017*, pages 207–218. World Scientific, 2017.

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. https://arxiv.org/abs/1412.6980, 2014. arXiv:1412.6980.

[32] Wei-Chao Lin, Chih-Fong Tsai, and Jia Rong Zhong. Deep learning for missing value imputation of continuous data and the effect of data discretization. *Knowledge-Based Systems*, 239:108079, 2022.

[33] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[34] Yoshua Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[35] L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4037–4058, 2020.

[36] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1735–1742, 2006.

[37] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. https://arxiv.org/abs/1807.03748, 2018. arXiv:1807.03748.

[38] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[39] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, pages 1597–1607, 2020.

[40] Kaiming He, Hao Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[41] Tian Gao, Xu Yao, and D. Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Met hods in Natural Language Processing*, pages 6894–6910, 2021.

[42] Cheng Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, others, and Quoc V. Le. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4904–4916, 2021.

[43] Steffen Schneider, Alex Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech*, pages 3465–3469, 2019.

[44] Petar Veličković, William Fedus, Will L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations*, 2019.

[45] Y. You, T. Chen, Z. Wang, and Y. Shen. When does self-supervision help graph convolutional networks? In *Proceedings of the International Conference on Machine Learning*, pages 10871–10880, 2020.

[46] N. Wu, J. Phang, J. Park, Y. Shen, K. Kim, J. Sohn, others, and P. Rajpurkar. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Medical Imaging*, 41(5):1200–1210, 2022.

[47] Takuya Akiba et al. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[48] Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. Temperature schedules for self-supervised contrastive methods on long-tail data. *arXiv preprint arXiv:2303.13664*, 2023.