

SemiConLens: Visual Analytics for 2D Semiconductor Discovery

Kavinda Athapaththu^{1,2}, Shiwei Chen¹, Yuan Fang², Sanchali Mitra³, Yee Sin Ang³, Yong Wang^{1,2†}

¹Nanyang Technological University

²Singapore Management University

³Singapore University of Technology and Design

Abstract

The past few years have witnessed vibrant efforts in discovering new two-dimensional (2D) semiconductor materials from both academia and the industry, due to their promising potential in resolving the severe performance deterioration of traditional semiconductors resulting from condensed silicon thickness. However, existing methods (e.g., Density Functional Theory (DFT) or machine-learning-based approaches) suffer from various challenges such as small datasets, and reliability and trustworthiness issues. To bridge this gap, we propose SemiConLens, a visual analytics approach to combine human expertise with the power of ML to enable effective and reliable 2D semiconductor discovery. Specifically, we first develop a new Correlation Aware Multivariate Imputation (CAMI) method and use ML models like autoencoder, which can better learn from limited data and reveal uncertainty, to address the challenge of sparse data in semiconductor prediction. Built upon this, our visualization module, consisting of three visualization views with linked interactions, allows material researchers to interactively filter, discover and compare 2D semiconductor candidates. A novel circular glyph design and a new cluster-aware layout optimization approach are proposed to effectively display all the user-configurable key attributes and possible prediction uncertainties of each semiconductor candidate, ensuring a reliable and trustable 2D semiconductor discovery. We assess SemiConLens through quantitative evaluations, expert interviews, and use cases. The results demonstrate SemiConLens's capability to help material researchers conduct effective discovery of desirable 2D semiconductors.

CCS Concepts

• *Human-centered computing* → *Visualization systems and tools*;

1. Introduction

Semiconductors are essential to modern technologies in areas such as quantum computing, communication, health care, energy, transportation, and defense, with the global market reaching 588.4 billion USD in 2024 and projected to grow to 654.7 billion USD in 2025 [Sem23, Sem24]. A major challenge is the performance degradation of silicon transistors at nanometer-scale thicknesses [IDK*04, LSWL19]. The discovery of graphene in 2004 [NGM*04] showed that two-dimensional semiconductors, semiconductors only a few atoms thick, can overcome these limits and offer superior performance [LSWL19]. As a result, discovering new 2D semiconductor materials with desirable properties has become a highly active research area [GTR*21, ZWML23, LLW*21], with major industry players such as TSMC, Intel, Nvidia, IBM, and Samsung investing heavily in this direction [Wor23].

However, 2D semiconductor discovery is inherently a multi-criteria and high-dimensional problem [JOH*13]. Existing methods for 2D semiconductor material discovery can be categorized

into two groups: *traditional experimental approaches* [BDC*18, OM18, CBHvD*18] and *computational approaches* [LLW*21, GTR*21].

Traditional experimental workflows depend on researchers synthesizing candidate materials based on domain knowledge and then measuring their properties to evaluate suitability. Because this process is fundamentally trial-and-error, it is slow and requires extensive manual effort [CZZ20, CBHvD*18, LZJS17, PSG*21]. To mitigate this issue, materials science researchers have also developed a series of computational approaches for 2D material discovery, which leverage theoretical models, simulations, and Machine Learning (ML) techniques to predict new 2D materials with desirable properties before experimental synthesis [GTR*21, ZSC*19, CTP*21, LLW*21, BDC*18, KGK*19]. Representative computational approaches include: 1) using Density Functional Theory (DFT) for calculating material attributes, and 2) employing ML to predict 2D material attributes by training on existing material data. DFT is a quantum mechanical simulation technique that can predict the electronic, optical and structural attributes of materials. However, with increasing material size, the methods become *computationally expensive and slow* [LLW*21, ZWML23].

† Corresponding author: yong-wang@ntu.edu.sg

ML-based approaches have been increasingly popular for 2D semiconductor attribute prediction [MAK*14, BDC*18, LXL*22], but ML-based approaches often require a large amount of labeled data to train the ML models to be effective.

A fundamental challenge in 2D material discovery is **data sparsity**, as materials science datasets often contain **limited labeled samples** [SCR*24, BDC*18, ZL18, CWE22], restricting the prediction performance of ML-based approaches. For example, the C2DB dataset [GTR*21], one of the most popular 2D material datasets, has a total number of 16,905 compounds, but there are only 339 compounds with a given value for CBM GW and VBM GW, two important attributes to indicate the efficiency of semiconducting material in high-speed electronic devices. Also, existing machine learning approaches, especially the deep neural network models, often work like a black box, negatively affecting the **reliability and trustworthiness** of these approaches [MQL*24].

To address the above challenges, we propose *SemiConLens*, a novel visual analytics approach to help material researchers with effective 2D semiconductor discovery by seamlessly combining the power of ML with their materials science expertise. *SemiConLens* consists of two modules: **sparsity-aware prediction** and **visualization**. Our semiconductivity prediction module explicitly considers the sparsity of material data labels by developing a Correlation Aware Multivariate Imputation (CAMI) method to fill the missing attribute values.

Built upon this, our visualization module allows material researchers to interactively analyze a large number of 2D compound materials, and easily compare and identify high-potential semiconductors. The system comprises three coordinated views: *Filter View* to display the compound attribute distribution and user exploration history; *Discovery View* to visualize filtered compounds and the similarities among them; and *Comparison View* to facilitate detailed comparison and investigation of selected 2D semiconductor candidates. Specifically, we propose a novel circular glyph to encode material attributes together with associated prediction uncertainties, and a new cluster-aware layout optimization technique to preserve inter and intra cluster similarity while improving readability in dense layouts. Together, these components support informed interactive exploration and transparent decision making in 2D semiconductor discovery. We evaluate the effectiveness and usability of *SemiConLens* through quantitative metric evaluations, expert interviews, and two use cases. The major contributions of this paper are as follows:

- We develop a new Correlation Aware Multivariate Imputation (CAMI) technique to impute the missing attribute values in the material dataset, reducing the influence of data sparsity challenge in the materials science field.
- We propose *SemiConLens*, a visual analytics approach for 2D semiconductor discovery, which integrates sparsity-aware prediction and interactive visualizations for effective and trustworthy 2D semiconductor discovery. A novel circular glyph design and cluster-aware glyph layout optimization are presented.
- We evaluate *SemiConLens* via quantitative metric evaluations, expert interviews, and use cases. The results demonstrate that *SemiConLens* is both effective and user-friendly for aiding material experts in discovering high-potential 2D semiconductors.

2. Background

In this section, we introduce materials science background knowledge essential to understand this research including desirable semiconductor attributes and different methods to calculate the above.

2.1. Desirable Semiconductor Properties

Carrier Mobility shows how fast an electron (*Electron Mobility*) or a hole (*Hole Mobility*) can travel in a semiconductor when an external electric field is applied. It determines properties of semiconductor devices like switching frequency in transistors [PLRG20].

For any semiconductor, there is an energy region where there can be no states. The energy bands below are known as Valence Bands while energy bands above are known as Conduction Bands. The difference between the lowest conduction band, which is the *Conduction Band Minimum (CBM)* and the highest valence band, which is the *Valence Band Maximum (VBM)*, is known as the *Bandgap* [SLN21]. Electrons in the valence band are bound to atoms while the electrons in the conduction band are free to conduct electricity [SB*00]. Suitable bandgap, high carrier mobility and stability at room temperature are the desirable properties of semiconductors [ZWML23].

2.2. Density Functional Theory (DFT) and beyond DFT

DFT is a widely used computational method for studying electron distribution in materials. Rather than solving the full Schrödinger equation for every electron, it models the overall electron density, making calculations more feasible while remaining accurate for many ground-state properties [Mar20]. Using DFT, semiconductor properties such as carrier mobility and bandgap can be estimated, but the computations are both expensive and time consuming.

Within DFT, the **Perdew–Burke–Ernzerhof (PBE)** functional [PBE96] is widely used, improving upon earlier methods by accounting for spatial variations in electron density and yielding more realistic results than the local density approximation. The **Heyd–Scuseria–Ernzerhof (HSE)** functional [HSE03] further enhances accuracy by mixing a portion of exact exchange from Hartree–Fock theory with DFT, allowing short- and long-range electron interactions to be treated differently. For even higher accuracy, particularly for electronic excitations, the **GW approximation** [SZL10] extends beyond standard DFT; here, “G” describes electron or hole motion while “W” captures their interactions, producing results that often match experimental measurements more closely than PBE or HSE [SZL10, HSE03, PBE96].

3. Related Work

Our work is related to prior research on ML for material discovery, visualization for material discovery, and high-dimensional data visualization.

3.1. ML for Material Discovery

ML has accelerated materials discovery by enabling rapid screening and prediction of novel compounds. ML approaches are commonly categorized by their reliance on labeled data, as supervised,

semi-supervised, and unsupervised learning [CZZ20], or by the primary data mining task such as dimension reduction, clustering, classification, and correlation [ERT*17]. Among these, supervised learning is most widely used for predicting material properties, identifying stable compounds, and designing new materials [CZZ20, LZJS17].

Supervised algorithms including Random Forest (RF) [LXL*22], Support Vector Regression (SVR) [CZZ20], Multi-Layer Perceptron (MLP) [LZO*18], Decision Tree (DT) [ZWML23, LZO*18], and Gradient Boost (GB) [KKG*19, LZO*18] have been applied to tasks such as predicting ionic conductivity [PSG*21], designing lead-free perovskites [LZO*18], and classifying crystal structures [OM18]. Ensemble methods like GB are particularly effective for uncovering nonlinear relationships in high-dimensional datasets [KKG*19]. However, no single model consistently outperforms others due to the complexity of materials data [CZZ20].

Unsupervised and semi-supervised methods help leverage large unlabeled datasets in materials discovery. Techniques like k-means and t-SNE group materials with similar properties [GTR*21], while Auto Encoders (AE) and other neural networks have shown advantages in extracting complex structures from high-throughput materials datasets and enabling accelerated screening of candidate materials [YAK*23, LLW*21]. Semi-supervised methods further combine small labeled datasets with abundant unlabeled data to predict properties such as phase stability and thermodynamics [CTP*21].

Although previous studies have explored various applications of machine learning in the materials science domain, none have been specifically tailored for the discovery of 2D semiconductors. Furthermore, to the best of our knowledge, the use of few-shot learning with autoencoders has not been investigated in this context.

3.2. Visualization for Material Discovery

Visualization offers valuable support in materials discovery, where researchers must navigate complex, high-dimensional datasets to identify promising candidates [ERT*17, GBM*19]. While still emerging in this domain, visual analytics approaches have shown potential to enhance the exploration and analysis of material properties [MNS*23, BXB*18, PSG*21].

Early visualization systems mainly supported basic property inspection and relationship discovery. Martínez et al. [MNS*23] employed scatter plots and scatter matrices, but these techniques were insufficient for high-dimensional materials data. Visualization also assisted model interpretation, as seen in Phase Mapper [BXB*18], which connected model predictions with expert assessment using simple visual tools. MatExplorer [PSG*21] further integrated predictive modeling with interactive exploration for solid-state electrolytes. However, it remains inadequate for semiconductor discovery because it does not enable progressive exploration of large-scale datasets and lacks support for in-depth examination of influential attributes.

To overcome these limitations, this paper introduces a visual analytics system that enables systematic exploration of 2D semiconductor materials through interactive visualizations, progressive filtering, and uncertainty representation. The system incorporates a

glyph-based design for examining semiconductor properties and a cluster-aware layout optimization method that combines cartograms with force-directed layouts to minimize overlap while preserving cluster structure.

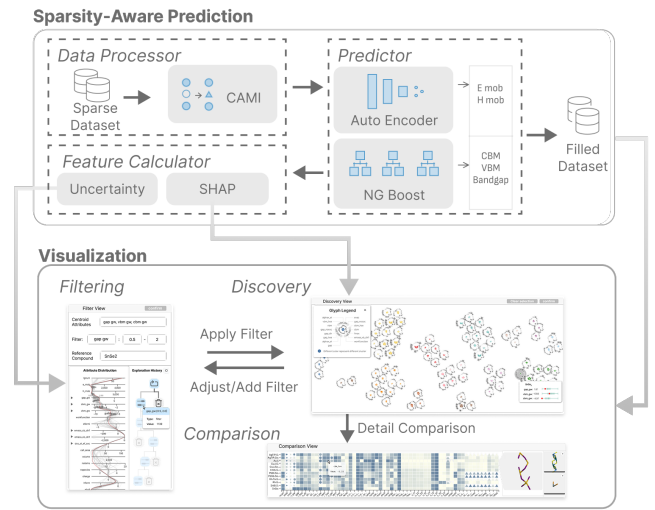


Figure 1: The workflow consists of two main components: (1) a sparsity-aware prediction pipeline that processes sparse datasets through CAMI data processing, followed by prediction using Auto Encoder and NG Boost models, with uncertainty and SHAP-based feature calculation for interpretability; (2) a visual analysis system, which enables interactive *filtering*, *exploration* and *comparison* of the filled dataset.

3.3. High Dimensional Data Visualization

High-dimensional data visualizations aim to improve the interpretability of complex multivariate datasets. Among them, glyph-based approaches are mainstream techniques, offering effective information preservation while reducing cognitive load.

For glyph design, multiple dimensions are encoded into a single visual object using properties such as color, shape, and size [CHUW08, BKC*13, ERT*17, SUS*20]. Different designs serve specific analytical purposes: star plots highlight extremal values, Chernoff faces exploit humans' sensitivity to facial feature differences, and radial encodings represent cyclical or hierarchical relationships [FFM*13]. Inspired by existing glyph designs, we adopted a radial encoding to illustrate key attributes and corresponding influencing factors in our visualizations.

For glyph layout, Dimension Reduction (DR) techniques are widely applied to project high-dimensional data into lower-dimensional spaces while preserving key relationships [BWT*25]. Common DR methods such as PCA [ERT*17], t-SNE [KKG*19], and UMAP [CZZ20] are widely used for glyph layout. However, DR approaches treat data points as dimensionless, ignoring the spatial requirements of glyphs, which causes overlap in dense regions, particularly for large datasets. Existing solutions such as force-directed layouts [Hu05], constraint-based techniques [ZYC*24], and space partitioning methods [HME*24] primarily address overlap reduction but do not explicitly preserve cluster structures in DR

projections. This paper presents a novel cluster-aware glyph layout optimization approach to address this challenge.

4. Tasks Analysis

Over more than 12 months, we worked closely with two domain experts, who are also co-authors, to identify challenges in 2D semiconductor discovery and iteratively refine our prototypes. Their continued feedback allowed us to distill six core analytical tasks that shaped the system's design.

T1. Identify high-potential models to predict compound attributes that are hard to calculate using DFT or other traditional methods. Traditional methods such as DFT struggle to compute several semiconductor attributes efficiently [LLW*21, ZWML23]. Thus, researchers have explored multiple ML models to predict these properties [MAK*14]. Since model performance varies across datasets and target attributes [PSG*21], it is essential to determine which ML models offer the highest predictive potential for these attributes.

T2. Visualize the uncertainty and rationales of the predictions. ML-based predictions inherently involve uncertainty, making it essential for material scientists to understand the reliability of the results [KGG*19, CTP*21]. This requires examining both the uncertainty in the predicted values and the underlying rationale for the model's decisions. Providing clear insight into model confidence and reasoning enables users to make informed judgments when interpreting prediction outcomes.

T3. Select key attributes flexibly and show influencing factors. Material scientists often prioritize different attributes depending on their application needs [PSG*21, LLW*21, KMM20]. Thus, enabling flexible selection and visualization of relevant attributes is essential. Moreover, examining relationships among these attributes helps identify key factors that influence material performance, supporting deeper analysis and more informed decision-making.

T4. Provide an overview of the attributes distributions. Understanding how attributes are distributed across compounds is vital for material scientists [PSG*21]. Distribution overviews enable the identification of outliers and provide insight into dataset-wide patterns. Clear visualizations of these distributions help users assess ranges, frequencies, and central tendencies, offering context for evaluating individual compounds. Such insights support the detection of promising regions in the material space and facilitate the development of data-driven hypotheses for further design and investigation.

T5. Enable progressive filtering exploration. For large datasets, material scientists must progressively narrow the search space through systematic compound filtering [PSG*21, BXB*18]. When clear criteria are available, users should be able to iteratively apply and refine multiple filter parameters. When such criteria are absent, the system should support non-parametric exploration and guide users by suggesting potential filtering strategies informed by attribute distributions and statistical patterns derived from exploratory analyses.

T6. Compare possible candidate compounds. After narrowing the search space, material scientists must compare candidate compounds across multiple attributes. The system should support both overall evaluation to determine the best-performing

compound and detailed, attribute-level comparisons to highlight individual strengths and weaknesses.

5. Sparsity-aware Prediction

This section summarizes our approach for predicting compound attributes, covering model selection, missing-value imputation, prediction and uncertainty estimation, and the data sources and processing procedures.

5.1. Correlation Aware Multivariate Imputation

The correlation between compound attributes and the similarity across different compound samples can facilitate the estimation of missing compound attributes, which, however, is not fully considered in by existing data imputation techniques. We propose a novel imputation method, Correlation Aware Multivariate Imputation (CAMI), which considers the above to iteratively impute the missing values. A sketch of the above method is as described in Algorithm 1 in Appendix A.7. Fig. 2 demonstrates a summarized example of the behavior of the CAMI algorithm (See Appendix A.6 for high resolution version). We used CAMI to impute missing data in the dataset, except for the desirable semiconductor properties discussed in section 2.1, which are predicted.

CAMI begins with a preprocessing step (Fig. 2^{A2}) where it selects columns that have less than a specified threshold $A\%$ of missing values and drops any columns exceeding this threshold to prevent noise from features that are largely incomplete, and the dropped columns are not imputed and will not be used in the system. Then, rows with less than $B\%$ missing values are selected for the same reason, removing and never filling rows that exceed this threshold. After preprocessing, CAMI computes the correlation matrix (c_m) using Pearson Correlation with pairwise complete observations, on the remaining data (Fig. 2^{A3}). For each target column Y containing missing values, the algorithm identifies the top C most correlated columns (K) using c_m (Fig. 2^{B1}). The missing values in these selected columns are then imputed using the mean of the observed values in each respective column (Fig. 2^{B2}).

Then, for each target column Y , the algorithm employs a similarity-based row selection and imputation. First, from X , the most correlated columns (K) and the column Y are selected (X') (Fig. 2^{B3}) and the subset of the dataset (X') is split into two subsets: L , containing rows where Y is missing, and M , containing rows where Y is observed (Fig. 2^{B4}). For each row P in L , the similarity to every row in M is computed based on the available features, using K-Nearest Neighbors (Fig. 2^{C1}). The algorithm then selects the top D most similar rows (O) from M (Fig. 2^{C2}) and estimates the missing value of Y in P using the mean value of Y from the selected similar rows (Fig. 2^{C3}). This ensures that missing values are imputed using contextually relevant instances, rather than relying solely on a global mean. However, this estimated value will not be used for subsequent similarity searches to prevent compounding errors. The similarity-based row selection and imputation is repeated to fill all the missing values from remaining columns and rows. The hyperparameters were tuned specifically for the dataset through an iterative exhaustive search procedure.

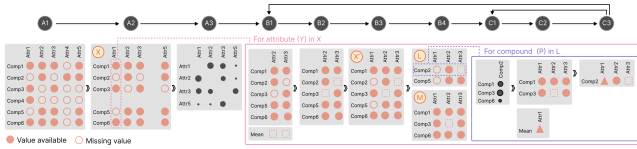


Figure 2: An illustration of the proposed Correlation Aware Multivariate Imputation. Step A_1 : Load the dataset. Step A_2 : Select only columns and rows with less than $A\%$ and $B\%$ missing values respectively. Step A_3 : Calculate the correlation matrix. Step B_1 : For each attribute (Y) in the filtered dataset X , select top C most correlated columns using the correlation matrix. Then calculate the mean for each selected column. Step B_2 : Fill the missing values of each column with the corresponding calculated mean. Step B_3 : Create X' , a subset of X , using the Y and its most correlated columns mentioned above. Step B_4 : split X' into L and M , where column Y is missing and not missing respectively. Step C_1 : For each compound (P) in L , compute similarity matrix with compounds in M . Step C_2 : Select the top D most similar compounds from the subset and calculate the mean for column Y . C_3 : Fill the missing value of P with the mean calculated above.

5.2. Model Choice

Based on the comparative evaluation results discussed in Table 1, AE was selected for predicting electron mobility and hole mobility, and Natural Gradient Boosting (NGBoost) model for CBM GW, VBM GW and Bandgap GW (T1). This decision was supported by the fact that the authors of NGBoost successfully tested the datasets with as few as 309 data points [DAD*20].

AE is a specialized neural network trained to compress and then reconstruct the inputs [RHW86]. It captures underlying patterns from large unlabeled datasets and also supports few-shot learning for predicting attributes when limited data is available [BBB*24]. NGBoost is a probabilistic prediction algorithm that extends gradient boosting to model full probability distributions rather than single point estimates, demonstrating strong performance on smaller datasets [DAD*20]. Both AE and NGBoost inherently provide uncertainty measures: AE through reconstruction error and NGBoost through predicted standard deviation. All models were trained using the preprocessed dataset, where missing values were imputed using the CAMI method.

5.3. Uncertainty and Feature Importance Estimation

As discussed in the introduction, the reliability and trustworthiness of ML models remain a concern. During the task analysis, we also identified the need to visualize uncertainty and the rationale behind predictions (T2). To support this, we compute both the prediction uncertainty and the feature importance for each model output.

Predicting semiconductor attributes. According to the previous results, two AE models were fine-tuned for the prediction of electron mobility and hole mobility. Bandgap GW, VBM GW, and CBM GW were predicted using three natural gradient boosting models.

Calculating uncertainty. Uncertainty was computed in two

ways. For AE based predictions, we use the reconstruction error prior to finetuning as the measure of uncertainty [PCCT14]. This error indicates how well the model reproduces the original input from its latent representation. Given an input vector $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and its reconstruction using the AE model $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n)$, the reconstruction error using Mean Absolute Percentage Error (MAPE) is defined as $\text{MAPE} = \text{Reconstruction error}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n} \sum_{i=1}^n \left| \frac{x_i - \hat{x}_i}{x_i} \right| \times 100\%$.

For natural gradient boosting models, we leverage Relative Standard Deviation (RSD) also known as the coefficient of variation, to calculate uncertainty of the predictions [CJS19, AZER21]. The NGB model inherently provides both the mean (μ) and standard deviation (σ) as outputs, representing the central estimate and associated uncertainty, respectively. The RSD can be defined as $\text{RSD} = \left(\frac{\sigma}{|\mu|} \right) \times 100\%$.

The uncertainty is not estimated for the imputed values, but they are clearly marked as imputed values in the visualization.

Calculating feature importance. To interpret model predictions and understand which features contribute most, we use SHAP (SHapley Additive exPlanations) [LL17], a game-theoretic approach to explain the output of machine learning models. First, local feature importance, specific to each compound, is calculated using SHAP values. Global feature importance is then derived by averaging these SHAP values across all compounds. This process is repeated for each of the five models. Detailed discussion about data sources and preprocessing can be found in Appendix A.2.

6. Visual Design

Based on the design tasks in section 4, we developed *SemiConLens* (Fig.3), an interactive visual analytics system with three coordinated views: Filter View, Discovery View, and Comparison View. Fig.1 showcases the overall workflow of our system (See Appendix A.3 for high resolution version). In this section, we present detailed descriptions of each view's design and interactions. A detailed discussion on interaction design can be found in Appendix A.1.

6.1. Filter View

The Filter View (Fig. 3A) is designed to enable flexible exploration, and provide an overview of the compound space (T3, T4, T5). The Filter View comprises three sub-views: Control Panel, Attribute Distribution, and Exploration History.

Control Panel (Fig. 3A1) serves as an interactive interface for attribute selection, filtering configuration, and reference semiconductor selection. It consists of three main components: (1) an attribute selector that enables users to choose multiple key attributes for detailed analysis in the Discovery View (T3), (2) a filtering criteria configuration panel that provides precise controls for configuring range filtering (T5), and (3) a reference semiconductor selector that allows users to incorporate known semiconductor materials as benchmarks. These components work in concert to support flexible exploration and precise filtering of the compound space.

Attribute Distribution (Fig. 3A2) integrates parallel coordinates plot and uncertainty visualization to reveal attribute relationships and prediction confidence. The visualization comprises two

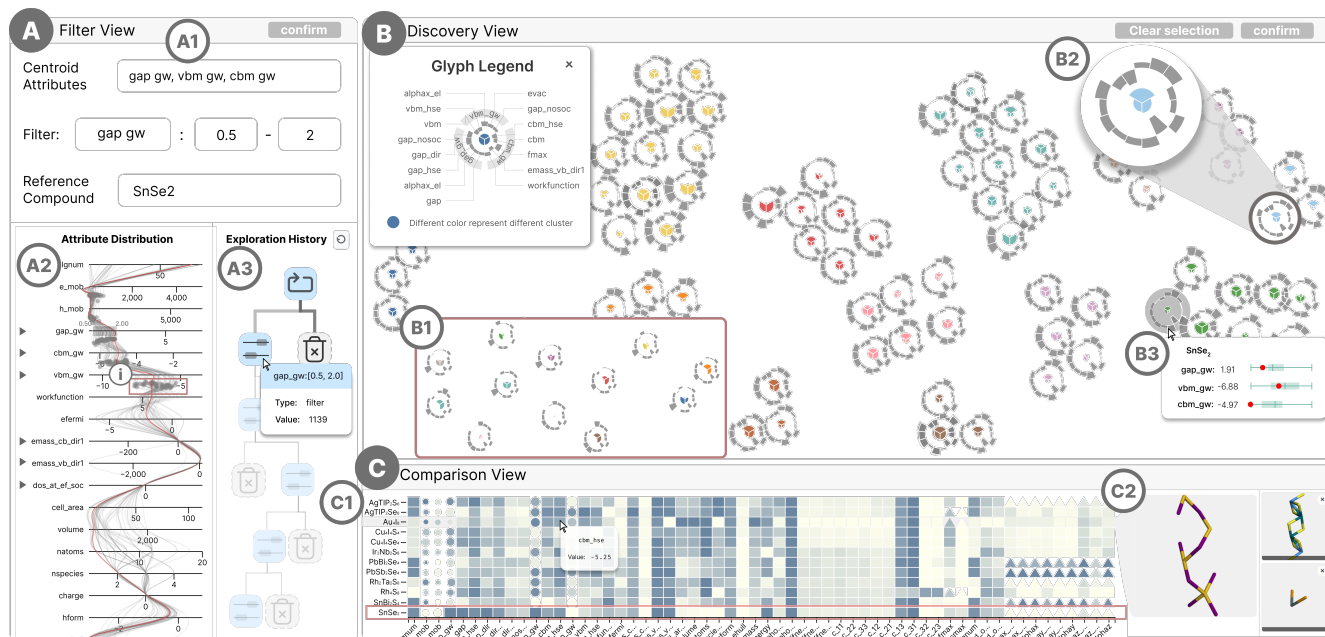


Figure 3: System interface for *SemiConLens*. Filter View **A** provides an overview of the compound space, applies range-based filters, and tracks exploration history for a more targeted and efficient analysis. Discovery View **B** visualizes clusters and influencing factors of the key attributes, enabling users to discover patterns in the compound space. Comparison View **C** facilitates the comparison of compounds across all attributes and displays their structures. See Appendix A.8 for high resolution version.

integrated parts: (1) A parallel coordinates plot that reveals distribution patterns and relationships among attributes. Each axis represents one attribute, and polylines connecting these axes illustrate potential correlations between attributes (T4). Users can interactively brush along any axis to define specific value ranges, dynamically updating the range filtering criteria (T5). To facilitate efficient exploration, we group attributes based on their physical representations (e.g., bandgap-related attributes like *gap*, *gap_gw*, and *gap_hse* are grouped together). Additionally, to maintain system responsiveness with large datasets, we adopted a density-based sampling strategy, which will be further discussed in section 6.2. (2) An embedded uncertainty visualization using scatter plots between adjacent axes (Fig. 3(i)). Each point in these plots represents a compound's uncertainty values (discussed in section 5.3) for the corresponding attribute pair, where the distance from the origin indicates uncertainty; points closer to the origin suggest higher model confidence. This uncertainty representation enables users to assess prediction reliability during their filtering process (T2).

The Exploration History (Fig. 3(A3)) supports tracking and navigation of previously applied filtering states (T5) using a tree diagram. Each filtering step is visualized as a transition, with edge color encoding the number of compounds retained. The system enables three filtering strategies: range filtering in the Filter View, and cluster and reference filtering in the Discovery View. Distinct icons represent range filtering (≡) and cluster/reference filtering (🔍).

To facilitate statistical assessment, Analysis of Variance (ANOVA) is integrated to determine the attribute with the highest variation when applying cluster and reference filtering, ensuring consistency across the three filtering strategies. Hover interactions on the tree diagram highlight the corresponding attribute distribu-

tions in the Attribute Distribution view, enabling users to compare distributions across different filtering stages and better understand the progressive refinement process.

6.2. Discovery View

The Discovery View (Fig. 3(B)) presents an overview of compounds within the current compound space, highlighting performance on user-selected key attributes. The compound space integrates original, AE-predicted, and CAMI-imputed values, focusing on similarity patterns, while detailed compound attribute comparison is addressed in the Comparison View (Sec. 6.3). This supports pattern exploration (T3, T4) and the identification of key attributes and their influencing factors (T3). A novel glyph-based representation is proposed to encode both key attribute values and influencing factors, and a new cluster-aware layout optimization technique is developed to reduce visual clutter and preserve the clustering of glyphs.

Glyph Design : Inspired by *matExplorer* [PSG*21], we also employed a glyph-based representation (Fig. 3(B2)) to visualize each compound in the Discovery View. The glyph design consists of two integrated components: an inner pie chart and an outer circular bar chart, providing a multi-faceted visualization of compound attributes. Unlike *matExplorer* [PSG*21], which primarily focuses on the comparison and analysis of different ML models, our design is centered around supporting and facilitating compound discovery, particularly in the context of identifying promising semiconductors. Our glyph aims to make key property's influencing factors interpretable to the user, enabling them to actively explore, compare, and filter compounds based on domain-relevant criteria (T3).

As illustrated in Fig. 4A, the inner pie chart encodes the user-selected key attributes from the Control Panel. Each segment corresponds to one attribute, where segment size reflects its value. The segment color indicates the compound's cluster membership, with distinct colors representing different clusters.

The outer circular bar chart extends this representation by encoding influencing factors through a systematic radial layout. In this way, glyph similarity and dissimilarity reveal not only how compounds compare in terms of values, but also the underlying reasons for these differences, thereby guiding users toward selecting further filtering attributes for more refined exploration. For each pie sector, we show the k most significant attributes ($k = \lfloor 15/n \rfloor$), where n is the number of sectors. The value 15 was empirically determined through iterative design and expert feedback to balance coverage and readability. The attributes are selected based on feature importance or correlation strength (T3). For model-predicted attributes, we use SHAP values to reflect feature importance, while for other attributes, we use Pearson correlation coefficients with the inner key attributes to indicate correlation strength. To ensure a consistent and comparable representation, we use a unified encoding for these influencing factors in the outer circular bars, regardless of whether they originate from prediction or imputation. This avoids extra visual complexity in the glyph, and the distinctions in data source (Predicted, Imputed or Retrieved) are clarified in the Comparison View (Sec. 6.3). The attributes are ordered by decreasing significance counterclockwise, bar height encodes normalized values, and bar orientation represents relationship direction (inward for negative, outward for positive)(Fig. 4A). The attributes influencing multiple inner attributes will appear accordingly in each relevant sector.

The glyph design allows users to focus on the selected attributes and also allows users to see the relationship between the attributes. Additionally, a tooltip showing the exact value and box plot of the selected attribute is displayed when users hover over a specific glyph, enabling detailed attribute exploration and comparison.

Alternative Designs. We explored several alternative glyph representations before finalizing our design. The first approach (Fig. 4B) used a circular bar chart with opposing bars along the circumference to show paired attributes like electron and hole mobility. While effective for highlighting relationships between paired properties, it became cluttered with more than three pairs and made cross-compound comparisons difficult. Our second design (Fig. 4C) employed an enhanced radar chart with multiple encoding channels, including vertices for attribute values and black/white dots for ground truth versus predicted values. This could display more information but required significant cognitive effort to interpret. These explorations led to our current dual-layer design, which separates primary attributes (inner pie chart) from their relationships (outer circular bars), balancing visual clarity with analytical capability.

Cluster-Aware Layout Optimization. A major challenge in the Discovery View is the visual clutter created by overlapping glyphs, especially in dense areas with many compounds. This is because projecting key attributes and influencing factors into a 2D t-SNE embedding results in high-density clusters, as many compounds are

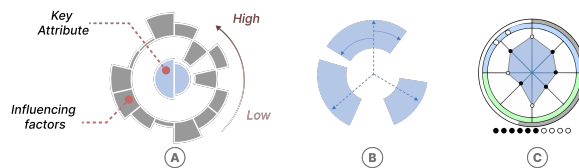


Figure 4: Design alternatives: (A) current glyph design; (B) a circular bar chart design; (C) a radar chart design.

closely related to each other. To address this, we use two strategies: density-based sampling and cluster-aware overlap removal.

Density-Based Sampling. We employ Kernel Density Estimation (KDE) to perform adaptive sampling while preserving the underlying data distribution. Based on the 2D embeddings obtained from t-SNE, we apply DBSCAN to identify clusters used for sampling. For clusters with more than three compounds, we compute a sampling rate proportional to their probability density

$$m_i = \max\left(3, \left\lfloor N \cdot \frac{n_i}{\sum_{k=1}^K n_k} \right\rfloor\right), \quad \mathbb{P}(x_{ij} \text{ is kept}) \propto \hat{f}_i(\mathbf{p}_{ij})$$

where m_i is the number of samples kept from cluster i , n_i is the number of compounds in cluster i , K is the total number of clusters, and N is the target sample size (set to 100 for optimal visualization). $\mathbb{P}(x_{ij} \text{ is kept})$ denotes the probability that compound x_{ij} is kept as a sample, \mathbf{p}_{ij} represents the 2D embedding of compound x_{ij} obtained by reducing the selected key attributes and corresponding influencing factors using t-SNE, and \hat{f}_i is the kernel density estimate computed over the 2D embeddings of compounds in cluster i , using a Gaussian kernel with a bandwidth of 0.5 to balance local detail and smoothness in density estimation. This approach ensures that small clusters maintain structural integrity while larger clusters are sampled proportionally to density distribution so that the overall data distribution is preserved.

Cluster-Aware Overlap Removal. While overlap removal algorithms like DGrid, SORBID [HME*24, GLGB24] offer systematic solutions, they usually ignore the cluster structure within the data which is essential for material analysis. Cluster-aware methods [ZYC*24] addressed this limitation in grid-based layouts, however gridifying the scatter plot can significantly distort the original spatial relationships. To overcome these limitations, we propose a two-step overlap removal approach that maintains cluster integrity while mitigating glyph overlaps.

1) *Space Optimization:* The first step transforms discrete point clusters into continuous spatial regions while preserving cluster relationships. Specifically, each cluster is represented as a convex hull polygon P_i , computed using the Quickhull algorithm [BDH96] on 2D embedding space. For the clusters with fewer than three points, we instead generate a circular buffer of fixed radius around the centroid to avoid degenerate polygons. Because clustering algorithms produce non-adjacent cluster polygons, dummy polygons are introduced to fill empty spaces for subsequent cartogram processing. These dummy regions are generated through a kd-tree inspired recursive partitioning approach. The global bounding box is initially divided into a 4×4 grid, and cells overlapping cluster hulls are recursively split along their longer axis until their area is below 0.1 (normalized units). The resulting regions are clipped to empty

spaces, balancing computational efficiency and geometric fidelity. With this continuous layout in place, we then employ the cartogram technique [GN04] to optimize polygon areas. Each cluster polygon is scaled proportional to its number of member compounds, while dummy polygons are assigned a weight of 0.1 to act only as spacers. This ensures sufficient space for expansion and approximately preserves relative topological relationships.

2) *Layout Refinement*: Within each optimized polygon, we use a force-directed algorithm to adjust compounds' positions. The total force \mathbf{F} acting on each compound is defined as $\mathbf{F} = \mathbf{F}_b + \mathbf{F}_o + \mathbf{F}_l + \mathbf{F}_{mb} + \mathbf{F}_c$, where \mathbf{F}_l , \mathbf{F}_{mb} , and \mathbf{F}_c are standard forces from D3's force-directed simulation, handling edge attractions, node repulsion, and collision avoidance, respectively. Here, edges are defined within each cluster by connecting every compound to its nearest neighbor in the 2D embedding space, which provides the basis for edge attractions. Additionally, we introduce two specialized forces to maintain cluster structure and spatial relationships:

- *Boundary Containment Force (\mathbf{F}_b)*: This force confines compounds within their designated cluster boundaries. If a compound drifts outside its cluster polygon, it is pulled back toward the nearest boundary point, given by $\mathbf{F}_b = k_b \cdot \frac{\mathbf{p}_{\text{nearest}} - \mathbf{p}}{\|\mathbf{p}_{\text{nearest}} - \mathbf{p}\|}$, where $\mathbf{p}_{\text{nearest}}$ is the nearest point on the polygon boundary to compound position \mathbf{p} , and $k_b(20)$ controls the force magnitude.
- *Position Preservation Force (\mathbf{F}_o)*: This force preserves spatial relationships by pulling compounds toward their initial positions, defined as $\mathbf{F}_o = k_o \cdot (\mathbf{p}_i - \mathbf{p})$, where \mathbf{p}_i is the initial position and $k_o(0.2)$ is the attraction strength.

The layout optimization proceeds iteratively until reaching equilibrium, with force magnitudes adjusted dynamically to balance position preservation, overlap prevention and cluster integrity.

6.3. Comparison View

Comparison View (Fig. 3c) facilitates detailed comparative analysis of candidate compounds (T6) through an interactive heatmap visualization. For each candidate compound, we use a heatmap (Fig. 3ci) to visualize the values of each attribute, with rows representing candidate compounds and columns representing attributes. The heatmap is color-coded to indicate the value of each attribute, with a gradient from light yellow (low) to blue (high), enabling rapid identification of patterns and outliers. Rows are ordered according to chemical formula similarity, ensuring that compounds with related compositions are positioned adjacently, while columns are grouped by predefined categories as described in section 6.1. Additionally, shapes encode the data source: circles ● indicate predicted values (AE & NGB), triangles ▲ represent imputed values (CAMI), and rectangles ■ denote ground truth. For predicted values, circle radius reflects uncertainty, with larger radius indicating lower uncertainty. To enhance analytical precision, a tooltip reveals the exact attribute value upon hovering over the cell.

For crystal structure, a non-numerical compound characteristic, *SemiConLens* integrates interactive 3D visualizations using 3DMol.js (Fig. 3c2). Hovering over the heatmap opens an interactive 3D molecular viewer with rotation, zooming, and screenshot capture interactions enabled, which helps users intuitively explore compound structures.

7. Evaluation

We evaluate our approach using both quantitative and qualitative approaches. The following section discusses the findings.

7.1. Quantitative Evaluation

We conducted two quantitative evaluations to validate model selection and assess the proposed imputation method.

Model Evaluation. We evaluated eight ML models, commonly used in material science, for predicting five semiconductor attributes using MAPE, including RF, SVR, LR, MLP, DT, GB, NGBoost, and a TensorFlow-based fine-tuned AE (See Appendix A.4 for architecture). As the dataset was small, we used leave-one-out cross-validation with three validation samples and repeated all experiments five times with different random seeds for robustness.

As shown in Table 1, the AE model performs best for carrier mobility, whereas NGBoost is superior for bandgap GW, VBM GW, and CBM GW predictions. These differences likely reflect the limited data availability described in Section 1.

Model	E Mob	H Mob	Bandgap	VBM	CBM
RF	109.00	203.38	7.25	2.71	6.29
SVR	104.25	242.39	41.53	4.74	5.68
LR	321.60	1,779.28	21.19	8.48	19.28
MLP	178.90	195.58	10.95	3.74	7.81
DT	122.35	186.01	9.49	3.71	8.24
GB	88.34	172.31	5.83	2.35	5.36
NGBoost	90.22	170.98	5.74	2.30	5.34
AE	71.39	102.70	27.17	16.95	24.75

Table 1: Comparison of MAPE for the evaluated models across five semiconductor properties. "E Mob" and "H Mob" refer to electron and hole mobility, respectively, while bandgap, VBM, and CBM use GW-calculated ground truth values.

Imputation Method Evaluation. We evaluate CAMI against the most popular imputation methods in materials science domain, such as mean imputation and zero imputation. These were selected as they were **commonly used in materials science workflows** [MXL*24, DWG*20]. Therefore, other multivariate imputations were not included in this comparison. According to table 2, CAMI has a significantly higher performance in E Mobility and Hole mobility prediction while it only shows some improvement over mean imputation for Bandgap GW VBM GW, CBM GW. The evaluation was conducted by comparing the prediction performance of the AEs for electron and hole mobility, and NGBoost for the other properties. We speculate this is a result of not filling all the missing values with the same value, which preserve the correlation between attributes of the compounds, especially when there are few ground truth data available.

7.2. Expert Interview

We conducted semi-structured interviews with 12 experts (Male = 8) including 8 PhD students, 1 postdoctoral researcher and 3 research fellows, averaging 5.2 years of research and engineering experience in semiconductors or other related materials.

Method	Desirable Semiconductor Attribute				
	E Mob	H Mob	Bandgap	VBM	CBM
Mean	71.39	102.70	5.74	2.30	5.34
Zero	68.96	116.78	7.19	2.68	5.36
MF	70.65	120.11	7.19	2.68	5.36
CAMI	63.53	98.17	5.67	2.28	5.26

Table 2: Comparison of imputation methods. E Mob and H Mob (electron and hole mobility) are predicted with AE, while Bandgap, VBM, and CBM are predicted with NGBoost, using GW-calculated values as ground truth. MF denotes Most Frequent.

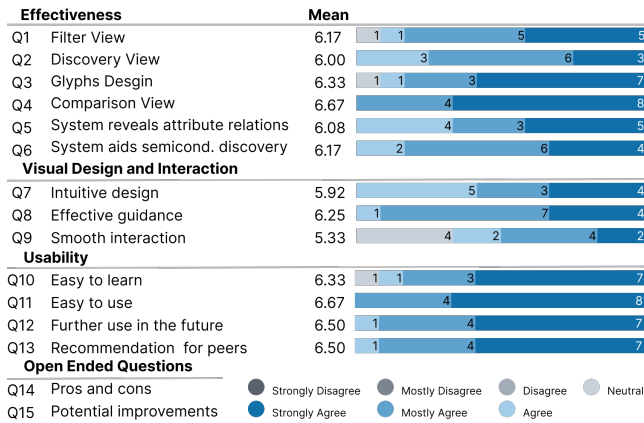


Figure 5: The expert interview questionnaire results.

Procedures: Each expert interview (approx. 60 minutes) comprised three phases: (1) a system introduction covering the research background, visualization design, system interactions, and a 15-minute usage scenario; (2) two assigned tasks (approx. 30 minutes); and (3) a post-study questionnaire (Q1-Q15), as shown in Fig. 5. Q1-Q13 used a 7-point Likert scale adapted from PSSUQ [Lew92] and prior work [WNR*24] to evaluate effectiveness (Q1-Q6), design and interaction (Q7-Q9), and usability (Q10-Q13). Q14-Q15 collected open-ended feedback. Participants received approximately 15 USD as compensation.

7.2.1. Results

The results are summarized in Fig. 5, with full questions in Appendix A.5. Overall, *SemiConLens* received high ratings, with no negative scores (1-3 on the Likert scale). Neutral ratings were given for Q1, Q3, Q9, and Q10 (1, 1, 4, and 1 respectively), notably by E7 for Q1 and Q3.

Effectiveness. According to scores from Q1-Q6, all but one of the participants agreed that the system is an effective tool for 2D semiconductor discovery. This is reflected in consistently high mean scores ranging from 6.00 to 6.67.

In terms of qualitative feedback, E7 rated neutral for Q1 and Q3, stating the main reason as “the amount of information displayed is overwhelming”. They further elaborated that the “text in the Filter View is too small” and that they “would like a feature where I

can toggle some of the view off to get more screen space for the view I’m interested in”. They also suggested to “add a tool tip for supporting attributes in addition to the glyph legend”.

Other experts emphasized effectiveness-related aspects, agreeing that the Filter View is “useful for the initial stage of material selection” and that the Discovery View can be utilized to “easily find related materials and alternative materials” and that the glyph design supports filtering by linking property values with their influencing factors. E7 highlighted that the filter history view is useful, while E11 and E12 appreciated the predicted values of the system, which mitigate the necessity of “hectic experimental procedures”.

Visual Design and Interaction. Most participants agreed that the system has a good visual design and interaction quality, with mean scores between 5.33 and 6.25. This indicates general satisfaction. Qualitative feedback confirmed this. Most participants agreed the system is “visually appealing”, which is consistent with quantitative ratings. Meanwhile, participants also provided suggestions for further improvement. For example, E4, E7, E8, and E11 provided a neutral score for Q9, stating that the system speed needs to be improved, as it takes about 7 seconds to update the Discovery View when there are a substantial number of elements due to computationally intensive cartogram calculations. E11 also stated that “the system needs more explanations such as more detailed labels and legends” and rated a neutral score for Q12. Other experts suggested improvements to the user interface, such as “providing more detailed legends and information panels” and “using larger font size and high-contrast colors”.

Usability. According to scores from Q11-Q13, all participants agreed that the usability of the system is excellent, with a mean rating of 6.50. This represents the highest overall score among the three categories. Despite the issues mentioned in the visual design and interaction section, all the experts agreed that the system is easy to use and that they would like to use it to identify potential semiconductors in the future. They also agreed that they would recommend this system to other material scientists.

7.3. Use Cases

Case 1: Exploring clusters - Identifying 2D Materials for Water Splitting

Photocatalytic water splitting refers to the light-driven decomposition of water into H_2 and O_2 using semiconductor photocatalysts, which can be achieved either in a single-step system with suitable band alignment or in a Z-scheme employing two semiconductors for separate reduction and oxidation [YSA20].

In this use case, we demonstrate how *SemiConLens* can be used to identify 2D semiconductors for photocatalytic water splitting. Effective splitting requires **proper CBM and VBM alignment with water redox potentials**, i.e., the CBM GW must lie above -4.44 eV for H^+/H_2 reduction, and the VBM GW below -5.67 eV for O_2/H_2O oxidation [WSBE24]. These criteria served as the initial filtering parameters (Fig. 6 A). Recognizing MoS_2 as a benchmark catalyst for Hydrogen Evolution Reaction (HER) [YZC*23], the user set it as a reference in the Filter View (Fig. 6 A1). Observing a cluster of compounds around MoS_2 (Fig. 6 A), the user lassoed this region to isolate candidates (Fig. 6 B). To further refine

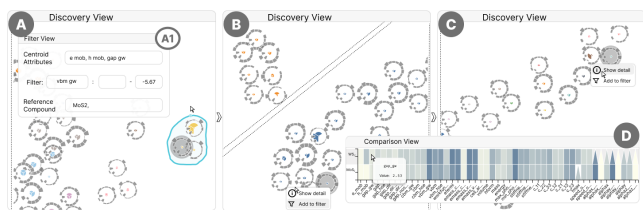


Figure 6: With *SemiConLens* an expert has identified WS_2 as a similar material to MoS_2 , suitable for hydrogen evolution reaction in water splitting. First, they applied 2 filters, CBM GW $> -4.44\text{eV}$ and VBM GW $< -5.67\text{eV}$ using the Filter View. Then they selected MoS_2 as the reference compound (A1). From the resulting Discovery View, they identified the cluster with MoS_2 and selected it by lassoing the cluster (A). They identified the MoS_2 from the resulting Discovery View (B) and add it as the filter for next step. Then they identified WS_2 as the nearest neighbor for MoS_2 from the Discovery View (C) and added both to comparison view for further exploration (D). Some compounds are omitted from the Discovery View, marked by dotted lines. See Appendix A.9 for the high resolution version.

the results, MoS_2 was selected as the reference material, yielding the 30 closest compounds (Fig. 6 C) (T5).

For comparative evaluation, MoS_2 and its nearest neighbor WS_2 were selected (Fig. 6 C) (T6). The user observed that their bandgaps (2.53 eV) slightly exceeded the ideal 1.5 to 2.5 eV range [ZTWS24], which balances visible light absorption with enough photovoltage to drive H_2/O_2 evolution, for visible-light photocatalysis but remained acceptable (Fig. 6 D). External validation also confirmed WS_2 as a promising HER catalyst [YZC*23], highlighting the system's effectiveness in rapidly identifying candidate 2D semiconductors for water splitting.

Case 2: Sequential Filtering - Discovering 2D Materials for Tunnel Field-Effect Transistors (TFETs)

A TFET is an electronic switch that leverages quantum tunneling rather than conventional current flow to enable lower-power operation than traditional transistors. TFETs require “**Type-II band alignment**”, where **the CBM of one material lies below the VBM of another to enable tunneling** [VDBK24]. This case shows how *SemiConLens* uses predicted CBM and VBM GW values to identify 2D donor-acceptor pairs for TFETs. First, the user selected *gap_gw*, *vbm_gw* and *cbm_gw* as the centroid attributes (T3). Initially, the user defined TFET-specific filters, selecting semiconductors with bandgaps of 0.5 to 2.0 eV, since bandgaps below 0.5 eV prevent turn-off and those above 2.0 eV hinder tunneling efficiency (Fig. 3 A1) [KCS*18]. Donor candidates were further filtered by low CBM value (typically around -4.5 eV or lower) using the control panel in Filter View (Fig. 3 A1) (T5). Recognizing SnSe_2 as a promising donor candidate [SND*21] based on prior domain knowledge, the user selected it for further analysis (Fig. 3 A2). In the Discovery View, SnSe_2 was found to have a **VBM of -6.88 eV and a CBM of -4.97 eV** (Fig. 3 B3)(T4).

To identify acceptors forming a Type-II heterostructure with SnSe_2 , the user removed the CBM filter and applied band alignment criteria of **CBM GW lower than -4.97 eV and VBM GW higher than -6.88 eV**. Using the Attribute Distribution view (Fig.

3 A2), candidates were further constrained to energy above hull values below 0.1 eV for thermodynamic stability and electron mobilities above $200\text{cm}^2\text{V}^{-1}\text{s}^{-1}$ for high performance suitability (T5). These criteria yields 11 potential candidates (Fig. 3 C1): AgTIP_2S_6 , $\text{AgTIP}_2\text{Se}_6$, Au_4I_8 , $\text{Cu}_4\text{I}_4\text{S}_4$, $\text{Cu}_4\text{I}_4\text{Se}_4$, $\text{Ir}_2\text{Nb}_2\text{S}_8$, PbBi_2Se_4 , PbSb_2Se_4 , $\text{Rh}_2\text{Ta}_2\text{S}_8$, Rh_4S_8 , and SnBi_2S_4 . The user was able to identify the confidence of the predicted electron mobility values that affect the performance, as the uncertainty for each predicted value was shown (Fig. 3 C1) (T2). The user used domain knowledge for further refinement. For instance, Au_4I_8 (Fig. 3 B1 C2) was excluded as a non-layered structure, unsuitable for high-quality van der Waals interfaces in TFETs. This domain-informed filtering shows the effectiveness of combining computational insights with expert knowledge to identify viable 2D semiconductors.

8. Discussion

In this section, we discuss lessons learned during the development and evaluation of *SemiConLens*, as well as its possible limitations.

Generalizability. Although *SemiConLens* uses C2DB data and targets 2D semiconductor discovery, its workflow, including CAMI, is database-independent and can be applied to other datasets. For instance, the Materials Project database could support bulk material discovery, but ML models and CAMI must be retrained with appropriate hyperparameter tuning.

Human-AI collaboration for new material discovery. As noted in the introduction, human intuition plays a vital role in the discovery of new materials. *SemiConLens* exemplifies how domain experts can harness the strengths of ML models to facilitate the material discovery process.

System speed and user perception. The system exhibited an average response delay of around 7 seconds due to the iterative overlap removal process (Section 6.2) and the complexity of high-dimensional datasets. Our participants rated Q9 lowest (mean 5.33), which reflects the relatively slow response. Response speed can be further improved, which is left for future work.

9. Conclusion and Future Work

We present *SemiConLens*, a visual analytics system to support discovery of new 2D semiconductors. Eight ML models were evaluated to identify suitable predictors for five hard-to-compute attributes, and a correlation- and similarity-based imputation method was proposed to handle missing values. Integrated into *SemiConLens*, these components form a comprehensive workflow. Two use cases and interviews with 12 material science experts evaluated effectiveness, design, interaction, and usability. Findings show that *SemiConLens* effectively supports semiconductor discovery.

Future work will extend *SemiConLens* to additional databases beyond 2D materials and experimentally validate shortlisted compounds in collaboration with material scientists.

Acknowledgements We used ChatGPT solely for grammar correction and language polishing. This project is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 grant (T2EP202220049), Tier 1 grant (22-SIS-SMU-054), and NTU Start Up Grant awarded to Yong Wang.

References

- [AZER21] ADDEPALLI S., ZHAO Y., ERKOYUNCU J. A., ROY R.: Quantifying Uncertainty in Pulsed Thermographic Inspection by Analysing the Thermal Diffusivity Measurements of Metals and Composites. *Sensors* 21, 16 (Aug. 2021), 5480. 5
- [BBB*24] BINDINI L., BERTAZZINI G., BARACCHI D., SHULLANI D., FRASCONI P., PIVA A.: Tiny Autoencoders are Effective Few-Shot Generative Model Detectors. In *2024 IEEE International Workshop on Information Forensics and Security (WIFS)* (Rome, Italy, Dec. 2024), IEEE, pp. 1–6. 5
- [BDC*18] BUTLER K. T., DAVIES D. W., CARTWRIGHT H., ISAYEV O., WALSH A.: Machine learning for molecular and materials science. *Nature* 559, 7715 (2018), 547–555. 1, 2
- [BDH96] BARBER C. B., DOBKIN D. P., HUHDANPAA H.: The quick-hull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)* 22, 4 (1996), 469–483. 7
- [BKC*13] BORGIO R., KEHRER J., CHUNG D. H., MAGUIRE E., LARAMEE R. S., HAUSER H., WARD M., CHEN M.: Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics (state of the art reports)* (2013), pp. 39–63. 3
- [BWT*25] BLUMBERG D., WANG Y., TELEA A., KEIM D. A., DENING F. L.: Multiinv: Inverting multidimensional scaling projections and computing classifier maps by multilateration. Available at SSRN 5108314 (2025). 3
- [BXX*18] BAI J., XUE Y., BJORCK J., LE BRAS R., RAPPAZZO B., BERNSTEIN R., SURAM S. K., VAN DOVER R. B., GREGOIRE J. M., GOMES C. P.: Phase mapper: Accelerating materials discovery with ai. *AI Magazine* 39, 1 (2018), 15–26. 3, 4
- [CBHvD*18] CORREA-BAENA J.-P., HIPALGAONKAR K., VAN DUREN J., JAFFER S., CHANDRASEKHAR V. R., STEVANOVIC V., WADIA C., GUHA S., BUONASSISI T.: Accelerating materials development via automation, machine learning, and high-performance computing. *Joule* 2, 8 (2018), 1410–1420. 1
- [CHUW08] CHEN C.-H., HÄRDLE W., UNWIN A., WARD M. O.: Multivariate data glyphs: Principles and practice. *Handbook of data visualization* (2008), 179–198. 3
- [CJS19] CHEN C., JIANG X., SUI Y.: Prediction of transport properties of fuels in supercritical conditions by molecular dynamics simulation. *Energy Procedia* 158 (Feb. 2019), 1700–1705. 5
- [CTP*21] CHOUBISA H., TODOROVIC P., PINA J. M., PARMAR D. H., LI Z., VOZNYI O., TAMBLYN I., SARGENT E.: Interpretable discovery of new semiconductors with machine learning. *arXiv preprint arXiv:2101.04383* (2021). 1, 3, 4
- [CWE22] CHANG R., WANG Y.-X., ERTEKIN E.: Towards overcoming data scarcity in materials science: unifying models and datasets with a mixture of experts framework. *npj Computational Materials* 8, 1 (2022), 242. 2
- [CZZ20] CHEN A., ZHANG X., ZHOU Z.: Machine learning: accelerating materials development for energy storage and conversion. *InfoMat* 2, 3 (2020), 553–576. 1, 3
- [DAD*20] DUAN T., AVATI A., DING D. Y., THAI K. K., BASU S., NG A. Y., SCHULER A.: NGBoost: Natural Gradient Boosting for Probabilistic Prediction, June 2020. arXiv:1910.03225 [cs]. 5
- [DWG*20] DUNN A., WANG Q., GANOSE A., DOPP D., JAIN A.: Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Computational Materials* 6, 1 (Sept. 2020), 138. 8
- [ERT*17] ENDERT A., RIBARSKY W., TURKAY C., WONG B. W., NABNEY I., BLANCO I. D., ROSSI F.: The state of the art in integrating machine learning into visual analytics. In *Computer Graphics Forum* (2017), vol. 36, Wiley Online Library, pp. 458–486. 3
- [FFM*13] FUCHS J., FISCHER F., MANSMANN F., BERTINI E., ISENBERG P.: Evaluation of alternative glyph designs for time series data in a small multiple setting. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2013), pp. 3237–3246. 3
- [GBM*19] GELLRICH S., BEGANOVIC T., MATTHEUS A., HERMANN C., THIEDE S.: Feature selection based on visual analytics for quality prediction in aluminium die casting. In *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)* (2019), vol. 1, pp. 66–72. 3
- [GLGB24] GIOVANNANGELI L., LALANNE F., GIOT R., BOURQUI R.: Overlap Removal by Stochastic Gradient Descent With(out) Shape Awareness. *IEEE Transactions on Visualization and Computer Graphics* 30, 12 (Dec. 2024), 7500–7517. 7
- [GN04] GASTNER M. T., NEWMAN M. E.: Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences* 101, 20 (2004), 7499–7504. 8
- [GTR*21] GJERDING M. N., TAGHIZADEH A., RASMUSSEN A., ALI S., BERTOLDO F., DEILMANN T., KNØSGAARD N. R., KRUSE M., LARSEN A. H., MANTI S., ET AL.: Recent progress of the computational 2d materials database (c2db). *2D Materials* 8, 4 (2021), 044002. 1, 2, 3
- [HME*24] HILASACA G. M., MARCÍLIO-JR W. E., ELER D. M., MARTINS R. M., PAULOVIK F. V.: A Grid-Based Method for Removing Overlaps of Dimensionality Reduction Scatterplot Layouts. *IEEE Transactions on Visualization and Computer Graphics* 30, 8 (Aug. 2024), 5733–5749. 3, 7
- [HSE03] HEYD J., SCUSERIA G. E., ERNZERHOF M.: Hybrid functionals based on a screened Coulomb potential. *The Journal of Chemical Physics* 118, 18 (May 2003), 8207–8215. 2
- [Hu05] HU Y.: Efficient, high-quality force-directed graph drawing. *Mathematica journal* 10, 1 (2005), 37–71. 3
- [IDK*04] IEONG M., DORIS B., KEDZIERSKI J., RIM K., YANG M.: Silicon Device Scaling to the Sub-10-nm Regime. *Science* 306, 5704 (Dec. 2004), 2057–2060. 1
- [JOH*13] JAIN A., ONG S. P., HAUTIER G., CHEN W., RICHARDS W. D., DACEK S., CHOLIA S., GUNTER D., SKINNER D., CEDER G., ET AL.: Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials* 1, 1 (2013). 1
- [KCS*18] KLINKERT Á. S. C., CAMPI D., STIEGER C., MARZARI N., LUISIER M.: Ab initio simulation of band-to-band tunneling FETs with single- and few-layer 2-D materials as channels. *IEEE Transactions on Electron Devices* 65, 10 (Oct. 2018), 4180–4187. 10
- [KKG*19] KAILKHURA B., GALLAGHER B., KIM S., HISZPANSKI A., HAN T. Y.-J.: Reliable and explainable machine-learning methods for accelerated material discovery. *npj Computational Materials* 5, 1 (2019), 108. 1, 3, 4
- [KMM20] KAHLE L., MARCOLONGO A., MARZARI N.: High-throughput computational screening for solid-state li-ion conductors. *Energy & Environmental Science* 13, 3 (2020), 928–948. 4
- [Lew92] LEWIS J. R.: Psychometric Evaluation of the Post-Study System Usability Questionnaire: The PSSUQ. *Proceedings of the Human Factors Society Annual Meeting* 36, 16 (Oct. 1992), 1259–1260. 9
- [LL17] LUNDBERG S. M., LEE S.-I.: A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (2017), Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R., Vishwanathan S., Garnett R., (Eds.), vol. 30, Curran Associates, Inc. 5
- [LLW*21] LUO S., LI T., WANG X., FAIZAN M., ZHANG L.: High-throughput computational materials screening and discovery of optoelectronic semiconductors. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 11, 1 (2021), e1489. 1, 3, 4
- [LSWL19] LI M.-Y., SU S.-K., WONG H.-S. P., LI L.-J.: How 2d semiconductors could extend moore’s law. *Nature* 567, 7747 (2019), 169–170. 1

- [LXL*22] LIU D.-Y., XU L.-M., LIN X.-M., WEI X., YU W.-J., WANG Y., WEI Z.-M.: Machine learning for semiconductors. *Chip J*, 4 (2022), 100033. 2, 3
- [LZJS17] LIU Y., ZHAO T., JU W., SHI S.: Materials discovery and design using machine learning. *Journal of Materiomics* 3, 3 (2017), 159–177. 1, 3
- [LZO*18] LU S., ZHOU Q., OUYANG Y., GUO Y., LI Q., WANG J.: Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nature communications* 9, 1 (2018). 3
- [MAK*14] MEREDIG B., AGRAWAL A., KIRKLIN S., SAAL J. E., DOAK J. W., THOMPSON A., ZHANG K., CHOUDHARY A., WOLVERTON C.: Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B* 89, 9 (2014), 094104. 2, 4
- [Mar20] MARTIN R. M.: *Electronic structure: basic theory and practical methods*. Cambridge university press, 2020. 2
- [MNS*23] MARTÍNEZ M. J., NAVEIRO R., SOTO A. J., TALAVANTE P., KIM LEE S.-H., GÓMEZ ARRAYAS R., FRANCO M., MAULEÓN P., LOZANO ORDÓÑEZ H., REVILLA LÓPEZ G., BERNABEI M., CAMPILLO N. E., PONZONI I.: Design of New Dispersants Using Machine Learning and Visual Analytics. *Polymers* 15, 5 (Mar. 2023), 1324. 3
- [MQL*24] MENG X., QIN C., LIANG X., ZHANG G., CHEN R., HU J., YANG Z., HUO J., XIAO L., JIA S.: Deep learning in two-dimensional materials: Characterization, prediction, and design. *Frontiers of Physics* 19, 5 (2024), 53601. 2
- [MXL*24] MA Y., XU P., LI M., JI X., ZHAO W., LU W.: The mastery of details in the workflow of materials machine learning. *npj Computational Materials* 10, 1 (July 2024), 141. 8
- [NGM*04] NOVOSELOV K. S., GEIM A. K., MOROZOV S. V., JIANG D., ZHANG Y., DUBONOS S. V., GRIGORIEVA I. V., FIRSOV A. A.: Electric field effect in atomically thin carbon films. *Science* 306, 5696 (2004), 666–669. 1
- [OM18] OLIYNYK A. O., MAR A.: Discovery of intermetallic compounds from traditional to machine-learning approaches. *Accounts of chemical research* 51, 1 (2018), 59–68. 1, 3
- [PBE96] PERDEW J. P., BURKE K., ERNZERHOF M.: Generalized Gradient Approximation Made Simple. *Physical Review Letters* 77, 18 (Oct. 1996), 3865–3868. 2
- [PCCT14] PIMENTEL M. A., CLIFTON D. A., CLIFTON L., TARASSENKO L.: A review of novelty detection. *Signal Processing* 99 (June 2014), 215–249. 5
- [PLRG20] PONCÉ S., LI W., REICHARDT S., GIUSTINO F.: First-principles calculations of charge carrier mobility and conductivity in bulk semiconductors and two-dimensional materials. *Reports on Progress in Physics* 83, 3 (2020), 036501. 2
- [PSG*21] PU J., SHAO H., GAO B., ZHU Z., ZHU Y., RAO Y., XIANG Y.: Matexplorer: Visual exploration on predicting ionic conductivity for solid-state electrolytes. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 65–75. 1, 3, 4, 6
- [RHW86] RUMELHART D. E., HINTON G. E., WILLIAMS R. J.: *Learning internal representations by error propagation*. MIT Press, Cambridge, MA, USA, 1986, p. 318–362. 5
- [SB*00] STREETMAN B. G., BANERJEE S., ET AL.: *Solid state electronic devices*, vol. 4. Prentice hall New Jersey, 2000. 2
- [SCR*24] SCHMIDT J., CERQUEIRA T. F., ROMERO A. H., LOEW A., JÄGER F., WANG H.-C., BOTTI S., MARQUES M. A.: Improving machine-learning models in materials science through large datasets. *Materials Today Physics* 48 (2024), 101560. 2
- [Sem23] SEMICONDUCTOR INDUSTRY ASSOCIATION (SIA): Factbook 2023, 2023. URL: [semiconductors.org/wp-content/uploads/2023/05/SIA-2023-Factbook_1.pdf](https://www.semiconductors.org/wp-content/uploads/2023/05/SIA-2023-Factbook_1.pdf). 1
- [Sem24] SEMICONDUCTOR INDUSTRY ASSOCIATION (SIA): Factbook 2024, 2024. URL: [semiconductors.org/wp-content/uploads/2024/05/SIA-2024-Factbook.pdf](https://www.semiconductors.org/wp-content/uploads/2024/05/SIA-2024-Factbook.pdf). 1
- [SLN21] SZE S. M., LI Y., NG K. K.: *Physics of semiconductor devices*. John Wiley & sons, 2021. 2
- [SND*21] SATO Y., NISHIMURA T., DUANFEI D., UENO K., SHINOKITA K., MATSUDA K., NAGASHIO K.: Intrinsic Electronic Transport Properties and Carrier Densities in PtS₂ and SnSe₂: Exploration of n⁺-Source for 2D Tunnel FETs. *Advanced Electronic Materials* 7, 12 (Dec. 2021), 2100292. 10
- [SUS*20] SABANDO M. V., ULBRICH P., SELZER M., BYŠKA J., MIČAN J., PONZONI I., SOTO A. J., GANUZA M. L., KOZLÍKOVÁ B.: ChemVA: Interactive Visual Analysis of Chemical Compound Similarity in Virtual Screening, Aug. 2020. arXiv:2008.13150 [cs]. 3
- [SZL10] SHI H., ZHANG P., LI S.-S.: Optical properties of CeO₂ using screened hybrid functional and GW+U methods, Jan. 2010. arXiv:1001.2899 [cond-mat]. 2
- [VDBK24] VERMA M., DUTTA P., BASAK A., KUMAR R.: A Comparative Study of Material Impact on Tunnel Field-Effect Transistor (TFET) Performance. In *2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)* (Windhoek, Namibia, July 2024), IEEE, pp. 472–477. 10
- [WNR*24] WEN X., NGUYEN T. D., RUAN S., SHEN Q., SUN J., ZHU F., WANG Y.: *PonziLens+*: Visualizing Bytecode Actions for Smart Ponzi Scheme Identification. *IEEE Transactions on Visualization and Computer Graphics* (2024), 1–14. 9
- [Wor23] WORLD TRADE ORGANIZATION: *Global Value Chain Development Report 2023: Resilient and Sustainable Global Value Chains in Turbulent Times*. WTO, Nov. 2023. 1
- [WSBE24] WANG Y., SORKUN M. C., BROCKS G., ER S.: ML-Aided Computational Screening of 2D Materials for Photocatalytic Water Splitting. *The Journal of Physical Chemistry Letters* 15, 18 (May 2024), 4983–4991. 9
- [YAK*23] YAMAGUCHI Y., ATSUMI T., KANAMORI K., TANIBATA N., TAKEEDA H., NAKAYAMA M., KARASUYAMA M., TAKEUCHI I.: Drawing a materials map with an autoencoder for lithium ionic conductors. *Scientific Reports* 13, 1 (Oct. 2023), 16799. 3
- [YSA20] YANG X., SINGH D., AHUJA R.: Recent Advancements and Future Prospects in Ultrathin 2D Semiconductor-Based Photocatalysts for Water Splitting. *Catalysis* 10, 10 (Sept. 2020), 1111. 9
- [YZC*23] YU H., ZHANG M., CAI Y., ZHUANG Y., WANG L.: The Advanced Progress of MoS₂ and WS₂ for Multi-Catalytic Hydrogen Evolution Reaction Systems. *Catalysis* 13, 8 (July 2023), 1148. 9, 10
- [ZL18] ZHANG Y., LING C.: A strategy to apply machine learning to small datasets in materials science. *Npj Computational Materials* 4, 1 (2018), 25. 2
- [ZSC*19] ZHOU J., SHEN L., COSTA M. D., PERSSON K. A., ONG S. P., HUCK P., LU Y., MA X., CHEN Y., TANG H., ET AL.: 2dmatpedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches. *Scientific data* 6, 1 (2019), 86. 1
- [ZTWS24] ZUO C., TANG X., WANG H., SU Q.: A Review of the Effect of Defect Modulation on the Photocatalytic Reduction Performance of Carbon Dioxide. *Molecules* 29, 10 (May 2024), 2308. 10
- [ZWML23] ZHANG C., WANG R., MISHRA H., LIU Y.: Two-dimensional semiconductors with high intrinsic carrier mobility at room temperature. *Phys. Rev. Lett.* 130 (Feb 2023), 087001. 1, 2, 3, 4
- [ZYC*24] ZHOU Y., YANG W., CHEN J., CHEN C., SHEN Z.: Cluster-Aware Grid Layout. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 240–250. 3, 7