Accelerating Dental Adhesive Innovations Through Active Learning and Bayesian Optimization

PREPRINT

Ramsey Issa
Department of Materials Science & Engineering
University of Utah
Salt Lake City, UT 84112
U1374011@utah.edu

Robert Sorenson
Department of R&D Chemical
Ultradent Products, Inc.
South Jordan, UT 84095
robert.sorenson@ultradent.com

Taylor D. Sparks
Department of Materials Science & Engineering
University of Utah
Salt Lake City, UT 84112
sparks@eng.utah.edu

May 17, 2024

Abstract

The discovery of new dental materials is typically a slow process due to high-dimensionality of the formulation space as well as the multiple competing objectives which must be optimized for a given application. Here, we lay out a strategy using active learning and Bayesian optimization that has led to the discovery of 3 new high-performing formulations for dental adhesives within 29 experiments. We utilize curated data from 91 experiments with 43 different components, to reduce the design space and incorporate domain knowledge into our search. The success of this machine learning approach can be adapted to a multitude of dental materials to allow for the fast and efficient discovery of optimal new formulations, leading to enhanced performance, reduced development times, and ultimately more cost-effective and innovative solutions in dental healthcare.

Keywords Machine Learning · Materials Science · Bayesian Optimization · Dental Adhesives · Active Learning

1 Introduction

Of late, applications of machine learning to materials science and engineering has emerged as a transformative tool for materials discovery[1, 2, 3]. These applications typically range from deep learning [4, 5], for image-based tasks [6, 7, 8], to more recent natural language processing adaptations to aid in design of new materials [9, 10, 11]. Although machine learning approaches to materials discovery have found applications across multiple materials domains, it has yet to make an impact in the field of dental materials [12, 13]. Instead, machine learning in dentistry has had modest impact in the areas of detection and diagnostics, with a few examples being the identification of dental caries, vertical root fractures, apical lesions, and diseases of the salivary glands [14].

Although these applications are very useful, they do not address a crucial need in developing next-generation dental materials themselves. This is a missed opportunity because there are many challenges in restorative materials, from dental composites to adhesives [15, 16], that could be addressed by machine learning including the following:

1. Identifying the best formulation within a large design space.
2. Understanding the contributions of each constituent on properties of interest particularly given that many resins are proprietary with undisclosed structures.

3. Constructing a design framework that is flexible to the reality of new resins coming to market and existing resins being discontinued.

First, new materials must have desirable properties, but the design space over which formulation scientists must search is daunting. Indeed, traditional development of new materials has been extremely slow and governed by trial-and-error based on local optimization of known formulations. The underlying reason for this slow development is due to the high dimensionality of these formulations consisting of resins, fillers, pigments, initiators, and inhibitors. Each of these constituents could include dozens to hundreds of candidate ingredients which can mixed with nearly infinite ratio combinations. To put this into perspective, if a formulations scientist were to choose 5 ingredients to incorporate into a formulation, of 100 ingredients on a bench top, the amount of unique combinations would be

$$C_k(n) = \frac{n!}{k!(n-k)!} = \frac{100!}{5!(100-5)!} \approx 10^7$$

and this assumes a fixed ratio of constituents. In reality, they will be allowed to vary resulting in an infinitely large design space; it is no wonder that new dental materials have been so challenging to find and optimize. The time is right to leverage data-driven optimization techniques to find new dental materials in treating dental diseases.

Bayesian Optimization (BO) presents an advanced alternative to traditional trial-and-error and design of experiments approaches for navigating design spaces. Notably, BO has been validated as a statistically robust method for optimizing material properties across a variety of materials [17, 18, 19, 20, 21, 22, 23, 24]. Central to BO are surrogate models, which are pivotal for modeling black-box functions that are expensive to evaluate and whose underlying functional forms remain unknown [25]. This capability makes BO exceptionally suitable for challenges in dental materials research, where precise mathematical models that map formulation space to property space are not well-defined. BO demonstrates its strength in situations where the material property or goal is explicitly specified; for instance, maximizing the yield strength of a dental composite or minimizing its shrinkage. Driving the success of BO are two critical components: the surrogate model, which estimates the black-box function, and the acquisition function, which strategically guides the selection of subsequent experiments in the design space for iterative exploration. This iterative approach, known as active learning, ensures the prioritization of the most informative experiments at each iteration, thereby enhancing the efficiency of the exploration process.

Bayesian Optimization (BO) presents a promising solution for both tasks at hand. Firstly, its data-driven approach accelerates the process of discovering novel formulations. Secondly, the inherent interpretability of the mathematical model utilized in BO enables scientists to grasp the contributions from each constituent with clarity. A third benefit that is particularly relevant to a formulation space relying on proprietary monomers is

A significant benefit of this approach is its applicability to the experimentation with new resins. Given that resins are often proprietary with undisclosed full structures, formulation scientists are constrained to trial-and-error methods due to limited chemical or domain-specific knowledge. BO circumvents this limitation by leveraging correlations between unknown components and experimental results, thereby streamlining the investigation of new resins as they become available.

The integration of machine learning with dentistry, particularly in the domain of dental materials, is progressing slowly due to a lack of widespread awareness and demonstrated applications. Our paper addresses this gap through a unified approach, structured into three core sections: data analysis, model construction, and experimental validation. In data analysis, we leverage domain knowledge to select optimal seed data and establish constraints, laying a foundation for focused exploration. The section on model construction dives into training our model within an active learning framework powered by Bayesian Optimization (BO), refining our search for optimal formulations. Finally, experimental validation tests the model's predictions against actual synthesis, using new data to iteratively enhance the model's accuracy and relevance. This concise yet comprehensive exploration aims to showcase the potential of BO and active learning in advancing dental material sciences, encouraging further exploration and adoption of these advanced methodologies in the field.

2 Methods

2.1 Data Analysis: Evaluate Curated Lab Experiments To Incorporate Domain Knowledge Into Model

Proprietary experimental dental adhesive data was collected from 91 experiments done at Ultradent Products, Inc. The target property for these adhesive materials is shear bond to dentin (MPa). The chemical space consisted of 43 different types of chemical ingredients that were sampled throughout 91 experiments and a breakdown of each type can be seen in Table 1.

Table 1: Count of Chemical Class In Curated Dataset

Chemical	**Count**
Resins	23
Fillers	4
Initiators	4
Inhibitors	1
Pigments	1
Solvents and Additives	3
Co-initiators/Catalysts	4
Plasticizers/Modifiers	1
Other Chemicals	2
Total	**43**

The mean value for shear bond across the dataset after 91 experiments was 43.02 MPa, with the 75th percentile of the values being less than or equal to 49.89 MPa. Further statistically relevant values can be found in Table 2. The formulation with the maximum shear bond identified in the dataset across the 91 experiments had a value of 63.07 MPa. Shear bond distribution across the curated dataset can be seen in Figure 1.

Table 2: Statistical Summary of Shear Bond In Curated Dataset

Statistic	**Value**
Count	91
Mean	43.02
Standard Deviation	10.29
Minimum	18.17
25th Percentile	37.81
Median (50th Percentile)	45.21
75th Percentile	49.89
Maximum	63.07

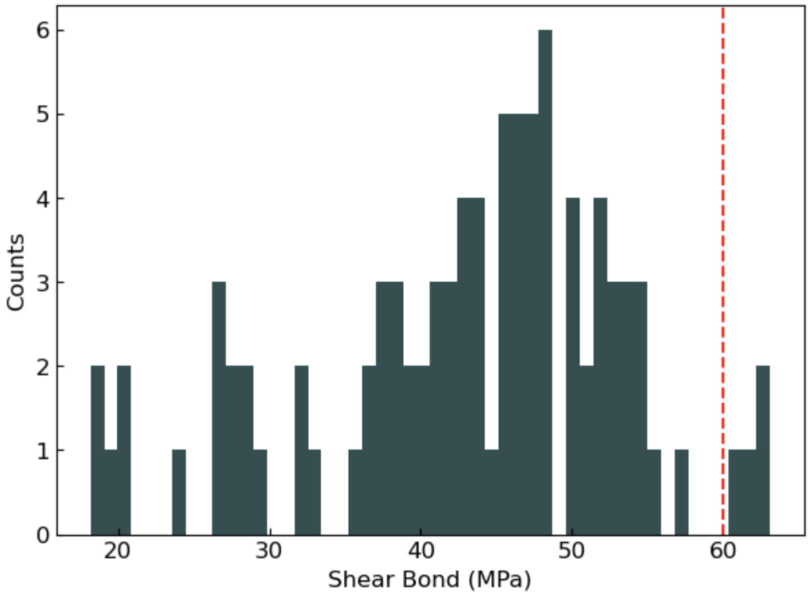

Figure 1: Distribution Of Shear Bond From 91 Curated Experiments.

To incorporate domain knowledge into our optimization runs, we further analyzed the 91 experiments. First, we sorted the dataset based on the top 10 highest performing adhesives with respect to shear bond, selecting the top 10 formulations, as can be seen in the python code below. We then removed any feature that did not contribute to the highest performing samples. This reduced the design space to a more viable search space of 9 features; 5 resins, 2 solvents, 1 initiator, 1 filler. To avoid random sampling, we initialized our model with 6 samples; the formulation with the highest shear bond along with 5 Sobol samples, where Sobol sampling has been shown to provide faster model convergence as compared to random sampling [26]. Next, we sought to constrain the design space by limiting certain features that would lead to unviable dental adhesives, for example limiting filler concentrations to avoid over filled dental adhesives that would lead to unworkable materials. The bounds on the each chemical parameter was identified through finding the min and max from the original experiments, and adding a buffer of 5%. Doing this incorporates the formulation scientists domain knowledge's expertise into the model.

```
# get top 10 highest performing formulations
top_ten_df = data.sort_values(by=["Shear Bond dentin (MPa)"], ascending=False).
    head(10)

# get col names of non-empty cols
top_ten_lst = [col for col in top_ten_df.columns if top_ten_df[col].sum() != 0]
# drop all zero cols in top 10 formulations
top_ten_df_no_zeros = top_ten_df[top_ten_lst]
```

2.1.1 Model Set-Up: Active Learning Via Bayesian Optimization

First, we understand that most dental researchers are not familiar with active learning and Bayesian Optimization. Thus, the code is provided in a GitHub repository in order to help introduce the dental research community to applying this method. Next we precisely layout our strategy and provide key insights into the model used to help provide a more detailed understanding of how the design space is modeled and how samples are selected.

Sample selection was conducted in a batch active learning cycle using Bayesian optimization [27]. Active learning is a process by which we iteratively update our surrogate model, the Gaussian process, at each experiment, leveraging the information gained from the latest batch of experiments to guide the selection of new experiments. By doing so, we reduce the surrogate models uncertainty when making new predictions on unseen samples. The selection of new samples is guided by an acquisition function, in our case we chose an expected improvement acquisition function that balances exploring regions of the design space that are unseen, with exploitation, areas that have yielded high performance in previous experiments.

In our approach, we utilize the Adaptive Experimentation Platform (Ax) to conduct Bayesian optimization. We selected a fully Bayesian model that enables inference of hyperparameters from their posterior distribution. This differs from more traditional approaches that rely on maximum a posteriori probability (MAP) estimates, where only a single value is determined for each hyperparameter. By sampling from the posterior distribution of the hyperparameters, we account for uncertainty in the hyperparameters, thereby constructing a more robust surrogate model.

Here we provide the model workings, but highly recommend the reader to [28] for a deeper dive into the inner workings of the model. In our optimization scheme, we sample functions from a Gaussian Process (GP) that is parameterized by a mean function ($m(\mathbf{x})$) and covariance function ($k(\mathbf{x}, \mathbf{x}')$):

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right) \tag{1}$$

We utilize a constant mean function $m(\mathbf{x})$ and a Matérn-5/2 kernel $k(\mathbf{x}, \mathbf{x}')$ for our GP:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_k^2 \left(1 + \sqrt{5}r + \frac{5}{3}r^2\right) \exp\left(-\sqrt{5}r\right) \tag{2}$$

Where r is given by:

$$r = \sqrt{\sum_{i=1}^{d} \frac{(x_i - x_i')^2}{\rho_i^2}} \tag{3}$$

The GP utilizes a Sparse Axis-Aligned Subspaces (SAAS) function prior that has been shown to outperform in high-dimensional spaces by "turning off" dimensions deemed unimportant. This is achieved via the global shrinkage parameter, τ, which is drawn from a Half-Cauchy distribution that tends to concentrate its mass near zero:

$$\tau \sim \text{HalfCauchy}(\alpha) \tag{4}$$

τ then serves as the scaling hyperparameter for the Half-Cauchy distribution from which the length scales ρ_i are drawn. Small values of τ encourage the length scales to be near zero, leading to a sparse covariance matrix and reducing the influence of certain dimensions in the model:

$$\rho_i \sim \text{HalfCauchy}(\tau) \text{ for } i = 1, \ldots, D \tag{5}$$

This sparsity in the covariance matrix helps the model focus on the most relevant dimensions for predicting material properties, improving interpretability and performance in high-dimensional settings.

As previously stated, the acquisition function used to guide the search for the next best points to sample was Expected Improvement (EI). We utilized a batched approach to generate 3 samples at each iteration. By using EI in a batch setting, we aimed to make informed decisions about which points to evaluate next, taking into account both the uncertainty and the potential improvement in the material property. After each round of batch trials were synthesized and characterized, the model was updated with the latest objective values, and a new batch was created.

2.1.2 Formulations: Synthesis & Characterization

Synthesis of the dental adhesive materials were done by Ultradent formulation scientists using 30g batches per sample. Formulations were weighed, and mixing was done using a FlackTek speedmixer to ensure uniformity. Seven samples from a 30g batch were taken for testing each new formulation. The mean value of the 7 samples used to measure shear bond to dentin was taken as the value of the materials shear bond. The standard followed to test these new formulations was done per the International Organization for Standardization 29022 (ISO 29022).

3 Results & Discussion

Our approach to finding new dental adhesive formulations yielded significant results, which we believe will open the door for formulations scientists to incorporate Bayesian optimization and domain knowledge with the dental adhesives framework for discovery. We begin by comparing the 91 experiments done in the curated dataset vs the trials done with Bayesian optimization as seen in Figure 2. The x-axis indicates the number of iterations / samples made in lab. The y-axis indicates the target property of interest, in our case shear bond. The formulations in dark green were done strictly using domain knowledge expertise and were sampled from 43 different possible inputs. We see that over the span of 91 experiments, the highest performing sample was yielded at experiment 80. In red is our active learning cycle via Bayesian optimization. The first value in our optimization scheme was seeded with the highest performing sample in the curated dataset, followed by 5 iterations of SOBOL samples. Here, after curating the legacy dataset and identifying the most influential combinations of chemicals we see that within 29 experiments performed via active learning, we find multiple comparable high performing formulations. The significance of this result is multi-faceted.

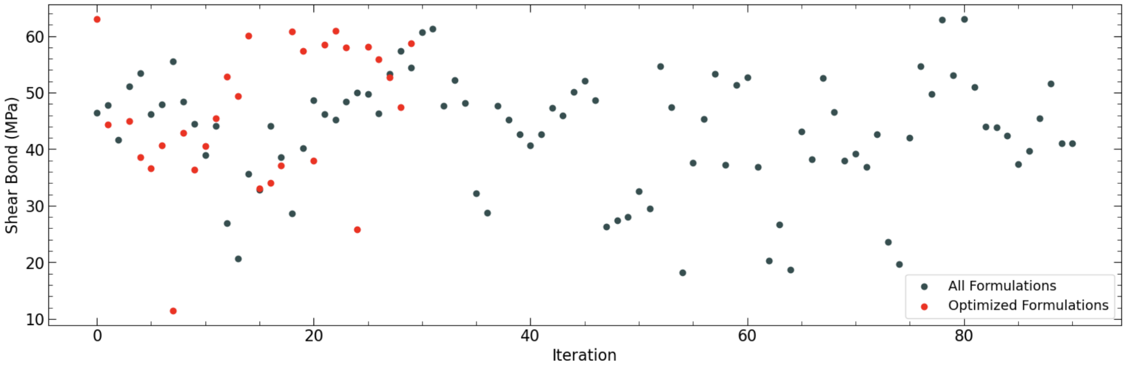

Figure 2: Original Experiments vs Active Learning.

First, we managed to find not just a single high performing formulation with respect to shear bond strength, but multiple high performing formulations. This is shown in Table 3, by the increase in the statistical value of 75th percentile, from 49.89 in the original curated dataset (seen in table 2), to 58.00. This indicates that 25% of the samples in the active learning cycle, within 29 trials were above 58.00 MPa in shear bond strength. Secondly, the time taken to realize these high performing samples was significantly decreased. Using active learning we managed to cut down the sample generation by 68%. To put this in perspective, each experiment roughly takes an 2 hours to go from weighing a sample, mixing, to characterization. By reducing the sample generation we managed to save 124 labor hours. Aside from labor hours, we also significantly reduce waste, leading to a greener process.

Table 3: Statistical Summary of Shear Bond In Active Learning Cycle.

Statistic	**Value**
Count	29
Mean	46.33
Standard Deviation	12.29
Minimum	11.50
25th Percentile	37.95
Median (50th Percentile)	45.43
75th Percentile	58.00
Maximum	63.07

Next we took a closer look at the design space. Here, we performed a dimensional reduction technique known as UMAP (Uniform Manifold Approximation and Projection). When using UMAP it is important to note when comparing the legacy data and the new samples in that the axis are arbitrary and non-interpretable. By reducing the design space this allows us to visualize our sample space on a 2 dimensional plot, preserving *local* and *global* structure. First we take a look at the original sampled space from the curated 91 sample dataset. Here we plot all formulations in teal, the top 5 formulations in orange, and the bottom 5 formulations in purple, as seen in Figure 3. As can be seen from the 91 original experiments, local formulation optimization was primarily used. One glaring issue with this, is despite having performed 91 experiments, the sample space is still relatively unexplored. Another important aspect of this figure is that a majority of smalls were made in regions that indicate low performance with respect to shear bond. This can be avoided when using an more probabilistic approach and having the model learn from areas that are low performing as will be seen in the UMAP figure for our Bayesian optimized sample set.

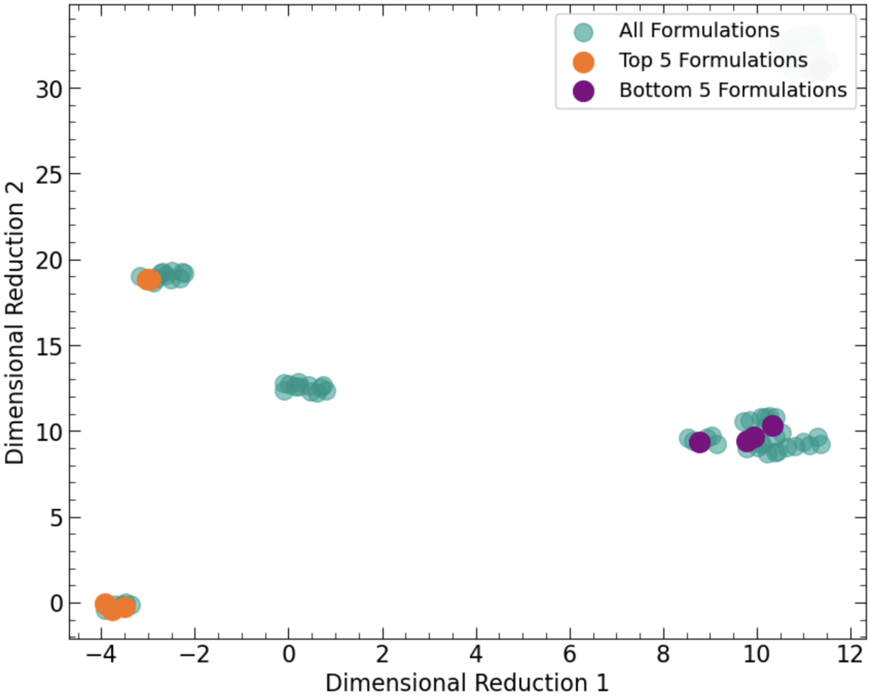

Figure 3: Uniform Manifold Approximation and Projection Of 91 Experiments.

Comparing this to the iterative active learning cycle, seen in Figure 4, where each point is carefully selected to choose the most optimal next sample, we see that within 29 experiments more of the design space was explored. Here, Each experiment in this figure is labeled with 2 numbers within the parenthesis. The first number indicates the sample number performed in the iteration loop, and the second is the shear bond value. From Figure 4, we see that we begin to identify high performance regions in the design space, as can be seen in the top left section of the plot. Another important factor

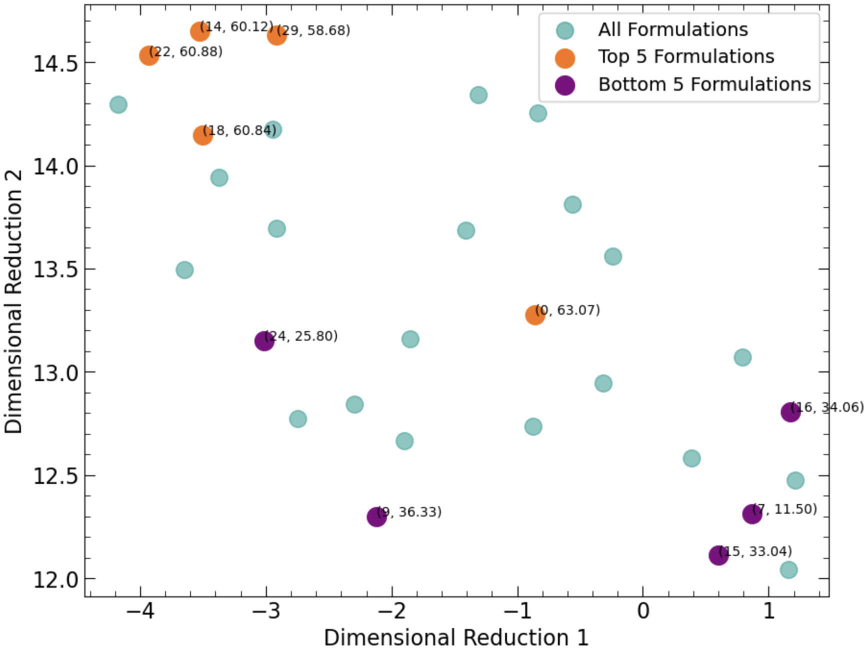

Figure 4: Uniform Manifold Approximation and Projection Of Active Learning Cycle.

identified in this figure, is the distance the high performing region is found from the initial seed labeled at (0,63.07). As a reminder, the initial seed was the highest performing sample found in the curated dataset. This distance of the high performing cluster from the original seed, indicates to us that we managed to find novel high performing formulations, rather than optimizing around the already known region. We also see that once a low performing region is identified, the model tends to move away and not resample multiple times from the same region. This in turn leads to a much more cost efficient process, as well yielding multiple high performing samples at a much faster rate.

Lastly, we take a look at the distribution of the 29 experiments done using Bayesian optimization in Figure 5. From this distribution we see that we have managed to successfully predict and synthesize 3 novel formulations (accounting for the removal of the initial high performing seed), that have exceptional shear bonding strength to dentin. This result is highly significant because we have shown the ability to find novel formulations that can allow dental materials researchers to find multiple exceptional materials at a faster rate. Another important factor is that this realization can bring dental materials to market at a cheaper cost due to the reduction in experiments. From this we hope to help propagate the use of machine learning and techniques such as Bayesian optimization coupled with active learning to drive dental materials research.

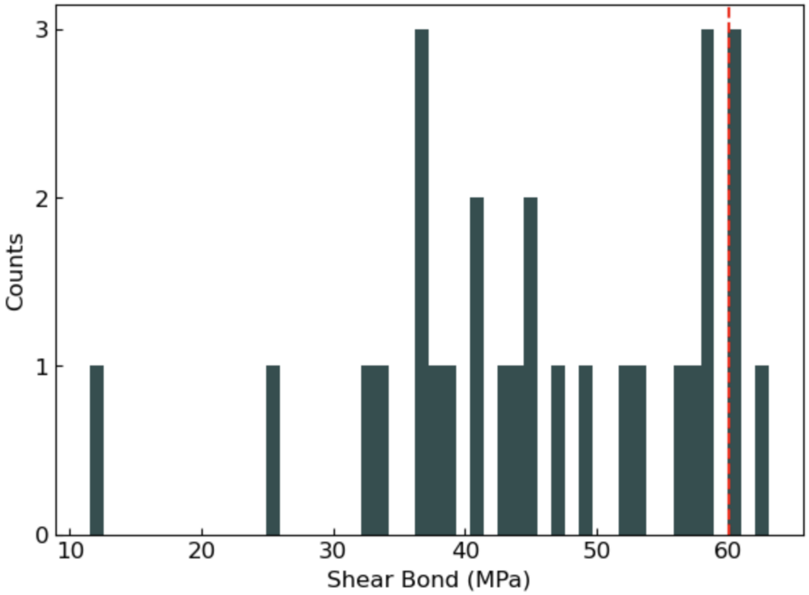

Figure 5: Distribution Of Shear Bond From 29 Active Learning Samples.

4 Highlights

- Dental adhesive formulations can be optimized rapidly by leveraging machine learning and Bayesian optimization.
- Leveraging experimental data from previous experiments can help reduce the design space and incorporate domain knowledge into the optimization cycle.
- A machine learning approach for optimizing dental materials properties adds a versatile component that can be adapted to discover a multitude of dental materials quickly and efficiently.
- Our approach leads to enhanced performance and reduced development times, resulting in more cost-effective and innovative solutions in dental healthcare.

5 Acknowledgements

The authors would like to thank the entire R&D chemical team at Ultradent Products Inc. for their support and collaboration. Additionally, we acknowledge Ultradent Products, Inc. for their generous support of this research.

References

[1] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim. Machine learning in materials informatics: recent applications and prospects. 3(1):1–13.

[2] Chen Li and Kun Zheng. Methods, progresses, and opportunities of materials informatics. 5(8):e12425.

[3] Sterling G. Baird, Marianne Liu, Hasan M. Sayeed, and Taylor D. Sparks. Data-driven materials discovery and synthesis using machine learning methods. pages 3–23.

[4] Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon J. L. Billinge, Elizabeth Holm, Shyue Ping Ong, and Chris Wolverton. Recent advances and applications of deep learning methods in materials science. 8(1):1–26.

[5] Ankit Agrawal and Alok Choudhary. Deep materials informatics: Applications of deep learning in materials science. 9(3):779–792.

[6] Angelo Ziletti, Devinder Kumar, Matthias Scheffler, and Luca M. Ghiringhelli. Insightful classification of crystal structures using deep learning. 9(1):2775.

[7] Lei Zhang and Shaofeng Shao. Image-based machine learning for materials science. 132(10):100701.

[8] M. Ge, F. Su, Z. Zhao, and D. Su. Deep learning analysis on microscopic imaging in materials science. 11:100087.

[9] Kamal Choudhary and Mathew L. Kelley. ChemNLP: A natural language-processing-based library for materials chemistry text data.

[10] Elsa A. Olivetti, Jacqueline M. Cole, Edward Kim, Olga Kononova, Gerbrand Ceder, Thomas Yong-Jin Han, and Anna M. Hiszpanski. Data-driven materials research enabled by natural language processing and information extraction. 7(4):041317.

[11] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. 571(7763):95–98.

[12] Satoshi Yamaguchi, Hefei Li, and Satoshi Imazato. Materials informatics for developing new restorative dental materials: A narrative review. *Frontiers in Dental Medicine*, 4, 2023.

[13] F. Schwendicke, W. Samek, and J. Krois. Artificial intelligence in dentistry: Chances and challenges. 99(7):769–774.

[14] Zinovia Surlari, Dana Gabriela BudalƒÉ, Costin Iulian Lupu, Carmen Gabriela Stelea, Oana Maria Butnaru, and Ionut Luchian. Current progress and challenges of using artificial intelligence in clinical dentistry,Äîa narrative review. 12(23):7378.

[15] Alireza Aminoroaya, Rasoul Esmaeely Neisiany, Saied Nouri Khorasani, Parisa Panahi, Oisik Das, Henning Madry, Magali Cucchiarini, and Seeram Ramakrishna. A review of dental composites: Challenges, chemistry aspects, filler influences, and future insights. *Composites Part B: Engineering*, 216:108852, July 2021.

[16] Jorge Perdig£o. Current perspectives on dental adhesion: (1) dentin adhesion ,Äì not there yet. 56(1):190–207.

[17] Lars Kotthoff, Hud Wahab, and Patrick Johnson. Bayesian optimization in materials science: A survey.

[18] Hud Wahab, Vivek Jain, Alexander Scott Tyrrell, Michael Alan Seas, Lars Kotthoff, and Patrick Alfred Johnson. Machine-learning-assisted fabrication: Bayesian optimization of laser-induced graphene patterning using in-situ raman analysis. 167:609–619. publisher: Elsevier Ltd.

[19] Yuki K. Wakabayashi, Takuma Otsuka, Yoshiharu Krockenberger, Hiroshi Sawada, Yoshitaka Taniyasu, and Hideki Yamamoto. Machine-learning-assisted thin-film growth: Bayesian optimization in molecular beam epitaxy of SrRuO3 thin films. 7(10). publisher: AIP Publishing, LLC.

[20] Kensaku Nakamura, Naoya Otani, and Tetsuya Koike. Search for oxide glass compositions using bayesian optimization with elemental-property-based descriptors. 128(8):569–572.

[21] Danial Khatamsaz, Brent Vela, Prashant Singh, Duane D. Johnson, Douglas Allaire, and Raymundo Arryave. Multi-objective materials bayesian optimization with active learning of design constraints: Design of ductile refractory multi-principal-element alloys. 236:118133.

[22] Trupti Mohanty, K. S. Ravi Chandran, and Taylor D. Sparks. Machine learning guided optimal composition selection of niobium alloys for high temperature applications. 1(3):036102.

[23] Cheng Wen, Yan Zhang, Changxin Wang, Dezhen Xue, Yang Bai, Stoichko Antonov, Lanhong Dai, Turab Lookman, and Yanjing Su. Machine learning assisted design of high entropy alloys with desired property. 170:109–117.

[24] Alireza Vahid, Santu Rana, Sunil Gupta, Pratibha Vellanki, Svetha Venkatesh, and Thomas Dorin. New bayesian-optimization-based design of high-strength 7xxx-series alloys from recycled aluminum. 70(11):2704–2709.

[25] Bowen Lei, Tanner Quinn Kirk, Anirban Bhattacharya, Debdeep Pati, Xiaoning Qian, Raymundo Arroyave, and Bani K. Mallick. Bayesian optimization with adaptive surrogate models for automated experimental design. 7(1):1–12.

[26] Marissa Renardy, Louis R. Joslyn, Jess A. Millar, and Denise E. Kirschner. To sobol or not to sobol? the effects of sampling schemes in systems biology applications. 337:108593.

[27] Francesco Di Fiore, Michela Nardelli, and Laura Mainini. Active learning and bayesian optimization: a unified perspective to learn with a goal.

[28] David Eriksson and Martin Jankowiak. High-dimensional bayesian optimization with sparse axis-aligned subspaces. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 493–503. PMLR.