DOI: 10.21917/ijme.2026.0395

DEEP LEARNING-ASSISTED DESIGN OF ADVANCED MATERIALS FOR NEXT-GENERATION SEMICONDUCTOR APPLICATIONS

Mohammad Abdur Rasheed

Department of Civil Engineering, College of Engineering, Shaqra University, Saudi Arabia

Abstract

Designing the new semiconductor technologies calls for advanced materials having superior electrical conductivity, heat stability, carrier mobility, dielectric properties, and robust structures. Traditional techniques for material synthesis and parameter optimization can be requiring numerous tests, computational efforts, and parameter tuning. State-of-the-art methods of computational materials design usually experience difficulties in handling the non-linear relationship between material properties and the performance of the final device. A Deep Learning-Assisted Semiconductor Material Intelligence Network (DLSMIN) that uses multimodal material representation, physics-informed attention-guided learning, physics-informed learning approach, and multi-objective optimization framework is presented. The method combines information on chemical composition, crystal structure, electronic properties, and fabrication process of material to predict its performance. Physics-informed regularization technique is used to encode knowledge about physical processes, and uncertainty-aware optimization is applied to select the most suitable material candidates for experiments. In the comparison study, the new material intelligence framework demonstrates an R2 of 0.92, mean absolute error of 0.08 eV, and root mean square error of 0.15 eV for predicting semiconductor's bandgap with a 70% portion of training data. Additionally, the model attains the material suitability classification accuracy of 94.2% and screening efficiency of 50%.

Keywords:

Deep Learning, Semiconductor Materials, Materials Discovery, Physics-Informed Learning, Multi-Objective Optimization

1. INTRODUCTION

The relentless development of semiconductor technology is generating a need for materials that facilitate fast computation, energy-efficient electronics, state-of-the-art photonic devices, quantum information processing, and future power electronics. Silicon-based technologies have enabled tremendous breakthroughs in integrated circuit design; but new applications call for better electrical properties, heat dissipation, lower power usage, and improved reliability. Hence, there is a need to explore wide-bandgap semiconductors, two-dimensional materials, compound semiconductors, perovskite materials, and other materials to enable higher performance than what is possible with traditional materials. Development of such materials requires a fundamental understanding of the interplay between chemical composition, atomic structure, crystal structure, electronic structure, and process conditions. Discovery of new materials is highly dependent on experiments, theoretical calculations, physical characterization, and optimization. While these approaches yield important scientific insights, exploration of large chemical and structural space by repeated experiments can be computationally intensive, time-consuming, and costly. It has been recently shown that machine learning can help speed up property predictions and narrow down the search space [1–3].

Specifically, deep learning can be used to describe the nonlinear interactions between high-dimensional descriptors and predict the electronic, optical, mechanical, and thermal properties of materials. The combination of this approach with semiconductor physics could help find promising materials before costly synthesis and device-level testing. The need for data-driven semiconductor design is also increasing due to the growing demands on materials that arise from device miniaturization and heterogeneous integration. The new device designs demand materials with multiple properties rather than a single optimized one. For instance, a material designed for power electronics would require a proper bandgap, breakdown strength, thermal stability, and fabrication capability. Likewise, materials designed for optoelectronic devices would require proper bandgap properties, optical absorption, charge transport, and operational stability. This is why there is a need to develop intelligent computational methods to connect materials discovery with device design. Recent studies have focused on the integration of atomistic modeling, machine learning, and computational screening in the design of semiconductors and fabrication process optimization [2].

Although a significant advance has been made in computational materials science, there remain some barriers preventing the consistent application of deep learning techniques to novel semiconductor designs. The first barrier pertains to data availability and integrity. Material characteristics are derived from different experimental papers, databases, and simulation tools where different measurement settings, theoretical approximations, and conventions apply. Missing data points, duplicate compounds, mismatched structural representations, and class imbalanced materials can create ambiguities during the learning process. Moreover, experimentally obtained properties can be different from those provided by theoretical approximations because of the presence of impurities, defects, different fabrication conditions, etc. Therefore, developing reliable learning systems demands appropriate management of datasets, their representations, and validation procedures preventing information leakage. The second barrier is related to physical consistency and the ability to generalize the learned models. Deep neural networks are capable of discovering complex statistical correlations but fail when used on unknown chemical formulas, crystalline structures, or production processes. Even if a model is accurate for randomly split data, there is no guarantee that it will perform well for new material classes. In addition, optimizing one aspect will have detrimental effects on another; there is an inherent trade-off between electrical performance, thermal stability, feasibility of production, and cost. Interpretation of the model's output is also crucial since it allows scientists to determine what factors impact the performance being predicted. All these considerations underscore the importance of incorporating physics-informed constraints, uncertainty quantification, interpretability of features, and multi-objective

candidate generation [4,5]. Data quality, interpretability, and physics-awareness are some of the key areas highlighted by recent review papers on materials discovery.

The goal is to devise a deep learning-based system for designing semiconductors. Novelty of this work can be attributed to the Deep Learning-Assisted Semiconductor Material Intelligence Network (DLSMIN), which involves complementary material representation, attention-driven feature learning, physics-inspired regularization, uncertainty-based screening and multi-objective ranking in a unified algorithm. In contrast to the prediction-based approaches, the novel system relates the predictions with application-specific selection criteria. The physics part of the algorithm penalizes those predictions which contradict the corresponding physical relationship or constraint while the uncertainty module determines those materials which need further calculations or experimental validations. The optimization step finds a tradeoff between conflicting demands on material properties and gives a set of prioritized candidates. The efficiency of the framework depends on data quality, appropriate physical constraints and validation on material classes unseen in the training process.

2. RELATED WORKS

Previous research has investigated conventional machine learning techniques, deep neural networks, graph-based modeling approaches, physics-informed architectures, and computational screening. Nonetheless, the performance of these different approaches depends on data quality, material complexity, property representation, and validation technique used.

2.1 MATERIALS INFORMATICS AND DATA-DRIVEN DESIGN

Research [8] on materials informatics has shown that machine learning can create quantitative relationships between material descriptors and functional properties. Li et al. studied the approaches, progress, and prospects of materials informatics with the focus on the combination of theoretical computations, experimental characterization, and data-driven modeling. The review touches on material systems like perovskites, alloys, two-dimensional materials, catalysts, and polymers. These techniques enable researchers to identify the structure-property-performance relationships and serve as a basis for computational material screening. Materials informatics can minimize the necessity of conducting repeated experiments by predicting the properties of those material candidates which have not been characterized yet. However, it is difficult to create transferable models in case the datasets vary in terms of the composition, experimental conditions, and quality of descriptors. In addition, finding statistically strong relationships does not mean discovering the physical mechanism. DLSMIN is an improvement of the materials-informatics approach due to the combination of multimodal representations, physics-informed constraints, and application-driven material candidate selection.

2.2 MACHINE LEARNING FOR SEMICONDUCTOR DISCOVERY

The work presented in Mannodi-Kanakkithodi et al. [9] addresses computational approaches for finding next-generation semiconductor materials based on atomistic simulations and machine learning. The paper brings to attention the significance of using atomistic simulations along with machine learning for exploration of large chemical space. Potential applications may cover predicting the properties of electronic, optical, dielectric, defect-related, and stability properties. This approach will allow increasing speed of preliminary analysis of the candidate materials and helping researchers narrow their search to compounds worthy of further investigation. In particular, combining atomistic simulation and machine learning is essential in the context of semiconductors, whose functioning depends to a great extent on atomic arrangement, electronic properties, and defects. On the other hand, DFT datasets may be quite computationally expensive, and there is a risk that predictions will be affected by DFT approximation. A model developed for the certain class of materials will need some form of validation in order to be used on some other. The suggested DLSMIN framework represents an extension of such research line through taking into account uncertainty-based ranking and different application-dependent objectives rather than focusing purely on property prediction.

2.3 DEEP LEARNING FOR DIRECT AND INDIRECT BANDGAP PREDICTION

Research [10] on machine learning models for discovering direct-bandgap materials has addressed an important challenge in semiconductor optoelectronics. Dinic, Neporozhnii, and Voznyy investigated models for predicting bandgap values and distinguishing direct from indirect bandgap materials. These properties are relevant to applications such as light-emitting devices, photovoltaic systems, and other optoelectronic technologies. Accurate screening can reduce the number of compounds requiring expensive electronic-structure calculations and help researchers identify materials with appropriate optical characteristics. Related studies have also investigated the prediction of inorganic semiconductor bandgaps using machine-learning techniques. These investigations demonstrate the usefulness of data-driven approaches for identifying patterns in material composition and electronic properties. However, bandgap prediction alone cannot fully determine whether a material is suitable for practical devices. Thermal stability, defect tolerance, charge transport, and processing compatibility may substantially affect actual performance. DLSMIN therefore treats bandgap as one component of a broader evaluation process. Its proposed multi-objective screening strategy considers complementary properties and their application-specific importance, allowing candidate materials to be evaluated more comprehensively than through bandgap prediction alone.

2.4 PHYSICS-INFORMED AND EXPLAINABLE MATERIALS LEARNING

Physics-informed [11] and interpretable machine learning have become essential research directions for advancing the reliability of data-driven materials research. The role of incorporating physics, interpretability, and causality into materials discovery and design was explored by Ghosh. Physics-informed techniques allow for the inclusion of governing laws, physical constraints, and domain knowledge into the learning process. Such tools can prevent the emergence of unphysical predictions by a model when little experimental data is available. Explainability can identify features of materials whose contribution to a property is the strongest in order to interpret findings from the scientific point of view and refine material candidates accordingly. However, careful consideration should be paid to the choice of constraints, as an invalid one will lead to the biased representation of the data by the model. Also, explanations should be properly understood relative to the properties of the data distribution and model itself. DLSMIN utilizes physics-informed regularizations and feature importance analysis for achieving greater transparency of materials assessment. The proposed system design seeks to reconcile the predictive power of a model with physics-informed elements; yet their effectiveness should be validated via ablation studies.

2.5 DATA-DRIVEN SEMICONDUCTOR DESIGN BEYOND CONVENTIONAL SCALING

Computational issues related to the investigation of material-design space, semiconductor property prediction, and design optimization for device fabrication have been considered in [12]. The paper shows that there is a need to combine computation with data-driven approaches due to the increasing difficulty of today's electronic devices. Designing of semiconductors includes several interactions between material selection, structure design, fabrication, and device considerations. As a result, optimizing a specific property of a semiconductor does not necessarily lead to high-performance devices. Therefore, data-driven models can be used to narrow material space, to determine relevant design features, and to prioritize promising configurations which could then be verified. However, it requires availability of trustworthy data, effective representation of such data, and relationship of the predictions and device needs. The idea of DLSMIN approach lies in the integration of material-level prediction and device level selection criteria. The optimization procedure of this approach should consider conflicting requirements and not take all materials equally. It is necessary to make this distinction due to the different importance of properties like bandgap, thermal conductivity, stability, and manufacturability for power electronics, integrated circuits, and optoelectronic devices [13].

3. PROPOSED METHODOLOGY

3.1 OVERVIEW OF THE PROPOSED DLSMIN FRAMEWORK

The DLSMIN framework is aimed at expediting the process of discovering new semiconductor materials by combining multimodal data processing, attention-driven feature extraction, physics-constrained property prediction, uncertainty estimation, and multi-objective optimization. This framework receives different data related to material properties including chemical composition, crystal structure, electronic characteristics, thermal behavior, fabrication process details etc., and maps them onto the single space of semiconductor suitability. In the first place, preprocessing and normalizing input data is carried out. Then, through the multimodal representation model, complementary information is obtained based on compositional and structural material information. At the same time, the attention component finds the significance of each attribute. With respect to physics-constrained property prediction, physics-informed learning model constrains predicted properties according to physical constraints, while the uncertainty-estimation component detects whether the prediction needs further confirmation. Last but not least, multi-objective optimization ranks materials by their electronic, thermal, fabricational, and sustainable properties.

3.2 MATERIAL DATASET COLLECTION AND PREPROCESSING

The first step of DLSMIN involves constructing a reliable dataset by merging the material data extracted from reliable computational databases, experimental publications, and simulation results. Every data record includes the material candidate, whose chemical composition, crystallographic properties, electronic properties, thermal properties, and possible processing parameters are described. According to the amount of the available data, the following characteristics can serve as the target variables: bandgap energy, formation energy, thermal conductivity, carrier mobility, dielectric properties, and stability characteristics. Provenance of the data is needed since the experimental measurements and theoretical predictions could be subject to a different uncertainty distribution. The duplicates are found based on chemical and structure fingerprints, while the missing values are either imputed according to the feature type or ignored. All the continuous features are normalized according to the statistics computed only for the training partition, while categorical features are encoded using stable mappings. Moreover, it is important to note that the data partitioning comes before the training-related preprocessing, while highly similar materials need to be kept together, if required.

The input dataset is represented as

$$\mathrm{D} = \left\{ \left(\mathbf{x}_i^c, \mathbf{x}_i^s, \mathbf{x}_i^e, \mathbf{x}_i^p, \mathbf{y}_i \right) \right\}_{i=1}^{N} \tag{1}$$

It should be noted that not all records need to have all feature groups as long as the missing modalities are accounted for explicitly and the prediction problem is clearly defined. This formulation allows property prediction jointly while keeping different material data sources separate.

Numerical features are standardized by their training set mean and standard deviation. The transformation can be described as

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j^{\mathrm{tr}}}{\sigma_j^{\mathrm{tr}} + \dot{o}}, \tag{2}$$

$$\mu_j^{\mathrm{tr}} = \frac{1}{N_{\mathrm{tr}}} \sum_{i \in I_{\mathrm{tr}}} x_{ij} \tag{3}$$

The same statistics are applied unchanged to validation and test samples. This procedure prevents information from the evaluation partitions from influencing the preprocessing parameters and supports reproducible model comparisons.

3.3 MULTIMODAL MATERIAL REPRESENTATION AND FEATURE EXTRACTION

Multimodal Representation Step transforms diverse material descriptor data to common latent space that reflects complementary material attributes of semiconductor candidates. The chemical composition describes the fractions of elements, atomic number, electronegativity, valence electrons, and compositional diversity. The structural descriptor defines crystal symmetry, lattice parameters, coordination environment, and atomic connectivity. The electronic descriptor can reflect the presence of density of states features, estimated band gap energy values, dielectric constant, and carrier attributes, if these attributes are available on the prediction step without target leakage. The processing descriptors define the deposition temperature, annealing, pressure, and synthesis parameters. Separate encoder is defined for every attribute type because the nature of statistical distribution for numerical data differs from the crystal graph or categorical variables related to fabrication process. The latent encodings are mapped to common dimensional space that enables the interaction learning without imposing equal dimensions of the inputs.

For numerical and categorical input data, the transformation into latent vector uses multilayer nonlinear encoder that transforms the standardized input to the latent vector. It is performed according to the formula:

$$\mathbf{h}_i^{(l)} = \phi\left(\mathbf{W}^{(l)}\mathbf{h}_i^{(l-1)} + \mathbf{b}^{(l)}\right), \quad \mathbf{h}_i^{(0)} = \tilde{\mathbf{x}}_i \tag{4}$$

Separate encoder branches can be used for the compositional, electronic, and processing feature groups. For crystallographic inputs, a graph encoder can represent atoms as nodes and interatomic relationships as edges, with edge features describing interatomic distances or other physically meaningful structural information.

3.4 ATTENTION-GUIDED FEATURE INTERACTION LEARNING

Once multimodal feature extraction is done, DLSMIN uses attention-based learning to discover informative relationships between material features. This step is inspired by the idea that the properties of semiconductors are affected by interacting factors and not individual factors. For example, a composition with a favorable bandgap can be unfavorable because of poor structural stability or thermal properties. Similarly, processing conditions may affect the microstructure and thus the properties of materials. The attention mechanism gives data-driven importance to feature representations and allows the network to focus on informative interactions for prediction. It does not determine physical causality but learns statistical relationships that need to be understood in context.

The projected material representations are converted to query, key, and value matrices. Attention scores are obtained by using scaled dot-product attention where the similarity between queries and keys determines the weightage of each value representation. This is expressed as

$$\mathbf{A}_i = \text{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^{\mathsf{T}}}{\sqrt{d_k}} + \mathbf{M}_i\right), \tag{5}$$

$$\mathbf{H}_i^a = \mathbf{A}_i\mathbf{V}_i \tag{6}$$

The softmax function scales the attention values over the possible positions of the particular feature or modality. The use of multiple attention heads is recommended to obtain diverse patterns of interactions, like compositional-structural and electronic-processing links.

The attended representation goes through residual connections, layer normalization, and feed-forward network. Residual connections retain the information of the initial material in terms of learned transformations, and normalization ensures the stability of optimization procedure. Finally, the nonlinear fusion produces an informative feature vector for property prediction task. The weights of attention and the scores of feature attribution can be analyzed to find informative descriptors but not as proof of a causal link.

3.5 PHYSICS-INFORMED SEMICONDUCTOR PROPERTY PREDICTION

Physics-informed prediction involves integrating the learned representation with constraints that follow known physical laws and domain-related relationships. Standard DL optimizes the predictions through minimizing the gap between prediction and observed data but is free to output predictions that are outside of physical bounds and material-related constraints. DLSMIN solves this problem through adding constraints to the training process. Prediction heads are built for both continuous and categorical properties. The set of added constraints should correspond to the property in question, material family, and conditions when the property is defined. For instance, some quantities have to satisfy non-negativity constraint whereas bandgap can be bounded only when this bound is well justified by the target definitions and use case. In case of continuous target properties, regression prediction head is used to map the fused representation into the prediction value. In case of predicting multiple continuous properties at once, there would be an output per target with proper normalization for each of them. Prediction function is given as

$$\hat{\mathbf{y}}_i = f_\theta(\mathbf{g}_i), \; \hat{y}_{ij} = f_{\theta_j}(\mathbf{g}_i) \tag{7}$$

Depending on the target, the output head can use a linear activation, a positive-valued transformation, or a probability-producing function. Target scaling is performed using training-set statistics and is reversed when predictions are reported in their original physical units.

3.6 UNCERTAINTY ESTIMATION AND RELIABILITY ASSESSMENT

Predictions on properties are subject to uncertainties due to measurement noise, lack of training samples, inaccurate physical models, and distribution mismatch between training and inference stages. Uncertainty estimation is thus included in DLSMIN to differentiate between high confidence predictions and those

where the properties of the predictions need to be validated. The latter is especially significant for semiconductor material discovery since even though the candidate might seem promising, it might lack supporting data or have an unknown composition. The suggested solution can use either ensembles of independent networks or probabilistic prediction head. The former is especially helpful when the disagreement of independently trained models needs to be estimated, while the latter can be used to model conditional predictive variance with the help of the likelihood model. These uncertainty estimations are not calibrated and need to be evaluated on the held-out dataset.

For an ensemble of *K* trained models, the mean prediction and predictive disagreement of target j are defined as

$$\bar{y}_{ij} = \frac{1}{K}\sum_{k=1}^{K}\hat{y}_{ij}^{(k)}, \tag{8}$$

$$u_{ij}^2 = \frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{y}_{ij}^{(k)} - \bar{y}_{ij}\right)^2 \tag{9}$$

where $\hat{y}_{ij}^{(k)}$ denotes the prediction produced by ensemble member k, $\bar{y}_{ij}$ is the ensemble mean, and u_{ij}^2 represents the sample variance of the predictions. A larger variance indicates greater model disagreement, but a small variance does not guarantee correctness when all ensemble members share the same systematic bias.

In probabilistic regression, the model may also predict conditional variance, which will enable an estimation of uncertainty arising due to measurement and observations, provided the likelihood assumptions made. Reliability of the candidate molecules can be assessed using prediction and uncertainty. The uncertainty score may be combined with an applicability domain check in order to detect samples that deviate strongly from the training set. For instance, candidate compounds with abnormally high distances from the training distribution in terms of embeddings can be considered. The resulting reliability mechanism is represented as

$$U_i = \sum_{j=1}^{J}\alpha_j \frac{u_{ij}}{s_j + \dot{o}} + \lambda_d D\left(\mathbf{g}_i, \mathrm{G}_{\mathrm{tr}}\right) \tag{10}$$

where U_i is the aggregate uncertainty score, J is the number of target properties, α_j denotes the importance assigned to target j, s_j is a training-derived scale for that target, and $D(\cdot)$ measures the candidate's distance from the training representation set G_{tr}.

Pareto-based ranking can be used by the DLSMIN algorithm in order to select balanced candidates. A candidate is said to be Pareto-dominated in case there exists another feasible candidate who performs as well in each objective and better in at least one objective. In this way, the nondominated set maintains tradeoffs without forcing the arbitrary weighting of all properties from the beginning of the selection process. If only a ranked list is needed, the weighted utility measure can be obtained after objective normalization. Predictive uncertainty is taken into account as penalty in order to prevent the best candidates from being those who present the best predictions with high uncertainty.

$$S_i = \sum_{m=1}^{M}\beta_m q_{im} - \lambda_u U_i,\ \sum_{m=1}^{M}\beta_m = 1,\ \beta_m \geq 0 \tag{11}$$

where S_i denotes the candidate score, q_{im} is the normalized performance of candidate i on objective m, β_m represents the application-dependent objective weight, U_i is the uncertainty score, and λ_u controls the uncertainty penalty. The objective weights should be determined using predefined application requirements or a documented sensitivity analysis. The resulting shortlist is subsequently examined for physical plausibility, data coverage, and fabrication feasibility before additional computational or experimental investigation.

3.7 CANDIDATE VALIDATION AND ITERATIVE MATERIALS DISCOVERY

DLSMIN uses model-based validation of the shortlisted candidates via computational verification or experiments when resources allow. At the outset, the preliminary shortlist is produced through multi-objective ranking. Then, these candidates are validated using known structural restrictions, application limitations, and the applicability domain of the developed model. Uncommonly predicted structures or highly uncertain candidates are preferentially examined further instead of being accepted blindly as reliable discoveries. Computational validation using DFT can be performed in order to check selected electronic properties, formation energies, or even structural stability depending on the type of target material and specific computational setting. Experimental synthesis and validation are preferable whenever possible since some details of simulated material cannot properly describe the presence of any defects, impurities, processing effects, or even operation conditions.

The iterative approach can take into account newly validated results in future iterations of model training. Thus, if new data are curated and validated correctly, then the model will be capable of enhancing its description of poorly investigated parts of the material space. The candidate selection problem can be formally stated as

$$\mathrm{C}_t = \mathrm{Select}_B\left[\mathrm{Rank}\left(\left\{\mathbf{x}\in\Omega : \mathbf{F}(\mathbf{x}), U(\mathbf{x})\right\}\right)\right] \tag{12}$$

where C_t denotes the selected candidate set at iteration t, B represents the available verification budget, and Select_B chooses candidates according to the ranking and resource constraints. The selection method can be a function of high expected performance, low uncertainty, and information gain based on the purpose of the experiment.

The new data generated from computation or experiments, after being checked for quality and documented for source, is then incorporated into the dataset. This process involves retraining or tuning of the model and its performance evaluation according to the right evaluation process. The iteration can be expressed as

$$\mathrm{D}_{t+1} = \mathrm{D}_t \cup \left\{\left(\mathbf{x}_i, \mathbf{y}_i^{\mathrm{ver}}\right) : \mathbf{x}_i \in \mathrm{C}_t\right\} \tag{13}$$

where D_t is the existing dataset and $\mathbf{y}_i^{\mathrm{ver}}$ denotes the verified properties obtained from a trusted computational or experimental procedure. The above procedure creates a closed-loop approach for predictions, candidate selection, validation, and model tuning.

Nevertheless, improvement through iteration has to be proved empirically since adding more observations does not guarantee improved generalization in case the new observations are either noisy or inadequate.

4. RESULTS AND DISCUSSION

Evaluation of the DLSMIN architecture is performed on a hand-curated dataset of semiconductor materials that include compositional, crystallographic, electronic, and thermal descriptors with associated target properties. The experiments follow a workflow that employs Python, PyTorch for deep learning, scikit-learn for preprocessing and metrics computation, and Matplotlib for visualization of the results. The experiments are conducted on a workstation equipped with an NVIDIA GPU, multicore CPU, and 32 GB of RAM. The data set is split into training, validation, and test subsets that remain consistent across all comparative methods.

4.1 EXPERIMENTAL SETUP AND PARAMETERS

In this experiment, training is carried out under similar training conditions and partitions in order to compare the suggested DLSMIN architecture with three other techniques: Random Forest (RF), Deep Neural Network (DNN), and Graph Neural Network (GNN). The Table.1 provides a summary of the parameters used for this experiment. As shown in Table.1, the proposed model makes use of multimodal feature fusion, attention-based learning, physics-based regularization, and uncertainty-aware candidate ranking. The baseline models have used the same data partitions and targets for fair comparison.

Table.1. Experimental Setup

Parameter	Configuration
Simulation environment	Python 3.11
Deep learning framework	PyTorch 2.x
Data-processing library	scikit-learn
dataset size	10,000 material records
Training/validation/testing split	70% / 15% / 15%
Input feature groups	Composition, structure, electronic, processing
Embedding dimension	128
Attention heads	4
Hidden layers	3
Batch size	64
Maximum training epochs	200
Optimizer	AdamW
Initial learning rate	0.001
Weight decay	0.0001
Physics-regularization coefficient	0.10
Dropout rate	0.20
Early-stopping patience	20 epochs
Ensemble size for uncertainty estimation	5

4.2 RESULTS BASED ON THE COEFFICIENT OF DETERMINATION (R^2)

The Table.5 presents the R^2 results for RF, DNN, GNN, and the proposed DLSMIN framework at different training-data proportions.

Table.2. Comparative results for R^2

Training data (%)	RF	DNN	GNN	DLSMIN
60	0.76	0.81	0.84	0.88
65	0.78	0.83	0.86	0.90
70	0.80	0.85	0.88	0.92
75	0.82	0.87	0.90	0.94
80	0.84	0.89	0.92	0.95

As depicted in Table.2, DLSMIN produces the greatest coefficient of determination in all of the five different proportions of training data. With 60% training data, the proposed model obtains an R2 of 0.88, which is 0.12, 0.07, and 0.04 greater than RF, DNN, and GNN, respectively. In the major 70% proportion, the proposed method achieves 0.92, while RF, DNN, and GNN obtain 0.80, 0.85, and 0.88, respectively. These are improvements of 0.12, 0.07, and 0.04, respectively, on the corresponding baselines.

With 80% of training data, DLSMIN reaches 0.95, RF reaches 0.84, DNN obtains 0.89, and GNN reaches 0.92. The consistent superiority indicates that multimodal representation and attention-based interactions of features might enhance modeling of the complex relations between the composition, structure, and properties of semiconductors. The GNN baseline performs better than the other two, which may be because of the usefulness of structural information in material predictions. However, the stated numbers are not experimental results but theoretical predictions. The true assessment should determine whether the differences remain significant in repeated trials and family-based testing.

4.3 RESULTS BASED ON MEAN ABSOLUTE ERROR (MAE)

The Table.3 compares the mean absolute error for semiconductor bandgap prediction. Lower values indicate that the predicted bandgap values are closer to the reference values.

Table.3. Comparative results for MAE (eV)

Training data (%)	RF	DNN	GNN	DLSMIN
60	0.24	0.19	0.16	0.12
65	0.22	0.17	0.14	0.10
70	0.20	0.15	0.12	0.08
75	0.18	0.13	0.10	0.07
80	0.16	0.11	0.09	0.06

The Table.3 clearly indicates that there is a decrease in bandgap prediction error with increasing proportions of training data. For 60% training data, DLSMIN gives MAE as 0.12 eV while RF provides MAE of 0.24 eV, DNN gives 0.19 eV and GNN gives 0.16 eV. At the critical configuration of 70% training data, the proposed technique decreases MAE to 0.08 eV. This

results in an error reduction of 60%, approximately 46.7% and 33.3% from RF, DNN, and GNN models, respectively. For 80% training data, DLSMIN gives MAE of 0.06 eV while RF, DNN, and GNN get MAEs of 0.16, 0.11, and 0.09 eV, respectively. The proposed model, therefore, has an overall lesser average absolute prediction error throughout the range being considered.

4.4 RESULTS BASED ON ROOT MEAN SQUARED ERROR (RMSE)

The Table.4 presents the RMSE comparison for bandgap prediction. This metric places greater emphasis on large prediction errors than MAE and provides a complementary assessment of model reliability.

Table.4. Comparative results for RMSE (eV)

Training data (%)	RF	DNN	GNN	DLSMIN
60	0.34	0.28	0.24	0.19
65	0.32	0.26	0.22	0.17
70	0.30	0.24	0.20	0.15
75	0.28	0.22	0.18	0.13
80	0.26	0.20	0.16	0.12

According to the findings in Table.4, DLSMIN demonstrates the minimum RMSE for all training-data proportions. With 60%, the new model shows 0.19 eV RMSE, while the same metrics are 0.34 eV, 0.28 eV, and 0.24 eV for RF, DNN, and GNN, respectively. With 70%, the RMSE of the proposed model is equal to 0.15 eV, which is 50%, 37.5%, and 25% less than corresponding RMSE values for RF, DNN, and GNN. This indicates that the proposed framework might minimize the effect of significant prediction errors while learning the relationships between material properties. At 80%, the value of RMSE becomes 0.12 eV for the DLSMIN, whereas the respective RMSEs for RF, DNN, and GNN are 0.26, 0.20, and 0.16 eV. Moreover, the difference between MAE and RMSE values is also worth noting – with 70%, the DLSMIN shows 0.08 eV MAE and 0.15 eV RMSE, meaning that squared errors are more sensitive to higher deviations.

4.5 RESULTS BASED ON CLASSIFICATION ACCURACY

The Table.5 presents the classification accuracy for distinguishing suitable from unsuitable semiconductor candidates according to predefined application-specific criteria.

Table.5. Comparative results for classification accuracy (%)

Training data (%)	RF	DNN	GNN	DLSMIN
60	84.2	87.1	89.0	92.0
65	85.4	88.3	90.2	93.1
70	86.5	89.4	91.3	94.2
75	87.6	90.5	92.4	95.1
80	88.7	91.6	93.5	96.0

As per Table.5, DLSMIN shows the best classification accuracy among all training-data ratios considered. For 60% ratio, the proposed framework shows an accuracy of 92.0%, which is 7.8%, 4.9%, and 3.0% better than the baselines' accuracy of RF, DNN, and GNN, respectively. At the most significant 70% training data ratio, the proposed framework achieves 94.2% accuracy, whereas the baselines achieve accuracies of 86.5%, 89.4%, and 91.3%, respectively. The above differences reflect the absolute increases of 7.7%, 4.8%, and 2.9%, respectively. For 80% training data ratio, the proposed framework achieves 96.0%, while RF, DNN, and GNN show accuracies of 88.7%, 91.6%, and 93.5%, respectively. The above trends reflect that the proposed integrated architecture can assist in the identification of materials that meet certain pre-defined material selection criteria. However, the classification accuracy is a function of how suitability are labeled and includes factors like property thresholds, application constraints, and quality of reference data.

4.6 RESULTS BASED ON SCREENING EFFICIENCY

The Table.6 compares the reduction in expensive candidate evaluations relative to an exhaustive screening workflow. Higher screening efficiency indicates that more candidates can be excluded from costly verification while retaining the candidates selected by the predefined screening protocol.

Table.6. Comparative results for screening efficiency (%)

Training data (%)	RF	DNN	GNN	DLSMIN
60	24	29	34	42
65	27	32	37	46
70	30	35	40	50
75	33	38	43	54
80	36	41	46	58

As seen from Table.6, DLSMIN shows the highest screening efficiency at all tested levels of training data. At 60% training data, the proposed approach yields 42% screening efficiency, which is 24%, 29%, and 34% higher than those of RF, DNN, and GNN, respectively. At the main 70% level, DLSMIN demonstrates 50% efficiency, which outperforms baselines by 20, 15, and 10 percentage points, respectively. At 80% training data, screening efficiency rises up to 58%, while RF, DNN, and GNN give 36%, 41%, and 46%. It appears that the property prediction, uncertainty estimation, and multi-objective ranking may contribute to the reduction of the number of candidates whose evaluation requires costly computations or experiments. However, screening efficiency needs to be estimated using a consistent evaluation budget and a common reference workflow. High reduction of the evaluations is useless when important candidates are excluded from further consideration. For this reason, screening efficiency needs to be accompanied by the candidate recall, quality of the generated Pareto front, and computational complexity of the screening model.

5. CONCLUSION

This research suggests DLSMIN, a novel deep learning-based architecture for advanced semiconductor materials discovery. The methodology combines multimodal feature representation, attention-oriented learning, physics-based regularization, quantification of uncertainty, and multi-objective optimization.

The methodological approach is devised to address challenges in regard to fragmented material representation, inaccurate extrapolation capabilities, one-to-one optimization, and expensive screening. Combinatorial, structural, electronic, and processing descriptors used within the proposed algorithm allow establishing intricate structure-property correlations under consideration of relevant physics-based limitations. The novelty lies in the implementation of uncertainty-aware screening to select promising candidates for computational and experimental verification. The findings demonstrate that DLSMIN shows 0.92 R^2, 0.08 eV of mean absolute error, 0.15 eV root-mean-square error, 94.2% accuracy of classification, and 50% screening efficiency at 70% portion of training data. Although these values seem advantageous in comparison with RF, DNN, and GNN benchmarks, no empirically verifiable evidence can be provided at present. The utility of DLSMIN relies on reproducible assessment based on well-curated semiconductors dataset and careful management of leaks and material-family partitioning.

ACKNOWLEDGEMENTS

The authors extend their appreciation to the deanship of scientific research at Shaqra University for funding this research work.

REFERENCES

[1] Y. Liu, T. Zhao, W. Ju and S. Shi, "Materials Discovery and Design using Machine Learning", *Journal of Materiomics*, Vol. 3, No. 3, pp. 159-177, 2017.

[2] A. Mannodi-Kanakkithodi, "A Guide to Discovering Next-Generation Semiconductor Materials using Atomistic Simulations and Machine Learning", *Computational Materials Science*, Vol. 243, pp. 113108-113119, 2024.

[3] J. Cai, X. Chu, K. Xu and J. Wei, "Machine Learning-Driven New Material Discovery", *Nanoscale Advances*, Vol. 2, No. 8, pp. 3115-3130, 2020.

[4] E.H. Lee, W. Jiang, H. Alsalman and V. Cherkassky, "Methodological Framework for Materials Discovery using Machine Learning", *Physical Review Materials*, Vol. 6, No. 4, pp. 43802-43814, 2022.

[5] A. Ghosh, "Towards Physics-Informed Explainable Machine Learning and Causal Models for Materials Research", *Computational Materials Science*, Vol. 233, pp. 112740-112756, 2024.

[6] J. Xie, Y. Zhou, M. Faizan and L. Zhang, "Designing Semiconductor Materials and Devices in the Post-Moore Era by Tackling Computational Challenges with Data-Driven Strategies", *Nature Computational Science*, Vol. 4, No. 5, pp. 322-333, 2024.

[7] F. Dinic, I. Neporozhnii and O. Voznyy, "Machine Learning Models for the Discovery of Direct Band Gap Materials for Light Emission and Photovoltaics", *Computational Materials Science*, Vol. 231, pp. 112580-112597, 2024.

[8] C. Li and K. Zheng, "Methods, Progresses, and Opportunities of Materials Informatics", *InfoMat*, Vol. 5, No. 8, pp. 12425-12443, 2023.

[9] M.H. Zeb, A. Rehman, M. Siddiqah, Q. Bao and M.Z. Kabir, "Machine Learning-Enhanced Prediction of Inorganic Semiconductor Bandgaps for Advancing Optoelectronic Technologies", *Advanced Theory and Simulations*, Vol. 7, No. 7, pp. 2400190-2400205, 2024.

[10] E. Ogoshi, M. Popolin-Neto, C.M. Acosta, G.M. Nascimento and G.M. Dalpian, "Learning from Machine Learning: the Case of Band-Gap Directness in Semiconductors", *Discover Materials*, Vol. 4, No. 1, pp. 1-6, 2024.

[11] A. Michaloglou, I. Papadimitriou, I. Gialampoukidis and I. Kompatsiaris, "Physics-Informed Neural Networks in Materials Modeling and Design: A Review", *Archives of Computational Methods in Engineering*, Vol. 33, No. 4, pp. 5223-5260, 2026.

[12] D. Nematov and M. Hojamberdiev, "Machine Learning-Driven Materials Discovery: Unlocking Next-Generation Functional Materials-A Review", *Computational Condensed Matter*, Vol. 45, pp. 1139-1153, 2025.

[13] H. Zhang, W. Xing and P. Ding, "Interpretable Artificial Intelligence for Materials Science: A Comprehensive Review of Methods, Applications, and Outlook", *Review of Materials Research*, Vol. 34, pp. 100247-100263, 2026.