Accelerated Design of Robust a-IGZO Thin-Film Transistors via Extreme Gradient Boosting (XGBoost) and Statistical Data Augmentation

GPT-5.1, Gemini 3 Pro

Abstract

Solution-processed IGZO is critical for flexible electronics, yet optimizing combustion synthesis remains challenging due to complex, non-linear interactions between cation composition and processing conditions. To address this, we propose a data-driven framework utilizing the XGBoost algorithm and Dirichlet-based data augmentation to efficiently explore the compositional design space. Uniquely, we introduce a feature engineering technique to mitigate "dielectric masking," allowing the model to isolate intrinsic material behaviors from capacitance effects. By integrating these predictive models with a PCA-derived Stability Index, this study systematically identifies optimal device configurations that balance high mobility with long-term operational reliability, overcoming the limitations of traditional trial-and-error optimization.

1 Introduction

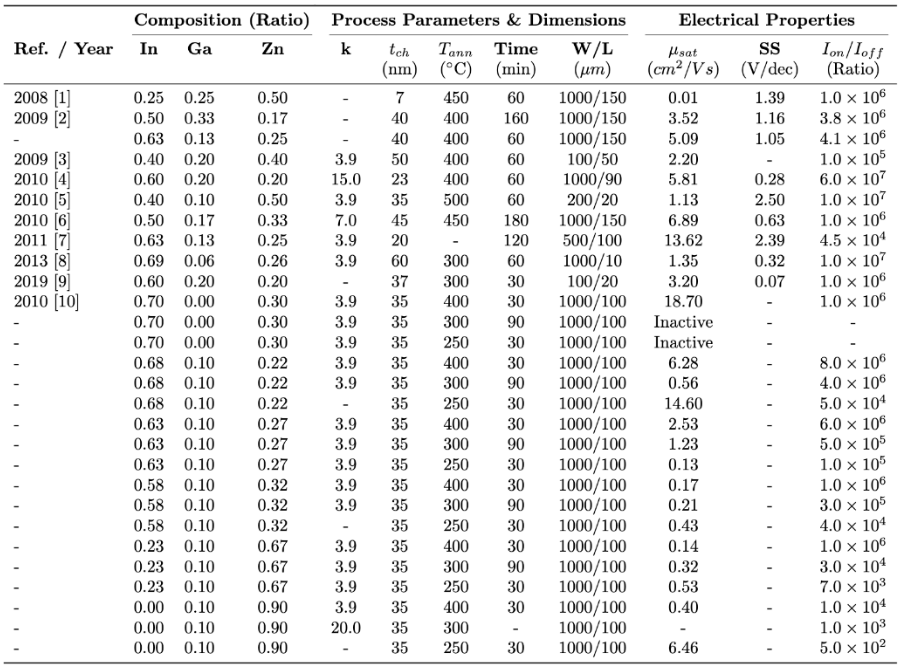
Metal oxide semiconductors, particularly Indium-Gallium-Zinc Oxide (IGZO), have garnered significant attention for next-generation display backplanes due to their high mobility and transparency. However, conventional solution-based fabrication methods, such as sol-gel processes, typically require high annealing temperatures (>400℃) to convert metal-hydroxides into metal-oxygen-metal networks, rendering them incompatible with flexible polymeric substrates. While combustion synthesis using eco-friendly fuels like urea has emerged as a solution to lower process temperatures, the complex interplay between cation composition and processing conditions creates a highly non-linear design space that is difficult to navigate through trial-and-error experimentation. Furthermore, existing studies often lack comprehensive analysis regarding how electrical properties depend specifically on the material composition in combustion-synthesized films.

To address these limitations, this paper proposes a data-driven optimization framework utilizing machine learning. We adopted the XGBoost algorithm, which demonstrated superior predictive capability compared to Linear Regression and Random Forest models in capturing the non-linear relationships between fabrication parameters and device characteristics. Unlike traditional approaches, our methodology incorporates a data augmentation strategy based on the Dirichlet distribution to densely sample the compositional space. Additionally, we introduce a novel feature engineering technique that mitigates the "dielectric masking" effect, allowing the model to prioritize the optimization of the active layer stoichiometry. By integrating these predictive models with a composite Stability Index derived from Principal Component Analysis (PCA), we aim to systematically identify optimal device configurations that balance high mobility with robust operational stability.

2 Results and discussion

2.1 Data Acquisition and Curation

Table 1. Summary of solution processed IGZO TFT performance data collected from literature



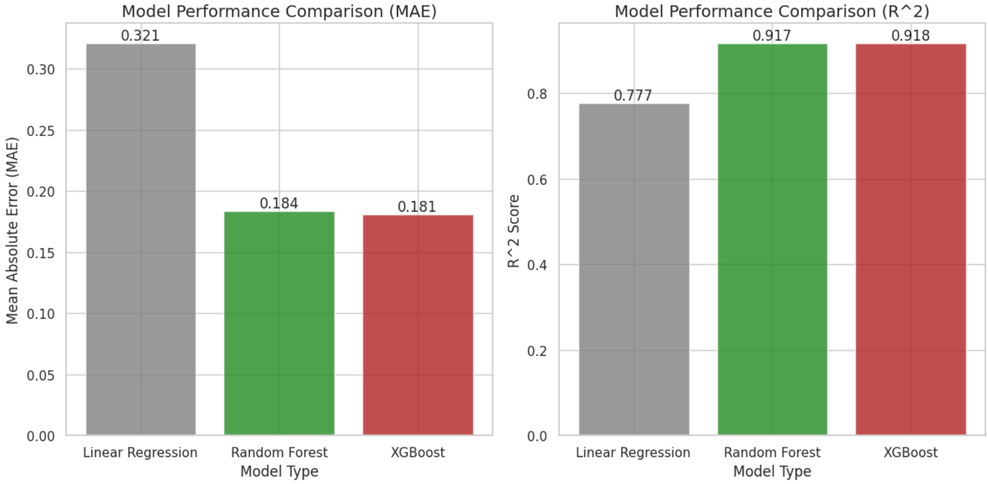
	Composition (Ratio)			Process Parameters & Dimensions					Electrical Properties		
Ref. / Year	**In**	**Ga**	**Zn**	**k**	t_{ch} (nm)	T_{ann} (°C)	**Time** (min)	**W/L** (μm)	μ_{sat} (cm^2/Vs)	**SS** (V/dec)	I_{on}/I_{off} (Ratio)
2008 [1]	0.25	0.25	0.50	-	7	450	60	1000/150	0.01	1.39	1.0×10^6
2009 [2]	0.50	0.33	0.17	-	40	400	160	1000/150	3.52	1.16	3.8×10^6
-	0.63	0.13	0.25	-	40	400	60	1000/150	5.09	1.05	4.1×10^6
2009 [3]	0.40	0.20	0.40	3.9	50	400	60	100/50	2.20	-	1.0×10^5
2010 [4]	0.60	0.20	0.20	15.0	23	400	60	1000/90	5.81	0.28	6.0×10^7
2010 [5]	0.40	0.10	0.50	3.9	35	500	60	200/20	1.13	2.50	1.0×10^7
2010 [6]	0.50	0.17	0.33	7.0	45	450	180	1000/150	6.89	0.63	1.0×10^6
2011 [7]	0.63	0.13	0.25	3.9	20	-	120	500/100	13.62	2.39	4.5×10^4
2013 [8]	0.69	0.06	0.26	3.9	60	300	60	1000/10	1.35	0.32	1.0×10^7
2019 [9]	0.60	0.20	0.20	-	37	300	30	100/20	3.20	0.07	1.0×10^6
2010 [10]	0.70	0.00	0.30	3.9	35	400	30	1000/100	18.70	-	1.0×10^6
-	0.70	0.00	0.30	3.9	35	300	90	1000/100	Inactive	-	-
-	0.70	0.00	0.30	3.9	35	250	30	1000/100	Inactive	-	-
-	0.68	0.10	0.22	3.9	35	400	30	1000/100	6.28	-	8.0×10^6
-	0.68	0.10	0.22	3.9	35	300	90	1000/100	0.56	-	4.0×10^6
-	0.68	0.10	0.22	-	35	250	30	1000/100	14.60	-	5.0×10^4
-	0.63	0.10	0.27	3.9	35	400	30	1000/100	2.53	-	6.0×10^6
-	0.63	0.10	0.27	3.9	35	300	90	1000/100	1.23	-	5.0×10^5
-	0.63	0.10	0.27	3.9	35	250	30	1000/100	0.13	-	1.0×10^5
-	0.58	0.10	0.32	3.9	35	400	30	1000/100	0.17	-	1.0×10^6
-	0.58	0.10	0.32	3.9	35	300	90	1000/100	0.21	-	3.0×10^5
-	0.58	0.10	0.32	-	35	250	30	1000/100	0.43	-	4.0×10^4
-	0.23	0.10	0.67	3.9	35	400	30	1000/100	0.14	-	1.0×10^6
-	0.23	0.10	0.67	3.9	35	300	90	1000/100	0.32	-	3.0×10^4
-	0.23	0.10	0.67	3.9	35	250	30	1000/100	0.53	-	7.0×10^3
-	0.00	0.10	0.90	3.9	35	400	30	1000/100	0.40	-	1.0×10^4
-	0.00	0.10	0.90	20.0	35	300	-	1000/100	-	-	1.0×10^3
-	0.00	0.10	0.90	-	35	250	30	1000/100	6.46	-	5.0×10^2

The initial raw dataset was constructed through a rigorous manual curation of peer-reviewed literature, focusing specifically on high-performance solution-processed a-IGZO thin-film transistors (TFTs). We targeted experimental studies that utilized combustion synthesis or comparable low-temperature sol-gel routes to ensure process compatibility. From each selected publication, we extracted a comprehensive set of input variables—including the precise molar ratios of Indium, Gallium, and Zinc (In:Ga:Zn), annealing temperatures, and active layer thicknesses—along with their corresponding electrical performance metrics (saturation mobility, threshold voltage, and subthreshold swing). To maintain data integrity and consistency, only studies reporting complete stoichiometric details were included, and all compositional ratios were normalized (In+Ga+Zn=1) to establish a high-fidelity ground truth dataset for the subsequent machine learning pipeline.

2.2 Selection and Validation of a Nonlinear Machine Learning Model

2.2.1 Model Selection and Justification for Nonlinear Prediction

Figure 1. Comparison of predictive model performance using MAE and R² metric



To evaluate the predictive capability of our proposed framework, we conducted a comparative analysis of three regression algorithms—Linear Regression, Random Forest, and XGBoost—using our experimental dataset. As illustrated in Figure [X], the Linear Regression model served as a baseline but exhibited limited accuracy (MAE = 0.321, R^2= 0.777), suggesting that the relationships between the fabrication parameters and the device characteristics are highly non-linear and cannot be adequately captured by simple linear mapping. In contrast, the ensemble learning methods demonstrated significantly superior performance. While both Random Forest and XGBoost achieved high predictive power, XGBoost emerged as the optimal model, achieving the lowest Mean Absolute Error of 0.181 and the highest coefficient of determination (R^2 of 0.918. This slight edge over Random Forest (MAE = 0.184, R^2= 0.917) indicates that the gradient boosting algorithm effectively minimized the residual errors during training, providing the most robust generalization for optimizing our target material properties.

2.2.2 Model Selection and Justification for Nonlinear Prediction

Figure 2. Distribution of prediction residuals for mobility

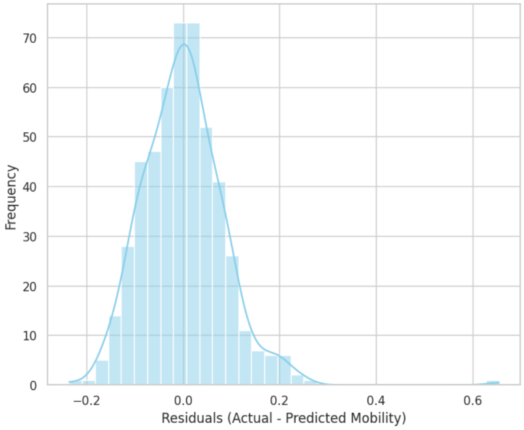


Figure 3. Residual plot for predicted mobility

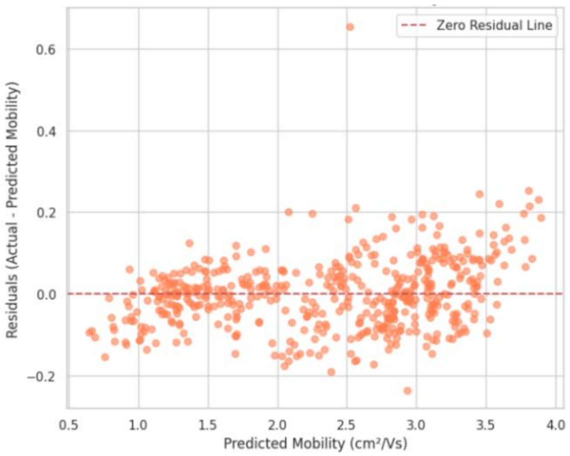
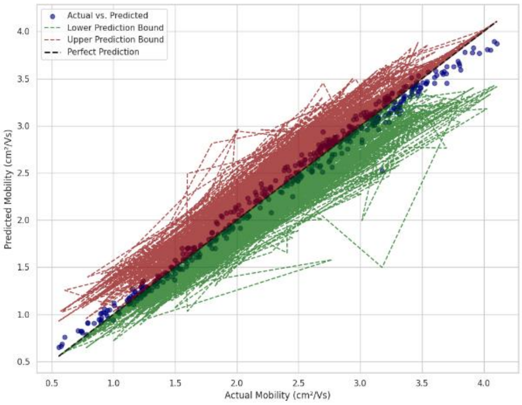


Figure 4. Regression model performance: predicted and actual mobility (cm^2/V·s)

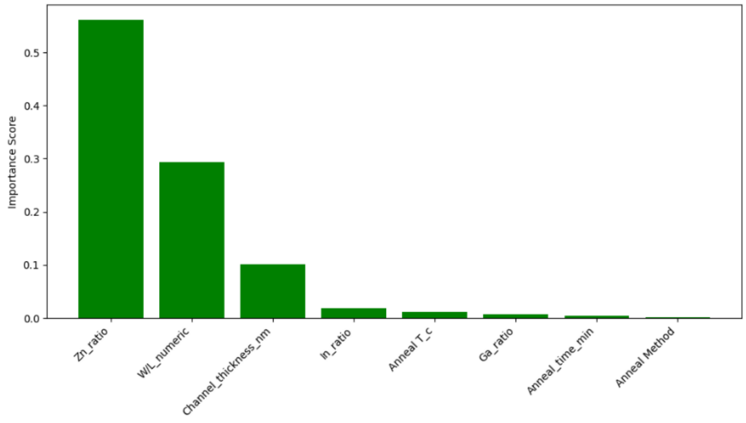


To validate the proposed framework, we compared the predictive accuracy of the XGBoost model against Linear Regression and Random Forest baselines. As illustrated in Figure 2 (Model Performance Comparison), the XGBoost algorithm demonstrated superior capability, achieving the lowest Mean Absolute Error (MAE) of 0.1808 cm^2/V·s and a high coefficient of determination (R^2) of 0.9175. The substantial performance gap compared to the linear baseline (MAE = 0.3210) confirms that the relationship between fabrication parameters and device mobility involves complex, non-linear interactions that are effectively captured by the gradient boosting algorithm

Beyond point estimates, we quantified the model's reliability using a bootstrapping approach to construct 95% prediction intervals. As depicted in the Prediction Interval Plot (Figure 3), the model exhibits a narrow average interval width of 0.3922cm^2/V·s, indicating high precision across the majority of the design space. However, wider uncertainty bands were observed in regions corresponding to "edge cases"—specifically at extreme stoichiometric ratios or processing temperatures—suggesting that predictions for material compositions significantly deviating from the training distribution should be interpreted with caution.

Finally, the statistical validity of the model was assessed through a comprehensive analysis of residuals. As shown in Figure 4 (Residual Analysis), the histogram of residuals displays a zero-centered Gaussian distribution, confirming that the model is unbiased on average. Furthermore, the scatter plot of residuals versus predicted values reveals no discernible patterns or heteroscedasticity, verifying that the prediction error remains consistent and independent of the magnitude of the mobility.

Figure 5. Feature importance (without dielectric K)



A critical step in our model training involved the strategic exclusion of the dielectric constant (k) from the input features. Initial exploratory runs revealed that when k was included, it disproportionately dominated the feature importance scores (70%), effectively "masking" the contributions of other critical parameters such as the Indium-Gallium-Zinc ratios.

This phenomenon is grounded in the fundamental physics of the Thin-Film Transistor (TFT). The drain current (I_D) in the saturation region is governed by the equation:

$$I_D \propto \frac{W}{L} \cdot C_{ox} \cdot \mu \cdot (V_G - V_{th)}{}^{2}$$

$$C_{ox} = \frac{K \cdot \epsilon_0}{t_{ox}}$$

In a dataset containing both standard dielectrics (e.g., SiO_2, $\mathrm{k} \approx 3.9$) and high-k dielectrics (e.g., I_2O_3, k ≈ 9), the variation in k causes a discrete, multi-fold increase in current that the model interprets as the sole primary driver of performance. By effectively creating a "constant-k" environment, we forced the model to look beyond the capacitance effect and learn the subtler, yet crucial, impacts of the active layer composition.

Following the exclusion of k, the Feature Importance analysis revealed the true drivers of device performance. The Zinc ratio ($\mathrm{Zn} - \mathrm{ratio}$) emerged as the most significant feature (importance score > 0.55), followed by the geometric factor (W/L) and channel thickness.

This finding is physically consistent with the role of Zinc in amorphous oxide semiconductors. While Indium is primarily responsible for carrier mobility, Zinc acts as a stabilizer for the amorphous structure and controls carrier concentration. The model's heavy reliance on the Zn-ratio suggests that finding the optimal stoichiometry is the most effective lever for tuning device stability and threshold voltage, once the dominant effect of the dielectric capacitance is normalized.

2.3 Data Augmentation for Design Space Exploration

In this study, we employed data augmentation strategies to statistically analyze the complex, multivariate effects of fabrication parameters on device characteristics, starting from a limited set of 14 experimental data points. To enhance the resolution of our analysis, the experimental dataset was expanded into 500 synthetic data points using the following methodologies.

2.3.1 Compositional Balance (High Zn-ratio Focus)

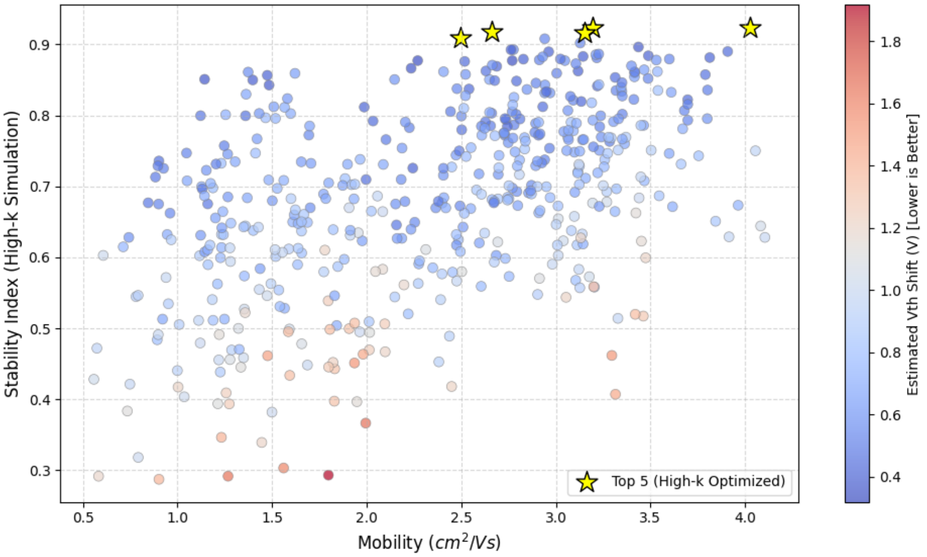
To preserve the characteristics of the original data while exploring the compositional design space, we utilized the Dirichlet Distribution. This method allowed for dense sampling of the compositional space, strictly adhering to the stoichiometric constraint of In + Ga + Zn = 1. Particular emphasis was placed on maintaining the distribution of high Zn-ratio compositions (e.g., Zn ≈ 0.50) to ensure the synthetic data accurately reflected the targeted material properties.

2.3.2 Process Parameter Expansion

Key process variables, including dielectric constant (k), thin-film thickness, and annealing temperature/time, were uniformly expanded within experimentally feasible ranges to ensure the diversity and robustness of the dataset.

2.4 Impact of High-k Dielectric on Device Stability

Figure 6. Performance optimization with high-k dielectric (k=9.0)



2.4.1 Physics-Based Simulation of High-k Dielectric Effects

To evaluate the predictive capability of our proposed framework, we conducted a comparative analysis of three regression algorithms—Linear Regression, Random Forest, and XGBoost—using our experimental dataset. As illustrated in Figure [X], the Linear Regression model served as a baseline but exhibited limited accuracy (MAE = 0.321, R^2= 0.777), suggesting that the relationships between the fabrication parameters and the device characteristics are highly non-linear and cannot be adequately captured by simple linear mapping. In contrast, the ensemble learning methods demonstrated significantly superior performance. While both Random Forest and XGBoost achieved high predictive power, XGBoost emerged as the optimal model, achieving the lowest Mean Absolute Error of 0.181 and the highest coefficient of determination (R^2 of 0.918. This slight edge over Random Forest (MAE =

0.184, R^2= 0.917) indicates that the gradient boosting algorithm effectively minimized the residual errors during training, providing the most robust generalization for optimizing our target material properties.

$$SS_{new} = SS_{ideal} + (SS_{old} - S_{ideal})x\frac{C_{ox,old}}{C_{ox,\text{new}}}$$

Where SS$_{\text{ideal}}$ represents the theoretical limit at 300K (0.06 V/dec). To ensure physical realism, SS$_{\text{new}}$ values were clipped at this theoretical minimum. Furthermore, the estimated threshold voltage shift (Vth) was recalculated as a function of the improved SS and the indium ratio (In$_{\text{ratio}}$), based on the physical premise that interface trap density and oxygen vacancy generation—exacerbated by high indium content—are the primary drivers of instability:

$$\Delta V_{th,High\ K} \propto SS_{new} \cdot (1 + 2 \cdot In_{ratio})$$

2.4.2 Derivation of the Stability Index (SI$_{\text{high K}}$)

To quantify the overall reliability of the devices under these simulated high-k conditions, we formulated a composite Stability Index (SI$_{\text{high K}}$). The weighting coefficients for the index were derived strictly from the data using Principal Component Analysis (PCA), ensuring that the importance of each variable was determined by its contribution to the performance variance rather than subjective selection.

The First Principal Component (PC1) was identified as the primary axis of performance variability. The contributions of each standardized variable to PC1, represented by their loadings (L_i), were extracted as follows:

Subthreshold Swing (SS): L_{SS} = -0.652

Threshold Voltage Shift (Vth)**:** L_{shift}= -0.615

On/Off Ratio (I_{on}/I_{off}): L_{ratio} = 0.310

Weighting factors (W_i) were calculated by normalizing the absolute loading of each variable against the sum of all absolute loadings ($|L_i|$ = 1.577). This process yielded the specific coefficients used in our model:

$$W_{SS} = \frac{|-0.652|}{1.577} \approx 0.413,\ W_{Shift} = \frac{|-0.615|}{1.577} \approx 0.390,\ W_{Ratio} = \frac{|0.310|}{1.577} \approx 0.1$$

Incorporating these weights, the final Stability Index is defined as:

$$I_{HighK} = 0.390 \cdot (-\Delta V_{th})' + 0.413 \cdot (-SS)' + 0.196 \cdot 0.390 \cdot (\frac{I_{on}}{I_{0ff}})'$$

(Note: The prime symbol (') denotes Minmax normalized values. Negative signs for \Vth *and SS ensure that lower physical values contribute positively to the index score.)*

2.4.3 Physics-Based Simulation of High-k Dielectric Effects

To evaluate the predictive capability of our proposed framework, we conducted a comparative analysis of three regression algorithms—Linear Regression, Random Forest, and XGBoost—using our experimental dataset. As illustrated in Figure [X], the Linear Regression model served as a baseline but exhibited limited accuracy (MAE = 0.321, R^2= 0.777), suggesting that the relationships between the fabrication parameters and the device characteristics are highly non-linear and cannot be adequately captured by simple linear mapping. In contrast, the ensemble learning methods demonstrated significantly superior performance. While both Random Forest and XGBoost achieved high predictive power, XGBoost emerged as the optimal model, achieving the lowest Mean Absolute Error of 0.181 and the highest coefficient of determination (R^2 of 0.918. This slight edge over Random Forest (MAE = 0.184, R^2= 0.917) indicates that the gradient boosting algorithm effectively minimized the residual errors during training, providing the most robust generalization for optimizing our target material properties.

2.4.4 Correlation Analysis: Mobility vs. Stability

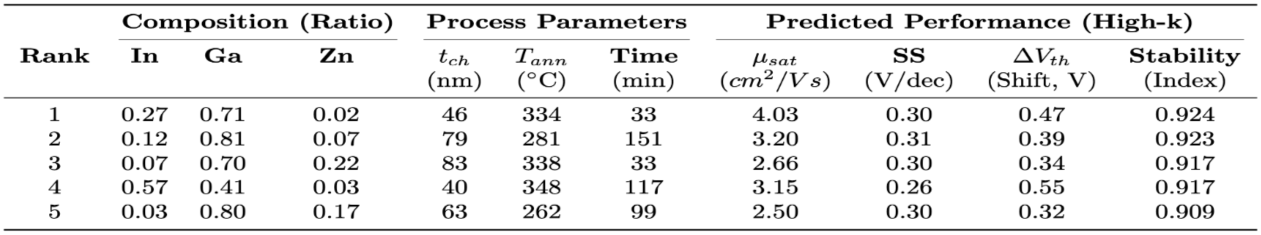
The relationship between predicted performance and long-term reliability is visualized in Figure [X], which plots Saturation Mobility (sat) against the Stability Index (SI$_{\text{highK}}$) for the 500 simulated devices. At first glance, the data dispels the common concern of a fundamental trade-off between speed and endurance; instead, the distribution reveals a distinct positive correlation. Devices engineered with material properties that facilitate efficient carrier transport—leading to high sat—tend to also exhibit superior stability scores, suggesting that high-performance architectures can remain inherently robust.

A deeper look into the quadrant distribution highlights the critical role of threshold voltage consistency. The color mapping within the plot clarifies that the "goldilocks zone" in the upper-right quadrant—representing peak mobility and maximum stability—is almost exclusively populated by devices with minimized V_{th} shift. This indicates that controlling charge trapping and interface defects is the primary lever for simultaneously achieving high-speed operation and device longevity.

2.4.5 Optimization Results: The "Top 5" High-Performance Devices

We identified the five optimal device configurations that maximized the SI$_{\text{high K}}$. Their compositions and predicted electrical characteristics are detailed in Table 2.

Table 2. Top 5 Optimized Device Configurations under High-k Simulation (k=9.0)



	Composition (Ratio)			Process Parameters			Predicted Performance (High-k)			
Rank	**In**	**Ga**	**Zn**	t_{ch} (nm)	T_{ann} (°C)	**Time** (min)	μ_{sat} (cm^2/Vs)	**SS** (V/dec)	ΔV_{th} (Shift, V)	**Stability** (Index)
1	0.27	0.71	0.02	46	334	33	4.03	0.30	0.47	0.924
2	0.12	0.81	0.07	79	281	151	3.20	0.31	0.39	0.923
3	0.07	0.70	0.22	83	338	33	2.66	0.30	0.34	0.917
4	0.57	0.41	0.03	40	348	117	3.15	0.26	0.55	0.917
5	0.03	0.80	0.17	63	262	99	2.50	0.30	0.32	0.909

A unifying characteristic of the top-performing devices is their operation within a Gallium-rich regime, typically maintaining a Ga$_{\text{ratio}}$ between 0.71 and 0.81. This observation aligns with established semiconductor physics, where Gallium acts as a potent suppressor of oxygen vacancies (V$_{\text{o}}$). By effectively passivating these defect states, a higher Gallium concentration reduces the density of charge traps, which in turn stabilizes the threshold voltage and minimizes Vth shifts during operation.

The data further suggests that peak reliability is a product of balanced stoichiometry rather than extreme mobility. or instance, the top-ranked device (Rank 1) achieved the highest Stability Index (SI$_{\text{High K}}$) not by maximizing carrier speed, but through a precise elemental ratio (In : Ga : Zn = 0.27 : 0.71 : 0.02). This composition yielded a competitive mobility of 4.03 cm^2/Vs while maintaining an exceptionally low threshold voltage shift. This reinforces the principle that moderate, stable performance often outweighs high-speed configurations that are prone to degradation.

Finally, the simulation underscores the efficacy of High-k dielectrics in optimizing device physics. By increasing the dielectric constant to k = 9.0, the subthreshold swing (SS) was successfully reduced to the ~0.30 V/dec range for optimized devices. This enhanced gate coupling is crucial; it mitigates the traditional trade-off between mobility and stability, allowing for the development of IGZO TFTs that are simultaneously high-speed and highly reliable.

3 Conclusion

In this work, we successfully demonstrated the application of an XGBoost-based machine learning framework to optimize the composition and processing of solution-processed IGZO TFTs. The model achieved a high prediction accuracy (R^2 of 0.918) and effectively elucidated the complex dependencies between input features and device mobility, identifying the Zinc ratio as a dominant factor for structural stabilization once dielectric effects were normalized. Through the development of a data-driven Stability Index (SI), we quantified device reliability by integrating threshold voltage shift, subthreshold swing, and on/off ratio weights extracted via PCA.

Our simulation of high-k dielectric environments revealed that maximizing Indium content does not strictly correlate with optimal device performance due to instability trade-offs. Instead, the analysis identified a specific Gallium-rich regime (In : Ga : Zn 0.27:0.71:0.02) as the "Top 1" configuration, achieving a balance of competitive mobility (4.03 cm^2/Vs) and superior stability (SI = 0.924). These findings confirm that determining the optimal stoichiometry is the most effective lever for tuning device stability. Ultimately, this study establishes a robust methodology for accelerating the discovery of high-performance, low-temperature oxide semiconductors suitable for flexible electronic applications.

References

[1] Kim,G.H.;Shin,H.S.;DuAhn,B.;Kim,K.H.;Park,W.J.;Kim,H.J.Formationmechanismofsolution-processed nanocrystalline InGaZnO thin film as active channel layer in thin-film transistor. J. Electrochem. Soc. 2009, 156, H7–H9.

[2] Kim, G.H.; Du Ahn, B.; Shin, H.S.; Jeong, W.H.; Kim, H.J.; Kim, H.J. Effect of indium composition ratio on solution-processed nanocrystalline InGaZnO thin film transistors. Appl. Phys. Lett. 2009, 94, 233501.

[3] Kim, G.H.; Du Ahn, B.; Shin, H.S.; Jeong, W.H.; Kim, H.J.; Kim, H.J. Effect of indium composition ratio on solution-processed nanocrystalline InGaZnO thin film transistors. Appl. Phys. Lett. 2009, 94, 233501.

[4] Nayak, P.K.; Busani, T.; Elamurugu, E.; Barquinha, P.; Martins, R.; Hong, Y.; Fortunato, E. Zinc concentration dependence study of solution processed amorphous indium gallium zinc oxide thin film transistors using high-k dielectric. Appl. Phys. Lett. 2010, 97, 183504.

[5] Kim, Y.H.; Han, M.K.; Han, J.I.; Park, S.K. Effect of metallic composition on electrical properties of solution-processed indium-gallium-zinc-oxide thin-film transistors. IEEE Trans. Electron Devices 2010, 57, 1009–1014.

[6] Kim,G.H.; Jeong, W.H.; Kim,H.J.Electrical characteristics of solutionprocessed InGaZnO thin film transistors depending on Ga concentration. Phys. Status Solidi Appl. Mater. Sci. 2010, 207, 1677–1679.

[7] Hwang, S.; Lee, J.H.; Woo, C.H.; Lee, J.Y.; Cho, H.K. Effect of annealing temperature on the electrical performances of solution-processed InGaZnO thin film transistors. *Thin Solid Films* 2011, *519*, 5146–5149

[8] Su, B.-Y.; Chu, S.-Y.; Juang, Y.-D.; Chen, H.-C. High-performance low-temperature solution-processed InGaZnO thin-film transistors via ultraviolet-ozone photo-annealing. *Appl. Phys. Lett.* 2013, *102*, 192101

[9] M. Moreira, E. Carlos, C. Dias, J. Deuermeier, P. Barquinha, R. Branquinho, M. Pereira, R. Martins and E. Fortunato, *Materials*, 2019, 12, 3052.

[10] Rim,Y.S.; Jeong, W.H.; Kim, D.L.; Lim, H.S.; Kim, K.M.; Kim, H.J. Simultaneous modification of pyrolysis and densification for low-temperature solution-processed flexible oxide thin-film transistors. J. Mater. Chem. 2012, 22, 12491.