Dreaming Up Novel Quantum Dyes using Inverse Machine Learning in MatFlow

Mahib H Ornob
Department of Computer Science
University of Idaho, USA
orno3564@vandals.uidaho.edu

Lan Li
Micron School of Materials Science and Engineering
Boise State University, USA
lanli@boisestate.edu

Hasan M Jamil ✉
Department of Computer Science
University of Idaho, USA
jamil@uidaho.edu

***Abstract*—Discovering novel molecules with targeted properties remains a formidable challenge in materials science, often likened to finding a needle in a haystack. Traditional experimental approaches are slow, costly, and inefficient. In this study, we present an inverse design framework based on a molecular graph conditional variational autoencoder (CVAE) that enables the generation of new molecules with user-specified optical properties, particularly molar extinction coefficient (ε). Our model encodes molecular graphs, derived from SMILES strings, into a structured latent space, and then decodes them into valid molecular structures conditioned on a target ε value. Trained on a curated dataset of known molecules with corresponding extinction coefficients, the CVAE learns to generate chemically valid structures, as verified by RDKit. Subsequent Density Functional Theory (DFT) simulations confirm that many of the generated molecules exhibit the electronic structures similar to those molecules with desired ε values. We have also verified the ε values of the generated molecules using a graph neural network (GNN) and the synthesizability of those molecules using an open-source module named ASKCOS. This approach demonstrates the potential of CVAEs to accelerate molecular discovery by enabling user-guided, property-driven molecule generation – offering a scalable, data-driven alternative to traditional trial-and-error synthesis.**

***Index Terms*—Inverse machine learning, variational autoencoder (VAE), DFT, graph neural network (GNN), synthesizability, quantum dye, extinction coefficient, materials design.**

I. Introduction

One of the central challenges in chemistry and materials science is the synthesis of novel molecules with targeted functional properties [5, 45]. Success in this area would unlock transformative advances in domains ranging from drug discovery and clean energy to next-generation electronics [10, 19]. Among such molecules, quantum dyes [40] stand out for their foundational role in technologies like organic photovoltaics (OPVs)[14], organic light-emitting diodes (OLEDs)[15], bio-imaging [52], and chemical sensing [47], where their performance directly affects efficiency and functionality [33, 51].

A key figure of merit for dye molecules is the molar extinction coefficient, which measures how effectively a molecule absorbs light at a given wavelength [28]. High extinction coefficients are crucial for optimal light harvesting, fluorescence, and optical sensitivity. However, designing dyes that simultaneously satisfy multiple constraints – such as solubility, stability, and spectral alignment – poses a complex multi-objective optimization problem [46].

The vastness of chemical space, estimated to exceed 10^{60} synthetically accessible small organic molecules [34], makes exhaustive experimental screening impractical [49]. While combinatorial chemistry and high-throughput screening (HTS) have increased efficiency, they remain constrained by established motifs and limited chemical intuition [8]. Moreover, physical synthesis remain costly and time-consuming.

Density Functional Theory (DFT) offers a computational alternative for accurate property prediction [26], but its high computational cost limits large-scale molecular exploration [7]. To address these barriers, data-driven methods leveraging advances in artificial intelligence and machine learning have emerged as promising alternatives [8]. By learning from existing molecular datasets, ML models can capture complex structure-property relationships and estimate properties of novel molecules with high accuracy. More significantly, the rise of deep generative models has ushered in a shift toward inverse design [13, 42], where new molecular structures are generated conditionally to meet predefined property profiles [17, 19, 25] – offering a powerful, scalable approach to chemical discovery.

II. Related Work

In molecular design, two distinct paradigms exist: direct and inverse design. Direct design involves constructing a molecule from its atomic structure and composition, then computing its properties post hoc using theoretical methods such as quantum chemistry [53]. This process is inherently nonlinear and computationally intensive, as properties like energy eigenvalues and wavefunctions are inferred only after a candidate structure is specified [27, 36]. In contrast, inverse design reverses the problem – starting from desired properties and working backward to identify structures in chemical space that fulfill them [42]. This approach reframes material discovery as an optimization problem over structure-property relationships [29] and offers a promising path to accelerate the design of functional molecules.

To explore the vastness of chemical space using inverse design, researchers have employed three primary strategies: high-throughput virtual screening (HTVS), global optimization algorithms, and generative models [9, 43]. HTVS involves computationally evaluating large material datasets to find candidates with desirable properties. For example, Jang et al.[22]

used DFT-based HTVS to predict inorganic materials, while Afzal et al.[1] identified high-refractive-index polyimides for optoelectronics. However, HTVS is often constrained by limited data coverage and cannot easily extrapolate to novel compounds outside known chemical databases.

To overcome this, global optimization methods such as Bayesian optimization (BO) [18], genetic algorithms (GA) [12], particle swarm optimization (PSO) [50], and simulated annealing have been applied to explore structure-property landscapes more flexibly [44]. Harper et al.[20] used BO to identify topologies for multifunctional optical materials, while Khadilkar et al.[23] coupled PSO with self-consistent field theory to model polymer morphology. Lee et al. [30] proposed a two-phase GA approach that encodes molecules with embedded strings and graphs to guide structural mutation and crossover. These methods are versatile and capable of generating viable candidate structures even when limited prior knowledge exists, though they may lack the expressiveness and scalability of data-driven techniques.

Generative models (GMs), especially deep generative architectures, have emerged as powerful tools for inverse molecular design [42]. These models embed high-dimensional chemical structures into low-dimensional latent spaces, enabling the creation of novel, property-optimized molecules. For instance, Kim et al.[24] developed a hybrid encoder-decoder model using deep neural networks (DNNs) and recurrent neural networks (RNNs) to reconstruct molecular structures with desired features. Popova et al.[35] employed deep reinforcement learning to generate new compounds by optimizing over property-based reward signals. Similarly, Geng et al. [16] applied generative adversarial networks (GANs) for the inverse design of meta-surfaces, using pretrained simulators to predict optical responses.

Variational autoencoders (VAEs), another widely used generative model, improve generalization by encoding molecular structures into probabilistic latent variables. Ma et al. [32] demonstrated this approach for meta-material design by modeling the joint distribution of latent variables, structural patterns, and spectral outputs. These architectures allow property conditioning during molecule generation, making them particularly well-suited for tasks such as designing dyes with specific optical properties.

Together, these approaches reflect a fundamental shift in materials discovery – from the deterministic construction of molecules toward intelligent, data-driven exploration of chemical space guided by target functionalities.

III. Methodology

The inverse molecular generation process shown in Figure 1 aims to produce molecules with a specific target extinction coefficient (ε). The first step is a *Molecule Generation Module*, which transforms molecular smiles obtained from the dye design dataset into molecular graph matrices. These matrices are fed into a variational autoencoder (VAE) together with the target ε (as a condition vector). This input is compressed into a latent space representation by the VAE's encoder. This latent space and the target ε are then used by the VAE's decoder to create new molecular graph matrices, which are subsequently transformed back into molecular structures.

These produced compounds are subsequently filtered by the *External Validation Module*. Several processes are involved in this process: 2D and 3D visualization, DFT calculations, RDKit analysis for molecular feasibility, SCScore for synthesizability evaluation, and a GNN prediction to confirm whether the produced molecules show the expected ε. The final generated molecule set is made up of molecules that pass each of these validation stages.

A. Data

In our dye design dataset, which comprises 8,816 molecules and 307 molecular features, we have completed our preliminary analysis and found this dataset is complete with no missing values. Our main target feature, Epsilon (ε), shows a wide dynamic range from a minimum of 9 to a maximum value of 5.8 million. In Figure 2(a) we can observe a right skewness of the data distribution where the mean of 55,815 is significantly higher than the median of 31,000 and the standard deviation of 157,138. This confirms that the majority of the molecules have lower Epsilon values. Another property, [Fig 2(b)] total atom count, ranges from 2 to 387 atoms per molecule, with a mean of 62 atoms and a median of 52 atoms. This distribution also shows right skewness, even though it is less dramatic than the Epsilon's. Most molecules in the dataset have a lower number of atoms, ranging from 38 to 74 atoms.

Another notable thing to observe in our dataset is that even though there are 8,816 total entries, there are only 4,299 unique smile strings present in the dataset. This suggests that multiple smile strings can have different Epsilon values, which may have occurred due to different conformational states or variations in experimental conditions or data aggregation from multiple sources.

The scatter plot [Fig 2(c)] describing the relation between Epsilon and Total Atom Count shows a positive monotonic trend as the value of the Total Atom Count increases; the Epsilon value also increases. This aligns with the Spearman correlation of 0.478 that shows a moderate positive relationship, and the weaker Pearson correlation value of 0.259 suggests that the trend is not strictly linear. Relation between Epsilon and SMILES String length shows a weak positive trend [Fig 2(d)], meaning longer SMILE strings show slightly higher Epsilon values, but this trend is less distinct than the total atom count one.

We have found twelve different atom kinds (B, C, N, O, F, Si, P, S, Cl, Br, Sn, and I) and four different bond types (single, double, triple, and aromatic) in our datasets. These compounds also have ClogP values, which are determined via RDKit [39] calculating methods. P is the ratio of the concentrations of a solute in two solvents [11]. The size and molecular weight of molecules are linked to CMR values, a crucial parameter for determining the steric factor [3]. Molecules appropriate for graph formation (such as SMILES lacking "+," "-," and ".") were extracted from the database for this investigation.

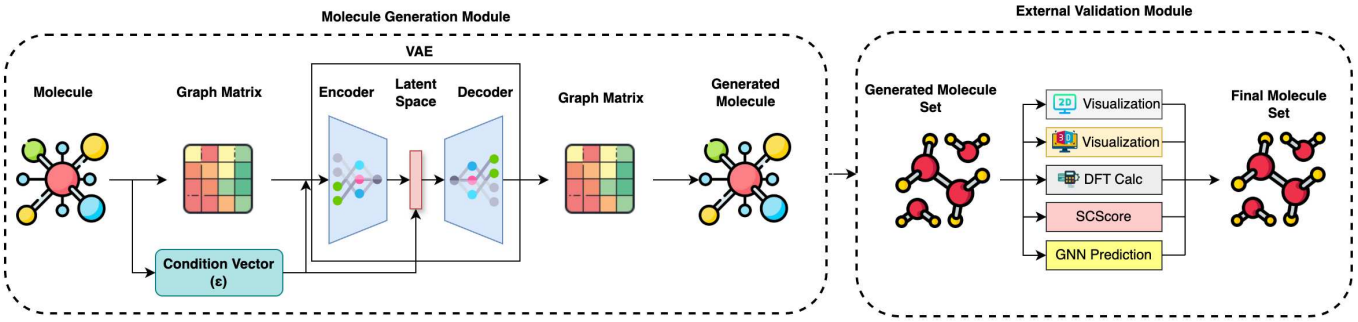


Fig. 1: Working Procedure of Inverse Molecular Generation.

A 9:1 ratio was used to split the data set into training and test sets, and the training procedure was evaluated. In order to achieve the greatest degree of closeness between the input and output initial graph matrices, the auto-encoder has been trained to minimize loss. Optimizations for the user-defined molar extinction coefficient (ε) value have been carried out if the model can be adequately trained when each property has a suitably large data distribution in the data set. Using StandardScaler, the Epsilon values were normalized (zero mean, unit variance). Both the training and test sets' Epsilon values were transformed after the scaler was fitted solely to the training set. For later usage, the training set's mean and scale factor were stored.

Epsilon	SMILES	Total Atom Count
3801.89	c1ccc2ccccc2c1	18
5370.31	C[Si](C)(C)c1cccc2ccccc12	30
5623.41	C[SiH](C)c1cccc2ccccc12	27

TABLE I: Sample Training Data

B. Molecular Graph

The graph representation of molecules uses annotation and adjustment matrices to present atoms as nodes and bonds as edges. Each row is represented as the one-hot encoding of atoms in the annotation matrix ($N \times X$, where N is the number of atoms and X is the number of types of atoms), and the adjacency matrix ($N \times N$) shows how each row and column corresponding to the atoms are binding. A complete molecular graph was created by reconstructing the original graph matrix of the present models into the adjacency and annotation matrices. The initial graph matrix has the structure $\{M, [1 + T + (M \cdot B)]\}$, where M is the maximum number of atoms (largest graph size), T is the number of atom types, and B is the number of bond types.

Next, for every atom position up to M, we create an atom feature matrix. We generate a feature vector of size T if the point matches an actual atom in the input molecule. Set all other elements to zero and the element that corresponds to the atom's type index to one. Make a feature vector in which the element corresponding to the specified padding atom type (index 0) is one and all other elements are zero if the position exceeds the actual atom count (i.e., padding). These vectors

Algorithm 1 Molecular Graph Construction from SMILES

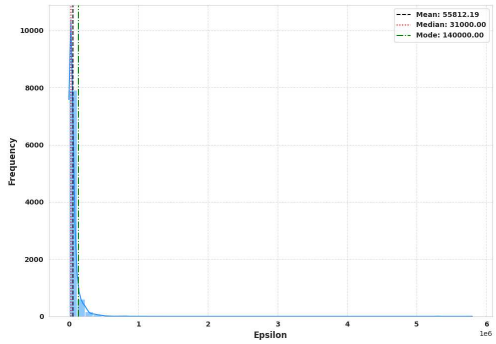
Input: SMILES string S
Output: Flattened molecular graph feature vector f

```
Parse S to extract atom list A and bond list B
N ← |A| {Number of atoms}
Define N_max (max atoms), A_t (atom types), B_t (bond types)
Initialize annotation matrix X ∈ {0,1}^(N_max × A_t) and adjacency tensor E ∈ {0,1}^(N_max × N_max × B_t) with zeros
for i = 1 to N do
    t ← atom type of A[i]
    X[i,:] ← one-hot(t)
end for
for each bond (i,j) in B do
    b ← bond type of (i,j)
    E[i,j,:] ← one-hot(b);
    E[j,i,:] ← E[i,j,:] {Undirected graph}
end for
f ← concatenate(flatten(X), flatten(E))
return f
```

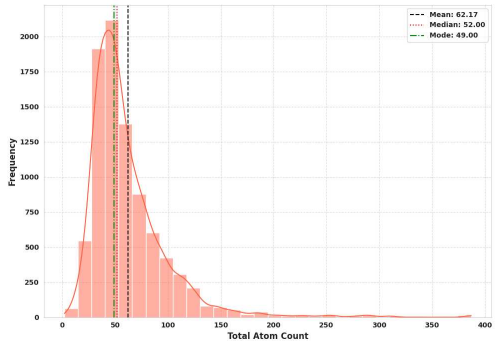
are then put together to create an atom feature matrix with the shape $[M, T]$.

Then, using the bond type encoding dictionary, the integer index corresponding to the bond type between each pair of possible atom positions i and j (up to M) is found, allocating the 'no bond' index where necessary. Following that, a temporary matrix of shape $[M, B]$ is produced for every possible source atom position i. The relevant element in the j-th row of this temporary matrix is set to one, while all other elements in that row are set to zero, for each potential target atom location j. This is done using the integer bond type index that was previously established for the pair (i, j). This temporary matrix is kept and represents all bonds that start at point i.

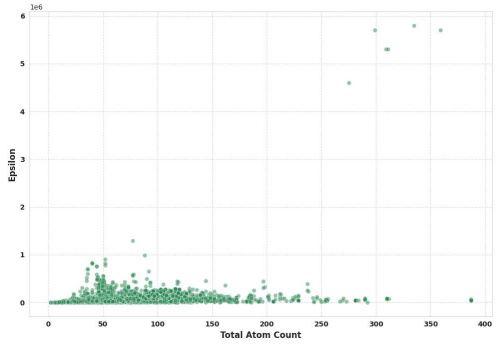
The final adjacency feature matrix, which has dimensions $[M, M \times B]$, is created by concatenating all stored temporary matrices horizontally (along the second dimension) after processing all source positions i. The final concatenated 2D tensor is then produced by horizontally combining the length



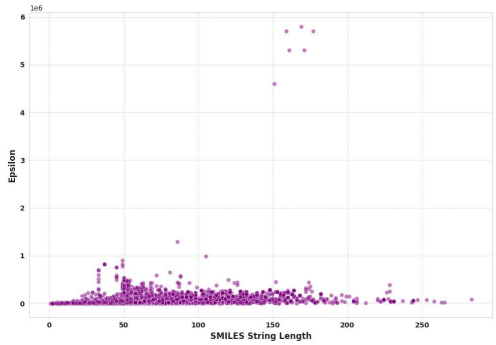
(a) Epsilon Distribution



(b) Total Atom Count Distribution



(c) Total Atom Count Vs Epsilon



(d) SMILES String Length Vs Epsilon

Fig. 2: Distribution of Dye Design Dataset

Algorithm 2 CVAE Training and Molecule Generation

Input: Dataset $\{(X^{(i)}, \varepsilon^{(i)})\}_{i=1}^{N}$, target condition ε^*
Output: Trained CVAE model (ϕ, θ) and predicted SMILES string S

Normalize graph matrices $X^{(i)}$
Initialize encoder $q_\phi(z|X, \varepsilon)$ and decoder $p_\theta(X|z, \varepsilon)$ networks
for each epoch $= 1$ to $T_{\max}$ **do**
  for each batch (X_b, ε_b) **do**
    Compute latent mean μ and log-variance σ from encoder
    Sample $z \sim \mathcal{N}(\mu, \sigma)$ using reparameterization trick
    Reconstruct $\hat{X}_b \leftarrow p_\theta(z, \varepsilon_b)$
    Calculate reconstruction loss L_{rec} between $\hat{X}_b$ and X_b
    Calculate KL divergence L_{KL} between $q_\phi(z|X_b, \varepsilon_b)$ and $\mathcal{N}(0, I)$
    Compute total loss $L \leftarrow L_{\text{rec}} + L_{\text{KL}}$
    Update model parameters ϕ, θ by minimizing L
  end for
end for
Sample $z \sim \mathcal{N}(0, I)$ from latent prior
Decode $\hat{X} \leftarrow p_\theta(z, \varepsilon^*)$
Post-process $\hat{X}$ to obtain valid molecular graph
Convert molecular graph $\hat{X}$ to SMILES string S
return Trained model (ϕ, θ) and predicted S

indicator vector, the atom feature matrix, and the adjacency feature matrix along the second dimension. The linear layers of the CVAE's encoder then processed this 2D tensor after it had been flattened into a 1D vector.

C. Conditional Variational Autoencoder

A CVAE is the main component of the generative process. An encoder and a decoder make up VAEs. The encoder converts a distribution in a lower-dimensional latent space z from the input data X (molecular graph matrix). A latent vector z taken from this distribution is used by the decoder $p(X|z)$ to recreate the input data. The Evidence Lower Bound (ELBO) is maximized when training VAEs. The model is dependent on the normalized molar extinction coefficient $c = \varepsilon_{\text{norm}}$ in order to allow guided generation. The encoder stays $q(z|X)$, while the decoder only incorporates this condition, becoming $p(X|z, c)$. A target condition c and a latent vector z sampled from the prior distribution $p(z)$ (usually $\mathcal{N}(0, I)$) are supplied to the decoder throughout the creation process.

A Multi-Layer Perceptron (MLP) network serves as the encoder. Its input is the flattened graph matrix X. ReLU activation is used in hidden layers. The latent mean μ and log-variance $\log \sigma^2$ are generated by the output layers. It is made up of layers that are fully connected ($x_{\text{dim}} \rightarrow 512 \rightarrow 256 \rightarrow z_{\text{dim}}$).

Another MLP network is used as a decoder. Its input is the concatenation of the condition vector c (normalized ε, dimension 1) and the latent vector z. ReLU activation is

ε	Dye#	SMILES	Total Energy	HOMO LUMO Gap	Predicted ε	SCScore
150,000 $M^{-1}cm^{-1}$	D_1	CC(C)c1ccc(N2C=CC=C3C=CC=C3C=C2)cc1	-780.42819	0.12075	104,347.89	3.2
	D_2	CCOC(=O)C1c2ccccc2C(=O)N1c1ccc2ooc2c1	-1072.36189	0.17016	204,250.05	3.0
	D_3	C=S(N)(=O)c1ccc(N2N=C(c3ccc(O)cc3)CC2c2ccc(Cl)cc2)cc1	-1998.97392	0.07048	275,141.50	4.1
200,000 $M^{-1}cm^{-1}$	D_4	C=S(N)(=O)c1ccc2c(c1)-c1ccccc1C1=C3C=CC(O)=CC=CC3=NC12	-1526.12519	0.04978	162,363.89	4.3
	D_5	C=S(N)(=O)c1ccc(N2N=C(c3ccc(O)cc3)CC2c2ccc(F)cc2)cc1	-1642.17399	0.07947	310,873.72	4.2
	D_6	CC1(C)CC(C=Cc2ccc(-n3cccc3)cc2)=CC(=C(C#N)C#N)C1	-1040.24235	0.12679	178,487.16	3.3

TABLE II: A sample of six dyes in two categories of extinction coefficients from a total of 75 unique predicted dyes.

used in hidden layers. A Sigmoid activation is used in the last layer to produce the reconstructed flattened graph matrix probabilities. It is made up of layers that are also fully connected, such as $(z_{\text{dim}} + 1) \rightarrow 256 \rightarrow 512 \rightarrow x_{\text{dim}}$.

The model is trained using the combined VAE loss (ELBO), given by:

$$L_{\text{total}} = L_{\text{BCE}} + D_{\text{KL}}(q(z|X)\|p(z))$$

where D_{KL} is the Kullback-Leibler divergence between the encoder's distribution $q(z|X)$ and the standard normal prior $p(z)$, and L_{BCE} is the Binary Cross-Entropy reconstruction loss between the input matrix X and the decoder's output X_{recon}. The encoder uses the μ and $\log \sigma^2$ output to compute the KLD term.

IV. Results

For our experiments, we have used Python version 3.9.6. The packages used include, but are not limited to, PyTorch (2.6.0), NumPy (1.26.4), matplotlib (3.9.4), pandas (2.2.3), RDKit (2024.9.6), scikit-learn (1.6.1), PubChemPy (1.0.4), PySCF (2.8.0), and Pillow (11.1.0). Training was performed on an Apple M3 Max device with 36 GB of memory.

A. Experimental Evaluation

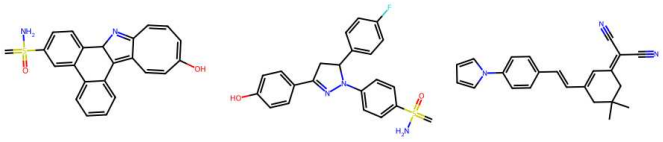
The CVAE model was trained on our dye design dataset using the Adam optimizer with an initial learning rate of $5\text{e}-5$. Training ran for 1,200 epochs with a batch size of 64. The latent space dimension was set to 256. Molecules were generated by sampling latent vectors z from the standard normal prior distribution $p(z) = \mathcal{N}(0, I)$ and selecting a target extinction coefficient $\varepsilon_{\text{target}}$. The target value was normalized using the mean and scale factor derived from the training data's Epsilon distribution:

$$c = \frac{\varepsilon_{\text{target}} - \varepsilon_{\text{mean}}}{\varepsilon_{\text{scale}}}$$

The output graph matrices are obtained by feeding pairs of (z, c) into the trained CVAE decoder, and the matrices are subsequently converted back to SMILES.

(a) D_1 (b) D_2 (c) D_3

Fig. 3: 2D structures of dyes D_1 through D_3 in Table II.



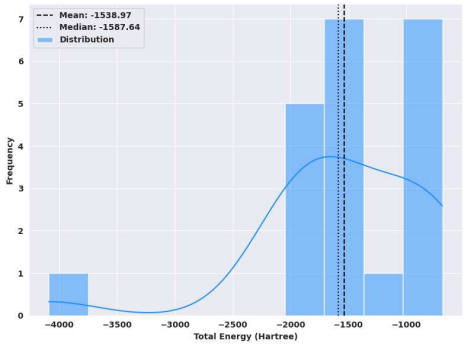
(a) D_4 (b) D_5 (c) D_6

Fig. 4: 2D structures of dyes D_4 through D_6 in Table II.

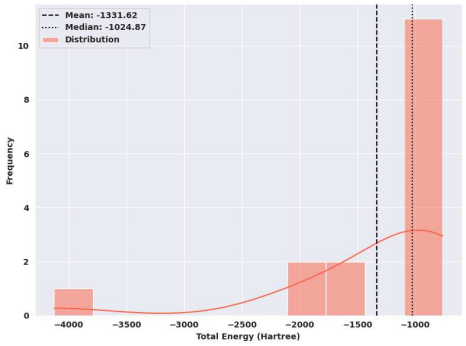
For each of the target generations (ε: 150,000 $M^{-1}cm^{-1}$, ε: 200,000 $M^{-1}cm^{-1}$), we initially created 5,000 molecules for each run. These molecules were then evaluated by checking the number of molecules that satisfy the target Epsilon values. We discovered 21, 8, and 7 valid molecules for various sets of hyperparameters, such as learning rate and maximum molecule size ($5\text{e}-5$, $4\text{e}-5$, and $3\text{e}-5$ and 60, 80, and 100, respectively), for the target value ε of 150,000 $M^{-1}cm^{-1}$. For the target value ε of 200,000 $M^{-1}cm^{-1}$, we discovered 22, 12, and 38 valid molecules for comparable sets of hyperparameters that meet the condition. We also checked the IUPAC names of these predicted molecules to assess if they already exist in the material database. We found that for $\varepsilon = 150,000$ $M^{-1}cm^{-1}$, eleven molecules are already present in the database, and for $\varepsilon = 200,000$ $M^{-1}cm^{-1}$, thirteen molecules are already present. We also performed DFT calculations on these two sets of predicted molecules. The data distribution of the calculated DFT values is given in Figure 5.

We measured the produced molecules' total energy and HOMO-LUMO gap in DFT calculations [6]. In this case, the total energy value is the molecule's total electronic energy as determined by DFT demonstrated in Table II. It is an essential indicator of the stability of the molecule in the gas phase at 0 Kelvin. A lower (more negative) energy typically denotes a more stable molecule. The HOMO-LUMO gap [41] is the energy difference between the lowest unoccupied molecular orbital (LUMO) and the highest occupied molecular orbital (HOMO). This gap is an important indicator of a molecule's kinetic stability [2], where a larger gap typically indicates greater kinetic stability; electronic excitations [31], where it is associated with the energy needed for the lowest electronic excitation (such as light absorption); and chemical reactivity [4], where a smaller gap typically implies higher reactivity because less energy is needed to excite an electron.

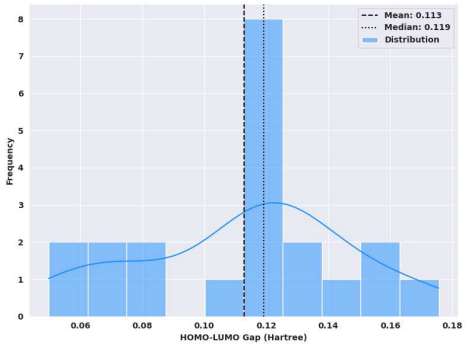
Both datasets (ε: 150,000 $M^{-1}cm^{-1}$, ε: 200,000 $M^{-1}cm^{-1}$) show very similar average values for both total energy and HOMO-LUMO gap. The shapes of their respective



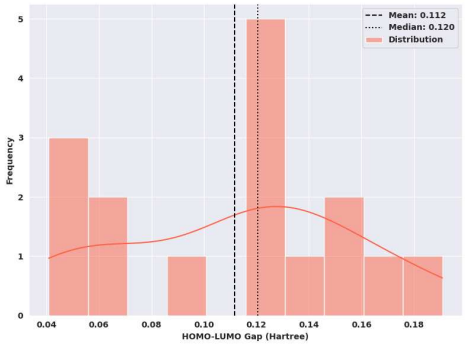
(a) Total Energy Distribution (ε: 200,000)



(b) Total Energy Distribution (ε: 150,000)



(c) HOMO-LUMO Gap Distribution (ε: 200,000)



(d) HOMO-LUMO Gap Distribution (ε: 150,000)

Fig. 5: Distribution of Total Energy and HOMO-LUMO Gap for $\varepsilon = 200,000\ M^{-1}cm^{-1}$ and $\varepsilon = 150,000\ M^{-1}cm^{-1}$.

distributions are also comparable, which suggests that the underlying generation process produced sets of molecules with broadly similar energetic and electronic profiles in both cases. A notable observation across both files is the wide energy distribution, which means the generation process is capable of producing diverse structures, from potentially smaller/less stable ones (higher energy) to larger/more stable ones (lower energy). Compared to the relatively tighter distribution around 0.11-0.12 Hartree for the HOMO-LUMO gap, this suggests the predicted molecules might share similar characteristics regarding chemical reactivity or suitability for applications sensitive to electronic excitation (like organic electronics or dyes). This strengthens our case of guided molecular generation for targeted Epsilon values.

A variety of structural modifications are displayed by the molecules in Figures 3 and 4, which are crucial for modulating optical characteristics. Diversity in ring systems (fused, acyclic, macrocyclic), conjugation, and the presence of heteroatoms for the ε value of 150,000 $M^{-1}cm^{-1}$ allows us to study how the model balances structural features to achieve the target Epsilon. Whereas the selection of molecules for the ε value of 200,000 $M^{-1}cm^{-1}$ highlights the model's ability to generate both planar and non-planar structures with a variety of functional groups (thioamides, sulfonamides), which enables a deeper understanding of the structure-property relationships at this higher Epsilon value.

B. Validation of Predicted Dyes

We have also used a graph neural network (GNN) model to externally validate the extinction coefficient of the predicted molecular SMILES from the CVAE model for a given ε value. This GNN acts as an independent prediction scheme where it takes a molecular SMILES as input, converts it to a graph representation, and predicts the ε value. The main purpose of using this model is to verify that the molecular SMILES predicted by the CVAE model indeed possess the optical property close to the target one.

The GNN model is based on graph convolutional network architecture, which processes molecular graphs where nodes represent the atoms and edges represent the bonds. It contains a total of 22 dimensions of atom features, which includes one-hot encoding for atom type (C, N, O, S, etc.). The features of an atom consist of 22 dimensions, which include one-hot encodings for atom type, degree, formal charge, hybridization, aromaticity, total hydrogen count, radical electron count, ring membership, and chirality. Bond features, involving six dimensions, capture bond type (single, double, triple, aromatic), conjugation, and ring membership; however, standard GCN convolutional layers mainly utilize node features and adjacency information. The network employs four GCN convolutional layers, with $256 \rightarrow 512 \rightarrow 1,024 \rightarrow 2,048$ hidden channels, increasing gradually. Additionally, $\mathrm{BatchNorm1D}$ is used for normalization, and ReLU is used as an activation function for each layer. Between these layers, a GNN dropout rate of 0.25 is also used to lessen overfitting. A global mean pool layer then combines all of the node embeddings to create

a single graph-level feature vector after graph convolutions. This feature vector is subsequently fed to two fully connected layers that have a higher dropout rate of 0.5, BatchNorm1D, and ReLU. The final output layer consists of a single neuron that uses the provided SMILES string to predict the ε value.

This model uses the same dataset, which includes SMILES strings and their corresponding ε values, as the CVAE procedure. The dataset was split into three sets: train (70%), test (15%), and validation (15%). The ε values were then scaled to a [0, 1] range using MinMaxScaller after being processed using NumPy log. This was done since the ε values found in the dataset had a wide range and a typical skewed distribution. Training was conducted using the Adam optimizer, which has a learning rate of 0.0005 and a weight decay of $1e-6$. With a batch size of 32 and an early stopping mechanism with a patience of 20 epochs, the model was trained for the maximum of 200 epochs. An R2 score of 0.8204 was obtained when the GNN model's performance was evaluated on the test set. This suggests that roughly 82% of the variance of the ε distribution was captured by the GNN model. The quantitative measures were further supported by a significant correlation between the expected and real ε values, as shown in Figure 6.

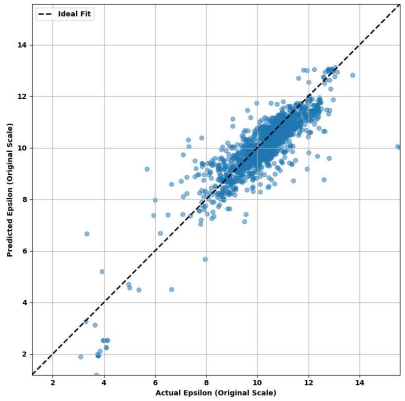


Fig. 6: Actual vs Predicted ε value distribution.

Even though the predicted molecules demonstrate significant structural diversity based on the targeted ε value, it is worth noting that not all of the predicted molecules can be mapped to a corresponding IUPAC chemical name available in the PubChem database. But this absence does not reduce the significance of our findings and our primary objective of utilizing the variational autoencoder's capacity to generate novel molecules with target properties. Rather, the absence of PubChem records might indicate that the model can explore uncharted chemical space. RDKit toolkit has been used to confirm the structure and chemical validity of the produced compounds, while DFT studies also shed light on their stability and electronic characteristics.

C. Synthesizability of Predicted Dyes

The synthesizability of AI-predicted dye molecules is a critical measure of their scientific validity and real-world applicability. While AI models such as MatFlow [21, 37, 38] can efficiently generate molecular structures with desirable optical properties, such as high molar extinction coefficients or tunable absorption spectra, these predictions must be chemically plausible and experimentally achievable to impact materials science meaningfully. Ensuring synthesizability bridges the gap between in silico discovery and laboratory implementation, enabling efficient validation, fabrication, and integration of novel dyes into quantum sensing, photonics, or biomedical imaging applications. Moreover, incorporating synthesizability constraints into AI design pipelines improves model robustness, reduces false positives, and accelerates the path from theoretical innovation to functional materials.

ASKCOS [48] is an open-source, AI-powered software suite for computer-aided synthesis planning (CASP), designed to help chemists evaluate and plan synthetic pathways for complex molecules. By integrating advanced machine learning models trained on large-scale reaction datasets, ASKCOS performs tasks such as retrosynthetic analysis, reaction condition recommendation, and reaction outcome prediction. This makes it particularly valuable for determining the synthetic feasibility of predicted molecules, like novel quantum dyes, by offering detailed, data-driven synthesis routes from commercially available starting materials.

The synthesizability of our predicted dye molecules was assessed using the ASKCOS platform, which assigns a Synthetic Complexity Score (SCScore) on a scale from 1 to 5. The SCScore values for the six dyes shown in Figures 3 and 4 are summarized in Table II. A lower SCScore implies higher synthesizability, and thus an ideal dye will have high ε and very low SCScore.

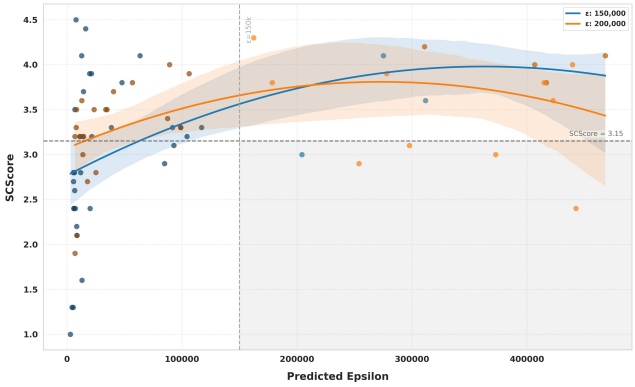
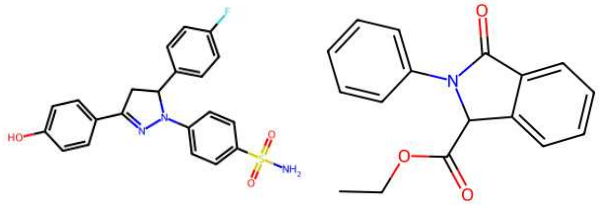


Fig. 7: Relationship between predicted SCScore and ε for the SMILES in target classes of $\varepsilon = 150,000\ M^{-1}cm^{-1}$ (blue line) and $\varepsilon = 200,000\ M^{-1}cm^{-1}$ (orange line). Likely best predicted SMILES are in the shaded quadrant.

The association between the created molecules' SCScore and predicted Epsilon by the GNN for two sets of ε: 150,000 and 200,000 has been demonstrated in Figure 7. The data points for both datasets are more densely clustered at lower Epsilon values and show notable vertical scatter, indicating variability in SCScore for similar Epsilon predictions. The general pattern of both datasets is that SCScore tends to

Dye#	SMILES	Target ε	Predicted ε	SCScore
D_7	**NS(=O)(=O)c1ccc(N2N=C(c3ccc(O)cc3)CC2c2ccc(F)cc2)cc1**	150,000	**468,121.00**	4.1
D_8	c1ccccccccccncccccccccc1	150,000	2,973.95	**1.0**
$D_9(D_2)$	**CCOC(=O)C1c2ccccc2C(=O)N1c1ccc2ooc2c1**	150,000	**204,250.25**	**3.0**
D_{10}	**NS(=O)(=O)c1ccc(N2N=C(c3ccc(O)cc3)CC2c2ccc(F)cc2)cc1**	200,000	**468,121.00**	4.1
D_{11}	N=C1C=CC=CC=CC=CC=CC=CC=c2ccccc2=CC=C1	200,000	7,013.13	**1.9**
D_{12}	**CCOC(=O)C1c2ccccc2C(=O)N1c1ccccc1**	200,000	**442,541.93**	**2.4**

TABLE III: Dyes with red bold fonts reflect highest predicted εs, and orange fonts mean lowest SCScores.



(a) D_7, D_{10}: $\varepsilon = 468,121.00$, SCScore = 4.1.

(b) D_{12}: $\varepsilon = 442,541.93$, SCScore = 2.4.

Fig. 8: 2D structures of molecules with the highest predicted ε and SCScores in each target extinction coefficient category.

increase with higher predictions ε before the trends diverge and plateau or even slightly decrease at higher ε values. The Pearson correlation ($r = 0.446$, $p = 0.00348$) for the ε: 150,000 data points shows a moderately positive linear relationship that is statistically significant. This suggests that although the general trend is curved, there is a definite underlying tendency for SCScore to increase with Epsilon. The ε: 200,000 data points, on the other hand, have a weaker positive linear correlation ($r = 0.330$) that is not statistically significant at the traditional 0.05 level ($p = 0.057$). This suggests that any linear trend is less clear and strong for this group, which is visually reflected in its flatter curve at higher ε values and then slightly declining curve at a wider confidence interval.

V. Discussion

Majority of the predicted dyes exhibited moderate complexity scores. Among the 41 and 34 unique SMILES predicted by MatFlow respectively for target extinction coefficients 150,000 $M^{-1}cm^{-1}$ and 200,000 $M^{-1}cm^{-1}$, dyes D_7 – D_{10} scoring the highest predicted extinction coefficients and SCScores are shown in Table III.

Among the predicted dye candidates, several noteworthy observations emerged. Structurally identical dyes D_7 and D_{10} were independently predicted under both target extinction coefficient categories. Interestingly, the median predicted extinction coefficients for the two target classes – 150,000 and 200,000 $M^{-1}cm^{-1}$ – were 16,230.09 and 88,297.20 $M^{-1}cm^{-1}$ respectively, reflecting the model's capacity to explore a range of absorption properties. Corresponding SCScores of 3.2 and 3.5 suggest moderate synthetic feasibility.

Dyes D_8 and D_{11} stand out for their very low SCScores (indicative of high synthesizability), making them attractive from a synthetic standpoint. However, their low predicted extinction coefficients place them below the desired threshold, rendering them suboptimal for applications demanding high optical density.

Conversely, D_7 (or equivalently D_{10}) emerges as a compelling candidate. Despite a slightly elevated SCScore of 4.1, it boasts an exceptionally high predicted molar extinction coefficient of 468,121 $M^{-1}cm^{-1}$, making it a strong contender for experimental synthesis in high-performance dye applications.

For applications where a target ε value near 150,000 $M^{-1}cm^{-1}$ is sufficient, D_9 (also listed as D_2 in Table II) offers an excellent balance, with a predicted ε near the design goal and a low SCScore of 3.0.

Ultimately, the standout candidate is D_{12}. With a remarkably low SCScore of 2.4 and a predicted ε of 442,541.93 $M^{-1}cm^{-1}$, it combines high synthesizability with outstanding optical performance, making it the most promising molecule for laboratory validation among those studied.

VI. Conclusion and Future Works

In this study, we developed and validated a graph-based CVAE for the inverse design of molecules with targeted molar extinction coefficients (ε). Our model successfully generated novel compounds clustered around user-defined ε values, as confirmed through comparison with the IUPAC database, DFT calculations, and synthesis feasibility assessment via ASKCOS. Additionally, a separate GNN model provided further validation of predicted optical properties. Importantly, the model demonstrated generalizability beyond ε, showing potential to design molecules for other target properties as well. This framework represents a significant step toward accelerated, property-driven discovery of functional molecules such as quantum dyes, surpassing the limitations of traditional screening methods.

Future work will focus on scaling the model to support larger molecular graphs (>400 nodes), refining CVAE architecture and hyperparameters, and enhancing molecular representations to optimize for multi-objective property profiles. Experimental validation will ensure practical chemical applicability. We also envision building a web-based e-Lab platform where materials scientists can upload custom datasets, define design goals, and interactively generate candidate molecules using trained inverse design models – bringing AI-assisted molecular discovery closer to real-world deployment.

Acknowledgement

This Research was supported in part by a National Institutes of Health IDeA grant P20GM103408, National Science Foundation CSSI grants OAC 2410667 and OAC 2410668, and a US Department of Energy grant DE-0011014.

References

[1] M. A. F. Afzal, M. Haghighatlari, S. P. Ganesh, C. Cheng, and J. Hachmann. Accelerated discovery of high-refractive-index polyimides via first-principles molecular modeling, virtual high-throughput screening, and data mining. *J. Phys. Chem. C*, 123:14610–14618, 2019.

[2] J.-i. Aihara. Reduced homo- lumo gap as an index of kinetic stability for polycyclic aromatic hydrocarbons. *The Journal of Physical Chemistry A*, 103(37):7487–7495, 1999.

[3] Z. Almi, S. Belaidi, T. Lanez, and N. Tchouar. Structure activity relationships, qsar modeling and drug-like calculations of tp inhibition of 1,3,4-oxadiazoline-2-thione derivatives. *Int. Lett. Chem., Phys. Astron.*, 37:113–124, 2014.

[4] L. Arnaut and H. Burrows. *Chemical kinetics: from molecular structure to chemical reactivity*. Elsevier, 2006.

[5] A. Aspuru-Guzik, R. Lindh, and M. Reiher. The matter simulation (r)evolution. *Annual Review of Physical Chemistry*, 69(1):497–522, Apr. 2018.

[6] D. Bagayoko. Understanding density functional theory (dft) and completing it in practice. *AIP Advances*, 4(12), 2014.

[7] K. Burke. Perspective on density functional theory. *The Journal of Chemical Physics*, 136(15):150901, Apr. 2012.

[8] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, Jul. 2018.

[9] L. Chen, W. Zhang, Z. Nie, S. Li, and F. Pan. Generative models for inverse design of inorganic solid materials. *J. Mater. Inform.*, 1:4, 2021.

[10] C. W. Coley, N. S. Eyke, and K. F. Jensen. Autonomous discovery in the chemical sciences part i: Progress. *Angewandte Chemie International Edition*, 59(51):22858–22893, Dec. 2020.

[11] J. Comer and K. Tam. *Lipophilicity Profiles: Theory and Measurement*, pages 275–304. 2007.

[12] K. Deb. An introduction to genetic algorithms. *Sadhana*, 24:293–315, 1999.

[13] D. C. Elton, Z. Boukouvalas, M. D. Fuge, and P. W. Chung. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering*, 4(4):828–849, Aug. 2019.

[14] S. R. Forrest. The path to ubiquitous and low-cost organic electronic appliances on plastic. *Nature*, 428(6986):911–918, Apr. 2004.

[15] R. H. Friend, R. W. Gymer, A. B. Holmes, J. H. Burroughes, R. N. Marks, C. Taliani, et al. Electroluminescence in conjugated polymers. *Nature*, 397(6715):121–128, Jan. 1999.

[16] Y. Geng, G. van Anders, and S. C. Glotzer. Predicting colloidal crystals from shapes via inverse design and machine learning, 2018.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014. NIPS 2014.

[18] J. Guo, B. Ranković, and P. Schwaller. Bayesian optimization for chemical reactions. *Chimia*, 77(1/2):31–38, 2023.

[19] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, S. Ekins, et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, Feb. 2018.

[20] E. Harper and M. Mills. Bayesian optimization of neural networks for the inverse design of all-dielectric metasurfaces. In *SPIE*, volume 11469, Bellingham, WA, USA, 2020.

[21] H. M. Jamil, L. Li, and A. Mirkouei. Matflow: A system for knowledge-based novel materials design using machine learning. In S. Tsumoto, Y. Ohsawa, L. Chen, D. V. den Poel, X. Hu, Y. Motomura, T. Takagi, L. Wu, Y. Xie, A. Abe, and V. Raghavan, editors, *IEEE International Conference on Big Data, Big Data 2022, Osaka, Japan, December 17-20, 2022*, pages 3423–3431. IEEE, 2022.

[22] J. Jang, G. H. Gu, J. Noh, J. Kim, and Y. Jung. Structure-based synthesizability prediction of crystals using partially supervised learning. *J. Am. Chem. Soc.*, 142:18836–18843, 2020.

[23] M. R. Khadilkar, S. Paradiso, K. T. Delaney, and G. H. Fredrickson. Inverse design of bulk morphologies in multiblock polymers using particle swarm optimization. *Macromolecules*, 50(17):6702–6709, 2017.

[24] K. Kim, S. Kang, J. Yoo, Y. Kwon, Y. Nam, D. Lee, I. Kim, Y.-S. Choi, S. Kim, et al. Deep-learning-based inverse design model for intelligent discovery of organic molecules. *npj Comput. Mater.*, 4:4, 2018.

[25] D. P. Kingma and M. Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013. Appeared at ICLR 2014.

[26] W. Koch and M. C. Holthausen. *A Chemist's Guide to Density Functional Theory (2nd ed.)*. Wiley-VCH, 2001.

[27] C. Kuhn and D. N. Beratan. Inverse strategies for molecular design. *J. Phys. Chem.*, 100:10595–10599, 1996.

[28] J. R. Lakowicz. *Principles of Fluorescence Spectroscopy (3rd ed.)*. Springer, 2006.

[29] T. Le, V. C. Epa, F. R. Burden, and D. A. Winkler. Quantitative structure–property relationship modeling of diverse materials properties. *Chemical reviews*, 112(5):2889–2919, 2012.

[30] Y. Lee, G. Choi, M. Yoon, and C. Kim. Genetic algorithm for constrained molecular inverse design, 2021.

[31] A. V. Luzanov. The structure of the electronic excitation of molecules in quantum-chemicalmodels. *Russian Chemical Reviews*, 49(11):1033, 1980.

[32] W. Ma, F. Cheng, Y. Xu, Q. Wen, and Y. Liu. Probabilistic representation and inverse design of metamaterials

based on a deep generative model with semi-supervised learning strategy. *Adv. Mater.*, 31:e1901111, 2019.

[33] J. Mei, N. L. C. Leung, R. T. K. Kwok, J. W. Y. Lam, and B. Z. Tang. Aggregation-induced emission: The whole is more brilliant than the parts. *Advanced Materials*, 27(35):5400–5400, Sep. 2015.

[34] P. G. Polishchuk, T. I. Madzhidov, and A. Varnek. Estimation of the size of drug-like chemical space based on gdb-17 data. *Journal of Computer-Aided Molecular Design*, 27(8):675–679, Aug. 2013.

[35] M. Popova, O. Isayev, and A. Tropsha. Deep reinforcement learning for de novo drug design. *Sci. Adv.*, 4, 2018.

[36] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vettering. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK, 2002.

[37] H. H. Rahman, J. Flores, L. Spear, L. Li, and H. M. Jamil. Potency of latent spaces in inverse quantum dye design. In S. Byna, A. Kougkas, S. Neuwirth, V. Vishwanath, J. Boukhobza, A. Cuzzocrea, D. Dai, and J. Bez, editors, *Proceedings of the 37th International Conference on Scalable Scientific Data Management, SSDBM 2025, Columbus, OH, USA, June 23-25, 2025*, pages 21:1–21:7. ACM, 2025.

[38] H. H. Rahman and H. M. Jamil. Toward knowledge engineering using matflow for inverse quantum dye design. In *IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application, HPCC/DSS/SmartCity/DependSys 2023, Melbourne, Australia, December 17-21, 2023*, pages 852–859. IEEE, 2023.

[39] RDKit. Rdkit. https://www.rdkit.org/, 2022. Accessed June 02, 2022.

[40] U. Resch-Genger, M. Grabolle, S. Cavaliere-Jaricot, R. Nitschke, and T. Nann. Quantum dots versus organic dyes as fluorescent labels. *Nature methods*, 5(9):763–775, 2008.

[41] Y. Ruiz-Morales. Homo- lumo gap as an index of molecular size and structure for polycyclic aromatic hydrocarbons (pahs) and asphaltenes: A theoretical study. i. *The Journal of Physical Chemistry A*, 106(46):11283–11308, 2002.

[42] B. Sanchez-Lengeling and A. Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, Jul. 2018.

[43] K. Sattari, Y. Xie, and J. Lin. Data-driven algorithms for inverse design of polymers. *Soft Matter*, 17:7607–7622, 2021.

[44] J. A. Scales, M. L. Smith, and T. L. Fischer. Global optimization methods for multimodal inverse problems. *J. Comput. Phys.*, 103:258–268, 1992.

[45] G. Schneider. Automating drug discovery. *Nature Reviews Drug Discovery*, 17(2):97–113, Feb. 2018.

[46] M. D. Segall. Multi-parameter optimization: identifying high quality compounds. *Drug Discovery Today: Technologies*, 9(1):e39–e45, Mar. 2012.

[47] T. M. Swager. The molecular wire approach to sensory signal amplification. *Accounts of Chemical Research*, 31(4):201–207, Apr. 1998.

[48] Z. Tu, S. J. Choure, M. H. Fong, J. Roh, I. Levin, K. Yu, J. F. Joung, N. Morgan, S. Li, X. Sun, H. Lin, M. Murnin, J. P. Liles, T. J. Struble, M. E. Fortunato, M. Liu, W. H. G. Jr., K. F. Jensen, and C. W. Coley. ASKCOS: an open source software suite for synthesis planning. *CoRR*, abs/2501.01835, 2025.

[49] W. P. Walters, M. T. Stahl, and M. A. Murcko. Virtual screening–an overview. *Drug discovery today*, 3(4):160–178, Apr. 1998.

[50] D. Wang, D. Tan, and L. Liu. Particle swarm optimization algorithm: an overview. *Soft computing*, 22(2):387–408, 2018.

[51] F. Würthner, T. E. Kaiser, and C. R. Saha-Möller. J-aggregates: from serendipitous discovery to supramolecular engineering of functional dye materials. *Angewandte Chemie International Edition*, 50(15):3376–3410, Mar. 2011.

[52] L. Yuan, W. Lin, K. Zheng, L. He, and W. Huang. Far-red to near-infrared analyte-responsive fluorescent probes based on organic fluorophore platforms for fluorescence imaging. *Chemical Society Reviews*, 42(2):622–661, Jan. 2013.

[53] A. Zunger. Inverse design in search of materials with target functionalities. *Nat. Rev. Chem.*, 2:0121, 2018.