AI-Driven Evaluation and Optimization of Bump Pitch Effects on Chiplet and Interposer Design Quality

Seungmin Woo
smwoo@gatech.edu
Georgia Institute of Technology
Atlanta, GA, USA

Pruek Vanna-iampikul
pruek.va@eng.buu.ac.th
Burapha University
Chonburi, Thailand

Sung Kyu Lim
limsk@ece.gatech.edu
Georgia Institute of Technology
Atlanta, GA, USA

ABSTRACT

2.5D integration is gaining popularity primarily due to its ability to facilitate intellectual property (IP) reuse. Unlike conventional 2D and 3D approaches, 2.5D integration requires a more complex design and analysis process and is highly sensitive to changes in design parameters. However, research on the sensitivity of 2.5D design parameters is notably scarce, with most studies still concentrating on 2D and 3D. In this paper, we propose an AI-driven model for predicting sensitivity and an optimization methodology for 2.5D parameters, with a particular focus on bump pitch. Our approach employs advanced machine learning models to accurately predict how variations in bump pitch impact the power, performance, and area of chiplets, as well as the footprint, signal integrity, power integrity, and thermal integrity of the interposer layout. We also utilize Bayesian optimization to identify the optimal bump pitch for specific design objectives. Experimental validation of our model demonstrates high accuracy, with average relative errors of 2.69% for interpolation and 2.7% for extrapolation. Furthermore, optimization results, tailored by adjusting weights for various potential design goals, show an average improvement of 11% in area, wire length, and signal integrity-driven optimization, and 9% in power and thermal integrity-driven optimization.

1 INTRODUCTION

Designing and analyzing in 2.5D differs significantly from conventional 2D and 3D methods, often requiring additional time and resources [1]. From a design perspective, 2.5D integration demands a unique set of tools and methodologies for effective implementation. For example, while designers can readily create chip layouts using an existing Process Design Kit (PDK) for both 2D and 3D, implementing a 2.5D design requires an additional interposer PDK as well as a chiplet PDK. This necessity arises because chiplet and interposer designs involve distinct components, which in turn necessitate separate design environments. Moreover, the 2.5D design process requires multiple commercial tools and significant manual effort to achieve an optimized layout, leading to extended design times. Additionally, compared to conventional methods, 2.5D analysis necessitates more detailed and precise examinations using a larger number of Electronic Design Automation (EDA) tools. Unlike 2D and 3D designs, which can straightforwardly assess power, performance, and area (PPA), 2.5D designs require more complex analyses, such as signal integrity (SI), power integrity (PI), and thermal integrity (TI). The increased number of required tools and the complexity of these analyses for 2.5D further complicate the process and significantly extend the time needed to complete designs.

Furthermore, the 2.5D design parameters depicted in Fig. 1 play a crucial role in determining the overall performance of the chip.

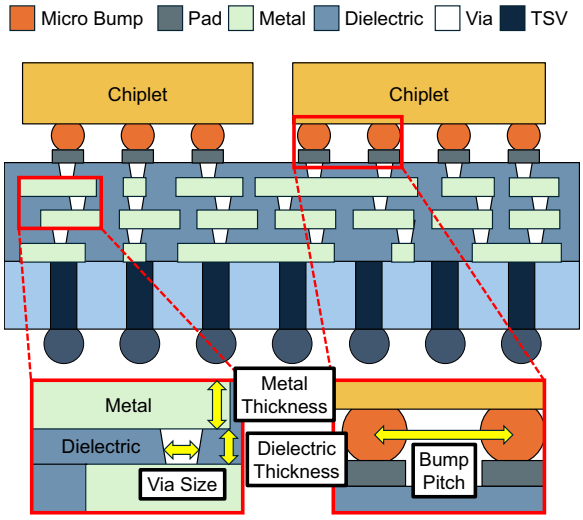

Figure 1: 2.5D silicon interposer and its various design parameters, including metal and dielectric thickness, via size and bump pitch. Every parameter significantly impacts on the PPA of the chiplet, as well as the SI, PI, and thermal dynamics of the interposer.

For example, changes in bump pitch can significantly improve the chiplet area, wire length, and SI, but may adversely affect the PI and thermal metrics of the interposer. Additionally, even relatively minor adjustments, such as altering the thickness of metal layers and dielectrics or the dimensions of vias, can lead to substantial and widespread impacts. Moreover, in addition to the changes in diemensions, the choice of substrate material—whether silicon, glass, or organic—significantly affects the chip and necessitates modifications in the overall design process.

Thus, it is crucial to thoroughly investigate and precisely predict the effects of parameter adjustments with a comprehensive and efficient 2.5D design and analysis. This step is essential, as it sets a valuable standard for those involved in semiconductor packaging processes and manufacturing. By evaluating the benefits and drawbacks that these adjustments bring to the entire system, it becomes possible to identify which parameters most effectively enhance performance, allowing for more targeted improvements. Moreover, accurately predicting these adjustments provides an opportunity to optimize parameters for specific design goals, enabling the identification of the most beneficial parameter value for critical aspects such as thermal or signal integrity, without the need for extensive preliminary designs.

This work is licensed under a Creative Commons Attribution International 4.0 License.

ICCAD '24, October 27–31, 2024, New York, NY, USA
© 2024 Copyright is held by the owner/author(s).
ACM ISBN 979-8-4007-1077-3/24/10.
https://doi.org/10.1145/3676536.3676817

Numerous research endeavors have explored various aspects of 2.5D design. For example, [2] examines the trade-offs between using silicon and organic interposers in 2.5D chiplet integration. Similarly, [1] introduces a novel methodology for designing with glass interposers, comparing its effectiveness against conventional materials. While these studies have advanced our understanding of material impacts on 2.5D designs, they do not fully address the influence of design parameters, such as bump pitch, on overall design metrics. Additionally, a growing body of work employs artificial intelligence (AI) to explore design spaces for heterogeneous integration. [3] introduces an AI model capable of predicting PPA for RISC-V architectures under specific configurations, while [4] extends this approach to 3D structures, optimizing accelerator configurations for defined design goals. However, these studies primarily focus on 2D and 3D environments, with limited exploration into 2.5D. [5] attempts to bridge this gap by presenting a methodology for predicting PPA in 2.5D-designed accelerators but lacks comprehensive prediction models for crucial aspects such as SI, PI, and TI. Moreover, it does not validate the accuracy of these prediction models and optimization methods through actual designs.

Therefore, we propose a new framework that utilizes machine learning to predict the sensitivity of 2.5D design parameters and identify the optimal configurations, with a particular focus on bump pitch. Based on an efficient design and analysis flow that extracts data from 2.5D designs, our framework provides initial insights into the impact of bump pitch on 2.5D technology. Moreover, we suggest suitable ML models for each design metric and present an appropriate optimization methodology. Our contributions are as follows:

(1) We propose an effective 2.5D co-design and co-analysis methodology specifically tailored for exploring the sensitivity of bump pitch.
(2) We demonstrate how adjustments in bump pitch affect the PPA of chiplets as well as wire length, SI, PI, TI of the interposer.
(3) To the best of our knowledge, we are the first to present a machine learning-based model that is capable of predicting the effects of bump pitch on 2.5D.
(4) We are also first to present an optimization methodology to find the optimal bump pitch for specific design goals.
(5) Our framework successfully predicts various design metrics with an average relative error of 2.69% in interpolation and 2.7% in extrapolation.
(6) Our framework also derives the desired points for all possible optimization configurations, exhibiting an 11% improvement for area, wire length, and signal integrity optimization, and a 9% improvement for power and thermal integrity optimization.

2 DESIGN FLOW AND BENCHMARK

2.1 Proposed Design and Analysis Flow

The design process is outlined in Fig. 2. We individually design and analyze both the chiplet and the interposer from various perspectives. First, the chiplet design and analysis process, illustrated in Fig. 2 (a), involves several key steps. The chipletization process is pivotal in 2.5D integration. Initially, we combine the register

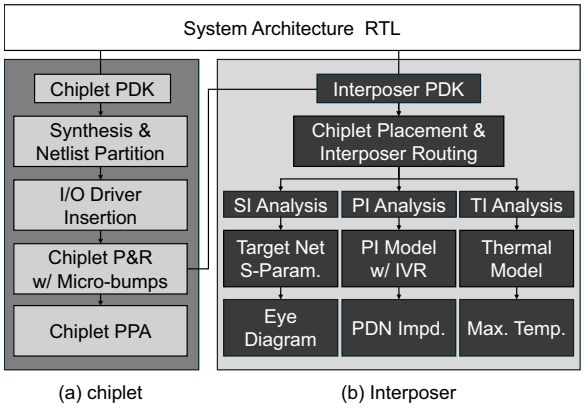

Figure 2: Proposed 2.5D design and analysis flow.

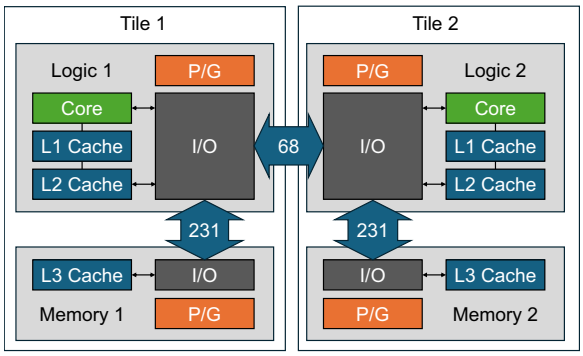

Figure 3: OpenPiton as a benchmark for 2.5D integration.

transfer level (RTL) with the chiplet's PDK specific to a particular technology node and partition the netlist. Synthesis is performed using Design Compiler. Subsequently, I/O drivers are integrated into the partitioned netlist to establish off-chip connectivity between chiplets. After successful implementation of micro-bumps and establishment of signal and power connections, we proceeded with chiplet placement and routing, followed by analysis of PPA metrics. Cadence Innovus was utilized for place and route (PnR) operations, and Cadence Tempus was used for PPA analysis.

In the interposer design depicted in Fig. 2 (b), we initially gather information such as die dimensions, bump locations, and port names. Subsequently, the interposer PDK, containing essential information such as material properties, dimensions of metal and dielectric layers, and wire width and spacing within the interposer, is linked. Connections between chiplets are defined based on the overall system architecture RTL, chiplets are placed in appropriate locations, and interposer routing is carried out. Siemens Xpedition software is used to create the interposer layout, with the autorouting function employed. In cases where automatic routing failed, manual routing is undertaken. Once the interposer layout is finalized and examined, SI, PI, and Thermal analyses were conducted. For SI analysis, Siemens HyperLynx is used to calculate the s-parameters of the target net, and Keysight Advanced Design System (ADS) is employed to simulate the eye diagram. For PI analysis, assuming an integrated voltage regulator (IVR) operating at

Table 1: Silicon interposer specification and the number of bumps used in this paper

	Logic	Memory
# Metal layer	4	
Metal thickness (μm)	1	
Dielectric thickness (μm)	1	
Min. wire W/S (μm)	0.4/0.4	
Via size (μm)	0.7	
Bump size (μm)	20	
Die-to-Die spacing (μm)	100	
Micro-bump pitch (μm)	40	
# Signal bump	299	231
# P/G bump	165	130
# Total Bump	464	361

Table 2: Material Properties for 2.5D Silicon Interposer.

	Substrate	Dielectric
Material	Silicon	SiO_2
Permittivity (Dk)	11.7	3.9
Loss tangent (Df)	-	0.001
Thermal cond. (W/mK)	148	1.5

125 MHz, the power delivery network's s- and z-parameters were obtained using Siemens Xpedition. Impedance is measured, and IVR is incorporated into the schematic in Keysight ADS to simulate settling time and voltage drop. Finally, for Thermal analysis, Ansys Redhawk is utilized to extract the chip thermal model (CTM) of the chiplet, which is combined with the thermal model of the interposer layout to measure the maximum temperature of both the chiplet and interposer.

2.2 Benchmark Architecture

We use OpenPiton as a benchmark with RISC-V Ariane Core for 2.5D implementation [6]. There are a total of four cores in each OpenPiton tile, and we utilized two tiles in total as shown in the Fig. 3. For cache capacity, we set L1 cache to 8KB, L2 cache to 16KB, and L3 cache to 1.8MB per tile. Then, we synthesize the configured RTL with chiplet PDK and partition the netlist by defining L3 cache and L3-related modules as 'memory chiplet' and other modules including cores as 'logic chiplet'. Thus, two logic chiplets and two memory chiplets are utilized. Also, the SerDes module is applied to the I/O driver design to reduce the 64-bit parallel communication to 8-bit serial communication. Thus, there are 68 logic-to-logic connections between tiles and 231 logic-to-memory connections. Furthermore, the number of power and ground bumps is assumed to be 165 for logic and 130 for memory. Therefore, the final number of bumps in each chiplet is 464 for logic and 361 for memory as indicated in Table 1.

3 INTERPOSER TECHNOLOGY SPECS

3.1 Geometries and Material Properties

The interposer specification utilized is as depicted in Table 1. The type of interposer is Silicon, and we use Chip-on-Wafer-on-Substrate

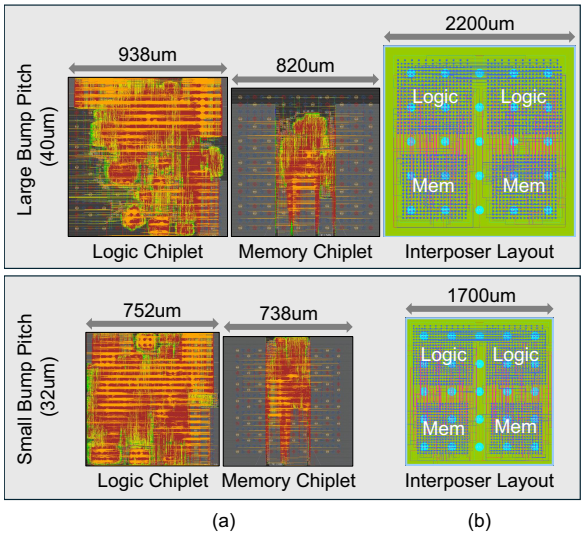

Figure 4: Impact of micro-bump pitch on chiplet and interposer sizes. (a) chiplet and (b) interposer. Smaller bump pitch leads to smaller chiplets and interposers.

(CoWoS) technology [7] as a default technology. According to this technology, the thickness of both metal and dielectric is 1um, and the wire width and spacing are assumed to be 0.4um. A total of four metal layers are used, two for signal and two for power. For our study, we only make adjustments to the bump pitch while maintaining the default settings for other parameters. We also employ material properties for accurate interposer simulation as indicated in Table 2. Silicon substrate features a permittivity of 11.7 and a thermal conductivity of 148 W/mK. SiO2, used as the dielectric material, has a permittivity of 3.9, complemented by a loss tangent of 0.001 and a thermal conductivity of 1.5 W/mK. By utilizing these specific material properties, we can successfully conduct precise interposer analysis.

3.2 Why Bump Pitch Sensitivity?

Parameter sensitivity plays a crucial role in 2.5D IC design, where even slight variances to design parameters can have a significant impact on the overall. In particular, bump pitch can significantly impact both the chiplet and interposer. To explore this phenomenon, we conduct experiments with bump pitches of 40μm and 32μm, analyzing their impacts from various perspectives. Fig. 4 depicts the layout design results in both the chiplet and interposer. It is noteworthy that reducing the bump pitch not only reduces the area of the chiplet but also effectively reduces the footprint of the interposer. Table 3 provides detailed numerical changes in the chiplet and interposer metrics. Our numerical investigation shows that there is a notable decrease in the chiplet's die width as the bump pitch is reduced. Additionally, there is a slight reduction in power while still achieving the target frequency.

A significant reduction in wire length and a decrease in footprint size are also observed in the analysis of the interposer. This reduction occurs despite the wire width and spacing remaining constant across both scenarios, suggesting that the reduction in chiplet area results in the decreased average distance between chiplets' bumps,

Table 3: Impact of Bump Pitch Reduction on Design Metrics. We observe improvement in chiplet power consumption, interposer wire length and eye diagram, while deterioration in PI and thermal metrics.

	40um Bump		32um Bump	
	Logic	Mem	Logic	Mem
Chiplet				
Area (μm^2)	938	820	**752**	**738**
Pwr (mW)	140.3	47.6	**137.7**	**45.4**
f_{max} (MHz)	689.2	669.3	687.8	654.9
Interposer				
Footprint (mm^2)	2.2 x 2.2		**1.7 x 1.7**	
Max WL (mm)	3.0		**2.4**	
Avg WL (mm)	1.6		**1.4**	
Tot WL (mm)	842.2		**750.7**	
Eye Width (ns)	0.90		**1.26**	
Eye Height (V)	0.31		**0.42**	
PDN Imp. (Ω)	**7.52**		12.53	
PI Time (μs)	**4.14**		4.15	
PI Drop (mV)	**27.30**		28.70	
Log Temp ($^\circ C$)	**30.15**		35.10	
Mem Temp ($^\circ C$)	**25.16**		27.49	

which in turn significantly impacts on wire length reduction. Furthermore, improvements in the eye diagram are noted as bump pitch decreased, indicating that shorter wire lengths have a more beneficial effect on signal integrity than the negative impacts of increased wire density. Conversely, power integrity metrics such as Power Delivery Network (PDN) impedance, PI settling time, and voltage drop deteriorates with smaller bump pitches, likely due to the reduced dimensions of the power plane which increase current density. Similarly, temperature results also deteriorate, suggesting that the chiplet's power density increases as its size decreases, exacerbating thermal challenges.

The variations in bump pitch have a significant impact on different design metrics, and the optimal bump pitch may vary depending on the design requirements. Hence, precise prediction and adjustment of these metrics are essential, necessitating the utilization of an appropriate regression model and optimization methods to ascertain the optimal bump pitch for attaining desired design goals.

4 PROPOSED AI FRAMEWORK

In this section, we describe our proposed framework for 2.5D sensitivity prediction model and optimization methodology. Section 4.1 gives an overview of the overall framework, while Section 4.2 describes how the training dataset is derived. Section 4.3 introduces the model required to predict various design metrics based on bump pitch, and Section 4.4 describes the optimization methodology to find the optimal bump pitch for a specific design goal based on the prediction model.

4.1 Overview

The overall framework for sensitivity prediction and optimization is depicted in Fig. 5. We first utilize the 2.5D design methodology

Table 4: Models employed for predicting (interpolation and extrapolation) each metric and their accuracy. We utilized Multi-layer Perceptron (MLP) for SI, PI and TI due to their non-linear attributes. We also show how each metric reacts to bump pitch reduction under "React".

	React	Pred. Model	R2	Inter. Err [%]	Extra. Err [%]
Chiplet Area	better	Linear	1.00	0	0
Interposer Area	better	Linear	0.96	1.84	0.35
Total WL	better	Elastic	0.96	0.35	2.09
Max. WL	better	Linear	0.91	0.18	2.74
Eye Height	better	MLP	0.86	2.6	0.1
Eye Width	better	MLP	0.78	2.6	0.2
PDN Impd	worse	MLP	0.72	7.76	0.07
Logic Temp.	worse	MLP	0.90	1.15	9.47
Mem. Temp.	worse	MLP	0.87	2.52	3.96
Interposer Temp.	worse	MLP	0.84	0.94	10.21
Average	-	-	0.88	**1.99**	**2.92**

Table 5: Boundary conditions for thermal analysis

	Type and Value
Simulation Env.	Air
Flow regime	Turbulent
Radiation model	Ray Tracing
# Iteration	10
Initial Temp. ($^\circ C$)	20
Fan speed (cfm)	2.12
Fan direction	negative Z
Fan radius (μm)	320

described in Section 2 to adjust the bump pitch and create a training dataset within the appropriate range. Subsequently, a machine learning-based prediction model is built for chiplet PPA, eye diagram, PDN impedance, and maximum temperature for each bump pitch. Using these prediction models, we establish an objective function and perform suitable optimization.

4.2 Training Dataset

2.5D design and analysis process is highly time-consuming. In particular, unlike the relatively automated process of chiplet design, the interposer design and analyses of SI, PI and TI necessitate substantial manual effort. This makes the processes not only time-consuming but also challenging to execute in parallel. For instance, if auto-routing fails for interposer layouts, manual routing is required. In SI analysis, the aggressor with the most overlap with the victim net must be manually identified in the visualized layout. Additionally, the thermal analysis involves extracting the CTM from chiplets and integrating these into a new 3D model for every bump pitch. Given these complexities, constructing a large dataset is impractical due to inefficiencies in 2.5D design and analysis.

Therefore, our strategy to construct training dataset focuses on developing a highly representative dataset for 2.5D designs. We follow the design flow outlined in Section 2 to generate a training dataset. We also establish a suitable range of the bump pitch. The

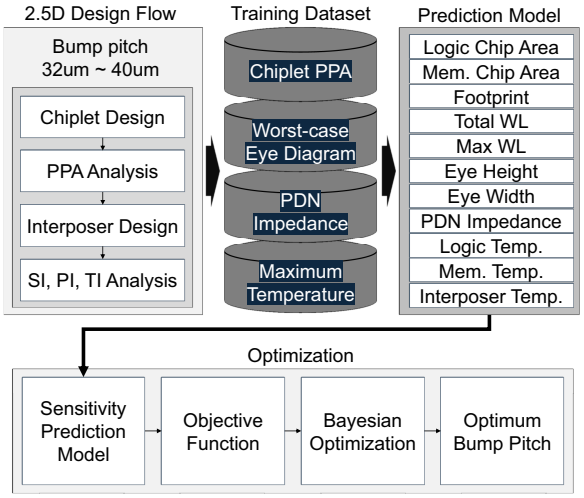

Figure 5: Our AI-driven bump pitch prediction and optimization framework.

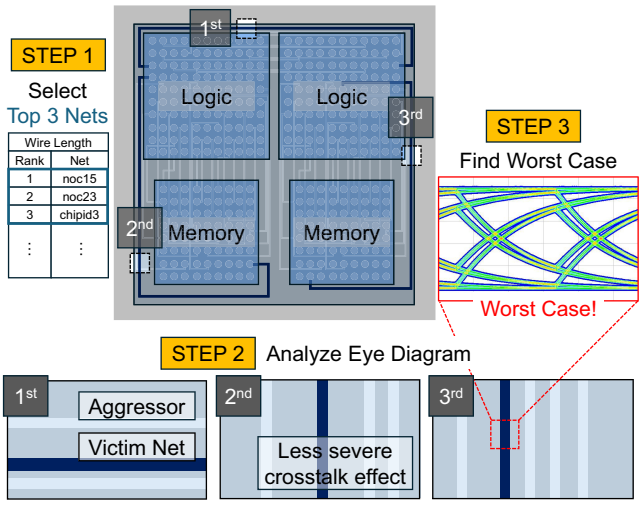

Figure 6: Our approach for interposer signal integrity (SI) analysis to find the worst case eye diagrams.

bump pitch was adjusted by increasing it from 32μm to 40μm. In chiplet's PPA analysis, we focus on areas where the bump pitch exhibited significant variation. We not only determine the area with the bump pitch but also verify the feasibility of that chiplet area by scrutinizing the actual PPA. Meanwhile, for the interposer, we extract wire length data from the layout after finishing the routing. While wirelength typically has the most substantial impact on the eye diagram, it is crucial to consider the aggressor, a net significantly affecting crosstalk. As depicted in Fig. 6, in order to efficiently determine the worst-case eye diagram, we rank the nets based on wirelength and analyze the eye diagrams of the top three nets. Subsequently, we identify the most severe net as the worst-case scenario. This approach allows us to effectively analyze the signal integrity (SI) of the design for each bump pitch, despite the time-consuming nature of the process. For PI analysis, we assume the utilization of a 125MHz IVR as mentioned in Section 2. Consequently, we analyze the Power Delivery Network (PDN) impedance across a frequency range of 1MHz to 1GHz and determine the impedance specifically at 125MHz. Finally, for thermal analysis, we initially build a 3D structure of the interposer as depicted in Fig. 7. Using this model and boundary condition in the Table 5, we simulate the maximum temperature of the interposer as well as the logic and memory chiplets.

4.3 Prediction Models

Utilizing the previously extracted dataset, our framework construct a model to predict various design metrics. As stated in Section 4.2, the unique characteristics of 2.5D design and analysis present difficulties in creating large datasets. In order to address these challenges, our framework employs a method to determine the most suitable model for predicting sensitivity. In addition, we implemented an advanced model evaluation approach to account for the inaccuracies resulting from the limited dataset. The machine learning model we suggest for each metric can be found in Table 4.

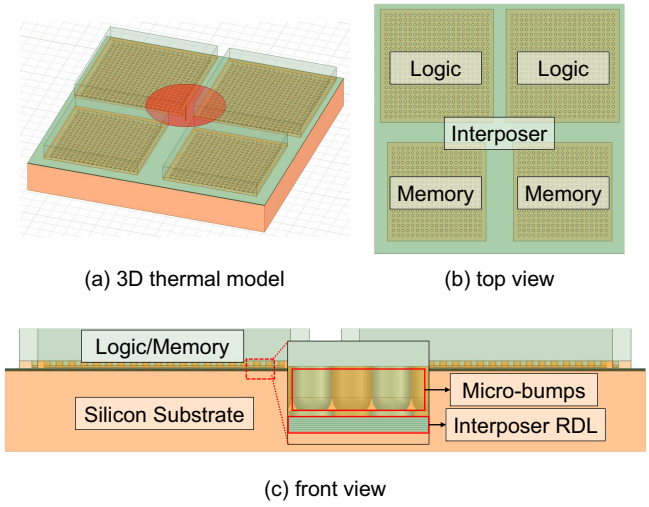

Figure 7: Thermal model for interposer layout: (a) 3D thermal model structure, (b) top view and (c) front view of the model.

4.3.1 Area, Footprint, Wire Length. Linear and elastic net models are employed to predict the values of Area, Footprint, and Wire Length. Due to the lower variability of these metrics in relation to bump pitch compared to other metrics, we can make relatively accurate predictions using straightforward models. Also, to figure out the exact input range of model, we investigate the range of bump pitches that are suitable for ensuring the validity of the area on our current benchmarks by analyzing the overall PPA of the chiplets. Our investigation shows that chiplet size cannot be further reduced for bump pitches below 32μm or above 100μm, as doing so would result in negative slack exceeding 100% of the target frequency. Therefore, we devised a linear regression model to predict the area within the constrained span of 32μm to 100μm. This range acts as the input range for all prediction models. Linear regression is also utilized for the footprint and maximum wire length. As for Total

Table 6: Time taken for data construction (build time), training and inferencing

Model	# Data	Range	Build Time	Training Time	Infer. Time
Chiplet Area	18	$32\mu m$ - $40\mu m$	3h	<1min	<1min
Footprint	9		4h	<1min	<1min
Wirelength	9		4h	<1min	<1min
Eye Diag.	27		12h	5min	<1min
PDN Impd.	9		7h	3min	<1min
Max. Temp.	27		15h	6min	<1min

wire length, we construct a prediction model utilizing an elastic net. The model's non-linearity stems from the fact that the reduction in bump pitch does not result in a proportional decrease in the total wire length, owing to the constant width and spacing of the wire.

4.3.2 Eye Diagram, PDN Impedance, Maximum Temperature. The datasets for eye diagram, PDN impedance, and maximum temperature exhibit a non-linear tendency. Since we vary bump pitch with other parameters unchanged, the impact of this overhead is not straightforward. For example, especially concerning SI, multiple factors influence the eye diagram, such as the overlap between aggressor and victim nets, as well as the route taken by the victim net. These variations cannot be accurately predicted using a basic model. Consequently, we employ the Multi-layer Perceptron (MLP) technique in this case. The input values range from 32μm to 100μm, consistent with the previous model. The layer count is set to 1000, and the alpha value is adjusted to improve the model's accuracy.

4.3.3 Evaluation. To enhance the reliability and accuracy of our machine learning models for 2.5D design predictions, we implement a comprehensive evaluation strategy. This includes not only assessing R2 scores but also conducting practical tests through interpolation and extrapolation at some bump pitches, such as 36.5μm and 45μm. By comparing the models' predictions against actual metrics from implemented designs, we can adjust the fitting of the models to minimize discrepancies. This methodology ensures that our models are both robust within training dataset and practically applicable in complex 2.5D design environments. In our study, when using 36.5μm and 45μm as an evaluation bump pitch, our models achieve an average R2 score of 0.88, with mean relative errors of 1.99% for interpolation and 2.92% for extrapolation as demonstrated in Table 4. This underscores the models' predictive accuracy and their capacity to adapt to diverse scenarios outside of training dataset, which is critical for bump pitch optimization.

4.4 Optimization Method

Based on the prediction model, we propose a method for identifying the most optimal bump pitch. Bayesian optimization is utilized to accomplish this goal once an objective function has been defined.

4.4.1 Objective Function. The objective function is created by normalizing each metric, dividing it by the maximum value. The normalized result is subsequently transformed into a scalar value through a weighted sum method, in which metrics with comparable impacts are combined into a single weight. Furthermore, if a decrease in metric results in an improvement, such as area or wire length, the negative weight is applied. As a result, based on the "react" in Table 4, five crucial weight variables is constructed, which together constitute an expression shown in Eq. 1. The variables in this function are defined as follows: α represents the area, β represents the wire length, γ represents the SI, δ represents the PI, and ϵ represents the thermal weight. The function takes into account various metrics, such as Chiplet Area (A), Interposer Footprint (IF), Total Wirelength (TW), Maximum Wirelength (MW), Eye Height (EH), Eye Width (EW), PDN Impedance (PDI), Logic Temperature (LT), Memory Temperature (MT), and Interposer Temperature (IT).

$$\begin{aligned} f = &-\alpha\left(\frac{A}{A_{\max}}+\frac{IF}{IF_{\max}}\right)-\beta\left(\frac{TW}{TW_{\max}}+\frac{MW}{MW_{\max}}\right) \\ &+\gamma\left(\frac{EH}{EH_{\max}}+\frac{EW}{EW_{\max}}\right)-\delta\left(\frac{PDI}{PDI_{\max}}\right) \\ &-\varepsilon\left(\frac{LT}{LT_{\max}}+\frac{MT}{MT_{\max}}+\frac{IT}{IT_{\max}}\right) \end{aligned} \tag{1}$$

4.4.2 Optimization Methodology. Utilizing the objective function, we employ Bayesian optimization. Bayesian optimization is known to be particularly beneficial in cases where the precise objective function is unknown, but can obtain samples of the objective function value at specific points [8]. In a similar vein, we cannot represent the exact expression of the objective function in terms of the bump pitch we established in the previous section, but we know the value of the function at a particular bump pitch from the output of the prediction model. Thus, we make use of Bayesian optimization, the most advantageous method of optimization for our specific problem. Furthermore, by employing a variable weight and allowing the user to determine the weight of the objective function, we can identify a bump pitch that satisfies a particular design objective.

5 EXPERIMENTAL RESULT AND DISCUSSION

We perform the optimization process and validated the accuracy of both the prediction model and optimization results by implementing them in the actual design. Python is utilized for implementing both the prediction model and optimization. In Section 5.1 and Section 5.2, we present and discuss the time taken for data construction and for training and inferencing. In Section 5.3, we scrutinize the outcomes of executing the optimization process using two distinct weight arrangements. Section 5.4 utilizes the optimal bump pitch values determined by the optimizer to conduct the actual design and verify the accuracy of the prediction model. Finally, in Section 5.5, we compare the two actual designs and demonstrate that the optimization functions as intended.

5.1 Dataset Construction

Table 6 presents the time commitments required for the dataset, as outlined in Section 4.2. We develop the training dataset by adjusting the bump pitch between 32μm and 40μm. Specifically, the chiplet design process requires about 3 hours, while the interposer layout design which provides footprint and wirelength data takes approximately 4 hours. More time-consuming are the analyses for the eye diagram, PDN impedance, and maximum temperature, which take 12, 7, and 15 hours, respectively. As discussed in Section 4.2, our

Table 7: AI-driven optimization results under two different goals. Our database covers pitch values in [32um, 40um] with 1um increment. We also show the model accuracy against the actual designs for the two optimized solutions.

			Area+WL+SI-driven optimization (pitch chosen=38.5um, interpolation)				PI+Thermal-driven optimization (pitch chosen=43um, extrapolation)			
Metric	**Model**	**R2**	**Predicted**	**Actual**	**Abs. Err.**	**Rel. Err.**	**Predicted**	**Actual**	**Abs. Err.**	**Rel. Err.**
Logic Area (μm^2)	Linear	1.00	904.19	904.75	0.56	**0.06%**	1009.34	1010.50	1.16	**0.11%**
Mem. Area (μm^2)	Linear	1.00	789.72	789.25	0.47	**0.06%**	876.00	881.50	5.50	**0.62%**
Footprint (mm^2)	Linear	0.96	2076.25	2000.00	76.25	**3.81%**	2302.68	2300.00	2.68	**0.12%**
Total WL (mm)	Elastic	0.96	818.00	830.98	12.98	**1.56%**	867.08	889.34	22.26	**2.50%**
Max WL (mm)	Linear	0.91	3.06	2.92	0.14	**4.79%**	3.31	3.19	0.12	**3.76%**
Eye Height (ns)	MLP	0.86	0.39	0.37	0.02	**6.25%**	0.33	0.32	0.01	**2.17%**
Eye Width (V)	MLP	0.78	1.02	0.99	0.03	**3.04%**	0.90	0.89	0.02	**1.80%**
PDN Impd. (Ω)	MLP	0.72	8.47	8.94	0.47	**5.30%**	6.45	6.77	0.32	**4.69%**
Logic Temp. ($^\circ C$)	MLP	0.90	32.11	30.74	1.37	**4.47%**	27.82	29.82	2.00	**6.69%**
Mem. Temp. ($^\circ C$)	MLP	0.87	25.62	25.62	0.00	**0.01%**	24.74	24.81	0.07	**0.28%**
Int. Temp. ($^\circ C$)	MLP	0.84	29.47	29.52	0.05	**0.18%**	26.01	27.96	1.95	**6.96%**
Average	-	0.88	-	-	-	**2.69%**	-	-	-	**2.70%**

Table 8: Full-chip comparison between the two optimal solutions from Table 7.

	small pitch	**large pitch**	**who is better?**	**% gain**
Logic Area (μm^2)	904.75	1010.50	small	10.47%
Memory Area (μm^2)	789.25	881.50	small	10.47%
Footprint (mm^2)	2000	2300	small	13.04%
Total WL (mm)	830.98	889.34	small	6.56%
Max WL (mm)	2.92	3.19	small	8.46%
Eye Height (ns)	0.37	0.32	small	14.29%
Eye Width (V)	0.99	0.89	small	11.27%
PDN Impedance (Ω)	8.94	6.77	large	24.28%
Logic Temp. ($^\circ C$)	30.74	29.82	large	3.00%
Memory Temp. ($^\circ C$)	25.62	24.81	large	3.17%
Interposer Temp. ($^\circ C$)	29.52	27.96	large	5.28%
Average Imprv.	**10.65%**	**8.94%**	-	-

strategy has been to focus on developing a representative dataset for 2.5D designs, while supporting the accurate prediction models.

5.2 AI Model Training and Inferencing

As demonstrated in Table 6, although the construction of the dataset is time-consuming, the training and inferencing of our models are remarkably swift, each taking only several minutes. This efficiency is largely due to the small size of our dataset. As mentioned in Section 4.3, we strategically choose machine learning models and validate their effectiveness through interpolation and extrapolation to overcome the potential accuracy issue. associated with a small dataset. As a result, this ensures that models deliver reliable performance despite the inherent constraints of the dataset.

5.3 Pitch Optimization Results

We conduct two weight configurations and derive the optimal bump pitch, utilizing optimization methods. Table 7 illustrates the correlation between bump pitch size and area, wirelength, and eye diagram metrics. A reduction in bump pitch size generally results in enhancements in these metrics, while PDN impedance and temperature tend to worsen. An ideal optimizer should be able to prioritize one of two sides and achieve desired optimization. Therefore, we validate two scenarios: 1) area, wirelength, and eye diagram-driven, and 2) PDN impedance and temperature-driven optimization. In the former scenario, we assign the weights α (Area), β (wire length), and γ (SI) a value of 2, while the weights δ (PI) and ϵ (Thermal) were assigned a value of 1. In contrast, for the later scenario, we assign a value of 3.5 to δ and ϵ, while assigning a value of 1 to α, β, and γ. Consequently, we assign a weight of 75% to a selected set of metrics for each case. The optimization range is delimited from 32μm to 100μm, aligning with the input range of the prediction model. Moreover, the total number of iterations is set at 50, with an initial point of 5. Consequently, the optimizer suggests a bump pitch of 38.5μm as the optimal point for area, wirelength, and eye diagram, while a bump pitch of 43μm is chosen for PDN impedance and temperature-driven optimization.

5.4 Model Accuracy Evaluation

Based on the derived optimum bump pitch, the actual design is carried out and analyzed as shown in Fig. 8. Also, the prediction error is calculated by comparing the analyzed results with the predicted values as indicated in the Table 7. The area, wirelength, and eye diagram-driven optimization yield a value of 38.5μm, which is in the range of 32μm and 40μm in the training dataset. Hence, this value falls under interpolation. Conversely, PDN impedance and temperature-driven optimization yields a bump pitch of 43μm, which falls into the category of extrapolation. Consequently, we evaluate the model's performance in terms of both interpolation and extrapolation. the interpolation displays an average error of 2.69%, while the extrapolation exhibits an average error of 2.7%. Overall, the errors deem to be small in both cases, suggesting a high accuracy of the prediction model. Furthermore, the outcome demonstrates the validity of our design process and the methodology employed in generating the training dataset.

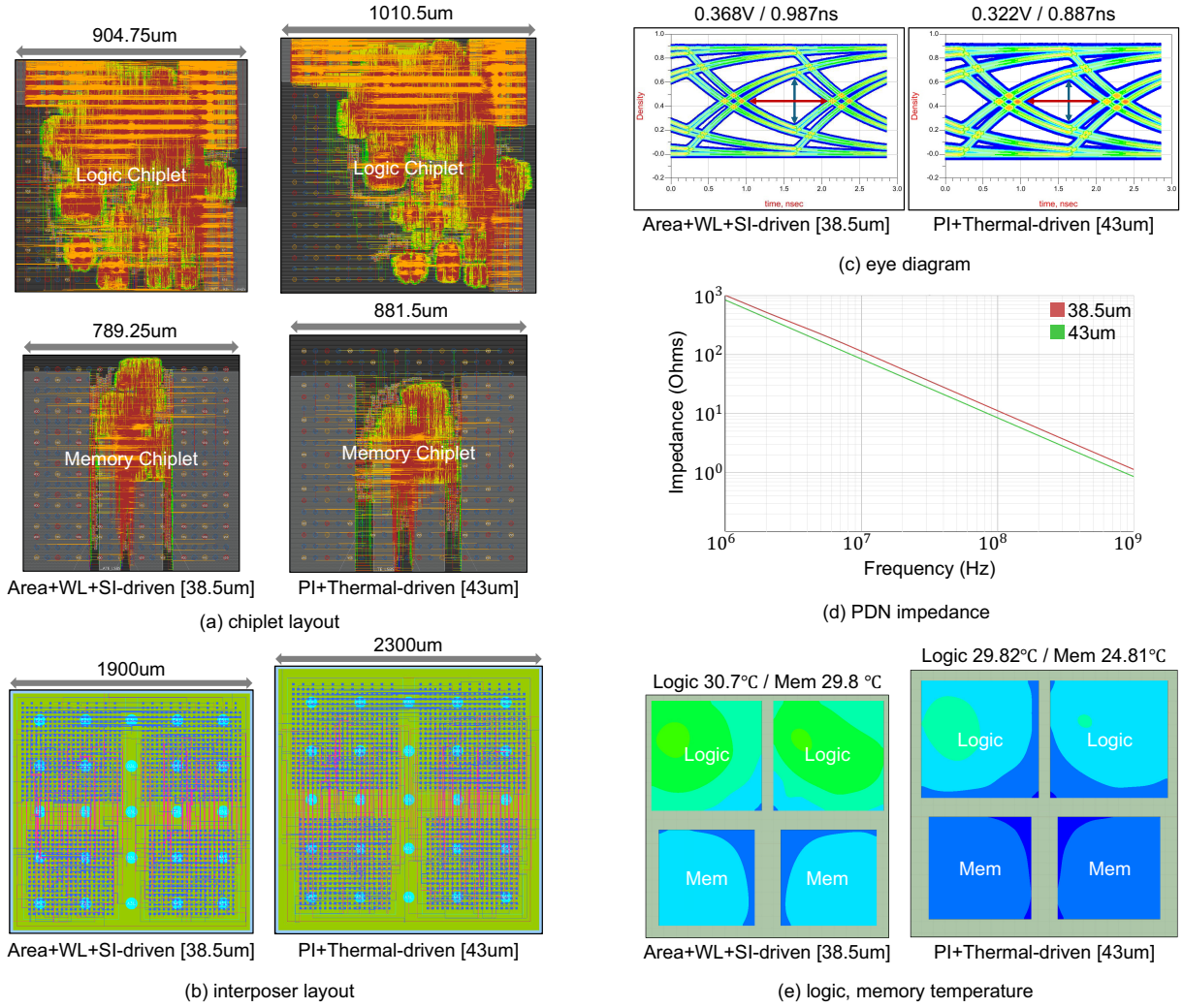

Figure 8: Actual design for optimum bump pitches: (a) chiplet and (b) interposer layout, (c) the worst eye diagram, (d) PDN impedance, and (e) thermal distribution.

5.5 Design Validation of Optimized Parameters

The validation of the optimization method is conducted by comparing the actual analysis results. Table 8 illustrates the extent of improvement when comparing the actual analysis results with their counterparts. For a bump pitch of 38.5μm, there is a 10% decrease in area, a 13% decrease in interposer footprint, a 7% decrease in total wirelength, and an 8% decrease in maximum wirelength compared to 43μm. From a SI perspective, we observe a 14% increase in eye height and an 11% increase in eye width. Conversely, 43μm demonstrates a 24% improvement in PDN impedance compared to 38.5μm. Logic, memory, and interposer temperatures show improvements of 3%, 3%, and 6%, respectively. Consequently, area, wire length, and signal integrity-driven optimization results in an average improvement of 11%, while power and thermal integrity-driven optimization yields an average improvement of 9%. Thus, this demonstrates that our proposed optimization method offers an optimal bump pitch that satisfies the design objective.

6 CONCLUSION

We present a novel machine learning-based framework for predicting and optimizing the sensitivity of 2.5D parameters, focusing specifically on bump pitch. By leveraging our efficient 2.5D design and analysis flow for sensitivity exploration, we develop machine learning models to predict key metrics such as area, wire length, eye diagram, PDN impedance, and temperature. Our framework also includes an optimization method to determine the optimal bump pitch for specific design objectives. Experimental results show that our models achieve high accuracy, with a relative error of about 2.7%, and demonstrate strong optimization performance.

ACKNOWLEDGEMENTS

This work was supported by the Semiconductor Research Corporation (CHIMES 3136.002), the Ministry of Trade, Industry & Energy of South Korea (1415187652, RS-2023-00234159), and the National Science Foundation (CNS-2235398).

REFERENCES

[1] Pruek Vanna-Iampikul, Lingjun Zhu, Serhat Erdogan, Mohanalingam Kathaperumal, Ravi Agarwal, Ram Gupta, Kevin Rinebold, and Sung Kyu Lim. Glass interposer integration of logic and memory chiplets: Ppa and power/signal integrity benefits. In *2023 60th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6, 2023.

[2] Jinwoo Kim, Gauthaman Murali, Heechun Park, Eric Qin, Hyoukjun Kwon, Venkata Chaitanya Krishna Chekuri, Nael Mizanur Rahman, Nihar Dasari, Arvind Singh, Minah Lee, Hakki Mert Torun, Kallol Roy, Madhavan Swaminathan, Saibal Mukhopadhyay, Tushar Krishna, and Sung Kyu Lim. Architecture, chip, and package codesign flow for interposer-based 2.5-d chiplet integration enabling heterogeneous ip reuse. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 28(11):2424–2437, 2020.

[3] Xin Zheng, Mingjun Cheng, Jiasong Chen, Huaien Gao, Xiaoming Xiong, and Shuting Cai. Bsse: Design space exploration on the boom with semi-supervised learning. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, pages 1–10, 2024.

[4] Gauthaman Murali, Aditya Iyer, Navneeth Ravichandran, and Sung Kyu Lim. 3dnn-xplorer: A machine learning framework for design space exploration of heterogeneous 3d dnn accelerators. In *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, pages 1–9, 2023.

[5] Kaniz Mishty and Mehdi Sadi. System and design technology co-optimization of chiplet-based ai accelerator with machine learning. In *Proceedings of the Great Lakes Symposium on VLSI 2023*, GLSVLSI '23, page 697–702, New York, NY, USA, 2023. Association for Computing Machinery.

[6] Jonathan Balkind, Michael McKeown, Yaosheng Fu, Tri Nguyen, Yanqi Zhou, Alexey Lavrov, Mohammad Shahrad, Adi Fuchs, Samuel Payne, Xiaohua Liang, Matthew Matl, and David Wentzlaff. Openpiton: An open source manycore research framework. *SIGPLAN Not.*, 51(4):217–232, mar 2016.

[7] Raghunandan Chaware, Kumar Nagarajan, and Suresh Ramalingam. Assembly and reliability challenges in 3d integration of 28nm fpga die on a large high density 65nm passive interposer. In *2012 IEEE 62nd Electronic Components and Technology Conference*, pages 279–283, 2012.

[8] Jakang Lee, Jaeseung Lee, Seonghyeon Park, and Seokhyeong Kang. Multi-source transfer learning for design technology co-optimization. In *2023 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pages 1–6, 2023.