

Machine Learning-guided Prediction of Ionizable Amphiphilic Janus Dendrimers for mRNA Nanomedicine

Wan-Ting Cheng, Heng-Li Zheng, Peng-Yu Zhu, Ji-Na Hao, Da-Peng Zhang*, and Yong-Sheng Li*

Lab of Low-Dimensional Materials Chemistry, Key Laboratory for Ultrafine Materials of Ministry of Education, Frontier Science Center of the Materials Biology and Dynamic Chemistry, School of Materials Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

 Electronic Supplementary Information

Abstract The efficient and safe delivery of messenger RNA (mRNA) therapeutics remains a critical challenge for clinical translation, driving the need for advanced carrier design. Ionizable amphiphilic Janus dendrimers (IAJDs) represent a promising class of carriers; however, their structural complexity and limited available datasets hinder systematic exploration and optimization. In this study, we established a tailored machine-learning framework to investigate the structure-function relationships of IAJDs under a constrained data regime ($n=231$). Conventional molecular fingerprints were found to be suboptimal for representing these macromolecules, motivating the adoption of count-based descriptors and systematic ablation analyses to disentangle the contributions of the substructural features. These experiments identified key functional motifs underlying transfection performance and provided interpretable insights into the IAJD design principles. Complementing these handcrafted descriptors, we further applied deep learning-based molecular embeddings, which captured higher-order chemical semantics and significantly improved predictive accuracy. Collectively, these advances demonstrate that both refined fingerprinting and representation learning approaches can overcome data limitations, enabling the reliable prediction of IAJD activity while offering mechanistic interpretability. This study illustrates the potential of data-driven strategies as hypothesis-generation and prioritization tools for the design of next-generation mRNA delivery systems.

Keywords Janus dendrimer; Machine learning; mRNA delivery; Molecular modeling

Citation: Cheng, W. T.; Zheng, H. L.; Zhu, P. Y.; Hao, J. N.; Zhang, D. P.; Li, Y. S. Machine learning-guided prediction of ionizable amphiphilic Janus dendrimers for mRNA nanomedicine. *Chinese J. Polym. Sci.* 2026, 44, 1154–1164.

INTRODUCTION

Messenger RNA (mRNA) has emerged as a transformative therapeutic modality^[1] owing to its ability to transiently express target proteins without integration into the host genome, thereby minimizing mutagenesis risks.^[2] The unprecedented success of COVID-19 mRNA vaccines has highlighted the critical role of delivery platforms in enabling clinical translation.^[3,4] Among currently available carriers, lipid nanoparticles (LNPs) represent the most established and have been widely employed in commercial mRNA vaccines.^[5] Nevertheless, their inherent limitations, such as multi-component formulation complexity, instability requiring strict cold-chain storage, and predominant liver tropism, restrict broader applications. These challenges have motivated the development of alternative delivery vectors. Recently, ionizable amphiphilic Janus dendrimers (IAJDs),^[6] a class of one-component multifunctional lipid-like macromolecules, have attracted increasing attention. Distinguished by their well-defined molecular architecture, facile self-assembly, and tunable

ionizable domains, IAJDs offer unique advantages over conventional LNPs, including simplified preparation, enhanced structural-function correlation, and the potential for organ-targeted delivery beyond the liver.^[7] Collectively, these properties enable IAJDs as a promising next-generation platform for mRNA vaccine and therapeutic development.

However, despite the remarkable potential of IAJDs as next-generation delivery vectors, the precise relationship between their molecular structures and functional performance remains insufficiently understood. As is well recognized, subtle variations in structural features, —such as ionizable group density, hydrophobic domain configuration, and branching topology, —can profoundly influence self-assembly, cellular uptake, and endosomal escape.^[8,9] These structure-function correlations are critical, since effective structural designs directly translate into higher transfection efficiency (TE) and improved therapeutic outcomes.^[10] Yet, systematic studies that quantitatively link IAJD substructures to their delivery efficacy are still scarce, creating a significant barrier to rational design and optimization of this promising molecular platform.

In recent years, machine learning (ML) has rapidly permeated the fields of chemistry and materials science,^[11] offering a powerful tool for revealing complex structure-property relationships and accelerating molecular discovery. By leverag-

* Corresponding authors, E-mail: dapengzhang@ecust.edu.cn (D.P.Z.)
E-mail: ysli@ecust.edu.cn (Y.S.L.)

Received December 3, 2025; Accepted January 6, 2026; Published online March 5, 2026

Machine Learning-guided Prediction of Ionizable Amphiphilic Janus Dendrimers for mRNA Nanomedicine

Wan-Ting Cheng, Heng-Li Zheng, Peng-Yu Zhu, Ji-Na Hao, Da-Peng Zhang*, and Yong-Sheng Li*

Lab of Low-Dimensional Materials Chemistry, Key Laboratory for Ultrafine Materials of Ministry of Education, Frontier Science Center of the Materials Biology and Dynamic Chemistry, School of Materials Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

Electronic Supplementary Information

Abstract The efficient and safe delivery of messenger RNA (mRNA) therapeutics remains a critical challenge for clinical translation, driving the need for advanced carrier design. Ionizable amphiphilic Janus dendrimers (IAJDs) represent a promising class of carriers; however, their structural complexity and limited available datasets hinder systematic exploration and optimization. In this study, we established a tailored machine-learning framework to investigate the structure-function relationships of IAJDs under a constrained data regime ($n=231$). Conventional molecular fingerprints were found to be suboptimal for representing these macromolecules, motivating the adoption of count-based descriptors and systematic ablation analyses to disentangle the contributions of the substructural features. These experiments identified key functional motifs underlying transfection performance and provided interpretable insights into the IAJD design principles. Complementing these handcrafted descriptors, we further applied deep learning-based molecular embeddings, which captured higher-order chemical semantics and significantly improved predictive accuracy. Collectively, these advances demonstrate that both refined fingerprinting and representation learning approaches can overcome data limitations, enabling the reliable prediction of IAJD activity while offering mechanistic interpretability. This study illustrates the potential of data-driven strategies as hypothesis-generation and prioritization tools for the design of next-generation mRNA delivery systems.

Keywords Janus dendrimer; Machine learning; mRNA delivery; Molecular modeling

Citation: Cheng, W. T.; Zheng, H. L.; Zhu, P. Y.; Hao, J. N.; Zhang, D. P.; Li, Y. S. Machine learning-guided prediction of ionizable amphiphilic Janus dendrimers for mRNA nanomedicine. *Chinese J. Polym. Sci.* 2026, 44, 1154–1164.

INTRODUCTION

Messenger RNA (mRNA) has emerged as a transformative therapeutic modality^[1] owing to its ability to transiently express target proteins without integration into the host genome, thereby minimizing mutagenesis risks.^[2] The unprecedented success of COVID-19 mRNA vaccines has highlighted the critical role of delivery platforms in enabling clinical translation.^[3,4] Among currently available carriers, lipid nanoparticles (LNPs) represent the most established and have been widely employed in commercial mRNA vaccines.^[5] Nevertheless, their inherent limitations, such as multi-component formulation complexity, instability requiring strict cold-chain storage, and predominant liver tropism, restrict broader applications. These challenges have motivated the development of alternative delivery vectors. Recently, ionizable amphiphilic Janus dendrimers (IAJDs),^[6] a class of one-component multifunctional lipid-like macromolecules, have attracted increasing attention. Distinguished by their well-defined molecular architecture, facile self-assembly, and tunable

ionizable domains, IAJDs offer unique advantages over conventional LNPs, including simplified preparation, enhanced structural-function correlation, and the potential for organ-targeted delivery beyond the liver.^[7] Collectively, these properties enable IAJDs as a promising next-generation platform for mRNA vaccine and therapeutic development.

However, despite the remarkable potential of IAJDs as next-generation delivery vectors, the precise relationship between their molecular structures and functional performance remains insufficiently understood. As is well recognized, subtle variations in structural features, —such as ionizable group density, hydrophobic domain configuration, and branching topology, —can profoundly influence self-assembly, cellular uptake, and endosomal escape.^[8,9] These structure-function correlations are critical, since effective structural designs directly translate into higher transfection efficiency (TE) and improved therapeutic outcomes.^[10] Yet, systematic studies that quantitatively link IAJD substructures to their delivery efficacy are still scarce, creating a significant barrier to rational design and optimization of this promising molecular platform.

In recent years, machine learning (ML) has rapidly permeated the fields of chemistry and materials science,^[11] offering a powerful tool for revealing complex structure-property relationships and accelerating molecular discovery. By leverag-

* Corresponding authors, E-mail: dapengzhang@ecust.edu.cn (D.P.Z.)
E-mail: ysli@ecust.edu.cn (Y.S.L.)

Received December 3, 2025; Accepted January 6, 2026; Published online March 5, 2026

ing high-dimensional descriptors and data-driven modeling, ML approaches have demonstrated superior performance in predicting chemical reactivity, material properties, and drug-target interactions.^[12] More recently, these methods have been extended to the design of mRNA delivery systems. For instance, artificial intelligence-guided optimization of LNPs has enabled efficient pulmonary gene therapy,^[13,14] and combinatorial ML-assisted design has facilitated the development of nanoparticles for pulmonary mRNA delivery and genome editing.^[15] These pioneering studies underscores the transformative potential of ML in advancing nucleic acid therapeutics. However, to the best of our knowledge, very few investigations^[16] have applied ML methodologies to IAJDs. This gap limits the systematic understanding and rational design of these unique molecular platforms.

In this study, we addressed the methodological challenge of modeling IAJDs, a class of structurally intricate macromolecules, under a constrained data regime ($n=231$). Traditional molecular fingerprints were initially considered as baseline representations; however, their limited capacity to capture the diverse substructural patterns of IAJDs highlights the need for methodological refinement. To this end, we introduced an improved count-based fingerprinting strategy

and performed a systematic ablation analysis to disentangle the contributions of fingerprint dimensionality, parameterization, and the learning paradigm. This stepwise evaluation not only enabled the identification of an optimal configuration that balanced predictive accuracy with computational tractability but also provided insight into the specific molecular substructures that serve as key determinants of biological activity.

Beyond handcrafted descriptors, we further integrated molecular embeddings derived from a pre-trained transformer-based language model. By leveraging the analogy between SMILES strings and natural language sequences, this approach yields contextualized representations that extend beyond local structural fragments to encompass higher-order semantic information. Embedding-based representations substantially improved predictive performance, demonstrating the value of transferring knowledge from large-scale chemical corpora to low-data regimes, such as the IAJD system.

Taken together, this work contributes to a comprehensive machine-learning framework that combines systematic feature evaluation with state-of-the-art representation learning (Fig. 1). By bridging traditional molecular fingerprints with transformer-based embeddings, the framework achieves both robustness and adaptability, while offering chemically

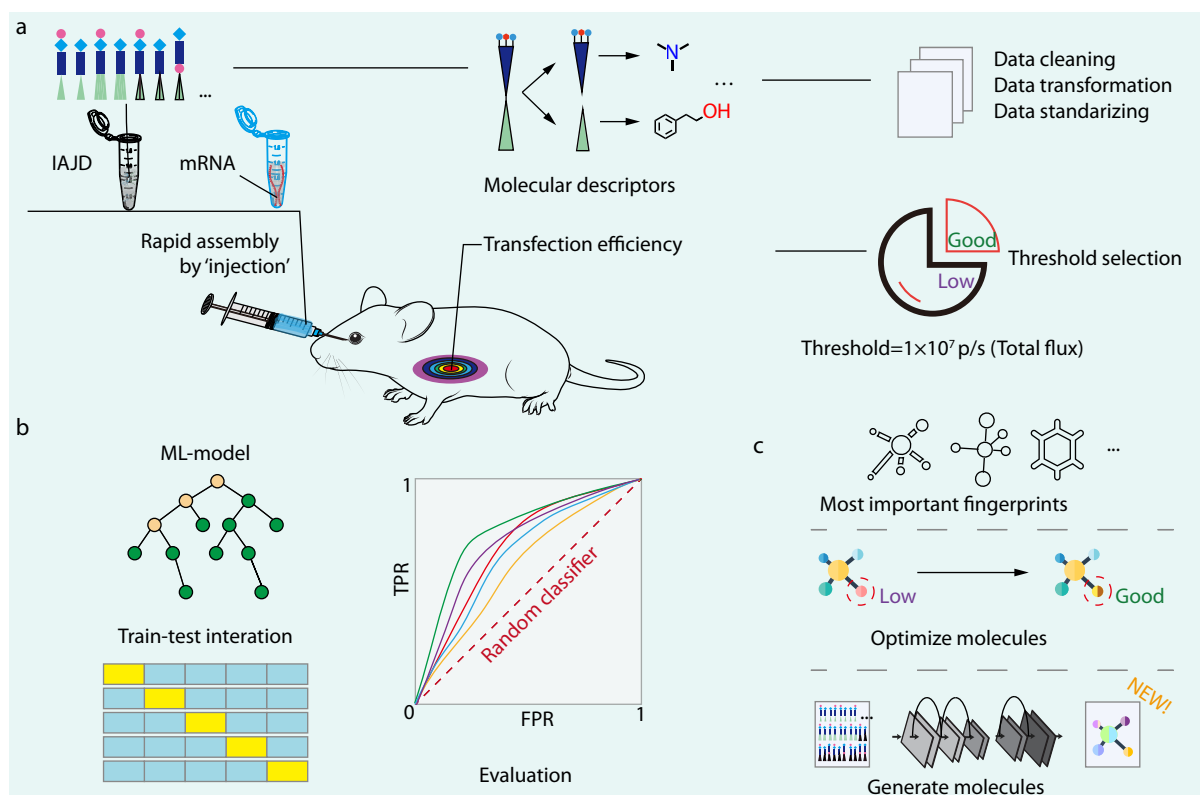


Fig. 1 Overall workflow for IAJD-based molecular analysis and optimization. (a) Experimental and data preprocessing pipeline. IAJDs were assembled with firefly luciferase mRNA via rapid injection method, and their transfection efficiency was measured *in vivo*. Molecular descriptors were calculated from chemical structures and subsequently cleaned, transformed, and standardized. Activity thresholds were defined based on total flux (1×10^7 p/s) to classify samples into Good and Low transfection groups. (b) Machine learning-based modeling and evaluation. Various algorithms were trained using cross-validation, and model performance was assessed by receiver operating characteristic (ROC) analysis to identify optimal predictive configurations. (c) Downstream model interpretation and molecular design. The most informative molecular fingerprints were extracted to reveal key structure-activity relationships. The trained model was further employed for fragment-based molecular optimization and generative design, yielding new candidates with improved predicted transfection efficiency.

interpretable insights into the structural determinants of IA-JD function. Such an integrative strategy highlights the potential of machine learning not only as a predictive tool, but also as a hypothesis-generating framework to guide the rational design and optimization of dendrimer-based mRNA delivery systems.

EXPERIMENTAL

Data Collecting and Preprocessing

The dataset used in this study was sourced from dendrimer-some nanoparticle (DNP) samples assembled from IAJDs reported by the Percec and Weissman laboratories.^[6,17–20] Owing to their distinctive amphiphilic architecture and tunable ionization properties, IAJDs demonstrate favorable biocompatibility and mRNA delivery efficiency and have consequently emerged as a focal point of research in gene and drug delivery. Comprehensive experimental data and structural information have been documented across multiple high-impact publications, thereby providing this work with a robust foundation of prior knowledge and empirical support.

The target compounds were initially drawn in ChemDraw (version 20.0) and exported as SMILES strings,^[21] providing a uniform digital representation of molecular structures. Morgan fingerprints (ECFP)^[22] were then computed to encode local atomic environments into fixed-length numerical vectors, serving as descriptors for machine learning models. Compounds that could not be successfully parsed or contained missing information were excluded to preserve the data integrity and maintain robustness in downstream feature extraction and model training. Preliminary SMILES-level data augmentation was evaluated but did not yield performance gains for IAJDs, likely owing to the structural redundancy inherent to dendritic Janus architectures.

To establish the classification labels, the transfection efficiency (TE) reported in the literature was employed as a quantitative measure of molecular functional activity (Table S1 in the electronic supplementary information, ESI). Given that certain molecules lacked explicit total flux values due to absent fluorescence responses, a threshold of 1×10^7 p/s was applied to binarize the dataset: molecules with total flux values exceeding this threshold were designated as “Good,” whereas those below the threshold or without measurable response were classified as “Low.” This process yielded 125 positive and 103 negative samples, forming a supervised dataset for the model development (Fig. 2a).

We selected a total-flux cutoff of 1×10^7 p/s as a pragmatic operating point because it lies in the mid-range of reported *in vivo* signals for the curated Percec/Weissman IAJD datasets and yields a relatively balanced class split, which improves model stability under small-sample learning. Molecules with no measurable fluorescence/flux were conservatively assigned to the Low class, consistent with a “below-detection” interpretation. To assess the robustness to binarization, we performed a cutoff sensitivity analysis by sweeping the threshold across 1×10^6 , 5×10^6 , 1×10^7 , 5×10^7 , and 1×10^8 p/s. For each cutoff, we re-trained the same modeling pipeline and reported the Area Under the Receiver Operating Characteristic Curve (ROC-AUC), area under the precision-recall curve (PR-AUC), balanced accuracy, and Matthews Correlation Coefficient (MCC) (Fig. S1 and Table S2 in ESI).

Feature Representation

Traditional molecular fingerprints (ECFP)

Extended connectivity fingerprints (ECFPs) are widely used molecular descriptors that encode local atomic environments using the Morgan algorithm. As a baseline, we implemented a k-nearest neighbors (KNN) classifier^[23] using binary ECFP fingerprints, where each molecule was represented as a 2048-bit Morgan fingerprint^[24] with a radius of 2. Classification was performed by weighted majority voting of the k-nearest neighbors in the fingerprint space, with weights proportional to the Tanimoto similarity.^[25]

Improved molecular fingerprints (Count-ECFP)

Despite that ECFPs are widely employed in conventional molecular feature extraction, the repetitive substructures present in IAJD molecules often lead to fingerprint representations with substantial redundancy, limiting their ability to highlight distinctive structural characteristics and consequently diminishing model discriminability and predictive accuracy. To capture the structural features of dendrimer molecules more comprehensively, this study adopted count-based Extended Connectivity Fingerprints (Count-ECFPs)^[22] as molecular descriptors.

In contrast to traditional binary fingerprints, Count-ECFPs quantify the frequency of each substructural motif within a molecule, thereby providing a more information-rich representation. This enhanced encoding improves the ability of machine-learning models to discriminate between structural features. The resulting Count-ECFPs were represented as fixed-length integer vectors, followed by variance-threshold filtering to remove non-informative zero-variance features, which further improved model efficiency and robustness (Fig. 2b). Variance-threshold filtering was implemented within the cross-validation pipeline and fitted on the training folds to prevent information leakage.

ChemBERTa-based molecular embedding

With the rapid advancement of deep learning, pretrained molecular language models have emerged as a powerful paradigm for molecular representation. The SMILES notation, which resembles natural language sequences, provides a natural foundation for transformer architecture. Among these, ChemBERTa^[26] is a Transformer-based molecular language model pretrained on large-scale chemical corpora, such as PubChem and ZINC, enabling the extraction of high-dimensional embeddings that capture both structural and semantic information. The ChemBERTa model was employed as a frozen feature extractor without task-specific fine-tuning, and embedding inference was accelerated using an NVIDIA Tesla T4 GPU. This frozen-encoder strategy was deliberately adopted because recent studies have shown that pretrained molecular language models often achieve superior data efficiency and robustness under low-data regimes compared with task-specific fine-tuning.

Evaluation Criteria

Model performance was primarily evaluated using the area under the receiver operating characteristic curve (AUC-ROC). The area under the curve (AUC) provides a threshold-independent assessment of a classifier's discriminative ability by integrating sensitivity (true positive rate, TPR) and specificity (false positive rate, FPR) across all possible classification thresholds.

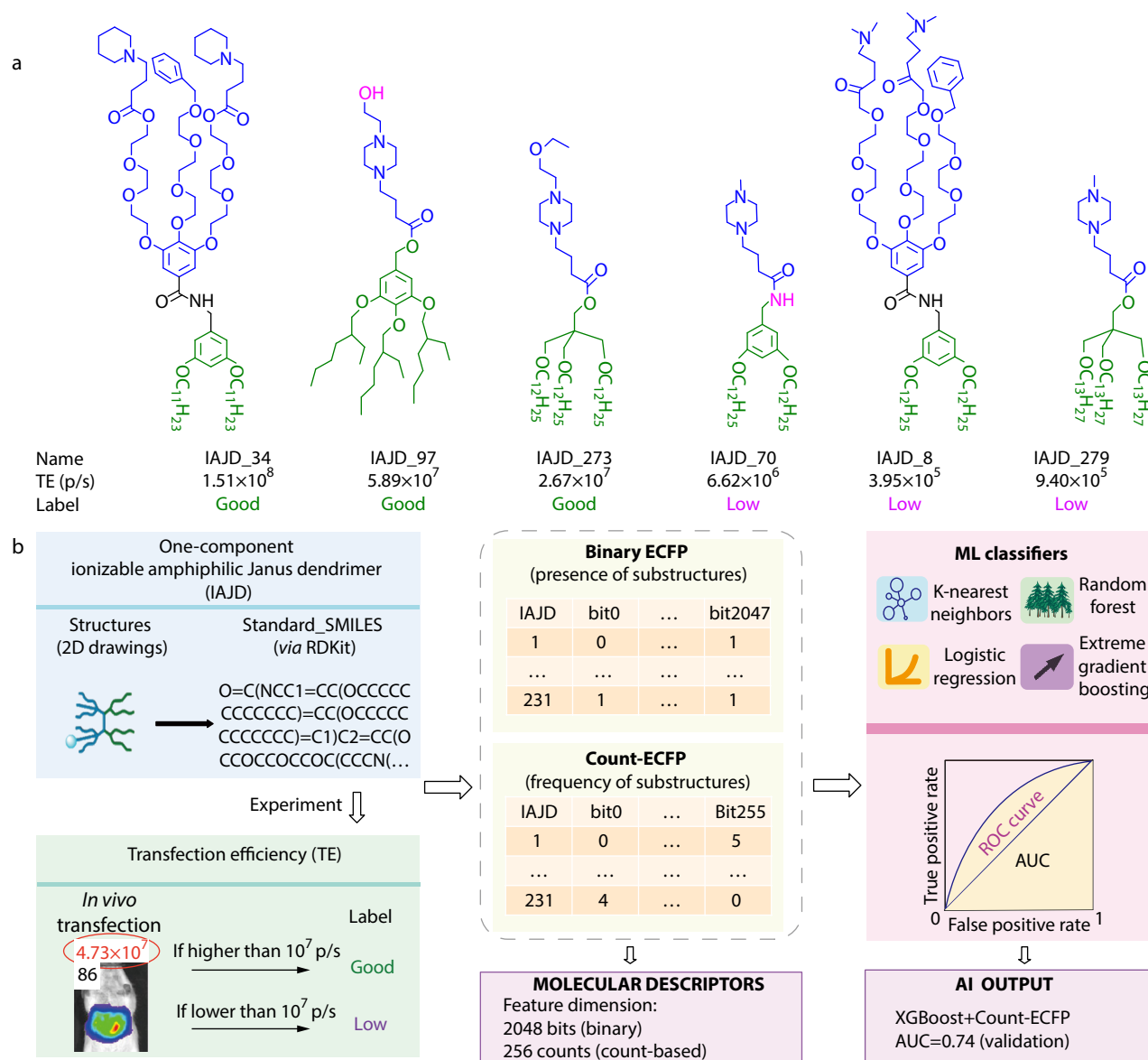


Fig. 2 (a) Examples of IAJD molecules and the resulting transfection efficiency along with their label; (b) Molecular structures were converted into standardized SMILES and featured as binary or count-based ECFPs. These descriptors were used as inputs for machine learning models evaluated under five-fold stratified cross-validation.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (2)$$

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})d(\text{FPR}) \quad (3)$$

where TP represents true positive, FN represents false negative, FP represents false positive, and TN represents true negative.

Compared with accuracy or F1-score, which are tied to a single decision boundary and may be biased under class-imbalanced conditions, AUC offers a more robust and generalizable measure of predictive performance. Model performance was primarily evaluated using the ROC-AUC. In addition, we report PR-AUC to reflect the performance under varying class prevalence, as well as balanced accuracy and MCC as complementary threshold-dependent metrics. This property is partic-

ularly advantageous in molecular classification tasks, where the number of active and inactive compounds is often imbalanced, and the key objective is to distinguish promising candidates from non-functional ones, regardless of the decision threshold. Consequently, the AUC was adopted as the primary metric to ensure a comprehensive and reliable evaluation of model performance.

RESULTS AND DISCUSSION

Baseline Model Performance

To evaluate the predictive capacity of the different molecular representations, we first established baseline models using traditional ECFP fingerprints and simple classifiers. Parameter scans across multiple neighborhood sizes revealed that intermediate values provided the most balanced performance, whereas overly small neighborhoods resulted in unstable predictions. Under

an outer cross-validation protocol, the KNN baseline achieved a mean ROC-AUC of 0.56 ± 0.11 , indicating limited but non-random discriminative ability. Overall, although the KNN baseline establishes a simple and intuitive benchmark, its sensitivity to hyperparameters and limited generalization underscores the need for more expressive molecular representations and advanced modeling approaches.

Comparison of ML Models

In addition to fingerprint parameterization, the influence of classification algorithms was systematically assessed. Three representative machine learning approaches, —Random Forest (RF),^[27] Logistic Regression (LR),^[28] and XGBoost (XGB)^[29], —were benchmarked using a nested stratified cross-validation

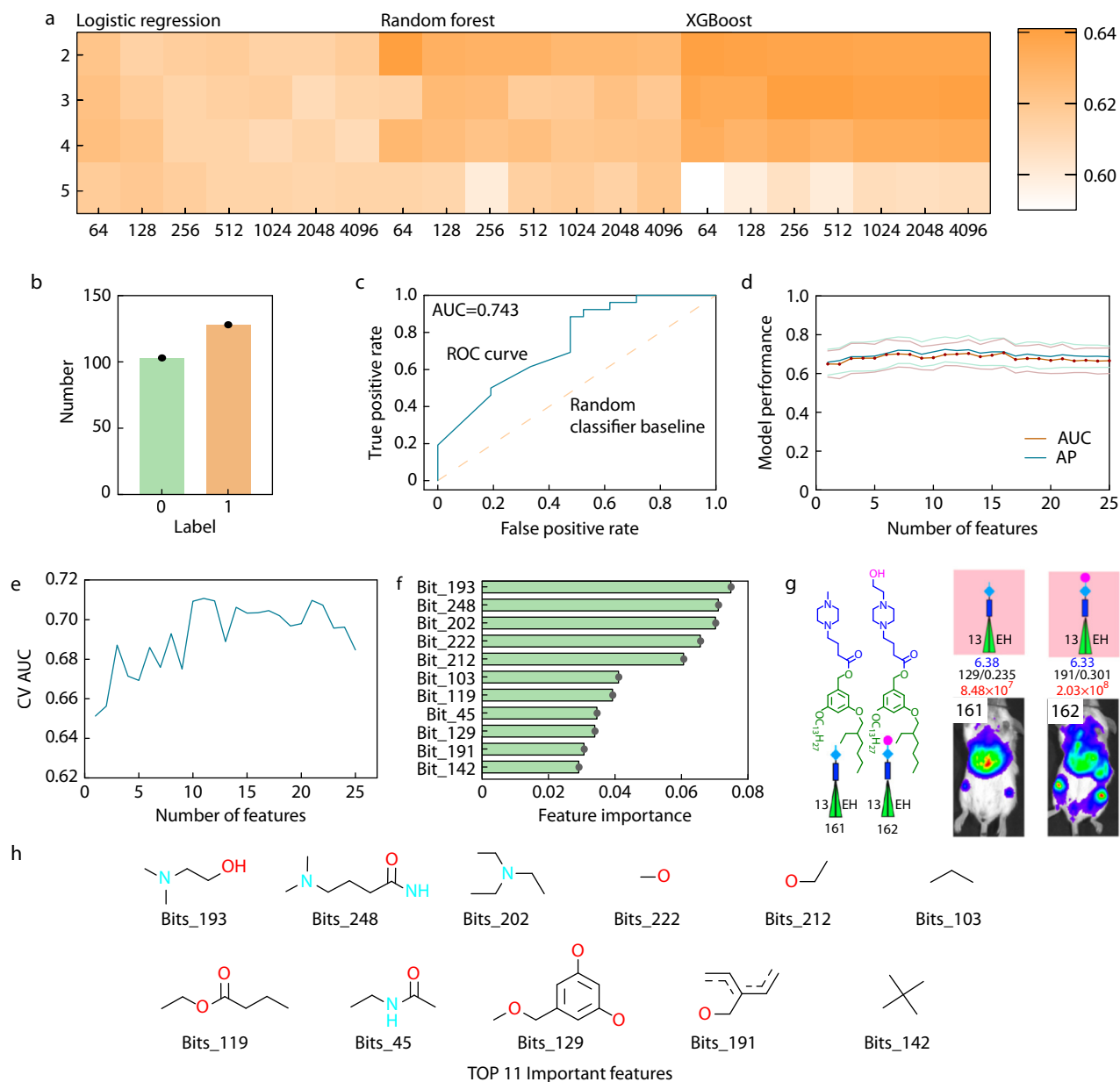


Fig. 3 Feature selection, model performance, and structural interpretation of machine learning models for IAJDs. (a) Comparative performance heatmap of logistic regression, random forest, and XGBoost classifiers under different fingerprint radius and bit lengths, showing that XGBoost achieves superior discriminative ability; (b) Distribution of labeled samples used for model training; (c) Receiver operating characteristic (ROC) curve of the optimized XGBoost model with an area under the curve (AUC) of 0.743, indicating strong predictive capacity; (d) Cross-validation results showing AUC and average precision (AP) as a function of the number of selected features. Shaded bands indicate ± 1 SD across cross-validation folds; (e) Five-fold cross-validation AUC revealing that model performance plateaus when approximately 11 features are used. (f) Ranked feature importance of top molecular substructures (Count-ECFP bits) contributing to classification; (g) Representative comparison of IAJDs with or without a key substructure (Bit_193). Molecules incorporating this feature (e.g., IAJD_162) exhibited markedly higher *in vivo* transfection signals compared to those lacking the feature (e.g., IAJD_161),^[18] which highlights the biological relevance of model-identified motifs and supports the integration of computational interpretation with experimental validation; (h) Chemical structures of the top 11 important substructures identified by the model, highlighting recurrent functional motifs related to transfection efficiency.

tion protocol (inner 5-fold for hyperparameter tuning; outer 5-fold for performance estimation). Hyperparameters were optimized within the inner loop using grid/randomized search strategies to ensure a robust evaluation (Fig. 3a and Table S3 in ESI).

The comparative results demonstrated that XGBoost consistently exhibited superior discriminative capacity across most settings, which is attributable to its ability to capture nonlinear decision boundaries and effectively manage high-dimensional feature spaces. Random Forest achieved stable performance, particularly with lower-dimensional fingerprints, while Logistic Regression benefited from higher-dimensional descriptors in combination with regularization.

Fig. 3(b) provides an overview of the class distribution in the curated dataset, confirming a relatively balanced representation between the two activity labels. Importantly, the optimal configuration was identified using XGBoost at $R=3$ and nBits=256, yielding an AUC of 0.7427 (Fig. 3c). This setting outperformed smaller fingerprints, which lacked representational richness, as well as larger fingerprints, which were hindered by dimensional redundancy. Collectively, these findings underscore that moderate fingerprint complexity, when paired with a powerful classifier, such as XGBoost, achieves the most favorable balance of accuracy, robustness, and computational efficiency (Table S4 in ESI).

Substructure Importance and Interpretability

In addition to predictive accuracy, interpretability is essential for elucidating the structural determinants underlying molecular classification. Accordingly, the trained XGBoost classifier was subjected to feature importance analysis based on the Count-ECFP descriptors (Figs. 3d and 3e). Relative contribution scores were extracted and ranked, and subsets of the top-ranked features were iteratively used for model retraining (Figs. 3f and 3g). Five-fold stratified cross-validation with AUC as the evaluation metric enabled the identification of the point of diminishing returns, that is, the minimal feature subset sufficient to preserve predictive performance while minimizing redundancy.

To link these abstract fingerprint features to chemically interpretable motifs, the most influential bits were mapped to their corresponding molecular fragments using the RDKit. Each bit encodes an atomic environment defined by a central atom and a radius. The associated fragments were reconstructed and visualized as two-dimensional structures. Representative examples of top-ranked bits (e.g., 212, 202, and 103) are shown in Fig. 3(h), revealing functional motifs such as hydroxyl groups, carbonyl fragments, and ionizable tertiary amine substructures. These results demonstrate that chemically meaningful fragments emerged as key determinants of molecular activity, thereby corroborating the biological relevance of the model.

Collectively, ablation experiments provide a comprehensive assessment of how fingerprint representation, model architecture, and substructural features jointly govern predictive performance. Variations in the fingerprint radius and vector length confirmed that moderate complexity affords an optimal trade-off between expressiveness and efficiency. The comparative model evaluation highlighted the superior robustness of XGBoost over Random Forest and Logistic Regression. Importantly, substructure-level interpretability re-

vealed functionally relevant motifs, bridging predictive accuracy with mechanistic understanding and offering actionable knowledge for rational molecular modification.

ChemBERTa Embedding Performance

In contrast to rule-based descriptors such as ECFPs, ChemBERTa representations are data-driven and can capture richer, context-dependent molecular information (Fig. 4a). To assess their utility for classification, we applied principal component analysis (PCA) to 768-dimensional ChemBERTa embeddings to reduce

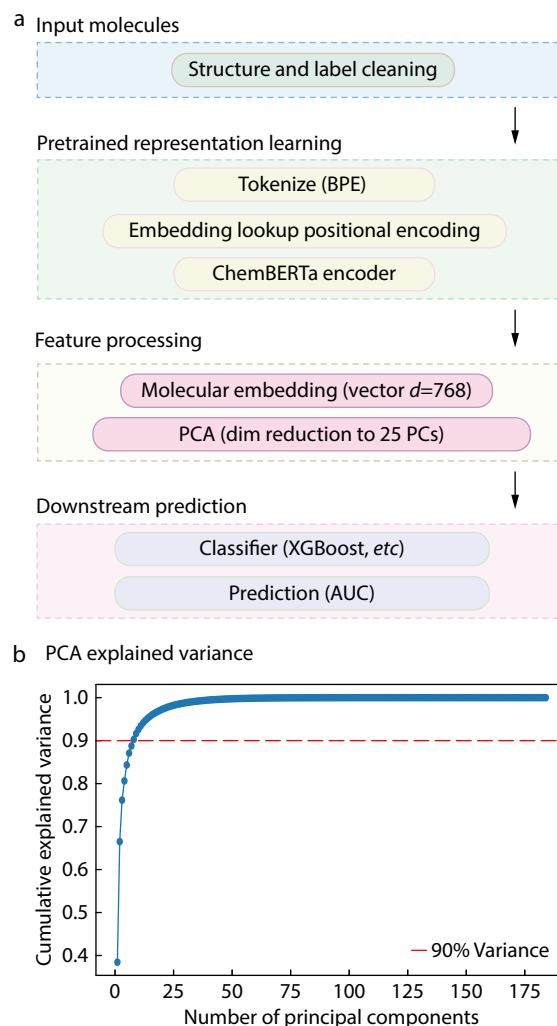


Fig. 4 (a) Workflow of the ChemBERTa-based molecular embedding framework. SMILES strings are tokenized using byte-pair encoding (BPE), followed by embedding lookup, positional encoding, and encoding with the pretrained ChemBERTa Transformer. The resulting 768-dimensional molecular embeddings are reduced to 25 principal components (PCs) *via* PCA for improved efficiency. These processed features are then used by downstream classifiers for prediction, with model performance evaluated using AUC as the primary metric. (b) Cumulative explained variance of PCA applied to ChemBERTa embeddings. The first 25 principal components capture over 90% of the variance (red dashed line), justifying dimensionality reduction prior to downstream classification. In model evaluation, PCA is fit within each training fold and applied to the held-out fold to prevent information leakage.

dimensionality (Fig. 4b). The first 25 principal components explained over 90% of the cumulative variance and were retained as the final feature set, providing a practical trade-off between information preservation and model complexity.

Using these features, we benchmarked multiple classifiers using a leakage-free nested cross-validation protocol. All fitted preprocessing steps (including standardization and PCA) were implemented within a single scikit-learn pipeline and were fit on the training split of each fold only; hyperparameters were tuned in an inner 5-fold loop and performance was estimated in an outer 5-fold loop. Under this protocol, the linear and nonlinear models (LR and RBF-SVM) achieved comparable performance, while the tree-based models achieved slightly lower performance (Table 1).

Overall, these findings demonstrate that ChemBERTa embeddings, when coupled with gradient boosting models, constitute a robust and generalizable framework for predicting molecular properties. This integration of pretrained language-model representations with advanced machine-learning classifiers offers a promising direction for the rational design of dendrimer-related functional materials.

Molecule Generation and Activity Prediction

To further demonstrate the potential of our pipeline for molecular design, we trained a recurrent neural network (RNN) on SELFIES-encoded^[30] representations of 125 positive molecules and sampled novel candidates. The generated SMILES were standardized and six physicochemical descriptors, molecular weight (MW), LgP, topological polar surface area (TPSA), hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), and rotatable bonds (RotB) were computed. Rather than imposing hand-crafted thresholds, admissible ranges were derived from the positive set using the 5th–95th percentile (P5–P95) window for each descriptor, defining an IAJD-compatible property region (Fig. 5a). A novelty-diversity filtering stage is applied to limit overfitting and control memorization. Molecules with high fingerprint similarity to the training positives (Tanimoto $\geq$ 0.85) or those sharing identical Murcko scaffolds were excluded (Fig. 5b). This step ensures that the retained candidates are structurally distinct while remaining proximal to the known active regions (Table S5 in ESI).

Overall, this workflow integrates constrained molecular generation, data-driven property filtering, and ChemBERTa-guided scoring to prioritize plausible and novel Janus dendrimer candidates. Rather than aiming at exhaustive chemi-

cal-space exploration, the generation module functions as a local proposal mechanism, producing a shortlist suitable for downstream synthesis and experimental evaluation.

Molecule Optimization and Filtering Pipeline

A computational framework was established to explore structure-activity relationships and to propose optimized IAJD candidates with predicted enhanced transfection efficiencies. To explore the local chemical space around the inactive candidates, fragment-based single-step edits were applied. A hybrid fragment library was constructed by combining BRICS-decomposed fragments^[31] enriched in active molecules with a set of chemically meaningful substituents. BRICS fragments were selected based on their frequency of occurrence and relative enrichment in experimentally validated active IAJDs, thereby prioritizing substructures that are empirically associated with high transfection efficiency. This activity-guided fragment selection was designed to explore local structure-activity relationships within the chemically coherent IAJD space, rather than to enumerate all synthetically accessible variants. Each fragment contained a defined anchor atom for targeted attachment to the feasible substitution sites on the parent molecule. Candidate structures were automatically generated through combinatorial fragment attachment and were sanitized to ensure chemical validity.

Each generated molecule was evaluated using a multi-objective scoring function integrating both model-predicted activity and physicochemical feasibility. The scoring function is defined as follows:

$$\text{Score} = w_1 P_{\text{active}} - w_2 \text{SA} - w_3 \text{PAINS} + w_4 \text{QED} - w_5 \text{Lipinski} - w_{\text{prop}} \text{Penalty} \quad (4)$$

where P_{active} denotes the calibrated activity probability, SA is the synthetic accessibility score,^[32] quantitative estimate of drug-likeness (QED)^[33] represents drug-likeness, and the penalty term measures physicochemical deviations from the active-molecule distribution.^[34] And the weights were determined through iterative tuning guided by the following design principles: predicted activity ($w_1=1.0$) was assigned unit weight as the primary optimization objective; PAINS filter ($w_3=1.0$) was weighted equally to activity prediction to strictly exclude promiscuous binders, as PAINS-flagged molecules are unsuitable for therapeutic development regardless of predicted activity; synthetic accessibility ($w_2=0.3$) was assigned moderate weight to penalize synthetically inaccessible candidates without overly constraining chemical

Table 1 Performance comparison of classifiers with different fingerprints and embeddings.

Classifier	Fingerprint				Outer-CV ROC-AUC (mean $\pm$ SD)
	ECFP	Count-ECFP	ChemBERTa	Setting	
KNN	✓	×	×		0.559 $\pm$ 0.113
LR	×	✓	×	R_3 / N_256	0.6819 $\pm$ 0.0958
RF	×	✓	×	R_3 / N_256	0.6962 $\pm$ 0.0772
XGBoost	×	✓	×	R_2/N_64	0.7091 $\pm$ 0.0807
XGBoost	×	✓	×	R_3/N_256	0.7319 $\pm$ 0.0709
XGBoost	×	✓	×	R_3/N_2048	0.6877 $\pm$ 0.0533
XGBoost	×	×	✓		0.684 $\pm$ 0.080
SVM	×	×	✓		0.718 $\pm$ 0.053
RF	×	×	✓		0.706 $\pm$ 0.041
LR	×	×	✓		0.712 $\pm$ 0.025

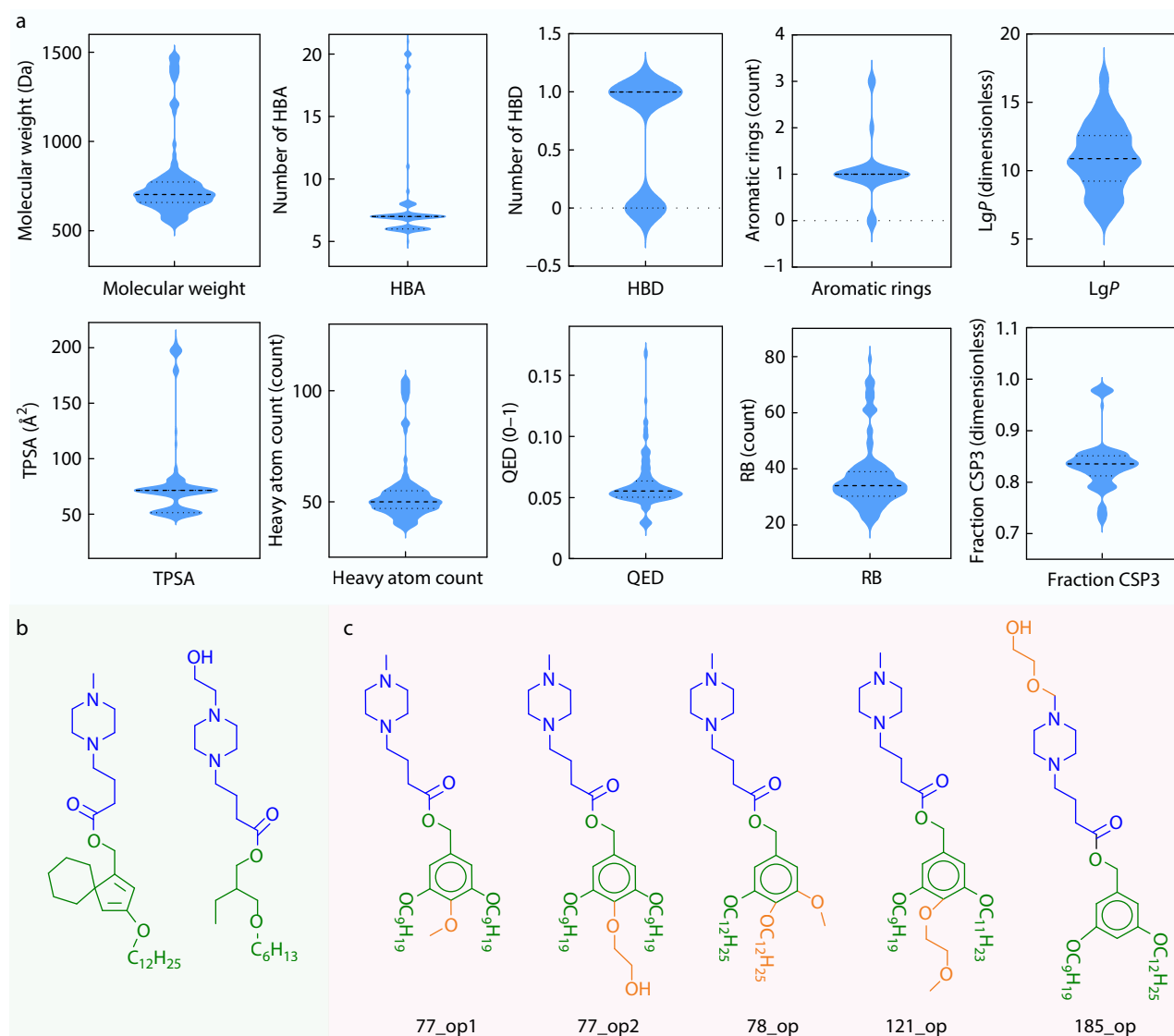


Fig. 5 Physicochemical properties and optimized structures of IAJDs. (a) The distributions of key molecular descriptors in the IAJD dataset, including molecular weight (MW), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), aromatic ring count, LgP , topological polar surface area (TPSA), heavy atom count, quantitative estimate of drug-likeness (QED), rotatable bonds (RB), and fraction of sp^3 carbons (Fraction CSP3); (b) Example molecules generated by the RNN-based proposal model after novelty and property filtering; (c) Optimized IAJD candidates (77_op1, 77_op2, 121_op, 78_op, 185_op) generated via fragment-based molecular editing and multi-objective scoring. The newly introduced fragments are highlighted in orange, corresponding to structural modifications predicted to enhance transfection efficiency while maintaining synthetic accessibility and physicochemical feasibility.

space exploration; QED ($w_4=0.2$) was included as a secondary desirability term to favor candidates with favorable physicochemical profiles; lipinski violations ($w_5=0.0$) were assigned zero weight, as IAJDs are dendritic macromolecules that inherently violate Lipinski's Rule of Five ($MW>500$ Da, $HBA>10$) due to their branched architecture—applying traditional drug-like filters would incorrectly penalize the entire IAJD chemical space; property deviation penalty ($w_{prop}=0.5$) was calibrated to ensure candidates remained within the physicochemical distribution of experimentally validated active IAJDs (5th–95th percentile windows).

To assess the robustness of candidate ranking to weight perturbations, we performed a systematic sensitivity analysis

(Table S6 in ESI). Each weight was independently varied by $\pm 30\%$ while holding others constant, and the candidate molecules were re-scored and re-ranked. Spearman rank correlations between the original and perturbed rankings remained above 0.96 for all perturbations, and the top-10 candidate set exhibited 80%–100% overlap across all weight variations. These results confirm that the prioritization of high-quality candidates is robust to reasonable variations in weight selection, indicating that the scoring function captures stable structure-activity relationships rather than artifacts of parameter tuning (Fig. S2 in ESI).

Candidates that violated predefined thresholds were excluded. The remaining molecules were ranked according to

their composite scores, and structurally diverse top-ranked candidates were selected using a greedy max-min algorithm to avoid redundancy. The final optimized molecules were exported for visualization and further experimental consideration (Fig. 5c). These optimized molecules are computationally proposed candidates intended for prioritization and hypothesis generation, but their transfection efficiency has not yet been experimentally validated.

CONCLUSIONS

In this study, we demonstrate the feasibility and effectiveness of integrating cheminformatics descriptors, pretrained molecular embeddings, and generative modeling to advance the study of IAJDs. By systematically evaluating multiple feature extraction strategies and machine learning models, we established a comprehensive computational framework for predicting and optimizing IAJD activities. Specifically, fingerprint-based descriptors remained robust baselines when carefully parameterized. A count-based ECFP with an intermediate radius ($R=3$) and moderate bit length (256) achieved an optimal balance between representational richness and efficiency. Nonlinear ensemble models, particularly XGBoost, outperformed linear classifiers, highlighting their capability to capture the substructural and physicochemical nonlinearities intrinsic to dendritic architectures. Feature importance analysis identified several recurring motifs that were closely correlated with biological activity, thus providing mechanistic interpretability and rational guidance for molecular design.

Building upon these predictive insights, a machine-learning-driven molecular optimization workflow was developed to explore new IAJD candidates. The framework integrates calibrated XGBoost predictions with fragment-based single-step editing and multi-objective scoring, incorporating synthetic accessibility, physicochemical feasibility, and structural diversity constraints. The pipeline successfully generated chemically valid, diverse, and novel IAJD derivatives, several of which exhibited high predicted transfection probabilities and favorable property profiles. This demonstrates the potential of computational design to guide molecular innovation in data-limited regimes. Beyond traditional fingerprints, transformer-based ChemBERTa embeddings provide superior performance by directly learning the SMILES-level semantics. After dimensionality reduction, these embeddings enabled robust classification and revealed transferable representations of complex nanocarrier systems. Together, these results establish a unified “analyze-predict-design” paradigm, in which molecular representation learning and generative modeling jointly accelerate the discovery of next-generation mRNA delivery materials.

Despite these promising results, the dataset size (231 labeled molecules) limits the model generalization and exploration of the full IAJD chemical space. Structural repetitiveness may also introduce representational redundancy, but the generated candidates remain to be experimentally validated. Future studies should expand the dataset, explore graph-based or multimodal molecular representations, and incorporate multi-objective optimization considering stability, toxicity, and tissue specificity. Importantly, these results reaffirm the exceptional potential of the IAJDs. Their intrinsic

amphiphilicity, dendritic topology, and tunable ionization collectively provide unique advantages for mRNA delivery, combining efficient encapsulation, structural rigidity, and controlled release. The observed link between molecular architecture and function underscores the rational design principles underlying IAJDs and highlights their dual role as both promising delivery vehicles and ideal model systems for data-driven material discovery.

In conclusion, this study established a machine learning framework tailored to the analysis of IAJDs under a limited-data regime. By systematically comparing the molecular representations, we found that conventional ECFP fingerprints are suboptimal for structurally complex macromolecules, whereas count-based descriptors yield more informative encodings. Furthermore, ChemBERTa-derived embeddings demonstrated superior predictive performance by capturing high-level structural and semantic information, whereas substructure importance analysis provided interpretable insights into key functional motifs. Collectively, these findings highlight the potential of machine learning to enhance both predictive accuracy and mechanistic understanding in the study of IAJDs, thereby laying the groundwork for the data-driven optimization of next-generation mRNA delivery systems.

Conflict of Interests

The authors declare no interest conflict.

Electronic Supplementary Information

Electronic supplementary information (ESI) is available free of charge in the online version of this article at <http://doi.org/10.1007/s10118-026-3563-3>.

Data Availability Statement

Data supporting the findings of this study are available from the corresponding author upon reasonable request.

ACKNOWLEDGMENTS

This work was financially supported by the National Key Research and Development Program of China (No. 2024YFF0508600), the National Natural Science Foundation of China (Nos. 22305081, 32571562 and 52572303), Leading Talents in Shanghai in 2018, Shanghai Sailing Program (No. 23YF1408600), 111 project (No. B14018), and Fundamental Research Funds for the Central Universities of China.

REFERENCES

- 1 Sahin, U.; Karikó, K.; Türeci, Ö. mRNA-based therapeutics — developing a new class of drugs. *Nat. Rev. Drug. Discov.* **2014**, *13*, 759–780.
- 2 Pardi, N.; Hogan, M. J.; Porter, F. W.; Weissman, D. mRNA vaccines—a new era in vaccinology. *Nat. Rev. Drug. Discov.* **2018**, *17*, 261–279.
- 3 Polack, F. P.; Thomas, S. J.; Kitchin, N.; Absalon, J.; Gurtman, A.;

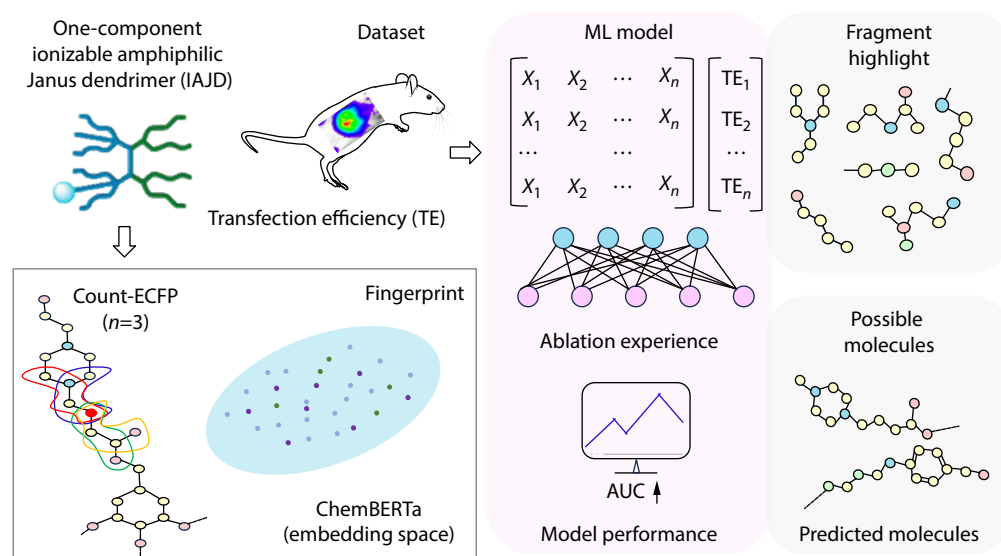
Graphical Abstract

Machine Learning-guided Prediction of Ionizable Amphiphilic Janus Dendrimers for mRNA Nanomedicine

Wan-Ting Cheng, Heng-Li Zheng, Peng-Yu Zhu, Ji-Na Hao, Da-Peng Zhang, and Yong-Sheng Li

East China University of Science and Technology

A tailored machine-learning (ML) framework was developed to investigate the relationship of molecular structures of ionizable amphiphilic Janus dendrimers (IAJDs) and their mRNA delivery effects. Count-based fingerprints and ChemBERTa embeddings improve accuracy and interpretability, revealing key motifs and enabling ML-guided design and optimization of IAJDs for mRNA delivery.



Chinese J. Polym. Sci. 2026, 44, 1154–1164

<https://doi.org/10.1007/s10118-026-3563-3>

- Lockhart, S.; Perez, J. L.; Pérez, Marc, G.; Moreira, E. D.; Zerbini, C.; Bailey, R.; Swanson, K. A.; Roychoudhury, S.; Koury, K.; Li, P.; Kalina, W. V.; Cooper, D.; Frenck, R. W., Jr.; Hammitt, L. L.; Türeci, Ö.; Nell, H.; Schaefer, A.; Ünal, S.; Tresnan, D. B.; Mather, S.; Dormitzer, P. R.; Şahin, U.; Jansen, K. U.; Gruber, W. C. Safety and efficacy of the BNT162b2 mRNA COVID-19 vaccine. *N. Engl. J. Med.* **2020**, 383, 2603–2615.
- 4 Baden, L. R.; El Sahly, H. M.; Essink, B.; Kotloff, K.; Frey, S.; Novak, R.; Diemert, D.; Spector, S. A.; Rouphael, N.; Creech, C. B.; McGettigan, J.; Khetan, S.; Segall, N.; Solis, J.; Brosz, A.; Fierro, C.; Schwartz, H.; Neuzil, K.; Corey, L.; Gilbert, P.; Janes, H.; Follmann, D.; Marovich, M.; Mascola, J.; Polakowski, L.; Ledgerwood, J.; Graham, B. S.; Bennett, H.; Pajon, R.; Knightly, C.; Leav, B.; Deng, W.; Zhou, H.; Han, S.; Ivarsson, M.; Miller, J.; Zaks, T. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *N. Engl. J. Med.* **2021**, 384, 403–416.
 - 5 Hou, Xucheng; Zaks, Tal; Langer, Robert; Dong, Yizhou. Lipid nanoparticles for mRNA delivery. *Nat. Rev. Mater.* **2021**, 6, 1078–1094.
 - 6 Zhang, D.; Atochina-Vasserman, E. N.; Maurya, D. S.; Huang, N.; Xiao, Q.; Ona, N.; Liu, M.; Shahnawaz, H.; Ni, H. P.; Kim, K.; Billingsley, M. M.; Pochan, D. J.; Mitchell, M. J.; Weissman, D.; Percec, V. One-component multifunctional sequence-defined ionizable amphiphilic Janus dendrimer delivery systems for mRNA. *J. Am. Chem. Soc.* **2021**, 143, 12315–12327.
 - 7 Zhu, P.; Li, Y.; Zhang, D. One-component ionizable amphiphilic Janus dendrimers for targeted mRNA delivery. *Angew. Chem. Int. Ed.* **2025**, 64, e202505304.
 - 8 Semple, S. C.; Akinc, A.; Chen, J.; Sandhu, A. P.; Mui, B. L.; Cho, C. K.; Sah, D. W. Y.; Stebbing, D.; Crosley, E. J.; Yaworski, Ed; Hafez, I. M.; Dorkin, J. R.; Qin, J.; Lam, K.; Rajeev, K. G.; Wong, K. F.; Jeffs, L. B.; Nechev, L.; Eisenhardt, M. L.; Jayaraman, M.; Kazem, M.; Maier, M. A.; Srinivasulu, M.; Weinstein, M. J.; Chen, Q.; Alvarez, R.; Barros, S. A.; De, S.; Klimuk, S. K.; Borland, T.; Kosovrasti, V.; Cantley, W. L.; Tam, Y. K.; Manoharan, M.; Ciufolini, M. A.; Tracy, M. A.; de Fougères, A.; MacLachlan, I.; Cullis, P. R.; Madden, T. D.; Hope, M. J. Rational design of cationic lipids for siRNA delivery. *Nat. Biotechnol.* **2010**, 28, 172–176.
 - 9 Kulkarni, J. A.; Witzigmann, D.; Chen, S.; Cullis, P. R.; van der Meel, R. Lipid nanoparticle technology for clinical translation of siRNA therapeutics. *Acc. Chem. Res.* **2019**, 52, 2435–2444.
 - 10 Arshad, M.; Atochina-Vasserman, E. N.; Maurya, D. S.; Lu, J.; Ona, N.; Vasserman, J. A.; McWilliams, B. C.; Ni, H.; Berkihsier, S.; Park, W.; Weissman, D.; Percec, V. Harnessing the electron-withdrawing inductive effect of one-component ionizable amphiphilic Janus dendrimers unveils cation-π interactions and their important

- roles to targeted mRNA delivery. *J. Am. Chem. Soc.* **2025**, *147*, 21347–21356.
- 11 Feng, W. X.; Zhang, S. Q.; Xu, Y. Y.; Ye, X. F.; Xu, X. Y.; Wang L. Q.; Lin, J. P.; Cai C. H.; Du, L. Machine-learning-assisted materials genome approach for designing high-performance thermosetting polyimides. *Chinese J. Polym. Sci.* **2025**, *43*, 1718–1729.
 - 12 Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.
 - 13 Cheng, Q.; Wei, T.; Farbiak, L.; Johnson, L. T.; Dilliard, S. A.; Siegwart, D. J. Selective organ targeting (SORT) nanoparticles for tissue-specific mRNA delivery and CRISPR–Cas gene editing. *Nat. Nanotechnol.* **2020**, *15*, 313–320.
 - 14 Witten, J.; Raji, I.; Manan, R. S.; Beyer, E.; Bartlett, S.; Tang, Y.; Ebadi, M.; Lei, J.; Nguyen, D.; Oladimeji, F.; Jiang, A. Y.; MacDonald, E.; Hu, Y.; Mughal, H.; Self, A.; Collins, E.; Yan, Z.; Engelhardt, J. F.; Langer, R.; Anderson, D. G. Artificial intelligence-guided design of lipid nanoparticles for pulmonary gene therapy. *Nat. Biotechnol.* **2025**, *43*, 1790–1799.
 - 15 Li, B.; Manan, R. S.; Liang, S.; Gordon, A.; Jiang, A.; Varley, A.; Gao, G.; Langer, R.; Xue, W.; Anderson, D. Combinatorial design of nanoparticles for pulmonary mRNA delivery and genome editing. *Nat. Biotechnol.* **2023**, *41*, 1410–1415.
 - 16 Kim J. Optimization of IAJD design for DNP mRNA delivery. [accessed 2025, August 1] <https://isef.net/project/bchm013-optimization-of-iajd-design-for-dnp-mrna-delivery>.
 - 17 Zhang, D.; Atochina-Vasserman, E. N.; Maurya, D. S.; Liu, M.; Xiao, Q.; Lu, J.; Lauri, G.; Ona, N.; Reagan, E.; Ni, H.; Weissman, D.; Percec, V. Targeted delivery of mRNA with one-component ionizable amphiphilic Janus dendrimers. *J. Am. Chem. Soc.* **2021**, *143*, 17975–17982.
 - 18 Zhang, D.; Atochina-Vasserman, E. N.; Lu, J.; Maurya, D.; Xiao, Q.; Liu, M.; Adamson, J.; Ona, N.; Reagan, E. K.; Ni, H. P.; Weissman, D.; Percec, V. The unexpected importance of the primary structure of the hydrophobic part of one-component ionizable amphiphilic Janus dendrimers in targeted mRNA delivery activity. *J. Am. Chem. Soc.* **2022**, *144*, 4746–4753.
 - 19 Lu, J.; Atochina-Vasserman, E. N.; Maurya, D. S.; Sahoo, D.; Ona, N.; Reagan, E. K.; Ni, H. P.; Weissman, D.; Percec, V. Targeted and equally distributed delivery of mRNA to organs with pentaerythritol-based one-component ionizable amphiphilic Janus dendrimers. *J. Am. Chem. Soc.* **2023**, *145*, 18760–18766.
 - 20 Sahoo, D.; Atochina-Vasserman, E. N.; Maurya, D. S.; Arshad, M.; Chenna, S. S.; Ona, N.; Vasserman, J. A.; Ni, H. P.; Weissman, D.; Percec, V. The constitutional isomerism of one-component ionizable amphiphilic Janus dendrimers orchestrates the total and targeted activities of mRNA delivery. *J. Am. Chem. Soc.* **2024**, *146*, 3627–3634.
 - 21 Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
 - 22 Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
 - 23 Cover, T. M.; Hart, P. E. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27.
 - 24 Morgan, H. L. The generation of a unique machine description for chemical structures—A technique developed at chemical abstracts service. *J. Chem. Doc.* **1965**, *5*, 107–113.
 - 25 Bajusz, D.; Rácz, A.; Héberger, K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations. *J. Cheminform.* **2015**, *7*, 20.
 - 26 Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. *arXiv* **2020**, arXiv:2010.09885.
 - 27 Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
 - 28 Hosmer, D. W.; Lemeshow, S.; Sturdivant, R. X. Applied logistic regression. In *Applied Logistic Regression*, 3rd ed.; Wiley: Hoboken, NJ, **2013**; pp 89–151.
 - 29 Chen, T.; Guestrin, C. XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16); ACM: New York, **2016**, pp. 785–794.
 - 30 Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.
 - 31 Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the art of compiling and using “drug-like” chemical fragment spaces. *ChemMedChem* **2008**, *3*, 1503–1507.
 - 32 Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*, 8.
 - 33 Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90–98.
 - 34 Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26.